Smart Agricultural Technology 1 (2021) 100010

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.elsevier.com/locate/atech

UAV-based coffee yield prediction utilizing feature selection and deep learning

Brenon Diennevan Souza Barbosa[a], Gabriel Araújo e Silva Ferraz[a,*], Lucas Costa[b], Yiannis Ampatzidis[b,*], Vinay Vijayakumar[b], Luana Mendes dos Santos[a]

[a] Department of Agricultural Engineering, Federal University of Lavras (UFLA), Lavras, Minas Gerais 37200-900, Brazil
[b] Department of Agricultural and Biological Engineering, Southwest Florida Research and Education Center, University of Florida, Immokalee, FL 34142, United States

ARTICLE INFO

Keywords:
Deep-learning
Remote sensing
UAV imagery
Yield prediction

ABSTRACT

Unmanned Aerial Vehicles (UAVs) combined with machine learning have a great potential for crop yield estimation. In this study, a UAV equipped with an RGB (Red, Green, Blue) camera and computer vision algorithms were used to estimate coffee tree height and crown diameter, and for the prediction of coffee yield. Data were collected for 144 trees between June 2017 and May 2018, in the Minas Gerais, Brazil. Six parameters (leaf area index - LAI, tree height, crown diameter, and the individual RGB band values) were used to develop UAV-based yield prediction models. First, a feature ranking was performed to identify the most significant parameter(s) and month(s) for data collection and yield prediction. Based on the feature rankings, the LAI and the crown diameter were determined as the most important parameters. Five algorithms were used to develop yield prediction models: (i) linear support vector machines (SVM), (ii) gradient boosting regression (GBR), (iii) random forest regression (RFR), (iv) partial least square regression (PLSR), and (v) neuroevolution of augmenting topologies (NEAT). The mean absolute percentage error (MAPE) was used to evaluate the yield prediction models. The best result was obtained by the NEAT algorithm (MAPE of 31.75%) for a reduced dataset containing only the most important features (LAI and the crown diameter) and the most important months (December 2017 and April 2018). The results suggest that a dataset of the most important month (December) could be used for the yield prediction model, reducing the need for extensive data collection (e.g., monthly data collection).

Introduction

Brazil is responsible for one-third of the world's coffee production, surpassing other countries such as Colombia and Vietnam in the production of this bean. The production for the year 2020 is estimated between 3,420 and 3,720 tons of coffee. These estimates are based on an increase in the productive area of about 4% (1,885 million ha) compared to the 2019 harvest. The state of Minas Gerais alone has an estimated production of 30 to 32 million bags (approximately 50% of the national production). This increase in production compared to the 2019 harvest is related to the negative biennial cycle of coffee [25]. This biennial rhythm is due to the plant's resource allocation: in the productive year, the coffee tree prioritizes crop production over vegetative; in the subsequent year, the plant must compensate for the vegetative shortfall by producing flowers and fruits. During the negative cycle, growers introduce new management practices, such as pruning, expecting a crop recovery in two years (Silva et al. 2016).

The increase in the productivity of crops can be achieved with the help of new precision agriculture techniques and technologies. Precision agriculture provides the agricultural manager with accurate information so that decisions can be made to optimize resources, especially in the current scenario of complex territorial expansion, high cost of inputs, multiple management practices and treatments, and preservation of the environment. The application of precision agriculture for coffee shows high potential, especially with geostatistical techniques, which can characterize the spatial distribution of the fruit's detachment force and generate maps of selective harvesting of fruits at the appropriate place and time [8].

The constant evolution of technology in the agricultural environment and the availability of remote sensing data has reinforced the popularity of unmanned aerial vehicles (UAVs), which are easy to operate and efficient in obtaining high spatial and temporal resolution images [37]. According to Deng et al. [30], UAVs have become more accessible to farmers because of advancements in technologies associated with this equipment and a reduction in the acquisition costs over the years. In

* Corresponding authors.
E-mail addresses: bdiennevan@estudante.ufla.br (B.D.S. Barbosa), gabriel.ferraz@ufla.br (G.A.e.S. Ferraz), i.ampatzidis@ufl.edu (Y. Ampatzidis), luana.goncalves1@estudante.ufla.br (L.M. dos Santos).

https://doi.org/10.1016/j.atech.2021.100010
Received 13 July 2021; Received in revised form 9 September 2021; Accepted 9 September 2021

coffee crops, dos Santos et al. [31] show the potential of UAV to obtain biophysical characteristics of the plant (height and diameter) quickly and non-destructively. The results can support decision-making for accurate coffee management.

In agricultural monitoring, machine learning (ML) algorithms, such as neural networks are in constant use, mainly associated with precision agriculture application techniques [52]. For example, Costa et al. (2021a) developed machine vision tools utilizing ML for measuring pecan nut growth, and Abdulridha et al. [1–3] developed disease detection techniques using UAV hyperspectral imaging and ML. Ampatzidis and Partel et al. [6] and Ampatzidis et al. [4] developed artificial intelligence (AI) based algorithms to process, analyze, and visualize data collected from UAVs for citrus. These algorithms can detect, count, and geolocate trees, categorizing them based on their health ([5, 26]a), plant nutrient concentration ([26]b), and yield potential (Vijayakumar et al., 2021).

Mathematical models to predict the behavior and productivity of the crop by biophysical and climatic variables are excellent tools for the agricultural markets to predict the risk of financial loans, crop estimates, and future markets. Several studies to estimate the productivity in crops such as onion [9], beet [49], soy [44], corn [39, 45], and forage crops [33] have been developed mainly with the use of aerial images from UAVs.

For coffee in Brazil, several studies utilized remote sensing and image processing techniques for monitoring diseases ([10]a; [11, 22]), water stress assessment [23], detection of planting failures [48], estimation of the volume of harvested fruits [18], plant volume estimation [27], and nematode detection [47].

For coffee yield prediction, Carvalho et al. [20] developed techniques for estimating crop productivity (yield) utilizing genomic-based models to minimize the biennial effect on the production cycle. De Oliveira Aparecido et al. [28] created a model for forecasting the production of coffee crops, using agrometeorological data, in different regions of Minas Gerais, finding results considered satisfactory for forecasts with a period of five months before harvest. Kouadio et al. [41] developed machine learning models to estimate coffee yield based on soil fertility maps, achieving a root mean squared error of 496.35 kg/ha. However, studies that explore the potential of UAVs to predict the productivity of a coffee crop are scarce in the specialized literature.

This study aims at estimating coffee productivity by using data from aerial images obtained by UAVs. Computer vision algorithms were developed to estimate tree height and crown diameter from the data collected by UAVs and compared with manual field measurements. To understand the level of reliability of the UAV-based predictive models, the models were compared with the field measurements, and a measure of agreement was estimated using difference plots. Then, the UAV data were used to develop models to predict yield. The yield prediction data along with the actual yield values were used to generate a feature ranking to determine the features that most influence the yield prediction model. Once the most important (for yield prediction) feature(s) and month(s) were determined, multiple yield prediction models were generated and compared to find the best model with the lowest error percentage.

Materials and Methods

Experimental design

A UAV (DJI Phantom 3 professional, DJI, China) equipped with an RGB camera (IMX147, Sony, Japan) containing a GPS sensor was used to collect images of the coffee trees. The images were then processed and stitched to create an orthomosaic image. The manual field measurements of tree height and crown diameter were taken on 144 sample trees, and the leaf area index (LAI) was determined using these two values based on a methodology developed by Favarin et al. [32]. The Crown Height Model (CHM) was determined from the orthomosaic using the Digital Height Model (DHM) and the Digital Terrain Model (DTM) information. Data were collected monthly between June 2017 and May 2018.

The UAV measurements were compared to the field data to evaluate if UAVs were reliable substitutes for manual field labor for collecting tree data from the field. Classical statistical techniques were used to analyze the dataset for creating a prediction model. A feature selection process was used to rank all parameters (measurements) of the dataset. This ranking shows the effect of a specific parameter of a given month on the final yield prediction. The feature rankings provided the most important parameters and the most significant months for the yield prediction model. Multiple regression algorithms were tested to generate a model for yield prediction.

Fig. 1 presents the workflow of this study. The data collection step is presented in blue, which includes both manual (ground-truth) and UAV-based data collection. The second step of dataset analysis (in green) evaluates the UAV measurements compared to manual field measurements. Statistical analyses such as the mean error and the difference plots generate a comparison between both methodologies. The third step (in yellow) is the process to build a prediction model. The feature selection stage is an analysis made to identify which parameters and measurements of the crop are important for yield predictions. With those identified parameters, different algorithms are tested in their ability to generate accurate yield prediction models using the mean average percentage error (MAPE) score to compare the results (in red).

Study site

An area located on the Federal University of Lavras (23K 502906. 23m E, 7652838.84m S, 936 m altitude), in the state of Minas Gerais, Brazil, was chosen as the study site. The species Coffea arabica L. with Travessia cultivar was implanted in the area in February 2009 with planting spacing equal to 2.60×0.60 m (Fig. 2). The crop underwent pruning (skeleton) in July 2016. Productivity recovery is expected within an average time of two years after pruning (Silva et al., 2016). The climate of the region, according to the Köppen classification, is of the 'Cwa' type, characterized by a dry season in winter and a rainy season in summer [12]. A total of 144 plants (Fig. 2) were selected for this study according to the methodology described by Ferraz et al. [34].

UAV-based sensing system

Image collection was performed using a small UAV model DJI Phantom 3 professional (DJI, Shenzhen, China). The UAV was equipped with a digital RGB camera (Red-R, Green-G, Blue-B), Sony brand, model EXMOR 1 / 2.3 ", with a resolution of 4000×3000 pixels, a sensor size 6.16×6.62 mm, 94°FOV, and a sample rate of 0.5 frames per second equipped with an internal GPS receiver. The UAV control system included a remote controller and a ground control station connected to a smartphone device in which an application for flight planning and control was installed. This application collected information and photo parameters from the UAV during the mission that was later used by image processing software for generating the orthomosaic [57].

In this study, the Drone Deploy application (DroneDeploy, San Francisco, CA, USA) was used in all the missions. The parameters used for flight planning were: an altitude of 30 m, a speed of 3 m/s, and a frontal and lateral overlap of 80% between images. The overlap and flight speed were selected based on the study by Torres-Sánchez et al. [58], where the overlap was determined as a factor that interfered with the precision and quality of the final stitched image, the orthomosaic. The flight parameters entered in the application were programmed to be constant throughout the flight time. The images were georeferenced using the coordinates obtained by the UAV GPS at each waypoint. The return period for image collection was 30 days. The time of capture of the images was defined between 11:00 to 14:00 hours. Two people were involved in all the missions: the pilot responsible for taking off and landing the UAV,

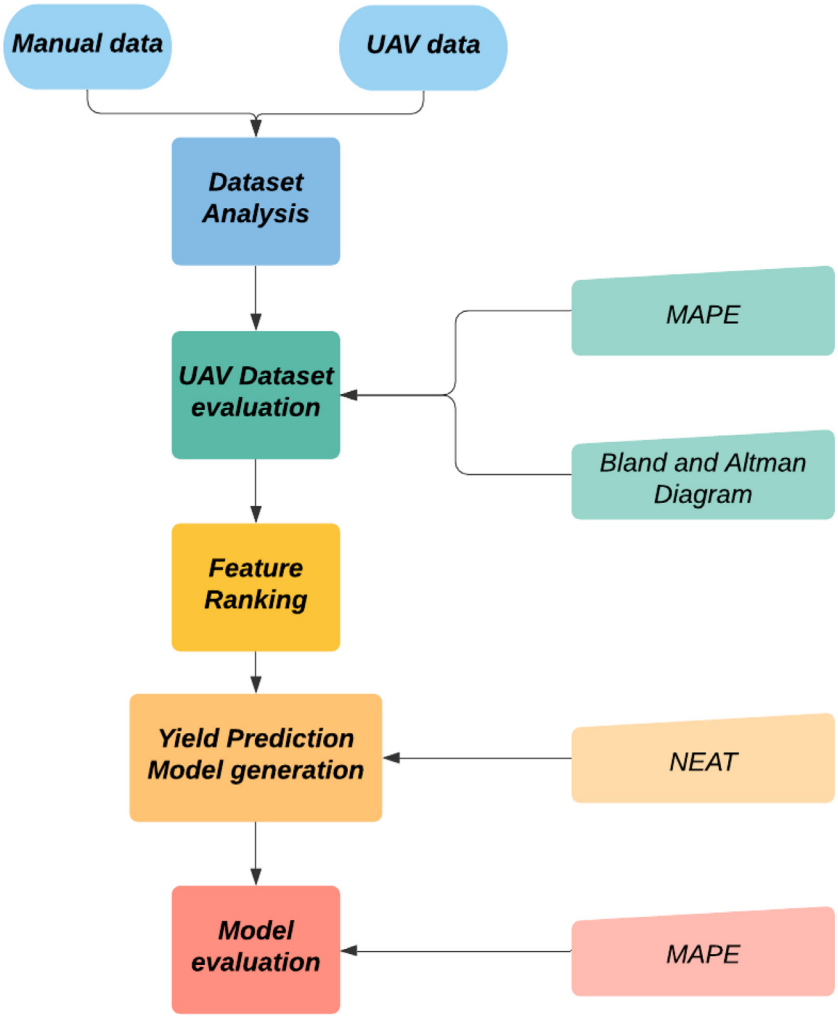

Fig. 1. Workflow of the study.

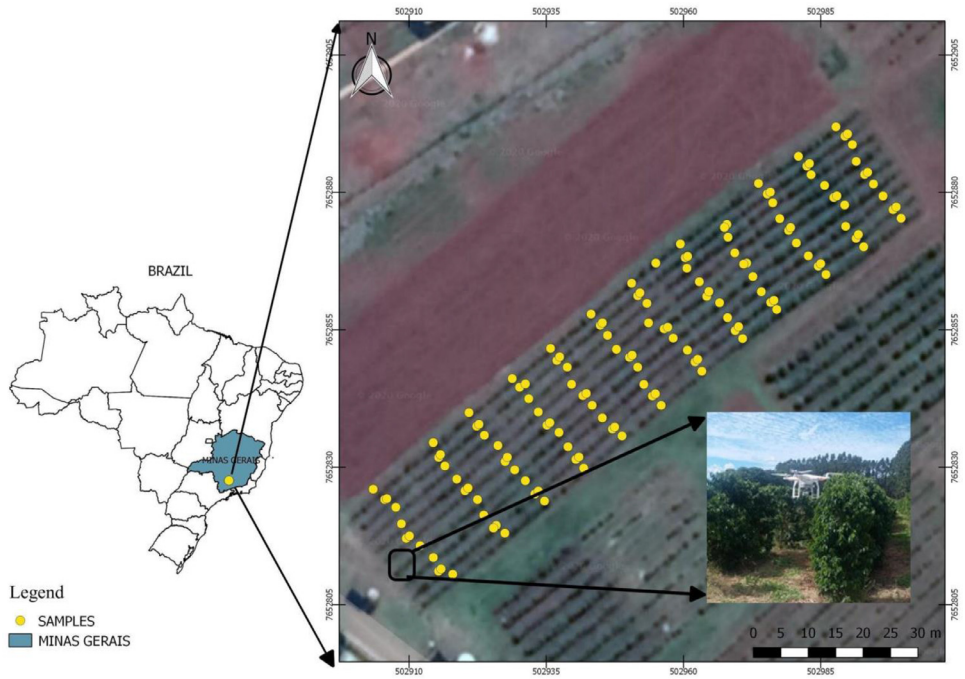

Fig. 2. Study site location with marked sampled trees.

and the observer, whose function was to alert the pilot about potential obstacles during the flight period.

Image processing

The images collected by the RGB camera were processed using Photoscan Professional software version 1.2.4 (Agisoft LLC, St. Petersburg, Russia). The geomatic products were obtained according to the methodology described by Jiménez-Brenes et al. (2019), which involved: i) generation of a three-dimensional (3D) point cloud to apply the Structure-from-Motion (SfM) technique; ii) generation of the digital terrain model (DTM) and digital surface model (DSM) from the 3D point cloud, using a maximum angle of 15° and a maximum distance of 0.1 m as parameters for a cell size of 40 m; and iii) creation of the orthomosaic of the area. 'High' quality and 'moderate' depth filtering were used as parameters in the software for calculating image position, orientation, correlation with neighboring images, and overlap. This ensured that the dense points cloud had greater accuracy, and outliers were removed [36].

For georeferencing of the images, six control points were fixed, with four at the ends and two in the central portion of the study area. The ground control points (GCP's) coordinates were obtained by a signal receiver that received signals from a global positioning differential system (DGPS; Trimble Navigation Limited, Sunnyvale, California, USA) with a horizontal and vertical accuracy of 0.007 m. These GCP's were manually identified in the images before the mosaic was made.

After the geometric correction of the images, the average spatial resolution of the orthomosaic in the evaluation period was 13 mm. The generated MDT, DSM, and orthomosaics were exported in Geotiff format to the geoprocessing software Quantum Gis ver. 2.16.3 (QGIS Development Team, Open Source Geospatial Foundation) in a GeoTiff file, in the Universal Transversal Mercator (UTM) projection, in the SIRGAS 2000 / UTM 23S zone.

Data collection

Using the methodology described by Ferraz et al. [34], 144 plants were selected for manual field data collection and ground-truthing. The trees were georeferenced using the same equipment described for the GCP's georeferencing. The plant height and crown diameter data were collected using a measuring tape. The tree height is the distance between the ground and the top of the tree, ignoring small branches as these are considered outliers. The diameter of the tree was measured in the middle third of each plant by averaging the diameter in two perpendicular axes (North-South and East-West), while also ignoring outlier branches. Outlier branches are small branches identified manually by the operator that grows outwards from the trend of the crop. The plant yield information was collected by manually harvesting each tree, and the volume was measured in a 20 L volume-measuring container.

The LAI was estimated based on the methodology described by Favarin et al. [32], as it is a fast and non-destructible methodology that can be used with the crown diameter (D) and plant height (H) (Eq. 1).

$$\mathrm{LAI} = 0.0134 + 0.7276 * D^2 * H \quad \text{(Eq. 1)}$$

The estimation of plant height from the UAV images was performed following the workflow described by Panagiotidis et al. [51] and Caruso et al. [19] (Eq. 2), where the difference between DSM and DTM estimated plant height. Each crown diameter (D) was estimated in the orthomosaic itself in the QGIS software at all sample points by manually selecting the crown bounding box and extracting the average diameter. The height values were extracted with the QGIS Point Sampling Tool to differentiate the DSM and DTM pixels.

$$\mathrm{H} = \mathrm{DSM} - \mathrm{DTM} \quad \text{(Eq. 2)}$$

Dataset analysis

Classical statistical techniques such as mean, minimum, maximum, standard deviation, skewness, and coefficient of variation were calculated for the yield dataset. Skewness is a measure of the symmetry of the data around its mean, where zero corresponds to a symmetric distribution. Negative skewness indicates that the data is skewed left (with the mean value less than the median). Positive skewness represents a clustered distribution on higher values and is skewed right (with the mean value higher than the median). The coefficient of variation is a measurement of the variance around the mean value.

Feature selection

Feature selection is a data preprocessing strategy that is effective and efficient in preparing datasets for multiple machine learning problems. The objectives of feature selection include building more straightforward and comprehensible models, improving data-mining performance, and preparing clean, understandable data [42]. Feature selection works by removing irrelevant and redundant data in a dataset.

Feature selection algorithms work by giving weights and ranking the best features (measurement) from a dataset that generates an output (prediction) of the target (yield). With these weights, it is possible to determine what features to keep for the model and what to ignore. This study's dataset includes monthly spectral data collected for coffee, and the feature selection algorithm provides the best month(s) and features for the yield prediction models. To perform the feature selection, either one or multiple algorithms can be used. This study used multiple algorithms, described below, that rank each feature and evaluate the resulting final ranking. The Pearson correlation coefficient [55] is a statistical measurement of the linear correlation between two variables. In this instance, for all the features of the model, the coefficient is calculated and used to rank their importance in a prediction. Spearman's rank correlation coefficient [29] is a nonparametric measurement of statistical dependence between the rankings of two variables. It assesses the relationship between two variables using a monotonic function. The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).

The F-test is a scoring function used in the feature selection procedures. This is done in two steps: i) the correlation between each feature and the target is computed as Eq. 3, where Xi is feature i, Y is the target and s is the standard deviation; ii) The correlations are converted to an F score value and ranked.

$$Correlation = (Xi - mean(Xi)) * (Y - mean(Y)) \,/\, (s(Xi) * s(Y)) \quad \text{(Eq. 3)}$$

Mutual Information (MI) is a powerful method for detecting relationships between data sets. It estimates mutual information for a continuous target variable. MI between two random variables is equal to zero if two random variables are independent, and higher values means higher dependency. The function relies on nonparametric methods based on entropy estimation from k-nearest neighbors distances described in Kraskov et al. [40] and Ross [54]. Mutual information methods can capture any kind of statistical dependency.

Recursive feature elimination with cross-validation (RFECV) is a recursive function based on eliminating features to rank their importance to an estimator (regression). It works by eliminating a feature from the dataset, running a regression algorithm, and evaluating the impact on the estimation. This effect is used to rank the importance of each feature. In this instance, the following regression algorithms are used: support vector machine [21], gradient boosting regression trees [35], and random forest regressor [15].

Table 1
Parameters used for each regression model and NEAT.

Regression Model	Parameters
SVM	Kernel : 'linear'
PLS	Number of components : 6, Maximum iterations : 500
Gradient boosting	Number of estimators : 500, Maximum depth : 4 Minimum samples split : 2, Learning rate : 0.01, Loss : 'ls'
Random forest regressor	Number of estimators = 80, Criterion : 'mae', Minimum samples split : 2
NEAT	Population size : 50, Elitism : 5, Maximum stagnation : 20, Maximum Generations : 500

Yield prediction models

For this study, multiple regression algorithms such as the linear SVM, partial least squares (PLS) regression, random forest, and gradient boosting regression were evaluated on their ability to generate a yield prediction model in such small datasets. All regression models were implemented using the Scikit-learn machine learning in Python library [53]. The used parameters are described in Table 1, with any other parameter not presented being the library's default values.

The linear support vector machine (SVM) algorithm segregates classes using a line or a hyper-plane. It classifies them by maximizing the distance between the nearest data points (margins) separated by the line or the hyperplane. A higher margin usually points to higher confidence in classification [16]. The random forest regression method is an ensemble-based learning method, where multiple weak base models are combined to generate an optimal ensemble model. Random forest regression runs efficiently on large datasets, is robust to outliers, and is less sensitive to overfitting [56]. A gradient boosting regression tree is also an ensemble-based method where base models are generated sequentially, and complex training cases are emphasized more to improve the prediction accuracy. PLS regression model links a dependent variable to a set of independent variables and can derive a sound and robust model from a large dataset [46]. A genetic algorithm called neuroevolution of augmenting topologies (NEAT) was implemented on the dataset. NEAT is an approach to artificial intelligence which uses both topology and weight parameters to evolve the artificial neural network [38]. NEAT was used to adjust the weight and topology of the artificial neural network (ANN), which generates the yield prediction model based on the UAV data.

Evaluation Metrics

The following evaluation metrics were used in this study: i) mean absolute percentage error (MAPE) and ii) measure of agreement. MAPE was used for comparison and evaluation of the yield prediction models. The measure of agreement was used for comparison of the UAV-based data to the field data to see if the UAV data could be a substitute to field measurements for the yield prediction models.

Table 2
Dataset analysis of the yield dataset.

Statistic	Value
Mean	3.23
Min-Max	0.5-8
Std. deviation	1.67
Skewness	0.62
CV	51.70%

Mean absolute percentage error (MAPE)

The mean absolute percentage error is a commonly used statistical tool. It is a measure of the accuracy of a forecast system. The MAPE is given by the average of the ratio of the absolute difference between the ground truth (Gt) and the prediction (P) to the ground truth. The formula for the MAPE is given below (Eq. 4), where n represents the number of individual items.

$$MAPE = \frac{1}{n}\Sigma\left|\frac{Gt - P}{Gt}\right| \quad \text{(Eq. 4)}$$

Measure of agreement

The agreement between measurements refers to the degree of agreement between two or more sets of measurements of a dataset by the same individual or two different individuals using similar methodologies. In this study, the difference plot was used to measure agreement instead of the Pearson correlation coefficient, which is often inappropriately used as a measure for agreement [59]. The difference plot, also known as Bland and Altman diagram [14], displays the pattern and agreement of one variable measured by two different methodologies [59]. The diagram plots the difference between a measurement pair (in our case, the difference between the UAV-based and field measurements) on the vertical axis and the pair's mean on the horizontal axis. To determine the repeatability of the proposed approach, the method assumes a normal distribution of differences, where 95% of them are expected to lie between $d \pm 1.96s$, where d is the mean of observed differences, and s is the standard deviation. This can be used as a range of error in application, where the top and bottom ranges define the limits to which to expect the measurement error to be included.

RESULTS

Dataset analysis

The yield dataset was analyzed for mean, minimum-maximum, standard deviation, skewness, and coefficient of variance (CV). Table 2 presents this analysis. The CV of 51.70% obtained represents a good variation for the dataset.

This variability in productivity may be associated with the coffee biennial production yield, which exhibits high and low values in alternated years [7]. The range between maximum and minimum values found can be explained by the fact that when the experiment was conducted, some plants located at the south end of the area did not show a recovery in their vegetative area, unlike the other plants in the northern region of the field.

UAV-based plant measurement evaluation

The pairs of observations between values measured with the UAV-based (height and diameter of the tree; collected monthly) and field measurements were compared with the twelve months data. The relative error of the data concerning each other was measured using minimum and maximum error, standard deviation, and MAPE. Table 3 shows the evaluation of the UAV-based measurements compared to the field measurements.

Table 3
Dataset analysis of the UAV-based data vs. field data.

Parameter	Min error	Max error	Standard Deviation	MAPE
Tree height	0.7%	63.42 %	4.42 %	5.88 %
Crown diameter	0.1%	72.61 %	7.43 %	6.83 %

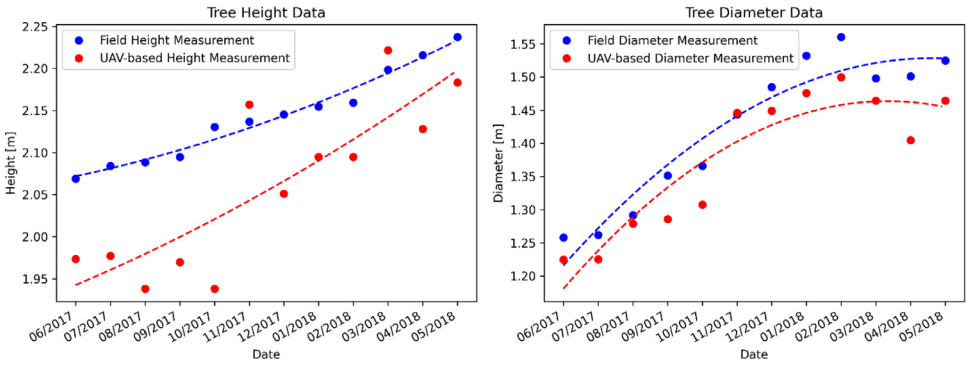

Fig. 3. Comparison of the monthly average UAV-based and field measurements for tree diameter and tree height for the 144 selected plants.

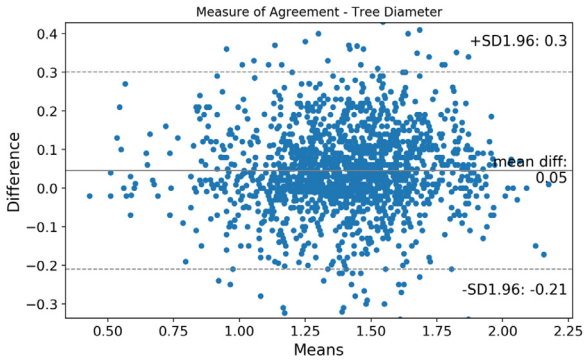

Fig. 4. Difference plot for tree diameter measurements.

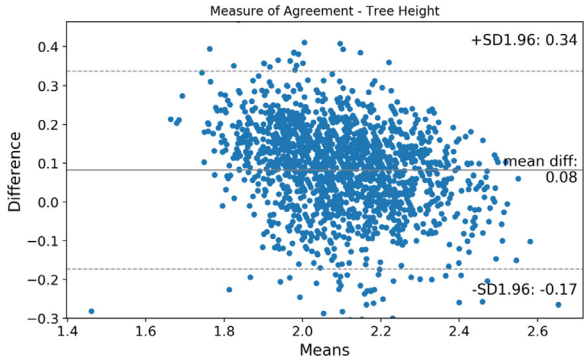

Fig. 5. Difference plot for tree height.

The maximum error for tree height was 63.42% and for tree diameter was 72.61%. Significant errors such as these may happen either during the stitching process of the maps due to a lack of key points while creating an orthomosaic, or due to errors in the 3D point cloud generation. However, the MAPE was under 6% and 7% for the tree height and diameter measurements, respectively. This means that the maximum error for both height and diameter could be safely considered mere outliers and not the norm.

Fig. 3 presents the comparison by month of the average values of height and diameter between the UAV-based and field measurements. For tree diameter, the maximum variation in measurement was seen in September (4.86%), October (4.27%), and April (6.42%), while every other month had less than 4% variation in the average values, with August showing the least variation in measurement (0.18%). Such low errors in diameter can be explained by minor errors in manual measurements, both in the field and virtual environments. Most variations were seen in July-October for the tree height measurement, with the maximum in October (9.05%). The small variation can explain such a low difference on field and UAV-based techniques for measuring the tree crown, such as small branches affecting either measurement. All other months had less than a 5% variation in the measurements. The comparison of the plots for the tree height and diameter showed more variation in the measurement of tree height (average variation of 4.21%) than in the measurement of tree diameter (average variation of 3.20%).

Using the difference plots presented in Figs. 4 and 5, the agreement between the UAV-based and field plant height and diameter values was determined. The difference of the values obtained through UAVs and manual field data was plotted on the Y-axis, while the mean of the data was plotted on the X-axis. The upper dashed line represents the upper limit of agreement given by $d + 1.96s$ and the lower dashed line represents the lower limit of agreement given by $d - 1.96s$, where d is the mean of observed differences and s is the standard deviation. The solid black line through the center represents d.

Fig. 4 presents the diagram for the tree diameter measurements. This showed an error of operation of +0.3 to -0.21 m. Since the manual in-field measurements could be considered perfect measurements, this is essentially the error expected from the UAV on this measurement.Fig. 5 presents the difference plot for the tree height measurements. It shows an upper limit of error of +0.34 and a bottom limit of -0.17 m.Field measurements can have similar errors for height and diameter. For manual in-field plant height measurement, there is an associated error with the measuring tape inclination to the soil surface and the vertical axis of the plant. For the plant diameter measurements, this factor may also be associated with the plant's topology (branches and leaves) that did not allow ideal parallelism during measurement for some sample points.

Considering the ranges of operation derived from the difference plots (Figs. 4 and 5), it was assumed that the UAV-based measurements were reliable enough to measure tree height and diameter. An expected error for tree height between -0.17 m to +0.34 m for crops with heights around 1.97 m is an acceptable error for large scale measurements. Similarly for crown diameter, with average values of 1.28 m presenting errors ranging from -0.21 m to +0.30 m, we can assume that these are acceptable error ranges. Assuming that these measurements were reli-

Table 4
Top 10 features to estimate yield in coffee.

Components	Final rank	Pearson	Spearman	F-test	Mutual info	SVM	Gradient boosting	Random rorest
LAI (12/2017)	1	1	1	1	7	1	1	1
LAI (04/2018)	2	2	2	2	3	10	1	1
Diameter (12/2017)	3	3	3	3	8	11	1	1
Diameter (04/2018)	4	8	6	8	10	13	1	1
Height (07/2017)	5	16	18	16	4	1	1	1
LAI (01/2018)	6	4	5	4	1	31	1	1
LAI (05/2018)	7	5	4	5	6	20	1	1
LAI (10/2017)	8	7	8	7	13	25	3	2
LAI (03/2018)	9	6	7	6	5	19	6	1
LAI (11/2017)	10	9	14	9	22	22	1	1

able, we can substitute a slow and laborious manual task in the field with a fast and precise tool.

Feature selection and ranking

The feature rankings for the yield prediction models were determined for the LAI, tree diameter, tree height, and the individual bands of the RGB data. This was done for all the months under consideration. Seven different algorithms - Pearson, Spearman, F-test, Mutual Info, SVM, Gradient Boosting, and Random Forest - were used to determine the rankings, and then the cumulative effect was calculated to determine the final rankings. The final rankings show the effect of a specific parameter of a given month on the final yield prediction. The higher the ranking of a component, the higher is its effect on the yield prediction model. Table 4 shows the top 10 components ranked based on their performance when tested using the seven algorithms. The most significant parameters in terms of the effect on yield prediction were the LAI and the tree diameter. The LAI of two months - December and April - had the most effect on the yield prediction model. The tree height showed up only once in the top 10 rankings, with the effect of July more prominent than the other months, but LAI was the most dominant feature considering 70% of the top 10 was made up of LAI. During December and April, the tree's diameter was also a part of the top 10 features, making the months of December and April highly important months in terms of the yield prediction contribution. The rankings of the RGB bands were also calculated in these rankings. It was observed that because of their low rankings (they occupied the lower half of the rankings table), their significance in the yield prediction model was very little.

With the ranking values, weights were generated as the inverse of the sum of the rankings for each feature. Fig. 6 presents these weights by month. The rankings of RGB were summed together. December 2017 and April 2018 were determined to be the most prominent months for this dataset.

Yield prediction models

After the feature selection step, the chosen regressors were used to predict the yield. We evaluated the use of the whole dataset and the features from the most critical months based on the feature ranking. The models were evaluated in a 5-fold cross-validation setting, and the overall MAPE was calculated. Table 5 presents the MAPE for the models using different feature selections.

The GBR, RFR, and PLSR models had similar results using different numbers of features that could be explained by these algorithms' ability to work with high-dimensionality data and small datasets. The MAPE values for the Linear SVM and NEAT-based yield prediction models showed that reducing the number of features improves the prediction. The NEAT algorithm, being a convergent genetic algorithm, did not converge in the entire dataset. It only started converging correctly from the top 3 months' data. The graph of Feature Contribution can explain the small differences in the MAPE between the top 3, top 2 months, and topmost month by month (Fig. 6), where the topmost month carried almost double the importance of the second most crucial month. This means that using data from 3 months or 1 month has small differences that can be neglected for this dataset.

Discussion

The feature selection process is an important step for optimizing data collection procedures, especially when involving laborious tasks such as field measurements. These feature rankings improve this process by highlighting important dates and variables/features, while also being beneficial to complex regression algorithms. A regression algorithm can become unstable and lose precision with a dataset containing a large number of features, as some least important information can confuse the model's training stage.

This study shows that even though tree height and diameter is fed into the algorithms, the LAI was more important feature in predicting coffee yield. This is a clear showcase of how adding a new feature to the dataset from the data itself can improve a regression. The results presented demonstrate that the LAI, although an equation that includes tree height and diameter, presents a better feature for the prediction model than both the other measurements.

The results described in Table 4 showed that the LAI feature was ranked first in the performance of the yield prediction model. Results found in the study by Chu et al. [24] reinforce the potential of using biophysical parameters of cotton derived from RGB images coupled to UAV to predict yield. The ranking of variables (Table 4 and Fig. 6) showed the more promising dates for estimating coffee productivity by UAV imagery being December and April. These results show that it is possible to reduce the monitoring time of the crop during the production year. According to Camargo & Camargo [17], the phenological cycle of Coffee arabica can explain these results. This study showed that in the second phenological year (evaluation period of this study), the month of December marks the end of the third phase of this cycle: the flowering and the beginning of the fruit granulation phase. For the month of April, the fifth and last phase of the cycle, the fruit matures, ending in June (harvest date). The results obtained by Aparecido and Rolim [7] enhance those found in this study (Table 4), where the months of December (flowering) and April (fruit granulation) are decisive times in coffee productivity.

The accuracy of the prediction models presented, with MAPE of 31.75%, showcases the possible application of this methodology. The process of feature selection to identify the best predictors achieved good results for yield prediction tasks, as it can be further enhanced. A further study using more collected data or using other sensors such as multi-

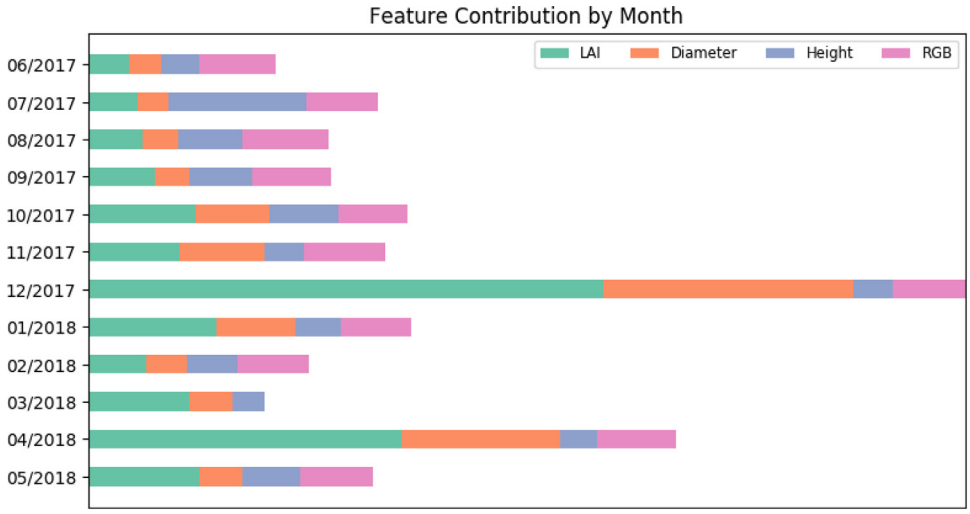

Fig. 6. Feature weights for the yield prediction model based on ranking for all months of data collected.

Table 5
The models MAPE for yield prediction are based on different feature selections.

	MAPE				
Algorithm*	**All months**	**Top 4 months**	**Top 3 months**	**Top 2 months**	**Top 1 month**
GBR	36.79%	37.83%	36.99%	35.71%	37.39%
Linear SVM	47.68%	37.76%	35.79%	35.05%	32.99%
RFR	36.13%	37.83%	35.05%	34.07%	37.13%
PLSR	34.80%	35.84%	36.63%	35.92%	33.48%
NEAT	100.00%	56.86%	31.91%	32.18%	**31.75%**

* GBR: Gradient Boosting Regression, Linear SVM: Linear Support Vector Machine, RFR: Random Forest Regression, PLSR: Partial Least Square Regression

spectral cameras can use the presented methodology to achieve higher prediction accuracies.

Conclusion

A methodology was proposed for estimating coffee productivity (yield) by applying machine learning techniques on the data (RGB images) obtained from UAVs. The RGB data were used to estimate tree height and crown diameter. Then, the crown diameter and the tree height were used to estimate the LAI. Seven different regression algorithms were used to select the best feature(s) out of LAI, tree height, crown diameter, and the RGB values to determine the feature that had the maximum influence on the yield prediction for all the months under consideration. Using the feature selection, it was determined that the LAI and the crown diameter were the most dominant features. The LAI, in particular, was the most dominant feature, contributing to 70% of the top 10 feature rankings. The other important aspect of the feature ranking was the importance of two months, December 2017 and April 2018, to the yield prediction. Five regression algorithms were used to generate the yield prediction models, and MAPE was used as the evaluation parameter for these algorithms. The models were used to predict yield using the whole dataset and the most critical months, which were determined from the feature rankings. For most of the regression algorithms used, reducing the number of features to include primarily the most important features instead of the whole dataset improved the MAPE and hence, the yield prediction. Regression algorithms decide the weight of each feature for estimating yield at the training step. The feature selection and thus removing low descriptive features from the data helps the model converge to the solution with fewer data and achieve higher precision. Although the NEAT model achieved the best results, all algorithms used achieved under 40% MAPE when using the same number of features. But it was also observed that some models retained similar errors for all features, showing that the feature selection improvement varies per algorithm. It also suggested that since the difference in MAPE for the top month vs. the top three months was very little, a dataset of parameters collected during just one month could be used satisfactorily for yield prediction. This is a significant result for future studies because it reduces the need for extensive year-round data collection and allows researchers to focus on the dominant parameters of certain most important months. Although the results obtained in this study show promise, there are still opportunities for improvement. Future studies could use spectral data collected from multispectral and hyperspectral sensors, and add vegetation indices in the feature selection process.

Declaration of Competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Abdulridha, Y. Ampatzidis, J. Qureshi, P. Roberts, Laboratory and UAV-based identification and classification of tomato yellow leaf curl, bacterial spot, and target spot diseases in tomato utilizing hyperspectral imaging and machine learning, Remote Sensing 12 (17) (2020) 2732, doi:10.3390/rs12172732.
[2] J. Abdulridha, Y. Ampatzidis, S.C. Kakarla, P. Roberts, Detection of target spot and bacterial spot diseases in tomato using UAV-based and benchtop-based hyperspectral imaging techniques, Precision Agriculture 21 (2020) 955–978 doi.org/10.1007/s11119-019-09703-4.
[3] J. Abdulridha, Y. Ampatzidis, P. Roberts, S.C. Kakarla, Detecting powdery mildew disease in squash at different stages using UAV-based hyperspectral imaging and artificial intelligence, Biosystems Eng. 197 (2020) 135–148 doi.org/10.1016/j.biosystemseng.2020.07.001.
[4] Y. Ampatzidis, V. Partel, L. Costa, Agroview: Cloud-based application to process, analyze and visualize UAV-collected data for precision agriculture applications utilizing artificial intelligence, Comput. Electron. Agric. 174 (2020) 105457.
[5] Y. Ampatzidis, V. Partel, B. Meyering, U. Albrecht, Citrus rootstock evaluation utilizing UAV-based remote sensing and artificial intelligence, Comput. Electron. Agric. 164 (2019) 104900, doi:10.1016/j.compag.2019.104900.

[6] Y. Ampatzidis, V. Partel, UAV-based high throughput phenotyping in citrus utilizing multispectral imaging and artificial intelligence, Remote Sensing 11 (4) (2019) 410, doi:10.3390/rs11040410.

[7] Lucas Eduardo de Oliveira Aparecido, Glauco de Souza Rolim, Forecasting of the annual yield of Arabic coffee using water deficiency, Pesquisa Agropecuária Brasileira 53 (12) (2018) 1299–1310, doi:10.1590/s0100-204X2018001200002.

[8] G. Araújo e Silva Ferraz, F.M. da Silva, M. de Carvalho Alves, R.F. Bueno, P.A.N. da Costa, Geostatistical analysis of fruit yield and detachment force in coffee, Precision Agriculture 13 (2012) 76–89, doi:10.1007/s11119-011-9223-8.

[9] R. Ballesteros, J.F. Ortega, D. Hernandez, M.A. Moreno, Onion biomass monitoring using UAV-based RGB imaging, Precision agriculture 19 (5) (2018) 840–857, doi:10.1007/s11119-018-9560-y.

[10] J.G.A. Barbedo, A review on the main challenges in automatic plant disease identification based on visible range images, Biosystems Eng. 144 (2016) 52–60, doi:10.1016/j.biosystemseng.2016.01.017.

[11] J.G.A. Barbedo, L.V. Koenigkan, T.T. Santos, Identifying multiple plant diseases using digital image processing, Biosystems Eng. 147 (2016) 104–116, doi:10.1016/j.biosystemseng.2016.03.012.

[12] J. Barbosa, E.A. Pozza, P.E. de Souza, M.D.L. Oliveira e Silva, A.A.A. Pozza, R.J. Guimarães, Irrigation drip and phosphorus managements in the rust coffee progress, Coffee Sci. 12 (2) (2017) 187–196 doi: doi:10.25186/cs.v12i2.1214.

[14] J.M. Bland, D.G. Altman, Measuring agreement in method comparison studies, Stat. Methods Med. Res. 8 (2) (1999) 135–160.

[15] Breiman, L. (2001). Random forests. Machine learning, 45(1), pp.5-32. https://doi.org/10.1023/A:1010933404324.

[16] D. Bzdok, M. Krzywinski, N. Altman, Machine learning: supervised methods, 2018.

[17] Â.P.D. Camargo, M.B.P.D. Camargo, Definition and schematization of the phenological phases of Arabica coffee in tropical conditions in Brazil, Bragantia 60 (1) (2001) 65–68 http://dx.doi.org/10.1590/S0006-87052001000100008.

[18] G.L. Carrijo, D.E. Oliveira, G.A. de Assis, M.G. Carneiro, V.C. Guizilini, J.R. Souza, Automatic detection of fruits in coffee crops from aerial images, in: Robotics Symposium (LARS) and 2017 Brazilian Symposium on Robotics (SBR), 2017 Latin American, IEEE, 2017, pp. 1–6. doi.org/10.1109/SBR-LARS-R.2017.8215283.

[19] G. Caruso, P.J. Zarco-Tejada, V. González-Dugo, M. Moriondo, L. Tozzini, G. Palai, G. Rallo, A. Hornero, J. Primicerio, R. Gucci, High-resolution imagery acquired from an unmanned platform to estimate biophysical and geometrical parameters of olive trees under different irrigation regimes, PLoS One 14 (1) (2019) e0210804, doi:10.1371/journal.pone.0210804.

[20] F.H. Carvalho, G. Galli, L.F.V. Ferrão, J.V.A. Nonato, L. Padilha, M.P. Maluf, M.F.R. Resende Júnior, O.G. Filho, R. Fritsche-Neto, The effect of bienniality on genomic prediction of yield in arabica coffee, Euphytica 216 (2020) 101, doi:10.1007/s10681-020-02641-7.

[21] Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27. https://doi.org/10.1145/1961189.1961199

[22] A. Chemura, O. Mutanga, J. Odindi, D. Kutywayo, Mapping spatial variability of foliar nitrogen in coffee (Coffea arabica L.) plantations with multispectral Sentinel-2 MSI data, ISPRS J. Photogramm. Remote Sens. 138 (2018) 1–11.

[23] A. Chemura, O. Mutanga, M. Sibanda, P. Chidoko, Machine learning prediction of coffee rust severity on leaves using spectroradiometer data, Tropical Plant Pathol. 43 (2) (2018) 117–127.

[24] T. Chu, R. Chen, J.A. Landivar, M.M. Maeda, C. Yang, M.J. Starek, Cotton growth modeling and assessment using unmanned aircraft system visual-band imagery, J. Appl. Remote Sens. 10 (3) (2016) 036018, doi:10.1117/1.JRS.10.036018.

[25] Abastecimento CONAB- Companhia Nacional de, Acompanhamento da Safra Brasileira de café. Monitoramento agrícola (Monitoring of the Brazilian coffee crop: Agricultural monitoring) 6– Safra 2020, Brasília, 2020 - First survey.

[26] L. Costa, L. Nunes, Y. Ampatzidis, A new visible band index (vNDVI) for estimating NDVI values on RGB images utilizing genetic algorithms, Comput. Electron. Agric. 172 (2020) 105334 doi.org/, doi:10.1016/j.compag.2020.105334.

[27] J.P. da Cunha, S. Neto, A. Matheus, S. Hurtado, Estimating vegetation volume of coffee crops using images from unmanned aerial vehicles, Engenharia Agrícola 39 (SPE) (2019) 41–47 doi.org/10.1590/1809-4430-eng.agric.v39nep41-47/2019.

[28] L.E. de Oliveira Aparecido, G. de Souza Rolim, R.A. Camargo Lamparelli, P.S. de Souza, E.R. dos Santos, Agrometeorological models for forecasting coffee yield, Agron. J. 109 (1) (2017) 249–258 doi.org/, doi:10.2134/agronj2016.03.0166.

[29] J.C. de Winter, S.D. Gosling, J. Potter, Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data, Psychol. Methods 21 (3) (2016) 273 https://doi.org/10.1037/met0000 https://doi.org/10.1037/met0000079 079.

[30] L. Deng, Z. Mao, X. Li, Z. Hu, F. Duan, Y. Yan, UAV-based multispectral remote sensing for precision agriculture: A comparison between different cameras, ISPRS J. Photogramm. Remote Sens. 146 (2018) 124–136.

[31] L.M. dos Santos, G.A.E.S. Ferraz, B.D.D.S. Barbosa, A.V. Diotto, D.T. Maciel, L.A.G. Xavier, Biophysical parameters of coffee crop estimated by UAV RGB images, Precision Agriculture (2020) 1–15, doi:10.1007/s11119-020-09716-4.

[32] J.L. Favarin, Dourado Neto, García y García D, Villa Nova A, A. N, M.D.G.G.V Favarin, Equations for estimating the coffee leaf area index, Pesquisa Agropecuária Brasileira 37 (6) (2002) 769–773, doi:10.1590/S0100-204X2002000600005.

[33] L. Feng, Z. Zhang, Y. Ma, Q. Du, P. Williams, J. Drewry, B. Luck, Alfalfa Yield Prediction Using UAV-Based Hyperspectral Imagery and Ensemble Learning, Remote Sensing 12 (12) (2020) 2020 doi.org/10.3390/rs12122028.

[34] G.A. Ferraz, F.M.D. Silva, M.S.D. Oliveira, A.A.P. Custódio, P.F.P. Ferraz, Spatial variability of plant attributes of a coffee crop, Revista Ciência Agronômica 48 (1) (2017) 81–91 http://dx.doi.org/10.5935/1806-6690.20170009.

[35] J.H. Friedman, Stochastic gradient boosting, Comput. stats. data anal. 38 (4) (2002) 367–378.

[36] M. Hobart, M. Pflanz, C. Weltzien, M. Schirrmann, Growth Height Determination of Tree Walls for Precise Monitoring in Apple Fruit Production Using UAV Photogrammetry, Remote Sensing 12 (10) (2020) 1656, doi:10.3390/rs12101656.

[37] F.H. Holman, A.B. Riche, A. Michalski, M. Castle, M.J. Wooster, M.J. Hawkesford, High throughput field phenotyping of wheat plant height and growth rate in field plot trials using UAV based remote sensing, Remote Sensing 8 (12) (2016) 1031.

[38] M.Y. Ibrahim, R. Sridhar, T.V. Geetha, S.S. Deepika, Advances in Neuroevolution through Augmenting Topologies–A Case Study, in: 2019 11th International Conference on Advanced Computing (ICoAC), IEEE, 2019, pp. 111–116.

[39] T. Kharel, S. Swink, C. Youngerman, A. Maresma, K.J. Czymmek, Q.M. Ketterings, P. Kyveryga, J. Lory, T.A. Musket, V. Hubbard, Processing/cleaning corn silage and grain yield monitor data for standardized yield maps across farms, fields, and years, Cornell University, Nutrient Management Spear Program, Department of Animal Science, Ithaca NY, 2018.

[40] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (6) (2004) 066138, doi:10.1103/PhysRevE.69.066138.

[41] L. Kouadio, R.C. Deo, V. Byrareddy, J.F. Adamowski, S. Mushtaq, Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties, Comput. Electron. Agric. 155 (2018) 324–338.

[42] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: A data perspective, ACM Comput. Surveys (CSUR) 50 (6) (2017) 1–45, doi:10.1145/3136625.

[44] M. Maimaitijiang, V. Sagan, P. Sidike, S. Hartling, F. Esposito, F.B. Fritschi, Soybean yield prediction from UAV using multimodal data fusion and deep learning, Remote Sens. Environ. 237 (2020) 111599 doi.org/, doi:10.1016/j.rse.2019.111599.

[45] A. Maresma, L. Chamberlain, A. Tagarakis, T. Kharel, G. Godwin, K.J. Czymmek, E. Shields, Q.M. Ketterings, Accuracy of NDVI-derived corn yield predictions is impacted by time of sensing, Comput. Electron. Agric. 169 (2020) 105236.

[46] K. Meacham-Hensold, C.M. Montes, J. Wu, K. Guan, P. Fu, E.A. Ainsworth, T. Pederson, C.E. Moore, K.L. Brown, C. Raines, C.J. Bernacchi, High-throughput field phenotyping using hyperspectral reflectance and partial least squares regression (PLSR) reveals genetic modifications to photosynthetic capacity, Remote Sens. Environ. 231 (2019) 111176.

[47] A.J. Oliveira, G.A. Assis, V. Guizilini, E.R. Faria, J.R. Souza, Segmenting and Detecting Nematode in Coffee Crops Using Aerial Images, in: International Conference on Computer Vision Systems, Springer, Cham, 2019, pp. 274–283. doi.org/10.1007/978-3-030-34995-0_25.

[48] Oliveira, H. C., Guizilini, V. C., Nunes, I. P., & Souza, J. R. (2018). Failure detection in row crops from UAV images using morphological operators. IEEEGeoscience and Remote Sensing Letters, 15(7), 991-995. doi.org/ 10.1109/LGRS.2018.2819944.

[49] D. Olson, A. Chatterjee, D.W. Franzen, Can We Select Sugarbeet Harvesting Dates Using Drone-based Vegetation Indices? Agron. J. 111 (5) (2019) 2619–2624.

[51] D. Panagiotidis, A. Abdollahnejad, P. Surový, V. Chiteculo, Determining tree height and crown diameter from high-resolution UAV imagery, Int. J. Remote Sens. 38 (8-10) (2017) 2392–2410, doi:10.1080/01431161.2016.1264028.

[52] V. Partel, L. Nunes, P. Stansly, Y. Ampatzidis, Automated vision-based system for monitoring Asian citrus psyllid in orchards utilizing artificial intelligence, Comput. Electron. Agric. 162 (2019) 328–336.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, Scikit-learn: Machine learning in Python. the, J. mach. Learning res. 12 (2011) 2825–2830.

[54] B.C. Ross, Mutual information between discrete and continuous data sets, PLoS One 9 (2) (2014) e87357, doi:10.1371/journal.pone.0087357.

[55] P. Schober, C. Boer, L.A. Schwarte, Correlation coefficients: appropriate use and interpretation, Anesthesia & Analgesia 126 (5) (2018) 1763–1768, doi:10.1213/ANE.0000000000002864.

[56] B. Singh, P. Sihag, K. Singh, Modelling of impact of water quality on infiltration rate of soil by random forest regression, Modeling Earth Syst. Environ. 3 (3) (2017) 999–1004.

[57] J. Torres-Sánchez, F. López-Granados, I. Borra-Serrano, J.M. Peña, Assessing UAV–collected image overlap influence on computation time and digital surface model accuracy in olive orchards, Precision Agric. 19 (1) (2018) 115–133.

[58] J. Torres-Sánchez, F. López-Granados, N. Serrano, O. Arquero, J.M. Peña, High--throughput 3-D monitoring of agricultural-tree plantations with unmanned aerial vehicle (UAV) technology, PLoS One 10 (6) (2015) e0130479.

[59] P.F. Watson, A. Petrie, Method agreement analysis: a review of correct methodology, Theriogenology 73 (9) (2010) 1167–1179.