PREDICTING SOIL ORGANIC CARBON IN NORTHWESTERN BRAZIL USING VISIBLE, NEAR-INFRARED, AND MID-INFRARED SPECTROSCOPY AND MACHINE LEARNING ALGORITHM

Borges Marfrann Dias Melo[1], Miguel Palacio Pelaez Carvalho Neto[1], Jean de Jesus Macedo Novais[1], Bruno dos Anjos Bartch[1], Samuel Silva de Moraes[1], Nícolas Augusto Rosin[1], Merilyn Taynara Accorsi Amorim[1], José A. M. Demattê[1]

[1]Department of Soil Science, "Luiz de Queiroz" College of Agriculture - University of São Paulo, Piracicaba, Brazil, borgesmelo@usp.br; miguel-palacio@usp.br; jeannovais@usp.br; brunobartch@usp.br; samuels.moraes@usp.br; narosin@usp.br; merylyn.accorsi@usp.br; jamdemat@usp.br

ABSTRACT

Soil organic carbon (SOC) modeling is essential for assessing soil fertility, soil health, and carbon sequestration potential. Visible and near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy have gained global popularity for SOC prediction as a cost-effective method in socioeconomic and environmental terms. This study explores the use of Vis-NIR and MIR spectroscopy to model organic matter (OM) content in soils from northeastern Brazil. The dataset included 629 samples from 7 of the 9 states in the region. We used the Cubist model to predict OM content, with R software aiding in statistical processing. Our results showed that the MIR technique performed better, with an R^2 of 0.78 for validation data and 0.92 for training data than Vis-NIR range. This suggests the potential of spectroscopy techniques to support or even replace conventional physical-chemical analysis methods, presenting a viable alternative, especially when comprehensive and optimized calibration databases are available.

Keywords **—** *Cross-validation, soil health, diffuse reflectance spectroscopy, organic matter.*

1. INTRODUCTION

Soil organic carbon (SOC) is important driver for maintaining soil quality and has a significant impact on the global carbon cycle. It directly affects the biological, chemical, and physical properties of soils and soil functions [1], which involves ecosystem services [2]. SOC content has been measured, traditionally, using dry combustion or chemical analysis. However, both methods have drawbacks, including the use of hazardous chemicals, high costs, and being time-consuming [3]. In comparison, visible, near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy offer a more efficient alternative, enabling the estimation of multiple soil attributes with a single measurement [4].

Studies using spectroscopy to evaluate soils in semi-arid regions worldwide are limited [5, 6], and research on tropical soils under various land uses is also scarce [7]. In particular, there is a significant gap in data on soils from Brazil's Northeast region, which features the country's highest climatic variability and a remarkable diversity of soil types[11]. Soil spectroscopy is an emerging tool that has potential to significantly reduce the cost and time involved in routine soil analyses [9]. Specifically, diffuse reflectance spectroscopy in combination with statistical modeling methods has been shown to provide good predictions of a number of soil properties [4, 10]. This study explores the use of Vis-NIR and MIR infrared spectroscopies data to predict SOC content in northeastern Brazil.

2. MATERIAL AND METHODS

2.1. Study site and sample processing

The study was carried out with soil samples from the northeast region of Brazil (Figure 1). A total of 629 soil samples from agricultural and native areas were selected for this study.

The Brazilian Northeast has a hot, humid tropical climate along the coast. The terrain is composed of plateaus and depressions, and its geology includes crystalline rocks and sedimentary formations. The soils are fertile in the coastal areas but poorer in the semi-arid inland. Vegetation ranges from the Atlantic Forest to the Caatinga, adapted to the dry hot climate. The economy is mainly based on agriculture, such as grains, cassava, sugarcane and fruit cultivation, as well as extensive livestock farming [11].

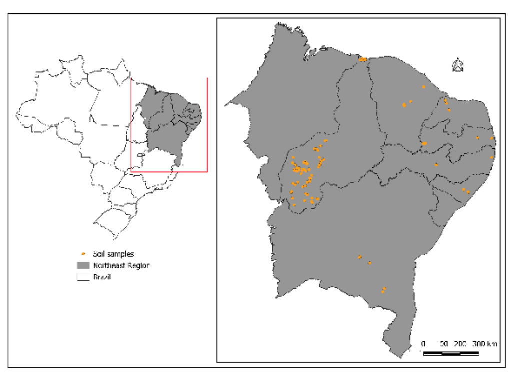

Figure 1 – Soil sampling locations in the northeast region of Brazil.

The selection process of soil samples involved the collection of points in surface (up to 20 cm) and subsurface (80 to 100 cm) horizons georeferenced using Global Positioning System (GPS) coordinates, mixed, and a representative sample was taken for laboratory OM and spectral analysis. All soil samples were oven-dried (45 °C) and sieved to pass a 2 mm mesh. Organic matter (OM) content was determined using the Walkley and Black method [12, 13].

2.2. Data analysis and spectroscopy measurements

MIR and Vis-NIR infrared spectra were obtained from soil aliquots sieved to 2 mm for Vis-NIR and 0.149 mm for MIR. MIR infrared spectra were captured using a Fourier-transform infrared (FTIR) spectrometer, scanning from 4000 cm^{-1} to 600 cm^{-1} at a resolution of 1 cm^{-1}, with each spectrum derived from an average of 32 scans. A background spectrum (from an empty sample compartment) was recorded as the mean of 32 scans before sample measurements. Vis-NIR infrared spectra were recorded in triplicate, with three separate subsamples collected from each soil sample, each placed in a glass Petri dish. Vis-NIR infrared spectral reflectance data were acquired with a FieldSpec Pro 3 spectroradiometer, using a 9 cm diameter sensor to form a 1.5 cm thick soil layer and capturing data over a wavelength range of 350 to 2500 nm with 1 nm intervals. The average of the three acquisitions was used for subsequent classification analysis [14].

2.3 Spectra preprocessing

Techniques of spectral preprocessing were evaluated using R software to enhance the quality of the data set. Root mean square error (RMSE) was used to quantify prediction accuracy. We also calculated coefficients of determination (R^2) values to compare our estimates with those found in other studies.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \qquad \text{Equation 1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \text{Equation 2}$$

Where, y_i are the observed values, $\hat{y}_i$ are the predicted values by the model, $\bar{y}_i$ is the mean of the observed values, and n is the total number of observations.

3. RESULTS

The results indicate that MIR spectroscopy technique showed the best performance, with an R^2 of 0.78 in the model Cubist with the total database, followed by the native area samples with an R^2 of 0.72.

Vis-NIR spectroscopy achieved a performance of 0.75, with agricultural areas reaching an R^2 of 0.71. The RMSE for the MIR technique was 5.72 $g.kg^{-1}$ in the general dataset, 5.48 $g.kg^{-1}$ for the agricultural area, and 7.46 $g.kg^{-1}$ for the native area. In the Vis-NIR processing, the RMSE model was 6.88, 4.88, and 8.74 $g.kg^{-1}$, respectively (Figure 2). Although our study presents an interesting sample set, some areas of the Northeast were not included. Adding samples from other specific regions and soil classes could strengthen our database and provide a more accurate assessment for the entire study area.

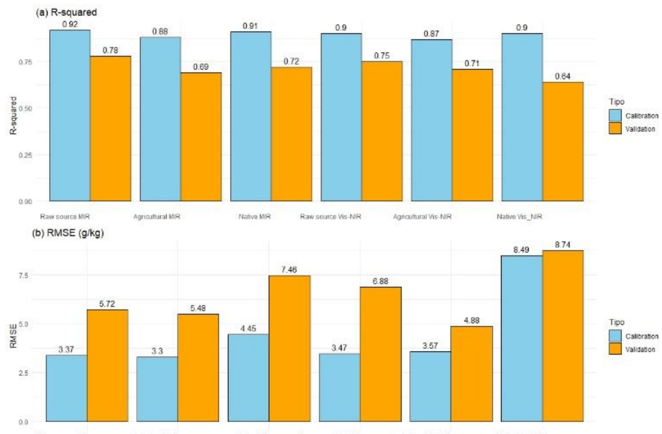

Figure 2 - Comparison of R² (a) and RMSE (Bb of Cubist performance for the raw database, agricultural and native area. The shaded areas in blue and orange indicate the best and worst statistics.

These sampling gaps, however, do not diminish the advantages of spectroscopy, which offers a simpler, cleaner, and more cost-effective alternative to the dry combustion method and is more accurate than the chemical methods currently used in the region (Figure 3).

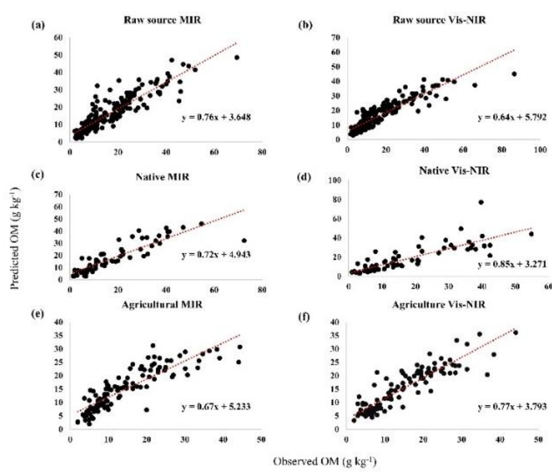

Figure 3 - Validation data for observed and predicted of OM using Cubist model in different ecosystems.

4. DISCUSSION

4.1. Implications of using soil spectroscopy for carbon study

Overall, better predictions of C concentration were achieved using the MIR region compared to the Vis-NIR. The lower predictive capabilities of Vis-NIR spectra are likely due to the region's low absorption, where overlapping spectral bands create a weak signal for chemical compounds [15]. In contrast, the MIR region captures stronger absorption signals from well-defined organic molecules, such as aliphatic hydrocarbons, aromatic compounds, amines, amides, and carboxylic acids, leading to enhanced predictive accuracy [11].

Spectroscopy applied to the study of soil carbon offers practical and strategic implications for agriculture and climate change mitigation. With techniques such as mid-infrared (MIR) and visible-near infrared (Vis-NIR), it is possible to quickly and cost-effectively estimate SOC content over extensive areas. This reduces the need for conventional physicochemical analyses, which are more expensive and time-consuming. The use of these spectral methods is especially advantageous in regions like northeastern Brazil, where soil type diversity and climatic variability are high, and traditional analysis costs can be prohibitive [1,6,10]. Thus, spectroscopy can provide a viable and accessible alternative, promoting continuous monitoring of soil organic matter and identifying more sustainable management practices.

4.2. Perspectives

Vis-NIR and MIR spectroscopy hold promising prospects for studying SOC, particularly in regions like northeastern Brazil, where conventional analysis methods may be less accessible [11]. These techniques enable large-scale, cost-effective approaches for modeling and continuously monitoring OM in agricultural and natural soils, which is essential for sustainable soil management. MIR spectroscopy, in particular, has demonstrated superior performance (with an R^2 of up to 0.92 in training data), indicating that, with robust spectral libraries and models such as Cubist model, it can provide precise and replicable SOC estimates. Integrating these tools into agricultural and environmental monitoring has the potential to accelerate the development of sustainable management practices and to expand understanding of soil carbon sequestration.

4.3. Limitations

Spectroscopy has limitations that must be considered. First, the method requires rigorous calibration and the availability of large spectral databases specific to the region and soil type to ensure accuracy [10,11]. This calibration process can be challenging, costly, and time-consuming, especially in areas with high soil type variability, such as northeastern Brazil. Additionally, spectral interferences and the method's sensitivity to environmental changes may affect estimate accuracy, necessitating continuous adjustments and model re-evaluations. Spectroscopy's limitation in distinguishing specific forms of C is another factor that may compromise analysis and result applicability.

4.4. Social and environmental impacts

Socially, adopting these techniques can democratize soil monitoring access, particularly in economically disadvantaged regions, such as certain areas of northeastern Brazil. With a more accessible method, farmers and environmental managers can monitor soil fertility and health, fostering agricultural practices that regenerate soil and maintain productivity [14]. Environmentally, spectroscopy enables more agile and detailed monitoring of organic carbon, which is crucial for developing carbon sequestration practices and mitigating climate change. By facilitating the adoption of more sustainable management, the technique also contributes to soil biodiversity conservation and ecosystem resilience.

5. CONCLUSION

We found strong relationships between MIR reflectance with an R^2 of 0.78 for validation data. The worst performance was observed for Vis-NIR spectroscopy techniques in native areas, with an RMSE of 8.74 $g.kg^{-1}$and an R^2 of 0.64. The recommended procedure is to obtain the spectra in the MIR spectral range, which provided better results than spectra in the Vis-NIR range.

The fact that information on fundamental soil properties, such as OM, can be obtained simultaneously with spectroscopy techniques offers a promising perspective for the use of sensors in many practical agronomic and environmental applications, including the development and monitoring of soil health indicators.

6. REFERENCES

[1] A. B. McBratney, U. Stockmann, D. Angers, B. Minasny, D. Field, 2014. Challenges for soil organic carbon research. In: Alfred, E., Hartemink, A.E., McSweeney, K. (Eds.), Soil Carbon. *Springer*, New York, pp. 3–16.

[2] U. Stockmann, J. Padarian, A. B. McBratney, B. Minasny, D. de Brogniez, L. Montanarella, S. Y. Hong, B. G. Rawlins, D. J. Field, 2015. Global soil organic carbon assessment. *Global Food Security*. 6, 9–16. http://dx.doi.org/10.1016/j.gfs.2015.07.001

[3] D. J. Brown, K. D. Shepherd, M. G. Walsh, M. Dewayne Mays, T. G. Reinsch, 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*. 132, 273–290.

[4] R. A. Viscarra Rossel, D. J. J. Walvoort, A. B. McBratney, L. J. Janik, J. O. Skjemstad, 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous

assessment of various soil properties. *Geoderma*. 131(1), 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007.

[5] Y. Ostovari, S. Ghorbani-Dashtakia, H.A. Bahramib, M. Abbasic, J.A.M. Demattê, E. Arthurd, P. Panagos, 2018. Towards prediction of soil erodibility, SOM and CaCO3 using laboratory VisNIR spectra: a case study in a semi-arid region of Iran. *Geoderma* 314, 102–112. https://doi.org/10.1016/j.geoderma.2017.11.014

[6] B. Wang, C. Waters, S. Orgill, A. Cowie, A. Clark, D. Li Liu, D T. Sides, 2018. Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. *Ecol. Indic*. 88, 425–438. https://doi.org/10.1016/j. ecolind.2018.01.049

[7] J.M. Moura-Bueno, R.S.D. Dalmolin, A. T. Caten, A.C. Dotto, J.A.M. Demattê, 2019. Of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions. *Geoderma* 337, 565–581. https://doi.org/10.1016/j. geoderma.2018.10.015

[8] V.D.P.R. Silva, 2004. On climate variability in Northeast of Brazil. *J. Arid Environ*. 58 (4), 575–596. https://doi.org/10.1016/j.jaridenv.2003.12.002

[9] J. Sanderman, K. Savage & S.R. Dangal (2020). Mid-infrared spectroscopy for prediction of soil health indicators in the United States. *Soil Science Society of America Journal*, 84(1), 251-261. DOI: 10.1002/saj2.20009.

[10] R.A. Viscarra Rossel, R. Webster, E.N. Bui, J.A. Baldock, 2014. Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. *Global Change Biology*, 20(9), 2953–2970. https://doi.org/10.1111/gcb.12569

[11] U.J. Santos, J.A.M. Demattê, R.S.C. Menezes, A.C. Dotto, C.C.B. Guimarães, B.J.R. Alves, D. C. Primo, E.V.S.B. Sampaio. Predicting carbon and nitrogen by visible near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy in soils of Northeast Brazil *Geoderma Regional*, Volume 23, 2020,e00333,ISSN 2352-0094,https://doi.org/10.1016/j.geodrs.2020.e00333.

[12] A. Walkley, I.A. Black, 1934. An estimation of the Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. *Soil Science* 37, 29–37.

[13] P. Teixeira, G. Donagemma, A. Fontana, W. Teixeira. Manual de métodos de análise de solo. 3. ed. rev. Brasília, DF: *Embrapa*, 2017. 574 p.

[14] E. Ben Dor, C. Ong, I. C. Lau, Reflectance measurements of soils in the laboratory: Standards and protocols, *Geoderma*, Volumes 245–246, 2015, Pages 112-124, ISSN 0016-061,https://doi.org/10.1016/j.geoderma.2015.01.002.

[15] B. Stenberg, R.A. Viscarra Rossel, A.M. Mouazen, J. Wetterlind, 2010. Visible and near infrared spectroscopy in soil science. *Adv. Agron*. 107, 163–215. https://doi.org/10.1016/ S0065-2113(10)07005-7.