Agrociencia

CALIBRATION ALGORITHMS ASSESSMENT FOR SOIL NITROGEN PREDICTION WITH NEAR-INFRARED SPECTROSCOPY AND DATA AUGMENTATION

Alejandro Eric **Reyes-Rivera**[1], Gilberto de Jesús **López-Canteñs**[1,2*], Pedro **Cruz-Meza**[2], Noel **Chávez-Aguilera**[2]

[1]Universidad Autónoma Chapingo. Posgrado en Ingeniería Agrícola y Uso Integral del Agua. Carretera Mexico-Texcoco km 38.5, Chapingo, Texcoco, State of Mexico, Mexico. C. P. 56227.
[2]Universidad Autónoma Chapingo. Departamento de Ingeniería Mecánica Agrícola. Carretera Mexico-Texcoco km 38.5, Chapingo, Texcoco, State of Mexico, Mexico. C. P. 56227
* Author for correspondence: alelopez10@hotmail.com

Citation: Reyes-Rivera AE, López-Canteñs G de J, Cruz-Meza P, Chávez-Aguilera N. 2024. Calibration algorithms assessment for soil nitrogen prediction with near-infrared spectroscopy and data augmentation. **Agrociencia.** https://doi.org/ 10.47163/agrociencia.v58i6.3074

Editor in Chief:
Dr. Fernando C. Gómez Merino

Received: September 30, 2023.
Approved: September 12, 2024.
Published in Agrociencia: September 18, 2024.

ABSTRACT

Spectroscopy and machine learning are crucial in smart farming, enhancing soil variability management through predictive spectral models. Choosing suitable regression algorithms is essential due to complex soil-reflection relationships. Additionally, algorithms require a large amount of data to reach good performance, which can be challenging for researchers. Through specific metrics such as R^2, root mean square error, and residual predictive deviation (RPD), this study evaluates four regression algorithms for soil nitrogen prediction: Partial Least Squares (PLS), Extreme Learning Machine (ELM), Support Vector Machine (SVM), and Random Forest (RF). Models were built using near-infrared (NIR) spectroscopy and artificial data augmentation through generative adversarial networks. Spectral preprocessing was performed using a moving average smoothing and Savitzky-Golay derivative filter. The selection of spectral variables was carried out using a genetic algorithm. Artificial data augmentation improved model performance, with SVM and RF outperforming PLS and ELM, achieving RPD > 2, R^2 > 0.8, and lower error rates.

Keywords: Remote sensing, regression models, machine learning, artificial data, soil nutrients.

INTRODUCTION

In precision agriculture, soil nitrogen variability management plays a vital role. Specifically, ammonium (NH_4^+) and nitrate (NO_3^-) are available nitrogen forms that can be absorbed and used by plants for protein, nucleic acid, and pigment synthesis (Patel *et al*., 2020). In addition, accurately determining soil available nitrogen content is critical for precise fertilizer application, resource optimization, and environmental impact reduction. The use of multi- and hyperspectral images for remote sensing is a promising approach in precision agriculture. It can extract valuable soil nutrient data using the appropriate methods. In this sense, spectroscopy is a vital non-destructive analytical tool for studying the interactions between matter and radiation

within spectra. It allows the correlation of spectral responses, such as reflectance or absorbance, with individual elements in a homogeneous mixture due to the unique spectra produced by different chemical species, driven by their bonds and molecular structures (Stenberg *et al*., 2010).

When combined with machine learning models, spectroscopy excels in accurately predicting various soil attributes, particularly when utilizing near-infrared (NIR) and visible (Vis) spectra (Liu *et al*., 2020). Modern spectral instruments offer high-resolution, information-rich spectra but suffer from drawbacks like redundancy, computational costs, noise, and spectral disturbances due to radiation. Therefore, spectral model development requires variable selection (Yun *et al*., 2019) and spectra preprocessing (Burger and Geladi, 2007) techniques to reduce these effects. In addition, spectral models mainly use supervised learning, requiring multivariate calibration methods to extract complex patterns and correlate with soil properties.

Approaches like Principal Component Regression (PCR) and Partial Least Squares (PLS) assume linear relationships and are widely used in soil science (Barra *et al*., 2021). However, the interest in non-linear calibration techniques grows due to the typically non-exclusively linear nature of spectral-soil relationships. Properties such as nitrogen and other elements could be related to spectra in both linear and non-linear ways (Qi *et al*., 2018). Nonlinear machine learning techniques include Artificial Neural Networks (ANN), Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM) regression, Extreme Learning Machine (ELM) networks, and cubist models (Liakos *et al*., 2018). For soil sciences, scientific literature comprehends many studies dedicated to understanding, analyzing, and comparing algorithms with different regression approaches to allow a better relationship between spectra and soil elements. However, proposed calibration techniques lack widespread acceptance, as they may not be universally applicable (Xu *et al*., 2018).

Although machine learning has shown great potential in studies related to remote sensing in soils, the effectiveness of artificial intelligence models such as SVM, RF, ELM, or PLS is closely related to the quantity and quality of the spectral data used in training (Barra *et al*., 2021). As an alternative, generative models based on deep learning, such as generative adversarial networks (GANs), have been proposed to address quantity and diversity scarcity issues. GANs can learn real data distributions and generate synthetic examples similar to original data (Goodfellow *et al*., 2020). They have been implemented to increase training datasets in several studies related to remote sensing and spectroscopy (Lv *et al*., 2021; Wu *et al*., 2021). In addition, they do not require much training data and have proven to be viable for reproducing soil spectra to improve machine learning model performance (Jiang *et al*., 2023). The study aims to evaluate the predictive capabilities of PLS, ELM, SVM, and RF algorithms for soil available nitrogen prediction and analyze the impact of artificial GAN-generated spectral data on these models.

MATERIALS AND METHODS

Study area, sampling, and lab analysis

Sampling was carried out in "La Xerona," an experimental field belonging to the Department of Agricultural Mechanical Engineering at Chapingo Autonomous University. It is located at the geographic coordinates 19° 29′ 2.3′′ N and 98° 53′ 59.8′′ W and has an area of 9.31 ha, made up of four predominant soil textures: Loam-silt, Loam-clayey, Loamy, and Clayey.

Soil (0–30 cm depth) from 129 sampling points was cleaned, dried in an Imperial V model 3481M mechanical convection oven at 90 °C for 36 hours, ground, and sieved through a number 10 sieve (2 mm opening) for particle size homogenization. Each sample was then divided using the quartering method into two portions: one for laboratory nutritional analysis and the other for spectral analysis. Samples were sent to the National Laboratory for Research and Agrifood and Forestry Service (LANISAF) for nitrogen and spectral analysis. The standard Kjeldahl method was used, which is a wet oxidation procedure that indicates available nitrogen for plants in nitrate (NO_3^-) and ammonium (NH_4^+) forms.

Reflectance spectra were obtained using a Bruker® MPA model (122000) NIR spectrophotometer, covering a range of 800–2700 nm with 0.125 nm spectral resolution and calibrated using a Spectralon® panel (99 % reflectance). To ensure representative spectra, three spectral response measurements were averaged.

Spectral models development and assessment

To achieve study goals, NIR and existing nitrogen data were used to train regression models and GANs for dataset augmentation. The process involved preprocessing and splitting spectral data, with 95 of them designated for the initial training set and the remainder for validation. The training set was then used to select variables and train GANs to generate available nitrogen and spectral data. Real and artificial data were used to create training sets with different numbers of samples and to calibrate the algorithms. Finally, the data was used to make predictions with trained models, and their performance was evaluated using the prediction metrics listed in the next section. Models were implemented in Matlab R2020B and GAN networks in Python 3 on a PC with 32GB of RAM and a 2.90 GHz Intel Core i7 processor.

Spectral preprocessing and variable selection

Spectrum preprocessing methods reduce digital noise and interference in spectral data due to dispersion and non-systematic factors, improving predictive algorithm accuracy. In this study, moving average smoothing (MAS) was used for high-frequency noise removal (Guiñón *et al*., 2007). MAS corrects reflectance for each spectral variable by averaging a specified number of neighboring wavelengths within a moving window. Likewise, for dispersion correction and to highlight spectral characteristics, first derivative of reflectance (FDR) was used, applying the Savitzky-Golay derivative

(SGD) filter. Parameters were as follows: window size of 10 for MAS, and a second order polynomial with window size of 7 in smoothing stage of SGD. In addition, this study employed a fixed-variable GA (Fei *et al*., 2009) coupled with PLS regression, as variable selection method to select key wavelengths for AN, setting GA parameters at 1000 generations, 50 chromosomes, mutation, and crossover rates of 0.001 and 0.05, with the root mean square error of cross validation (RMSEcv) as the objective function.

Regression algorithms

Partial Least Squares (PLS). This is one of the most popular multivariable methods for predictive spectral models in soils (Barra *et al*., 2021). To establish the regression model, PLS projects the original data onto a set of new inferred variables known as latent variables through a linear transformation. These are mutually orthogonal and unrelated, effectively eliminating multicollinearity. In this study, the number of latent variables was calculated by cross-validation on the training set of the model.

Extreme Learning Machine (ELM). This algorithm consists of a single-layer feedback neural network with random input weights and bias, while output weights are computed using the least-squares method. Its single-iteration process offers fast learning and mitigates overfitting issues (Vestergaard *et al*., 2021). Hidden layer neuron numbers must be established; however, there is no fixed theory to determine this parameter. In this study, a stepwise test method was adopted. With this method, the number of neurons was set to 15.

Support Vector Machine (SVM). This method, along with the radial basis function (RBF) kernel, is commonly used in soil nutrient prediction (Qi *et al*., 2018; Xu *et al*., 2018). These models project input data into a feature space, handling linear and non-linear relationships efficiently. The SVM calibration process involves choosing an appropriate kernel function and determining optimal model parameters, that is, the penalty (C) and kernel function scale (γ) parameters. In this study, the RBF kernel function was used, and C and γ were set to 1000 and 0.001, respectively, by cross-validation.

Random Forest (RF). This approach corresponds to assembly learning and is based on two key concepts: decision trees (Franklin, 2005) and bagging (Brieman, 1996). Basically, RF constructs uncorrelated trees, averaging their outputs for robust models, reducing ensemble variance via variable subsets. RF has been widely used in spectral analysis for nutrient assessment, yielding good accuracy results (Ding *et al*., 2018; Vestergaard *et al*., 2021). The selection of an appropriate number of trees is a crucial consideration. Based on the RMSEcv, the number of trees that produced the best results for this research was 300.

Assessment

The predictive performance of the model was assessed using the determination prediction coefficient (R^2), root mean square error of prediction (RMSEp), and residual predictive deviation (RPD). Higher R^2 values and lower RMSEp values indicate more accurate models. RPD, calculated as the quotient of the validation data standard deviation and RMSEp, was used as an additional indicator. In the context of agricultural applications, RPD values between 2 and 3 indicate strong predictive capability, values between 1.5 and 2 suggest models with room for improvement, and values below 1.5 signify poor predictive performance (d'Acqui *et al*., 2010). Each regression model underwent 10 runs and was evaluated by five-fold cross-validation.

Generative adversarial networks and artificial data augmentation

Generative adversarial networks (Goodfellow *et al*., 2020) are a deep learning approach that consists of two neural networks (generator and discriminator) trained in parallel. Generator (G) creates synthetic data, and discriminator (D) distinguishes real from artificial, both using the following cross-entropy function to minimize the gap between real and fake distribution:

$$\arg \min_{G} \max_{D} E_{z,x} [logD\,(G(z)) + log\,(1\text{-}D(x))]$$

where $G(z)$ are the generated data, and $D(G(z))$, $D(x)$, and $(1\text{-}D(x))$ represent the estimated probabilities of true negatives, false negatives, and true positives of classifier D, respectively. The generator's goal is to minimize cross entropy, approximating at each epoch Gaussian noise (z) more and more to the real data, while D aims to maximize it for better discrimination, enhancing sample quality. Due to the computational cost of this process and the available computer equipment, in this study, the maximum value of epochs was established at 300.

To assess the impact of data augmentation, spectral models were trained using a combination of 95 real spectra plus an increasing number of GAN-generated spectra 'n'. This resulted in seven additional sets beyond the original real dataset. Subsequently, models were employed to predict the same validation set. For each set, the added samples ('n_i') were as follows: $n_1 = 0$, $n_2 = 61$, $n_3 = 122$, $n_4 = 183$, $n_5 = 244$, $n_6 = 305$, $n_7 = 366$, $n_8 = 427$.

Artificial data analysis

Statistics such as maximum, minimum, median, percentiles, mean, and standard deviation were used to assess the similarity between artificial and real available nitrogen data. Artificial spectral data underwent principal component analysis (PCA) for dimensionality reduction, facilitating the comparison of generated and real spectra. In this study, the original 4615-dimensional spectra were reduced to 2

components for simplicity. The comparison was conducted using divided violin plots and curve characteristics (mean and standard deviation) to visualize the distribution of generated and real data.

RESULTS AND DISCUSSION

Statistics for real and artificial nitrogen data

Available nitrogen descriptive statistics show similarities in central tendency measures between real and artificial data (Table 1), since the means and medians have low variation percentages. The distribution shapes are similar since both have positive skewness. However, larger differences can be seen in dispersion measures. The standard deviation in real data is 46 %, greater than in artificial data. In addition, the variation coefficient differs in a similar proportion, indicating that artificial values tend to be closer to their mean than the real ones. This indicates that real data have a greater dispersion in nitrogen concentration, which is also reflected in a wider interquartile range as well as the presence of outliers.

Table 1. Descriptive statistics of real and artificial available soil nitrogen (AN) data.

Statistics	Real AN	Artificial AN
	(mg kg^{-1})	
Mean	25.60	25.71
Standard deviation	7.64	5.21
Skewness	0.98	0.68
Variation Coefficient (%)	29.85	20.30
Min	13.34	17.96
Max	47.20	37.50
p25	20.26	22.08
p50	22.94	23.90
p75	29.75	28.7

In general, artificial data shows a similar distribution to real data in terms of central tendency, coinciding in this sense with the results of Jiang *et al*. (2023), but with less variability. This could be an effect of simplification or smoothing in original distribution during the GAN generation process, due to a lack of diversity and training data bias.

Real and artificial spectra analysis

Real spectra

Reflectance spectra were obtained in the NIR spectral range (800–2700 nm), containing 4615 variables. The general trend of NIR spectra was similar, which can be attributed

to the presence of the same spectrally active properties in all samples collected in the field (Xu *et al*., 2018). In both the original spectra (Figure 1A) and derivatives (Figure 1B), noise is evident, particularly in the 800–1000 nm range, with a sloped baseline probably due to non-systematic factors in spectral measurement equipment.

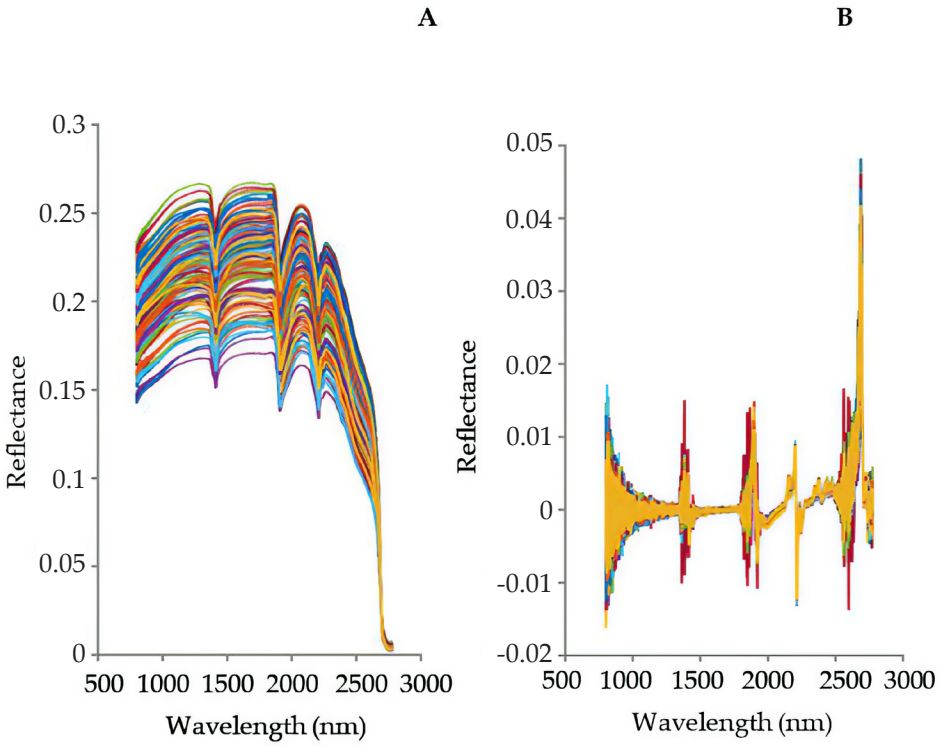

Figure 1. Spectral data pretreatment comparison. A: original spectra; B: first derivative of reflectance.

Field-collected samples showed three prominent absorption features at 1400, 1900, and 2200 nm wavelengths, approximately. The first two absorption regions are attributed to water molecules trapped in the crystal lattice of the soil. At 1400 nm, the absorption is caused by stretching harmonics in oxygen and hydrogen (O-H) bonds; at 1900 nm, these harmonics are combined with H-O-H bending harmonics (Clark *et al*., 1990). These characteristics reflect humid field conditions in samples, which could have led to spectral disturbances, which can make it challenging to establish a nitrogen content and reflectance relationship. The absorption peak at approximately 2200 nm is related to the absorption of Al-OH by clays in the soil (Clark and Roush, 1984), suggesting a high proportion of small-sized particles (silt and clay) in the samples.

Spectral principal components (PC) analysis

The contribution rates of PC1 and PC2 corresponding to real and artificial spectra (Table 2) suggest that both data sets have a similar structure in terms of the first two PCs importance, with a total contribution of 98 %. However, there are differences in the individual contributions of the components, which indicate some discrepancies in the underlying structure of the data.

Table 2. Principal components contribution rates for spectral data.

Spectral set	Contribution rate (%)	
Real	94.67	4.47
Artificial	90.43	8.29

To visualize these similarities and discrepancies, the two-dimensional data probability density distribution was projected using split violin plots (Figure 2A). The similarity among shapes of both distributions suggests that sets have common underlying features or patterns. However, certain discrepancies were observed in distribution

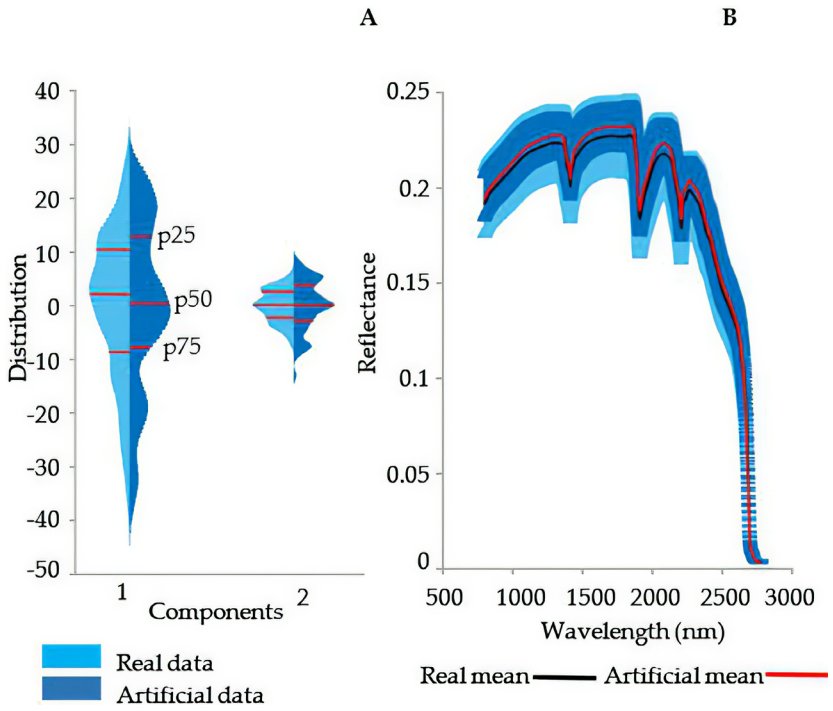

Figure 2. Comparative analysis for real and artificial spectra. A: Divided violin plots for two principal components in real and artificial spectral distributions. B: Graphical comparison for mean and standard deviation reflectance.

lengths, due to differences in variability being greater for real data. In addition, median and interquartile distance variations also indicate a systematic difference in the central and extreme values of each set. Likewise, the above is reflected in a great standard deviation for real data (Figure 2B).

Artificial spectra exhibit a broad resemblance to the original spectra in terms of their overall shape. Common absorption features are discernible at 1499, 1900, and 2200 nm in both datasets. However, these characteristics are accompanied by reduced spectral scattering, along with fluctuations in mean reflectance in artificial spectra.

Data augmentation impact on regression algorithms

To carry out regressions, 10 wavelengths were selected using GA-PLS. These variables were located mainly in the spectral range of 900–1010 nm, a spectral zone strongly related to nitrogen content in soil (Zhang *et al*., 2019). Similarly, wavelengths from 1315 to 1400 nm were correlated with nitrogen, consistent with Zhou *et al*. (2018).

The RMSEp and R^2 graphs show changes for each model (Figure 3). Each point represents a model performance, with the first point corresponding to the original set

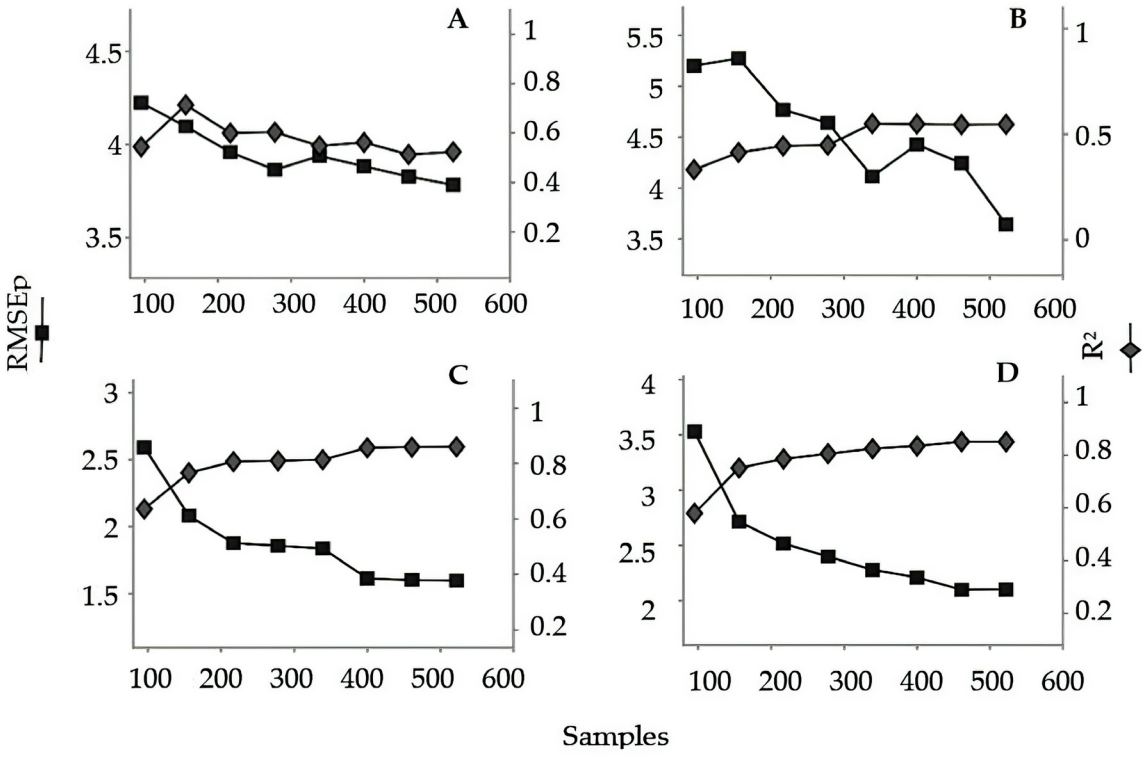

Figure 3. Effect of increased data on prediction parameters for regression models. A: Partial Least Square; B: Extreme Learn Machine; C: Support Vector Machine; D: Random Forest.

and the remaining points representing the seven data augmentation levels. When only real data was used for calibration, R^2 was low for most of models, with the best being SVM, with $R^2 = 0.63$ and RMSEp = 2.59.

The models' performance after data augmentation revealed that artificial data contributed positively to algorithm learning, which is reinforced by Jiang *et al.* (2023). More samples led to improved R^2 and reduced RMSEp values. However, beyond a certain point, model performance plateaued or declined. Finding the optimal data augmentation level requires a balance between performance gains and added complexity from artificial data. In this research, there were levels of increase in which the algorithms performed key behaviors. For example, the fourth level proved crucial for PLS and ELM learning. Beyond 244 training samples, further increases had minimal impact on reducing RMSEp for PLS, with R^2 remaining similar to the non-increased model (Figure 3A).

On the other hand, although RMSEp values decreased at later levels, ELM's explanatory capacity for total variability plateaued after level four, resulting in stabilized R^2 values (Figure 3B). SVM and RF (Figures 3C and 3D) showed consistent behavior during training data augmentation. R^2 and RMSEp values indicated continuous model improvement and generalizability, with a direct relationship between errors and the R^2 coefficient. At levels five and six for SVM and RF, RMSEp and R^2 remained stable or worsened, indicating a point of model stability where additional data didn't significantly benefit. Further data addition could have introduced noise and ambiguity, leading to metric stabilization or deterioration as observed by Jiang *et al*. (2023).

The impact of data aggregation on accuracy and maximum RPD scores (Table 3) helps to identify the optimal augmentation levels for each model. PLS performed best with 122 samples but achieved a lower RPD value compared to similar studies (Li *et al*., 2019b) with a similar sample size, likely due to data quality and noise. Similarly, ELM showed its highest performance with 522 training samples but fell short of achieving the high RPD scores seen in studies with fewer samples (Li *et al*., 2019a). SVM achieved peak performance with 400 samples, which is consistent with the nitrogen prediction study by Xu *et al*. (2018). For RF, the sixth level with 461 samples achieved optimal stability and balance between all metrics.

Model performance

Calibration and prediction accuracy

In general, a minimal gap between calibration and prediction errors (RMSEt and RMSEp) signifies a well-balanced model, combining fitting capacity for training data with generalization ability for new data. This is desirable since it indicates model reliability and robustness (Montesinos-López *et al*., 2022).

Considering the boundaries of each model's learning capacity, the PLS model consistently adapts to new data as artificial data increases. Minor discrepancies between RMSEt and RMSEp errors were observed across most augmentation levels.

Table 3. Metrics for regression model performance through training data augmentation.

Metrics	Algorithm	Real data	Data augmentation levels						
		(95)	1(+61)	2(+122)	3(+183)	4(+244)	5(+305)	6(+366)	7(+427)
RMSEt	PLS	3.76	2.78	3.19	3.10	3.47	3.40	3.54	3.51
	SVM	1.22	0.94	0.84	0.78	0.76	0.76	0.74	0.72
	ELM	3.85	2.86	3.62	4.09	4.13	4.31	4.23	4.26
	RF	3.03	2.81	2.46	2.10	2.14	2.07	2.15	2.10
RMSEp	PLS	4.22	4.09	3.95	3.86	3.93	3.88	3.82	3.78
	SVM	2.59	2.08	1.87	1.85	1.83	1.61	1.60	1.59
	ELM	5.20	5.27	4.76	4.64	4.11	4.42	4.54	3.64
	RF	3.53	2.71	2.51	2.39	2.27	2.21	2.09	2.10
R^2	PLS	0.54	0.71	0.6	0.60	0.54	0.56	0.51	0.52
	SVM	0.63	0.76	0.80	0.81	0.81	0.85	0.85	0.86
	ELM	0.33	0.41	0.44	0.44	0.55	0.54	0.54	0.54
	RF	0.58	0.75	0.78	0.80	0.82	0.83	0.85	0.85
RPD	PLS	1.05	1.36	1.40	1.44	1.41	1.43	1.45	1.47
	SVM	1.78	2.21	2.51	2.55	2.57	2.89	2.90	2.90
	ELM	1.08	1.05	1.18	1.20	1.39	1.28	1.32	1.54
	RF	1.57	2.04	2.20	2.31	2.43	2.51	2.64	2.63

Performance obtained by five-folds cross-validation (reported metrics are average values). RMSEt: root mean square error for calibration; RMSEp: root mean square error for prediction; RPD: residual predictive deviation.

SVM initially overfitted more than PLS in early augmentation levels, but a significant decrease in this disparity in later levels indicated improved model generalizability with more data. ELM showed higher error differences compared to PLS in the first two augmentation levels but reduced differences thereafter. Both models were more sensitive to artificial data at the first level. Finally, RF consistently showed minimal error differences across all levels, highlighting superior generalization and effective use of artificial data for improved prediction.

Prediction Fittings

All four methods achieved peak prediction accuracy at different augmentation levels. The SVM model, with 400 samples (95 real and 305 artificial) at level five (Figure 4A), produced the best results: $R^2 = 0.857$ and RPD = 2.89, demonstrating strong fit to artificial data. Moreover, SVM outperformed linear methods, as noted by Zhang *et al*. (2016) for soil nitrogen. RF (Figure 4B) performed well at level six (461 samples), with $R^2 = 0.859$ and RPD = 2.64. After SVM, RF had the lowest errors, indicating good adjustability and accuracy. Conversely, both ELM and PLS displayed poorer fit and limited predictive ability (Figures 4C and 4D). They had larger errors and low RPD values, making them less feasible.

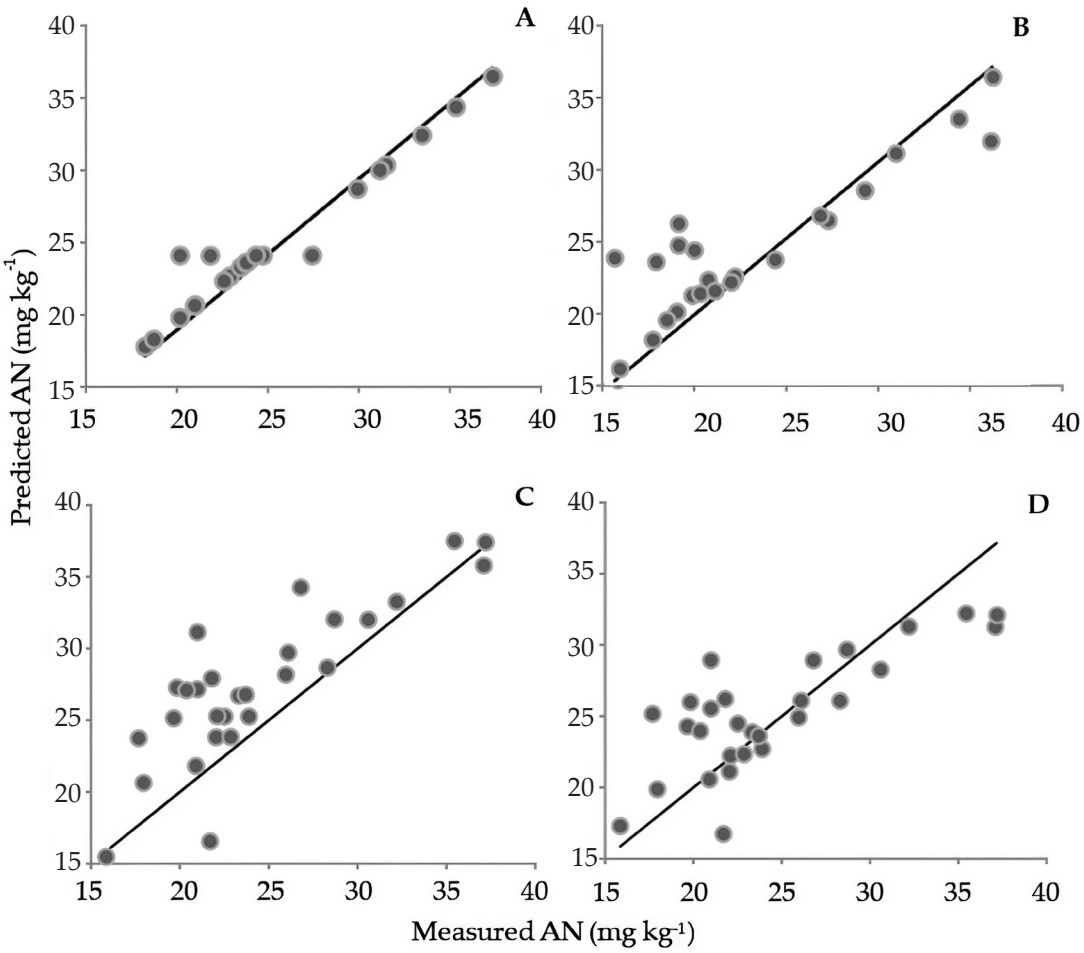

Figure 4. Scatter diagrams for available nitrogen content prediction showing best fits for each model with augmented data. A: Support Vector Machine (SVM) on the fifth level; B: Random Forest (RF) on the sixth; C: Extreme Learning Machine (ELM) on the seventh; D: Partial Least Squares (PLS) on the second.

These findings align with spectral studies on soil available nitrogen prediction, as observed by Xu *et al*. (2018), where nonlinear multivariate methods like SVM outperform linear models such as PLS. ELM, despite being the weakest among nonlinear methods, still outperforms the linear approach, as supported by Li *et al*. (2019a). Additionally, RF's strong performance in soil nitrogen prediction is consistent with Ding *et al*. (2018) and Vestergaard *et al*. (2021), affirming its suitability for such applications.

Differences in performance result from algorithm structural diversity. SVM and RF, as deeper nonlinear methods, capture complex relationships better, leading to superior prediction compared to ELM and PLS. SVM employs kernel functions to map data nonlinearly onto a high-dimensional hyperplane, enhancing its ability to capture nonlinear relationships. On the other hand, RF, as an ensemble of decision trees,

leverages diverse random data samples in training. This diversity enables it to capture non-linear relationships and interdependencies in predictor variables. Combining tree predictions reduces bias and model variance, enhancing generalization capacity (Cootes *et al*., 2012).

ELM, though non-linear, has a simpler structure than SVM and RF. It employs a single hidden layer with non-linear mapping via an activation function. However, random weight selection and a lack of iterative fitting may limit its complex nonlinear relationship capture. PLS, unlike other methods, is a linear approach aiming to maximize covariance between predictors and target variables, assuming linearity. This may limit modeling complex nonlinear data relationships.

CONCLUSIONS

Based on the employed evaluation metrics, the Support Vector Machine (SVM) and Random Forest (RF) models demonstrated superior predictive performance, effectively capturing the relationships between soil reflectance and soil available nitrogen. Although SVM and RF models achieved higher accuracy and robustness, the Extreme Learning Machine (ELM) model and the linear approach Partial Least Squares model (PLS) exhibited limitations in their predictive capacity. This highlights the importance of selecting appropriate algorithms for soil nitrogen prediction.

In general, the utility of generative adversarial networks (GANs) is reaffirmed as a viable alternative for augmenting and enhancing NIR reflectance spectral databases. Using GANs to augment NIR spectral data improved the performance of the regression models, enhancing their accuracy and generalization capabilities.

ACKNOWLEDGEMENTS

We thank the National Council of Humanities, Sciences, and Technologies (CONAHCYT) for the scholarship grant to A.E.R.-R., as well as the Soil Physics and Mechanics Laboratory of the Department of Agricultural Mechanical Engineering (DIMA) and the National Laboratory for Agricultural, Food and Forestry Research and Service (LANISAF) for having provided the necessary tools for the development of this study.

REFERENCES

Barra I, Haefele SM, Sakrabani R, Kebede F. 2021. Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–A review. TrAC Trends in Analytical Chemistry 135: 116166. https://doi.org/10.1016/j.trac.2020.116166

Brieman L. 1996. Bagging predictors. Machine Learning 24 (2): 123–140. https://doi.org/10.1007/bf00058655

Burger J, Geladi P. 2007. Spectral pre-treatments of hyperspectral near infrared images: Analysis of diffuse reflectance scattering. Journal of Near Infrared Spectroscopy 15 (1): 29–37. https://doi.org/10.1255/jnirs.717

Clark RN, King TV, Klejwa M, Swayze GA, Vergo N. 1990. High spectral resolution reflectance spectroscopy of minerals. Journal of Geophysical Research 95 (B8): 12653–12680. https://doi.org/10.1029/jb095ib08p12653

Clark RN, Roush TL. 1984. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. Journal of Geophysical Research 89 (B7): 6329–6340. https://doi.org/10.1029/JB089iB07p06329

Cootes TF, Ionita MC, Lindner C, Sauer P. 2012. Robust and accurate shape model fitting using random forest regression voting. *In* Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C. (eds.), Computer Vision – ECCV 2012. Lecture Notes in Computer Science, vol 7578. Springer: Berlin, Germany. https://doi.org/10.1007/978-3-642-33786-4_21

d'Acqui LP, Pucci A, Janik L J. 2010. Soil properties prediction of western Mediterranean islands with similar climatic environments by means of mid-infrared diffuse reflectance spectroscopy. European Journal of Soil Science 61 (6): 865–876. https://doi.org/10.1111/j.1365-2389.2010.01301.x

Ding J, Yang A, Wang J, Sagan V, Yu D. 2018. Machine-learning-based quantitative estimation of soil organic carbon content by VIS/NIR spectroscopy. PeerJ 6: e5714. https://doi.org/10.7717/peerj.5714

Fei Q, Li M, Wang B, Huan Y, Feng G, Ren Y. 2009. Analysis of cefalexin with NIR spectrometry coupled to artificial neural networks with modified genetic algorithm for wavelength selection. Chemometrics and Intelligent Laboratory Systems 97 (2): 127–131. https://doi.org/10.1016/j.chemolab.2009.03.003

Franklin J. 2005. The elements of statistical learning: Data mining, inference and prediction. The Mathematical Intelligencer 27 (2): 83–85. https://doi.org/10.1007/BF02985802

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2020. Generative adversarial networks. Communications of the ACM 63 (11): 139–144. https://doi.org/10.1145/3422622

Guiñón JL, Ortega E, García-Antón J, Pérez-Herranz V. 2007. Moving average and Savitzki-Golay smoothing filters using Mathcad. *In* International Conference on Engineering Education. Coimbra, Portugal. 4 p.

Jiang C, Zhao J, Ding Y, Li G. 2023. Vis-NIR spectroscopy combined with GAN data augmentation for predicting soil nutrients in degraded alpine meadows on the Qinghai-Tibet plateau. Sensors 23 (7): 3686. https://doi.org/10.3390/s23073686

Li H, Jia S, Le Z. 2019a. Quantitative analysis of soil total nitrogen using hyperspectral imaging technology with extreme learning machine. Sensors 19 (20): 4355 https://doi.org/10.3390/s19204355

Li Y, Yang Q, Chen M, Wang M, Zhang M. 2019b. An ISE-based on-site soil nitrate nitrogen detection system. Sensors 19 (21): 4669. https://doi.org/10.3390/s19214669

Liakos KG, Busato P, Moshou D, Pearson S. 2018. Machine learning in agriculture: A review. Sensors 18 (8): 2674. https://doi.org/10.3390/s18082674

Liu J, Xie J, Han J, Wang H, Sun J, Li R, Li S. 2020. Visible and near-infrared spectroscopy with chemometrics are able to predict soil physical and chemical properties. Journal of Soils and Sediments 20 (7): 2749–2760. https://doi.org/10.1007/s11368-020-02623-1

Lv N, Ma H, Member S, Chen C, Member S. 2021. Remote sensing data augmentation through adversarial training. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14: 9318–9333. https://doi.org/10.1109/jstars.2021.3110842

Montesinos-López OA, Montesinos-López A, Corssa J. 2022. Over fitting, model tuning, and evaluation of prediction performance. *In* van Euwijk F. (ed.), Multivariate Statistical Machine Learning Methods for Genomic Prediction. Springer Nature: Cham, Switzerland, pp: 109–139. https://doi.org/10.1007/978-3-030-89010-0_4

Patel AK, Ghosh JK, Sayyad SU. 2020. Fractional abundances study of macronutrients in soil using hyperspectral remote sensing. Geocarto International 37 (2): 474–493. https://doi.org/10.1080/10106049.2020.1720315

Qi H, Paz-Kagan T, Karnieli A, Jin X, Li S. 2018. Evaluating calibration methods for predicting soil available nutrients using hyperspectral VNIR data. Soil and Tillage Research 175: 267–275. https://doi.org/10.1016/j.still.2017.09.006

Stenberg B, Viscarra Rossel RA, Mouazen AM, Wetterlind J. 2010. Visible and near infrared spectroscopy in soil science. Advances in Agronomy 107: 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7

Vestergaard RJ, Adamchuk V, Biswas A. 2021. Evaluation of optimized preprocessing and modeling algorithms for prediction of soil properties using VIS-NIR spectroscopy. Sensors 21 (20): 6745. https://doi.org/10.3390/s21206745

Wu M, Wang S, Pan S, Terentis AC, Strasswimmer J. 2021. Deep learning data augmentation for Raman spectroscopy cancer tissue classification. Scientific Reports 11 (1). https://doi.org/10.1038/s41598-021-02687-0

Xu S, Zhao Y, Wang M, Shi X. 2018. Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by Vis–NIR spectroscopy. Geoderma 310: 29–43. https://doi.org/10.1016/j.geoderma.2017.09.013

Yun Y, Li H, Deng B, Cao D. 2019. An overview of variable selection methods in multivariate analysis of near-infrared spectra. TrAC Trends in Analytical Chemistry 113: 102–115. https://doi.org/10.1016/j.trac.2019.01.018

Zhang Y, Li M, Zheng L, Qin Q, Suk W. 2019. Spectral features extraction for estimation of soil total nitrogen content based on modifi ed ant colony optimization algorithm. Geoderma 333: 23–34. https://doi.org/10.1016/j.geoderma.2018.07.004

Zhang Y, Li MZ, Zheng LH, Zhao Y, Pei X. 2016. Soil nitrogen content forecasting based on real-time NIR spectroscopy. Computers and Electronics in Agriculture 124: 29–36. https://doi.org/10.1016/j.compag.2016.03.016

Zhou P, Yang W, Li M, Yao X, Liu Z. 2018. Performance analysis of vehicle-mounted soil total nitrogen detector. IFAC-PapersOnLine 51 (17): 51–56. https://doi.org/10.1016/j.ifacol.2018.08.071