**Abstract:** The coffee plant is one of the main crops grown in Brazil. However, strategies to estimate its yield are questionable given the characteristics of this crop; in this context, robust techniques, such as those based on machine learning, may be an alternative. Thus, the aim of the present study was to estimate the yield of a coffee crop using multispectral images and machine learning algorithms. Yield data from a same study area in 2017, 2018 and 2019, Sentinel 2 images, Random Forest (RF) algorithms, Support Vector Machine (SVM), Neural Network (NN) and Linear Regression (LR) were used. Statistical analysis was performed to assess the absolute Pearson correlation and coefficient of determination values. The Sentinel 2 satellite images proved to be favorable in estimating coffee yield. Despite the low spatial resolution in estimating agricultural variables below the canopy, the presence of specific bands such as the red edge, mid infrared and the derived vegetation indices, act as a countermeasure. The results show that the blue band and green normalized difference vegetation index (GNDVI) exhibit greater correlation with yield. The NN algorithm performed best and was capable of estimating yield with 23% RMSE, 20% MAPE and R² 0.82 using 85% of the training and 15% of the validation data of the algorithm. The NN algorithm was also more accurate (27% RMSE) in predicting yield.

**Keywords:** coffee crop; yield; prediction models; spatial distribution of yield

---

1. **Introduction**

Coffee production is a significant driver of industrial and commercial activities in Brazil [1,2] creating a multitude of jobs in the growing regions [3]. The variety and extent of producing regions make Brazil the world’s largest coffee producer and second largest consumer market [4,5]. The Coffee Exporters Council of Brazil, CECAFÉ, estimates that by 2030 global coffee consumption will rise by around 30%, reaching 205 trillion kg. Thus, in order for Brazil to maintain its market position, national production must increase by 980 million kg by 2030 [6].

In order to increase yield, it is essential to know the variables that facilitate and cause the variability of coffee production [7]. According to Durand-Bessart [8], a strategy to increase yield is smaller spacing between rows and plants. Ref. [9] reported that when the aim is to increase yield, it is important to know the soil nutrient level. The authors of the study assessed the variability of phosphorous and potassium in a plantation, observing that soils in the most productive areas contained satisfactory nutrient levels. By contrast, the presence of pathogens in coffee crops is one of the main causes of a decline in production [10,11]. Among phytoparasites, nematodes can cause considerable losses in the planted area and consequent drop in yield [12,13].
Remote sensing is an alternative to determine agronomic parameters, such as yield [14,15]. Specifically, the interaction between electromagnetic radiation (EMR) and agricultural parameters has proved to be efficient in estimating the yield of different agricultural crops, primarily using multispectral images obtained from orbital platforms [16,17]. In pioneering studies, [18] used Moderate Resolution Imaging Spectroradiometer (MODIS) images to determine sugarcane yield in the 2004/2005 and 2005/2006 growing seasons. Using images from the same sensor, ref. [19] estimated the yield of soybean and maize crops in Paraná state, Brazil. The results showed that the methodology is highly efficient and can be replicated to map these crops. In addition to the abovementioned crops, the sensor is efficient in estimating coffee yields [20].

Coffee exhibits a number of unique characteristics that make estimating its yield difficult [21]. The bienniality of coffee results in alternating years of low and high yield [22,23]. Thus, yield estimation also depends on the analysis of a series of historical data, which incurs systematic operational costs [24]. In this respect, the correlation between agricultural yield and multispectral images shows significant potential, primarily in reducing cost and increasing producer profits [25]. Thus, in addition to obtaining quality multispectral images, the information obtained must be processed using robust techniques capable of coping with large volumes of data and generating reliable results [26]. Given this demand, machine learning has proved to be efficient in solving problems related to the processing of agronomic parameter data, and has been increasingly applied in precision agriculture [27,28]. Furthermore, multispectral sensors also offer other benefits such as facilitating the differentiation of some types of plants from their environment, thanks to the different vegetation indices such as the chlorophyll absorption index and the cellulose absorption index, among others. Moreover, they provide more information during the feature extraction processes, such as those used in the object-based image analysis methods and machine learning classification and regression techniques [14,17,27].

However, despite the existence of spectral models based on machine learning algorithms to estimate the yield of different crops, such as coffee, a number of methodological issues and shortcomings remain to be assessed by the specialized scientific community. For example, although coffee is a biannual crop, there are still no studies that use purely spectral and remote prediction models based on the yield of earlier harvests, that is, the literature is limited to spectral models to estimate current year crops and hybrid predictive models, combining multispectral images, meteorological data and agronomic parameters obtained in situ. Another point to be questioned are the agricultural conditions considered to create coffee yield prediction models. In many cases, the multispectral images taken are of experimental areas with uniform management of the entire area and controlled biotic and abiotic factors, not reflecting a real commercial coffee crop scenario.

Thus, given the lack of robust methodologies to estimate coffee yield from spectral models and the hypothesis that a relation can be established between the energy reflected by the coffee canopy and yield, the aim of the present study was to assess the potential of medium spatial resolution multispectral images and machine learning-based models in predicting coffee crop yield. To that end, in order to create spectral prediction models, we considered a historic three-year series of yield and Sentinel 2 satellite image data, as well as the following machine learning-based algorithms: RF, NN and SVM. In addition, a commercial coffee crop was used to create and assess the models.

2. Materials and Methods
2.1. Study Area

The study area is located at 18°40′7″ S and 47°35′22″ W, in the city of Monte Carmelo, in the state of Minas Gerais, Brazil. Red latosol predominates in the region and the crop grown was Coffea arabica L. cv. Catuaí 144, planted at the end of 2013 with 4.0 m between rows and 0.5 m between plants, totaling 5000 plants/ha. The total area planted was 54 ha, located at an altitude of approximately 820 m.

The Triângulo Mineiro and Alto Paraíba mesoregion is one of the main coffee producers in Minas Gerais state. —1. Climate is one of the factors that favors coffee plantations...
in this region. Average temperatures vary between 18 and 21 °C in winter, characterized by a hot dry climate, and the plants are grown at an altitude of approximately 850 m.

The production area underwent drip irrigation consisting of single-line drippers spaced 0.5 m apart, with a flow rate of 2.3 L/h and 100% of the irrigation depth. For better irrigation efficiency, the area was divided into eight sectors, where for reasons of topography and relief, the amount of water flow (m³/ha) did not remain constant, as shown in Figure 1B.

![Figure 1. (A) Map of the study area and yield sample sites for 2017 to 2020; (B) spatial distribution map of different irrigation sectors; (C) soil distribution map of soil nematodes.](image)

The soil physical and hydrological characteristics were determined in three equidistant layers in the 0.6 m profile, obtaining average overall density of 1100 kg/m³, soil volumetric humidity at field capacity and permanent wilting point of $3.5 \times 10^{-4}$ and $2.1 \times 10^{-4}$ m³, respectively, and water capacity available to plants of 0.082 m. One of the conditions that influences the health of a number of plants is the presence of soil nematodes (i.e., *Meloidogyne* species) (Figure 1C), which, when spatially distributed in large concentrations, affect the yield of some regions. Thus, crop and phytosanitary treatments were applied as needed.

### 2.2. Coffee Yield Data Collection

Coffee bean samples were used to determine yield, obtained at previously determined points in three field campaigns undertaken in May 2017, 2018 and 2019. The number of collection points in each campaign was 64, 64 and 80, respectively (Figure 1A). These points were georeferenced using a GNSS receiver, which was preceded by installing the base for the GNSS receiver pair, for the application of the survey by the relative positioning technique. The base was fixed to a strategic point, located in an area as open as possible, in order to mitigate possible errors and inaccuracies in the co-ordinates. Next, the receiver was used to measure the co-ordinates of the points analyzed, using the relative positioning method.

Beans from five coffee plants were collected at each sampling point, with two plants selected to the right and two to the left of the previously marked point. Thus, all the fruits were collected from the selected plants, using manual strip harvesting on canvas. The onset of harvesting was determined by the lowest possible percentage of green fruits on the plant (<10%). Next, the fruits were weighed to obtain total variable weight, then ripened
and classified into four classes: green, sugarcane green, cherry and dry-raisin. Yield was estimated in the laboratory after the drying stages and bean processing.

After the data were obtained, scattering analysis was conducted, where the yield samples and the respective multispectral image data by satellite were submitted to descriptive statistical analysis to obtain the mean, standard deviation and the coefficient of variation.

2.3. Multispectral Data Collection

In order to construct multispectral models that estimate yield, Sentinel 2 satellite images captured by the MSI (multi-spectral instrument) sensor were used. The models have a 12-bit radiometric resolution, with two focal planes based on a complementary metal oxide semiconductor (CMOS) monolithic detector, which records the radiation from the visible and near-infrared (NIR) channels, and a mercury-cadmium-telluride (MCT) detector hybridized in a CMOS to capture shortwave infrared radiation (SWIR). In addition, the sensor exhibits 13-band spectral resolution, 10-m spatial resolution for the bands $\rho_{490}$, $\rho_{560}$, $\rho_{665}$, $\rho_{842}$, and 20-m spatial resolution for the bands $\rho_{490}$, $\rho_{560}$, $\rho_{665}$, $\rho_{842}$, and 20-m spatial resolution for the bands $\rho_{705}$, $\rho_{740}$, $\rho_{783}$, $\rho_{865}$, $\rho_{1610}$, $\rho_{2190}$ (where the subscripts are the center wavelength in nm).

The images used were those from capture dates nearest the harvest time, which occurred on May 27 and 28, 2017, 2018 and 2019, totaling 3 images.

2.4. Multispectral Data Processing

The Sentinel 2 level 2A images were acquired and atmospherically corrected. The correction was performed using the SNAP software via the sen2cor plugin (version 2.10), which makes atmospheric, terrain and cirrus corrections in the upper atmosphere, transforming the images to level 1C (https://step.esa.int/main/snap-supported-plugins/sen2cor/ accessed on 1 July 2021). In the atmospheric correction process, the algorithm calculated the concentrations of aerosols and water vapor contained in the atmosphere based on the brightness values of the cirrus band [29].

After atmospheric correction, the 20 m spatial resolution images ($\rho_{705}$, $\rho_{740}$, $\rho_{783}$, $\rho_{865}$, $\rho_{1610}$ and $\rho_{2190}$) were resampled to standardize them at a 10 m spatial resolution. This procedure was carried out in the SNAP software. In order to calculate the new values of the resampled pixels, the algorithms of the nearest neighbor interpolation method were used. Down sampling and flag aggregation were used, whereby every pixel value in the output image is set to the nearest input pixel value.

In order to standardize the upper and lower radiometric thresholds of the images used after atmospheric correction and resampling, radiometric normalization was carried out with the 2017 image used as the standard. Processing was conducted in the ENVI 5.1 software, using the band math tool, following the methodology proposed by [30].

The linear transformation parameters were obtained by Equation (1):

$$T_i = m_i x_i + b_i$$

where $m_i = (Bri - Dri)/(Bsi - Dsi)$; $b_i = (Dri \times Bsi - Dsi \times Bri)/(Bsi - Dsi)$; $T_i =$ FRB of the reference image; $x_i =$ FRB of the image to be normalized; $Bri =$ average of the bright set of reference images; $Dri =$ average of the dark set of reference images; $Bsi =$ average of the bright set to be normalized; $Dsi =$ average of the dark set to be normalized and $i =$ sensor bands under study.

2.5. Remote Sensing Data Extraction and Calculation of Vegetation Indices

In order to increase spectral variability to estimate coffee yield from multispectral models, vegetation indices with visible and multispectral wavelength were calculated (Table 1). Vegetation indices, calculated for possible inclusion in prediction models, were selected according to their sensitivity to the agronomic parameters of coffee most correlated with yield, such as chlorophyll (NDRE, CI-RE, TCARI, CVI and CI-G), biomass (NDVI, RVI and SAVI), nitrogen (GNDVI), and leaf area index (MCARI). The indices were based on the
original bands of Sentinel 2 images and calculated using the ENVI 5.1 software, using the band math tool.

Table 1. Equations and references to calculate the vegetation indices derived from the original Sentinel 2.

<table>
<thead>
<tr>
<th>Index</th>
<th>Equation</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDRE</td>
<td>( \left( \frac{\rho_{842} - \rho_{705}}{\rho_{842} + \rho_{705}} \right) )</td>
<td>[31]</td>
</tr>
<tr>
<td>NDVI</td>
<td>( \left( \frac{\rho_{842} - \rho_{665}}{\rho_{842} + \rho_{665}} \right) )</td>
<td>[32]</td>
</tr>
<tr>
<td>GNDVI</td>
<td>( \left( \frac{\rho_{842} - \rho_{560}}{\rho_{842} + \rho_{560}} \right) )</td>
<td>[33]</td>
</tr>
<tr>
<td>CI-RE</td>
<td>( \frac{\rho_{842}}{\rho_{705}} - 1 )</td>
<td>[34]</td>
</tr>
<tr>
<td>MCARI</td>
<td>( [(a) - 0.2 \times (b)] \times \left( \frac{\rho_{320}}{\rho_{842}} \right) )</td>
<td>[35]</td>
</tr>
<tr>
<td>TCARI</td>
<td>( 3 \times [(a) - 0.2 \times (b)] \times \left( \frac{\rho_{320}}{\rho_{842}} \right) )</td>
<td>[36]</td>
</tr>
<tr>
<td>RVI</td>
<td>( \frac{\rho_{665}}{\rho_{842}} )</td>
<td>[37]</td>
</tr>
<tr>
<td>SAVI</td>
<td>( \left( \frac{\rho_{842} - \rho_{665}}{\rho_{842} + \rho_{665} + \text{C}} \right) \times \left( \frac{1}{1 + \text{C}} \right) )</td>
<td>[38]</td>
</tr>
<tr>
<td>CVI</td>
<td>( \frac{\rho_{842} \times \rho_{665}}{\rho_{560}^2} )</td>
<td>[39]</td>
</tr>
<tr>
<td>CI-G</td>
<td>( \frac{\rho_{842}}{\rho_{705}} - 1 )</td>
<td>[34]</td>
</tr>
</tbody>
</table>

NDVI: normalized difference vegetation index; GNDVI: green normalized difference vegetation index; CI-RE: chlorophyll index—red edge; MCARI: modified chlorophyll absorption in reflectance index, where \( a = R_{700} - R_{670}, b = R_{670} - R_{550} \); TCARI: transformed chlorophyll absorption in reflectance index, where \( a = R_{700} - R_{670}, b = R_{700} - R_{550} \); RVI: ratio vegetation index (also called simple ratio); SAVI: soil adjusted vegetation index; CVI: chlorophyll vegetation index; CI-G: chlorophyll index—green.

The estimation models were constructed using the surface reflectance values and digital numbers extracted from the original bands and the vegetation indices derived from the geographic position of the points sampled in the field. For each coffee yield sampling point, information on only one pixel per co-ordinate was extracted, given that the spatial resolution of the sensor is greater than the total area, which contains the 5 plants used to obtain the yield of each point (P).

2.6. Descriptive and Exploratory Yield Analysis

In order to analyze yield behavior and the value of the Sentinel 2 satellite bands, descriptive analysis of these parameters was carried out. A priori, the mean, standard deviation and coefficient of variation were calculated for these variables.

In addition, exploratory analysis was conducted for the yield parameter, where the maximum, minimum, 1st quartile, 2nd quartile, median and coefficient of variation were estimated. The aim was to determine the variability of coffee yield values, as a function of its bienniality, and whether the data demonstrate any trend.

2.7. Generation of Prediction Models and Quality Control

In a situation similar to that of studies presented in the literature, this study is limited to local prediction models, given that the data used are restricted to a planted area of 54 ha. However, the aim was to create a unique model that would be useful in predicting and forecasting yield, regardless of the bienniality of the coffee crop.

The spectral values extracted were used to conduct statistical analyses to determine which image information was best related to yield and, as such, could be used in prediction models. To that end, the correlation between yield and radiometric data was determined and analysis of the significance of each attribute was analyzed as a function of the regression models used.
Thus, the six indices/bands that contributed significantly to model generation were separated, that is, the vegetation indices and bands with the highest correlation with yield were selected.

These values were used to establish which bands/vegetation indices would be used in the model generation process. Given the large number of parameters available to estimate yield, tests were carried out to determine the ideal number of variables that would be used. Thus, the primary aim was to avoid overfitting the model created to estimate yield, which may occur when the number of parameters used is high.

The selection of which parameters were used was based on prioritizing the highest absolute Pearson correlation and coefficient of determination values. This criterion was used to select the parameters that exhibited both a negative and positive correlation with the study variable. In this respect, it was possible to provide a greater total contribution of the parameters used.

Next, the yield estimation models were created, where two methodologies were applied, based on non-parametric RF, NN, SVM and parametric regression models (Simple and Multiple Linear Regression-LR), the main algorithms used to estimate agricultural variables from data obtained by Remote Sensing [27,28]. This stage was carried out in the Weka 3.9.5 software (https://waikato.github.io/weka-wiki/downloading_weka/ accessed on 1 August 2021).

After one series of tests in the present study, the RF algorithm was configured into 100 iterations, one seed and null depth, the NN into one intermediate layer, three neurons, learning rate of 0.3 and momentum of 0.2, and SVM as a Pearson-type kernel classifier with a coefficient c of 0.5.

Processing using parametric multiple linear regression consists of using one or more independent parameters able to describe the study variable. Thus, the response variable is determined by creating a model that best fits the data used [40,41].

In the present study, the linear regression algorithm belonging to the Weka software was used, whereby the multiple forward stepwise linear regression technique was applied to calculate the model. This initially considers a simple regression model, using the variable that exhibited the highest correlation coefficient with the response variable (yield) as an auxiliary variable (bands and multispectral indices). The process continues for as long as a new auxiliary variable is incorporated into the model, and stops when no new variable is included.

In order to create the models and analyze their accuracy, three scenarios were established for both algorithms, considering the number of training and validation samples, as follows: first scenario—80% training, 20% validation; second scenario—85% training, 15% validation; third scenario—90% training, 10% validation. The training and validation samples were randomly separated, with no distinction made for sampling year in the field. The different scenarios considered for the number of training and validation samples made it possible to analyze the interference of test/validation sample size in estimating yield. In relation to creating models and analyzing the accuracy of multi-temporal data, it is essential to establish a methodology to predict yield considering the temporal variability of coffee.

The models were validated by calculating the root mean square percentage error (RMSE%) (Equation (1)), mean absolute percentage error (MAPE%) (Equation (2)) and coefficient of determination ($R^2$) between measured and estimated productivity. Finally, yield maps for 2017, 2018 and 2019 were created for model spatialization and to represent yield distribution.

\[
\text{MAPE} (%) = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{y_i} \times 100
\]  
\[
\text{RMSE} (%) = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} y_i}} \times 100
\]  

where $\hat{y}_i$ are the predicted values, $y_i$ measured values and $n$ total number of observations.
2.8. Analysis of Multispectral Model Accuracy in Predicting Yield

In order to assess the accuracy of yield prediction models calibrated with data from 2017, 2018 and 2019, a Sentinel 2 image of the same area taken two months before the 2020 harvest was used, that is, the coffee crop already had a significant load of unripe fruit. To that end, an image taken in May 2020 was used, where RMSE% analysis was conducted on 64 control points with yield measured in situ (Figure 1A). Finally, yield was mapped for the entire study area.

3. Results

3.1. Characterization of the Study Area: Exploratory Analysis of Yield Data

Table 2 presents the minimum/maximum/Q1/Q3/median yield values, separating the values by harvest year. Analysis of the values exhibited in the table shows that the years of lowest crop yield were 2017 and 2019, and 2018 the highest.

<table>
<thead>
<tr>
<th>Year</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
<th>CV%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017</td>
<td>95.4</td>
<td>169.2</td>
<td>265.2</td>
<td>445.2</td>
<td>802.2</td>
<td>91.5%</td>
</tr>
<tr>
<td>2018</td>
<td>3.900</td>
<td>4.830</td>
<td>5.454</td>
<td>6148.8</td>
<td>7319.4</td>
<td>90.2</td>
</tr>
<tr>
<td>2019</td>
<td>78.9</td>
<td>472.8</td>
<td>1003.2</td>
<td>1852.2</td>
<td>2672.4</td>
<td>92.3</td>
</tr>
</tbody>
</table>

Q1: Quartile 1; Q3: Quartile 3.

Given the physiological traits of the crop and different phytosanitary, management and irrigation conditions in the area for the three years, the coefficient of variation (CV) was greater than 90%. In 2019, the lowest and highest values were 78.9 and 2672.4 kg/ha, respectively. Q1 and Q3 obtained 472.8 and 1852.2 kg/ha, respectively, with an interquartile range of 1379.4 kg/ha. The median for 2019 was 1003.2 kg/ha, indicating an increase in yield over 2017.

3.2. Statistical Analysis of the Agronomic Parameters

The means, standard deviations and the coefficients of variation (CV) are shown in Table 3.

Table 3. Means, standard deviations and coefficients of variation (CV) of the coffee yield and reflectance of the bands of the Sentinel 2 sensor.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Means</th>
<th>Standard Deviations</th>
<th>CV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yield</td>
<td>42.98</td>
<td>39.17</td>
<td>91.14</td>
</tr>
<tr>
<td>$\rho_{490}$</td>
<td>173.53</td>
<td>53.46</td>
<td>30.81</td>
</tr>
<tr>
<td>$\rho_{560}$</td>
<td>345.20</td>
<td>72.42</td>
<td>20.98</td>
</tr>
<tr>
<td>$\rho_{665}$</td>
<td>303.56</td>
<td>130.21</td>
<td>42.89</td>
</tr>
<tr>
<td>$\rho_{705}$</td>
<td>688.40</td>
<td>146.90</td>
<td>21.35</td>
</tr>
<tr>
<td>$\rho_{740}$</td>
<td>2078.00</td>
<td>119.10</td>
<td>5.73</td>
</tr>
<tr>
<td>$\rho_{783}$</td>
<td>2811.60</td>
<td>140.90</td>
<td>5.01</td>
</tr>
<tr>
<td>$\rho_{842}$</td>
<td>2766.90</td>
<td>127.90</td>
<td>4.62</td>
</tr>
<tr>
<td>$\rho_{865}$</td>
<td>3035.50</td>
<td>157.10</td>
<td>5.18</td>
</tr>
<tr>
<td>$\rho_{1610}$</td>
<td>1561.70</td>
<td>288.10</td>
<td>18.45</td>
</tr>
<tr>
<td>$\rho_{2190}$</td>
<td>822.70</td>
<td>264.00</td>
<td>32.09</td>
</tr>
</tbody>
</table>

CV = coefficient of variation.

According to Table 3, the highest CV was for productivity, 91.14, in contrast to the high yield variability, data variability for surface reflectance spectral was significantly lower,
where the highest CV values were recorded for the visible and near and mid-infrared bands, varying from 18.45 ($\rho_{1610}$) to 42.89% ($\rho_{665}$). The lowest CVs were obtained for the red edge bands, where variability was between 4.62 ($\rho_{783}$) and 5.73% ($\rho_{740}$).

3.3. Analysis of the Correlation between Yield and Multispectral Bands and Their Derived Indices

Table 4 presents the correlation and coefficient of determination between original bands and vegetation indices obtained from the Sentinel 2 sensor (reflectance extracted from the training data (90% of the total sample)).

Table 4. Correlation between the vegetation bands/indices and yield.

<table>
<thead>
<tr>
<th>Band/Index</th>
<th>Correlation</th>
<th>Coefficient of Determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{490}$</td>
<td>0.72</td>
<td>53.00</td>
</tr>
<tr>
<td>$\rho_{560}$</td>
<td>0.65</td>
<td>43.00</td>
</tr>
<tr>
<td>MCARI</td>
<td>0.55</td>
<td>29.30</td>
</tr>
<tr>
<td>$\rho_{705}$</td>
<td>0.50</td>
<td>24.60</td>
</tr>
<tr>
<td>$\rho_{740}$</td>
<td>0.46</td>
<td>24.60</td>
</tr>
<tr>
<td>$\rho_{1610}$</td>
<td>0.41</td>
<td>13.80</td>
</tr>
<tr>
<td>$\rho_{2190}$</td>
<td>0.40</td>
<td>12.80</td>
</tr>
<tr>
<td>$\rho_{565}$</td>
<td>0.33</td>
<td>10.50</td>
</tr>
<tr>
<td>$\rho_{842}$</td>
<td>0.16</td>
<td>2.10</td>
</tr>
<tr>
<td>$\rho_{865}$</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>$\rho_{783}$</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>SAVI</td>
<td>−0.34</td>
<td>10.70</td>
</tr>
<tr>
<td>NDVI</td>
<td>−0.34</td>
<td>10.70</td>
</tr>
<tr>
<td>RVI</td>
<td>−0.37</td>
<td>13.20</td>
</tr>
<tr>
<td>TCARI</td>
<td>−0.45</td>
<td>20.00</td>
</tr>
<tr>
<td>NDRE</td>
<td>−0.48</td>
<td>22.70</td>
</tr>
<tr>
<td>CI_RE</td>
<td>−0.49</td>
<td>23.40</td>
</tr>
<tr>
<td>CVI</td>
<td>−0.52</td>
<td>26.60</td>
</tr>
<tr>
<td>CI_G</td>
<td>−0.60</td>
<td>35.90</td>
</tr>
<tr>
<td>GNDVI</td>
<td>−0.62</td>
<td>38.30</td>
</tr>
</tbody>
</table>

Table 4 shows that the deterministic coefficients of determination are not above 53%. In relation to correlation, in absolute values, Pearson’s correlations varied from 0.05 ($\rho_{783}$) to 0.72 ($\rho_{490}$). The lowest limit for significant correlations was 0.40 ($\rho_{2190}$), that is, the correlations with a $p$-value < 0.05 (>95% significance level). Among the original bands, the highest significant correlations were recorded primarily in visible spectral bands, such as, $\rho_{490}$ (0.72), $\rho_{560}$ (0.64), $\rho_{705}$ (0.50), $\rho_{740}$ (0.46) and at a lower magnitude, $\rho_{1610}$ (0.42) and $\rho_{2190}$ (0.41). With respect to vegetation indices, the significant correlations were predominantly negative, varying from −0.45 (TCARI) to −0.62 (GNDVI), except for the MCARI index, which was 0.55.

3.4. Analysis of the Performance of Predictive and Forecast Models

Table 5 presents the accuracy (RMSE%) and trend level (MAPE%) of the predictive and forecast models based on the RF, NN, SVM and LR algorithms. In order to construct the models, the six highest absolute correlation values were used: $\rho_{490}$, $\rho_{560}$ MCARI, $\rho_{705}$, $\rho_{740}$ and NDRE.
Table 5. Performance of algorithms in estimating yield.

<table>
<thead>
<tr>
<th>P.S.</th>
<th>RF</th>
<th>SVM</th>
<th>LR</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE%</td>
<td>MAPE%</td>
<td>R²</td>
<td>RMSE%</td>
</tr>
<tr>
<td>Test 80%</td>
<td>27</td>
<td>23</td>
<td>0.81</td>
<td>37</td>
</tr>
<tr>
<td>Test 85%</td>
<td>27</td>
<td>23</td>
<td>0.81</td>
<td>35</td>
</tr>
<tr>
<td>Test 90%</td>
<td>28</td>
<td>24</td>
<td>0.80</td>
<td>36</td>
</tr>
<tr>
<td>Y.P. 2020</td>
<td>32</td>
<td>38</td>
<td>41</td>
<td>41</td>
</tr>
</tbody>
</table>

PS = percent splitage; Y.P. = yield prediction; RMSE = root mean squared error; MAPE = mean absolute percentage error; R² = coefficient of determination; RF = random forest; SVM = support vector machine; LN = linear regression; NN = neural network.

Table 5 shows that for all the algorithms, the highest RMSE%, MAPE% and correlation occurred in conditions in which the models were constructed with 85% of the samples. For this situation, NN and RF obtained lower RMSE% (23 and 27%, respectively), that is, the models exhibited an accuracy of 77 and 73%, respectively. The highest errors were for the LR and SVM algorithms, which displayed RMSE% of 39 and 36%, that is, accuracy of 61 and 64%, respectively.

In relation to MAPE%, for 85% sample training, the model generated by NN obtained the lowest trend (20%) and LR the highest (34%). In relation to R², for 85% sample training, the model generated by NN obtained the highest trend (0.85) and LR the lowest (0.67). In relation to forecast accuracy, the NN algorithm performed best, where it was possible to predict yield with an RMSE% of 27%, that is, model accuracy was 73%.

3.5. Time-Space Distribution of Yield

Figure 2 shows the maps created as a function of the yield estimated for 2017, 2018 and 2019.

Figure 2. Spatial distribution of yield in the study area, derived from the RF algorithm applied to the bands/indices obtained from the Sentinel 2 satellite. (a) Yield for 2017; (b) yield for 2018; and (c) yield for 2019.

The maps in Figure 2 show the space-time evolution of yield, demonstrating the biennial factor, where low yield occurred in 2017 (660–1800 kg) and 2019 (900–1800 kg) and high yield in 2018 (4500–5400 kg). Rows in the lower portion of the study area were always the least productive, irrespective of temporal crop growth. This is because of the...
high concentration of nematodes in the soil in this area (Figure 1C) and recent crop renewal in some locations.

4. Discussion

With respect to yield variability in the area between 2017 and 2019, despite returning to the low biennality year, there was a considerable rise in yield values. This increase may be associated with an improvement in the productive conditions of the study area, given that the rise occurred systematically in all the data used.

The highest CV was found for the variable yield, which may be due to the particularity of the crop, which exhibits product biennality, that is, alternating years of low and high yields. It is important to underscore that this scenario is normal for this crop, given that in studies aimed at estimating coffee yield using different remote sensing approaches, ref. [15,42] also reported high yield variability in different coffee cultivars.

In addition, crop management contributed to the high yield variability, since in addition to being a productive commercial area, it is used as an experimental area for studies on the application of agrochemicals. Among these applications is the management of different irrigation depths and magnetized water in specific rows of coffee (Figure 1B).

Specifically in this area, variability of visible wavelengths is due to the different leaf pigment levels resulting from the application of variable amounts of fungicides per sector. Visible wavelength variability is explained by plant structure at different phenological stages, given that small parts of the study area were replanted. The variability in mid-infrared bands is related to the amount of water in plants resulting from the different sectors in the study area. Although nematodes have been detected in the soil of the study area, the low spectral variability in red edge bands is due to the low disease and pest incidence in the area.

Other factors may also be related to the spectral variability of the data, according to [43], who attributed visible band variability to the nutritional conditions of the coffee, where variations in nitrogen caused high variability in the blue and red bands, due to chlorophyll a and b. Ref. [42] reported that multispectral data, specifically NIR bands, are also sensitive to the volumetric variability of coffee, primarily near harvest times, that is, conditions in which coffee exhibits a larger volume due to the presence of fruits on the branches.

In relation to mid-infrared variability, ref. [44] reported the sensitivity of infrared bands to the stress on coffee caused by diseases. With respect to the spectral variability in red edge bands, ref. [45] found that only under high pathogen infestation is it possible to discriminate healthy from infected vegetation. Thus, sample heterogeneity becomes extremely high, which could hinder the accuracy of purely radiometric models in estimating predictive models.

Although the correlation results were smaller than those found in studies with annual crops, such as that conducted by [46], the findings suggest a dependence of leaf area and pigmentation in relation to coffee productivity in the same year. Coffee is a perennial crop and takes two years to complete its phenological cycle, unlike most other crops, which complete their reproductive cycle in one year. Thus, this crop represents a unique set of problems, as it follows a biannual phenological cycle and presents high and low production in alternate years. This feature has been reported as an important factor to be incorporated into spectral models to estimate coffee productivity [7]. An effective tool to assess this pattern and estimate it in the spatial domain could significantly improve coffee productivity modeling.

The coefficients of determination between bands, vegetation indices and yield demonstrated the complexity and influence of different environmental variables that affect coffee yield and at the same time, are not sensitive to the Sentinel 2 bands. Ref. [47] underscored that these low coefficients indicate that, although vegetation indices and bands may express coffee crop biomass, yield is a more complex factor, which depends on leaf biomass and numerous environmental conditions. In addition, the effect of biomass on yield is an
indirect result of increased flowering. High yield values are a result of suitable biomass conditions; however, suitable biomass alone does not ensure high yields, especially in years with water stress or extreme minimum temperatures during critical phenological phases.

The high correlation with coffee yield in visible bands occurs because these ranges are highly correlated with fruit volume and load [46]. According to [47] for parameters related to coffee yield such as biomass and height, the green and red bands of the GeoEye sensor also exhibited significant correlations (>0.70). The high correlations found mainly in indices such as GNDVI are due to the fact that they are sensitive to the cell structure and phenological stage of coffee plants in harvest conditions [46].

Ref. [14] related that vegetation indices consisting of NIR bands derived from the OLI/Landsat8 sensor exhibited a correlation of up to 0.89. This was also observed by [47], who underscored that even for sensors with low spatial resolution, such as MODIS, vegetation indices such as GNDVI show a high correlation with coffee yield and may display different magnitudes as a function of the bienniality of coffee.

Furthermore, ref. [47] used lagged correlation analysis and annual cycle deviations to relate yield to accumulated deviations in fractional vegetation. The MODIS vegetation indices were spatially aggregated over the municipality of Monte Santo de Minas, Minas Gerais, Brazil. The authors observed that data from MODIS vegetation indices converted to fractional vegetation indicate trends in coffee productivity. As the correlation between vegetation indices (GNDVI, CI_G and MCARI) and productivity is significant in this study, the alternating pattern in coffee productivity is also true for vegetation indices. Thus, it is possible to infer the biennial effect through vegetation indices.

In contrast, the correlation of NDVI (0.34) and SAVI (0.34) indices was low which, in different studies, exhibited a high correlation with yield [14,47]. Ref. [14] related the low correlation to crops submitted to different irrigation systems or some type of leaf stress caused by pests and pathogens. Thus, for the area of this study, managing irrigation under specific conditions was the primary factor that influenced the correlation between bands, indices and yield.

In regard to the indices and bands used to create the spectral estimation and predictive models, the correlation analysis showed that the vegetation indices and bands do not fully explain the yield variation because there are many factors responsible for the final yield, but these indices can be useful as indicators of the biennial coffee yield. Based on the RMSE%, MAPE% and $R^2$ values obtained, spectral model performance reflects the possibility of estimating and predicting coffee yield remotely, despite the high spectral variability of the canopy caused by the bienniality of the crop and the biotic and abiotic factors in the production environment. Thus, the methodology breaks the paradigm of studies such as that by [48], who report that remote sensing can only be used to estimate coffee yield with hybrid models that consider the association between spectral data and leaf macro and micronutrient measures.

In general, the highest accuracy and lowest trend for the non-parametric NN and RF models were expected, given that coffee yield does not follow a normal distribution and exhibits high variability, since under crop conditions, plants were observed under different physiological conditions, such as agronomic parameters related to plant health and growth. Another factor that explains the better performance of non-parametric models in estimating yield is the fact that multi-temporal yield was considered, that is, given the bienniality of production, parametric models such as LR were only highly accurate under annual prediction conditions [47]. This premise was also observed by [14], where coffee yield was estimated using vegetation indices derived from the OLI/Landsat8 sensor associated with linear regression models. However, the present study shows the need for different models associated with years of low and high yield. The high accuracy of non-parametric models in estimating coffee yield has also been observed in other studies. Ref. [15] reported the possibility of estimating coffee yield with RMSE % and MAPE% of 16 and 13%, respectively, from algorithms based on Baysian inference and random forests based on Landsat 8 satellite images.
The high correlations between yield measured in loco and estimated by spectral models demonstrate the capacity to accurately map areas with low and high coffee yield based on Sentinel 2 images. The decline in forecast compared to prediction accuracy may be due to the fact that the algorithm was trained with only the yield data of previous years and because the crop exhibited different biotic e abiotic factors.

From the yield distribution maps, the biennality effect was observed primarily in the coffee plant rows in the upper section of the study areas, where yield alternated according to natural crop development. The challenges to implementing the models in areas neighboring the study area require indexing more yield data from other agricultural regions, that is, coffee growing areas subject to different management techniques and environmental conditions (Figure 1B). In relation to the spatial distribution pattern of yield classes, an irregular geometric pattern can be observed, that is, the map does not significantly reflect any spectral model sensitivity to the rectilinear geometry defined by the different irrigation sectors (Figure 1B). Thus, the spatial distribution of yield is more conditioned to the biotic and abiotic factors of the crop itself (Figure 1C), that is, the yield classes do not exhibit the influence of the anthropic actions applied in the experimental area.

With respect to biennality data, in order to improve prediction models, yield data should be continuously inserted over the years, since a historical series will reflect natural yield variations related to environmental conditions, that is, crops compromised by factors other than management, such as pest and pathogen attacks, drought and likely water stress.

5. Study Limitations and Future Perspectives

A number of studies are being conducted to estimate coffee yield from multispectral images. Most of these propose local temporal spectrum models limited to the biennality of coffee crops, that is, different predictive models should be created for years of high and low yields. In a situation similar to that of studies presented in the literature, this study is limited to local prediction models, given that the data used are restricted to a planted area of 54 ha. However, the aim was to create a unique model that would be useful in predicting and forecasting yield, regardless of the biennality of the coffee crop. The challenges to implementing the models in areas neighboring the study area require indexing more yield data from other agricultural regions, that is, coffee growing areas subject to different management techniques and environmental conditions.

From an economic standpoint, there is also a need to apply prediction models to multispectral sensors with better spatial resolutions. This is because the prediction models generated by images with average spatial resolution provide only general information of real yield observed in loco. Coffee is the most valuable agricultural crop in the world, and the more specific the information, the more the implementation of this technology for crop management is justifiable.

6. Conclusions

The results of the present study showed that the possibility of estimating the yield of a nematode-infested commercial coffee crop under irrigation based on spectral models constructed by machine learning-based algorithms. To that end, with the addition of historical yield data, it was possible to model yield biennality, estimate and predict future crop yield in a 54-ha experimental area. These results showed that it is possible to reduce the monitoring time of the crop during the production year.

The main findings of the paper were:

1. The Sentinel 2 satellite images were favorable in estimating coffee yield. Despite their low spatial resolution in estimating agricultural variables below the canopy, specific bands such as the red edge, mid-infrared and derived vegetation indices act as a countermeasure to this limitation.

2. The blue band and GNDVI1 vegetation index showed the highest correlation with yield, but the low accuracy exhibited by the spectral models demonstrated the need to predict yield from non-parametrized algorithms. Additionally, other indices that also
displayed significant correlations (CI_G and MCARI) indicate that leaf pigmentation and biomass remain a reasonable method to estimate coffee yield.

3. After a data mining process, the NN algorithm exhibited the highest and lowest trend in predicting and forecasting yield.

4. The high coefficients of determination showed the capacity of the spectral model to accurately estimate the spatial distribution of high and low-yield areas.

5. The yield distribution maps demonstrated the sensitivity of spectral models to the biotic and abiotic factors present in the crop.


Funding: This research received no external funding.

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

References
8. Durand-Bessart, C.; Tixier, P.; Quinteros, A.; Andreotti, F.; Rapidel, B.; Tauvel, C.; Allinne, C. Analysis of interactions amongst shade trees, coffee foliar diseases and coffee yield in multistrata agroforestry systems. *Crop Prot.* 2020, 133, 105137. [CrossRef]