Soil Organic Carbon Estimation Using NIRS and MIRS Spectroscopy in the Senegalese Groundnut Basin: A Rapid Approach for Sustainable Soil Management

.

ABSTRACT

|  |
| --- |
| Senegalese agriculture relies heavily on peanut cultivation, but agricultural intensification has led to soil degradation and a decline in fertility. Soil organic carbon (SOC) is a key indicator of soil quality, influencing its structure and fertility. However, conventional SOC analysis methods are costly and time-consuming. Infrared spectroscopy (NIRS and MIRS) offers a fast and non-destructive alternative, allowing SOC estimation based on the soil’s spectral properties.The study, conducted in the Senegalese Peanut Basin, involved the analysis of 240 soil samples at two depths (0–10 cm and 10–30 cm). Spectra were acquired using NIRS and MIRS, then calibrated with reference measurements obtained through CHNSO analysis. Various spectral preprocessing techniques (SNV, SG, MSC, etc.) and machine learning models (PLSR, SVM, Random Forest, XGBoost) were tested to optimize SOC prediction.The results show that the SVM and Random Forest models offer the best performance, particularly with NIRS spectra preprocessed using Savitzky-Golay, achieving a coefficient of determination (R²) above 0.8 and an RPD > 2, indicating sufficient accuracy for soil management applications. This study highlights infrared spectroscopy as a promising tool for the rapid and cost-effective mapping of SOC, contributing to improved agricultural soil fertility management. |

*Keywords: Soil organic carbon (SOC); Near-infrared spectroscopy (NIRS); Mid-infrared spectroscopy (MIRS); Machine learning; Predictive modeling; Senegalese Groundnut Basin Senegalese Groundnut Basin*

1. INTRODUCTION

Senegalese agriculture relies heavily on peanut cultivation, which plays a central role in the country's economy and food security. However, the intensification of agricultural practices, combined with a decrease in organic matter inputs, has led to a gradual degradation of soils and a decline in their fertility (Tondoh et al., 2013; Lal, 2014). Soil organic carbon (SOC) is a key indicator of soil quality, influencing its structure, water retention capacity, and fertility (Six et al., 2002). Its accurate assessment is therefore essential for guiding sustainable soil management practices and mitigating the effects of degradation (Batjes, 2014).

Conventional methods for measuring organic carbon, based on chemical oxidation (Walkley & Black, 1934) or elemental CHNSO analysis (Nelson & Sommers, 1996), are costly, time-consuming, and require the use of chemical reagents. To overcome these limitations, near-infrared spectroscopy (NIRS, 350–2500 nm) and mid-infrared spectroscopy (MIRS, 4003–648 cm⁻¹) have emerged as rapid and non-destructive alternatives for SOC quantification (Viscarra-Rossel et al., 2006; Barthès et al., 2016).

Infrared spectroscopy is based on the selective absorption of electromagnetic radiation by the chemical bonds of soil components. This technique captures information on functional groups associated with organic matter and minerals (Ng et al., 2019). By using machine learning models, it is possible to establish a relationship between measured spectra and SOC content, paving the way for precise mapping of soil organic matter (Grinand et al., 2016; Minasny et al., 2011).

However, the effectiveness of infrared spectroscopy strongly depends on the predictive model used and the preprocessing applied to the spectra (Ji et al., 2015). Spectral correction techniques such as Savitzky-Golay (SG), Standard Normal Variate (SNV), or Detrending (DT) help improve signal quality and optimize model accuracy (Conforti et al., 2015). Among the most efficient machine learning models are: Support Vector Machine (SVM): widely used in soil spectroscopy for its robustness in handling nonlinear data (Roudier et al., 2017), Random Forest (RF): effective in capturing complex interactions between spectral variables (Liu et al., 2020), XGBoost: a recent approach that enhances performance through the aggregation of multiple decision trees (Chen & Guestrin, 2016), Partial Least Squares Regression (PLSR)  (Wold et al., 2001)

The objective of this study is to assess the feasibility of using infrared spectroscopy to estimate the organic carbon content of soils in the Senegalese Peanut Basin by testing different combinations of machine learning models and spectral preprocessing techniques. The performance of the models will be compared using classical indicators such as the coefficient of determination (R²), the root mean square error of prediction (RMSEP), and the Ratio of Performance to Deviation (RPD).

By providing a fast and cost-effective method to assess soil fertility, this study aims to contribute to the improvement of sustainable soil management strategies and support the transition towards agriculture that is more resilient to environmental and climatic challenges.

2. methodology

2.1 Study Area

The study area is located in the Peanut Basin (in central Senegal) within the Sudano-Sahelian zone. It lies between the coordinates 14°27'-14°31'N and 16°25'-15°30'W (Figure 2) and covers an area of 91.7 km². According to the Köppen-Geiger classification, this climate is classified as hot semi-arid (Köppen: BSh). It is divided into a dry season from November to May and a wet season from June to October. Annual rainfall in the area has shown a decreasing trend, averaging 807 mm from 1918 to 1972, compared to 556 mm from 1972 to 2018 (ANACIM, 2020). Annual precipitation in the study area generally varies between 300 and 850 mm. The average annual temperature is 27.5°C, with a maximum temperature of 38°C and a minimum of 15°C. The terrain is almost flat, with an elevation gradient of 10 meters.

The different soil types, according to the WRB classification (World Reference Base, 2015), are mainly Arenosols, locally known as “Dior,” which dominate the area, Lixisols, called “Deck-Diors” in the local language, and Gleysols, referred to as “Decks,” which are nearly insignificant

2.2 Sampling and Sample Preparation

A total of 240 soil samples were collected at two stratified depths to assess both the vertical and horizontal distribution of organic carbon:

* 0–10 cm: Surface horizon, influenced by biomass inputs and agricultural practices. This depth is generally used to estimate active organic carbon (Janik et al., 2007).
* 10–30 cm: Underlying horizon, allowing analysis of organic carbon dynamics at depth and evaluating the impact of leaching and soil management practices.

The samples were air-dried, homogenized, and sieved to 2 mm for physico-chemical analyses. For spectroscopic analyses, a fraction of each sample was finely ground to 0.2 mm to improve spectral signal quality and calibration model accuracy (Bellon-Maurel & McBratney, 2011).

2.3 Spectroscopic Analyses and Model Calibration

Spectra were acquired using two spectroscopic technologies:

* Near-Infrared Spectroscopy (NIRS, 350–2500 nm, 1 nm resolution): Conducted with the LabSpec4 Standard RES spectrometer, capturing molecular vibrations of C-H, N-H, and O-H bonds, which are commonly associated with soil organic matter (Shepherd & Walsh, 2002).
* Mid-Infrared Spectroscopy (MIRS, 4003–648 cm⁻¹, 7.45 cm⁻¹ resolution): Performed with the Agilent Technologies 4300 Handheld FTIR spectrometer, which is more sensitive to carbon functional groups and soil minerals (Ng et al., 2019).

2.4 Calibration of Predictive Models

Model calibration was performed by associating spectral data with reference measurements of soil organic carbon. To this end, 240 samples were analyzed using a CHNSO autoanalyzer, a reference technique providing accurate total organic carbon measurements through high-temperature combustion and elemental detection (Nelson & Sommers, 1996).

2.5 Prediction Models and Performance Indicators

The spectroscopic data used in this study were derived from NIRS and MIRS spectra of soil samples. Several preprocessing techniques were applied, including: SG: Savitzky-Golay filtering for spectral smoothing (Savitzky & Golay, 1964), Log: Logarithmic transformation to reduce extreme variations (Stevens et al., 2013), SNV: Standard normal variate normalization (Barnes et al., 1989), MSC: Multiplicative scatter correction (Geladi et al., 1985), Detrend: Linear trend removal (Rinnan et al., 2009), Raw: Unprocessed spectra.

The evaluated machine learning models included: SVM: Support Vector Machine, a robust nonlinear regression model (Vapnik, 1995), RandomForest: An ensemble model based on decision trees (Breiman, 2001), XGBoost: A gradient boosting model (Chen & Guestrin, 2016), PLSR: Partial Least Squares Regression (Wold et al., 2001).

The performance of the models was assessed using the following indicators:

Coefficient of determination (R²): Measures the proportion of variance explained by the model, reflecting its predictive ability (Draper & Smith, 1998).

Root Mean Squared Error of Calibration (RMSEC): Quantifies model error on the training set.

Root Mean Squared Error of Prediction (RMSEP): Evaluates model error on the validation set, estimating generalization ability (Williams & Norris, 2001).

Mean Absolute Error (MAE): Measures the average deviation between predicted and observed values, providing an alternative to RMSEP that is less sensitive to outliers (Willmott & Matsuura, 2005).

Ratio of Performance to Deviation (RPD): The ratio of the standard deviation of calibration data to RMSEP. A higher RPD indicates better model accuracy (Chang et al., 2001).

Ratio of Performance to Interquartile Range (RPIQ): The ratio of the range of observed values to RMSEP, offering an additional measure o=f prediction quality (Bellon-Maurel et al., 2010).

According to Chang et al. (2001), models are classified into three categories:

A: Excellent models, with RPD > 2 and 0.8 < R² < 1.

B: Moderately accurate models, with 1.4 < RPD < 2 and 0.5 < R² < 0.8.

C: Unreliable models, with RPD < 1.4 and R² < 0.5.

3. results and discussion

3.1 Results

### **3.1.1 Model Performance Analysis**

The performance of the models was evaluated based on two main factors: the spectra used (NIRS and MIRS) and the impact of preprocessing (with and without preprocessing). Performance indicators such as RMSEP, MAE, R², RPD, and RPIQ were calculated for each combination of spectrum and preprocessing. The tables present the performance of the PLSR, Random Forest, SVM, and XGBoost models based on the preprocessing techniques applied to the NIRS and MIRS spectra (Fig.1)



**Fig.1.** **performance of the PLSR, Random Forest, SVM, and XGBoost models based on the R2 applied to the NIRS and MIRS spectra**

#### 3.1.1.1 Model Performance

**PLSR**: This model exhibits modest performance, with R² values below 0.6. The best results are obtained with the Detrend preprocessing for NIRS (R² = 0.54) and SG for MIRS (R² = 0.59). SNV and MSC slightly improve accuracy but do not significantly optimize predictions (table1)

**Table 1: Performance statistics of PLSR models based on spectrum and preprocessing.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| IR | Preprocessing | RMSEP | MAE | R2 | RPD | RPIQ |
| NIRS | Raw | 0.29 | 0.19 | 0.21 | 1.13 | 1.11 |
| NIRS | SNV | 0.23 | 0.19 | 0.51 | 1.41 | 1.39 |
| NIRS | Detrend | 0.22 | 0.17 | 0.54 | 1.47 | 1.45 |
| NIRS | SG | 0.24 | 0.19 | 0.43 | 1.33 | 1.31 |
| NIRS | MSC | 0.23 | 0.19 | 0.51 | 1.41 | 1.39 |
| NIRS | Log | 0.28 | 0.19 | 0.22 | 1.14 | 1.12 |
| MIRS | Raw | 0.22 | 0.17 | 0.52 | 1.44 | 1.42 |
| MIRS | SNV | 0.23 | 0.17 | 0.50 | 1.41 | 1.39 |
| MIRS | Detrend | 0.23 | 0.17 | 0.50 | 1.41 | 1.39 |
| MIRS | SG | 0.21 | 0.16 | 0.59 | 1.55 | 1.53 |
| MIRS | MSC | 0.23 | 0.17 | 0.50 | 1.41 | 1.39 |
| MIRS | Log | 0.23 | 0.17 | 0.48 | 1.39 | 1.36 |

Random Forest**: This model performs better than PLSR. Applying the SG preprocessing on NIRS yields the best results (R² = 0.79, RPD = 2.21). However, performance on MIRS remains lower, with R² not exceeding 0.74. (table2)**

**Table 2: Descriptive statistics of the RandomForest model performance based on spectrum and preprocessing.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| IR | Preprocessing | RMSEP | MAE | R2 | RPD | RPIQ |
| NIRS | Raw | 0.17 | 0.12 | 0.72 | 1.91 | 1.88 |
| NIRS | SNV | 0.16 | 0.11 | 0.76 | 2.05 | 2.02 |
| NIRS | Detrend | 0.16 | 0.10 | 0.75 | 2.03 | 2.00 |
| NIRS | SG | 0.15 | 0.10 | 0.79 | 2.21 | 2.18 |
| NIRS | MSC | 0.16 | 0.11 | 0.76 | 2.04 | 2.01 |
| NIRS | Log | 0.17 | 0.13 | 0.72 | 1.90 | 1.87 |
| MIRS | Raw | 0.21 | 0.15 | 0.63 | 1.57 | 1.55 |
| MIRS | SNV | 0.17 | 0.13 | 0.72 | 1.89 | 1.86 |
| MIRS | Detrend | 0.19 | 0.14 | 0.67 | 1.72 | 1.70 |
| MIRS | SG | 0.17 | 0.13 | 0.74 | 1.90 | 1.87 |
| MIRS | MSC | 0.17 | 0.13 | 0.72 | 1.89 | 1.86 |
| MIRS | Log | 0.21 | 0.16 | 0.63 | 1.56 | 1.54 |

**SVM**: This model demonstrates the best overall performance, with R² reaching 0.83 for SG on NIRS (RPD = 2.37, RPIQ = 2.33). The SNV and MSC preprocessings improve accuracy compared to raw data, although slightly less than SG. On MIRS, the best R² is 0.72 with SNV and MSC (table3)

**Table 3 : Statistiques descriptives des performances du modèle SVM en fonction du spectre et du prétraitement**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| IR | Preprocessing | RMSEP | MAE | R2 | RPD | RPIQ |
| NIRS | Raw | 0.16 | 0.11 | 0.76 | 2.00 | 1.97 |
| NIRS | SNV | 0.16 | 0.10 | 0.77 | 2.02 | 1.99 |
| NIRS | Detrend | 0.16 | 0.11 | 0.75 | 1.96 | 1.93 |
| NIRS | SG | 0.14 | 0.09 | 0.83 | 2.37 | 2.33 |
| NIRS | MSC | 0.16 | 0.10 | 0.77 | 2.02 | 1.99 |
| NIRS | Log | 0.16 | 0.11 | 0.77 | 2.02 | 1.99 |
| MIRS | Raw | 0.18 | 0.12 | 0.69 | 1.82 | 1.79 |
| MIRS | SNV | 0.17 | 0.12 | 0.72 | 1.92 | 1.89 |
| MIRS | Detrend | 0.17 | 0.12 | 0.72 | 1.89 | 1.86 |
| MIRS | SG | 0.19 | 0.13 | 0.67 | 1.73 | 1.70 |
| MIRS | MSC | 0.17 | 0.12 | 0.72 | 1.92 | 1.89 |
| MIRS | Log | 0.18 | 0.12 | 0.70 | 1.84 | 1.81 |

**XGBoost**: XGBoost's performance is lower than that of Random Forest and SVM, although some preprocessing combinations improve results. The best R² is 0.74 with SG on NIRS, but performance remains moderate on MIRS (R² = 0.67) (table4).

##### **Table 4: Descriptive statistics of XGBoost model performance based on spectrum and preprocessing**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| IR | Preprocessing | RMSEP | MAE | R2 | RPD | RPIQ |
| NIRS | Raw | 0.22 | 0.14 | 0.59 | 1.49 | 1.46 |
| NIRS | SNV | 0.19 | 0.11 | 0.74 | 1.73 | 1.70 |
| NIRS | Detrend | 0.18 | 0.12 | 0.71 | 1.75 | 1.72 |
| NIRS | SG | 0.16 | 0.11 | 0.74 | 1.97 | 1.94 |
| NIRS | MSC | 0.19 | 0.11 | 0.74 | 1.73 | 1.70 |
| NIRS | Log | 0.22 | 0.14 | 0.59 | 1.49 | 1.46 |
| MIRS | Raw | 0.24 | 0.17 | 0.52 | 1.32 | 1.30 |
| MIRS | SNV | 0.19 | 0.14 | 0.64 | 1.66 | 1.64 |
| MIRS | Detrend | 0.19 | 0.14 | 0.66 | 1.70 | 1.67 |
| MIRS | SG | 0.19 | 0.14 | 0.67 | 1.70 | 1.68 |
| MIRS | MSC | 0.19 | 0.14 | 0.64 | 1.66 | 1.64 |
| MIRS | Log | 0.24 | 0.17 | 0.52 | 1.32 | 1.30 |

#### 3.1.1.2 Prediction Error (RMSEP and MAE)

The analysis of RMSEP values reveals a general similarity between the NIRS and MIRS spectra. The average RMSEP value for NIRS is 0.1986, while for MIRS, it is slightly lower at 0.1894. In contrast, the standard deviations for both spectra (0.0345 for NIRS and 0.0294 for MIRS) are relatively low, indicating that the prediction errors are moderate and fairly homogeneous within each spectrum. These results suggest that both spectra are interchangeable in terms of prediction error, and no statistically significant difference has been identified for this measure, which is reflected in the relatively close distribution of RMSEP values for both spectra.

The MAE, which measures the mean absolute prediction error, follows a similar trend. The averages for NIRS and MIRS are 0.1318 and 0.1441, respectively, revealing a slight tendency in favor of NIRS, although this difference is not practically significant. The standard deviations (0.0294 for NIRS and 0.0279 for MIRS) suggest that the variation in absolute errors is relatively low, implying that while NIRS has a slight edge in prediction accuracy, both spectra offer similar performance.

### *3.1.1.3 Explanatory Power of the Model (R², RPD, and RPIQ)*

R², the primary indicator of the model's explanatory power, shows that NIRS holds a slight advantage over MIRS. The average R² for NIRS is 0.6537, while for MIRS it is 0.6275. Although NIRS seems to explain slightly more of the variance in the data, the difference between the two spectra is minimal, suggesting that overall, both spectra can be considered to have comparable explanatory capacity. The low standard deviations (0.136 for NIRS and 0.137 for MIRS) further reinforce this observation, indicating that the models are relatively stable.

The RPD, which measures the relative performance of the model in relation to the data variability, is higher for NIRS (1.7731) than for MIRS (1.6503), although this difference is marginal. This means that NIRS exhibits better relative performance in terms of accuracy when compared to the variability of the data. However, the difference between the two spectra does not have a significant impact on the practical application of the models.

The RPIQ follows a similar trend to that of the RPD. For NIRS, the RPIQ is 1.7462, while for MIRS it is 1.6253. This slight difference suggests that both spectra show comparable performance in terms of prediction relative to the interquartile range of the data.

### **3.1.2 Effect of Spectral Type and Preprocessing on Model Performance**

ANOVA tests were conducted to analyze the effect of the factors Spectral Type (NIRS vs MIRS) and Preprocessing (with vs without preprocessing), as well as their interaction, on performance indicators.

*3.1.2.1 Effect of Spectral Type (NIRS vs MIRS)*

The results of the ANOVA tests reveal no statistically significant difference between NIRS and MIRS for the indicators RMSEP, MAE, R², RPD, and RPIQ (p > 0.05). This suggests that, although NIRS shows slight improvements for certain indicators (R², RPD, RPIQ), these differences are not substantial enough to conclude a marked superiority of one spectrum over the other. In other words, the choice between NIRS and MIRS can be made based on other factors such as the ease of spectrum acquisition or associated costs.

*3.1.2.2 Effect of Preprocessing*

The Savitzky-Golay (SG) preprocessing method showed positive effects on model performance, particularly for NIRS. The SG preprocessing improved the performance of SVM and Random Forest models for NIRS by reducing RMSEP and MAE values, suggesting an enhancement in the quality of the spectral data. In contrast, for MIRS, no preprocessing yielded a significant improvement, which could be due to the nature of the data or the specific characteristics of the MIRS spectrum.

*3.1.2.3 Interaction Between Spectral Type and Preprocessing*

The interaction between spectral type and preprocessing showed no significant effect on performance indicators (p > 0.05). This suggests that the effect of preprocessing is independent of the spectral type used. Thus, the impact of the Savitzky-Golay preprocessing on model performance is not modulated by the spectral type (NIRS or MIRS), but rather depends solely on the effectiveness of the preprocessing in handling the data.

### **3.1.3 Comparison of Machine Learning Models**

The performance of different machine learning models was compared based on the RMSEP, MAE, R², RPD, and RPIQ indicators. The best-performing models were identified for each spectrum (Fig.2), and a classification of the models was made according to the criteria of Chang et al. (2001)



**Fig.2. compares NIRS and MIRS based on the RMSEP, MAE, R², RPD, and RPIQ indicators**.

**3.1.3.1 Best-Performing Models for NIRS**

The SVM and RandomForest models demonstrated the best performance for NIRS, achieving R² values as high as 0.83 with Savitzky-Golay preprocessing. These models also exhibited RPD values greater than 2.0, indicating very strong predictive ability. In contrast, XGBoost produced less optimal results, with R² values not exceeding 0.74, reflecting the lower precision of this model compared to the others.

**3.1.3.2 Best-Performing Models for MIRS**

For MIRS, the overall performance was lower compared to NIRS. The SG preprocessing slightly improved the performance of SVM and RandomForest models, but PLSR yielded weaker results, with R² values ranging from 0.50 to 0.59. The XGBoost model was particularly ineffective for MIRS, with R² values dropping as low as 0.52.

### **3.1.4 Categorization of Models According to Chang et al. (2001)**

Using the criteria outlined by Chang et al. (2001), the models were classified into three categories based on their performance. The results showed that NIRS combined with SVM and RandomForest, especially with Savitzky-Golay preprocessing, provided the best performance. In contrast, MIRS showed overall lower performance, particularly with PLSR and XGBoost (Table 5 and Figure 3). However, no significant difference was detected between the spectra, suggesting that the choice between NIRS and MIRS may be more influenced by practical considerations, such as ease of spectrum acquisition, rather than by marked performance differences.

**Table 5: Categorization of Models Based on Performance Indicators (R² and RPD) According to Chang et al. (2001)**

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Models | R² | RPD |
| Category A (Excellent models) | SVM and RandomForest (NIRS + SG) | R² ≈ 0.83 | RPD > 2 |
| Category B (Moderately accurate models) | PLSR (MIRS) | R² ≈ 0.55 - 0.59 | RPD ≈ 1.5 - 1.7 |
|  | XGBoost (NIRS) | R² ≈ 0.74 | RPD ≈ 1.8 |
| Category C (Unreliable models) | XGBoost (MIRS) | R² ≈ 0.52 | RPD ≈ 1.3 |



**Fig.3.classication of Models(NIRS vs MIRS) Based on Performance Indicators (R² and RPD) According to Chang et al. (2001)**

**3.2. Discussion**

### **3.2.1 Influence of Preprocessing on Model Performance**

The results indicate that preprocessing, particularly Savitzky-Golay (SG) filtering, significantly improves the model performances for NIRS spectra, especially for SVM and Random Forest. This improvement can be attributed to SG’s ability to reduce noise while preserving the spectral structure, making it easier to extract relevant information for prediction (Savitzky & Golay, 1964; Rinnan et al., 2009). In contrast, applying various preprocessing techniques to MIRS spectra did not result in significant improvements in model performance, suggesting that this technique may be less suitable for quantifying organic carbon in certain soil types (Stenberg et al., 2010).

**3.2.2 Comparison of Model Performance by Spectroscopic Technique**

Among the models tested, SVM performed best for predicting soil organic carbon, particularly with NIRS spectra, achieving R² values greater than 0.80 in several preprocessing configurations. This result is consistent with other studies highlighting the effectiveness of SVM for near-infrared spectral analysis due to its ability to handle complex relationships between variables (Vapnik, 1995; Wu et al., 2009). However, for MIRS, the performance was less optimal, with R² values often below 0.75, which can be attributed to higher spectral variability in certain pedological contexts (Viscarra Rossel & Behrens, 2010).

**3.2.3 Comparison of Overall Performance of NIRS and MIRS**

The results show that NIRS and MIRS spectra offered similar performances in terms of RMSEP, MAE, R², RPD, and RPIQ, suggesting that neither technique significantly outperforms the other. This finding is consistent with several comparative studies that have shown that the relevance of each method strongly depends on the soil type and the range of organic carbon concentrations being studied (Shepherd & Walsh, 2002; Minasny et al., 2011). Additionally, preprocessing had little effect on model performance for these different indicators, suggesting that the influence of preprocessing may be dependent on the dataset and spectroscopic technique used (Ji et al., 2015).

**3.2.4 Limitations and Future Perspectives**

Although the observed differences between means are not large enough to have a significant practical impact, it is important to note that these results do not exclude situations where data quality or other input variables could more strongly influence model performance. Previous studies have shown that factors such as soil moisture, grain size, or the presence of specific minerals can affect the spectral response and, thus, the prediction accuracy (Gomez et al., 2008; Knadel et al., 2015).

It would, therefore, be beneficial to further explore these factors in other contexts or using alternative preprocessing methods, including more advanced approaches such as diffusion correction techniques or deep learning methods to optimize spectral information extraction (Baath et al., 2018; Shrestha et al., 2021).

This study provides a solid foundation for comparing NIRS and MIRS as tools for predicting soil organic carbon. It also highlights the importance of other factors, such as data quality and model optimization, that may have a more significant impact on prediction accuracy. Future work could focus on integrating new machine learning approaches and evaluating these techniques in more varied environmental contexts.

4. Conclusion

The results of this study emphasize the importance of selecting appropriate learning models and preprocessing techniques to improve the accuracy of predictions for soil organic carbon based on NIRS and MIRS spectra. Among the models tested, SVM and Random Forest, combined with Savitzky-Golay (SG) preprocessing, displayed the best performance for NIRS, achieving R² values over 0.83 and RPD values above 2.0. In contrast, performance for MIRS remained generally lower, with R² values not exceeding 0.75 and RPD often below 1.8, indicating more limited prediction capacity.

However, these results do not allow for a definitive conclusion regarding the superiority of one spectroscopic technique over the other. The observed variability in performance suggests that other factors, such as soil mineral composition, grain size, and moisture, may influence the spectral response and model accuracy. Therefore, further studies with larger (>1000 samples) and more diverse datasets are necessary to deepen the understanding of the underlying mechanisms behind these differences and optimize preprocessing and modeling strategies.
Finally, exploring new approaches, such as integrating deep neural networks or using advanced spectral interference correction techniques, could further enhance the robustness of predictive models. Extending the study to tropical or arid soils would also help better define the optimal conditions for using near and mid-infrared spectroscopy for soil organic carbon quantification.

References

Baath, R., Wadoux, A. M. J. C., & Minasny, B. (2018). Improving soil carbon predictions with ensemble learning and spatial covariates. Geoderma, 337, 229-239. https://doi.org/10.1016/j.geoderma.2018.09.003

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard Normal Variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied Spectroscopy, 43(5), 772-777. https://doi.org/10.1366/0003702894202201

Barthès, B. G., Brunet, D., Hien, E., et al. (2016). Improving the predictions of soil carbon and nitrogen contents by infrared reflectance spectroscopy. Geoderma, 276, 41-50. https://doi.org/10.1016/j.geoderma.2016.04.020

Batjes, N. H. (2014). Total carbon and nitrogen in the soils of the world. European Journal of Soil Science, 65(1), 10-21. https://doi.org/10.1111/ejss.12114

Bellon-Maurel, V., & McBratney, A. (2011). Near-infrared (NIR) and mid-infrared (MIR) techniques for assessing the soil properties. Journal of Near Infrared Spectroscopy, 19(3), 141-158. https://doi.org/10.1255/jnirs.946

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J. M., & McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. Trends in Analytical Chemistry, 29(9), 1073-1081. https://doi.org/10.1016/j.trac.2010.05.006

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chang, C. W., Laird, D. A., Mausbach, M. J., & Hurburgh, C. R. (2001). Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Science Society of America Journal, 65(2), 480-490. https://doi.org/10.2136/sssaj2001.652480x

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

Conforti, M., et al. (2015). Comparing different preprocessing techniques and models to predict soil properties from NIR spectroscopy. Geoderma, 267, 54-63. https://doi.org/10.1016/j.geoderma.2015.12.020

Draper, N. R., & Smith, H. (1998). Applied Regression Analysis. John Wiley & Sons. https://doi.org/10.1002/9781118625590

Feller, C., Albrecht, A., & Tessier, D. (2001). Influence des pratiques culturales sur la dynamique du carbone dans les sols tropicaux. Agronomie, 21(4), 343-364. https://doi.org/10.1051/agro:2001134

Geladi, P., MacDougall, D., & Martens, H. (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat. Applied Spectroscopy, 39(3), 491-500. https://doi.org/10.1366/0003702854248656

Gomez, C., Rossel, R. V., & McBratney, A. B. (2008). Soil organic carbon prediction by hyperspectral remote sensing and field VIS-NIR spectroscopy: An Australian case study. Geoderma, 146(3-4), 403-411. https://doi.org/10.1016/j.geoderma.2008.06.011

Grinand, C., et al. (2016). Estimating soil properties in tropical regions using NIRS. Geoderma Regional, 7(1), 131-141. https://doi.org/10.1016/j.geodrs.2016.02.002

Ji, W., Viscarra Rossel, R. A., Shi, Z., & Zhou, Y. (2015). Accounting for the effects of water and the environment on proximally sensed vis-NIR soil spectra and their calibrations. European Journal of Soil Science, 66(3), 555-567. https://doi.org/10.1111/ejss.12242

Janik, L. J., Merry, R. H., & Skjemstad, J. O. (2007). Can mid-infrared diffuse reflectance analysis replace soil extractions? Australian Journal of Soil Research, 45(8), 613-622. https://doi.org/10.1071/SR07085

Knadel, M., Rossel, R. A. V., Deng, F., Thomsen, A., & Greve, M. H. (2015). Comparing predictive abilities of three quantitative soil organic carbon mapping methods using visible-near-infrared spectroscopy. Geoderma, 239, 229-239. https://doi.org/10.1016/j.geoderma.2014.10.019

Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. Science, 304(5677), 1623-1627. https://doi.org/10.1126/science.1097396

Lal, R. (2014). Soil carbon sequestration to mitigate climate change. Geoderma, 241, 1-14. https://doi.org/10.1016/j.geoderma.2014.01.016

Liu, X., et al. (2020). Random Forest modeling for predicting soil organic carbon stocks. Soil Science Society of America Journal, 84(1), 123-134. https://doi.org/10.1002/saj2.20002

Minasny, B., McBratney, A. B., & Viscarra Rossel, R. A. (2011). Soil spectroscopy: A high-throughput analytical technique for soil science. Geoderma, 163(3-4), 147-156. https://doi.org/10.1016/j.geoderma.2011.05.002

Nelson, D. W., & Sommers, L. E. (1996). Total carbon, organic carbon, and organic matter. Methods of Soil Analysis, 9, 961-1010. https://doi.org/10.2136/sssabookser5.3.c34

Ng, W., Minasny, B., & Montazerolghaem, M. (2019). Mid-infrared spectroscopy for soil analysis: Recent advances and future challenges. Soil Systems, 3(2), 21. https://doi.org/10.3390/soilsystems3020021

Rinnan, Å., van den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. Trends in Analytical Chemistry, 28(10), 1201-1222. https://doi.org/10.1016/j.trac.2009.07.007

Roudier, P., et al. (2017). Predicting soil properties using SVM: a comparison with PLSR. Geoderma, 289, 98-109. https://doi.org/10.1016/j.geoderma.2016.11.021

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36(8), 1627-1639. https://doi.org/10.1021/ac60214a047

Shepherd, K. D., & Walsh, M. G. (2002). Development of reflectance spectral libraries for characterization of soil properties. Soil Science Society of America Journal, 66(3), 988-998. https://doi.org/10.2136/sssaj2002.9880

Shrestha, R., Mousavi, S. M., Rossel, R. A. V., & McBratney, A. B. (2021). Deep learning for soil organic carbon estimation. Geoderma, 383, 114753. https://doi.org/10.1016/j.geoderma.2020.114753

Six, J., et al. (2002). Soil organic matter, aggregation, and aggregate turnover. Soil Science Society of America Journal, 66(4), 1367-1376. https://doi.org/10.2136/sssaj2002.1367

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., & van Wesemael, B. (2013). Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. Geoderma, 223, 61-67. https://doi.org/10.1016/j.geoderma.2013.10.002

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer. https://doi.org/10.1007/978-1-4757-2440-0

Viscarra Rossel, R. A., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158(1-2), 46-54. https://doi.org/10.1016/j.geoderma.2009.12.025

Walkley, A., & Black, I. A. (1934). An examination of the Degtjareff method for determining soil organic matter. Soil Science, 37(1), 29-38. https://doi.org/10.1097/00010694-193401000-00003

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), 79-82. https://doi.org/10.3354/cr030079