Journal of Technology and Innovation June 2024 Vol.11 No.28 1-13
Article

Machine-Learning model for estimating sugarcane production at crop level

Modelo Machine-Learning para estimación de índice de producción de caña de azúcar a nivel de cultivo

Lárraga-Altamirano, Hugo René* [a], Hernández-López, Dalia Rosario [b], Piedad-Rubio, Ana María [c] and Blanco-Martínez, José Ramón[d]

[a] Tecnológico Nacional de México - Instituto Tecnológico de Ciudad Valles • T-2296-2018 • 0000-0001-8258-9418 • 626539
[b] Tecnológico Nacional de México - Instituto Tecnológico de Ciudad Valles • T-2470-2018 • 0000-0002-2751-5886 • 536472
[c] Tecnológico Nacional de México - Instituto Tecnológico de Ciudad Valles • T-2477-2018 • 0000-0003-1258 • 732279
[d] Tecnológico Nacional de México - Instituto Tecnológico de Ciudad Valles • KBB-6715-2024 • 0009-0005-2456-4673 • 1348457

CONAHCYT classification:

Area: Engineering
Field: Engineering
Discipline: System engineer
Subdiscipline: Computer Sciences

https://doi.org/10.35429/JTI.2024.28.11.1.13
History of the article:
Received: January 15, 2024
Accepted: June 30, 2024

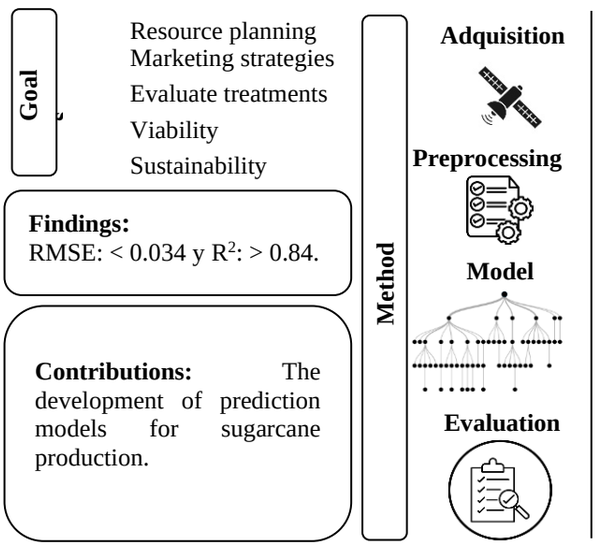
* ✉ [dalia.hernandez@tecvalles.mx]

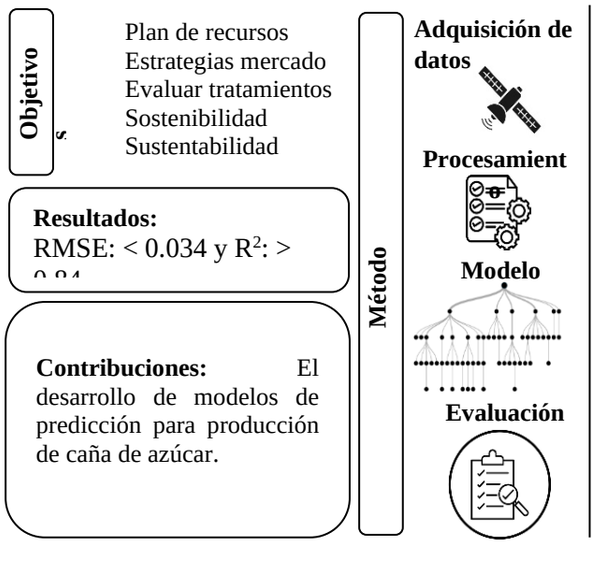
Abstract

Yield maps provide essential information for those who manage the field. The anticipated production data will be able to make better decisions on how resources should be used in harvesting, define market strategies and, above all, it will help evaluate treatments used on the crop. Sugar cane is the predominant crop in Huasteca Potosina, Mexico. The proposed Machine Learning model based on Random Forest Regressor integrates time series of vegetation indices extracted from Sentinel-2 images and meteorological data. The R^2 and RMSE metrics (0.84 y 0.034) show the effectiveness of the model for prediction.

Resumen

Los mapas de producción de cultivo proveen información esencial para quienes administran el campo. El dato anticipado de la producción permitirá tomar mejores decisiones sobre los recursos a ocupar en la cosecha, definir estrategias de mercado y, sobre todo, servirá para evaluar tratamientos utilizados sobre el cultivo. La caña de azúcar es el cultivo predominante en la Huasteca Potosina, México. El modelo Machine Learning propuesto basado en un Random Forest Regressor integra series de tiempo de índices de vegetación extraídas de imágenes Sentinel-2, y datos meteorológicos. Las métricas R^2 y RMSE (0.84 y 0.034) muestran la efectividad del modelo para la predicción.





Yield estimation, sugarcane, random forest regressor

Estimacion de la producción, caña de azúcar, random forest regressor

Citation: Lárraga-Altamirano, Hugo René, Hernández-López, Dalia Rosario, Piedad-Rubio, Ana María and Blanco-Martínez, José Ramón.Machine-Learning model for estimating sugarcane production at crop level. Journal of Technology and Innovation. 2024. 11-28:1-13.

Introduction

Crop production maps provide essential information for field managers, providing indicators to guide decisions towards best practice in agriculture. The importance of knowing the production rate in advance of harvest lies in better understanding the variability of each part of the crop over a complete cycle, and thus zoning plots and improving the application of site-specific strategies (Xu et al., 2020).

Sugarcane is the crop with the highest production worldwide; in Mexico alone, 772,003 hectares were harvested in 2017, which produced 56,954,993 tonnes of this grass, mostly used by the sugar industry (FAO, 2019). The Huasteca Potosina is the regional context where the present project is developed, the cultivation of sugar cane predominates over others, contributing significantly to the national production of this grass. The national production index per hectare from 2007 to 2016 decreased by 4.2%, showing very marked ups and downs and closing at approximately 70 ton/ha (SIAP, 2019).

Therefore, in the particular case of sugar cane in the Huasteca Potosina region, Mexico, the trend is to better manage resources under a sustainability scheme. The efficient and sustainable production of sugar cane is an important issue, since on the one hand there is a tendency to better manage the field with fewer resources, and on the other, to reduce the negative impact of agricultural activities on the environment (Said Mohamed et al., 2021).

Information and communication technologies play a central role in building smart farm models that provide the farmer with information acquired from different media, allowing real-time monitoring of plots to plan their activities in response to changing circumstances (O'Grady & O'Hare, 2017).

The study of agricultural fields through Remote Sensing (RP) techniques has empowered Smart Agriculture, being the basis for the development of technological solutions that favour the implementation of activities such as crop monitoring. RP methods that provide satellite images as a basis for statistical analysis are potentially suitable for producing data of good accuracy on some biophysical variables, from which reliable, efficient and timely estimates of crop condition can be obtained.

This potential is largely due to the information that is generated on a frequent basis from revisits of satellite platforms (Tovar Blanco et al., 2020).

From satellite images it is possible to construct time series based on the calculation of vegetation indices (VI), another important component that adds value to crop characterisation. IVs correlate with crop development, examples of these are: Normalized Difference Vegetation Index (NDVI), Green Normalized Difference Vegetation Index (GNDVI), Normalized Difference Water Index (NDWI), Green-Red Vegetation Index (GRVI), Leaf Area Index (LAI), among others (K.C. et al., 2021).

Machine Learning (ML) methods, either supervised or unsupervised, have been used to transform the large amount of spectral time series data into information useful to the producer, through classification or regression models that can estimate crop production. The ML method integrates a learning process from a set of examples (training), each example describing crop attributes also known as variables or characteristics. In other words, the ML method has the ability to relate each input characteristic to a corresponding output value. Of the techniques that have demonstrated greater accuracy in prediction compared to traditional statistical methods are Radom Forest (RF), Artificial Neural Network (ANN) and Support Vector Machine (SVM). Highlighting that RF in agricultural data analysis applications has proven to be one of the best non-parametric statistical methods for classification and regression due to its high estimation accuracy, high computational speed, robustness and ability to predict important variables (Felipe Maldaner et al., 2021).

The importance of knowing the production index in advance of harvest lies in better understanding the variability of each part of the crop during a complete cycle, and thus zoning plots and improving the application of site-specific strategies. There are several methods of estimating crop production rate: 1) using growth models (CGMs) that simulate the physical process of the crop; 2) through high and low resolution satellite imagery with which it is possible to cover larger areas; 3) through light detection and ranging (LiDAR) data with which height information is obtained due to its very high spatial resolution (Xu et al., 2020).

ISSN: 2410-3993
RENIECYT-CONAHCYT: 1702902

Lárraga-Altamirano, Hugo René, Hernández-López, Dalia Rosario, Piedad-Rubio, Ana María and Blanco-Martínez, José Ramón.Machine-Learning model for estimating sugarcane production at crop level. Journal of Technology and Innovation. 2024. 11-28:1-12. https://doi.org/10.35429/JTI.2024.28.11.1.13

The present research project is focused on the development of a technological solution that supports the decision making of sugarcane producers in the region, in other words, producers must make decisions based on a greater amount of information that gives certainty to the economic viability without leaving aside the environmental friendliness. To this end, it is proposed to design a sugar cane production estimation model based on Machine Learning (ML) algorithms to predict the number of tonnes per hectare expected to be harvested during the season. The anticipated production data will allow better decisions to be made about the resources to be used in the harvest, define market strategies and, above all, serve as a parameter for evaluating the chemical or organic treatments used in the season.

Background

The estimation of the sugar cane production index has been a subject addressed in different research studies around the world, which show the diversity of methods applied for its calculation. Of the methods applied, those that use remote sensing for the study of agricultural surfaces stand out, and currently, in conjunction with machine learning techniques, have managed to improve the estimation models. This section mentions work carried out by different authors who contributed to this project.

The production prediction model presented in (Arab et al., 2021), oriented to grape cultivation, exposes the importance of the crop phases in the production season, as well as the usefulness of production maps for the producer, showing the variability along the field to determine the best harvesting time and market strategies. An ML model was implemented together with time series satellite imagery based on NDVI, LAI and NDWI. Statistical methods were used to determine the different crop growth stages, particularly in the study of NDVI behaviour, this seasonality was treated by a representative mean for each crop. A regression model was implemented through an artificial neural network (ANN). The prediction results were compared against field data, finding that the model can be applied to estimate yield indices.Meanwhile (Canata et al., 2021), highlights the importance of the use of production maps in decision making in the context of Precision Agriculture (PA). Using multi-temporal orbital imagery and ML techniques, a sugar cane production estimation model was proposed.

The satellite platform used was Sentinel 2, obtaining images from sowing to harvesting of the crops studied. The model based on RF (Random Forest) and MLR (Multiple Linear Regression) used the dataset of satellite images and crop data filtered and interpolated to the same spatial resolution as the images, was divided into a training and test set. The near infrared band showed a large contribution to the yield estimate.

In (Singla et al., 2020), the use of ML algorithms and the use of remote sensing to extract agricultural information was proposed for the construction of sugarcane prediction models. As a first attempt, the parameters to be used as model input were determined through the Mean Decline Accuracy and Mean Gini Decline metrics of the Random Forest (RF) algorithm. It was noted that GNDVI, NDVI and Land Surface Water Index obtained the best results among other indices. The objective of the proposed work focused on machine learning methods to optimise the correlation of historical crop yield values with spectral information. The RF method shows significant performance compared to other methods such as classification and regression tree, support vector regression and nearest neighbour.

In (Jeena Jacob et al., 2021, Chapter 58) a crop yield estimation model was implemented using a multilayer Perceptron neural network and Random Forest Regression, trained with data from 4 crops, weather information and yield data. Climatological information included maximum, minimum and average values of temperature, humidity and pressure. The metrics used to evaluate the models were: absolute error (MAE), mean square error (MSE) and root mean square error (RMSE). For real-time prediction, a web application was created using Python, Flask where the user accesses the trained model to predict performance.

Methodology

Study area

The study area is a commercial crop with an area of 100 ha of which 80 ha are used for production, divided into 17 plantations, Figure 1. Located in the municipality of Ciudad Valles, San Luis Potosi, Mexico (22.0038508° N -99.0496236° W 78.0256153 m), whose climate is warm sub-humid with summer rains.

ISSN: 2410-3993
RENIECYT-CONAHCYT: 1702902

Journal of Technology and Innovation June 2024 Vol.11 No.28 1-13
Article

Box 1

Figure 1
Study crop, "El Tuzo" ranch, Ciudad Valles, S.L.P., México

Source: Google Earth
https://earth.google.com/web/@22.00020428,-99.05020569,78.37226312a,4280.28218377d,30y,0h,0t,0r

All fields within the crop have the same cane variety CP 722086, the production season lasts 12 months, the harvest months are from October to December each year.

Data acquisition

The data collection includes variables that affect the sugar cane production rate, such as spectral information from satellite images, climatological data measured by weather stations and the production data of the studied crop. The multispectral images are obtained from the Sentinel-2 platform, whose constellation consists of two satellites (A and B), equipped with an MSI (MultiSpectral Instrument) sensor with 13 spectral resolution bands, with spatial resolutions of 10 m, 20 m and 60 m and a revisit frequency of 10 days for each satellite with a difference of 5 days between them. The images are downloaded through the free Copernicus Hub repository for the period 2018-2023, each image covers an area of 110 km x 110 km with a UTM/WGS84 projection system (Canata et al., 2021).

The climatological data collected must coincide with the period of the image download, so that the variables of precipitation, evapotranspiration, average, maximum and minimum temperatures of each production cycle are known (Yu et al., 2020). In Mexico, there are weather stations administered by CONAGUA (National Water Commission), which makes this information available on its official website. The stations located near the study crop are the following (Government of Mexico, 2023):

- 24076 Santa Rosa.
- 24012 Ciudad Valles.
- 24043 Micos.

Once the data have been homogenised they must be geospatially interpolated within the crop area, considering as reference the known locations of the meteorological stations, applying the kriging and IDW interpolation algorithms (Shukla et al., 2020). On the other hand, production measurements for each crop fraction (block) are collected through the sugarcane producer in tonnes/hectare (ton/ha). Additionally, data related to the supply of fertilisers, herbicides and pesticides, as well as irrigation management and timing, crop age and finally sowing and harvesting dates are obtained for each block (Hammer et al., 2020).

Data adequacy

Feature extraction

The vegetation indices known as NDVI, GNDVI, LAI and NDWI are calculated, these characterise the sugar cane crop in its different phenological phases through the production cycles, making it possible to extract statistical data from each image over time. The mean, maximum and minimum value, standard deviation, quartiles 2 and 3 are obtained. Through the NDVI statistics and extracting only the cultivated surface area (segmentation), the time series is visualised over the period 2019-2023, thus, the behaviour of production with respect to the mentioned index is observed. Based on the NDVI time series, phenological characteristics related to the crop cycle are calculated, taking advantage of the advantages of this index in identifying the density of the vegetation with high values. Phenological characteristics such as (Dimov et al., 2022):

- Start of season (SOS).
- Date of highest NVDI peak of the season (POS).
- End of Season (EOS)
- Maximum NDVI value POS
- Average sum of NDVI values
- Average sum of maximum NDVI values POS
- Length of season in days
- First half of the season, days between SOS and POS
- Second half of the season, day between POS and EOS

ISSN: 2410-3993
RENIECYT-CONAHCYT: 1702902

The yield dataset was obtained from the transformation of EVI (Enhanced Vegetation Index) values to the yield index reported by the crop manager (Ji et al., 2021).

Selection of characteristics

The number of features infers in the estimation process, variables with a low correlation with respect to the sugarcane yield index cause an unreliable prediction, also, the computation time is increased due to the high dimensionality of the model. This is why feature selection is desirable for better interpretation and increased estimation efficiency. The study of correlation between independent and dependent variables is of importance for feature selection, alternatively there are models such as PCA (Principal Component Analysis) and EFA (Exploratory Factor Analysis) to identify those independent variables with a high correlation with respect to the dependent variable (Singla et al., 2020).

Model design

In this study, the Random Forest Regressor (RFR) algorithm is proposed to analyse the non-linearity of the predictor variables in relation to the dependent variable (production index), the model can be tested using a varied set of input data. For each characteristic, the RFR identifies the degree of importance of all selected predictor variables. The RFR is comprised of multiple independent decision trees, which average their estimates to minimise absolute error and handle a high dimensionality data set such as time series. The tuning of the algorithm considers the adjustment of hyperparameters to achieve a higher degree of prediction efficiency. The number of decision trees, the depth of the tree and the maximum number of features to be analysed are the parameters to be tuned. The model training estimates that 70% of the data is used for the learning process, while the remaining 30% is used for validation (Everingham et al., 2016).

Model evaluation

To measure acceptable model accuracy, the evaluation metrics Rr2, root mean square error RMSE and mean absolute deviation MAE are applied, using equations 01, 02 and 03.

$$R^2 = 1 - \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n}(y_i - \overline{y_i})^2} \quad [1]$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad [2]$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad [3]$$

Where n is the number of samples and yi and ^y the observed and estimated values by the model respectively and ---y the means of all observations. The higher the R2 value, the lower the RMSE and MAE, which can be concluded as an effective prediction of the model (Wang et al., 2022).

Results

Data acquisition

The images were acquired from the Copernicus platform managed by the European Space Agency (ESA). The search configuration specified the study crop, the query period, the 2A and 2B satellites that form the Sentinel-2 constellation, see Figure 2.

Box 2

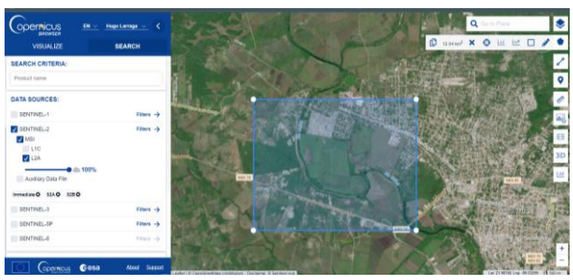


Figure 2
Copernicus Hub Platform

Source: ESA https://dataspace.copernicus.eu

Also, the product level was determined as L2A, i.e. orthorectified images with reflectance levels below the atmosphere. The cloud cover percentage was kept at 100%, as it is not known with certainty whether the crop surface will be covered by clouds. Table 1 shows the number of images downloaded from the above-mentioned platform.

ISSN: 2410-3993
RENIECYT-CONAHCYT: 1702902

Box 3

Table 1

Sentinel-2 images downloaded

Year	S2 scenes
2019	71
2020	64
2021	46
2022	47
2023	62

Source: Own elaboration

The pre-processing of this image gallery was carried out in 3 stages:

1) Image cropping using a vector format mask to obtain only the area where the study crop is located, as presented in Figure 3. It is worth remembering that the Sentinel-2 images have an extension of 110 x 110 km, therefore, it is necessary to extract only the crop image. The download package consists of a set of images in different regions of the electrogenic spectrum with different spatial resolutions, the average data weight is 1Gb.

Box 4

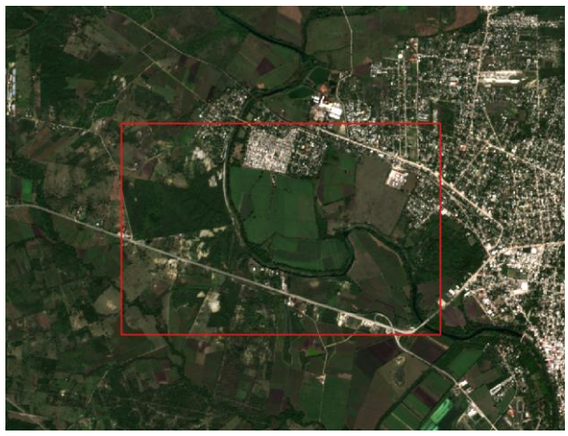


Figure 3

Sentinel-2 image crop

Source: Own elaboration

2) Increase the resolution of bands 5, 6, 7, corresponding to the Red Edge region; band 12 shortwave infrared and the SCL (Scene Classification Map) product. These images are used for the calculation of vegetation indices and cloud masking operations. It is desirable to work with the highest resolution spectral information, therefore, the aforementioned bands are transformed from 20m to 10m resolution. The operations implemented for this purpose, occupy a bilinear interpolation, taking band 4 (visible red) as the basis of transformation.

3) Calculation of the vegetation indices NDVI, GNDVI, LAI and NDWI. These combinations of spectral bands allow the study of crop conditions of interest to the grower. An example of the NDVI index is shown in Figure 4.

Box 5

Figure 4

Example of NDVI vegetation index

Source: Own elaboration

Time series

The time series that were analysed are 2, according to the type of data they contain: the vegetation index series and the climatological data series.

The information of the latter was provided by the CONAGUA agency of the state of San Luis Potosí, comprised the period 2018-2022, with the following variables:

- Monthly average temperature
- Date of the extreme minimum temperature
- Date of the extreme minimum temperature
- Maximum daily precipitation in the month
- Date of the maximum daily precipitation in the month
- Date of the maximum daily precipitation in the month

Total monthly evaporation

From the above list, only two meteorological characteristics were considered for study, those that have the greatest influence on crop development, namely temperature and precipitation.

ISSN: 2410-3993
RENIECYT-CONAHCYT: 1702902

The providers of this information were 5 meteorological stations distributed in the municipality of Ciudad Valles, S.L.P., identified by key and name according to the following list:

- 24076 Santa Rosa.
- 24012 Ciudad Valles.
- 24043 Micos.
- 24065 San Felipe.
- 24028 El Tigre.

The homogenisation of the climatological information was carried out by means of the CLIMATOL software. The cleaning of the data foresees that there are values affected by instrumentation failures, omission, human error, etc. Stations 24065 and 24028 were discarded due to lack of information on the variables of interest.

The Kriging and IDW algorithms and the QGIS geographic information system were used to interpolate the temperature and precipitation values over the crop space, see Figure 5.

Box 6

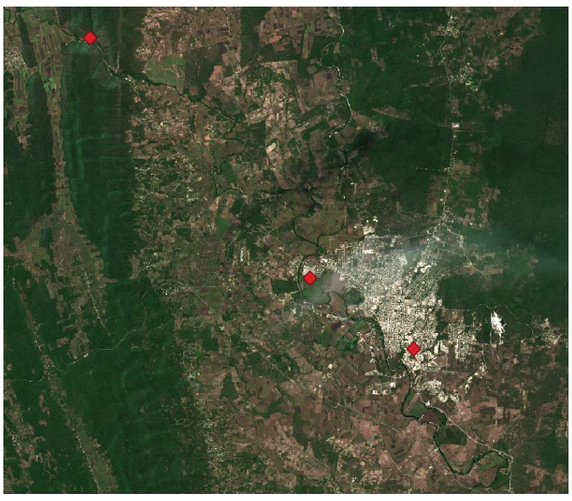


Figure 5
Weather stations near the study crop
Source: Own elaboration

The result of this operation was 12 files in raster format with interpolated monthly temperature values and 12 with monthly precipitation values. In total 96 rasters were processed for the seasons 2019-2022.

For the vegetation index time series we first calculated the overall statistics for each image, selecting only the pixels belonging to the crop by means of a binary mask presented in Figure 6.

Box 7

Figure 6
Binary mask of the study crop
Source: Own elaboration

The statistics calculated from the pixel collection were:

- Number of crop pixels.
- Percentage of the crop classified as vegetation according to the SCL.
- Minimum value.
- First quartile ['25%'].
- Minimum value.
- Mean.
- Third quartile ['75%'].
- Maximum value.
- Standard deviation.
- Histogram of pixels in the ranges -1 to 1.

Statistical values were stored in CSV format for quick retrieval. The graphical representation of the NDVI time series using the general statistics is shown in Figure 7. Cloud occlusion is a common phenomenon in satellite images, for this case the SCL product was used to filter out classes 4 and 5 corresponding to vegetation and classes 8, 9 to low and high probability of clouds. The scenes affected in a percentage greater than 30% of the crop surface by the cloud masking operation were recalculated by cubic interpolation. Subsequently, to plot the time series, the lowest, average and maximum values were smoothed by applying a moving average and spline interpolation to minimize the overall curvature of the plot, see Figure 8.

ISSN: 2410-3993
RENIECYT-CONAHCYT: 1702902

Box 8

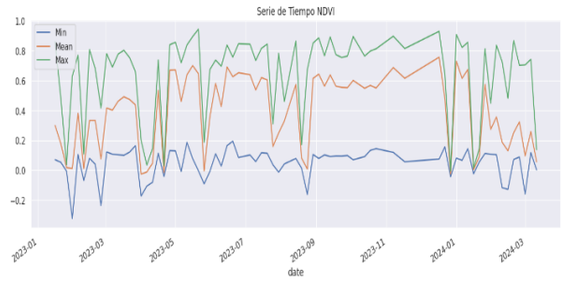


Figure 7

NDVI time series plot with statistical values (2023-2024)

Source: Own elaboration

Box 9

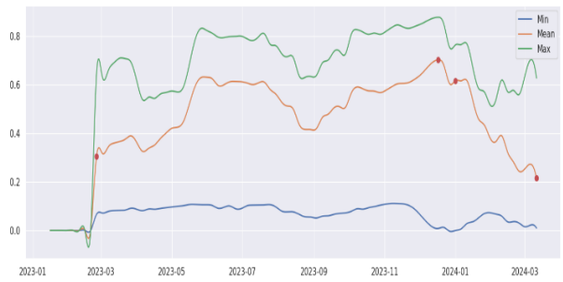


Figure 8

NDVI time series plot with smoothing and phenological data (2023-2024)

Source: Own elaboration

Once the time series has been processed it is possible to identify phenological variables. In Figure 8 the beginning of the season, the maximum NDVI value in the season and the end of the season are marked with red dots. Table 2 contains the details of the phenological information.

Box 10

Table 2

Phenological variables of the NDVI time series

CICLO	SOS	SOS_VAL	POS	POS_VAL	EOS
2019	2019-02-06	0.219916	2019-11-08	0.719613	2020-03-17
2020	2020-03-17	0.302193	2020-10-08	0.654029	2021-02-15
2021	2021-02-15	0.265945	2021-08-19	0.652035	2022-03-02
2022	2022-03-02	0.307735	2022-10-28	0.681958	2023-02-25
2023	2023-02-25	0.302675	2023-12-17	0.703349	-

Source: Own elaboration

Input data

The input data consisted of six sets of characteristics for each season. Each set is composed of the seasonal production index as the dependent variable and the features or independent variables that are hypothesised to be related to production.

The spectral and climatic data were transformed to CSV format along with the coordinates (x,y) so as not to lose the location of each element on the 2D image. These sets were labelled for control as follows:

DS1: time series of vegetation indices between SOS and EOS period.

DS2: 4 randomly selected scenes of the growth phase, between the SOS and POS period.

DS3: 4 randomly selected scenes from the senescence phase, between the POS and EOS period.

DS4: DS3 + Sentinel-2 bands 10m resolution (2, 3, 4, 5, 6, 7, 8, 12).

DS5: DS3 smoothed by sliding window averaging operation with 3x3 kernel.

DS6: DS3 + weather variables January-December.

The production index for each season is provided by the crop manager, based on the record of the sugar mill where the sugar cane enters. Table 3 shows the production for each season.

Box 11

Table 3

Seasonal production rates

Year	ton/ha
2019	89
2020	70
2021	70
2022	78

Source: Own elaboration

To estimate the production index at crop image level, the spatial resolution and the EVI vegetation index were considered. The process starts with the selection of a pre-harvest scene to estimate the EVI index, Figure 9 presents the histogram which like NDVI index the range of values is from -1 to 1.

ISSN: 2410-3993
RENIECYT-CONAHCYT: 1702902

Journal of Technology and Innovation June 2024 Vol.11 No.28 1-13
Article

Box 12

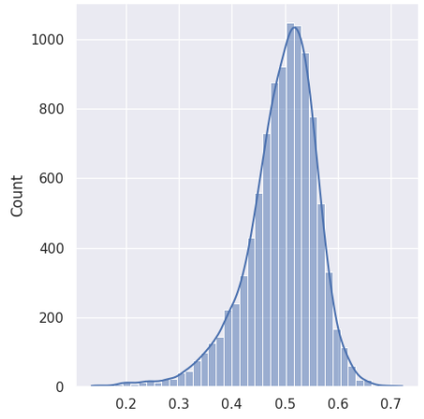


Figure 9
EVI vegetation index histogram
Source: Own elaboration

Outliers were identified as shown in Figure 10, which were removed from the set.

Box 13

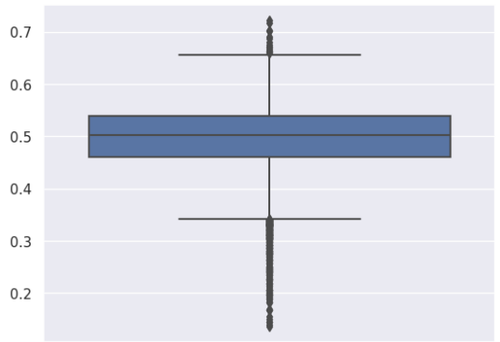


Figure 10
Box plot of EVI values
Source: Own elaboration

After data cleaning, a scaling operation was performed for each of the EVI values and the range determined by the minimum and maximum value of the set, so that the mean of the new set is the expected production index.

Training and evaluation

Hyperparameter tuning of the RFR model was achieved using the Cross Validation technique implemented in a Grid Search.

The number of subsets tested was k = 5, the result is shown in Table 4.

Box 14

Table 3
Best hyperparameters, crossvalidation

Hyperparameters	Value
Number of trees	500
Branch depth	50
Maximum Characteristics	sqrt
Random_State	18
RMSE	-0.01667882

Source: Own elaboration

A significance analysis of the spectral variables was carried out to determine their relationship with the production index and to determine the most appropriate time frame for estimation.

Figure 11 shows the most suitable period for estimation, the dotted line represents the POS of the season. The months of November and December show a higher importance, therefore, it is the time where the prediction will have a higher accuracy.

Box 15

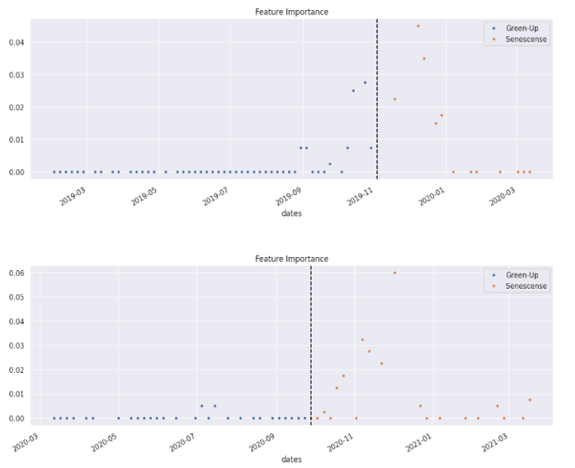


Figure 11
Spectral scene importance analysis
Source: Own elaboration

The RFR training was implemented through Google Colab and the Python programming language.

The 6 input datasets were evaluated for the 2019, 2020 and 2021 seasons. The data are cleaned of outliers, the pixels for each crop that were trained are shown in Table 5.

ISSN: 2410-3993
RENIECYT-CONAHCYT: 1702902

Box 16

Table 5
Data available for training

Training	Test	Total
6422	2753	9175
6722	2881	9603
6483	2779	9262

Source: Own elaboration

The results of the RFR model are shown in Table 6, metrics such as R^2 and RMSE are considered to be the most effective indicators of model performance.

Box 17

Table 6
RFR model results

Ciclo	Input	Variables	MAE	MAPE	RMSE	R^2
2019	1	240	0.02	97.21	0.03345754	0.85
2019	2	16	0.04	95.02	0.05588784	0.59
2019	3	16	0.03	96.72	0.03892498	0.8
2019	4	48	0.02	97.66	0.02675273	0.91
2019	5	16	0.03	96.23	0.04462116	0.74
2019	6	40	0.03	97.07	0.03486532	0.84
2020	1	192	0.02	97.21	0.02617325	0.89
2020	2	16	0.03	94.92	0.04768077	0.65
2020	3	16	0.02	96.34	0.03397908	0.82
2020	4	48	0.02	97.53	0.02298952	0.92
2020	5	16	0.03	96.24	0.03617964	0.8
2020	6	40	0.02	97.16	0.02671924	0.89
2021	1	140	0.01	98.23	0.01865486	0.92
2021	2	16	0.03	95.54	0.04167443	0.61
2021	3	16	0.02	96.87	0.03212605	0.77
2021	4	48	0.01	98.06	0.01864354	0.92
2021	5	16	0.02	96.66	0.03353144	0.75
2021	6	40	0.02	97.65	0.02365406	0.87

Source: Own elaboration

Discussion of results

Table 6 shows that set 4, integrated by the vegetation indices of the senescence phase plus the Sentinel-2 bands with resolution at 10m obtained an R2 of 0.91, 0.92 and 0.92 for the studied seasons, at the same time they present the lowest RMSE errors, being 0.0267, 0.0229 and 0.0186 respectively. 0267, 0.0229 and 0.0186 respectively, however, there are other sets with encouraging results such as 1 and 6. Considering that these are the 3 data sets with the highest number of variables, they explain more effectively the relationship with the production index.

The analysis of the importance of the characteristics of set 1 can be seen in Figure 12 for the 2019 season.

Box 18

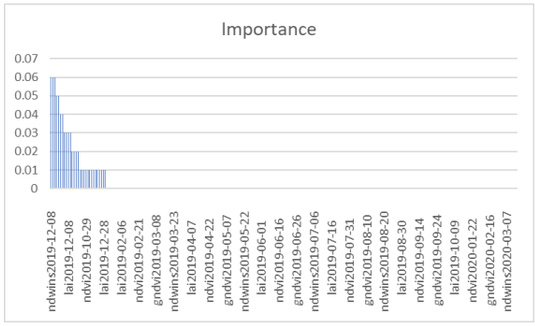


Figure 12
Importance of characteristics set 1 2019 season
Source: Own elaboration

This behaviour is the same for the remaining seasons, so that even when the model performs well, not all variables contribute value, making unnecessary computation with risk of failure. 80% of the characteristics of model 6 were important in the calculation of the estimate (Figure 13).

Box 19

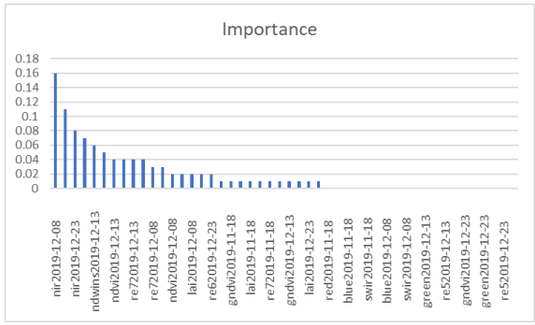


Figure 13
Importance of characteristics set 6 2019 season
Source: Own elaboration

Box 20

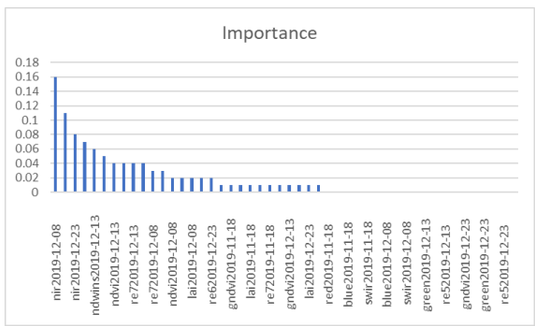


Figure 14
Importance of characteristics set 4 2019 season
Source: Own elaboration

Model 4 registers importance for only 58% of its characteristics (Figure 14). This behaviour is the same for the 2020 and 2021 seasons. The prediction behaviour of the models can be studied in Figure 15.

ISSN: 2410-3993
RENIECYT-CONAHCYT: 1702902

Box 21

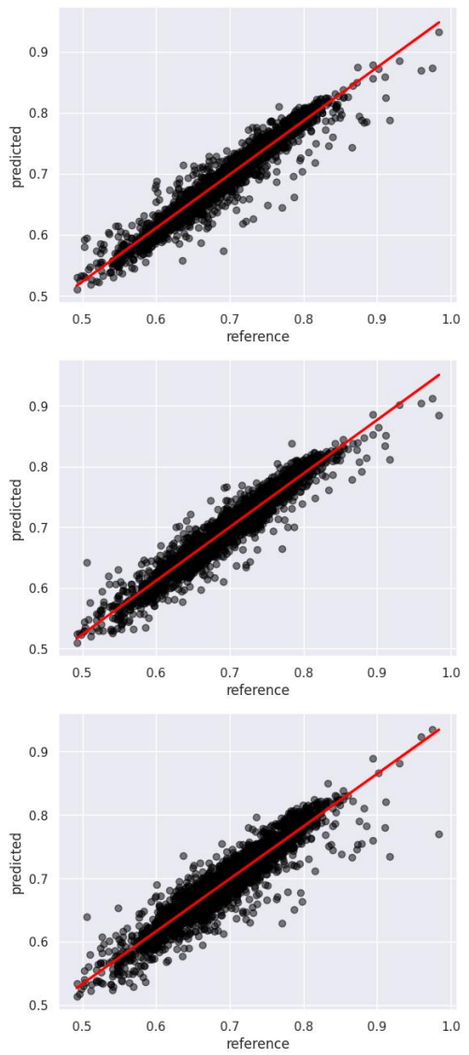


Figure 15

15 2021 season forecast performance graphs

Source: Own elaboration

The previous figure shows in red the estimation function that the model has built to predict the sugar cane production index. Corresponding to the 2021 season, set 1, 4 and 6 appear in first place.

Conclusions

This study demonstrates the potential of time series formed by Sentinel-2 multispectral satellite images with a spatial resolution of 10m in the development of predictive models for sugar cane production.

The Random Forest Regressor is a Machine Learning algorithm that has demonstrated efficiency in regression tasks, reaching in this particular case, metrics $R^2 > 0.84$ and RMSE < 0.034, when working with spectral samples of the phenological cycle of the crop included in the senescence stage. It also highlights the importance of identifying the appropriate time frame to carry out the prediction exercise, which for the crop under study is the months of November and December. The selection of features proves the importance of spectral information such as vegetation indices to characterise crops, however, the intervention of other factors such as Sentinel-2 bands or climatological data can enhance the accuracy of the prediction.

Future work includes the possibility of exploring techniques other than cloud masking by SCL, as this product is inaccurate in class classification. Therefore, the processing of the time series and the identification of phenological variables could be improved. With the prediction of the sugar cane production index, it is possible to plan the harvest and optimise resources.

Conflict of interest

The authors declare no interest conflict. They have no known competing financial interests or personal relationships that could have appeared to influence the article reported in this article.

Authors' Contribution

The contribution of each researcher in each of the points developed in this research, was defined based on:

Lárraga-Altamirano, Hugo René: Contributed to the project idea, research method and technique. He supported the design of the field instrument. He carried out the data analysis and systematisation of results, as well as writing the article.

Hernández-López, Dalia Rosario: Carried out the systematisation of the background for the state of the art. She supported the design of the field instrument. She also contributed to the writing of the article.

Piedad-Rubio, Ana María: contributed to the research design, the type of research, the approach, the method and the writing of the article.

ISSN: 2410-3993
RENIECYT-CONAHCYT: 1702902

Lárraga-Altamirano, Hugo René, Hernández-López, Dalia Rosario, Piedad-Rubio, Ana María and Blanco-Martínez, José Ramón.Machine-Learning model for estimating sugarcane production at crop level. Journal of Technology and Innovation. 2024. 11-28:1-12. https://doi.org/10.35429/JTI.2024.28.11.1.13

Blanco-Martínez, José Ramón: worked on the application of the field instrument, data collection and systematisation of the results. He also worked on the writing of the paper.

Availability of data and materials

The satellite images for the integration of the time series were obtained from the free Copernicus platform managed by the European Space Agency. Climatological data measured by the EMAS were requested from the National Water Commission of the State of SLP.

Funding

The research did not receive any funding.

Abbreviations

ANN	Artificial Neural Network
AP	Precision agriculture
CONAGUA	National Water Commission
EMAS	Automatic Weather Stations
EOS	End of season
ESA	European Space Agency
EVI	Enhanced Vegetation Index
GNDVI	Green Normalised Difference Vegetation Green
GRVI	Green-Red Vegetation Index
IV	Vegetation Indices
LAI	Leaf Area Index
MAE	Absolute Error
ML	Machine Learning
MSE	Mean square error
NDVI	Normalised Difference Green Vegetation Index
NDWI	Normalized Difference Water Index
POS	Highest peak of the season
PR	Remote Sensing
R2	Completion Coefficient
RF	Random Trees
RFR	Random Forest Regressor
RMSE	Root Mean Square Error
SCL	Scene classification map
SCV	Comma Separated Values
SOS	Season Start

References

Basics

FAO. (2019). FAOSTAT.

Felipe Maldaner, L., De Paula Corrêdo, L., Fernanda Canata, T., & Paulo Molin, J. (2021). Predicting the sugarcane yield in real-time by harvester engine parameters and machine learning approaches. Computers and Electronics in Agriculture, 181, 105945.

K.C., S., Ninsawat, S., & Som-ard, J. (2021). Integration of RGB-based vegetation index, crop surface model and object-based image analysis approach for sugarcane yield estimation using unmanned aerial vehicle. Computers and Electronics in Agriculture, 180, 105903.

O'Grady, M. J., & O'Hare, G. M. P. (2017). Modelling the smart farm. Information Processing in Agriculture, 4(3), 179-187.

Said Mohamed, E., Belal, Aa., Kotb Abd-Elmabod, S., El-Shirbeny, M. A., Gad, A., & Zahran, M. B. (2021). Smart farming for improving agricultural management. The Egyptian Journal of Remote Sensing and Space Science, 24(3), 971-981.

SIAP. (2019). Sistema de Información Agroalimentaria y Pesquera.

Supports

Arab, S. T., Noguchi, R., Matsushita, S., & Ahamed, T. (2021). Prediction of grape yields from time-series vegetation indices using satellite remote sensing and a machine-learning approach. Remote Sensing Applications: Society and Environment, 22, 100485.

Jeena Jacob, I., Kolandapalayam Shanmugam, S., Piramuthu, S., & Falkowski-Gilski, P. (Eds.). (2021). Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2020. Springer Singapore.

Tovar Blanco, A. L., Lizarazo Salcedo, I. A., & Rodríguez Eraso, N. (2020). Estimación de biomasa aérea de Eucalyptus grandis y Pinus spp. Usando imágenes Sentinel1A y Sentinel2A en Colombia. Colombia forestal, 23(1).

Xu, J.-X., Ma, J., Tang, Y.-N., Wu, W.-X., Shao, J.-H., Wu, W.-B., Wei, S.-Y., Liu, Y.-F., Wang, Y.-C., & Guo, H.-Q. (2020). Estimation of Sugarcane Yield Using a Machine Learning Approach Based on UAV-LiDAR Data. Remote Sensing, 12(17), 2823.

Differences

Canata, T. F., Wei, M. C. F., Maldaner, L. F., & Molin, J. P. (2021). Sugarcane Yield Mapping Using High-Resolution Imagery Data and Machine Learning Technique. Remote Sensing, 13(2), 232.

ISSN: 2410-3993
RENIECYT-CONAHCYT: 1702902

Dimov, D., Uhl, J. H., Löw, F., & Seboka, G. N. (2022). Sugarcane yield estimation through remote sensing time series and phenology metrics. Smart Agricultural Technology, 2, 100046.

Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for Sustainable Development, 36(2), 27.

Hammer, R. G., Sentelhas, P. C., & Mariano, J. C. Q. (2020). Sugarcane Yield Prediction Through Data Mining and Crop Simulation Models. Sugar Tech, 22(2), 216-225.

Ji, Z., Pan, Y., Zhu, X., Wang, J., & Li, Q. (2021). Prediction of Crop Yield Using Phenological Information Extracted from Remote Sensing Vegetation Index. Sensors, 21(4), 1406.

Shukla, K., Kumar, P., Mann, G. S., & Khare, M. (2020). Mapping spatial distribution of particulate matter using Kriging and Inverse Distance Weighting at supersites of megacity Delhi. Sustainable Cities and Society, 54, 101997.

Singla, S. K., Garg, R. D., & Dubey, O. P. (2020). Ensemble Machine Learning Methods to Estimate the Sugarcane Yield Based on Remote Sensing Information. Revue d'Intelligence Artificielle, 34(6), 731-743.

Wang, Z., Lu, Y., Zhao, G., Sun, C., Zhang, F., & He, S. (2022). Sugarcane Biomass Prediction with Multi-Mode Remote Sensing Data Using Deep Archetypal Analysis and Integrated Learning. Remote Sensing, 14(19), 4944.

Yu, D., Zha, Y., Shi, L., Jin, X., Hu, S., Yang, Q., Huang, K., & Zeng, W. (2020). Improvement of sugarcane yield estimation by assimilating UAV-derived plant height observations. European Journal of Agronomy, 121, 126159.

ISSN: 2410-3993
RENIECYT-CONAHCYT: 1702902