Spectral Index-Based Estimation of Total Nitrogen in Forage Maize: A Comparative Analysis of Machine Learning Algorithms

Aldo Rafael Martínez-Sifuentes, Ramón Trucios-Caciano, Nuria Aide López-Hernández, Enrique Miguel-Valle and Juan Estrada-Ávalos *

National Institute of Forestry, Agricultural and Livestock Research, National Center for Disciplinary Research on Water, Soil, Plant and Atmosphere Relationships, Gómez Palacio 35150, Mexico; martinez.aldo@inifap.gob.mx (A.R.M.-S.); trucios.ramon@inifap.gob.mx (R.T.-C.); lopez.nuria@inifap.gob.mx (N.A.L.-H.); enrique.miguel89@hotmail.com (E.M.-V.)

* Correspondence: estrada.juan@inifap.gob.mx

Abstract: Nitrogen plays a fundamental role as a nutrient for the growth of leaves and the process of photosynthesis, as it directly influences the quality and yield of corn. The importance of knowing the foliar nitrogen content through Machine Learning algorithms will help determine the efficient use of nitrogen fertilization in a context of sustainable agronomic management by avoiding Nitrogen loss and preventing it from becoming a pollutant for the soil and the atmosphere. The combination of machine learning algorithms with vegetation spectral indices is a new practice that helps estimate parameters of agricultural importance such as nitrogen. The objective of the present study was to compare random forest and neural network algorithms for estimating total plant nitrogen with spectral indices. Five spectral indices were obtained from remotely piloted aircraft systems and analyzed by mean, maximum and minimum from each sample plot to finally obtain 15 indices, and total nitrogen was estimated from the georeferenced points. The most important variables were selected with backward, forward and stepwise methods and total nitrogen estimates by laboratory were compared with random forest models and artificial neural networks. The most important indices were NDREmax and TCARImax. Using 15 spectral indices, total nitrogen with a variance of 79% and 81% with random forest and artificial neural network, respectively, was estimated. And only using NDREmax and TCARImax indices, 73% and 79% were explained by random forest and artificial neural network, respectively. It is concluded that it is possible to estimate nitrogen in forage maize with two indices and it is recommended to analyze by phenological stage and with a greater number of field data.

Keywords: artificial neural network; random forest; sustainability; environment

1. Introduction

Remote sensing in agriculture has become an important tool to assist agronomic activities at a low cost through nondestructive testing. Remote sensing enables the faster and more frequent acquisition of crop data in comparison to conventional methods [1]. In crop nutrition, various remote sensing techniques have been used and evaluated. However, it has been found that spectral analysis for crops is a viable alternative for estimating plant health conditions [2,3].

A crucial aspect of agricultural crop management involves monitoring the nitrogen (N) levels in plants. Nitrogen is a key nutrient essential for leaf growth and photosynthesis, exerting a direct impact on both quality and yield [4]. A deficiency of nitrogen in plants leads to an ineffective photosynthesis process, as it is the essential element that forms the chlorophyll molecule [5].

Over-application of fertilizer on crops is still a common activity today. This practice has a direct negative impact on crops because it causes intoxication and affects the environment.
through leaching and volatilization of non-absorbed plants [6]. It is for this reason that monitoring plant nutrients becomes an important activity in agronomic management. Conventional approaches for assessing nitrogen levels rely on chemical analysis of leaf tissue, a process typically characterized by its complexity, time consumption and high cost, often generating environmentally hazardous waste [7].

The conventional laboratory procedure to obtain total nitrogen in the plant is Kjendahl; however, this process is time-consuming as it requires drying and grinding of the plant sample (3 to 4 days) for US$7 per sample without considering labor, with an accuracy of 0.003% detection, 0.01% quantification with recoveries above 90%. Therefore, if more extensive monitoring is required to know the nitrogen content in a plot, this becomes costly and not viable for many producers with small areas.

Optical sensors mounted on both unmanned and manned aerial vehicles have been specifically developed to capture extensive high-resolution data concerning the dynamic characteristics of crops. Furthermore, satellite imagery serves as an excellent choice for monitoring nitrogen levels across vast expanses of vegetated areas [8]. Even though thematic maps derived from satellite images offer adequate accuracy despite their low spatial resolution and high temporal resolution, manned aerial vehicles can capture images with superior spatial and temporal resolution when compared to satellites [9].

Currently, remote sensing studies employ vegetation indices associated with chlorophyll or plant N content to indirectly estimate their content [2,10]. Among these indices is the GNDVI (Green Normalized Difference Vegetation Index), CI_green (Chlorophyll Index Green), NDVI (Normalized Difference Vegetation Index), RVI (Ratio Vegetation Index), NDRE (Normalized Difference Red Edge), CCCI (Canopy Chlorophyll Content Index), CI_rededge (Red edge Chlorophyll Index), MCARI/OSAVI (Modified Chlorophyll Absorption in Reflectance Index/Optimized Soil-Adjusted Vegetation Index), MCARI/OSAVI RE (Red Edge-based Modified Chlorophyll Absorption in Reflectance Index/Optimized Soil Adjusted Vegetation Index), TCARI/OSAVI (Transformed Chlorophyll Absorption in Reflectance Index/Optimized Soil Adjusted Vegetation Index), TCARI/OSAVI RE (Red Edge-based Transformed Chlorophyll Absorption in Reflectance Index/Optimized Soil Adjusted Vegetation Index), and MTCI (MERIS Terrestrial Chlorophyll Index), among others.

These studies use a variety of methods to relate the field-measured value and the spectral index. Recently, autonomous learning algorithms have been used with various approaches and applications [11–13]. Algorithms such as artificial neural networks (ANN), support vector machine (SVM), decision trees (DT) and random forest (RF) are powerful tools to assist in analyses with remotely piloted aircraft imagery [14]. Combining machine learning algorithms with vegetation spectral indices is a relatively new practice, as these algorithms ensure good model performance even with diverse variables as input components [15].

The objectives of the present study were (i) to select the best spectral indices through forward, backward and stepwise methods, and (ii) to compare machine learning algorithms such as random forest and artificial neural networks to estimate total nitrogen in forage maize through spectral indices obtained from remotely piloted aircrafts in northern Mexico.

2. Materials and Methods

2.1. Study Area

The study area is located in northern Mexico, on the property called “Granja Palestina” in the municipality of Francisco I. Madero, Coahuila, Mexico, with extreme coordinates of 103°18’34.33” to 103°19’15.24.24” West longitude and 25°46’39.61” to 25°46’57.49.49” North latitude (Figure 1). The cultivable study area included 26.14 ha. The predominant climate is dry semi-warm (BWH), with average annual temperature and precipitation of 20.9 °C and 260.7 mm, respectively [16]. The hottest months are from May to August and the coldest months are December and January, with a convective rainfall regime in summer as a result of the North American Monsoon [17].
Nitrogen 2024, 5

Figure 1. Geographic location of Granja Palestina in northern Mexico.

2.2. Description of Sampling Sites

The study was performed in the 2020 summer agricultural cycle. The soil texture is clay loam, has a good water retention capacity, is slightly alkaline with a moderate amount of salts, and contains considerable levels of organic matter and essential nutrients such as nitrogen (in the form of nitrate and ammonium), phosphorus and potassium. These data suggest that the soil can be quite fertile, although the high alkalinity and conductivity may affect certain crops sensitive to these conditions. The corn hybrid Syngenta n8305 was used. The planting density was 93,100 plants per ha. The agronomic management in the plots of interest included three pre-planting activities. First, a cross subsoiling was carried out at a depth of 40 to 50 centimeters. Once this activity was completed in each plot, fallowing was carried out to a depth of 20 centimeters. Subsequently, the soil was harrowed to remove the large clods. Once this activity was completed, the borders of the plots were bordered with 25-m wide borders, according to an agronomic design that corresponds to the flow rate of the irrigation system, which contemplates a unit flow rate of 4 to 7 liters per second per square meter, thus avoiding soil degradation. The irrigation system used is called “alfalfa valves” and consists of transferring water through pipes that deliver the water to the plots through hydrants. This system can irrigate a maximum of one hectare per hydrant. Four irrigations were carried out in the plots of interest, starting with watering irrigation. Finally, the crop was harvested when it reached 35% maturity, with a total duration of 95 to 105 days after sowing. The chemical fertilization rate was 200 kg ha$^{-1}$ of monoammonium phosphate and 100 kg ha$^{-1}$ of Solub 45 (nitrogen fertilizer with nitrification inhibitor; 45% nitrogen).

2.3. Field Sampling

Plant samples were collected starting at the V6 phenological stage of corn, for each collection the site was georeferenced with a Garmin Etrex 20 GPS. For each sampling point, a 1 m$^2$ plot was made and a representative plant was selected.
Before flowering, samples were collected from a single plant, comprising both stem and leaves, within the designated plot. Following flowering, samples were obtained from the leaf positioned opposite the ear of corn. Thirty-two sampling points were performed, under the following scheme: 34 days after sowing (das) (two samplings), 35 das (two samplings), 36 das (two samplings), 44 das (two samplings), 46 das (two samplings), 55 das (two samplings), 56 das (two samplings), 57 das (two samplings), 65 das (two samplings), 66 das (two samplings), 67 das (two samplings), 75 das (one sampling), 85 das (two samplings), 86 das (one sampling), 87 das (two samplings), 95 das (two samplings), and 97 das (two samplings) of the 2020 summer cycle.

The plant samples were transported to the laboratory for additional examination.

2.4. Laboratory Analysis

The estimation of total nitrogen (TN) was carried out at the Water and Soil Laboratory of the National Center for Disciplinary Research on Water, Soil, Plant and Atmosphere Relations of the National Institute of Forestry, Agriculture and Livestock Research. To obtain the percentage of TN, plant samples were analyzed by thermal conductivity using the Dumas method [18]. In this technique, nitrogen undergoes a conversion into gaseous form through the calcination process. The resultant gases are then reduced by copper and dehydrated, while carbon dioxide emissions are captured, facilitating the subsequent quantification of TN.

2.5. Aerial Images of Remotely Piloted Aircraft System

The Remotely Piloted Aircraft System (RPA) flight missions were conducted when the corn plants had four fully expanded leaves. Seventeen flights were conducted according to the post-planting sampling program with an eBee plus fixed-wing drone equipped with Parrot sequoia multispectral sensor, which can record images in four bands, 530–570 nm (green), 640–680 nm (red), 730–740 nm (red edge) and 770–810 nm (near infrared). The flight missions were performed with 60% lateral overlap and 80% longitudinal overlap at an altitude of 191.0 m, which helped obtain a pixel size of 18 cm/px with the eMotion 3 software. The processing to obtain the orthomosaic and the bands with reflectance values was carried out with the Pix4DMapper software version 4.5. To have greater coverage in the electromagnetic spectrum, the maximum, minimum and mean values of spectral indices generated with QGIS software version 3.28 were obtained for each 1 km² sampling plot, spectral indices that are described in Table 1.

<table>
<thead>
<tr>
<th>Index</th>
<th>Equation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDVI: determines the greenness of vegetation.</td>
<td>( \frac{NIR - R}{NIR + R} )</td>
<td>[19]</td>
</tr>
<tr>
<td>CCCI: estimates chlorophyll in leaves.</td>
<td>( \left( \frac{NIR}{NDVI} \right) )</td>
<td>[20]</td>
</tr>
<tr>
<td>CIGreen: estimates chlorophyll in leaves.</td>
<td>( \frac{NIR}{G} - 1 )</td>
<td>[21]</td>
</tr>
<tr>
<td>NDRE estimates chlorophyll in leaves.</td>
<td>( \frac{NIR - \text{Red Edge}}{NIR + \text{Red Edge}} )</td>
<td>[20]</td>
</tr>
<tr>
<td>TCARI: indicates the relative abundance of chlorophyll. It is affected by the reflectance of the underlying soil, especially in vegetation with low leaf area index.</td>
<td>( 3\left{ (NIR - \text{RE}) - 0.2(NIR - G)\left( \frac{NIR}{RE} \right) \right} )</td>
<td>[22]</td>
</tr>
</tbody>
</table>

2.6. Variable Selection

For the selection of the best variables in the models, the AIC criterion of information loss was used, which indicates how much of the information is lost or not, being explained by the model according to the variables that construct it. Additionally, the coefficient of determination (R²), root mean square error (RMSE) and mean squared error (MSE) were considered for the selection of the vegetation indices. Therefore, forward, backward and
stepwise models were used, which refer to the process of testing and discarding variables from the whole set to obtain the best possible model [23].

The forward method consists of generating a model whose response is the dependent variable and which is explained only by its average \( y_i = \mu + e_i \). From this “empty” model, variables from the set are added and it is taken into account whether the variables added to the model help to better explain the response variable by verifying the AIC criterion or, on the contrary, whether they worsen the model.

The backward method consists of performing the process of the forward model but in reverse, i.e., starting from the saturated model that considers all the variables in the data set to explain the dependent variable and then starting to eliminate variables and corroborate using the AIC criterion to determine whether removing a variable improves or worsens it to decide on whether to include or eliminate the variable.

The stepwise model is a combination of the forward and backward methods, since for the selection of the most significant variables, it makes its way forward and backward. Therefore, it is necessary to use the empty model and the saturated model to indicate where to start and where to stop in your search.

2.7. Machine Learning Models

The total nitrogen content in forage corn was estimated using artificial neural network (ANN) and random forest (RF) algorithms. ANN is a computational system inspired by the interconnected structure of neurons in the human brain. An artificial neuron, a fundamental component of a neural network, is a mathematical entity that processes information [24].

An ANN comprises interconnected artificial neurons organized into layers. Each layer contains neurons with an activation function. Typically, a neural network includes an input layer, an output layer, and one or more hidden layers. The input layer receives external inputs and transmits them, weighted, to the hidden layers [25].

RF is a type of ensemble learning algorithm that creates numerous decision trees randomly, which are then utilized with training samples [26]. RF algorithm exhibits both high accuracy and stability, enabling it to effectively handle input samples with large datasets and numerous dimensional features. Both algorithms were implemented using R software version 4.2.1 [27].

Nonparametric regression methods are optimized through a learning process that utilizes training data to establish a model. The model parameters are tuned to minimize estimation errors. To assess the performance of both predictive models, a k-fold cross-validation method with five subsets was employed. This technique, widely utilized for validating machine learning models such as RF and ANN, involves dividing the entire dataset into two subsets: one for training or model adjustment and the other for validation. Consequently, a model is exclusively created using the training data, which is then compared with the reserved validation data not utilized during model development. In this study, 70% of the data was utilized for fitting, while the remaining 30% was allocated for validation purposes [28].

For the neural network algorithm, the “resilient backpropagation” algorithm was used, as it allows error detection in the learning process. After discovering the first fault in the original location, the algorithm can move backward and point out other anomalies affecting both nodes and previous layers. Similarly, it is capable of providing solutions in subsequent elements. Two hidden layers were used with a “threshold” value of 0.05 for the partial derivatives of the error function as a stopping criterion. The “max_step” parameter was set to \( 1 \times 10^8 \), and the “rep” was set to one for the network training. The logistic function was used to smooth the result of the cross-product of the covariate or neurons and the weights.

For the random forest algorithm, regression was used with the parameters “keep_forest” to maintain the output object, number of trees “ntree” set to 550, and “mtry” set to two, which is the number of predictors to be randomly sampled at each split when building the tree models. The “importance” parameter was considered in order to generate a matrix
where the first column represents the average decrease in accuracy, and the second column represents the average decrease in mean squared error.

In order to enhance the models’ performance, it is essential to adjust the hyperparameters. Various combinations of these hyperparameters are tested during the training phase to achieve optimal performance. However, this process may lead to overfitting, where the model performs well on the training set but poorly on the test set. One of the most commonly used techniques for testing the effectiveness of a Machine Learning model is cross-validation. This method is also a re-sampling procedure that allows a model to be evaluated even with limited data. To perform a cross-validation, a part of the data must be removed from the training data set beforehand. This data will not be used to train the model, but later to test and validate it. Cross-validation is often used in Machine Learning to compare different models and select the most suitable one for a specific problem. K-fold cross-validation is a method that ensures the representativeness of all observations in the training and test sets. It is particularly effective when there are limitations in the input data. The procedure involves randomly dividing the data into K groups, where K is a parameter that determines the number of splits. The choice of K, usually between 5 and 10, depends on the scale of the data. A higher value of K reduces the model bias but may increase the variance and lead to over-fitting, while a lower value resembles the Train-Test Split method. The model is then fitted using K-1 groups and validated with the remaining group. This process is repeated until each group has served once as a test set. The model performance metric is calculated as the mean of the recorded scores. Hence, it is crucial to conduct k-fold cross-validation [29].

For the present analysis, the dataset was initially split into training and testing sets. For the k-fold method, solely the training set was utilized, which was then partitioned into five subsets.

The subsets were cycled through utilizing one-fifth of the samples for model validation and the remaining four-fifths for training. In the initial iteration, the first subset was designated for validation, while the subsequent four iterations involved using the remaining subsets for training. This process continued iteratively, with each subset taking turns for validation until all five iterations were completed.

This approach involves conducting five separate training runs to validate the model, with the final accuracy being the average of the accuracies obtained from these five runs.

2.8. Evaluation of Model Performance

The performance in the estimation of TN through the models was carried out using the coefficient of determination ($R^2$, Equation (1)), root mean square error (RMSE, Equation (2)) and mean squared error (MSE, Equation (3)). The formulas are as follows:

$$R^2 = \left( \frac{\sum_{i=1}^{n} (O_i - \bar{O})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n} (O_i - \bar{O})^2}\sqrt{\sum_{i=1}^{n} (S_i - \bar{S})^2}} \right)^2 \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (O_i - S_i)^2} \quad (2)$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (O_i - S_i)^2 \quad (3)$$

where, $O_i$, $S_i$, $\bar{O}$, $\bar{S}$, and $n$ represent the observed data, estimated data, average value of the observed data, average value of the estimated data, and the total number of samples, respectively.
3. Results
3.1. Laboratory-Estimated Nitrogen and RPA Spectral Indexes

The range of laboratory-estimated nitrogen was from 1.20 to 5.66% with an average of 3.22% and a median of 3.15%. Of the values obtained from remotely piloted aircraft, the CLGmean, CLGmin and CLGmax indices were those that presented the greatest variability, showing a larger interquartile range compared to the rest of the spectral indices analyzed (Figure 2).

\[
R^2 = \frac{\sum(O_i - \bar{O}_t)(S_i - \bar{S})^2}{\sum(O_i - \bar{O}_t)^2}
\]

\[
RMSE = \sqrt{\frac{\sum(O_i - S_i)^2}{n}}
\]

\[
MSE = \frac{\sum(O_i - S_i)^2}{n}
\]

where, \(O_i\), \(S_i\), \(\bar{O}_t\), \(\bar{S}\), and \(n\) represent the observed data, estimated data, average value of the observed data, average value of the estimated data, and the total number of samples, respectively.

Figure 2. Box plot with mean, maximum and minimum values of vegetation indices and percentage of nitrogen.

3.2. Variable Selection

The forward method for variable selection explains 67.26% of the variance of the data, with MSE of 0.52 and RMSE of 0.72, where all the variables are statistically significant in their contribution to the model. The backward model, when considering all the variables, presents a variance of 68.56% with MSE of 0.38 and RMSE of 0.61, and apparently, none of the coefficients of the variables is different from 0 (based on the Student’s t statistic), which may be due to the presence of the multicollinearity problem in the data. In the stepwise model, the results are the same as the backward model, where the variables that are significant for the model are NDREmax and TCARImax with an explained variance of 67.26% with MSE of 0.52 and RMSE of 0.72.

Due to the similarity of the stepwise and backward model found and the fact that it consists of only three variables (two independent and one response variable), besides being parsimonious and explanatory, it satisfies the assumptions; therefore, it is a very good candidate for modeling the total nitrogen data with random forest and artificial neural network algorithms. The scatter plot between the NDREmax and TCARImax indices with the TN is shown in Figure 3a. The Shapiro–Wilk normality test showed \(p = 0.52\); therefore, it shows a normal or parametric behavior, and a lack of homoscedasticity in the data is not observed (Figure 3b).

Individually, both indices present an \(R^2\) value of 0.05 and 0.64 for the variable NDREmax and TCARImax, respectively, with the functions \(TN = -13.113\) (NDREmax) + 7.4046 and \(TN = 2.5145\) (TCARImax) + 2.9181 (Figure 4). However, together, the plant
nitrogen content can be reliably estimated using random forest and artificial neural network algorithms.

![Image](attachment:image_url)

**Figure 3.** Scatter plot between TCARImax, NDREmax indices and TN (a). Homoscedasticity test of the data (b).

![Image](attachment:image_url)

**Figure 4.** Scatter plot between the NDREmax index and TN (a). Scatter plot between the TCARImax index and TN (b).

### 3.3. Estimation of TN Using Machine Learning Algorithms

Applying the random forest model with all the variables shows an explained variance of 79.04%, MSE of 0.35 and RMSE of 0.59, and the variables NDREmean and NDREmax were the most important in the model (Figure 5). On the other hand, when considering only the TCARImax and NDREmax variables, resulting from the variable selection models, the variance explained by RF was 73.04% with MSE of 0.51 and RMSE of 0.72, where NDREmax was the most important variable (Figure 6). A total of 550 trees were required in the RF model to obtain the lowest mean absolute error (Figure 7).

Applying neural networks with all the variables, we obtained an explained variance of 81.33%, an MSE of 0.32 and RMSE of 0.56 (Figure 8a). On the other hand, when applying the ANN contemplating the NDREmax and TCARImax indices, an explained variance of 79.35%, MSE of 0.20 and RMSE of 0.44 were found (Figure 8b). The above shows similarity to the RF model, where considering all variables explains more of the model, compared to using only two variables. However, in both cases, the increase of 6.02 and 1.98 for RF and ANN, respectively, is not significant. The conceptualization of the ANN model, for both cases, of all variables and only NDREmax and TCARImax is shown in Figure 9a,b.
3.3. Estimation of TN Using Machine Learning Algorithms

Applying the random forest model with all the variables shows an explained variance of 79.04%, MSE of 0.35 and RMSE of 0.59, and the variables NDREmean and NDREmax were the most important in the model (Figure 5). On the other hand, when considering only the TCARImax and NDREmax variables, resulting from the variable selection models, the variance explained by RF was 73.04% with MSE of 0.51 and RMSE of 0.72, where NDREmax was the most important variable (Figure 6). A total of 550 trees were required in the RF model to obtain the lowest mean absolute error (Figure 7).

Figure 5. Importance of random forest model variables when considering all variables (a). Scatter plot between measured and estimated nitrogen from the random forest model with all variables (b).

Figure 6. Importance of variables of the random forest model using only the variables NDREmax and TCARImax (a). Scatter plot between measured and estimated nitrogen from the model with two predictor variables (b).

Figure 7. Accuracy of the random forest model.
Applying neural networks with all the variables, we obtained an explained variance of 81.33%, an MSE of 0.32 and RMSE of 0.56 (Figure 8a). On the other hand, when applying the ANN contemplating the NDREmax and TCARImax indices, an explained variance of 79.35%, MSE of 0.20 and RMSE of 0.44 were found (Figure 8b). The above shows similarity to the RF model, where considering all variables explains more of the model, compared to using only two variables. However, in both cases, the increase of 6.02 and 1.98 for RF and ANN, respectively, is not significant. The conceptualization of the ANN model, for both cases, of all variables and only NDREmax and TCARImax is shown in Figure 9a,b.

**Figure 8.** Scatter plot when applying artificial neural networks with all variables (a). Using only the NDREmax and TCARImax indices in the neural network algorithm (b).

**Figure 9.** Conceptualization of the artificial neural network algorithm with all variables (a). Conceptualization of the artificial neural network algorithm with NDREmax and TCARImax indices (b).
4. Discussion

To estimate in-plant nitrogen, geospatial technologies have become more robust and accurate [30,31]. Accuracy in nitrogen estimation involves a combination of methods and processes, as well as the analysis of a large amount of data. In this sense, the combination of geospatial technologies and artificial intelligence algorithms has allowed revolutionizing the agricultural sector by improving decision-making in hours or days, compared to weeks or months of analysis, reducing costs and increasing yields [32].

Correlations have been found between vegetation indices and agronomic parameters, which makes these analyses an alternative to the estimation of agricultural traits [33]. In this study, high-resolution indices proved valuable for predicting total nitrogen (TN) levels in forage maize, a crop highly responsive to this element’s content. Additionally, the crop is susceptible to various biotic and abiotic stressors, particularly during the flowering stage. Various studies have employed machine learning algorithms to estimate nitrogen levels in crops. Wang et al. [34], applied several machine learning algorithms such as Support Vector Machine and Random Forest with hyperspectral data to estimate nitrogen levels in corn leaves. The results showed that machine learning models significantly outperformed traditional nitrogen estimation methods. Multispectral images have been used to predict foliar nitrogen levels in wheat, and machine learning algorithms such as Convolutional Neural Networks and Gradient Boosting Machines were implemented to develop highly accurate predictive models [35]. Another study by Liu et al. [36], utilized a combination of remote sensing data, such as satellite images and weather data, along with machine learning algorithms to estimate nitrogen levels in rice crops, applying techniques such as Random Forest and Support Vector Regression to develop predictive models. The results highlighted the utility of this approach for monitoring and managing nitrogen fertilization in rice crops.

The spectral indices were analyzed separately under two criteria: first, the RF and ANN models were generated with all the variables and then only with the NDREmax and TCARImax indices, which showed greater importance according to the stepwise and backward tests. Although a higher explained variance was obtained by using all the indices in the models, a greater analysis and geoprocessing effort is required; therefore, it is feasible to use the NDREmax and TCARImax indices for TN estimation. In this sense, the NDRE index has been used to estimate agronomic parameters such as the yield of some crops such as rice [12] or nitrogen content in sorghum reaching a variance of 41% [37].

The described approach is viable because red-edge bands derived from the normalized difference red-edge (NDRE) can reliably detect crop traits, exhibiting a stronger correlation with indicators like nitrogen accumulation in plants. This capability aids in addressing the saturation issue [38]. It has also been employed for the estimation of characteristics such as canopy coverage, leaf area index and leaf chlorophyll content [39]. The TCARI index has been used to monitor the yield and physiological response of sweet corn [40] because it is sensitive to leaf chlorophyll content [41]. It has also been found that it can be used to estimate plant nutrition with model variance values up to 0.83% [42].

Because TCARI’s importance value significantly contrasts with that of the other assessed indices, a distinct pattern emerges in the prediction of TN in forage maize compared to indices related to crop structure like NDVI. These findings suggest that structural indices have limited predictive capability for estimating TN, possibly due to their tendency to exhibit saturation values under moderate to high soil coverage conditions [43].

The outcomes of the current investigation validate the conclusions of Hunt et al. [44] in wheat cultivation; this suggests a stronger correlation between chlorophyll content and indices utilizing reflectance in the near red bands compared to those that do not incorporate such data. Previous studies on phenotypic characterization and monitoring of maize crops have also highlighted the close association between the near red band and chlorophyll content [45,46].

Currently, traditional multivariate methods are still used to estimate plant nitrogen content [47]. However, artificial intelligence algorithms have become a viable option for...
data implementation and processing [48]. Several studies have analyzed the performance of artificial intelligence algorithms for different agronomic characteristics such as total nitrogen in soil [49] or yield prediction [50], while the Random Forest algorithm has shown good predictive ability for wheat yield ($R^2 = 0.89$) [44]. However, it has been found that the neural network model has been able to estimate corn plant nitrogen with $R^2$ values of 0.86–0.97 [51].

The application of Random Forest algorithms in our study, complemented by insights from previous research [52–54], proved instrumental in accurately estimating total nitrogen levels in forage maize. Our findings align with similar investigations in related agricultural contexts, where Random Forest models demonstrated efficacy in predicting key agricultural parameters [52,55]. Additionally, the incorporation of Partial Least Squares-Discriminant Analysis (PLS-DA) and the Debiased Sparse Partial Correlation (DSPC) algorithm, as described by Rodriguez et al. [55] and Rey et al. [56] respectively, offers promising avenues for further refining our predictive models. Our study underscores the interdisciplinary nature of agricultural research and highlights the potential for leveraging advanced statistical techniques to enhance sustainability and productivity in crop management.

5. Conclusions

Artificial intelligence algorithms have become important tools for agronomic decision making. In the present study, significant relationships were found for estimating nitrogen in corn with random forest and artificial neural network algorithms. The analysis showed that the difference in variance when using all the indices and only the most important ones is reduced; therefore, it is feasible to use only two indices to estimate nitrogen in corn, allowing for savings in data processing and analysis. For future studies in northern Mexico, it is important to analyze the nitrogen content by phenological stage, to more accurately find the nitrogen for each stage of corn, in addition to having a greater number of field data for a better and more detailed estimation of nitrogen content. Although the processing cost to obtain the spectral indices NDREmax and TCARImax is the same because with a single image all indices can be obtained with a multispectral sensor, however, the analysis time is optimized by using only the two indices. On the other hand, estimating nitrogen content through remote sensing, allows savings in plant sample processing costs, since by the traditional laboratory method, these have a cost of up to 10 dollars per sample and the cost could increase if more plant samples were taken, exceeding 120 dollars per sampling plot throughout the phenological cycle. Field calibration of this type of studies plays a key role in the validation of machine learning models, as it allows to deepen the knowledge for agronomic crop management and, consequently, to focus it towards sustainability and savings in laboratory processing costs for the farmer.

Author Contributions: Conceptualization, A.R.M.-S. and R.T.-C.; methodology, A.R.M.-S. and E.M.-V.; software, A.R.M.-S.; formal analysis, A.R.M.-S. and N.A.L.-H.; resources, J.E.-Á.; writing—original draft preparation, A.R.M.-S., R.T.-C. and N.A.L.-H.; writing—review and editing, E.M.-V., J.E.-Á. and R.T.-C.; project administration, J.E.-Á. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study. Requests to access the datasets should be directed to estrada.juan@inifap.gob.mx.

Conflicts of Interest: The authors declare no conflicts of interest.
References


32. Tsouros, D.C.; Bibi, S.; Sarigiannidis, P.G. A Review on UAV-Based Applications for Precision Agriculture. *Information* 2019, 10, 349. [CrossRef]


42. Sharifi, A. Remotely sensed vegetation indices for crop nutrition mapping. *J. Sci. Food Agric.* 2020, 100, 5191–5196. [CrossRef]


48. Ma, J.; Wang, L.; Chen, P. Comparing Different Methods for Wheat LAI Inversion Based on Hyperspectral Data. *Agriculture* 2022, 12, 1353. [CrossRef]


52. Olivares, B.O.; Vega, A.; Rueda Calderón, M.A.; Montenegro-Gracia, E.; Araya-Almán, M.; Marys, E. Prediction of Banana Production Using Epidemiological Parameters of Black Sigatoka: An Application with Random Forest. *Sustainability* 2022, 14, 14123. [CrossRef]

53. Rodríguez-Yzquierdo, G.; Campos, B.O.; Silva-Escobar, O.; González-Ulloa, A.; Soto-Suárez, M.; Betancourt-Vásquez, M. Mapping of the Susceptibility of Colombian Musaceae Lands to a Deadly Disease: *Fusarium oxysporum* f. sp. cubense Tropical Race 4. *Horticulturae* 2023, 9, 757. [CrossRef]


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.