Computers and Electronics in Agriculture 178 (2020) 105791

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices

Ana Paula Marques Ramos[a,*], Lucas Prado Osco[b], Danielle Elis Garcia Furuya[a], Wesley Nunes Gonçalves[c,h], Dthenifer Cordeiro Santana[d], Larissa Pereira Ribeiro Teodoro[e], Carlos Antonio da Silva Junior[g], Guilherme Fernando Capristo-Silva[i], Jonathan Li[j], Fábio Henrique Rojo Baio[e], José Marcato Junior[c], Paulo Eduardo Teodoro[e], Hemerson Pistori[f,h]

[a] Postgraduate Program of Environment and Regional Development, University of Western São Paulo, Rodovia Raposo Tavares, km 572 – Limoeiro, 19067-175, Presidente Prudente, São Paulo, Brazil
[b] Faculty of Engineering and Architecture and Urbanism, University of Western São Paulo, Rodovia Raposo Tavares, km 572 – Limoeiro, 19067-175, Presidente Prudente, São Paulo, Brazil
[c] Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul, Cidade Universitária, Av. Costa e Silva – Pioneiros, MS 79070-900, Brazil
[d] State University of Mato Grosso do Sul, Rodovia Graziela Maciel Barroso, Km 12, Zona Rural, Aquidauana, MS 79200-000, Brazil
[e] Department of Agronomy, Federal University of Mato Grosso do Sul, Rodovia MS 306, km. 305, Caixa Postal 112, 79560000 – Chapadão do Sul, MS, Brazil
[f] Inovisão, Universidade Católica Dom Bosco, Av. Tamandaré, 6000, Campo Grande, MS 79117-900, Brazil
[g] Department of Geography, State University of Mato Grosso, Av. dos Ingas, 3001 – Jardim Imperial, Sinop – MT 78555-000, Brazil
[h] Faculty of Computing, Federal University of Mato Grosso do Sul, Cidade Universitária, Av. Costa e Silva – Pioneiros, MS 79070-900, Brazil
[i] Federal University of Mato Grosso, Postgraduate Program in Agronomy, Sinop, Mato Grosso, Brazil
[j] Department of Geography and Environmental Management and Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada

ARTICLE INFO

Keywords:
Precision agriculture
Multispectral images
Shallow learner
Random Forest

ABSTRACT

Random Forest (RF) is a machine learning technique that has been proved to be highly accurate in several agricultural applications. However, to yield prediction, how much this technique may be improved with the adoption of a ranking-based strategy is still an unknown issue. Here we propose a ranking-based approach to potentialize the RF method for maize yield prediction. This approach is based on the correlation parameter of individual vegetation indices (VIs). The VIs were individually ranked based on a merit metric that measures the improvement on the Pearson's correlation coefficient by using RF against a baseline method. As a result, only the most relevant VIs were considered as input features to the RF model. We used 33 VIs extracted from multispectral UAV-based (unmanned aerial vehicle) imagery. The multispectral data were generated with two different sensors: Sequoia and MicaSense; during the 2017/2018 and 2018/2019 crop seasons, respectively. Amongst all the evaluated indices, NDVI, NDRE, and GNDVI were the top three in the ranking-based analysis, and their combination with RF increased the maize yield prediction. Our approach also outperformed other known machine learning methods, like support vector machine and artificial neural network. Additive regression, using the RF as the base weak learner, provided a higher accuracy with a correlation coefficient and MAE (Mean Absolute Error) of 0.78 and 853.11 kg ha^{-1}, respectively. We conclude that the ranking-based strategy of VIs is appropriate to predict maize yield using machine learning methods and data derived from multispectral images. We demonstrated that our approach reduces the number of VIs needed to determine a high accuracy and relative low MAE, and the approach may contribute to decision-making actions, resulting in accurate management of maize fields.

* Corresponding author.
E-mail addresses: anaramos@unoeste.br (A.P. Marques Ramos), lucasosco@unoeste.br (L. Prado Osco), wesley.goncalves@ufms.br (W. Nunes Gonçalves), larissa_ribeiro@ufms.br (L. Pereira Ribeiro Teodoro), carlosjr@unemat.br (C. Antonio da Silva Junior), junli@uwaterloo.ca (J. Li), fabiobaio@ufms.br (F. Henrique Rojo Baio), jose.marcato@ufms.br (J. Marcato Junior), paulo.teodoro@ufms.br (P. Eduardo Teodoro), pistori@ucdb.br (H. Pistori).

https://doi.org/10.1016/j.compag.2020.105791
Received 13 July 2020; Received in revised form 30 July 2020; Accepted 12 September 2020

1. Introduction

Machine learning techniques associated with UAV-based (unmanned aerial vehicle) images are been continuously applied for different areas, including precision agriculture. A critical topic in agricultural applications refers to crop yield prediction due to its dependence on multiple factors, such as crop genotype, environmental factors, and management practices (Khaki et al. 2020). Adopting machine learning methods to support this task is currently a promising strategy since information derived from analyses may support farmers to decide, for example, which crop management provides the maximum yield. This could be achieved by considering multiple factors, like temperature, rainfall, area, among others.

Yield prediction with machine learning techniques is a recent topic in literature, and was considered for multiple cultivars, like a cherry tree (Amatya et al., 2016); sugar-cane (Vani et al. 2015); wheat (Jeong et al. 2016; Pantazi et al. 2016; Hunt et al., 2019); potato (Jeong et al. 2016); tomatoes (Senthilnath et al., 2016); coffee (Ramos et al. 2017); rice (Su et al. 2017; Zhang et al., 2019); groundnut (Shah and Shah, 2018). In this regard, maize (*Zea mays* L.) is an important economic crop for multiple countries that could also benefit from this type of approach. On a world scale, Brazil is the third market producer preceded only by United States (US) and China (Conab, 2020). In Brazil, the states of Mato Grosso do Sul, Paraná, and Goiás are responsible for most of the maize-production rates. For the 2019/2020 crop season, the estimated rate is to achieve record production of 100.6 million tons (Conab, 2020). Technological solutions may contribute to improving yield rates, and the artificial intelligence area, combined with remote sensing data, has an essential role in this context.

Performing grain-yield prediction with only the spectral information of a plant is a challenging scientific task. Related to maize-crops, the existing literature presents some information about the yield estimation with machine learning techniques and remote sensing imagery (Serele et al., 2000; Uno et al. 2005; Khanal et al. 2018). One research (Serele et al., 2000) tested four ANN (Artificial Neural Networks) models in aerial images with a spatial resolution of 1.5 m to predict maize yield. The input variables of the ANN models consisted of the topographic data (elevation, slope, and aspect), vegetation indices - VIs (NDVI - Normalized Difference Vegetation Index, SAVI - Soil Adjusted Vegetation Index, TSARV - Transformed Soil Atmospherically Resistant Vegetation Index, and WDVf - Weight Difference Vegetation Index), and textural indices - TIs (homogeneity, contrast, entropy, and ASM - Angular Second Moment). When combined with VIs, TIs, and topographic data, the investigated ANN models presented better accuracy to predict maize yield.

Another related study (Uno et al. 2005), adopting hyperspectral images (spectral range of 408–947 nm and spatial resolution of 2 m), acquired with a Compact Airborne Spectrographic Imager, estimated maize yield prediction using machine learning techniques with multispectral data. The authors explored three VIs (NDVI, SR - Sample Ratio, and PRI - Photochemical Reflectanceindex), and determined that ANN models were efficient in capturing the complex relationship between crop yield and spectral reflectance values. A more recent research (Khanal et al. 2018) integrated five field-based soil properties (soil organic matter, cation exchange capacity, magnesium, potassium, and pH) with multispectral aerial images and topographic data (digital elevation model) to predict soil properties and maize yield applying multiple machine learning algorithms (Random Forest; Neural Network; Support Vector Machine with Radial and Linear Kernel Functions; Gradient Boosting Model; and Cubist). For maize yield, they determined that Random Forest (RF) consistently outperformed other models.

In recent years, machine learning techniques such as deep neural networks (known as deep learning) have also been applied to estimate crop yield (Nevavuori et al. 2019; Barbosa et al. 2020; Khaki et al. 2020). Even so, it should be highlighted that these approaches required an extensive number of datasets to return high performances on the aforementioned papers. Besides, deep learning can be a high-computational cost technique, which may not be the best alternative to be implemented in some of the agriculture applications. A recent study (Barbosa et al. 2020), aiming to estimate maize yield production based on deep learning approach, adopted five input variables (nitrogen rate, seed rate, elevation map, soil's electroconductivity, and the NDVI index), and the results were compared with shallow machine learning methods. Although it was verified a reduction in the RMSE (Root Mean Square Error) up to 29%, compared to the Random Forest, it should be mention that a total of 1800 plots was necessary to define the deep learning model, while related research (Sérélé et al. 2003; Uno et al. 2005; Khanal et al. 2018) used fewer data to perform the same task.

A revision study (Liakos et al. 2018) on applications of machine learning in agricultural production systems concluded that, by applying machine learning to sensor data, farm management systems provide rich recommendations and insights for farmer decision support and action. Another revision study (Chlingaryan et al., 2018), specifically about machine learning approaches for crop yield prediction, concludes that the rapid advances in sensing technologies and artificial intelligence techniques will provide a cost-effective, efficient and comprehensive solution for better crop management and decision-making tasks shortly. As mentioned, yield inference using only remote sensing data is a challenging task. Offering farmers both novel and low-cost computational and simpler in-field data collection applications is still a high topic in the current literature. A possible approach to this is to get the maximum potential of machine learning algorithms, to perform such tasks. As an alternative for improving accuracy, combining machine learning techniques with vegetation spectral indices (VIs) extracted from multispectral images appears to offer the potential for precision agriculture-related practices. Osco et al. (2019a) analyzed the individual contribution of multiple VIs for the Random Forest model to estimate canopy nitrogen (N_2) content in citrus-trees context.

Random Forest is a machine learning technique that has been proved to be highly accurate in several agricultural applications, including maize prediction (Jeong et al. 2016). When evaluated to predict the yield of different crops (wheat, maize, and potato) with climate and biophysical variables, at global and regional scales, the Random Forest algorithm outperformed multiple linear regressions models, adopted as a benchmark, in all considered performance statistics (Jeong et al. 2016). Random Forest is an effective and versatile machine-learning method for crop yield predictions because of its high accuracy and precision, ease of use, and utility in data analysis (Jeong et al. 2016). However, specifically in the precision agriculture context, how much of this technique may be improved with the adoption of a ranking-based strategy is an unsolved issue.

In previous research (Osco et al., 2019a; Osco et al., 2019b) we evaluated the importance of different inputs into remote sensing datasets related to plant analysis. Still, to the best of our knowledge, no literature evaluated the impact on the performance of a ranking-based strategy to estimate yield-prediction in maize or a similar culture. To help fulfill this gap, we propose a Random Forest ranking-based approach to predict maize yield using only multispectral UAV-imagery. A ranking strategy using the correlation parameter was initially applied, considering a group of VIs. As a result, only the most relevant indices were considered as input to the Random Forest model. We compared our approach with state-of-the-art machine learning algorithms, showing its potential to maize yield prediction. The main contribution of this study is to propose an alternative to help in potentialize the Random Forest technique to crop yield prediction tasks. To the best of our knowledge, this is the first tentative in the precision agriculture domain using remote sensing imagery. Our paper is organized as follows: Section 2 describes the proposed method; Section 3 and Section 4 presents and discusses the results, respectively; and Section 5 concludes this research.

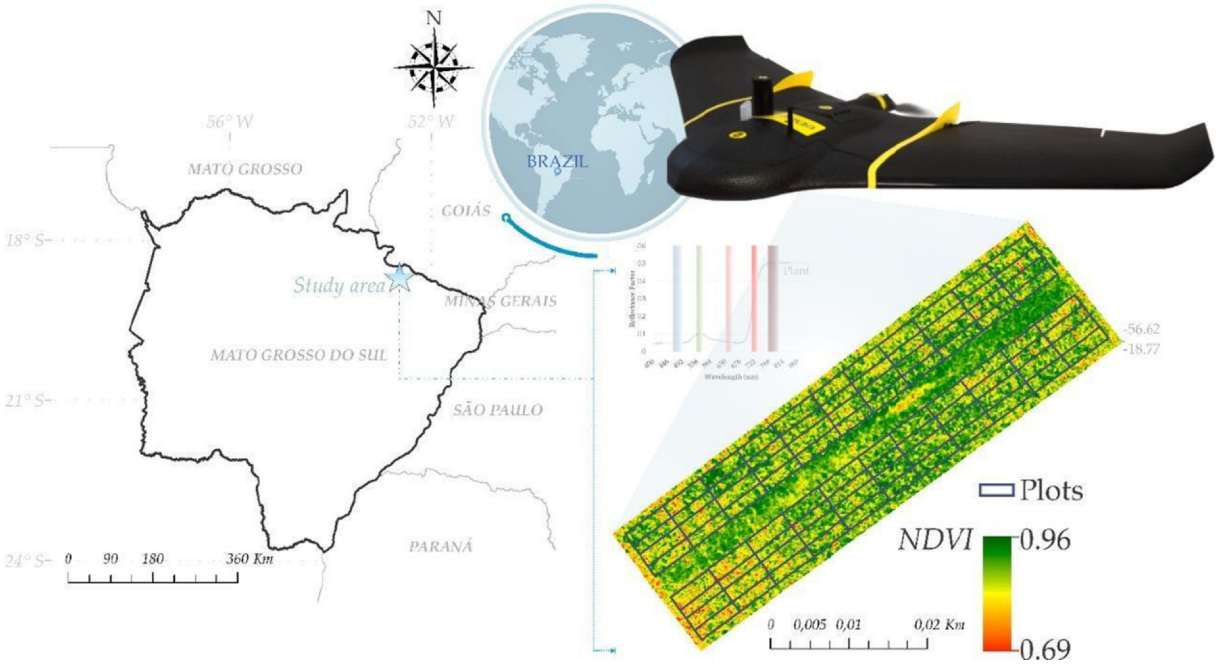

Fig. 1. Location of the study area in Chapadão do Sul, Brazil.

2. Materials and methods

2.1. Field trials

We experimented with 11 maize cultivars grown under two doses of Nitrogen (60 and 180 kg ha^{-1}) in top dressing for two crop seasons (2017/18 and 2018/19), with four replicates of each experimental unit. Each plot was 5 m long with five rows spaced at 0.45 m each. Fig. 1 shows the location of the experimental area on the campus of the Federal University of Mato Grosso does Sul, in the municipality of Chapadão do Sul. The grain yield of each experimental unit was obtained by harvesting, tracing the three central rows, and correcting for 13% humidity. The values were extrapolated to kg ha^{-1}.

2.2. Aerial multispectral image acquisition and vegetation indices

A total of 33 Vegetation indices (VIs) was calculated from aerial multispectral orthoimages. These orthoimages were generated from aerial scenes acquired by the Sensefly Parrot Sequoia ® and MicaSense ® multispectral sensors during the 2017/2018 and 2018/2019 crop seasons, respectively. VIs represent the mean of each plot (5 m × 0.45 m). Therefore, firstly, for each spectral band, we computed the mean values of each plot. Later, using these bands, the VIs were computed with the formulations presented in Table 1. The group of VIs included in this work is based on the Osco et al. (2019a) approach. A MicaSense Red-Edge multispectral sensor was installed on the rotary-wing UAV model X800 manufactured by the company XFly ®. The Sequoia multispectral sensor was embedded on the Sensefly eBee ® RTK fixed-wing developed by the company Sensefly. The spectral regions registered by both sensors are similar and correspond to green (550 nm), red (660 nm), red-edge (735 nm), and near-infrared (NIR) (790 nm). MicaSense Red-Edge sensors also register a blue-band, however, to perform our tests appropriable, we did not consider it since we aimed to produce a similar data-set for both periods. Furthermore, both multispectral sensors have a luminosity sensor allowing for an in-field calibration of the registered values. The flyover was carried out with the crop at 50 days after emergence (DAE) during the first crop season (2017/2018). In the second crop season, the overflight was carried out by the eBee RTK.

2.3. Statistical analysis

Pearson's correlation coefficients were initially estimated to verify the association between grain yield (GY) and vegetation indices (VIs), and we used the correlation network to graphically express these results (Fig. 2). In this procedure, green lines link variables with positive correlation, and red lines join negatively correlated variables. The line thickness is proportional to the magnitude of the correlation. We choose to present Fig. 2 and discuss the outcomes of this procedure into the manuscripts' next sections.

To rank the VIs, a Random Forest learner was trained using each vegetation index individually. The performance for each trained Random Forest using only one VI at a time was compared against a simpler regression method using the same VI. The percentual gain in performance of the Random Forest compared to the simpler regression method, using each VI, was used as a reference to rank the VIs. This performance was measured using Pearson's correlation coefficient. To achieve a less biased rank, this process was repeated 10 times using 10-fold cross-validation, and the mean rank position of each VI was used as the final position. The simpler regression method used in the experiments was the ZeroR (0-R) that only uses the mean value of the target variable, calculated using the training set, as a predictor. After that, n sets of VIs are created, where n is the number of available VIs. The first set contains all the n VIs and the other sets are formed by removing one VI at a time, always the last in the merit ranking among the available VIs. Mean Absolute Errors (MAE) and Pearson's correlation coefficients (r) were then obtained over randomized 10-fold cross-validation with 10 repetitions for each set using the RF model. This ranking strategy has also been applied to the other two learners: Support Vector Machines (SVM) and Artificial Neural Networks (ANN).

We also compared the RF overall performance against four other machine learning models (Table 2): SVM using Sequential Minimal Optimization (SMO) and a polynomial Kernel, Linear Regression (LR) with Akaike information criterion for attribute selection; k-Nearest Neighbors (KNN) with $k = 5$, and an ANN using sigmoid as the activation function for the hidden neurons and a linear activation function for the output neuron. The hidden layer has 17 neurons (Weka's default heuristic value that corresponds to the number of classes plus outputs divided by 2). For the RF approach, we also tested two meta learners;

Table 1
The vegetation indices used in the experiments.

Index	Equation
ARVI2 (Atmospherically Resistant Vegetation Index 2)	$-0.18 + 1.17 * [(R\lambda_{nir} - R\lambda_{red})/(R\lambda_{nir} + R\lambda_{red})]$
ATSAVI (Ajusted Transformed soil-ajusted VI)	$1.22 * \left[\frac{(R\lambda_{nir} - 1.22 * R\lambda_{red} - 0.03)}{(1.22 * R\lambda_{nir} + R\lambda_{red} - 1.22 * 0.03 + 0.08(1 + 1.22^2)}\right]$
BWDRVI (Blue-wide dynamic range vegetation index)	$0.1 * (R\lambda_{nir} + R\lambda_{red})$
CCCI (Canopy Chlorophyll Content Index)	$\frac{(R\lambda_{nir} - R\lambda_{rededge}) / (R\lambda_{nir} + R\lambda_{rededge})}{(R\lambda_{nir} - R\lambda_{red}) / (R\lambda_{nir} + R\lambda_{red})}$
CIgreen (Chlorophyll Index Green)	$(R\lambda_{nir}/R\lambda_{green}) - 1$
CIrededge (Chlorophyll Index RedEdge)	$(R\lambda_{nir}/R\lambda_{rededge}) - 1$
CVI (Chlorophyll Vegetation Index)	$R\lambda_{nir} * (R\lambda_{red}/R\lambda_{green}{}^2)$
DVI (Difference Vegetation Index)	$R\lambda_{nir}/R\lambda_{red}$
EVEI2 (Enhanced Vegetation Index 2)	$2.5 * (R\lambda_{nir} - R\lambda_{red})/(R\lambda_{nir} + 2.4 * R\lambda_{red} + 1)$
GDVI (Difference NIR/Green Difference Vegetation Index)	$R\lambda_{nir} - R\lambda_{green}$
GEMI (Global Environment Monitoring Index)	$2 * (1 - 0.25 * 2) - ((R\lambda_{red} - 0.125)/(1 - R\lambda_{red}))$
GNDVI (Green Normalized Difference Vegetation Index)	$(R\lambda_{nir} - R\lambda_{red})/(R\lambda_{nir} + R\lambda_{red})$
GRNDVI (Green-Red NDVI)	$[R\lambda_{nir} - (R\lambda_{green} + R\lambda_{red})]/[R\lambda_{nir} + (R\lambda_{green} + R\lambda_{red})$
GRVI (Green-Red Vegetation Index)	$(R\lambda_{green} - R\lambda_{red})/(R\lambda_{green} + R\lambda_{red})$
GSAVI (Green Soil Adjusted Vegetation Index)	$[(R\lambda_{nir} - R\lambda_{green})/(R\lambda_{nir} + R\lambda_{green} + 0.5)] * 1.5$
GTVI (Green Triangle Vegetation Index)	$(NDVI + 0.5)/(NDVI + 0.5) * [(\sqrt{NDVI} + 0.5)]$
IPVI (Infrared Percentage Vegetation Index)	$R\lambda_{nir}/((R\lambda_{nir} + R\lambda_{red})/2) * (NDVI + 1)$
LogR (Log Ratio)	$\mathrm{Log}(R\lambda_{nir}/R\lambda_{red})$
MSAVI (Modified Soil Adjusted Vegetation Index)	$[2 * R\lambda_{nir} + 1 - \sqrt{(2 * R\lambda_{nir} + 1)^2 - 8 * (R\lambda_{nir} - R\lambda_{red})}]/2$
MSRNir_Red (Modified Simple Ratio NIR/RED)	$(R\lambda_{nir}/R\lambda_{red} - 1)/\sqrt{(R\lambda_{nir}/R\lambda_{red} + 1}$
NDRE (Normalized Difference Red-Edge Index)	$(R\lambda_{nir} - R\lambda_{rededge})/\lambda_{nir} + R\lambda_{rededge}$
NDVI (Normalized Difference Vegetation Index)	$(R\lambda_{nir} - R\lambda_{red})/(R\lambda_{nir} + R\lambda_{red})$
NGRDI (Normalized Green-Red Difference Index)	$(R\lambda_{green} - R\lambda_{red})/R\lambda_{green}/(R\lambda_{nir} + R\lambda_{red} + R\lambda_{green})$
NormR1 (Normalized G)	$R\lambda_{green}/(R\lambda_{nir} + R\lambda_{red} + R\lambda_{green})$
NormR2 (Normalized NIR)	$R\lambda_{nir}/(R\lambda_{nir} + R\lambda_{red} + R\lambda_{green})$
NormR3 (Normalized R)	$R\lambda_{red}/(R\lambda_{nir} + R\lambda_{red} + R\lambda_{green})$
RGR (Red Green Ratio Index)	$R\lambda_{red}/R\lambda_{green}$
RI (Redness Index)	$R\lambda_{red} - R\lambda_{green}/R\lambda_{red} + R\lambda_{green}$
RRI 1	$R\lambda_{nir}/R\lambda_{rededge}$
SAVI (Soil-Adjusted Vegetation Index) II	$[(1 + 0.5) * (R\lambda_{nir} - R\lambda_{red})]/(R\lambda_{nir} + R\lambda_{red} + 0.5)$
SRQT_IR_R	$\sqrt{R\lambda_{nir}/R\lambda_{red}}$
SRRed_NIR	$R\lambda_{red}/R\lambda_{nir}$
WDRVI (Wide Dynamic Range Vegetation Index)	$(0.1 * R\lambda_{nir} - R\lambda_{red})/(0.1 * R\lambda_{nir} + R\lambda_{red})$

NIR = near-infrared.

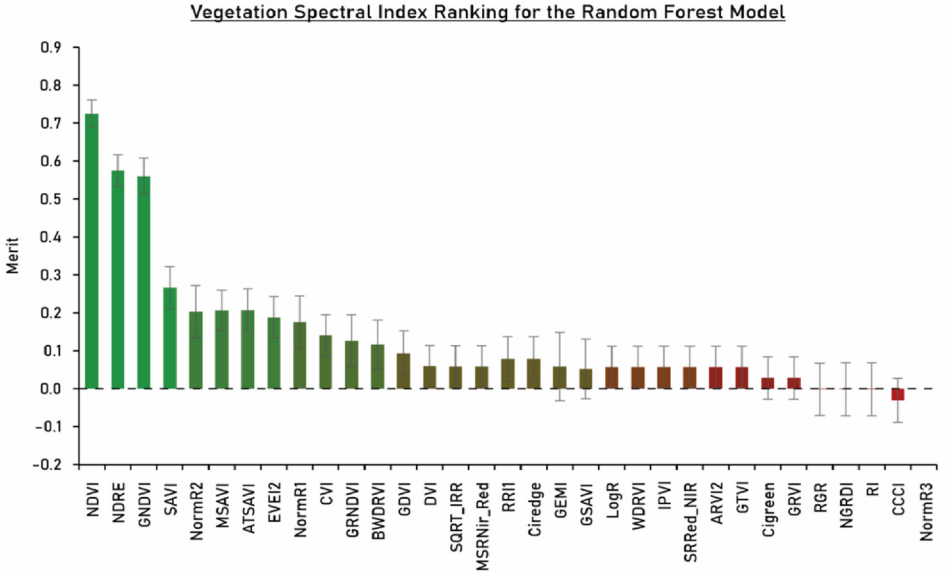

Fig. 2. Pearson's correlation coefficient improvement using Random Forest against a baseline method to predict maize production with each vegetation index individually. VIs is ordered by the average ranking position achieved over 10-fold cross-validation using the merit metric.

Table 2
Machine learning and *meta*-learning (#5 and #6) algorithms used in this study.

Test Order	ML Model	Reference
#1	Random Forests - RF	Belgiu and Drăgu (2016)
#2	Support Vector Machine - SVM	Nalepa and Kawulok (2019)
#3	Linear Regression - LR	Štepanovský et al. (2017)
#4	K-Nearest Neighbours - KNN	Ali et al. (2019)
#5	Bagging - RF + BAG	Breiman (1996)
#6	Additive Regression - RF + AR	Friedman (2002)
#7	Artificial Neural Networks - ANN	Egmont-Petersen et al. (2002)

the Bagging algorithm (RF + BAG) and the Additive Regression (RF + AR), using RF as the base weak learner. We organized the test order of each machine learning model according to Table 2.

Comparisons were made using all available VIs (33 in total) and using the best set of VIs found for RF with the ranking-based approach. Additionally, we tested the best set of VIs against three other sets: (1) the VIs that use only visible bands – (thus, referring to it as RGB); (2) the VIs using RGB plus NIR bands, and finally; (3) the VIs calculated over RGB, NIR and Red-Edge spectral bands groups, that correspond to all of the available indices in this study.

The experiments were run on an Intel® Core™ i7 CPU with 12 Gb RAM and all hyperparameters were set according to Weka 3.9.4 default library. Boxplots for all the configurations evaluated are presented together with the Scott-Knott test results.

3. Results

3.1. Ranking of Vegetation indices

Fig. 2 shows the merits and rank positions of each VI with the standard deviation over the 10-fold cross-validation, ordered by the average rank position. NDVI is the highest-ranked VI, followed by NDRE and GNDVI, whereas NormR3 is the last-ranked VI. Among all tested VIs, twelve (36%) presented a higher than 10% improvement over the baseline method (average merit $>$ 0.1), and three from this subgroup was able to provide an improvement upper to 50% (average merit $>$ 0.5) compared to the baseline method in the maize yield prediction task using RF. According to Figs. 2, 5 VIs showed a proximal to zero improvements (average merit also negative) over the baseline, indicating that besides not helping to improve performance in this task they may be disturbing the algorithm's performance.

Fig. 3 shows the correlation coefficient (r) and MAE variations as the size of the VI sets is reduced, one VI at a time, using the ranking shown in Fig. 2. The increase in r and decrease in MAE is not steady and a peak happens with the best 3 VIs: NDVI, NDRE, and GNDVI. The same strategies were applied to SVM and ANN methods (Figs. 4 and 5).

When RF adopts these three VIs (NDVI, NDRE, and GNDVI) to perform the maize yield prediction task, the model can explain > 72% (MAE = 950 kg ha^{-1} approximately) of GY, whereas this estimation decrease to 65% (MAE above 1000 kg ha^{-1}) when GNDVI is removed from the model. The results for both SVM (Fig. 4) and ANN (Fig. 5) models, using the VIs ranking-based strategy to maize yield prediction, are extremely poor when compared to the RF.

In this sense, RF seems to be potentialized the with VIs ranking strategy application, but the same was not observed for the other machine learning algorithms; SVM and ANN. The best result for the SVM (Fig. 4) model occurs when 8 VIs (r = 0.35 and MAE = 1450 kg ha^{-1} approximately) are adopted. For the ANN model, it was necessary 11 VIs for it to achieve the best performance in predicting maize-yield, with a correlation coefficient of 0.5 and MAE equals 1550 kg ha-1 approximately (Fig. 3). Nonetheless, these results are at least 31% worse than those obtained with RF in the same task.

3.2. Machine learning Models' performances

Table 3 shows the results of grouping the Scott-Knott test for Pearson's correlation coefficient (r) and Mean Absolute Error (MAE) obtained with machine learning models using all VIs and the three best VIs. The *meta*-learner RF + AR, using RF as the base machine learning model, was the technique that stood out; statistically presenting the highest averages of r (0.78) and the lowest MAE (853.11 kg ha^{-1}). It is important to mention that the usage of the three best VIs provided satisfactory results only for the AR and RF techniques. The other techniques only showed better results when considering all VIs.

Pearson's correlation coefficient and MAE obtained with machine learning models using all VIs and three best VIs are displayed in Figs. 6 and 7, respectively. When using all VIs (Fig. 4), the RF, RF + AR, and RF + BAG techniques stand out for presenting the highest average for r and the lowest for MAE. Although ANN is one of the techniques that provided the highest value for r, the ones obtained for MAE were higher with also high variability.

When we used the proposed ranking strategy to select the three best VIs (NDVI, NDRE, and GNDVI) (Fig. 8), the AR and RF techniques stood out for presenting the highest median values for r and the lowest for MAE. The boxplots comparing RF in four different configurations of VI sets are shown there (Fig. 8). The set automatically constructed with the ranking-based approach returned the highest average values for r and the lowest MAE.

Fig. 8 shows that, by using only the RGB group of indices, a low accuracy is achieved in predicting grain yield. An improvement occurs when considering NIR and Red-edge indices in the RF model. However, the higher accuracy was obtained considering only the three indices (NDVI, NDRE, and GNDVI). These indices use Red, Green, NIR, and Red-Edge bands.

4. Discussion

The potential of the Random Forest (RF) ranking-based approach for predicting maize yield using multispectral UAV-imagery was evaluated with a robust dataset composed of 11 maize cultivars, under two rates of N_2 fertilization rates (60 kg ha^{-1} and 180 kg ha^{-1}) in four replicates of each plantation plot. These aerial images represent two crop seasons (2017/2018 and 2018/2019) acquired by two different multispectral sensors (MicaSense Red-Edge ® and Sensefly Parrot Sequoia ®) embedded in UAVs, characterizing more robustness to our experimental setup. The use of a vast number of vegetation spectral indices (VIs) (total of 33) in the maize yield prediction task was important to evaluate different configurations of sets using 7 distinct machine learning techniques. Additionally, the analysis with two meta learners, Bagging and Additive Regression, using the RF as the base weak learner, was essential to identify how much the RF algorithm can be potentialized with the ranking-strategy presented here.

Based on Pearson's correlation coefficient and Mean Absolute Error (MAE) values, our study demonstrates how feasible machine learning is to predict maize-yield using only VIs. In this manner, the strategy to individually rank each of the 33 VIs based on a merit metric that measures the improvement on the Pearson's correlation coefficient using RF against a baseline method was essential to improve the performance of the machine learning algorithms at the aforementioned task. When the seven learners were explored with all VIs, we identified a similar performance among the models. Two exceptions, however, were found to the SVM and KNN models, returning the lower correlations identified with the grain yield variable. Nonetheless, the use of the only the three best VIs found by the RF ranking-strategy reveals that machine learning algorithms do not present similar performance, being RF the best option, which was even better when the meta learner Additive Regression was implemented with. In this regard, the configurations were an indicator of the importance of this ranking-strategy to be applied before the machine learning method being applied for the

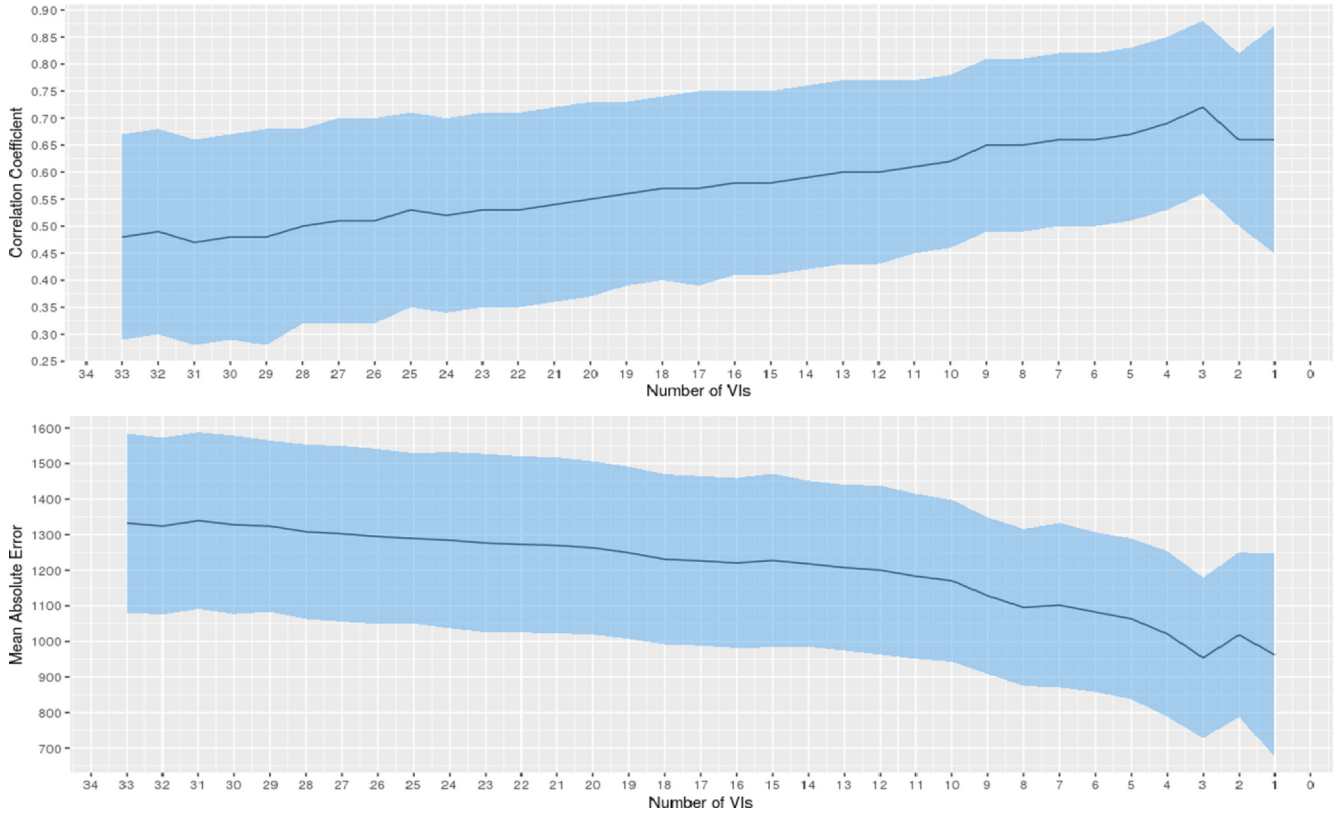

Fig. 3. Pearson's correlation coefficient (r) and mean absolute error (MAE) variation as the number of VIs is reduced, for the random forest model, using the VI ranking strategy proposed in this work.

yield prediction.

Concerning the accuracy of our model, it achieved similar metrics as others obtained by similar research. One study (Serele et al., 2000) adopted aerial multispectral imagery and demonstrated that the ANN model presented high performance (r = 0.95, RMSE = 365 kg ha^{-1}) for the maize-yield estimation, but this metric was achieved only when a total of 11 variables (three topographic features, four vegetation index, and four textural indices) were included as input data in the ANN

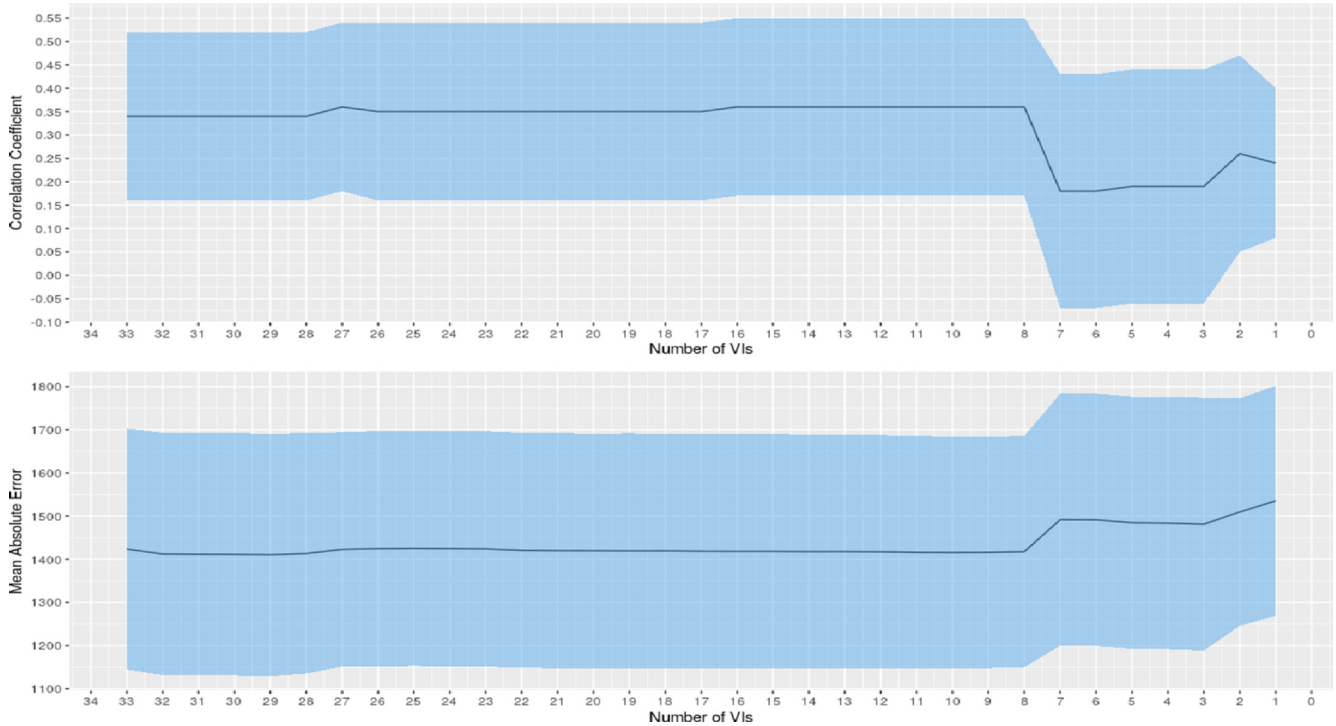

Fig. 4. Pearson's correlation coefficient (r) and mean absolute error (MAE) variation as the number of VIs is reduced, for the SVM model, using the VI ranking strategy proposed in this work.

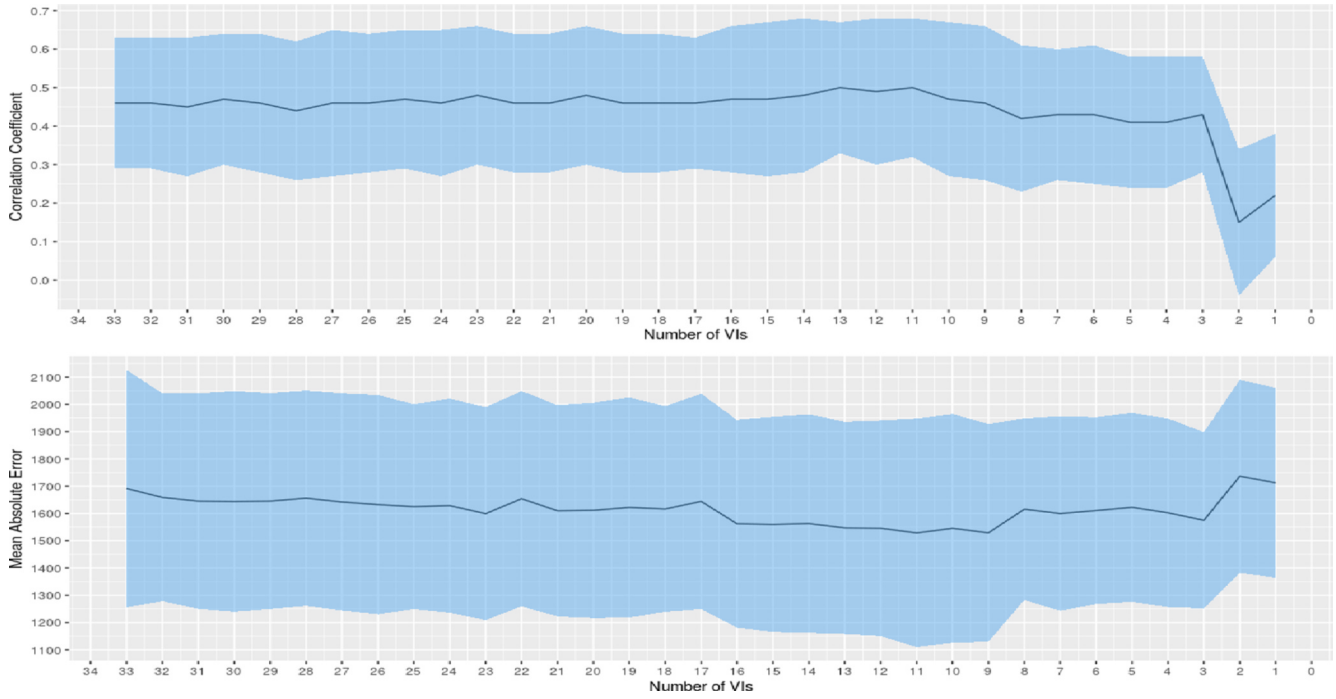

Fig. 5. Pearson's correlation coefficient (r) and mean absolute error (MAE) variation as the number of VIs is reduced, for the ANN model, using the VI ranking strategy proposed in this work.

Table 3
Grouping of means by the Scott-Knott test for Pearson's correlation coefficient (r) and mean absolute error (MAE) obtained with machine learning models using all VIs and the three best VIs identified using the ranking-based approach.

ML model	r		MAE	
	all VIs	three best VIs	all VIs	three best VIs
RF	0.48 a	0.72b	1332.47 d	953.86 d
RF + AR	0.51 a	0.78 a	1274.05 d	853.11 e
RF + BAG	0.47 a	0.69b	1373.30 d	1026.34 d
SVM	0.34c	0.31 d	1423.56b	1431.37b
LR	0.41b	0.31 d	1416.92c	1477.46b
KNN	0.21 d	0.54c	1502.71b	1224.16c
ANN	0.46 a	0.31 d	1691.14 a	1666.79 a

Note: Means followed by different letters in the same column differ by the Scott-Knott test at 5% probability.

model. Another research (Uno et al. 2005) used hyperspectral images in the aforementioned task and stated that the ANN model was also efficient (r = 0.76, RMSE = 19.7%). Nonetheless, the authors argued that the expected prediction errors of approximately 20% (RMSE for the test datasets) appear to be high for the creation of yield maps for precision agriculture, requiring new experiments. A more recent study (Khanal et al. 2018) also related to maize yield prediction using the RF algorithm combined multispectral aerial images with topographic data and stated that the RF consistently outperformed (r = 0.73, RMSE = 970 kg ha-1) other machine learning models.

Our experiment section using only UAV-images corroborates that the *meta*-learner RF + AR, using RF as the base machine learning model, (r = 0.78, MAE = 853.11 kg ha-1) has a high potential to support the maize yield prediction task. Moreover, it should be argued that when we adopted the RF ranking approach, only three VIs (NDVI, NDRE, and GNDVI) were required to execute this prediction-task with high performance, implying that our solution is a low-cost and simpler method than most approaches. Additionally, it is worth mentioning that when the ranking-based approach was applied with other machine learning algorithms, like SVM and ANN, we discovered that only the RF method strongly benefited from this strategy. We found out that the ranking-based approach potentialized the algorithm performance, increasing the correlation coefficient, and decreasing the MAE, but the same situation did not occur when other machine learning methods were tested.

Our analysis of the ranking-based approach demonstrates that some VIs contribute more and positively to improve the performance of machine learning algorithms in maize-yield prediction tasks, whereas other VIs are disturbing the RF algorithm's performance. We verified that among the three VIs that presented a higher than 50% improvement over the baseline (average merit $>$ 0.5), two are calculated using the Red or Green and NIR spectral bands. Moreover, we noted that the contribution of both NDRE (that requires Red-Edge spectral band in its calculation) and GNDVI (that uses the Green band) are similar according to the merits and rank positions analysis (Fig. 2). This could be interpreted that the VIs extracted from RGB and NIR images may be enough to perform the maize-yield prediction task. However, when we analyzed the RF's performance including the NDRE (that uses the Red-Edge band) index, we discovered that the correlation coefficient with grain yield increased from 65% to 78%. Studies (Abdel-Rahman et al. 2013; Ramoelo et al. 2015) have demonstrated that, in the RF model, VIs developed considering red-edge bands generally outperform other VIs for predicting crops' nutrients, like Nitrogen, which is related to crops yield rates.

The performance of each machine learning algorithm was, as presented, evaluated with a multiple set of configurations. This analysis returned interesting outcomes because while some learners improved their performance with the ranking strategy, other models had their performances degraded (Table 3; Figs. 6 and 8). Among all the explored algorithms, the RF was the only one that significantly improved its performance with the ranking-strategy. This finding was observed for all RF variations used in this study, such as RF + AF (r = 0.78; MAE = 853.11 kg ha^{-1}), RF + BAG (r = 0.69; MAE = 1026.34 kg ha^{-1}), and RF (r = 0.72; MAE = 953.86 kg ha^{-1}), compared to SVM (r = 0.31; MAE = 1431.37 kg ha^{-1}) or ANN (r = 0.31; MAE = 1666.79 kg ha^{-1}). It should be highlighted that our strategy to combine RF + AF (r = 0.78; MAE = 853.11 kg ha^{-1}) resulted in an improvement in MAE value of 100% approximately

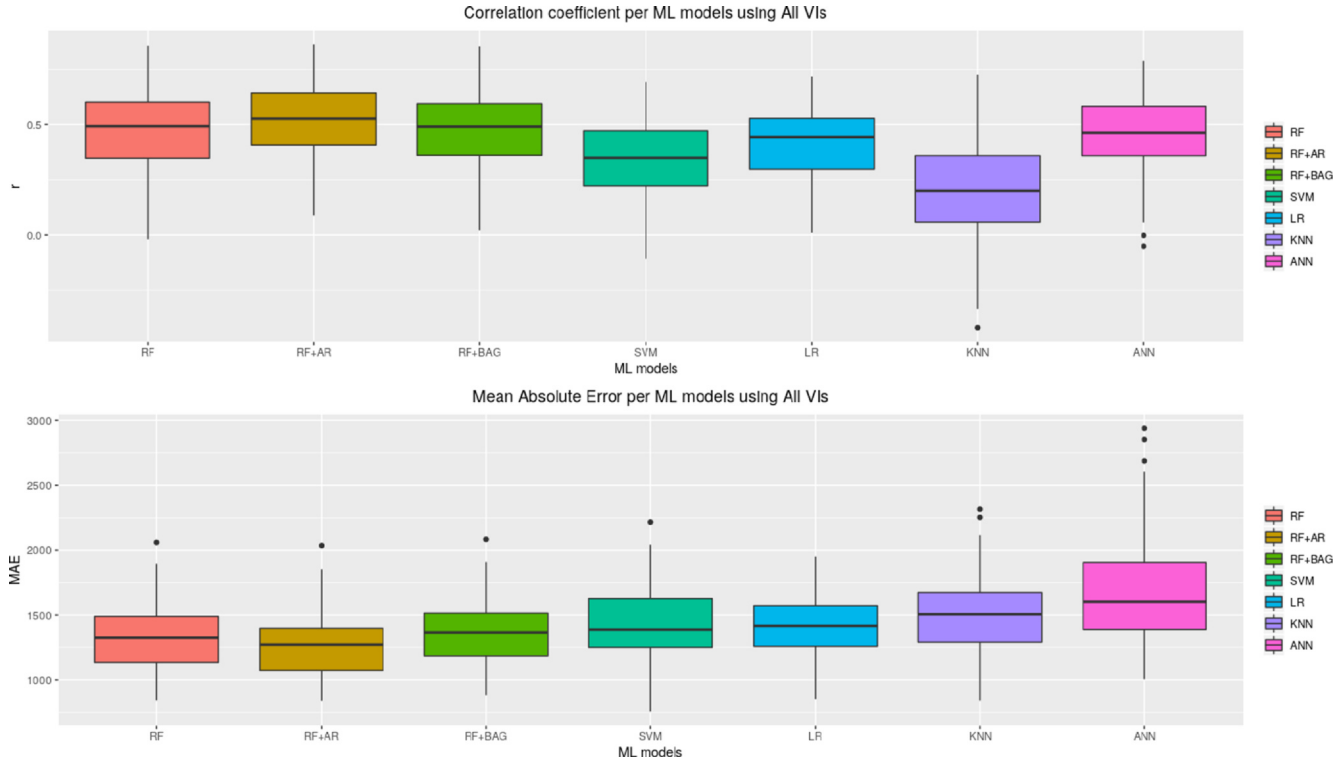

Fig. 6. Boxplots for Pearson's correlation coefficient (r) and mean absolute error (MAE) for different machine learning models evaluated using all VIs.

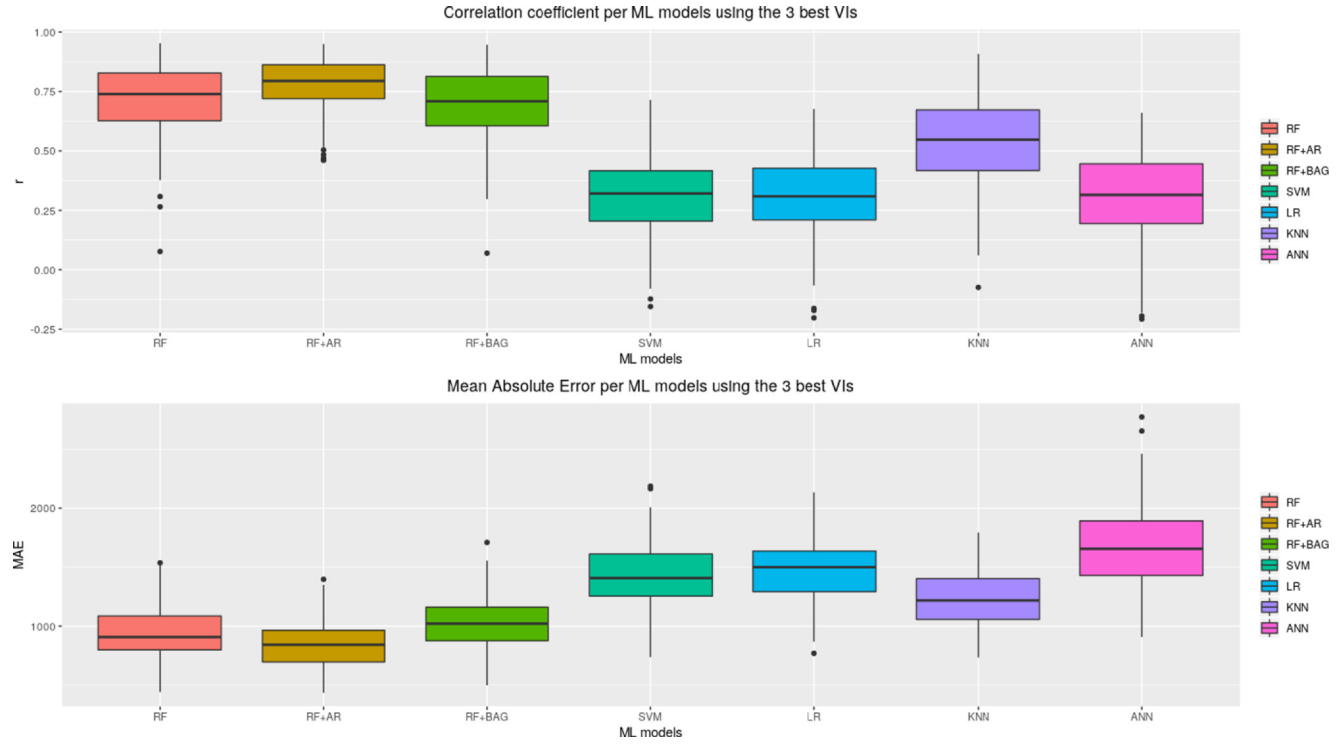

Fig. 7. Boxplots for Pearson's correlation coefficient (r) and mean absolute error (MAE) for different machine learning models evaluated using the three best VIs.

compared to the value obtained with the ANN model (r = 0.31; MAE = 1666.79 kg ha^{-1}) which has been adopted frequently by related studies (Serele et al., 2000; Uno et al. 2005). The highest performance with RF possibly occurred due to the internal structure of the algorithm which is based on decision tree sets. Other studies that implemented not only spectral information but also soil variables, climatological data, among others to estimate yield in their RF models, could be potentialized with the proposed ranking-based approach, achieving even higher accuracies.

The approach presented here may be implemented with different datasets over diverse cultivars. Here, our particular interest was to investigate whether a ranking-based approach can potentialize the RF algorithm to predict maize yield using VIs calculated from multispectral UAV-imagery since this learner is classified as an effective and versatile

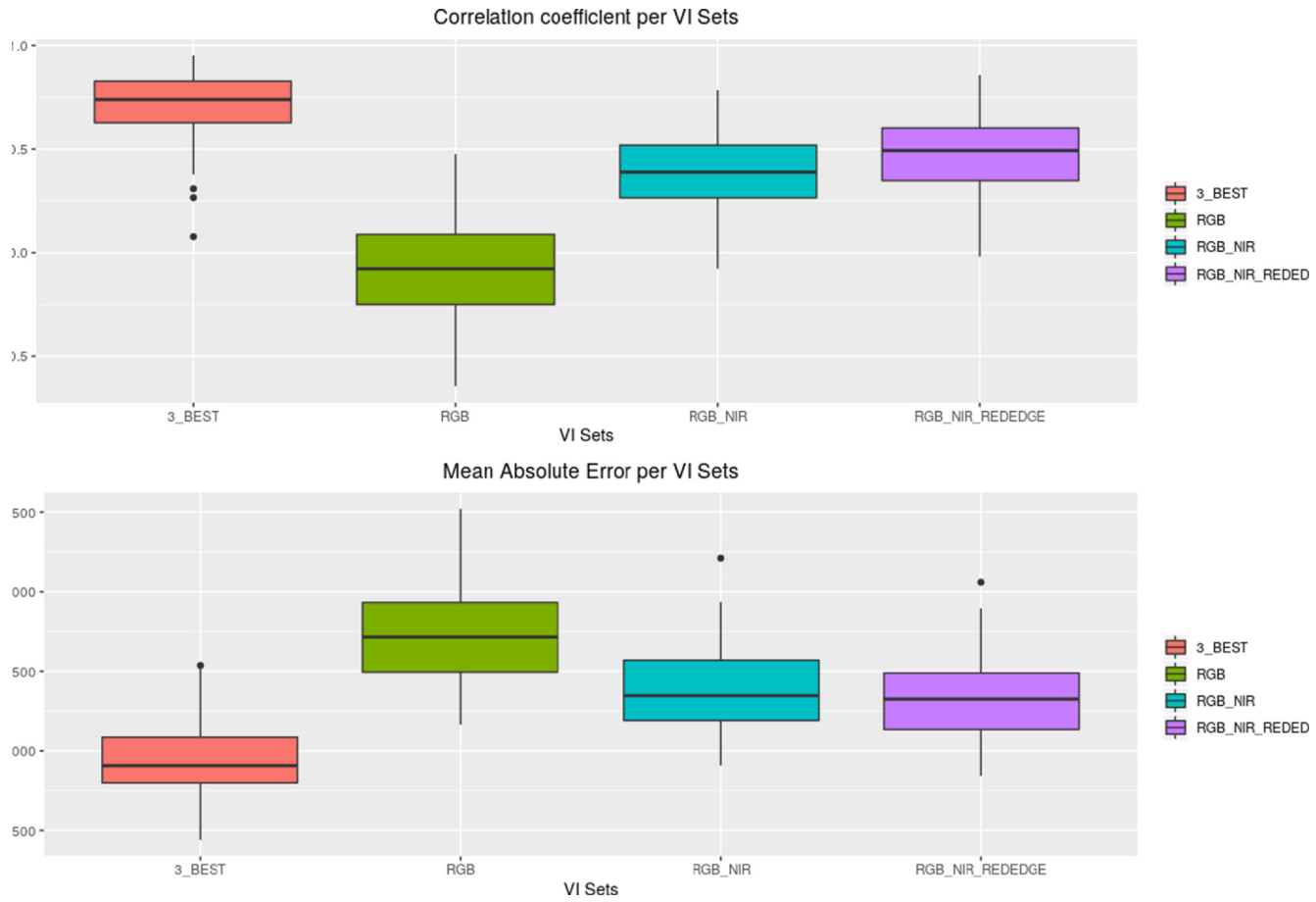

Fig. 8. Boxplots for Pearson's correlation coefficient (r) and mean absolute error (MAE) when comparing different sets of VIs using random forest. The sets compared are the ones resulting from the proposed approach, 3_BEST, and 3 other configurations: only RGB based VIs, RGB + NIR, and RGB + NIR + Rededge VIs.

machine-learning method for crop yield predictions (Jeong et al. 2016). The main advantage of our proposed procedure is that it allowed estimating maize yield with high accuracy, using a rapid and cost-efficient manner, which is essential to support the recurrent monitoring of the agricultural landscapes. In this regard, precision agriculture practices could benefit from our finding framework, supporting agricultural system management, and helping decision-making actions.

5. Conclusions

This study investigated whether the ranking-based strategy of vegetation indices (VIs) can potentialize the Random Forest (RF) algorithm to predict maize yield using only multispectral UAV-imagery. Up to the best of our knowledge, this refers to the first exploration of its kind, mainly in the precision agriculture context. Here, we tested several combinations of VIs using a group of machine learning techniques, and we pointed out which of the implemented learners is more suitable to predict maize-yield with VIs calculated from multispectral UAV-imagery. Our findings showed that the RF algorithm performed better in all of the configuration scenarios, especially for the ranking-based approach of VIs. Besides, we detected that some VIs contributed more to maize yield prediction than others. The integration between the three best VIs and a *meta*-learner, Additive Regression, using the RF as the base learner, improved maize yield prediction accuracy even more. The procedure developed during our analysis refers to a simpler strategy to estimate maize-yield prediction considering only UAV-based imagery. We conclude that the RF ranking-based approach is appropriate to predict this agronomic variable (grain yield). We suggest our method to be adopted in future research to evaluate different types of crop yield, as to assist proper management and be used in decision-making models in the precision agriculture domain.

Funding

This research was funded by CNPq, grant number 303559/2019–5, 433783/2018–4, 314902/2018–0, and 304173/2016–9; CAPES - Print, grant number 88881.311850/2018–01, and Fundect, grand number 59/300.066/2015, and 59/300.095/2015.

Declaration of Competing Interest

The authors declared that there is no conflict of interest.

Acknowledgments

The authors acknowledge the support of UFMS (Federal University of Mato Grosso do Sul), UCDB (Dom Bosco Catholic University), CNPq (National Council for Scientific and Technological) and CAPES (Coordination for the Improvement of Higher Education Personnel - Finance code 001).

References

Abdel-Rahman, E.M., Ahmed, F.B., Ismail, R., 2013. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. Int. J. Remote Sens. 34 (2), 712–728. https://doi.org/10.1080/01431161.2012.713142.

Ali, N., Neagu, D., Trundle, P., 2019. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. SN Appl. Sci. 1, 1–15. https://doi.org/10.1007/s42452-019-1356-9.

Amatya, S., Karkee, M., Gongal, A., Zhang, Q., Whiting, M.D., 2016. Detection of cherry tree branches with full foliage in planar architecture for automated sweet-cherry

harvesting. Biosyst. Eng. 146, 3–15. https://doi.org/10.1016/j.biosystemseng.2015.10.003.
Barbosa, A., Trevisan, R., Hovakimyan, N., Martin, N.F., 2020. Modeling yield response to crop management using convolutional neural networks. Comput. Electron. Agric. 170, 105197. https://doi.org/10.1016/j.compag.2019.105197.
Belgiu, M., Drăgu, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.
Breiman, L., 1996. Machine Learning 24 (2), 123–140. https://doi.org/10.1023/a:1018054314350.
Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Comput. Electron. Agric. 151, 61–69. https://doi.org/10.1016/j.compag.2018.05.012.
Zhang, J, Yuhang, He, Lian, Yuan, Peng, Liu, Xianfeng, Zhou, Yanbo, Huang, et al., 2019. Machine learning-based spectral library for crop classification and status monitoring. Agronomy 9 (9), 496. https://doi.org/10.3390/agronomy9090496. https://www.mdpi.com/2073-4395/9/9/496#cite.
CONAB, C.N. de A. Monitoring of the Brazilian harvest, Grains; Brasília, DF, 2020; ISBN 2318-6852. Available from https://www.conab.gov.br/info-agro/safras/graos/boletim-da-safra-de-graos.
Egmont-Petersen, M., de Ridder, D., Handels, H., 2002. Image processing with neural networks—a review. Pattern Recogn. 35 (10), 2279–2301. https://doi.org/10.1016/s0031-3203(01)00178-9.
Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38 (4), 367–378. https://doi.org/10.1016/s0167-9473(01)00065-2.
Hunt, M.L., Blackburn, G.A., Carrasco, L., Redhead, J.W., Rowland, C.S., 2019. High resolution wheat yield mapping using Sentinel-2. Remote Sens. Environ. 233, 111410. https://doi.org/10.1016/j.rse.2019.111410.
Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Kim, S.-H., 2016. Random Forests for Global and Regional Crop Yield Predictions. PLoS ONE 11 (6), e0156571. https://doi.org/10.1371/journal.pone.0156571.
Khaki, S., Wang, L., Archontoulis, S.V., 2020. A CNN-RNN Framework for Crop Yield Prediction. Front. Plant Sci. 10. https://doi.org/10.3389/fpls.2019.01750.
Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., Shearer, S., 2018. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. Comput. Electron. Agric. 153, 213–225. https://doi.org/10.1016/j.compag.2018.07.016.
Liakos, K., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine Learning in Agriculture: A Review. Sensors 18 (8), 2674. https://doi.org/10.3390/s18082674.
Nalepa, J., Kawulok, M., 2019. Selecting training sets for support vector machines: a review. **Artif. Intell. Rev.** 52, 857–900. https://doi.org/10.1007/s10462-017-9611-1.
Nevavuori, P., Narra, N., Lipping, T., 2019. Crop yield prediction with deep convolutional neural networks. Comput. Electron. Agric. 163, 104859. https://doi.org/10.1016/j.compag.2019.104859.
Osco, L.P., Paula, A., Ramos, M., Pereira, D.R., Akemi, É., Moriya, S., Matsubara, E.T., 2019. Predicting canopy nitrogen content in citrus-trees using random forest algorithm associated to spectral vegetation indices from UAV-imagery. Remote Sensing 11 (24), 2925–2942. https://doi.org/10.3390/rs11242925.
Osco, L.P., Ramos, A.P.M., Moriya, É.A.S., Bavaresco, L.G., de Lima, B.C., Estrabis, N., Pereira, D.R., Creste, J.E., Júnior, J.M., Gonçalves, W.N., Imai, N.N., Li, J., Liesenberg, V., de Araújo, F.F. Modeling hyperspectral response of water-stress induced lettuce plants using artificial neural networks. 2019b. Remote Sensing, 11(23). https://doi.org/10.3390/rs11232797.
Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R.L., Mouazen, A.M., 2016. Wheat yield prediction using machine learning and advanced sensing techniques. Comput. Electron. Agric. 121, 57–65. https://doi.org/10.1016/j.compag.2015.11.018.
Ramos, P.J., Prieto, F.A., Montoya, E.C., Oliveros, C.E., 2017. Automatic fruit count on coffee branches using computer vision. Comput. Electron. Agric. 137, 9–22. https://doi.org/10.1016/j.compag.2017.03.010.
Ramoelo, A., Cho, M.A., Mathieu, R., Madonsela, S., van de Kerchove, R., Kaszta, Z., Wolff, E., 2015. Monitoring grass nutrients and biomass as indicators of rangeland quality and quantity using random forest modelling and WorldView-2 data. Int. J. Appl. Earth Obs. Geoinf. 43, 43–54. https://doi.org/10.1016/j.jag.2014.12.010.
Senthilnath, J., Dokania, A., Kandukuri, M.K.N.R., Anand, G., Omkar, S.N., 2016. Detection of tomatoes using spectral-spatial methods in remotely sensed RGB images captured by UAV. Biosyst. Eng. 146, 16–32. https://doi.org/10.1016/j.biosystemseng.2015.12.003.
Serele, C.Z., Gwyn, Q.H.J., Boisvert, J.B., Pattey, E., McLaughlin, N., Daoust, G., 2000. Corn yield prediction with artificial neural network trained using airborne remote sensing and topographic data. International Geoscience and Remote Sensing Symposium. https://doi.org/10.1109/igarss.2000.860527.
Shah, Vinita, Shah, Prachi, 2018. Groundnut crop yield prediction using machine learning techniques. Int. J. Scient. Res. Comput. Sci. Eng. Inform. Technol. 3 (5), 1093–1097. https://www.researchgate.net/profile/Vinita_Shah/publication/326112319_Groundnut_Crop_Yield_Prediction_Using_Machine_Learning_Techniques/links/5c4e9d9b458515a4c74584c7/Groundnut-Crop-Yield-Prediction-Using-Machine-Learning-Techniques.pdf.
Štepanovský, M., Ibrová, A., Buk, Z., Velemínská, J., 2017. Novel age estimation model based on development of permanent teeth compared with classical approach and other modern data mining methods. Forensic Sci. Int. 279, 72–82. https://doi.org/10.1016/j.forsciint.2017.08.005.
Su, Y., Xu, H., Yan, L., 2017. Support vector machine-based open crop model (SBOCM): Case of rice production in China. Saudi Journal of Biological Sciences 24 (3), 537–547. https://doi.org/10.1016/j.sjbs.2017.01.024.
Uno, Y., Prasher, S.O., Lacroix, R., Goel, P.K., Karimi, Y., Viau, A., Patel, R.M., 2005. Artificial neural networks to predict corn yield from Compact Airborne Spectrographic Imager data. Comput. Electron. Agric. 47 (2), 149–161. https://doi.org/10.1016/j.compag.2004.11.014.
Vani, S., Sukumaran, R.K., Savithri, S., 2015. Prediction of sugar yields during hydrolysis of lignocellulosic biomass using artificial neural network modeling. Bioresour. Technol. 188, 128–135. https://doi.org/10.1016/j.biortech.2015.01.083.