Application of a Preproceessing Pipeline to VIS-NIR Data for Predicting Soil Nutrient Concentration Values

Eduardo Menezes de Souza Amarante[1*], Julian Santana Liang[1], Carlos Alberto Campos da Purificação[1], Rômulo Alexandrino Silva[1]

[1]Department of Software, SENAI CIMATEC University; Salvador, Bahia, Brazil

This paper compares the impact of each preprocessing step in predicting soil nutrient concentration values using the partial least squares technique (PLS). The preprocessing pipeline comprises log transformation of the output variable, determination of the optimal number of components, and feature engineering. An increase in the coefficient of determination (R^2) and an improvement in model stability were observed.

Keywords: VIS-NIR Data. Preprocessing Pipeline. Soil Nutrient. Partial Least Square Regression.

Standard procedures for measuring soil properties are time-consuming, complex, and expensive. Therefore, an analytical technique that is fast, precise, and affordable is necessary to determine soil fertility levels.

The conventional spectroscopic modeling procedure requires the pretreatment of soil samples, such as drying and sieving, before scanning with a spectrophotometer [1]. Near-infrared spectroscopy (NIR) has been widely used to meet these needs. In addition, it can analyze many constituents simultaneously, making it a viable alternative to conventional laboratory analyses for assessing and monitoring soil quality [2-5].

He and colleagues [6] predicted levels of nitrogen (N), phosphorus (P), potassium (K), soil organic matter (OM), and pH content from NIR spectroscopy data. Wetterlind and colleagues [7] determined the soil texture, SOM, total N, pH and plant-available P, K and Mg from visible and near infrared reflection. Jin and colleagues [8] tested twenty-nine preprocessing combination techniques with VIS-NIR data of yellow loam samples to find the best combination for predicting potassium levels.

Received on 28 September 2024; revised 18 November 2024.
Address for correspondence: Eduardo Menezes de Souza Amarante. Av. Orlando Gomes, 1845, Piatã. Zipcode: 41650-010. Salvador, Bahia, Brazil. E-mail: eduardo.amarante@fbter.org.br.

J Bioeng. Tech. Health 2024;7(4):369-374

This research aimed to determine whether the chained processing techniques improve the performance of the partial least squares regression model. The models were evaluated using the coefficient of determination (R^2) and root mean squared error (RMSE).

Materials and Methods

Materials

In total, 420 soil samples were collected from five soil classes at two depths: 0- 20 cm and 20-40 cm, with 210 samples for each depth. The collected samples were dried at room temperature for seven to fifteen days. Before absorbance measurement, the soil samples were ground and sieved using a 2 mm mesh size to remove the particle effect size on reflectance spectra.

The spectrometer used was a FieldSpec 3, with a spectral range between 350 and 2500 nm, a resolution of 8 nm, and a precision of +/- 1 nm. A Spectralon ceramic plate was used to calibrate the device before each measurement. Each soil sample was measured 30 times, and the mean spectrum for each sample was calculated.The mean reflectance values were converted into absorbance measures using the formula log(1/R), where R represents the reflectance. The concentration values of boron were extracted using Mehlich-1 extraction, measured in mg/dm^3.

Exploratory Data Analysis

Initially, we conducted an exploratory data analysis on the entire dataset to identify patterns in the input and output variables. Figure 1 shows the distribution of nutrients with the outliers highlighted.

We performed a descriptive statistic of the nutrient values to obtain more precise information about the dataset. Table 1 shows the number of samples, minimum and maximum values, mean, median, and standard deviation. The distribution curve is right-skewed, with values concentrated near zero. Therefore, we applied an asymmetrical correction with log(1+x). Figure 2 shows the distribution of B values before and after the correction.

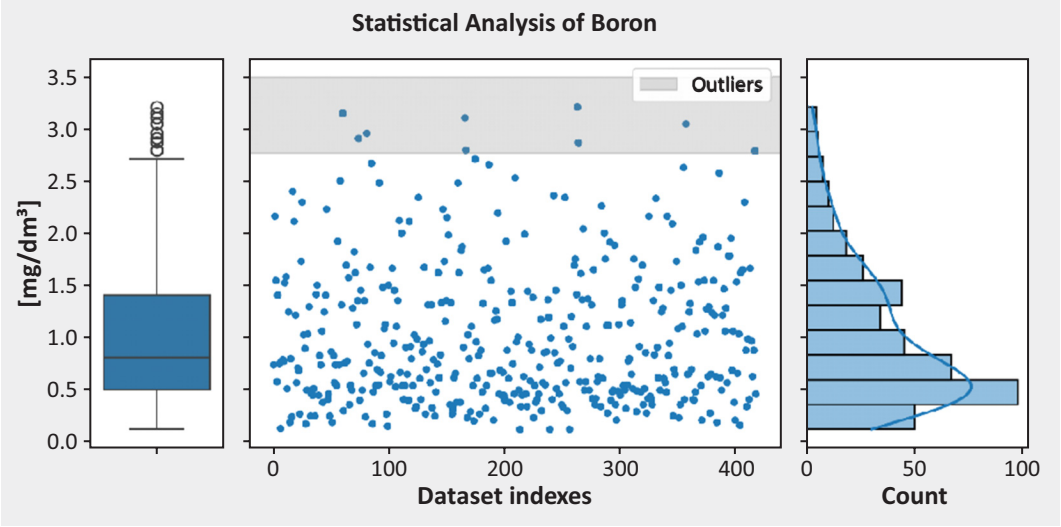
Figure 3 illustrates the average absorbance signal across the entire spectrum. In this picture, it is possible to identify the absorption region of mean peaks.

During the data exploration analysis for VIS-NIR data, we identified some records with a coefficient of absorption more significant than 1, which lacks physical meaning. A shift in absorbance values could correct them, but we kept them. We plotted the distribution curve from each input variable and realized that all of them are nearly bell-shaped, as shown in Figure 4. These outliers were kept, avoiding data shortage.

Pipeline

Before applying the processing pipeline, we randomly separated eighty-four soil samples to constitute the testing dataset, ensuring they were excluded from all preprocessing steps. The remaining dataset was split into training and

Figure 1. Statistical analysis of boron nutrient.



On the left, the boxplot shows the outliers values with concentration values above 2.76875 mg/dm³. In the middle, the scatterplot illustrates how the concentration values are distributed according to the dataset indexes. On the right, we have the histogram of the concentration values.

Table 1. Stats of nutrient concentration values.

Count	Min	Max	Mean	Median	Standard Deviation
420	0.1100	3.2100	1.0014	0.8050	0.6697

Figure 2. Correctness of distribution with log transformation applied. The dotted and dashed lines represent the median and mean values of distribution, respectively.

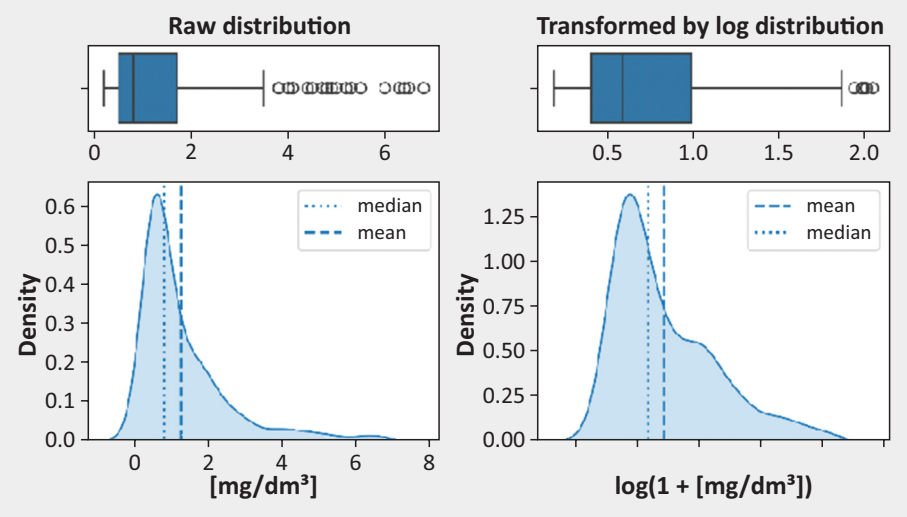


Figure 3. NIR spectrum mean.

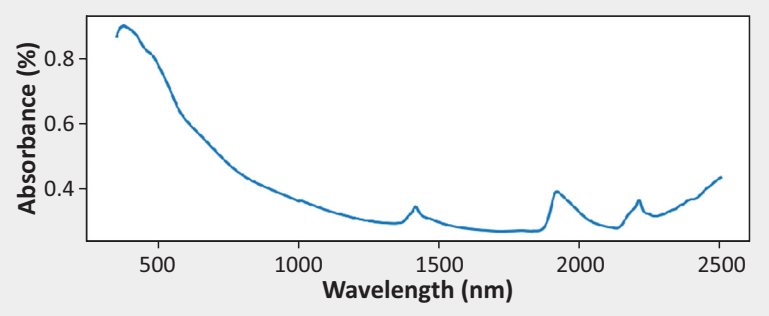
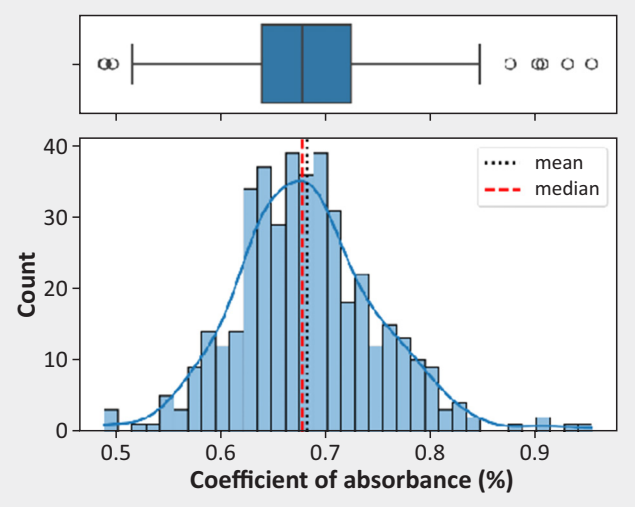


Figure 4. Distribution of absorption coefficients at 550 nm wavelength.



validation datasets, with 252 samples allocated for training, 84 for validation, and 84 for testing, respectively.

According to Figure 5, the pipeline starts with the baseline model, followed by applying three different preprocessing techniques. After each step, the coefficient of determination (R^2) and the root mean square error (RMSE) were calculated on the validation dataset, allowing us to assess the impact of each preprocessing step on the model's performance. Each step following the log transformation in this pipeline can be seen as an additional preprocessing layer sequentially added as the pipeline progresses. As a result, four different models were produced and will be evaluated using the testing dataset. Table 2 outlines the processing layers at each step. K-fold cross-validation with ten folds was applied to determine the number of components optimally.

The PLS regressor used the resulting value to obtain the R^2 and RMSE error. This value was then carried forward through the subsequent steps of the pipeline. The pipeline ends with a statistical feature engineering process applied to each row in the dataset, incorporating features such as Q3/Q1, Q3 x Q1, number of peaks, kurtosis, skew, Q1, mean, minimum, and maximum values. It's important to note that the outliers in the output variables, as shown in Figure 1, were not removed.

After each step, the models were evaluated using the validation dataset, allowing their results to be compared.

Results and Discussion

Partial least squares regression analysis results for the dry soil samples for boron determination (Table 3). The R^2 values in the validation dataset

Figure 5: Pipeline flows.

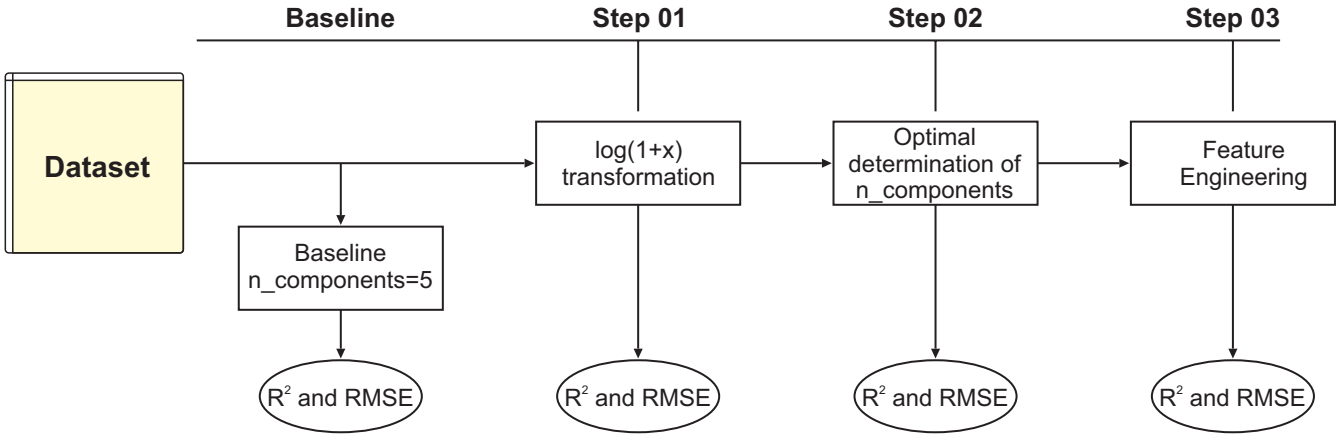


Table 2. Chained pipeline processing.

Steps	Processes		
	Log	N components optimization	Feature Engineering
Step 1	x		
Step 2	x	x	
Step 3	x	x	x

Table 3. Partial least squares result of the dry soil for validation and testing datasets.

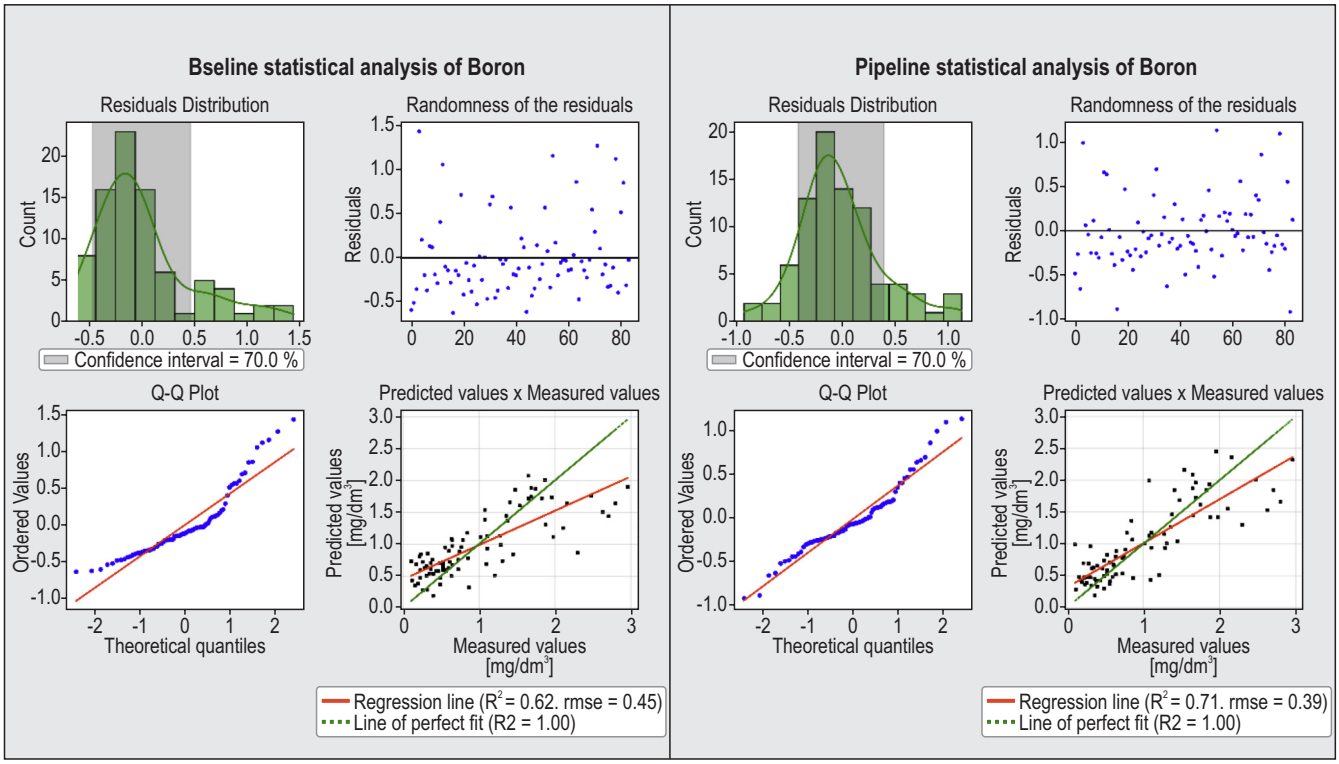
Steps	Validation		Test	
	R^2	RMSE	R^2	RMSE
Baseline	0.4937	0.4525	0.6150	0.4454
Step 1	0.5008	0.4493	0.6091	0.4489
Step 2	0.6304	0.3866	0.7175	0.3816
Step 3	0.6623	0.3696	0.7067	0.3884

for each step were 0.4937, 0.5008, 0.6304, and 0.6623, showing a relevant improvement. The models fitted with the entire dataset were tested on an unknown testing dataset. In the testing dataset, the coefficients of determination were 0.6150 for baseline, 0.6091 for step 01, 0.7175 for step 02, and 0.7067 for step 03, respectively. Figure 6 shows the statistical analysis of the baseline and step 03 regression residuals from the testing dataset. After passing through the entire pipeline, the residuals produced follow a normal distribution.

Conclusion

The preprocessing techniques applied in the dataset produced a relevant improvement in RMSE and R^2 metrics, starting with $R^2 = 0.6150$ and RMSE = 0.4454 mg/dm³ for the baseline model and ending up with $R^2 = 0.7067$ and RMSE

Figure 6. Statistical analysis of residuals produced by the regression. On the left is the baseline model, and on the right, after passing through the pipeline.



equal to 0.3884 mg/dm³. The R^2 values obtained for the testing dataset were more significant than those for the validation dataset, but the RMSE error values were similar. The differences between R^2 and RMSE values obtained from validation and testing datasets can be explained by a shortage of data at certain concentration levels resulting from different data distributions on training, validation, and testing sets. The findings indicate that preprocessing techniques are crucial in producing a useful predictive model for soil nutrient concentration from spectral data. In addition, the amount of data is also a critical factor in the model's performance. The reduced dataset, particularly with limited representation at certain concentration levels, introduces variability in the model's ability to generalize across the full spectrum of data. With fewer data points, the model might be capturing noise or specific patterns in the validation set that do not generalize well to the testing set. This reinforces the need for a more prominent and representative dataset for future works.

Acknowledgments

This research was executed in partnership between SENAI CIMATEC and ITECH startup. The authors would like to acknowledge the Brazilian Company for Industrial Research and Innovation (EMBRAPII) 's support and investments in RD&I.

References

1. Maleki MR et al. Phosphorus sensing for fresh soils using visible and near infrared spectroscopy. Biosystems Engineering 2006;95(3):425-436. doi: 10.1016/j.biosystemseng.2006.07.015.
2. Xu S. et al. Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by Vis–NIR spectroscopy. Geoderma 2018;310:29-43.
3. Vasques GM, Grunwald S, Sickman JO. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma, 2008;146:14–25.
4. Vohland M, Besold J, Hill J, Fründ H. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. Geoderma 2011;166:198–205.
5. Chang CW, Laird DA, Mausbach MJ, Hurburgh CR. Near infrared reflectance spectroscopy—principal components regression analysis of soil properties. Soil Science Society of America Journal 2001;65:480-490.
6. He Y, Huang M, García A, Hernández A, Song H. Prediction of soil macronutrients content using near-infrared spectroscopy. Computers and Electronics in Agriculture 2007;58:144-153.
7. Wetterlind J, Stenberg B, Söderström M. Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models. Geoderma 2010;156:152-160. Available at: http://www.elsevier.com/locate/geoderma.
8. Jin X et al. Prediction of soil-available potassium content with visible near- infrared ray spectroscopy of different pretreatment transformations by the boosting algorithms. Applied Sciences 2020;10(4):1-15. Available at: https://www.mdpi.com/2076-3417/10/4/1397.