Smart Agricultural Technology 5 (2023) 100292

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Soil and satellite remote sensing variables importance using machine learning to predict cotton yield

Franciele Morlin Carneiro [a], Armando Lopes de Brito Filho [b], Francielle Morelli Ferreira [b,c], Getulio de Freitas Seben Junior [c], Ziany Neiva Brandão [d], Rouverson Pereira da Silva [b], Luciano Shozo Shiratsuchi [e,*]

[a] Federal Technological University of Paraná (UTFPR), Santa Helena 85892-000, PR, Brazil
[b] São Paulo State University (UNESP), School of Agricultural and Veterinarian Sciences, Jaboticabal 14884-900, SP, Brasil
[c] State University of Mato Grosso (UNEMAT), Nova Mutum 78450-000, MT, Brazil
[d] Brazilian Agricultural Research Corporation (EMBRAPA Cotton), Campina Grande 58428-095, PB, Brasil
[e] School of Plant, Enviromental and Soil Sciences, Louisiana State University (LSU), Baton Rouge 70808, LA, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Gossypium hirsutum
Random forest
Satellite imagery

ABSTRACT

Remote sensing (RS) in agriculture has been widely used for mapping soil, plant, and atmosphere attributes, as well as helping in the sustainable production of the crop by providing the possibility of application at variable rates and estimating the productivity of agricultural crops. In this way, proximal sensors used by RS help producers in decision-making to increase productivity. This research aims to identify the best feature importance ranking to the Random Forest Classifier to predict cotton yield and select which one best correlates with cotton yield. This work was developed in four commercial fields on a Newellton, LA, USA farm. We evaluated the cotton in different years as 2019, 2020, and 2021. The variables evaluated were: soil parameters, topographic indices, elevation derivatives, and orbital remote sensing. The soil sensor used was: GSSI Profiler EMP400 (soil electromagnetic induction sensor) at a frequency of 15 kHz, and the RS data were collected from satellite images from Sentinel 2 (passive sensor) and active sensor from LiDAR (Light Detection and Ranging). For training (70%) and validation (30%) of dataset results, Spearman correlation was used between sensors and cotton yield data, machine learning (Random Forest Classifier and Regressor - RFC and RFR). The metric parameters were the coefficient of determination (R^2), the Mean Absolute Error (MAE), and the Root Mean Square Error (RMSE). This study found that profiler, Sentinel-2 (blue, red, and green), TPI, LiDAR, and RTK elevation show the best correlations to predicting cotton yield.

Abbreviation: RS, remote sensing; LA, Louisiana; LiDAR, Light Detection and Ranging; RFC, Random Forest Classifier; RFR, Random Forest Regressor; R^2, coefficient of determination; MAE, Mean Absolute Error; RMSE, Root Mean Square Error; LAI, leaf area index; TPI, Topographic Position Index; S2, Sentinel-2; ML, machine learning; WSS, Web Soil Survey; ShA, Sharkey clay; TeB, Tensas-Sharkey clays; DgB, Dundee-Goldman complex; DoA, Dowling clay; TnA, Tunica clay; CeA, Commerce silty clay loam; TeD, Tensas-Sharkey complex; TaA, Tensas clay; CEC, cation exchange capacity, B, boron; Ca, calcium; H, hydrogen; K, potassium; Mg, magnesium; Na, sodium; pH, soil pH; Cu, copper; Fe, iron; Mn, manganese; OM, organic matter; P, phosphorus; S, sulfur; Zn, zinc; TWI, Topographic Wetness Index; IDW, Inverse Distance Weighted; NIR, near-infrared; Colab, Google Colaboratory; VI, vegetation index; B2, Blue; B3, Green; B4, Red; B6, Rededge; B8, NIR; CM, confusion matrix; SR, Simple Ratio; NDVI, Normalized Dif-ference Vegeta-tion Index; NDRE, Normalized Difference Red Edge Index; NLI, Nonlinear Vegetation Index; NDWI, Normalized Dif-ference 860/1240 Normalized Dif-ference Water Index; DVI, Difference Vege-tation Index; BNDVI, Blue-Normalized Difference Vegetation Index; ISR, Inverse of the Simple Ratio; GNDVI, Green Normal-ized Difference Vegetation Index; RGBVI, Red Green Blue Vegetation Index; SAVI, Soil Adjusted Vegetation Index; CIRE, Chlorophyll Index-RedEdge; IPVI, Infrared Percent-age Vegetation Index; EVI, Ehanced Vegeta-tion Index; EVI 2, Ehanced Vegeta-tion Index 2; y_i, observed values; $\hat{y}_i$, estimated values; N, number of samples; $\overline{y_i}$, mean of the observed values; Orbital RS, satellite imageries; RTK elevation, Real-time kinematic positioning elevation; B, blue; R, red; G, green; NIR, near infrared; RE, red edge; CHIRPS, Climate Hazards Group InfraRed Precipitation with Stations; Hz, hertz.

* Correspondence author.
E-mail address: lshiratsuchi@agcenter.lsu.edu (L.S. Shiratsuchi).

https://doi.org/10.1016/j.atech.2023.100292
Received 20 April 2023; Received in revised form 9 June 2023; Accepted 17 July 2023
Available online 21 July 2023

Introduction

The early determination of crop yields is essential information for crop field management. Crop yield forecasts are fundamental to national policy formulation involving food security and personal living standards [1,2]. Therefore, in recent years remote sensing-based techniques have become increasingly crucial for monitoring crop growth, such as biomass [3], leaf area index – LAI [4], chlorophyll content [4,5], and yield estimation [6].

Remote sensing (RS) technologies are gaining prominence due to the expanding availability of massive time-series datasets, such as the availability of satellite information and high-quality image resolutions (radiometric, spectral, spatial, temporal) [7]. In addition, it is increasingly common to monitor crop physiological characteristics with spectral reflectance and combined vegetation indices. In such a context, numerous studies have applied orbital remote sensing to provide a fast, profitable, cost-effective, and non-destructive way for yield estimation of various crops.

Satellite imageries, orbital RS, allow users to monitor crops over large areas with image resolutions (spectral, spatial, radiometric, and temporal) [7,8]. The launch of the Sentinel-2 (S2) satellite allowed the collection of images with good frequency and medium spectral resolution, and a passive sensor with multispectral bands, being widely used in agriculture because it provides free images with a revisit time of five days [9]. In addition, another RS technology that has been used in agriculture is LiDAR (Laser Imaging Detection and Ranging). LiDAR is an active sensor installed on vehicles such as airplanes or helicopters and is mainly used for plant height, biomass estimation [10], and phenotyping [11]. However, other active sensors, such as proximal sensors, are used to make fertilizer maps, track the crop's growth stages and biophysical characteristics (plant height, biomass, yield), and monitor the spatio-temporal variability of crops.

Moreover, some studies compare the performance of active and passive sensors. The interaction between active and passive sensors was studied by many researchers with different objectives, such as crop yield, irrigation systems, detection of salinity on soil, etc. [12–14]. These studies used techniques for the fusion of satellite images and proximal sensor data to monitor irrigation scheduling and crop growth stage [15]. The author states that using both active and passive sensors allows adjustments for incongruent data due to the limitations of each technology.

Remote Sensing data can get extensive information, usually with a behavior that is not linear. Thus, the use of machine learning analysis demonstrates an excellent statistical analysis tool. Among the various ML, Random Forest is one of the most used to select important features [16]. In this study, we proposed to use Random Forest to choose the most informative variables on cotton yield prediction.

To estimate the yield, we use the biophysical characteristics of the crop, proposing a new approach, combining different scenarios using soil parameters (soil physical properties and fertility, and data from WSS), topographic indices, and elevation derivatives, in addition to orbital remote sensing data (Sentinel-2 spectral bands and Vegetation indices) to predict cotton yield before harvesting. Our study was motivated to identify the best attribute and the importance ranking to the Random Forest Classifier to predict cotton yield and select which ones better correlate with the cotton yield.

Materials and methods

Study area

The cotton yield work focused on four areas conducted on-farm in Newellton (32°10′58.7″N, 91°16′38.1″W), LA, USA. The study was realized in four fields and three different years (2019, 2020, and 2021). Information on the location of each field is in Fig. 1.

Remote sensing collection was realized before the emergence of the cotton crop (Table 1) in 2019, 2020, and 2021 years. This study focused on the soil variability analysis days after sowing cotton crops. In this farm, we evaluated just cotton crop, on which sowing year per field was field #37 (2019), #54 (2020), #34 (2019 and 2021), and #60 (2020). Table 1 describes soil type, sowing and harvesting data, and satellite imagery data.

The weather conditions focus on accumulative rainfall during 2019, 2020, And 2021 years. According to this data, there was greater rain in 2019 than in 2020 and 2021. However, comparing all the years of study, in 2020, the rain had a more uniform distribution (Fig. 2). The difference between wetter (2019) and drier (2020) years was 100.61 mm. The

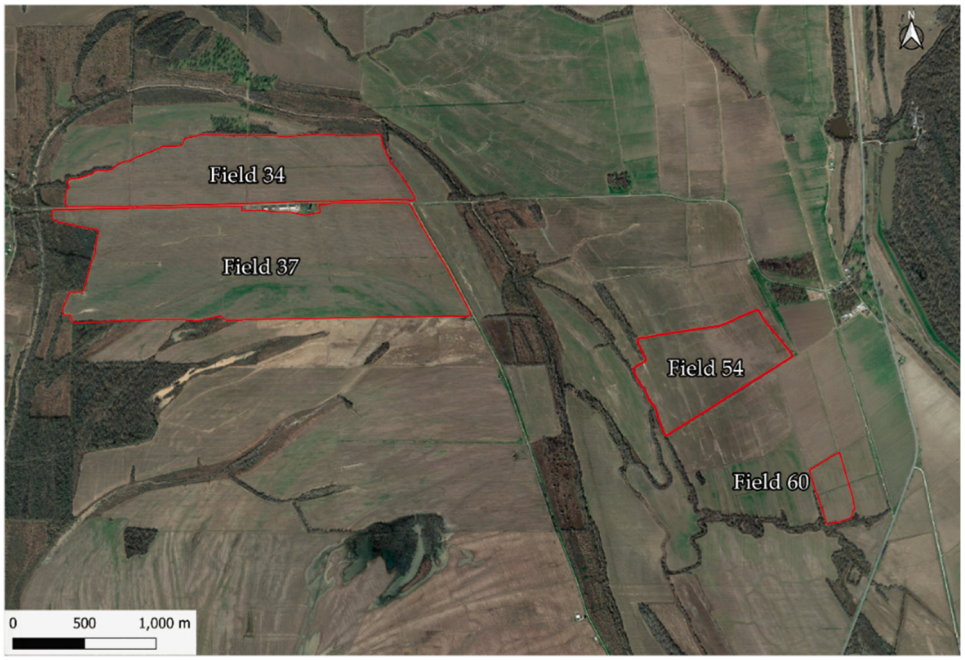

Fig. 1. Study area showing the experimental fields on a farm (Hardwick Planting Company) in Newellton, LA, USA.

Table 1
Description of soil type, sowing and harvesting date, and satellite imagery collection per field.

Field ID#	Soil Type*	Area (ha)	Sowing Date	Harvesting date	Satellite imagery date
37	ShA TeB DgB DoA	232.14	May 1st, 2019	Oct 8th, 2019	May 6th, 2019
54	TeB ShA TnA CeA	60.92	May 2nd, 2020	Oct 3rd, 2020	May 10th, 2020
34	ShA TeB TeD TaA	111.29	–	Nov 1st, 2019	May 6th, 2019
34	ShA TeB TeD TaA		May 16th, 2021	Oct 18th, 2021	April 25th, 2021
60	ShA TeB	11.64	May 4th, 2020	Sept 19, 2020	April 30th, 2020

* ShA: Sharkey clay, TeB: Tensas-Sharkey clays, DgB: Dundee-Goldman complex, DoA: Dowling clay, TnA: Tunica clay, CeA: Commerce silty clay loam, TeD: Tensas-Sharkey complex, TaA: Tensas clay.

average multi-year accumulative rainfall from 2019 to 2021 was around 128.94 mm. We collected weather data using CHIRPS (Climate Hazards Group InfraRed Precipitation with Stations) data. The CHIRPS data was obtained using the Google Earth Engine platform (https://earthengine.google.com/).

Datasets

Datasets for this work were compound per six different scenarios (Fig. 3). They were:

1 *All inputs:* soil analysis parameters (CEC, B, Ca, H, K, Mg, Na, pH, Cu, Fe, Mn, OM, P, S, and Zn), Web Soil Survey (WSS), RTK elevation, Orbital Remote Sensing, LiDAR, and profiler (GSSI Profiler EMP400 - soil electromagnetic induction sensor).
2 *Soil parameters* ± *elevation derivatives*: soil analysis parameters (CEC, B, Ca, H, K, Mg, Na, pH, Cu, Fe, Mn, OM, P, S, and Zn) + RTK elevation (slope, TPI - Topographic position index, TWI - Topographic Wetness Index), profiler, WSS (Web Soil Survey), and LiDAR.
3 *Proximal Soil Sensing*: profiler and Orbital Remote Sensing.
4 *Remote Sensing*: LiDAR and Orbital Remote Sensing.
5 *Remote Sensing*: Orbital Remote Sensing (Vegetation Indices and Sentinel-2 bands)
6 *Remote Sensing*: Orbital Remote Sensing (Sentinel-2 bands)

Data collection and processing

Fig. 4 shows that the methodology summary of this work was organized using workflow. The data collection was organized and processed in three steps:

1 *Data collection:* reported input we collected in the field as soil parameters, elevation derivatives, orbital and proximal remote sensing, and cotton yield.
2 *Data processing*: a Random Forest Classifier (RFC) and confusion matrix were used to select the five and fifteen inputs that have greater importance for estimating yield. After that, it was identified

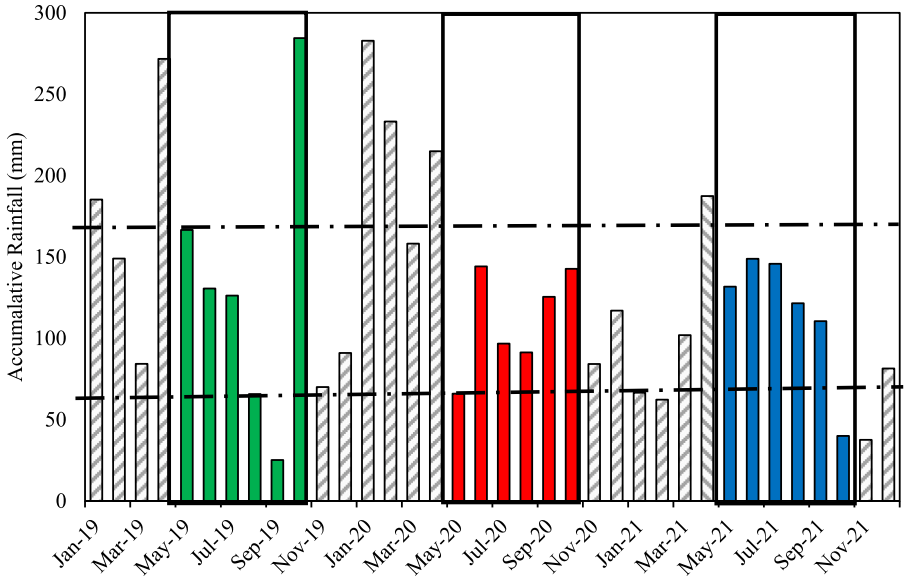

Fig. 2. Accumulative rainfall during 2019, 2020, and 2021, with emphasis on the crop planting period, for studies carried out in the Newellton, LA, USA.

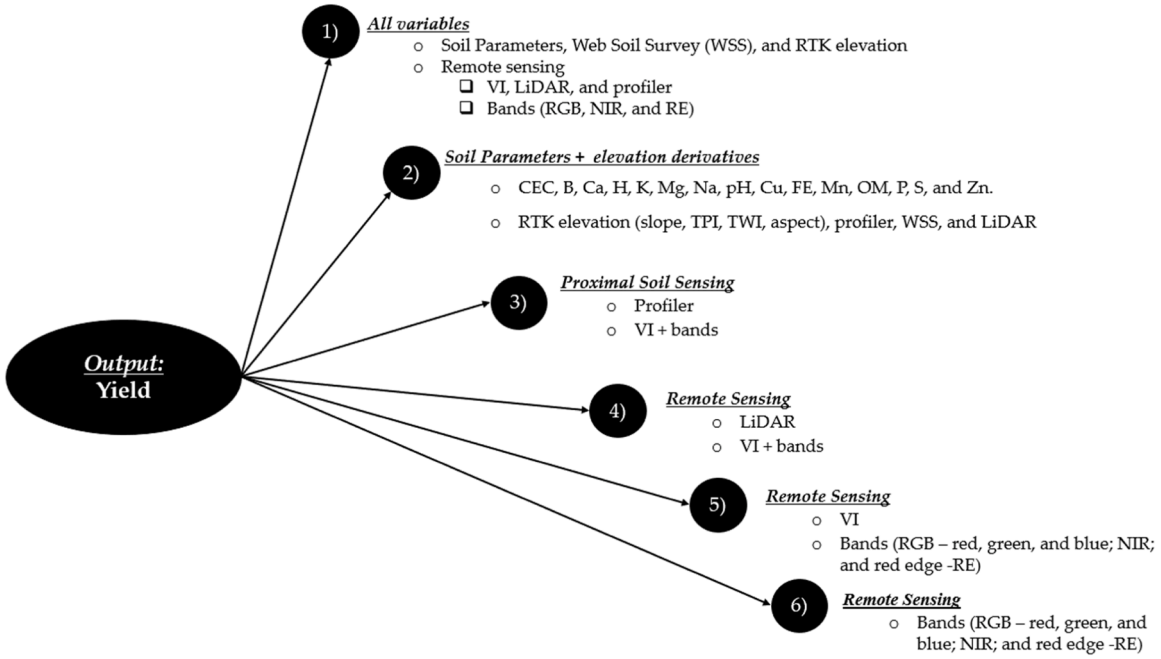

Fig. 3. Datasets were used with six different scenarios for this work.

which inputs presented the best correlation with yield using the Spearman coefficient correlation.

3 *Data Analysis*: it was selected the linear regression and Random Forest Regressor (RFR) to compare with the ones obtained by metric parameters (R2- determination coefficient, RSME - Root Mean Square Error, and MAPE - Mean Absolute Percentage Error) with the best to work on for these variables.

Soil parameters were generated from a soil fertility test, and this test was collected before the sowing date. WSS was acquired from the USDA webpage (https://websoilsurvey.nrcs.usda.gov/app/) [17], which provides the U.S. soil classification. The sower machine obtained RTK GPS elevations with a high-precision dataset. Topographic indices such as TPI and TWI, slope, and aspect terrain were generated from RTK GPS elevations.

The soil sensor was GSSI Profiler EMP400 (soil electromagnetic induction sensor) at 15 kHz. The profiler sensor collected the electrical conductivity of the soil, and about time collection was 1 Hz (1 cps). The remote sensing (RS) data was obtained from satellite images with Sentinel-2 (passive sensor) from the Copernicus Open Access Hub website (https://scihub.copernicus.eu/dhus/#/home), using the dataset from Sentinel-2 satellite missions. Another input used was LiDAR data acquired from Louisiana Statewide LiDAR - LSU Atlas (https://atlas.ga.lsu.edu/datasets/lidar2000/) [18], in which its database has high-resolution (5 m) elevation data from the state of Louisiana. The output was the cotton yield data, filtered by open-source software, MapFilter 2.0, with spatial dependence in 10 m and 25% variation of limits. After that, this output was processed and interpolated using Inverse Distance Weighting (IDW) in QGIS software.

For processing the dataset, we used GIS open-source software QGIS.

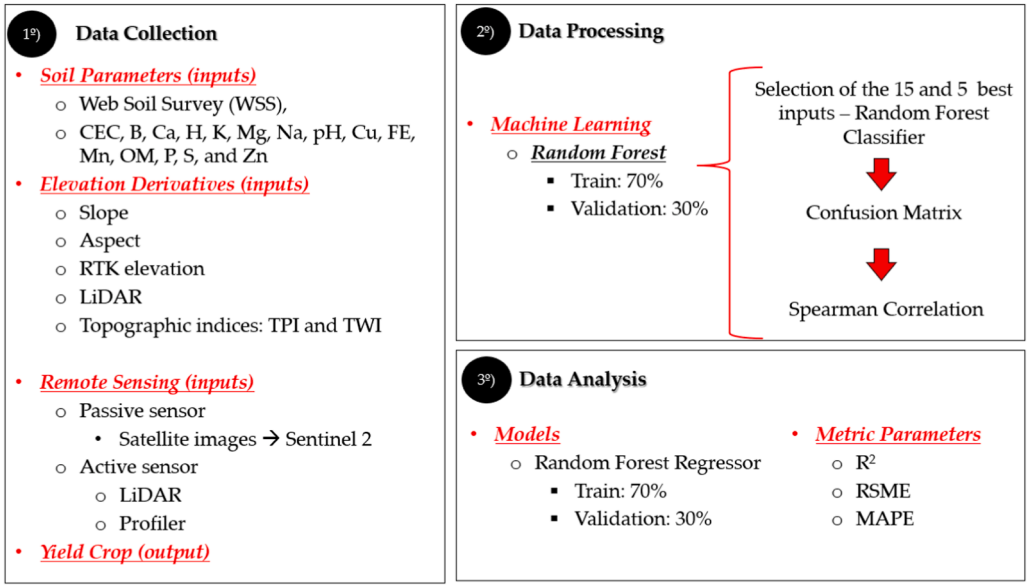

Fig. 4. Representation of the work-plan in steps used in this research.

Table 2
Sentinel-2 spectral bands used in this research.

Sentinel-2 band number	Band type	Central wavelength (nm)	Spatial resolution (m)	Abbreviation
B2	Blue	490	10	B
B3	Green	560	10	G
B4	Red	665	10	R
B6	Red Edge	740	20	RE
B8	Near-infrared	842	10	NIR

All input, except for soil fertility, were created maps and interpolated by IDW (Inverse Distance Weighted) method, with a grid of 3 m. After that, a 3 × 3 m grid was created, and a point sampling tool was used to join these inputs. For soil fertility, a buffer was created for each center point from soil test analysis, on which each buffer had around 25 m. Finally, to join the fertility layer with the others, we used joined layer with all input layers in just one file using QGIS software.

Remotely sensed data

Sentinel-2 satellite constellation

The Sentinel-2 (S2) imageries were collected 15 days after the sowing date, using just one imagery per the date on what was observed of spatial-temporal variability per field and date. In this research, Table 2 shows the S2 bands used blue, green, red, red edge, and NIR (near-infrared). Google Colaboratory, called Google Colab, was used to download and process these satellite images using Python language code using the Google Colab website (https://colab.research.google.com). Besides the six S2 bands in Table 2, we used 15 vegetation indices (VI) to predict cotton yield (Table 3). Table 1 shows the period when S2 imagery was selected, and the VI was calculated.

Statistical analysis and modeling

Accuracy evaluation model

The Random Forest Classifier (RFC) was used to classify the best inputs in six scenarios to predict cotton yield. The hyperparameters used were ntree (300), mtry (8), proximaty (True), importance (True), and Type of random forest (Classification). The Random Forest Regressor (RFR) was used to select which scenario and variable best-predicted cotton yield. The hyperparameters for RFR were used GridSearch. Verbose (0), random state (0), and the criterion (squared_error) were the same for all scenarios, while the other hyperparameters in Table 4 show those in detail. In both models, RFC and Random Forest Regressor (RFR), the database used 70% train and 30% test. R and Python languages were used to process data for RFC and RFR, respectively. In addition, for the running RFR, we used the best parameters you can see in detail in Supplementary Table S1.

The extraction accuracy of prediction cotton yield was evaluated using three metrics parameters: the coefficient of determination (R2), root mean square error (RMSE), and mean absolute percentage error (MAPE). The Equations 1 - 3 for each metric parameter in below. In addition, it was calculated confusion matrix (CM), also processed in the R language code of RFC. The accuracy calculation from CM was Kappa coefficient and precision.

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

$$R^2 = 1 - - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \quad (2)$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i - y_i|}{y_i} \times 100 \quad (3)$$

Table 3
Vegetation Indices (VI) evaluated for the prediction cotton yield using machine learning.

Vegetation Index	Abbreviation	Spectral Bands	Equation using Sentinel-2 bands	Source
Simple Ratio	SR	NIR and Red	$\frac{B8}{B4}$	[19]
Normalized Difference Vegetation Index	NDVI	NIR and Red	$\frac{(B8 - B4)}{(B8 + B4)}$	[20]
Normalized Difference Red Edge Index	NDRE	NIR and Rededge	$\frac{(B8 - B6)}{(B8 + B6)}$	[21]
Nonlinear Vegetation Index	NLI	NIR and Red	$\frac{(B8^2 - B4)}{(B8^2 + B4)}$	[22]
Normalized Difference 860/1240 Normalized Difference Water Index	NDWI	Green and NIR	$\frac{(B3 - B8)}{(B3 + B8)}$	[23]
Difference Vegetation Index	DVI	Red and NIR	B8 – B4	[24]
Blue-Normalized Difference Vegetation Index	BNDVI	NIR and Blue	$\frac{(B8 - B2)}{(B8 + B2)}$	[25]
Inverse of the Simple Ratio	ISR	Red and NIR	$\frac{B4}{B8}$	[26]
Green Normalized Difference Vegetation Index	GNDVI	NIR and Green	$\frac{(B8 - B3)}{(B8 + B3)}$	[21]
Red Green Blue Vegetation Index	RGBVI	Green, Blue, and Red	$\frac{B3^2 - (B2 \times B4)}{B3^2 + (B2 \times B4)}$	[27]
Soil Adjusted Vegetation Index	SAVI	NIR and Red	$\frac{(B8 - B4)}{(B8 + B4 + 1) \times (1 + 1)}$	[28]
Chlorophyll Index-RedEdge	CIRE	NIR and Rededge	$\frac{B8}{B6} - 1$	[29]
Infrared Percentage Vegetation Index	IPVI	NIR and Red	$\frac{B8}{(B8 + B4)}$	[30]
Ehanced Vegetation Index	EVI	NIR, Red, and Blue	$2.5 \times \frac{(B8 - B4)}{(B8 + 6 \times B4 - 7.5 \times B2) + 1}$	[31]
Ehanced Vegetation Index 2	EVI2	NIR and Red	$2.5 \times \frac{(B8 - B4)}{B8 + (2.4 \times B4) + 1}$	[32]

B2: Blue, B3: Green, B4: Red, B6: Rededge, and B8: NIR.

Table 4
Importance values of top 5 and 15 ranked variables for the yield prediction using Random Forest Classifier.

Field (year) All inputs	Top 5 – Importance inputs to estimate yield	Top 15 – Importance inputs to estimate yield
37 (2019)	TPI, aspect, slope, NDRE, and CIRE	Aspect, CIRE, slope, EVI, Elevation, profiler, NDWI, and IPVI
54 (2020)	Profiler, LiDAR, NDRE, aspect, and slope	Aspect, LiDAR, slope, profiler, NDRE, TWI, CIRE, RGBVI, SR, NDVI, elevation, IPVI, ISR, green, and red
34 (2019)	TPI, RGBVI, CIRE, slope, and TWI	Slope, TWI, TPI, RGBVI, CIRE, aspect, NDRE, elevation, NLI, profiler, LiDAR, NIR, P, Fe, and DVI
34 (2021)	Aspect, TPI, slope, NDRE, and TWI	Aspect, slope, TPI, NDRE, TWI, GNDVI, NDWI, BNDVI, LiDAR, elevation, NLI, RGBVI, NIR, profiler, and CIRE
60 (2020)	Profiler, LiDAR, elevation, TPI, and RGBVI	Elevation, LiDAR, profiler, TPI, RGBVI, BNDVI, CIRE, NDWI, red, slope, NIR, TWI, NLI, and NDRE
Soil parameters and elevation derivatives		
37 (2019)	Profiler, TPI, aspect, slope, and TWI	Aspect, slope, profiler, LiDAR, BS_Mg, K, B, and PH
54 (2020)	Profiler, LiDAR, TPI, aspect, and slope	aSpect, slope, profiler, LiDAR, TPI, elevation, TWI, K, Mg, Fe, Zn, BS, Mg, P, and S
34 (2019)	Elevation, TPI, slope, aspect, and TWI	TPI, slope, aspect, TWWI, elevation, profiler, LiDAR, Na, Ca, Fe, P, pH, K, OM, and S
34 (2021)	Elevation, TPI, slope, aspect, and TWI	Aspect, slope, TPI, TWI, elevation, LiDAR, profiler, Mn, BS_K, P, S, BS_Ca, BS_MG, NA, and BS_H
60 (2020)	Profiler, elevation, LiDAR, slope, and aspect	Profiler, LiDAR, elevation, aspect, slope, TWI, pH, WSS, B, Zn, K, OM, P, and CEC
Proximal soil sensing (profiler, VI, and bands)		
37 (2019)	Red edge, profiler, BNDVI, NDRE, and CIRE	Profiler, red edge, CIRE, BNDVI, NDRE, RGBVI, EVI, NDWI, GNDVI, blue, NIR, NLI, red, greem, and IPVI
54 (2020)	Profiler, blue, EVI, CIRE, and NDRE	Profiler, NDRE, CIRE, EVI, blue, BNDVI, RGBVI, red edge, green, SR, ISR, NDVI, IPVI, red, and EVI2
34 (2019)	Profiler, NDWI, red edge, RGBVI, and CIRE	Profiler, RGBVI, red edge, NDWI, CIRE, GNDVI, NDRE, BNDVI, greeen, NIR, blue, red, NLI, EVI, and EVI2
34 (2021)	Profiler, red edge, BNDVI, NDRE, and CIRE	Profiler, CIRE, NDRE, BNDVI, red edge, GNDVI, SAVI, RGBVI, NDWI, red, NLI, DVI, EVI, NIR, and EVI2
60 (2020)	Red edge, profiler, RGBVI, CIRE, and NDRE	Profiler, red edge, NDRE, CIRE, RGBVI, BNDVI, EVI, NDWI, GNDVI, DVI, NLI, EVI2, red, SAVI, and ISR
Remote Sensing (LiDAR, VI, and bands)		
37 (2019)	LiDAR, BNDVI, CIRE, EVI, and red edge	LiDAR, BNDVI, red edge, EVI, CIRE, NDWI, GNDVI, NDRE, NIR, blue, RGBVI, red, green, NLI, and SAVI
54 (2020)	LiDAR, EVI, vblue, CIRE, and NDRE	LiDAR, NDRE, CIRE, blue, EVI, BNDVI, RGBVI, red edge, green, NDVI, ISR, IPVI, NDWI, SR, and GNDVI
34 (2019)	LiDAR, red edge, BNDVI, NIR, and RGBVI	LiDAR, RGBVI, red edge, BNDVI, NIR, NLI, NDWI, CIRE, GNDVI, NDRE, red, green, EVI, blue, and DVI
34 (2021)	LiDAR, red edge, BNDVI, NDRE, and CIRE	LiDAR, CIRE, NDRE, BNDVI, reds edge, GNDVI, red, NDWI, RGBVI, SAVI, NLI, EVI, EVI2, DVI, and NIR
60 (2020)	Red edge, LiDAR, BNDVI, CIRE, and NDRE	LiDAR, BNDVI, red edge, CIRE, NDRE, RGBVI, EVI, GNDVI, red, NDWI, EVI2, NLI, DVI, SAVI, and SR
Remote Sensing (VI and bands)		
37 (2019)	Red edge, BNDVI, CIRE, NDWI, RGBVI	Red edge, RGBVI, BNDVI, CIRE, NDWI, NDRE, GNDVI, EVI, NIR, blue, NLI, green, red, SR, and DVI
54 (2020)	Blue, EVI, BNDVI, NDRE, and CIRE	BNDVI, NDRE, CIRE, EVI, blue, RGBVI, red edge, GNDVI, SR, NDWI, IPVI, ISR, NDVI, red, and green
34 (2019)	BNDVI, red edge, red, NDRE, and CIRE	BNDVI, red edge, NDRE, CIRE, red, RGBVI, NDWI, GNDVI, SAVI, EVI, NLI, EVI2, DVI, NIR, and blue
34 (2021)	Red, red edge, BNDVI, CIRE, and NDRE	BNDVI, red edge, NDRE, CIRE, red, RGBVI, NDWI, GNDVI, SAVI, EVI, NLI, EVI2, DVI, NIR, and blue
60 (2020)	Red edge, RGBVI, BNDVI, CIRE, and NDRE	NDRE, BNDVI, red edge, CIRE, RGBVI, EVI, GNDVI, red, NDWI, NLI, SAVI, EVI2. DVI, ISR, and NIR
Remote Sensing (bands)		
37 (2019)	Red edge, blue, NIR, red, and green	—
54 (2020)	Blue, red, green, red edge, and NIR	—
34 (2019)	Red edge, blue, red, NIR, and green	—
34 (2021)	Red, red edge, green, NIR, and blue	—
60 (2020)	Red, red edge, NIR, blue, and green	—

y_i and $\widehat{y_i}$ are the observed and estimated values, N is the number of samples, and $\overline{y_i}$ is mean of the observed values.

Results

Importance variables for the cotton yield prediction using Random Forest Classifier

The top 15 and 5 important features selected using RFC are in Table 4. These features include six scenarios in Fig. 3, i.e., all inputs; soil parameters and elevation derivatives; profiler and orbital RS; LiDAR and orbital RS; orbital RS (VI and S2 spectral bands); and S2 spectral bands.

Fig. 5 shows the summary of feature importance by RFC, the Top 5 variables that appear with more constancy among the six different scenarios. The most frequent variables that appear according to RFC were Red edge, NDRE, BNDVI, CIRE slope, red, aspect, LiDAR blue, TPI, profiler, TWI, nir, green, EVI, RTK elevation, and RGBVI. According to these results, the Spearman correlation coefficient was used to verify which variables best correlates for predicting cotton yield; Fig. 6 shows those.

Fig. 6 shows, for cotton yield prediction, according to results from Fig. 5, the inputs strongly correlated with the profiler, wherein profiler and yield are inversely proportional. Fig. 6 shows, for cotton yield prediction, according to results from Fig. 5, the inputs strongly correlated with the profiler, wherein profiler and yield are inversely proportional. Inputs with a moderate correlation with yield were: blue, green, red, TPI, RTK elevation, and LiDAR. In topographic indices correlation with yield data, we observed that the TPI has a more excellent correlation with the output than TWI.

Table 5 shows the confusion matrix for classifying the accuracy of the Top 15 and 5 feature importance by RFC. The overview of the results was that LiDAR, profiler, and Orbital RS data got greater accuracy values than soil and RTK elevation variables. We can use a small dataset to predict cotton yield using these inputs because the difference between the Top 5 and Top 15 has a few different values. For example, using orbital imagery and LiDAR data for the Top 5 was 0.80, and the Top 15 was 0.82. In this case, it is easier for the farmers to apply this methodology by using some inputs such as Top 5. They can obtain this data-free imagery and elevation high-resolution from LiDAR. In addition, Table 5 shows that the accuracy has improved from 0.75 to 0.90 for a few fields, and kappa also improved from 0.69 to 0.87. However, we focused on this work is reducing the number of inputs to be collected to make it easier for the farmers and for them to apply the methodology proposed in this work, maintaining the same quality, so we opted for five inputs.

Accuracy assessment for cotton yield prediction

RSME, R^2, and MAPE values show, in Table 6, that the best inputs classification using RFR were soil parameter and elevation derivatives, LiDAR, profiler, and orbital RS. In the overview for comparison between variables, the Top 5 got results more excellent than the Top 15 for the prediction of cotton yield. Using all inputs is unnecessary, and we observed that when using a few variables (Top 5) to get great results. It also allows farmers to adopt the accessible collection of these variables. In complement, in Table 6 RMSE value for the top 15 features is low compared to the Top 5. Therefore, we recommend using five inputs to predict productivity, not requiring many variables, facilitating data collection for producers, and making decision-making more practical.

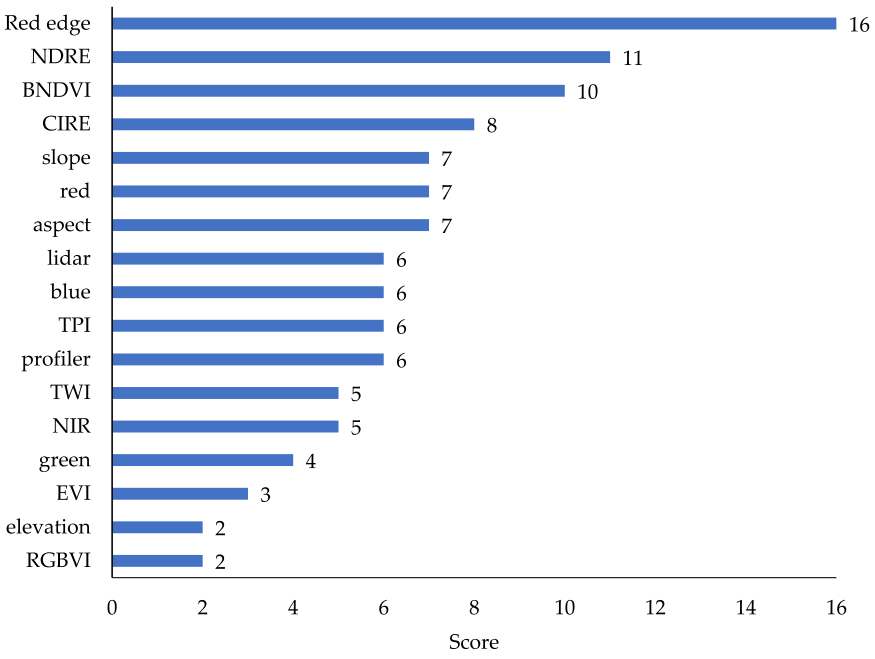

Fig. 5. Summary of feature importance by Random Forest Classifier, the Top 5 variables that appear with more constancy among six different scenarios.

Discussion

This work verified the best input variables related to soil parameters, elevation derivatives, topographic indices, and proximal and orbital remote sensing (VI and Sentinel-2 spectral bands) data for predicting cotton yield. Most studies have used crop parameters (e.g., soil type, plant, date, rainfall, fertilizer, seed variety, etc.) [33]. This work differs from others by using soil variables to predict cotton yield. Furthermore, the input variables as a profiler, S2 spectral bands (blue, red, and green), TPI, LiDAR, and RTK elevation had the best correlations with yield for the prediction of cotton yield.

According to RFC features importance, the best inputs to predict yield were orbital RS (red edge, NDRE, BNDVI, CIRE, red, blue, green, EVI, and RGBVI), soil parameters (slope, aspect, TPI, and TWI), RTK elevation, profiler, and LiDAR (Fig. 5). Using the Spearman correlation coefficient, the best correlation variable to predict cotton yield was a profiler, orbital RS (blue, red, and green), TPI, LiDAR, and RTK elevation (Fig. 6). In addition, S2 spectral bands were better than VI for correlation with yield. In the same way, Wang et al. [34] observed. They evaluated hail damage in cotton and predicted crop yield using RS, wherein the model prediction model of yield reduction due to hail damage was better for spectral than VI.

Orbital RS has been provided to monitor crops during growth state, collecting data over large agricultural areas with high and moderate resolution can be obtained with weekly and daily temporal resolution [35], for example, 16, 5, 1 day Landsat-8, Sentinel-2, and Planet Scope constellations, respectively.

Cotton yield prediction helped farmers make decision support tools

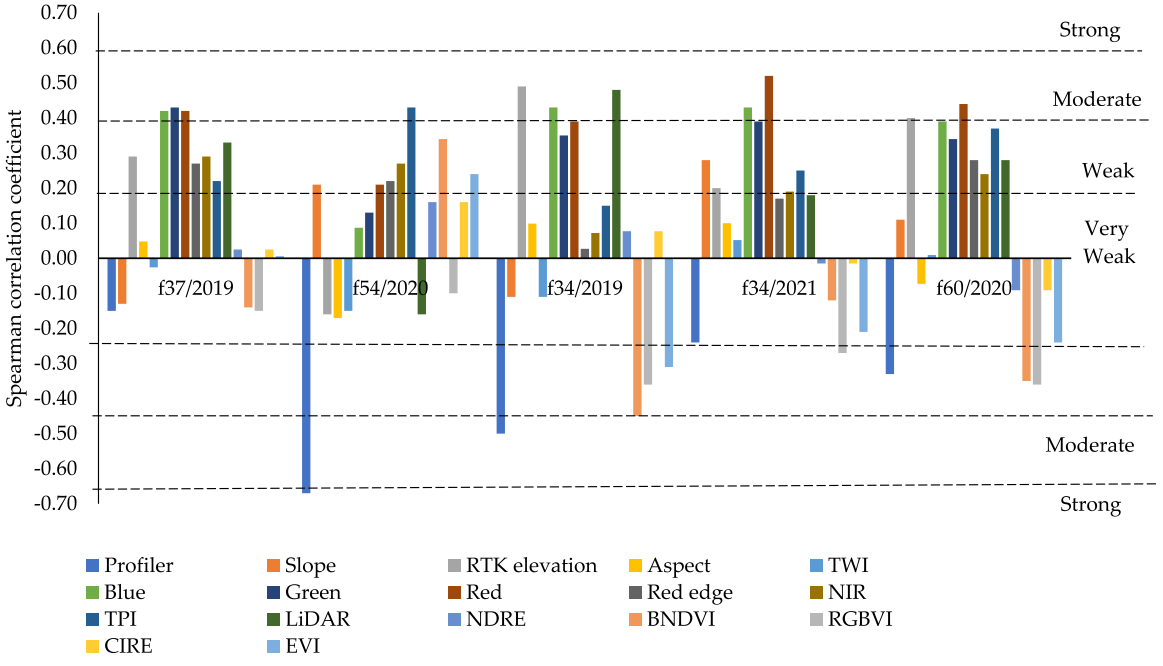

Fig. 6. Spearman coefficient correlation from the Top 5 variables appears more consistent among six different scenarios.

Table 5
Confusion matrix for classification accuracy of different scenarios using Random Forest Classifier for selection of Top 5 and Top 15 for prediction cotton yield.

Inputs*	Field/ year	Top 5 Overall Accuracy	Kappa	Top 15 Overall Accuracy	Kappa
All inputs	37	0.72	0.65	0.80	0.74
Soil	2019	0.68	0.60	0.78	0.72
Profiler + Orbital RS		0.76	0.70	0.78	0.73
LiDAR + Orbital RS		0.75	0.69	0.90	0.87
Orbital RS (VI + S2 bands)		0.76	0.70	0.76	0.70
Orbital RS (S2 bands)		0.76	0.70	—-	—-
All inputs	54	0.79	0.73	0.84	0.80
Soil	2020	0.79	0.73	0.82	1.0
Profiler + Orbital RS		0.78	0.72	0.82	0.78
LiDAR + Orbital RS		0.80	0.75	0.82	0.78
Orbital RS (VI + S2 bands)		0.80	0.75	0.80	0.75
Orbital RS (S2 bands)		0.80	0.75	—-	—-
All inputs	34	0.72	0.65	0.80	0.74
Soil	2019	0.68	0.60	0.78	0.72
Profiler + Orbital RS		0.76	0.70	0.78	0.73
LiDAR + Orbital RS		0.75	0.69	0.90	0.87
Orbital RS (VI + S2 bands)		0.76	0.70	0.76	0.70
Orbital RS (S2 bands)		0.76	0.70	—-	—-
All inputs	34	0.59	0.49	0.82	0.77
Soil	2021	0.67	0.59	0.79	0.73
Profiler + Orbital RS		0.76	0.70	0.78	0.73
LiDAR + Orbital RS		0.75	0.69	0.78	0.72
Orbital RS (VI + S2 bands)		0.76	0.70	0.76	0.70
Orbital RS (S2 bands)		0.76	0.70	—-	—-
All inputs	60	0.72	0.65	0.72	0.65
Soil	2020	0.71	0.64	0.70	0.62
Profiler + Orbital RS		0.68	0.60	0.70	0.63
LiDAR + Orbital RS		0.67	0.58	0.69	0.61
Orbital RS (VI + S2 bands)		0.67	0.58	0.67	0.58
Orbital RS (S2 bands)		0.67	0.58	—-	—-

* RS: Remote Sensing; S2: Sentinel-2 spectral bands; VI: vegetation Indice, Soil: Soil parameters and elevation derivatives.

using RS and machine learning to do this more accurately and before the harvesting date. Traditional techniques demand destructive samples and high costs over large areas because they need more labor to collect samples [8]. While using RS tools, you can collect data over large areas and do not need destructive samples.

Topographic indices such as TPI and TWI help generate management zones [36] and predict crop yield more accurately [37]. In this way, how observed in these works that TPI had more outstanding results among these indices evaluated than TWI for having a more excellent correlation with cotton yield. These topographic indices were generated using RTK elevation available in the sower and harvesting machine during this operation.

LiDAR is used to monitor plant height [10] and can be used to estimate biomass and yield. For the cotton crop, the variable plant height is an essential variable that allows knowing when applying desiccant at the right time that this variable can be collected from LiDAR.

Conclusions

Through brainstorming, we established the maximum value of 15 inputs, and the Random Forest Classifier (RFC) method allowed us to do this. We observed that the results of the models were promising. However, we focused on reducing the number of inputs to be collected to make it easier for the farmers and for them to apply the methodology proposed in this work, maintaining the same quality, so we opted for five inputs. We could work with various amplitudes within our dataset. However, at that moment, we believed these established limits were enough to reach our proposed objective.

This study used different scenarios to select the best inputs to predict cotton yield. RFC is a potential tool for identifying the best feature importance as an excellent filter and working with a large dataset. Spearman's coefficient correlation demonstrates an excellent statistical analysis for nonlinear relationships.

Table 6
Metrics parameters (RSME, R^2, and MAPE) evaluation of different scenarios for predicting cotton yield.

	Field/ year	Top 5 R^2	RSME	MAPE	Top 15 R^2	RSME	MAPE
All inputs	37	42.08	0.15	5.67	71.60	0.15	3.83
Soil	2019	54.78	0.13	4.87	78.22	0.09	3.30
Profiler + Orbital RS		55.10	0.13	4.94	54.43	0.13	5.04
LiDAR + Orbital RS		63.21	0.12	4.42	63.50	0.12	4.41
Orbital RS (VI + S2 bands)		53.12	0.13	5.05	54.40	0.13	5.08
Orbital RS (S2 bands)		52.47	0.13	5.16	—-	—-	—-
All inputs	54	77.26	0.17	4.33	80.84	0.16	4.01
Soil	2020	78.87	0.12	4.23	88.09	0.13	3.43
Profiler + Orbital RS		67.58	0.21	5.48	76.03	0.13	4.61
LiDAR + Orbital RS		71.76	0.19	4.98	73.11	0.19	4.87
Orbital RS (VI + S2 bands)		56.81	0.24	6.54	65.59	0.16	5.71
Orbital RS (S2 bands)		55.26	0.24	6.63	—-	—-	—-
All inputs	34	45.82	0.14	5.61	72.88	0.10	4.02
Soil	2019	62.06	0.12	4.74	76.64	0.09	3.68
Profiler + Orbital RS		66.58	0.11	4.48	71.94	0.10	4.06
LiDAR + Orbital RS		66.76	0.11	4.41	70.00	0.11	4.19
Orbital RS (VI + S2 bands)		51.13	0.14	5.24	56.71	0.13	4.96
Orbital RS (S2 bands)		55.80	0.13	4.98	—-	—-	—-
All inputs	34	32.07	0.20	7.31	64.06	0.15	5.44
Soil	2021	76.07	0.12	4.49	51.00	0.17	4.49
Profiler + Orbital RS		51.75	0.17	6.24	62.18	0.15	5.46
LiDAR + Orbital RS		52.21	0.17	6.16	64.41	0.14	5.22
Orbital RS (VI + S2 bands)		50.93	0.17	6.25	55.94	0.16	5.86
Orbital RS (S2 bands)		51.78	0.17	6.22	—-	—-	—-
All inputs	60	82.78	0.17	5.65	86.22	0.15	5.12
Soil	2020	81.90	0.17	5.77	84.25	0.16	5.38
Profiler + Orbital RS		79.30	0.19	6.37	81.94	0.17	5.87
LiDAR + Orbital RS		78.83	0.20	6.75	79.00	0.18	6.22
Orbital RS (VI + S2 bands)		76.06	0.20	6.72	79.00	0.19	6.22
Orbital RS (S2 bands)		73.57	0.21	7.11	—-	—-	—-

*RS: Remote Sensing; S2: Sentinel-2 spectral bands; VI: vegetation indices, Soil: Soil parameters and elevation derivatives

The inputs evaluated that demonstrated more excellent results to predict cotton yield was profiler, Sentinel-2 (blue, red, and green), TPI, LiDAR, and RTK elevation, which show good correlations to predict cotton yield. In this way, a farmer to collect this data need orbital imagery, can use free satellite imagery, RTK elevation from a sowing machine or LiDAR sensor, and generates topographic indices (TWI and TPI) from RTK elevation, and profiler from the electrical conductivity of the soil sensor.

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.

Data availability

The data that has been used is confidential.

Funding

This research was funded by the Soybean Board (GR-00010525); Cotton INC (GR-00010529); NIFA USDA Hatch (PG-005787); and Multistate Frontier (PG-005787).

Acknowledgments

We thank Hardwick Planting Company, Louisiana State University (LSU), Soybean Board, NIFA USDA Hatch, Multistate Frontier and Cotton INC for their support and collaboration.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.atech.2023.100292.

References

[1] B. Yang, W. Zhu, E.E. Rezaei, J. Li, Z. Sun, J Zhang, The optimal phenological phase of maize for yield prediction with high-frequency UAV remote sensing, Remote Sens. 14 (2022) 3–18, https://doi.org/10.3390/rs14071559.

[2] X. Zhou, H.B. Zheng, X.Q. Xu, J.Y. He, X.K. Ge, X. Yao, T. Cheng, Y. Zhu, W.X. Cao, Y.C. Tian, Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery, ISPRS J. Photogramm. Remote Sens. 130 (2017) 246–255, https://doi.org/10.1016/j.isprsjprs.2017.05.003.

[3] F.M. Carneiro, C.E.A. Furlani, C. Zerbato, P.C. Menezes, L.A.S. Gírio, M.F. Oliveira, Comparison between vegetation indices for detecting spatial and temporal variabilities in soybean crop using canopy sensors, Precis. Agric. 21 (2020) 979–1007, https://doi.org/10.1007/s11119-019-09704-3.

[4] A. Ali, M. Imran, Evaluating the potential of red edge position (REP) of hyperspectral remote sensing data for real time estimation of LAI & chlorophyll content of kinnow mandarin (*Citrus reticulata*) fruit orchards, Sci. Hortic. 267 (109326) (2020) 1–11, https://doi.org/10.1016/j.scienta.2020.109326.

[5] F.M. Carneiro, M.F. Oliveira, S.L.H. Almeida, A.L. Brito Filho, C.E.A. Furlani, G. S. Rolim, A.S. Ferraudo, R.P. Silva, Biophysical characteristics of soybean estimated by remote sensing associated with artificial intelligence, Biosci. J. 38 (e38024) (2022) 1–12, https://doi.org/10.14393/BJ-v38n0a2022-55925.

[6] Á. Maresma, M. Ariza, E. Martínez, J. Lloveras, J.A. Martínez-Casasnovas, Analysis of vegetation indices to determine nitrogen application and yield prediction in maize (*Zea mays* L.) from a standard UAV service, Remote Sens. 8 (12) (2016) 973, https://doi.org/10.3390/rs8120973.

[7] M. Vizzari, PlanetScope, Sentinel-2, and Sentinel-1 data integration for object-based land cover classification in google earth engine, Remote Sens. 14 (2022) 2628, https://doi.org/10.3390/rs14112628.

[8] L. Wang, J. Wang, X. Zhang, L. Wang, F. Qin, Deep segmentation and classification of complex crops using multi-feature satellite imagery, Comput. Electron. Agric. 200 (2022) 2–15, https://doi.org/10.1016/j.compag.2022.107249.

[9] A.F.S. Putri, W. Widyatmanti, D.A. Umarhadi, Sentinel-1 and Sentinel-2 data fusion to distinguish building damage level of the 2018 Lombok Earthquake, Remote Sens. Appl.: Soc. Environ. 26 (2022) 2–13, https://doi.org/10.1016/j.rsase.2022.100724.

[10] M. Gao, F. Yang, H. Wei, X. Liu, Individual maize location and height estimation in field from UAV-borne LiDAR and RGB images, Remote Sens. 14 (10) (2022) 2292, https://doi.org/10.3390/rs14102292.

[11] Yi-Chun; Lin, A. Habib, Quality control and crop characterization framework for multi-temporal UAV LiDAR data over mechanized agricultural fields, Remote Sens. Environ. 256 (2021), 112299, https://doi.org/10.1016/j.rse.2021.112299.

[12] S. Elsayed, P. Rischbeck, U. Schmidhalter, Comparing the performance of active and passive reflectance sensors to assess the normalized relative canopy temperature and grain yield of drought-stressed barley cultivars, Field Crop. Res. 177 (2015) 148–160, https://doi.org/10.1016/j.fcr.2015.03.010.

[13] A.A.A. Aldabaa, D.C. Weindorf, S. Chakraborty, A. Sharma, B. Li, Combination of proximal and remote sensing methods for rapid soil salinity quantification, Geoderma 239-240 (2015) 34–46, https://doi.org/10.1016/j.geoderma.2014.09.011.

[14] E. Erazo-Mesa, A. Echeverri-Sánchez, J.G. Ramírez-Gil, Advances in Hass avocado irrigation scheduling under digital agriculture approach, Revista Colombiana de Ciencias Hortícolas 16 (1) (2022) e13456, https://doi.org/10.17584/rcch.2022v16i1.13456.

[15] D. Benedetto, A. Castrignano, M. Diacono, M. Rinaldi, S. Ruggieri, R. Tamborrino, Field partition by proximal and remote sensing data fusion, Biosyst. Eng. 114 (4) (2013) 372–383, https://doi.org/10.1016/j.biosystemseng.2012.12.001.

[16] J.A. Fontes, M.J. Anzanello, J.B.G. Brito, G.B. Bucco, F.S. Fogliatto, F.P. Puglia, Combining wavelength importance ranking to the random forest classifier to analyze multiclass spectral data, Forens. Sci. Int. 328 (110998) (2021) 1–10, https://doi.org/10.1016/j.forsciint.2021.110998.

[17] USDA webpage, Web Soil Survey, 2023. Retrieved April 19,from, https://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx.

[18] LSU Department of Geography & Anthropology, *Atlas Louisina GIS:* Louisiana Statewide LiDAR, 2023. Retrieved April 19,from, https://atlas.ga.lsu.edu/datasets/LiDAR2000/.

[19] G.S. Birth, G. McVey, Measuring the color of growing turf with a reflectance spectrophotometer, Agron. J. 60 (1968) 640–643.

[20] J.W. Rouse Jr, R.H. Haas, J.A. Schell, D.W Deering, Monitoring vegetation systems in the great plains with ERTS, in: Proceeding pf the N. SP-351, Ed. Third ERTS symposium, 1, Washington, USA, NASA, 1974, pp. 309–317.

[21] A.A. Gitelson, Y. Kaufman, M.N. Merzlyak, Use of a green channel in remote sensing of global vegetation from EOS-MODIS, Remote Sens. Environ. 58 (3) (1996) 289–298, https://doi.org/10.1016/s0034-4257(96)00072-7.

[22] N. Goel, W. Qin, Influences of canopy architecture on relationships between various vegetation indices and LAI and Fpar: a computer simulation, Remote Sens. Rev. 10 (4) (1994) 309–347, https://doi.org/10.1080/02757259409532252.

[23] B.C. Gao, NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space, Remote Sens Environ 58 (3) (1996) 257–266, https://doi.org/10.1016/S0034-4257(96)00067-3.

[24] C.F. Jordan, Derivation of leaf area index from quality of light on the forest floor, Ecology 50 (1969) 663–666.

[25] F.-m. Wang, J.-f. Huang, Y.-l. Tang, X.-z. Wang, New vegetation index and its application in estimating leaf area index of rice, Rice Sci. 14 (3) (2007) 195–203, https://doi.org/10.1016/S1672-6308(07)60027-4.

[26] N.G. Silleos, T.K. Alexandridis, I...Z.. Gitas, K. Perakis, Vegetation indices: advances made in biomass estimation and vegetation monitoring in the last 30 years, Geocarto Int 21 (2006) 21–28, https://doi.org/10.1080/10106040608542399.

[27] J. Bendig, K. Yu, H. Aasen, A. Bolten, S. Bennertz, J. Broscheit, M.L. Gnyp, G. Bareth, Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley, Int. J. Appl. Earth Obs. Geoinf. 39 (2015) 79–87, https://doi.org/10.1016/j.jag.2015.02.012.

[28] A.R. Huete, A soil-adjusted vegetation index (SAVI), Remote Sens. Environ. 25 (1988) 295–309.

[29] A.A. Gitelson, Y. Gritz, M.N. Merzlyak, Relationships between leaf chlorophyll content and spectral reflectance algorithms for non-destructive chlorophyll assessment in higher plants, J. Plant Physiol. 160 (3) (2003) 271–282, https://doi.org/10.1078/0176-1617-00887.

[30] R. Crippen, Calculating the vegetation index faster, Remote Sens. Environ. 34 (1) (1990) 71–73, https://doi.org/10.1016/0034-4257(90)90085-Z.

[31] A. Huete, K. Didan, T. Miura, E.P. Rodriguez, X. Gao, L.G. Ferreira, Overview of the radiometric and biophysical performance of the MODIS vegetation indices, Remote Sens. Environ. 83 (1–2) (2002) 195–213, https://doi.org/10.1016/S0034-4257(02)00096-2.

[32] Z. Jiang, A.R. Huete, K. Didan, T Miura, Development of a two-band enhanced vegetation index without a blue band, Remote Sens. Environ. 112 (10) (2008) 3833–3845, https://doi.org/10.1016/j.rse.2008.06.006.

[33] D.M. Johnson, A. Rosales, R. Mueller, C. Reynolds, R. Frantz, A. Anyamba, E. Pak, C. Tucker, USA crop yield estimation with MODIS NDVI: are remotely sensed models better than simple trend analyses? Remote Sens. 13 (21) (2021) 4227, https://doi.org/10.3390/rs13214227.

[34] L. Wang, Y. Liu, M. Wen, M. Li, Z. Dong, Z. He, J. Cui, F. Ma, Using field hyperspectral data to predict cotton yield reduction after hail damage, Comput. Electron. Agric. 190 (2021), 106400, https://doi.org/10.1016/j.compag.2021.106400.

[35] U. Alganci, M. Ozdogan, E. Sertel, C. Ormeci, Estimating maize and cotton yield in southeastern Turkey with integrated use of satellite images, meteorological data and digital photographs, Field Crop. Res. 157 (2014) 8–19, https://doi.org/10.1016/j.fcr.2013.12.006.

[36] M.S. Mieza, W.R. Cravero, F.D. Kovac, P.G. Bargiano, Delineation of site-specific management units for operational applications using the topographic position index in La Pampa, Argentina, Comput. Electron Agric. 127 (2016) 158–167, https://doi.org/10.1016/j.compag.2016.06.005.

[37] M.F. Oliveira, *Forecast and estimation of cultivation variables using remote sensing levels and forms and machine learning techniques*. P.h.D degree. School of Agricultural and Veterinarian Sciences, São Paulo State University (Unesp, Jaboticabal, 2021. Retrieved December 12, 2022, from, https://repositorio.unesp.br/handle/11449/210982.