*remote sensing*

MDPI

*Article*

# Feasibility of Early Yield Prediction per Coffee Tree Based on Multispectral Aerial Imagery: Case of Arabica Coffee Crops in Cauca-Colombia

**Julian Bolaños *, Juan Carlos Corrales and Liseth Viviana Campo**

Telematics Engineering Group, University of Cauca, Street 5, No. 4-70, Popayan 190003, Colombia
* Correspondence: julianbolanos@unicauca.edu.co

**Abstract:** Crop yield is an important factor for evaluating production processes and determining the profitability of growing coffee. Frequently, the total number of coffee beans per area unit is estimated manually by physically counting the coffee cherries, the branches, or the flowers. However, estimating yield requires an investment in time and work, so it is not usual for small producers. This paper studies a non-intrusive and attainable alternative to predicting coffee crop yield through multispectral aerial images. The proposal is designed for small low-tech producers monitored by capturing aerial photos with a MapIR camera on an unmanned aerial vehicle. This research shows how to predict yields in the early stages of the coffee tree productive cycle, such as at flowering by using aerial imagery. Physical and spectral descriptors were evaluated as predictors for yield prediction models. The results showed correlations between the selected predictors and 370 yield samples of a Colombian Arabica coffee crop. The coffee tree volume, the Normalized Difference Vegetation Index (NDVI), and the Coffee Ripeness Index (CRI) showed the highest values with 71%, 55%, and 63%, respectively. Further, these predictors were used as the inputs for regression models to analyze their precision in predicting coffee crop yield. The validation stage concluded that Linear Regression and Stochastic Descending Gradient Regression were better models with determination coefficient values of 56% and 55%, respectively, which are promising for predicting yield.

**Keywords:** crop yield; coffee; image segmentation; multispectral; MapIR; predictor; UAV



**Citation:** Bolaños, J.; Corrales, J.C.; Campo, L.V. Feasibility of Early Yield Prediction per Coffee Tree Based on Multispectral Aerial Imagery: Case of Arabica Coffee Crops in Cauca-Colombia. *Remote Sens.* **2023**, *15*, 282. https://doi.org/10.3390/rs15010282

Academic Editors: Jianxi Huang and Javier J Cancela

Received: 1 October 2022
Revised: 28 October 2022
Accepted: 17 November 2022
Published: 3 January 2023



## 1. Introduction

Coffee is one of the most important products for the Colombian economy and represents a source of income for 540,000 medium and small producers' families. According to the Colombian Ministry of Finance, it is an important export product for economic recovery thanks to its international value. Coffee production in Colombia was 12.1 million 60 kg bags in 2021, falling by 11% compared to 2020. This diminution was due to unfavorable weather conditions for coffee crops. For this reason, it is of particular interest to focus efforts on investigating the coffee production process that allows for the optimization of production [1]. The crop yield is quantified by the number of coffee grains per unit area as a representative measure of productivity [2].

Early yield prediction can anticipate the nutritional requirements of the trees, and can also allow for optimizing irrigation and fertilizer use, improving production quality at a lower cost. To predict yield early, it is important to consider phenological cycles and production stages to support the farmers' decision-making as set out in [3]. The phenological cycle starts at the flowering stage [4]. The subsequent process is the filling of the coffee fruits, and it is fundamental to foresee the demanding nutritional requirements of the tree, before the maturation stage and picking. Furthermore, understanding crop yields allows support for decision-making in the other processes, such as picking, drying, and storage [4,5]. Crop yield is directly determined by the dynamics of the tree–soil–environment system, as defined by specific predictors in each element of the system. For example, temperature,

water availability, and precipitation are predictors that correspond to the environment [6]. Other variables may be edaphic predictors, such as the soil type and its composition, and tree characteristics, such as its age, variety, planting density, health, and physiology, which directly influence crop yield.

Predictors can be collected manually or by proximity sensors, satellites, cameras, or weather stations. Currently, the use of UAVs has countless advantages in precision agriculture. By capturing images, it is possible to obtain high-resolution data without the influence of atmospheric conditions, making this technology accessible to small and medium size growers.

One of the applications of aerial images is to obtain physical and spectral characteristics; for example, the work of dos Santos et al. [7] calculated the height and diameter variables from aerial images. The objective of this research was to use DSM from UAVs carrying conventional RGB cameras. That work [7] demonstrated the feasibility of obtaining the height and the diameter of a coffee tree with a correlation of 95% for diameter and 85% for height.

In our research of the literature, the work of Idol et al. in [8] exemplified the prediction of crop yields through manually obtaining information. This process estimates the number of nodes and fruits on all the visible sides of the tree. In the same way, Castro et al. in [9] studied collecting manual yield samples, showing the relationship between the lateral yield and total tree yield. Both works show the efforts in making yield estimates. However, neither does it in a non-early way. Similarly, Unigarro et al. [10] compared the phenotypic characteristics of coffee trees to yield, concluding that the leaf area is a determining predictor. The above methods detail manual data collection methodology. These models have a precision greater than 90% but the data collection is very difficult and is highly invasive. They are both costly and time-consuming, and do not allow early prediction of crop yields [8,9].

Works such as [11,12] that involved agrometeorological models likewise used satellite images and environmental variables at a regional level, but was not suitable for tree-centered analysis. For example, Picini et al. [11] studied a model to estimate potential coffee production based on the evapotranspiration of the planting and the previous year's production; this work obtained a $R^2$ of 0.9 in a deterministic model involving the relationship between potential yield and expected yield. However, Rosa et al. [12] obtained a non-conclusive result using the NDVI and its relationship with the LAI.

Barbosa et al. proposed another essential approach in [13], studying regression models on the basis of physiological characteristics, such as the height and diameter obtained from RGB images using UAVs for yield estimation, manually validating the results with the image. In this research, the total beans of the coffee tree were obtained by georeferencing each point with the GCP tool. The training process of regression models, such as SVM, PLS, gradient boosting, and RF, used 144 data records. The calculation of the height and diameter varied between 6% and 7%, and the MAPE measurement for the regression models was around 31%. This work did not consider the spectral analysis based on vegetation indices.

Kouadio et al. [14] has a relevant approach, where a machine learning model based on three different algorithms, EML, RF, and LR, was proposed. All of these models were trained using the soil nutrient characteristics which allowed yield calculations. It is important to note that the best performance model for yield calculation was EML, using organic material, phosphorus and sulphur predictors. This was validated by RMSE with ±13.6% and a MAE with ±7.9%. Similarly, Nguyen et al. [15] proposed a statistical model for the early prediction of coffee crop yields based on vegetation indexes at a regional level with Copernicus data for the NDVI, FAPAR, and LAI predictors, obtaining an Adj $R^2$ from 64 to 69% in regression models using the Crop Growth Monitoring System statistical tool, which allows early prediction of up to 6 months before the harvest.

Considering these previous works, this research focuses on obtaining reliable data to demonstrate the feasibility of early coffee yield prediction using low-cost tools, facilitating higher technology access to small and medium producers. The aerial images obtained were

processed and segmented to find the spectral and physical characteristics for each coffee tree. Manual methodology was used to collect yield data. This consisted of picking all the coffee cherries on one tree. Finally, a linear correlation analysis was performed, obtaining significant results, such as a 70% correlation between the tree volume and the crop yields. With these results, a regression model training process was carried out. It obtained a $R^2$ score of approximately 56% for some models, such as linear regression. The $R^2$ score or the determination coefficient defines the quality or the adjustment of a model using the percentage variance of a variable that is explained by another according to [16].

This work involved the following process: 1. obtaining multi-spectral aerial images during the flowering stage; 2. defining the post-processing process for the segmentation and individualization of the trees; 3. building a manual data collection interface; 4. the manual data collection; 5. the manual data analysis to calculate the yield by tree; 6. exploiting the regression-based prediction models; 7. analysis of the results.

## 2. Materials and Methods

Section 2 describes the image collection process and the task of obtaining manual coffee crop yield training information. This section begins with a description of the study area, the tools, the image processing, and the segmentation of the trees, and ends with the analysis of the predictors for the regression models.

### *2.1. Study Area*

This experiment was carried out on the "La Sultana" estate, which is a farm at the Universidad del Cauca located in the municipality of Timbio, Cauca, Colombia (Figure 1) (2°2′28.51″ N, 76°43′31.89″O) altitude: 1850 m.s.n.m.

Since 2006, La Sultana has carried out sustainable coffee production through its ecological processing and management [17]. In addition, good practices produce high-quality coffee with environmental, social, and economic sustainability.

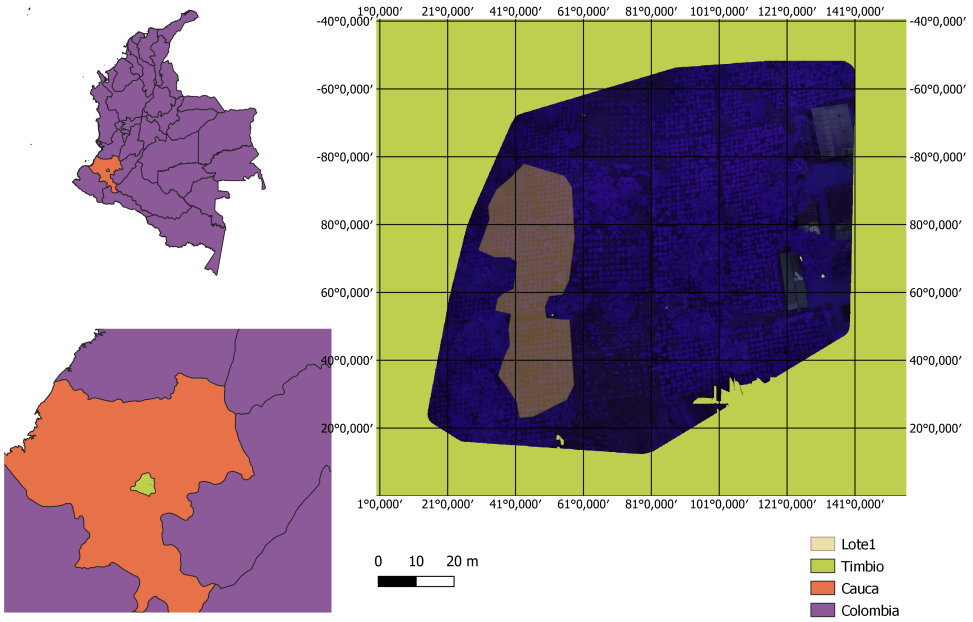


**Figure 1.** The map on the left shows the distribution by departments of Colombia, in which Cauca is located. On the left is the geolocated orthophoto of "La Sultana" farm.

On the other hand, the image taking was carried out during the flowering stage from July to August 2021. Manual yield samples were taken from three coffee plantations of the Castillo variety between November and December 2021.

### *2.2. Manual Yield Sampling of Coffee Trees*

The objective to obtain manual variables is to determine the real yield of a tree. The non-invasive data collection process in this paper was based on the work of Idol et al. [8]. It begins counting plagiotropic branches and then performs a sampling of nodes and beans in order to estimate the total yield per tree.

Manual data collection is a costly, complex, and slow process. This work does not have per tree information that would allow training the prediction models. Because of this, a web tool was implemented to expedite the process of obtaining the crop yield samples. The individualization of trees in the web application streamlines manual sampling. This process is explained in Section 2.4. Figure 2 shows the results of the massive segmentation carried out by the web tool developed to obtain manual variables. All of the tree information is stored in a Mysql relational database.

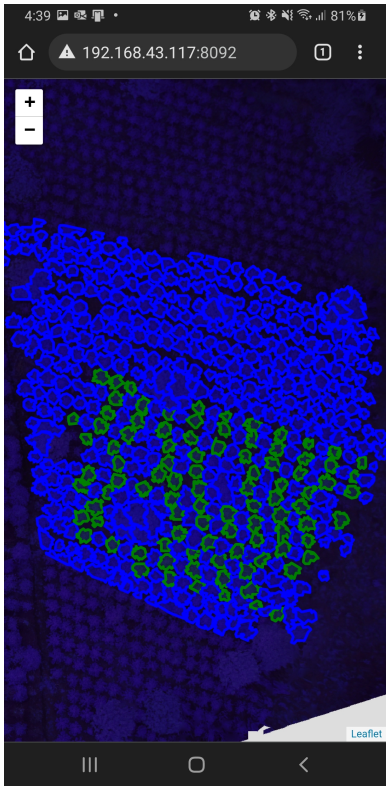


**Figure 2.** Web service interface for manual sampling, in green the plants with complete information, in blue the plants to be completed.

Figure 3 shows a representation of the client–server architecture of the web application. The interface allows us to save the information automatically to avoid any data loss. For the dynamic, georeferenced and agile load of the orthophotos, this research integrated the WEBodm mosaic service. This service, along with the GEOJson representation of tree edges, allows us to use the LeafletJS library that incorporates all the information on the same map [18]. The result is an interactive experience that is easy to carry out by people with no advanced knowledge of the sample collection process.

As shown in Figure 4, the height and diameter of the trees are required in the interface for comparison purposes. Here, the trees were divided into two independent samples, the header and the footer. Figure 5 shows the interface where the number of branches, nodes and cherries is recorded. Based on a statistical calculation of the sample size, this research sampled nodes and cherries with an 8% error and a 80% confidence level. The standard deviation was fixed at 0.5 [19].

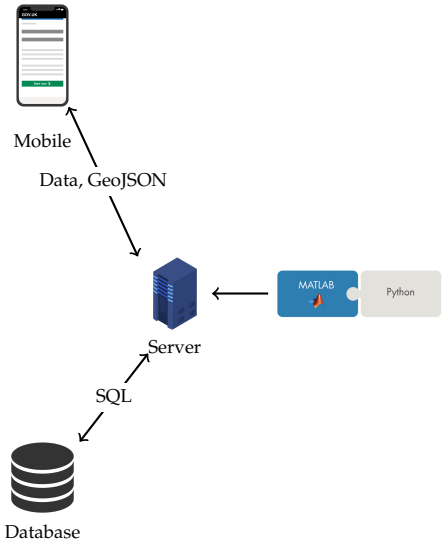


**Figure 3.** Software architecture for manual collection of crop yield samples.

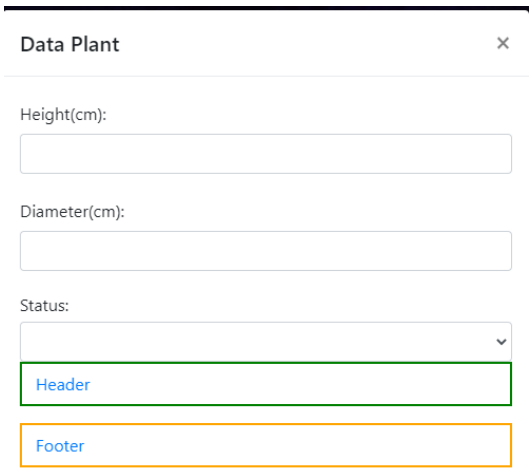


**Figure 4.** Height and diameter manual sampling interface.

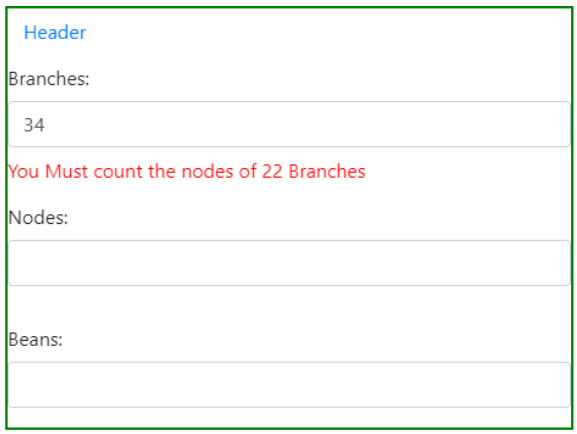


**Figure 5.** Header and footer tree sampling interface.

The manual sampling process is expensive, so the data volume cannot be high. In this work, it was possible to take 370 real data samples verified in 3 different fields randomly defined by the coffee pickers. The age of the trees in the fields were 2, 3, and 7 years, respectively.

### *2.3. Platform to Acquire Images*

The UAV platform used in this work was a Phantom 3 Standard four-bladed drone. Mission Planner was used for flight plans, the missions were run using the Litchi app The operating speed was 5 km/h with an 18-meter altitude. This research established longitudinal and lateral overlapping of 80%, capturing the images between 10:00 a.m. and 1:00 p.m. Colombian local time under a cloudy sky. The aerial image capture was carried out between July and August 2021 to perform the early yield predictions.

The images were obtained using a MAPIR Survey 3W camera. The modified RGN camera was equipped with a Sony Exmor R IMX117 sensor that has a 12-megapixel resolution using the default parameters recommended by Mapir [20]. All the configurations were the same for all flights. Finally, the resulting images were saved in the JPG and RAW format for subsequent processing.

Cloudy sky conditions were chosen to avoid saturating the camera's NGR channels. The NIR reflectance affects the red channel as established in [21]. Images were taken under cloudy and clear sky conditions to determine the best NDVI results, as shown in Figures 6 and 7. In the sunny image, the shadows directly influence the NDVI index behavior.

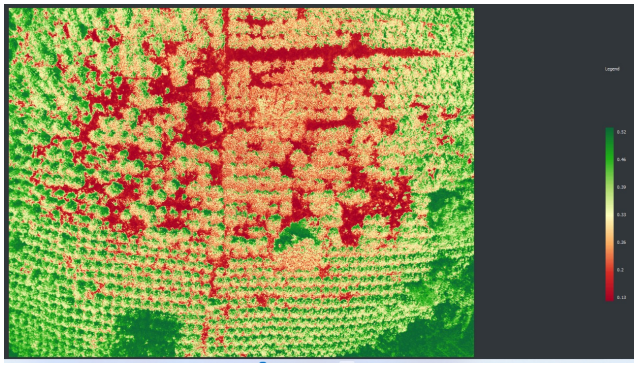

**Figure 6.** Sunny NDVI image map.

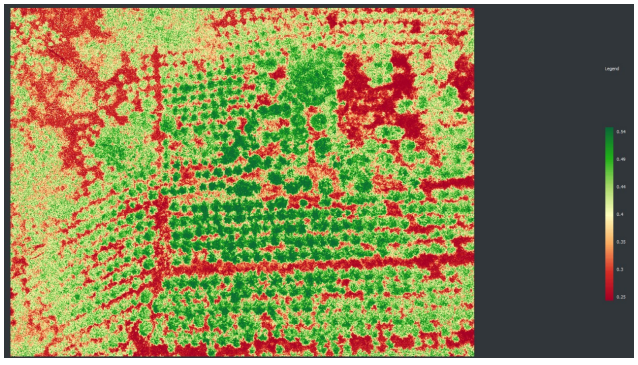

**Figure 7.** Cloudy NDVI image map.

This research pre-processed the images collected using the Mapir Camera Control for the radiometric calibration [22]. The GPS was set up with incorrect height tags; the ExifTool solved this. Later, the image mosaics were made using the WEBOdm tool generating the DSM and the orthophotos [23].

### *2.4. Processing Images*

To identify each single coffee tree, a watershed method based on the foreground markers was used. This algorithm separates and defines the edges between two elements of an image. It consists of simulating the watershed basins, which are filled based on local minimum levels until a limit between two or more watersheds is defined [24]. In Figure 8, a representation of this process is described.

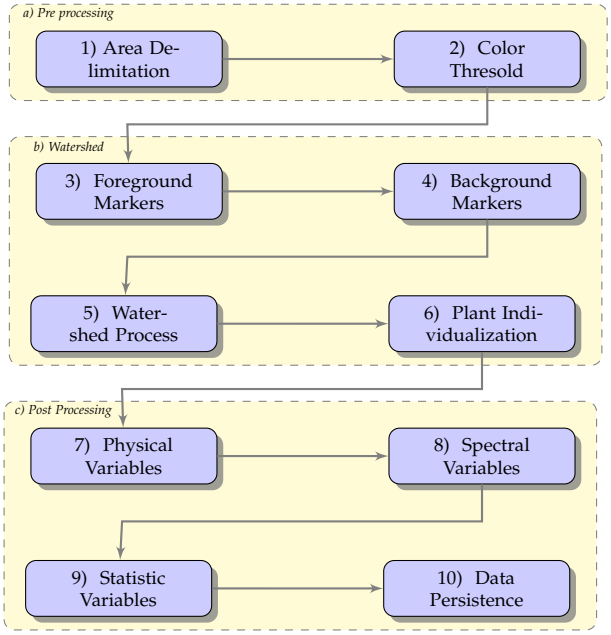


**Figure 8.** Tree individualization process: it involves the pre-processing of the image, the application of the watershed algorithm for segmentation, and the obtaining of variables and characteristics.

In the preprocessing stage, the target coffee field is delimited and segmented. This research carries out this process through a previously delimited polygonal mask. In step two, the background or the image floor is removed by a color threshold based on the LAB color space or by using 3 groups with the K-means algorithm. Both background removal methods were tested and the color threshold prevailed.

One of the ways to apply the Watershed algorithm is to use the foreground and background markers [25]. This work carried out multiple tests to find the correct configuration of the foreground markers, which led to a successive application of morphological operations to define them. The final result of the foreground marker definition is superimposed in blue in Figure 9.

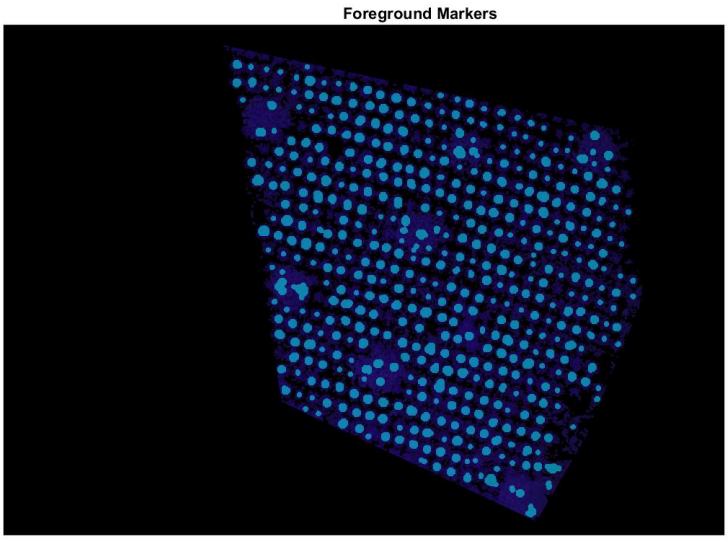


**Figure 9.** Foreground markers: in blue the coffee plants detected.

The background markers define what is not a tree. This process allows the algorithm to establish the limits based on a magnitude gradient. The background markers are shown in Figure 10.

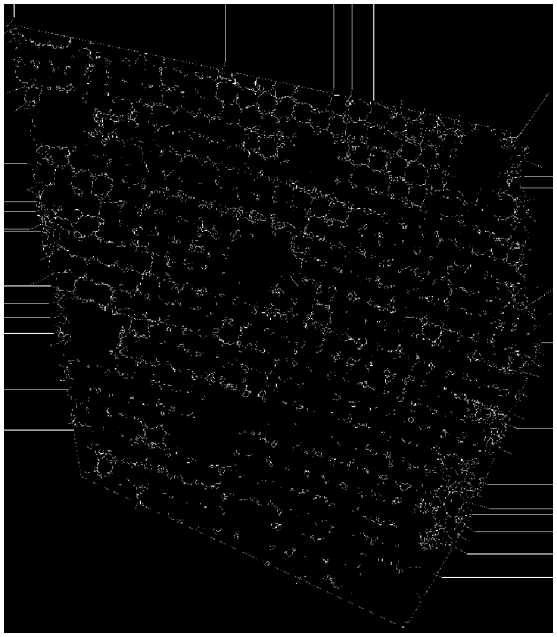

**Figure 10.** Background markers.

The marker-based watershed algorithm resulted in Figure 11. In this case, the edges of the limits of each individual tree were defined, accompanied by each foreground marker. These results were highlighted by applying a color map to each label to observe the segmentation quality, as in Figure 12.

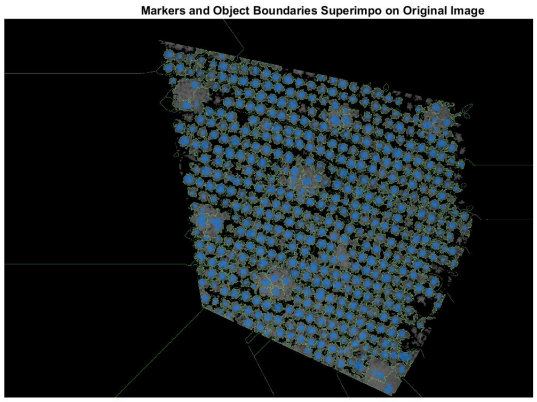


**Figure 11.** Watershed labeling.

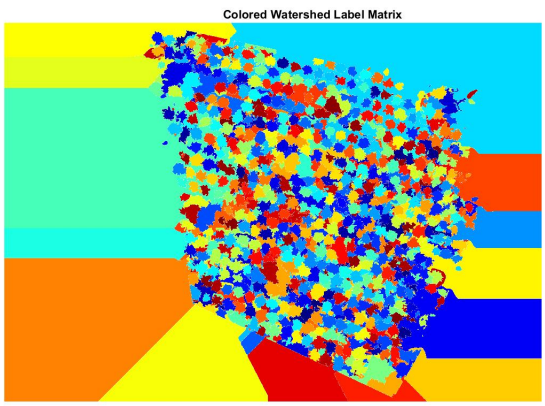


**Figure 12.** Watershed results: display all detected labels in a different color by applying a LUT color workspace.

Once the individualization of the trees was complete, the next step was the information extraction process. This process begins with the noise analysis to review the size of each tree by previously determined values of the lower and higher limits. The outline of the mask of Figure 13 is processed by changing the (x, y) points in their pairs (lat, long). This process is focused on calculating the WGS84 coordinates using the georeferenced information through a Python language process. With this information it is possible to obtain physical measurements of the coffee trees.

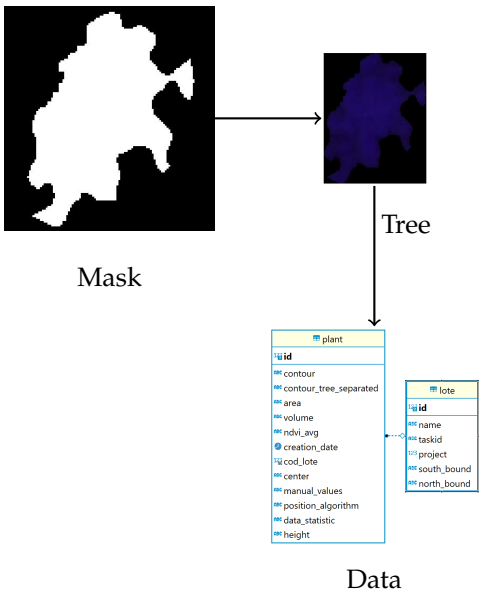


**Figure 13.** Extraction variables flow: the mask corresponds to the detection of the tree. The mask is applied to the RGB three-band image to extract the data and store it in the Mysql relational database.

The separate tree interest bands were defined by applying the figure mask Figure 13 to the original image. This process allows us to calculate statistical variables, such as the average of each band and their variance and the vegetation indices of Table 1, which were calculated by going over and operating each pixel of the image of the three bands of the individual tree. Each tree was processed so that it was possible to obtain its outline as geographical coordinates WGS84. With this GEOJson format contour, the area in square meters can be obtained using the python "area" package.

In short, this whole process is necessary to obtain the spectral and physical descriptors used to train and test the prediction models. The variables received were the area, height, and vegetation indices and the statistical variables of the spectral bands.

### *2.5. Physical Descriptors*

The physical descriptors of the coffee trees are characteristic measures that can be height or area. This work uses the georeferenced orthophoto and the DSM to generate this information. The tree height is measured by the subtracting of the minimum value of the maximum calculated by the geographical coordinates in the DSM. This work built a method that iterates through tree image pixels. This method shows the heights present in the DSM every 10 pixels in the individualized tree image. The pixel jump values were tested incrementally to increase speed without losing any precision; it is not necessary to transform all of the pixels of the tree mask. The tree image is dilated with a radius of 20 to include part of the ground. Finally, all the points are compared to extract the maximum and the minimum, and then the difference. The volume, lateral area, and LAI values were calculated by taking into account the work of Favarin et al. [26]; these values were obtained from Equations (1), (3) and (4), respectively:

$$volume = area * \frac{h}{200} * \frac{4}{3} \quad (1)$$

$$diameter = 2 * \sqrt{(\frac{area}{\pi})} \tag{2}$$

$$Al = \pi * \frac{diameter}{4} * \sqrt{(4 * h^2 + diameter^2)} \tag{3}$$

$$laiArea = -0.5786 + 0.7896 Al \tag{4}$$

The *area* and height ($h$) are previously known image values. The *laiArea* value corresponds to the leaf area index obtained from the lateral area based on the work of Favarin et al. [26], who successfully proposed a linear approximation to this expected value.

### 2.6. Spectral Descriptors

The spectral descriptors correspond to those obtained through aerial images, which will be the input for the prediction models. These were obtained considering the segmentation of Section 2.4. The vegetation indices were processed for each tree, obtaining medium, maximum, and minimum values.

Normally, the spectral reflectance of a tree changes according to the wavelength and its physiological state [22]. In this work, the value of the vegetation indices and the maximum, minimum, and average values of the red bands, green, NIR, and histogram were collected.

By using vegetation indices, different physical conditions can be inferred. Trees mostly reflect the NIR band and absorb the red band when they are in good condition, representing the physiological state. This condition is shown using the normalized difference vegetation index NDVI [12]. Other important indices are the NDWI, which can provide information on tree moisture, or the visible excess of green band EXG, which can be used for tree segmentation, as well as for improved NDVI versions, such as ENDVI, which is also used to minimize the effect of the ground in the final result [27].

Table 1 presents the summary of the vegetation indices involved in this work. This research was based on the vegetation indices of the results obtained by Rosas et al. [21], who carried out spectrometric analysis with the Survey3W camera. All the vegetation indices of the Table 1 are available using the RGN bands of the Survey3 camera.

**Table 1.** Vegetation indices.

| Index | Form | Description | Ref |
|---|---|---|---|
| CRI | $\frac{R}{R\_mean} * 100$ | Coffee Ripeness Index | [21] |
| GNDVI | $\frac{NIR-G}{NIR+G}$ | Green Normalized Difference Vegetation Index | [21] |
| MCARI1 | $1.2[2.5(NIR-R)-1.3(NIR-G)]$ | Modified Chlorophyll Absorption in Reflectance Index 1 | [28] |
| MTVI1 | $1.2[1.2(NIR-G)-2.5(R-G)]$ | Modified Triangular Vegetation Index 1 | [28] |
| NGRDI | $\frac{G-R}{G+R}$ | Normalized Green–Red Difference Index | [29] |
| NDVI | $\frac{NIR-R}{NIR+R}$ | Normalized Difference Vegetation Index | [21] |
| RVI | $\frac{R}{NIR}$ | Ratio Vegetation Index | [21] |
| NRVI | $\frac{RVI-1}{RVI+1}$ | Normalized Ratio Vegetation Index | [21] |

### 2.7. Prediction Models

The prediction models can be defined as a representation of the relationship between two or more variables [30]. The main objective of this work is to predict the crop yield at an early stage of the phenological cycle. Since the amount of data is limited, it is impossible to use neural network models, so a simple regression model is proposed for this work.

A cross-validation methodology was carried out to obtain an average result of the $R^2$ metric that measures how good an algorithm is in predicting a variable [31]. This methodology divides the dataset into data groups, performs a cross-validation of all groups and gives each iteration a value of $R^2$. The final result of the process corresponds to the average of all the values.

The tested models are the support vector regression or SVR with linear kernel, multiple and simple linear regression, random forest, and decision trees. All the models were implemented using the Scikit-learn library [30,32]. Cross-validation is a process where the number of iterations to be performed on the data is defined. In this case, a value of fourth data subsets was defined and each of the models was executed with their default settings.

## 3. Analysis and Results

This section analyzes the results obtained by reviewing the variables collected through graphs and correlations. The results obtained by the prediction models were exposed.

### *3.1. Predictors Selection*

With all the information collected, this research carried out a data cleaning, checking the ranges and assuring that all the manual sampling was complete. Initially, taking into account the correlation analysis of Pearson, the behavior of the available characteristics was verified in relation to the crop yield obtained manually in Section 2.2, representing the most relevant results in Figure 14, this analysis allows evaluating the relationship between the predictors and the crop yield [33].

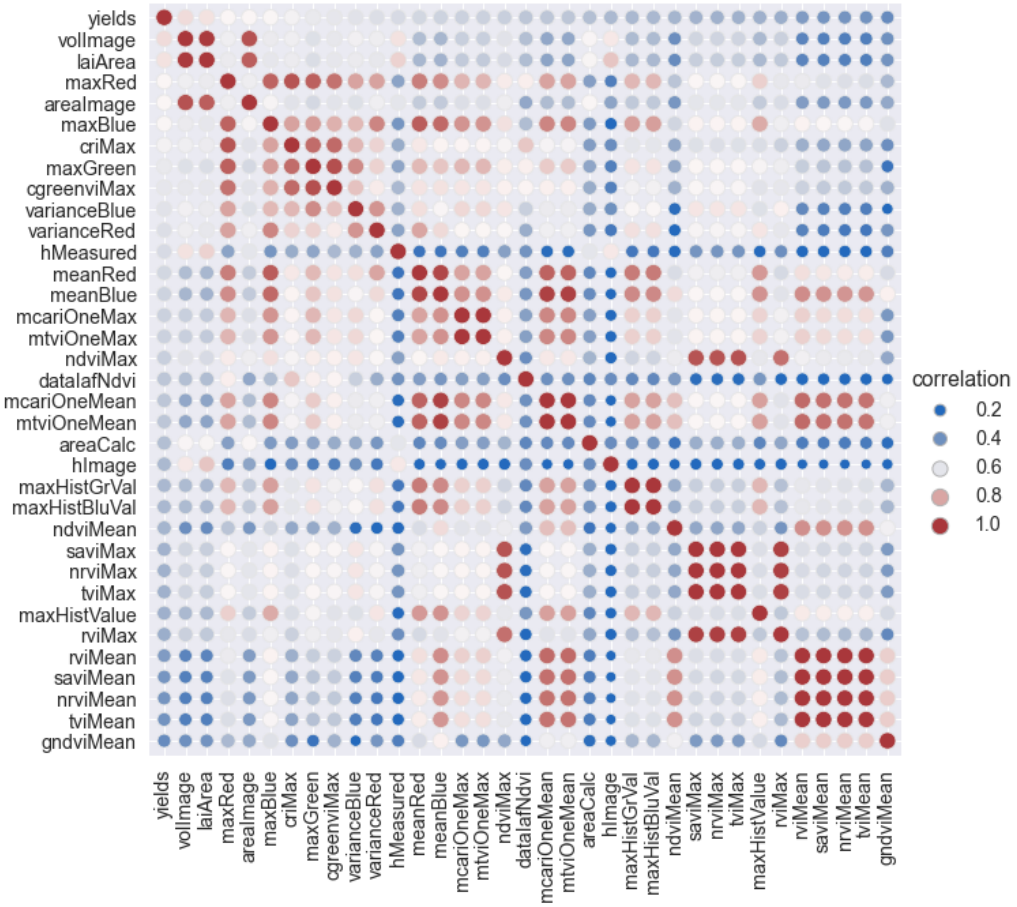


**Figure 14.** Correlation matrix.

The graph also shows a strong relationship between the height measure (hMeasured) and the calculated height (hImage) with the DSM since the correlation index was 75%, which validates the tree height calculation process based on the image information.

The tree volume values, as volImage calculated from the contour areas and heights, such as in Equation (1), gave a correlation of 76%, as well as the lateral area that was calculated with (3) gave 77% and an expected value according to [9] who in his work shows the relationship between physical characteristics, such as LAI which is strongly related to the lateral area and to the volume [26]. Regarding the vegetation indices, this research found different behavior. The NDVIMean had a correlation of 49%, the NDVimin and the NDVimax were −39% and 55%, which shows the importance of the NDVI that, according to [21], represents the greenness and the vigor of the plant [34]. The RVI, SAVI, NRVI, and TVI vegetation indices have a correlation of about 45%, so it is essential to take them into account in analyzing crop yield.

This work measured all of these variables at the flowering stage, which represents the state of the trees before the cherries fill out. The other indices with their median, maximum, and minimum values had slightly higher values but were less than 0.5, probably due to their low physiological relationship with crop yield. By including the maximum and minimum values in the analysis of the vegetation indices, it was possible to find that the NDVI had higher correlation values in the average of the maximums.

According to these correlations, the predictors were prepared in "dataframes" to carry out the regression models. Some of the variables involved were volume, NDVIMax, maxRed, maxBlue, and criMax.

Figure 15 shows a dispersion diagram with yield, volImage, maxRed, and maxBlue, which shows a linear relationship between all of these variables and the crop yield in column one. The tree volume is correlated 77% with the crop yields and has a lower dispersion compared to the other two variables. Max red and max blue versus yield have similar graphics, but with a lower concentration in their higher values. It is important to note that the crop yield values are concentrated in the first quarter of the standard range of 0 to 1, similar to the tree volume values.

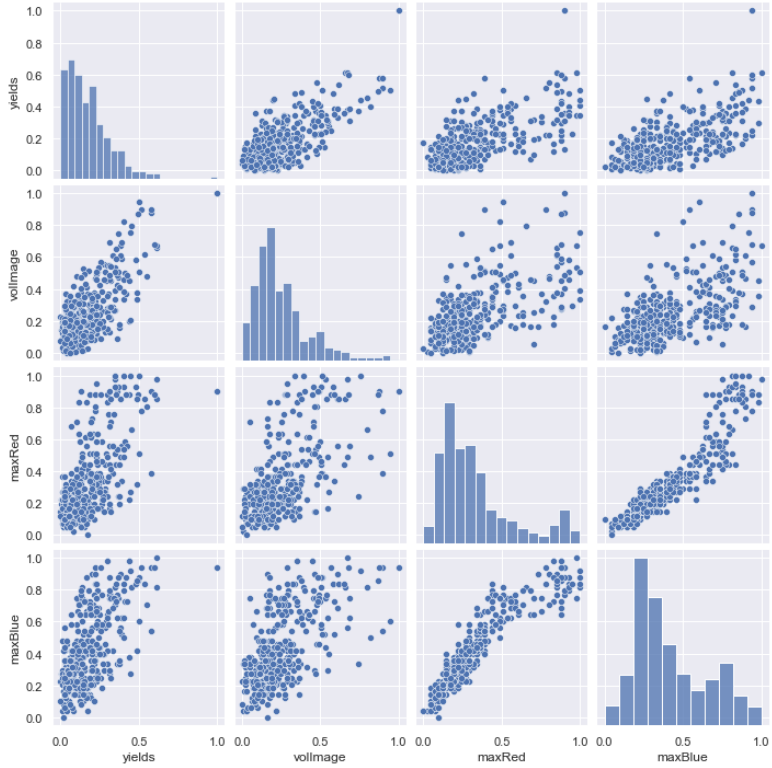


**Figure 15.** Scatter plot of crop yield, volume, maxRed, and maxBlue.

### *3.2. Performance of Prediction Models*

Table 2 presents the variables found by the characteristic selection method which obtained the best determination coefficient $R^2$. In this process, some correlated variables, such as criMax with a correlation of 63%, decreased the behavior of the $R^2$ coefficient, and the variables with a correlation of 45%, such as saviMax, tviMax, and ndviMean, increased the value of $R^2$ for linear regression.

**Table 2.** Variables for the multiple regression process.

| Variable | Resume |
|---|---|
| volImage | Volume obtained from the height and the area |
| maxRed and maxBlue | Maximum value of the red and blue bands of tree |
| varianceBlue | Variance value of Blue band (In this case NIR band) |
| LAILatArea | Leaf Area Index from Equation (4) |
| saviMax | Maximum of Soil Adjusted Vegetation Index |
| tviMax | Maximum of Triangular Vegetation Index |
| mtviOneMean | Mean of Modified triangular vegetation index |
| ndviMax and ndviMean | Mean and Maximum Normalized Difference Vegetation Index |

The prediction model application process is based on the SkLearn library using supervised learning regression models. This research selected the regression models based on the data quantity and the linear correlation of Figure 16.

The model entries were manual crop yield and the predictors of Table 2. This research tested a simple variable model with volume since it had the highest correlation with the crop yield and multiple models with the following variables to compare the simple and multiple regression models. Several iterations were performed using different configurations with the available regression models; for example, for SVR, the core was changed from RBF to linear. For SDG, the loss function was changed between huber, squared_error, or epsilon_insensitive, showing better results with the squared_error settings. For the other models, this research maintained its default settings.

The models were evaluated by cross-validation with a $R^2$ score that allowed us to determine the model adjustment percentage of the data. The predictors and the models with the highest $R^2$ score are shown in Table 3. The volume role in the regression models can be explained by the relationship between the tree architecture and its age [10]. In addition, it can be supplemented by a physiological status indicator, such as the NDVI [13].

The results of Table 3 show that for this process, the method that showed best results was the linear regression method with 57.6% and a 2361 RMSE, followed by Lasso with a $R^2$ of 55% and with 2442 RMS. The SVR model with linear kernel obtained 53%. In this model, when the kernel was changed to rbf, the result decreased to 46%. Since the previous models have a linear base, they adjust better to the data better. Still, the difference between Random Forest applied to the tree volume with 22% and multivariate with 52% is an interesting phenomenon.

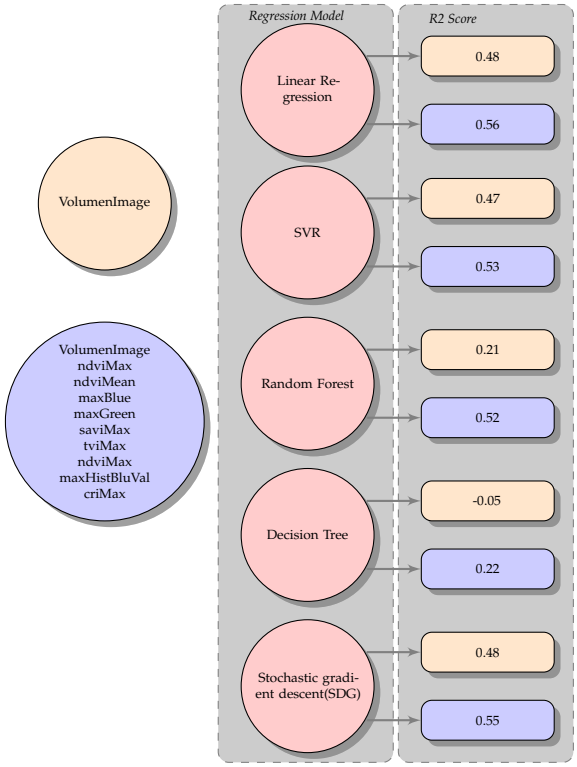


**Figure 16.** Regression models.

**Table 3.** Summary of regression models

| Model | Type | $R^2$ | RMSE |
|---|---|---|---|
| Linear Regression | Simple | 0.48 | 2648 |
| | Multiple | 0.576 | 2361 |
| Lasso | Simple | 0.48 | 2648 |
| | Multiple | 0.55 | 2442 |
| PLSRegression | Simple | 0.48 | 2648 |
| | Multiple | 0.544 | 2462 |
| SDG | Simple | 0.1 | 3474 |
| | Multiple | 0.51 | 2546 |
| SVR | Simple | 0.47 | 2660 |
| | Multiple | 0.53 | 2477 |
| Random Forest | Simple | 0.22 | 3226 |
| | Multiple | 0.50 | 2526 |
| Decision Tree | Simple | −0.05 | 3747 |
| | Multiple | 0.12 | 3333 |

The least suitable model for the data in this scenario was the decision tree. The random forest model is less affected by dispersion, and its performance is closer to the SVR, higher than this model with the RBF type kernel of 48%. Similar to [13], in this work, the height and the area are not relevant separately until they become volume. The two component PLSR model has a coefficient of 0.54 and a 2462 RMSE. The foregoing results define the linear behavior of the selected variables as related to crop yields.

During the cross-validation process, this research tested 10 regression models, and in all cases, the multiple regression model had better results than simple regression.

Table 4 compares the yield obtained by manual yield models. These values were randomly calculated by iterating the division of the training and test values The original

crop yield was presented in Section 2.2. The objective of this proposal was to evaluate the relevance of the coffee yield predictors, showing that the predictor behavior is linear, supported by regression models. In addition, when the predictors are combined, they have behave better when making the predictions.

**Table 4.** Comparison regression models. *Predicted and original values in grains per tree.

| Model | Type | Predicted | Original |
|---|---|---|---|
| Linear Regression | Simple | 5180 | 5144 |
| | Multiple | 5148 | 5144 |
| PLS Regression | Simple | 5159 | 5144 |
| | Multiple | 5106 | 5144 |
| Lasso | Simple | 5180 | 5144 |
| | Multiple | 5149 | 5144 |
| SVR | Simple | 4774 | 5144 |
| | Multiple | 5227 | 5144 |
| SDG | Simple | 5096 | 5144 |
| | Multiple | 5427 | 5144 |
| Random Forest | Simple | 5174 | 5144 |
| | Multiple | 5208 | 5144 |
| Decision Tree | Simple | 5179 | 5144 |
| | Multiple | 5226 | 5144 |

These results validate the approach proposed in this work; however, validations with more data need to be carried out to make the process conclusive.

Figure 17 shows the behavior of the spectral bands as presented by Rosas et al. [21] on the spectral reflectance in plants. This result validates the radiometric correction and shows that the camera reflectance agrees with what was expected, since for the plants, the NIR band has the lowest absorption and the greatest reflectivity [21].

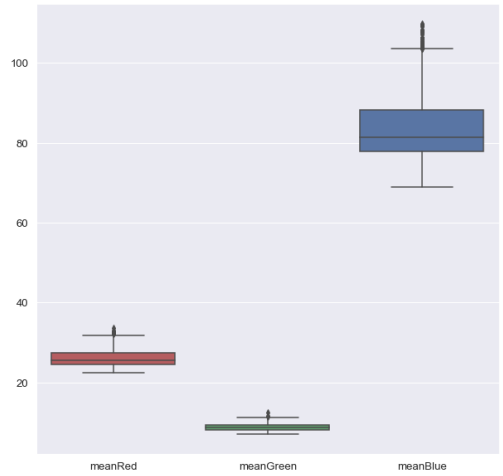


**Figure 17.** Range of means of the blue, green, and red bands.

## 4. Discussion

Coffee cultivation is extremely important for the Colombian economy. To ensure that these crops remain economically profitable, new cultivation techniques need to be adopted to improve the current processes. Several models are dedicated to crop yield estimation through climatic variables, flowering records, and soil factors, among others [4–6,35]. This study is oriented towards early prediction of coffee crop yields by using only multi-spectral image data obtained using low-cost tools. Multi-spectral images obtained from UAVs

allow a high precision, focused analysis which, for this research, led to precise, automatic individualization of coffee trees and generated physical and spectral descriptors with the potential to predict crop yields [28,36,37].

The vegetation indices have the potential to predict yield in other crops. In this sense, the work of Douglas et al. [27]. involves the $EVI_2$, NDWI, and NDVI indices based on MODIS data to use regression models with a $R^2$ of 0.70, 0.69, and 0.69, respectively, for soybeans. It also shows that in corn (maize), the $EVI_2$ obtains a $R^2$ of 0.73. The importance of using multi-spectral images in calculating yield in different crops is highlighted.

This research focused on early yield prediction by obtaining images at the phenological flowering stage. The $R^2$ determination coefficient of 0.54 presented in Table 3 shows a clear potential for calculating yield tree by tree as used by Barbosa et al. [13] but with physical variables and RGB images.

One of the limitations of this research obtaining crop yield data for model training. The manual data collection process is very expensive, both in terms of time and money. The multispectral image approach to facilitate manual collection of crop yield data is one of the contributions of this research. The plant segmentation process of this research defined a sample collection method guided by a web-based application that allows identifying the plants to be sampled.

In future perspectives, this application can be complemented by observation and surveillance. Hyperspectral cameras enable the capturing of many more bands than multi-spectral cameras. However, the multispectral cameras are more accessible for small and medium producers. This research studied arabica coffee varieties, the dominant varieties grown in Colombia. It would be interesting to apply this methodology to other varieties, such as Robusta. This research analyzes the data by assuming a linear relationship between crop yield data and linear predictors. However, additional analysis looking for non-linear behaviors and testing the method on different varieties of coffee of other ages would be desirable.

## 5. Conclusions

The proposed methodology allowed for early prediction of crop yield, which would facilitate decision-making. It analyzes the feasibility of predicting coffee crop yield by using multi-spectral aerial images. It begins with discriminating the conditions for optimal capture of photographs at the flowering stage. Subsequently, this work described image processing for segmentation and individualization of the coffee trees with the watershed algorithm produced the expected results. The development of this work resulted in a strong correlation between physical variables, such as tree volume, and spectral variables, such as maximum NDVI and crop yield, similar to [13]. The segmentation of the plants helps to avoid the influence of the ground and other plants in calculating the vegetation indices.

The individualization of image processing allows for automated classification and labeling of manual yield samples for regression model training. Considering the results of Table 3, it is possible to affirm that the best model for predicting coffee crop yields is linear regression.

For future research, capturing more manual samples in different crops to map the biennial coffee crop behaviors and to open the possibility of using deep learning tools.

**Author Contributions:** Conceptualization, J.B. and L.V.C.; methodology, J.B. and L.V.C.; software, J.B.; validation, J.B. and L.V.C.; formal analysis, J.B. and L.V.C.; investigation, J.B.; resources, J.B.; data curation, J.B.; writing—original draft preparation, J.B.; writing—review and editing, J.B. and L.V.C.; visualization, J.B.; supervision, J.C.C.; project administration, J.C.C.; funding acquisition, J.B., L.V.C. and J.C.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors thank the Ministry of Science, Technology, and Innovation (MINTIC)—Colombia under project "Estimación del rendimiento de un cultivo de café mediante imagenes aéreas tomadas con un uav multirrotor" Convocatoria 823-Formación de capital humano de alto nivel para las regiones-Cauca, Universidad del Cauca, especially the Telematics Engineering Research Group (GIT), and La Sultana's farm.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| NDVI | Normalized Difference Vegetation Index |
| CRI | Coffee Ripeness Index |
| LAI | Leaf Area Index |
| RGB | Red–Green–Blue |
| UAV | Unmanned Aerial Vehicle |
| GCP | Ground Control Position |
| SVM | Support Vector Machines |
| PLS | Partial Least Squares |
| MAPE | Mean or Average of the Absolute Percentage Errors |
| DSM | Digital Surface Model |
| Kc | Coefficient of Crop |
| EML | Extreme Machine Learning |
| SDG | Stochastic Gradient Descendent |
| RF | Random Forest |
| LR | Linear Regression |
| RMSE | Root Mean Square Error |
| MAE | Mean Absolute Error |
| RGN | Red–Green–NIR |
| NDWI | Normalized Difference Water Index |
| ExG | Excess Of Green |
| FAPAR | Fraction of absorbed photosynthetically active radiation |

## References

1. Cafeteros, F. Producción de Café de Colombia Cerróel 2019 en 14.8 Millones de Sacos. Available online: https://federaciondecafeteros.org/wp/listado-noticias/produccion-de-cafe-de-colombia-cerro-el-2019-en-148-millones-de-sacos (accessed on 15 August 2021).
2. Arcila, J.; Farfan, F.; Moreno, A.; Salazar, L.F.; Hincapié, E. Sistemas de Producción de Café en Colombia. Available online: https://biblioteca.cenicafe.org/bitstream/10778/720/1/Sistemas%20producci%C3%B3n%20caf%C3%A9%20Colombia.pdf (accessed on 15 August 2021).
3. Ramirez, V. *La fenología del Café una Herramienta para Apoyar la Toma de Decisiones*; Technical Report; Centro Nacional de Investigaciones de Café (Cenicafé): Chinchiná, Colombialombia, 2014.
4. Rendón, J.; Arcila, J.; Montoya, E. Estimación de la Producción de café con Base en los Registros de Floración. Available online: https://www.cenicafe.org/es/publications/arc059(03)238-259.pdf (accessed on 10 August 2021).
5. Miranda, J.M.; Reinato, R.A.; Silva, A.B.d. Modelo matemático para previsão da produtividade do cafeeiro. *Rev. Bras. Eng. Agrícola Ambient.* **2014**, *18*, 353–361. [CrossRef]
6. Montoya-Restrepo, E. *Modelo para Simular la Producción Potencial del Cultivo del café en Colombia*; Boletin Técnico; FNC-Cenicafé: Chinchiná, Colombia, 2009.
7. dos Santos, L.M.; de Souza Barbosa, B.D.; Diotto, A.V.; Maciel, D.T.; Xavier, L.A.G. Biophysical parameters of coffee crop estimated by UAV RGB images. *Precis. Agric.* **2020**, *21*, 1227–1241. [CrossRef]
8. Idol, T.W.; Youkhana, A.H. A rapid visual estimation of fruits per lateral to predict coffee yield in Hawaii. *Agrofor. Syst.* **2020**, *94*, 81–93. [CrossRef]
9. Castro-Tanzi, S.; Flores, M.; Wanner, N.; Dietsch, T.V.; Banks, J.; Ureña-Retana, N.; Chandler, M. Evaluation of a non-destructive sampling method and a statistical model for predicting fruit load on individual coffee (*Coffea arabica*) trees. *Sci. Hortic.* **2014**, *167*, 117–126. [CrossRef]
10. Muñoz, C.A.U.; Rivera, R.D.M.; Ramos, C.P.F. Relación entre producción y características fenotípicas en *Coffea arabica* L. *Cenicafé* **2017**, *68*, 62–72.

11. Picini, A.G.; Camargo, M.B.P.D.; Ortolani, A.A.; Fazuoli, L.C.; Gallo, P.B. Desenvolvimento e teste de modelos agrometeorológicos para a estimativa de produtividade do cafeeiro. *Bragantia* **1999**, *58*, 157–170. [CrossRef]
12. Rosa, V.G.C.d.; Moreira, M.A.; Rudorff, B.F.T.; Adami, M. Coffee crop yield estimate using an agrometeorological-spectral model. *Pesqui. Agropecu. Bras.* **2010**, *45*, 1478–1488. [CrossRef]
13. Barbosa, B.D.S.; e Silva Ferraz, G.A.; Costa, L.; Ampatzidis, Y.; Vijayakumar, V.; dos Santos, L.M. UAV-based coffee yield prediction utilizing feature selection and deep learning. *Smart Agric. Technol.* **2021**, *1*, 100010. [CrossRef]
14. Kouadio, L.; Deo, R.C.; Byrareddy, V.; Adamowski, J.F.; Mushtaq, S. Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. *Comput. Electron. Agric.* **2018**, *155*, 324–338. [CrossRef]
15. Thao, N.T.T.; Khoi, D.N.; Denis, A.; Viet, L.V.; Wellens, J.; Tychon, B. Early Prediction of Coffee Yield in the Central Highlands of Vietnam Using a Statistical Approach and Satellite Remote Sensing Vegetation Biophysical Variables. *Remote Sens.* **2022**, *14*, 2975. [CrossRef]
16. Ozer, D.J. Correlation and the coefficient of determination. *Psychol. Bull.* **1985**, *97*, 307. [CrossRef]
17. Ordonez, A. La Sultana Farm. Available online: https://faca.unicauca.edu.co/cienciasagrarias/infraestructura (accessed on 15 July 2022).
18. Crickard, P., III. *Leaflet. js Essentials*; Packt Publishing Ltd.: Birmingham, UK, 2014.
19. QuestionPro. Tamaño de Muestra. Available online: https://www.questionpro.com/es/tama%C3%B1o-de-la-muestra.html (accessed on 15 August 2021).
20. Mapir. Survey3: Multi-Spectral Survey Cameras. Available online: https://www.mapir.camera/pages/survey3-cameras#specs (accessed on 15 August 2021).
21. Rosas, J.T.F.; de Carvalho Pinto, F.d.A.; de Queiroz, D.M.; de Melo Villar, F.M.; Magalhaes Valente, D.S.; Nogueira Martins, R. Coffee ripeness monitoring using a UAV-mounted low-cost multispectral camera. *Precis. Agric.* **2021**, *23*, 300–318. [CrossRef]
22. Mapir. Calibrating Images in MAPIR Camera Control Application. Available online: https://www.mapir.camera/pages/calibrating-images-in-mapir-camera-control-application (accessed on 15 August 2021).
23. Map, O.D. WebODM. Available online: https://github.com/OpenDroneMap/WebODM (accessed on 15 August 2021).
24. Soetedjo, A.; Hendriarianti, E. Plant Leaf Detection and Counting in a Greenhouse during Day and Nighttime Using a Raspberry Pi NoIR Camera. *Sensors* **2021**, *21*, 6659. [CrossRef] [PubMed]
25. Matlab. Color Thresolder. Available online: https://www.mathworks.com/help/images/ref/colorthresholder-app.html (accessed on 15 August 2021).
26. Favarin, J.L.; Dourado Neto, D.; García y García, A.; Villa Nova, N.A.; Favarin, M.d.G.G.V. Equações para a estimativa do índice de área foliar do cafeeiro. *Pesqui. Agropecu. Bras.* **2002**, *37*, 769–773. [CrossRef]
27. Bolton, D.K.; Friedl, M.A. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agric. For. Meteorol.* **2013**, *173*, 74–84. [CrossRef]
28. Nogueira Martins, R.; de Carvalho Pinto, F.d.A.; Marçal de Queiroz, D.; Magalhães Valente, D.S.; Fim Rosas, J.T. A Novel Vegetation Index for Coffee Ripeness Monitoring Using Aerial Imagery. *Remote Sens.* **2021**, *13*, 263. [CrossRef]
29. Parreiras, T.C.; Lense, G.H.E.; Moreira, R.S.; Santana, D.B.; Mincato, R.L. Using unmanned aerial vehicle and machine learning algorithm to monitor leaf nitrogen in coffee. *Coffee Sci.* 2020, *15*, e151736.
30. Molina, G.; Rodrigo, M. El Modelo de Regresión Lineal. Available online: http://ocw.uv.es/ciencias-de-la-salud/pruebas-1/1-3/t_09nuevo.pdf (accessed on 15 January 2022).
31. Scikit-Learn. Cross Validation. Available online: https://scikit-learn.org/stable/modules/cross_validation.html (accessed on 15 January 2022).
32. Khosla, E.; Dharavath, R.; Priya, R. Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. *Environ. Dev. Sustain.* **2019**, *22*, 5687–5708. [CrossRef]
33. Taylor, R. Interpretation of the Correlation Coefficient: A Basic Review. *J. Diagn. Med. Sonogr.* **1990**, *6*, 35–39. [CrossRef]
34. Rousel, J.; Haas, R.; Schell, J.; Deering, D. Monitoring vegetation systems in the great plains with ERTS. In Proceedings of the Third Earth Resources Technology Satellite—1 Symposium, Washington, DC, USA, 10–14 December 1973; NASA SP-351; pp. 309–317.
35. Silva, S.d.A.; Lima, J.d.S.; de Oliveira, R. Agrometeorological model estimating the productivity of two varieties of Arabic coffee considering the spatial variability. *IRRIGA* **2011**, *16*, 1–10. [CrossRef]
36. Santana, L.S.; Ferraz, G.A.e.S.; Marin, D.B.; Faria, R.d.O.; Santana, M.S.; Rossi, G.; Palchetti, E. Digital Terrain Modelling by Remotely Piloted Aircraft: Optimization and Geometric Uncertainties in Precision Coffee Growing Projects. *Remote Sens.* **2022**, *14*, 911. [CrossRef]
37. Barbosa, B.D.S.; Araújo e Silva Ferraz, G.; Mendes dos Santos, L.; Santana, L.S.; Bedin Marin, D.; Rossi, G.; Conti, L. Application of RGB Images Obtained by UAV in Coffee Farming. *Remote Sens.* **2021**, *13*, 2397. [CrossRef]