ReGeo ISSN: 2177-3246

MODELOS PREDITIVOS DE QUANTIFICAÇÃO DE CARBONO ORGÂNICO POR ESPECTRORRADIOMETRIA EM SOLOS DA CAATINGA

PREDICTIVE MODELS FOR QUANTIFYING ORGANIC CARBON BY SPECTRORADIOMETRY IN CAATINGA SOILS

MODELOS PREDICTIVOS PARA LA CUANTIFICACIÓN DEL CARBONO ORGÁNICO MEDIANTE ESPECTRORRADIOMETRÍA EN SUELOS DE LA CAATINGA

https://doi.org/10.56238/revgeov17n11-155

Data de submissão: 24/10/2026 **Data de publicação:** 24/11/2026

Bruna Suellen Oliveira Mota
Mestra em Ciências Ambientais
Instituição: Universidade Estadual de Feira de Santana
E-mail: brunasuellen.om@gmail.com

Deorgia Tayane Mendes de Souza
Doutora em Geociências Aplicadas e Geodinâmica
Instituição: Universidade Estadual de Feira de Santana
E-mail: dtmsouza@uefs.br

Taciara Zborowski Horst
Doutora em Ciência do Solo
Instituição: Universidade Federal de Santa Maria
E-mail: tacihorst@gmail.com

Rodrigo Nogueira de Vasconcelos
Doutor em Ecologia
Instituição: Universidade Estadual de Feira de Santana
E-mail: rnv@uefs.br

Higo Batista Ferreira
Mestre em Ciências Ambientais
Instituição: Universidade Estadual de Feira de Santana
E-mail: higoferreira8@gmail.com

Lucas Teles de Oliveira
Engenheiro Agrônomo
Instituição: Universidade Estadual de Feira de Santana
E-mail: lucastelesdeoliveira3@gmail.com

Valdinéia Gusmão Silva
Mestra em Ciências Ambientais
Instituição: Universidade Estadual de Feira de Santana
E-mail: neiagusmaogeo@gmail.com

ReGeo

ISSN: 2177-3246

Washington de Jesus Sant'Anna da Franca Rocha
Doutor em Geologia
Instituição: Universidade Estadual de Feira de Santana
E-mail: wrocha@uefs.br

RESUMO
O carbono orgânico do solo (COS) é indicador primário de qualidade do solo e componente central do ciclo biogeoquímico do carbono, mas sua determinação por métodos laboratoriais convencionais é lenta e onerosa, o que limita o monitoramento em regiões extensas e pouco amostradas, como o semiárido brasileiro. A espectroscopia de reflectância difusa associada à quimiometria é uma alternativa rápida e de baixo custo, porém o desempenho dos modelos depende do pré-processamento espectral, do algoritmo de calibração e da representatividade amostral, aspectos ainda pouco avaliados em solos da Caatinga. Este trabalho testou modelos preditivos de COS em solos de áreas naturais e antropizadas (pastejo) da Caatinga com base em espectros Vis-NIR-SWIR. Foram utilizadas 52 amostras da camada de 0–5 cm, coletadas em 2013 pela Embrapa Semiárido em nove municípios de sete estados, lidas em laboratório com espectrorradiômetro FieldSpec 3 Hi-Res (350–2500 nm). Modelos de Regressão por Mínimos Quadrados Parciais (PLSR) e Random Forest (RF) foram calibrados com o espectro em reflectância (SPEC) e com seis pré-processamentos — absorbância (AB), remoção do contínuo (CRR), correção multiplicativa de espalhamento (MSC), primeira derivada de Savitzky-Golay (SG), suavização (SMO) e variação normal padrão (SNV) — e avaliados por validação cruzada leave-one-out (R^2, RMSE e RPIQ). Os teores de COS variaram de 0,34% a 4,64% (média de 2,04% nas áreas naturais e 1,17% nas antropizadas). O PLSR apresentou R^2 de 0,46 a 0,69, com melhor ajuste para SNV e MSC (R^2 = 0,69; RMSE = 0,56% e 0,57%; RPIQ = 2,38 e 2,37); o RF variou de 0,28 a 0,58, com melhor resultado para CRR (R^2 = 0,58; RMSE = 0,66%; RPIQ = 2,02). Os pré-processamentos de correção de espalhamento melhoraram o PLSR, enquanto o RF foi mais sensível à escolha da técnica e menos preciso. Ambos os algoritmos subestimaram os teores elevados de COS, predominantes nas áreas naturais, e superestimaram os teores baixos, comportamento compatível com a contração das predições em direção à média em calibrações com poucas amostras. O estudo demonstra o potencial da espectrorradiometria para estimar o COS em solos da Caatinga e indica o PLSR com SNV ou MSC como ponto de partida; o número reduzido de amostras, a ausência de conjunto de validação independente e a origem pretérita das amostras restringem a generalização e exigem bibliotecas espectrais regionais maiores.

Palavras-chave: Sensoriamento Proximal. Espectroscopia de Reflectância. Quimiometria. Pedometria. Semiárido.

ABSTRACT
Soil organic carbon (SOC) is a primary indicator of soil quality and a central component of the carbon biogeochemical cycle, but its determination by conventional laboratory methods is slow and costly, which limits monitoring in large and poorly sampled regions such as the Brazilian semiarid. Diffuse reflectance spectroscopy combined with chemometrics is a fast and low-cost alternative; however, model performance depends on spectral preprocessing, the calibration algorithm and sample representativeness, aspects that remain poorly evaluated in Caatinga soils. This study tested SOC predictive models for soils under natural and anthropized (grazed) Caatinga based on Vis-NIR-SWIR spectra. We used 52 samples from the 0–5 cm layer, collected in 2013 by Embrapa Semiárido in nine municipalities of seven Brazilian states and measured in the laboratory with a FieldSpec 3 Hi-Res spectroradiometer (350–2500 nm). Partial Least Squares Regression (PLSR) and Random Forest (RF) models were calibrated with the reflectance spectrum (SPEC) and six preprocessing techniques —

absorbance (AB), continuum removal (CRR), multiplicative scatter correction (MSC), Savitzky-Golay first derivative (SG), smoothing (SMO) and standard normal variate (SNV) — and assessed by leave-one-out cross-validation (R^2, RMSE and RPIQ). SOC contents ranged from 0.34% to 4.64% (mean of 2.04% in natural and 1.17% in anthropized areas). PLSR yielded R^2 from 0.46 to 0.69, with the best fit for SNV and MSC (R^2 = 0.69; RMSE = 0.56% and 0.57%; RPIQ = 2.38 and 2.37); RF ranged from 0.28 to 0.58, with the best result for CRR (R^2 = 0.58; RMSE = 0.66%; RPIQ = 2.02). Scatter-correction preprocessing improved PLSR, whereas RF was more sensitive to the choice of technique and less accurate. Both algorithms underestimated high SOC contents, which predominated in natural areas, and overestimated low contents, a behavior consistent with the shrinkage of predictions towards the mean in calibrations with few samples. The study demonstrates the potential of spectroradiometry to estimate SOC in Caatinga soils and points to PLSR with SNV or MSC as a starting point; the small number of samples, the lack of an independent validation set and the legacy origin of the samples limit generalization and call for larger regional spectral libraries.

Keywords: Proximal Sensing. Reflectance Spectroscopy. Chemometrics. Pedometrics. Semiarid.

RESUMEN

El carbono orgánico del suelo (COS) es un indicador primario de la calidad del suelo y un componente central del ciclo biogeoquímico del carbono, pero su determinación mediante métodos de laboratorio convencionales es lenta y costosa, lo que limita el monitoreo en regiones extensas y poco muestreadas, como el semiárido brasileño. La espectroscopía de reflectancia difusa asociada a la quimiometría es una alternativa rápida y de bajo costo; sin embargo, el desempeño de los modelos depende del preprocesamiento espectral, del algoritmo de calibración y de la representatividad de las muestras, aspectos aún poco evaluados en suelos de la Caatinga. Este trabajo evaluó modelos predictivos de COS en suelos de áreas naturales y antropizadas (pastoreo) de la Caatinga a partir de espectros Vis-NIR-SWIR. Se utilizaron 52 muestras de la capa de 0–5 cm, recolectadas en 2013 por Embrapa Semiárido en nueve municipios de siete estados brasileños y medidas en laboratorio con un espectrorradiómetro FieldSpec 3 Hi-Res (350–2500 nm). Se calibraron modelos de Regresión por Mínimos Cuadrados Parciales (PLSR) y Random Forest (RF) con el espectro de reflectancia (SPEC) y con seis preprocesamientos — absorbancia (AB), remoción del continuo (CRR), corrección multiplicativa de la dispersión (MSC), primera derivada de Savitzky-Golay (SG), suavizado (SMO) y variable normal estándar (SNV) — evaluados por validación cruzada leave-one-out (R^2, RMSE y RPIQ). Los contenidos de COS variaron de 0,34% a 4,64% (media de 2,04% en las áreas naturales y 1,17% en las antropizadas). El PLSR presentó R^2 de 0,46 a 0,69, con el mejor ajuste para SNV y MSC (R^2 = 0,69; RMSE = 0,56% y 0,57%; RPIQ = 2,38 y 2,37); el RF varió de 0,28 a 0,58, con el mejor resultado para CRR (R^2 = 0,58; RMSE = 0,66%; RPIQ = 2,02). Los preprocesamientos de corrección de la dispersión mejoraron el PLSR, mientras que el RF fue más sensible a la técnica elegida y menos preciso. Ambos algoritmos subestimaron los contenidos elevados de COS, predominantes en las áreas naturales, y sobrestimaron los contenidos bajos, comportamiento compatible con la contracción de las predicciones hacia la media en calibraciones con pocas muestras. El estudio demuestra el potencial de la espectrorradiometría para estimar el COS en suelos de la Caatinga e indica el PLSR con SNV o MSC como punto de partida; el número reducido de muestras, la ausencia de un conjunto de validación independiente y el origen pretérito de las muestras restringen la generalización y exigen bibliotecas espectrales regionales más amplias.

Palabras clave: Teledetección Proximal. Espectroscopía de Reflectancia. Quimiometría. Pedometría. Semiárido.

1 INTRODUÇÃO

O Carbono é um objeto de estudo cuja relevância abarca tanto a dinâmica climática do planeta quanto o solo, tendo em vista que os processos que compõem o Ciclo Biogeoquímico do Carbono são sistêmicos. Quando emitido para a atmosfera na forma de dióxido de carbono (CO_2), o carbono contribui para o efeito estufa; no solo, na forma de carbono orgânico do solo (COS), constitui um indicador primário de qualidade, associado à fertilidade, à estrutura e à retenção de água (Lal, 2004; Ribeiro et al., 2021). O solo armazena aproximadamente duas vezes mais carbono na fração orgânica do que a atmosfera, de modo que pequenas variações relativas nesse estoque têm efeito expressivo no balanço global de carbono (Scharlemann et al., 2014; Guo et al., 2019).

O carbono responde por cerca de 58% da matéria orgânica do solo (MOS) e influencia funções fundamentais para o sistema solo, desde a nutrição até a estruturação e a disponibilidade hídrica (Ker et al., 2012), com implicações para a segurança do solo, a produção de alimentos e o equilíbrio do ambiente (Lal, 2004). As características do solo bem como práticas de uso e manejo, condicionam o acúmulo de Carbono Orgânico que é lento e requer décadas. Todavia, sua liberação pode ocorrer de forma rápida e dificilmente é revertida em curto espaço de tempo (Lal, 2004). Acompanhar essas mudanças exige grande número de determinações de COS no espaço e no tempo, o que é dificultado pelos métodos de referência, tanto por combustão seca quanto por oxidação úmida, que são laboriosos, consomem reagentes e encarecem o monitoramento (Nocita et al., 2015; Das et al., 2023).

A espectroscopia de reflectância difusa na faixa do visível, infravermelho próximo e infravermelho de ondas curtas (Vis-NIR-SWIR) tornou-se a principal alternativa para essa demanda, pois possibilita a obtenção de inúmeras informações com uma única varredura, com pouco ou nenhum resíduo químico, baixo custo operacional e precisão compatível com muitas aplicações, em comparação aos métodos laboratoriais tradicionais (Viscarra Rossel et al., 2006; Stenberg et al., 2010; Nocita et al., 2015). Bibliotecas espectrais globais e nacionais, como a biblioteca espectral mundial (Viscarra Rossel et al., 2016) e a Biblioteca Espectral de Solos do Brasil (Demattê et al., 2019), consolidaram a técnica, e estudos em diferentes contextos mostram que o COS pode ser estimado com acurácia moderada a alta a partir de espectros de laboratório e de campo (Gomez; Viscarra Rossel; McBratney, 2008; Ji; Viscarra Rossel; Shi, 2015; Viscarra Rossel et al., 2022).

O desempenho desses modelos, contudo, não é garantido. Os espectros de solo são pouco específicos e resultam da sobreposição de sinais da matéria orgânica, dos minerais de argila, dos óxidos de ferro, da água e da granulometria, o que exige pré-processamentos espectrais e algoritmos de calibração multivariada capazes de extrair a informação relevante (Stenberg et al., 2010; Viscarra Rossel; Behrens, 2010). Não há consenso sobre a melhor combinação entre pré-processamento e

algoritmo: a eficiência relativa de técnicas como a variação normal padrão, a correção multiplicativa de espalhamento e as derivadas depende do conjunto de dados (Rinnan; Berg; Engelsen, 2009; Dotto et al., 2018), e algoritmos de aprendizado de máquina, como o Random Forest, nem sempre superam a regressão por mínimos quadrados parciais, sobretudo quando o número de amostras é pequeno (Kuang; Mouazen, 2012; Moura-Bueno et al., 2020; Wadoux; Minasny; McBratney, 2020).

Apesar de os estudos relativos ao COS apresentarem um crescimento considerável durante os últimos anos no Brasil, as áreas semiáridas possuem defasagem de pesquisas nesse campo, com poucas calibrações espectrais desenvolvidas para seus solos (Ribeiro et al., 2021; Mendes et al., 2021). Na Caatinga, os teores de COS são em geral baixos, os solos variam de rasos e pedregosos sobre o embasamento cristalino a arenosos e profundos sobre bacias sedimentares, e a conversão da vegetação nativa em pastagens e cultivos reduz os estoques de carbono (Giongo et al., 2011; Schulz et al., 2016; Menezes et al., 2021). Essa combinação de baixa amplitude de teores, diversidade mineralógica e alteração antrópica torna incerto o desempenho de modelos calibrados em outros biomas e justifica avaliar, com amostras da própria Caatinga, quais combinações de pré-processamento e algoritmo são mais adequadas.

Neste estudo, amostras superficiais de solos de áreas naturais e antropizadas, distribuídas em nove municípios do bioma, foram utilizadas para comparar dois algoritmos de natureza distinta — um linear baseado em variáveis latentes (PLSR) e um não linear baseado em conjuntos de árvores de decisão (RF) — aplicados ao espectro original e a seis pré-processamentos. A contribuição esperada é metodológica e regional: indicar a combinação mais promissora para a predição de COS em solos da Caatinga, identificar padrões sistemáticos de erro e apontar os requisitos amostrais para futuras bibliotecas espectrais do semiárido.

Desse modo, objetivou-se testar métodos preditivos de quantificação de COS em áreas naturais e antropizadas da Caatinga com base em dados espectrais Vis-NIR-SWIR, utilizados como preditores em modelos PLSR e RF associados a diferentes pré-tratamentos espectrais. De forma específica, buscou-se:

1.1 OBJETIVOS ESPECÍFICOS

a) caracterizar a variabilidade dos teores de COS e da granulometria das amostras de solos naturais e antropizados nos nove municípios amostrados;

b) calibrar modelos PLSR e RF com o espectro em reflectância e com seis pré-processamentos espectrais, avaliando-os por validação cruzada leave-one-out com as métricas R^2, RMSE e RPIQ;

c) comparar o desempenho dos dois algoritmos e dos pré-processamentos, identificando a combinação de melhor ajuste e a sensibilidade de cada algoritmo à técnica empregada;

d) analisar os resíduos das predições por amostra, por nível de COS e por condição de uso, identificando padrões sistemáticos de subestimação e superestimação e seus possíveis mecanismos.

2 REFERENCIAL TEÓRICO

2.1 CARBONO ORGÂNICO DO SOLO NO SEMIÁRIDO

A capacidade do solo de incorporar carbono é condicionada por uma série de variáveis, como por exemplo, clima, tipo de solo (considerando suas propriedades, dentre elas mineralogia e textura), tipo de vegetação ou manejo. Nas florestas secas e savanas semiáridas, a baixa e irregular precipitação limita a produção primária e, portanto, o aporte de resíduos ao solo, enquanto as temperaturas elevadas aceleram a decomposição da matéria orgânica; o resultado são estoques de carbono pequenos e de rápida ciclagem, sensíveis a perturbações (Tiessen et al., 1998; Menezes et al., 2012).

Para Giongo et al. (2011), a característica de semiaridez em conjunto com os demais fatores paisagísticos limita as taxas de produção e estoque de biomassa disponíveis, visto que o clima quente intensifica a mineralização da MOS, em contraste a algumas outras paisagens tropicais úmidas e temperadas, que somado ainda ao fator de forte degradação associado ao bioma, caracterizam os baixos teores de COS em maior parte da Caatinga. Assim, tais áreas tendem a incorporação lenta, em contrapartida à liberação rápida. Os mesmos autores relatam estoques de COS entre 20 e 48 t ha^{-1} na camada de 0–20 cm no semiárido brasileiro, com redução sistemática após mudança de uso da terra.

Além das condições do ambiente, Dias et al. (2017) retratam outros fatores que desfavorecem a incorporação de carbono, como por exemplo, o revolvimento do solo, que diminui a proteção dos agregados e aumenta a oxidação da matéria orgânica, e os teores elevados da fração areia, que aumentam a velocidade da mineralização. Nesse sentido, o manejo possui consequências variáveis em virtude da forma a qual é realizado, podendo gerar tanto contribuições a incorporação, quanto promover/ acelerar sua liberação.

A mudança de uso da terra também é um fator de influência para o armazenamento do carbono. Na Caatinga, o pastejo contínuo reduz a cobertura do solo e o aporte de serapilheira e diminui os estoques de COS em relação às áreas de vegetação preservada (Schulz et al., 2016), e a conversão da floresta seca em pastagens e cultivos reduz os estoques de carbono do solo e da vegetação (Menezes et al., 2021; Gava et al., 2022). Por outro lado, sistemas agroflorestais e práticas conservacionistas podem manter ou recuperar parte desses estoques (Sacramento et al., 2013). Portanto, é necessário

levar em consideração tais fatores, visto que áreas antrópicas também podem promover condições favoráveis a uma menor taxa de liberação de carbono.

2.2 FUNDAMENTOS DA ESPECTROSCOPIA DE REFLECTÂNCIA APLICADA AO COS

Os materiais possuem diferentes morfologias e composições o que possibilita a produção de feições e comportamentos espectrais característicos ao interagir com a energia eletromagnética incidente (Dalmolin et al., 2005). No visível e no início do infravermelho próximo (aproximadamente 350–1100 nm), predominam os processos eletrônicos, como as transições de campo cristalino e de transferência de carga dos íons de ferro, responsáveis pelas feições dos óxidos de ferro; no infravermelho próximo e de ondas curtas (aproximadamente 1100–2500 nm), predominam os processos vibracionais, representados por sobretons e combinações das ligações O–H, C–H, N–H e metal–OH, cujas vibrações fundamentais ocorrem no infravermelho médio (Hunt, 1977; Clark; Roush, 1984; Meneses; Madeira Netto, 2001).

A matéria orgânica atua no espectro do solo de duas formas: reduz a reflectância em toda a faixa, pelo escurecimento do solo, e produz feições de absorção associadas às ligações C–H, N–H e O–H, mais evidentes no SWIR (Ben-Dor, 1997; Stenberg et al., 2010). Entretanto, os dados espectrais são considerados inespecíficos e passíveis de sobreposição de informações para a análise de COS (Viscarra Rossel; Behrens, 2010), tendo em vista que o solo é um sistema complexo e compreende grande diversidade de constituintes, que influenciam diretamente sua assinatura espectral. As feições das assinaturas ocorrem em razão dos atributos presentes, tais como conteúdo de matéria orgânica, minerais associados, umidade e granulometria (Meneses; Almeida; Baptista, 2019), e o efeito de óxidos de ferro, umidade e textura sobre a predição de COS já foi demonstrado experimentalmente (Kuang; Mouazen, 2013; Aïchi et al., 2020). Desse modo, a utilização da quimiometria aliada ao Sensoriamento Proximal para análises quantitativas de propriedades do solo, possui inúmeros benefícios que tendem a otimizar as análises de modo a obter informações precisas, haja vista essa complexidade.

Durante as últimas décadas, houve um aumento significativo de trabalhos que aliam quimiometria e dados espectrais ao estudo do COS. Esses estudos abrangem desde frações do carbono orgânico (Cozzolino; Morón, 2006) e correção de espectros de campo (Gomez; Viscarra Rossel; McBratney, 2008; Ji; Viscarra Rossel; Shi, 2015) até a comparação entre algoritmos lineares e redes neurais (Souza; Madari; Guimarães, 2012; Tekin; Tümsavas; Mouazen, 2014), a fusão de sensores (Shi et al., 2023) e a estimativa da relação C:N em perfis florestais (Gholizadeh et al., 2023). No Brasil, calibrações foram desenvolvidas por Pinheiro et al. (2017) para solos da Amazônia central, por Mendes

et al. (2021) para solos piauienses e por Ribeiro et al. (2021) para o semiárido cearense, e bibliotecas tropicais mostraram que o desempenho depende da faixa espectral e da heterogeneidade das amostras (Terra; Demattê; Viscarra Rossel, 2015; Araújo et al., 2014). Parte desses autores destaca o visível e o infravermelho próximo como faixas informativas para o carbono, embora as feições orgânicas mais específicas se concentrem no SWIR (Stenberg et al., 2010).

2.3 PRÉ-PROCESSAMENTOS ESPECTRAIS E CALIBRAÇÃO MULTIVARIADA

As técnicas de pré-processamento espectral compreendem operações matemáticas realizadas sobre os espectros "brutos" antes da etapa de calibração (Viscarra Rossel et al., 2006; Dotto et al., 2018). Essas alterações nas medidas espectrais têm por objetivo realizar correções nos dados, visto que o processo de leitura se encontra suscetível a distorções como ruído de dispersão, falhas de medição, resolução do equipamento e até em razão da complexidade da composição do solo, que responde simultaneamente na assinatura. Dotto et al. (2018) classificam essas técnicas em dois grupos principais — correção de espalhamento (por exemplo, MSC e SNV) e derivadas espectrais (por exemplo, Savitzky-Golay) —, aos quais se somam transformações como a absorbância e a remoção do contínuo (Clark; Roush, 1984; Rinnan; Berg; Engelsen, 2009).

A calibração é um método quimiométrico considerado um dos mais utilizados em pesquisas referentes ao solo, que diz respeito à construção de equações matemáticas que relacionam os espectros a um atributo medido em um dado conjunto de amostras (Souza; Madari; Guimarães, 2012), visto que possibilita a análise correlacionada de múltiplas variáveis. A Regressão por Mínimos Quadrados Parciais (PLSR) projeta os espectros em poucas variáveis latentes que maximizam a covariância com o atributo, o que contorna a colinearidade entre comprimentos de onda (Wold; Sjöström; Eriksson, 2001). O Random Forest (RF), por sua vez, combina predições de muitas árvores de decisão construídas com subamostras das observações e das variáveis, capturando relações não lineares e interações (Breiman, 2001). Comparações em bibliotecas de solos mostram que os algoritmos de aprendizado de máquina podem superar o PLSR em conjuntos grandes e heterogêneos, mas perdem vantagem — e interpretabilidade — em conjuntos pequenos (Viscarra Rossel; Behrens, 2010; Araújo et al., 2014; Wadoux; Minasny; McBratney, 2020).

Portanto, o desenvolvimento de modelos preditivos por meio de dados espectrais é importante para estudos pedológicos, visto que possibilita inúmeras aplicações e inferências que abrangem tanto a quantificação, quanto a compreensão de fenômenos e processos.

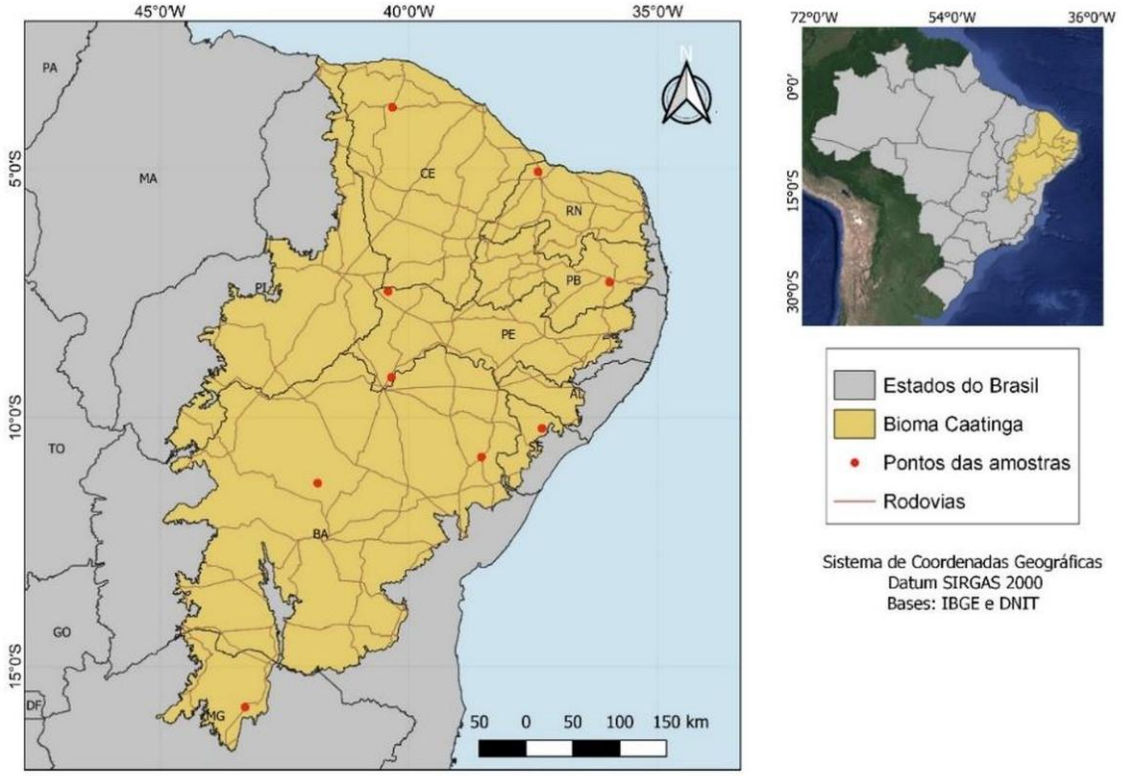
3 METODOLOGIA

3.1 ÁREA DE ESTUDO

A área de estudo em questão abrange o Bioma Caatinga, que possui uma extensão territorial de aproximadamente 862.818 km² (IBGE, 2019), o que corresponde a cerca de 10% do território nacional, e abrange estados da região Nordeste e o norte de Minas Gerais, na região Sudeste. Assim, em razão da vasta área que recobre, possui variadas paisagens e características geoambientais. Conforme a classificação de Köppen, o clima predominante é o BSh (semiárido quente), com ocorrência, em menor proporção, do tipo As (tropical com estação seca no verão) (Alvares et al., 2013).

Os pontos de coleta de solo distribuem-se em nove municípios de sete estados (Figura 1), o que consiste em um importante fator para a representatividade amostral do estudo, pois abrange diferentes contextos geológicos, climáticos e de uso da terra do bioma.

Figura 1 – Localização dos nove municípios de coleta das amostras de solo (pontos vermelhos) no Bioma Caatinga, com a malha rodoviária e o limite dos estados; no detalhe, a posição do bioma no Brasil. Sistema de coordenadas geográficas, Datum SIRGAS 2000.



Fonte: Elaborada pelos autores (2023), com bases do IBGE e do DNIT.

Aproximadamente 70% dos solos do Bioma possuem origem em substrato rochoso do embasamento cristalino, tendo como principal característica pedogenética serem rasos e pedregosos.

Dos 30% restantes, boa parte é proveniente de bacias sedimentares, sendo arenosos e mais profundos (Silva; Araújo; Ferraz, 2009), o que explica a grande diversidade de classes de solo amostradas, apresentadas na Tabela 1 segundo o Sistema Brasileiro de Classificação de Solos (Santos et al., 2018) e a World Reference Base (IUSS Working Group WRB, 2015).

Tabela 1 – Identificação das classes de solo das áreas amostradas, por município.

ID	Município-UF	Classe de solo (SiBCS)	Grupo de referência (WRB)
1	Petrolina-PE	ARGISSOLO VERMELHO-AMARELO Distrófico	Acrisols
2	Araripina-PE	LATOSSOLO AMARELO Distrófico psamítico	Ferralsols
3	Sobral-CE	LUVISSOLO CRÔMICO Órtico típico	Luvisols
4	Mossoró-RN	LATOSSOLO VERMELHO Distrófico típico	Ferralsols
5	Campina Grande-PB	NEOSSOLO LITÓLICO Distrófico típico	Leptosols
6	Feira Nova-SE	LUVISSOLO HÁPLICO Órtico típico	Luvisols
7	Ribeira do Pombal-BA	ARGISSOLO AMARELO Distrófico abrúptico	Acrisols
8	Irecê-BA	CAMBISSOLO HÁPLICO Ta Eutrófico típico	Cambisols
9	Nova Porteirinha-MG	LATOSSOLO VERMELHO Eutrófico típico	Ferralsols

Fonte: Embrapa Semiárido; classificação SiBCS (Santos et al., 2018) e WRB (IUSS, 2015). ID = primeiro dígito do código das amostras.

De acordo com o relatório do MapBiomas (2023), a Caatinga foi o terceiro bioma com maior área desmatada em 2022, precedido apenas pela Amazônia e pelo Cerrado. Dentre as principais atividades responsáveis pela degradação do bioma por meio da remoção de vegetação nativa, destaca-se a agricultura, pecuária e queimadas.

3.2 BASE DE DADOS E AMOSTRAGEM

As amostras utilizadas na pesquisa foram disponibilizadas pela Embrapa Semiárido, em Petrolina-PE. O conjunto amostral compreende, em cada município, três trincheiras em áreas de Caatinga natural e três trincheiras em áreas de Caatinga antropizada (áreas de pastejo), com exceção de um município que possui duas amostras para cada condição, totalizando 52 amostras (26 naturais e 26 antropizadas). Todas foram coletadas em 2013, na profundidade de 0–5 cm. O texto original atribui a exceção a Ribeira do Pombal; contudo, na base de dados das amostras, o município com quatro amostras é o de código 4 (Mossoró-RN), o que deve ser confirmado.

O código das amostras tem três dígitos: o primeiro refere-se ao município (Tabela 1), o segundo à condição da área (4 = antropizada; 5 = natural) e o terceiro à trincheira de coleta; assim, a amostra 853 corresponde à terceira trincheira da área natural de Irecê-BA. Os teores de COS (%) e as frações

granulométricas (g kg^{-1}) de referência foram fornecidos com as amostras; os métodos analíticos empregados na determinação do COS e da granulometria não foram informados pelos autores. Os dados de COS e granulometria correspondem, portanto, à condição dos solos em 2013, e não à situação atual das áreas.

3.3 LEITURAS ESPECTRAIS

Para obtenção das curvas espectrais, as amostras de solo foram submetidas previamente ao processo de pré-tratamento, conforme o protocolo de Ben-Dor, Ong e Lau (2015): foram peneiradas em malha de 2 mm e secas em estufa a 45 °C durante 24 horas, para que a presença de água não interferisse nos resultados e para uniformizar a granulometria.

As leituras foram realizadas pelo Laboratório de Sensoriamento Remoto e Espectrorradiometria (LABESPECTRO) do Programa de Pós-Graduação em Modelagem em Ciências da Terra e do Ambiente (PPGM) da Universidade Estadual de Feira de Santana. Foi utilizado o espectrorradiômetro FieldSpec® 3 Hi-Res, sensor hiperespectral que opera no intervalo de 350 nm a 2500 nm, abrangendo o visível, o infravermelho próximo e o infravermelho de ondas curtas (Vis-NIR-SWIR), com resolução espectral de 3 nm no VNIR e de 8 nm no SWIR. A iluminação foi fornecida por fonte artificial própria (lâmpada halógena), o que torna a medição independente da iluminação solar. A geometria de aquisição (sonda de contato ou iluminação externa), o número de leituras por amostra e o procedimento de calibração com painel de referência não foram informados.

Após a realização das leituras espectrais, as assinaturas foram exportadas para o *Software* ENVI 5.3, para criação das bibliotecas espectrais e, em seguida, foram realizadas algumas análises prévias. Para compreensão e inferência de atributos por meio das curvas espectrais é necessário considerar padrões de referência de reflectância/ absorção, ainda morfologia da curva, profundidade, e faixa de ocorrência. Visando caracterizar a mineralogia associada foi utilizada a biblioteca de referência da United States Geological Survey (USGS) para fins comparativos.

3.4 PRÉ-PROCESSAMENTOS ESPECTRAIS

A utilização da técnica adequada pode influenciar diretamente o resultado do modelo e existe na literatura uma ampla disponibilidade de opções. No *software* R (ambiente RStudio), além dos espectros em reflectância sem tratamento (SPEC), foram aplicadas seis técnicas de pré-processamento: suavização (SMO); primeira derivada de Savitzky-Golay (SG) (Savitzky; Golay, 1964); remoção do contínuo (CRR) (Clark; Roush, 1984); transformação em absorbância (AB), calculada como $\log_{10}(1/R)$, em que R é a reflectância; variação normal padrão (SNV) (Barnes; Dhanoa; Lister, 1989);

e correção multiplicativa de espalhamento (MSC) (Isaksson; Næs, 1988; Helland; Næs; Isaksson, 1995). A SNV centraliza e escalona cada espectro por sua própria média e desvio-padrão, enquanto a MSC regride cada espectro contra o espectro médio; ambas corrigem efeitos multiplicativos e aditivos do espalhamento da luz causados por diferenças de tamanho de partícula e empacotamento (Rinnan; Berg; Engelsen, 2009). O tamanho da janela e a ordem do polinômio da suavização e da derivada de Savitzky-Golay não foram informados.

3.5 MODELOS PREDITIVOS E VALIDAÇÃO

O PLSR é o método de regressão utilizado para modelar uma variável resposta, com base em um grande número de variáveis preditoras, nesse caso, utilizando espectros completos de 350 nm a 2500 nm, pois lida bem com a correlação entre preditores e reduz a dimensão dos dados a poucas variáveis latentes (Wold; Sjöström; Eriksson, 2001). É o método mais utilizado em espectroscopia de solos, com resultados satisfatórios na maioria das pesquisas em que foi aplicado (Stenberg et al., 2010). O número ótimo de variáveis latentes foi selecionado automaticamente pelo menor RMSE de validação cruzada, por meio do parâmetro *bestTune*, que corresponde ao objeto de saída da função de ajuste do pacote caret (Kuhn, 2008); o pacote empregado não foi declarado pelos autores.

RF consiste em algoritmo de aprendizagem de máquina, utilizado tanto para fins de previsão quanto para classificação e baseia-se em árvores de decisão de modo que é amplamente aplicado a áreas que envolvem mineração de dados. Possui as seguintes variáveis como parâmetros de ajuste: *mtry* (quantidade de variáveis que serão amostradas de modo aleatório), número de árvores e tamanho do nó (Breiman, 2001). Os valores adotados para esses hiperparâmetros não foram informados.

Todas as 52 amostras foram utilizadas na calibração de ambos os modelos, avaliados pela validação cruzada leave-one-out (LOOCV): a cada iteração, uma amostra é retirada, o modelo é ajustado com as demais e a amostra retirada é predita, até que todas tenham sido preditas uma vez (Cozzolino; Morón, 2006). Dentre as vantagens de aplicação, considera-se que tende a não superestimar o erro de teste do modelo e favorece conjuntos amostrais menores (Zeraatpisheh et al., 2019). Não houve conjunto de teste independente, de modo que as métricas apresentadas correspondem ao desempenho em validação cruzada.

O potencial de predição de cada modelo foi avaliado pelo coeficiente de determinação (R^2) (Equação 1), pela raiz do erro quadrático médio (RMSE) (Equação 2) e pela razão de desempenho do intervalo interquartil (RPIQ) (Equação 3), calculados com as predições da validação cruzada.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}i - yi)^2}{\sum_{i=1}^{n}(yi - \bar{y}i)^2} \quad (1)$$

em que R^2 é o coeficiente de determinação (adimensional); y_i é o teor de COS observado (de referência) na amostra *i*, em %; $\hat{y}_i$ é o teor de COS predito pelo modelo para a amostra *i* na validação cruzada, em %; $\bar{y}$ é a média dos valores observados, em %; e *n* é o número de amostras (n = 52). O R^2 expressa a fração da variância dos teores observados explicada pelas predições e é o indicador mais empregado em espectroscopia de solos (Bellon-Maurel et al., 2010).

$$RMSE = \sqrt[2]{\frac{1}{N}\sum_{i=1}^{n}(\hat{y}i - yi)^2} \quad (2)$$

em que RMSE é a raiz do erro quadrático médio, expressa na mesma unidade do COS (%); *N* é o número de amostras (igual a *n* na Equação 1, n = 52); e os demais termos são definidos como na Equação 1. O RMSE mede a magnitude média dos erros de predição, penalizando mais os erros grandes.

$$RPIQ = \frac{q3 - q1}{RMSE} \quad (3)$$

em que RPIQ é a razão de desempenho do intervalo interquartil (adimensional); q_1 e q_3 são, respectivamente, o primeiro quartil (25%) e o terceiro quartil (75%) dos teores de COS observados, em %; $q_3 - q_1$ é a distância interquartil; e RMSE é o definido na Equação 2. O RPIQ foi proposto por Bellon-Maurel et al. (2010) como alternativa à razão de desempenho do desvio (RPD), por ser mais adequado a variáveis de distribuição assimétrica, como o COS; quanto maior o RPIQ, menor o erro em relação à amplitude central dos dados. Os resíduos foram calculados como observado menos predito ($y_i - \hat{y}_i$): resíduos positivos indicam subestimação e negativos, superestimação.

4 RESULTADOS E DISCUSSÃO

4.1 VARIABILIDADE DOS TEORES DE COS E DA GRANULOMETRIA

Tendo em vista a área de estudo em análise, é importante considerar os aspectos fisiográficos característicos do Bioma, que além de possuir uma grande extensão territorial, apresenta clima, em geral, quente com índices pluviométricos baixos/irregulares, que influenciam diretamente nas características físicas, químicas e biológicas do solo, possuindo processos de intemperismo específicos. Essas condições se refletem na amplitude dos teores de COS das 52 amostras, que variaram de 0,34% a 4,64%, com mediana de 1,21% e distribuição assimétrica à direita: apenas dez amostras ultrapassaram 2,5% e dezesseis ficaram abaixo de 1,0% (Tabela 2).

Tabela 2 – Estatística descritiva dos teores de carbono orgânico do solo (COS, %) e mediana das frações granulométricas (g kg^{-1}) das amostras de 0–5 cm, por condição de uso.

Condição	n	Mín.	Q1	Mediana	Média	Q3	Máx.	DP	AIQ	Argila (med.)	Areia (med.)	Silte (med.)
Natural	26	0,47	1,06	1,92	2,04	2,60	4,64	1,13	1,54	137	675	242
Antropizada	26	0,34	0,74	1,08	1,17	1,40	3,51	0,70	0,65	89	635	306
Total	52	0,34	0,84	1,21	1,61	2,19	4,64	1,03	1,34	120	659	261

Fonte: Elaborada a partir da base de dados das amostras (coletadas em 2013). Q1 e Q3 = primeiro e terceiro quartis; DP = desvio-padrão; AIQ = amplitude interquartil (Q3 − Q1). As medianas granulométricas excluem a amostra 551, cuja soma das frações (351 g kg^{-1}) indica erro de registro.

As áreas naturais apresentaram teores médios de COS (2,04%) superiores aos das antropizadas (1,17%), e a média natural foi igual ou maior que a antropizada em todos os nove municípios, com diferenças pequenas em Petrolina-PE e Nova Porteirinha-MG e maiores em Irecê-BA e Feira Nova-SE (Tabela 3). Houve, contudo, amostras antropizadas com teores elevados, como 943 (3,51%) e 342 (2,51%), e amostras naturais com teores baixos, como 253 (0,47%), o que mostra que a condição de uso não explica sozinha a variação do COS. Esse padrão é coerente com a redução dos estoques de carbono após a substituição da vegetação nativa por pastagens relatada para a Caatinga (Schulz et al., 2016; Menezes et al., 2021) e com a influência do manejo sobre a magnitude dessa perda (Sacramento et al., 2013; Gava et al., 2022). Como as amostras não foram desenhadas para testar o efeito do uso da terra e não houve teste estatístico, essas diferenças devem ser lidas como descritivas.

Tabela 3 – Teores de COS (%) por município e condição de uso, granulometria média e resíduo médio (observado − predito, %) dos modelos de melhor ajuste.

Município	n (nat. + antr.)	COS natural: média (mín.–máx.)	COS antropizada: média (mín.–máx.)	Argila (g kg^{-1})	Areia (g kg^{-1})	Resíduo PLSR-SNV (nat. / antr.)	Resíduo RF-CRR (nat. / antr.)
Petrolina-PE	3 + 3	0,90 (0,68–1,04)	0,84 (0,63–1,18)	40	661	0,33 / −0,17	−0,16 / −0,24
Araripina-PE	3 + 3	1,68 (0,47–2,44)	0,90 (0,83–1,01)	173	758	0,45 / 0,05	0,32 / 0,05
Sobral-CE	3 + 3	2,57 (1,70–3,23)	1,89 (1,24–2,51)	142	556	0,76 / 0,06	0,34 / 0,00
Mossoró-RN	2 + 2	1,24 (1,05–1,43)	0,51 (0,34–0,68)	100	795	−0,42 / −0,22	−0,93 / −0,32
Campina Grande-PB	3 + 3	1,78 (1,16–2,54)	0,72 (0,53–0,82)	39	766	0,32 / −0,55	0,17 / −0,23
Feira Nova-SE	3 + 3	2,58 (2,39–2,89)	1,12 (1,08–1,15)	249	265	0,10 / −0,67	0,72 / 0,14
Ribeira do Pombal-BA	3 + 3	0,91 (0,80–1,08)	0,62 (0,38–1,07)	23	922	0,12 / −0,07	−0,33 / −0,97
Irecê-BA	3 + 3	4,30 (4,03–4,64)	1,65 (1,44–1,78)	349	200	0,81 / −0,99	1,04 / −0,84
Nova Porteirinha-MG	3 + 3	2,15 (1,88–2,62)	2,09 (1,35–3,51)	142	444	0,27 / −0,28	0,39 / 0,47

Fonte: Elaborada a partir da base de dados das amostras. Argila e areia: médias das seis (ou quatro) amostras do município; em Campina Grande-PB, a amostra 551 foi excluída do cálculo da granulometria. Resíduo positivo = subestimação; negativo = superestimação.

O COS correlacionou-se negativamente com o teor de areia (r = −0,70) e positivamente com os teores de silte (r = 0,55) e argila (r = 0,51), valores calculados nesta revisão a partir da base de dados das amostras (n = 51). Os maiores teores ocorreram nos solos mais finos de Irecê-BA (Cambissolo eutrófico) e Feira Nova-SE (Luvissolo), e os menores, nos solos arenosos de Ribeira do Pombal-BA,

Petrolina-PE e Mossoró-RN, o que é consistente com a proteção física e química da matéria orgânica pelas frações finas e com a mineralização mais rápida em solos arenosos (Dias et al., 2017).

Com base na análise das curvas espectrais, os autores inferiram cinco tipos de minerais associados — montmorilonita, ilita, caulinita, goethita e hematita —, a partir das absorções típicas, da morfologia, da profundidade e da assimetria das feições, por comparação com a biblioteca da USGS; as curvas e as feições diagnósticas não foram apresentadas, de modo que essa inferência é qualitativa. As características da área estudada de irregularidade pluviométrica associada a demais fatores fitofisiográficos, condicionam o baixo a médio grau de intemperismo com predominância do tipo físico, resultando na ocorrência frequente de solos rasos constituídos, em sua maior parte, por minerais primários, no qual é comum a forte presença de características associadas ao material de origem. Tais fatores justificam tanto as classes de solo encontradas, quanto a mineralogia inferida a partir da análise espectral, em conformidade com Meneses, Almeida e Baptista (2019), que discutem a relação entre o domínio climático e as características pedogenéticas e espectrais dos solos. A coexistência de argilominerais 2:1 e 1:1 e de óxidos de ferro com feições no Vis-NIR é relevante para a modelagem, pois essas feições se sobrepõem ao sinal da matéria orgânica (Stenberg et al., 2010; Aïchi et al., 2020).

4.2 DESEMPENHO DOS MODELOS PLSR E RF

Todas as amostras foram utilizadas para calibrar ambos os modelos, PLSR e RF, aplicados ao espectro em reflectância (sem pré-processamento) e aos seis pré-processamentos, o que totaliza sete modelos para cada algoritmo. No PLSR, o número de variáveis latentes selecionado variou de 2 a 5, conforme o pré-tratamento; o mesmo número foi observado para SG, CRR e SMO. Os valores por modelo não foram apresentados pelos autores. As estatísticas de validação cruzada estão reunidas na Tabela 4.

Tabela 4 – Estatísticas de validação cruzada leave-one-out dos modelos de predição de COS por PLSR e RF, para o espectro em reflectância (SPEC) e os pré-processamentos AB, CRR, MSC, SG, SMO e SNV (n = 52).

Pré-processamento	PLSR R^2	PLSR RMSE (%)	PLSR RPIQ	RF R^2	RF RMSE (%)	RF RPIQ
SPEC	0,46	0,75	1,79	0,28	0,87	1,55
SG	0,46	0,75	1,79	0,52	0,73	1,84
MSC	0,69	0,57	2,37	0,49	0,73	1,84
SNV	0,69	0,56	2,38	0,49	0,73	1,84
CRR	0,46	0,75	1,79	0,58	0,66	2,02
AB	0,54	0,69	1,95	0,31	0,85	1,58
SMO	0,46	0,75	1,79	0,32	0,84	1,60

Fonte: Elaborada pelos autores (2023). R^2 = coeficiente de determinação; RMSE = raiz do erro quadrático médio; RPIQ = razão de desempenho do intervalo interquartil.

Com base nos resultados demonstrados por meio da Tabela 4, os modelos que utilizaram PLSR obtiveram variação de R^2 de 0,46 a 0,69 na validação cruzada. O resultado mais satisfatório foi atribuído aos pré-processamentos MSC e SNV, que apresentaram o mesmo R^2 (0,69), com pequena variação no RMSE (0,57% e 0,56%, respectivamente) e no RPIQ (2,37 e 2,38). Para esses modelos, a razão entre o desvio-padrão dos teores observados (1,03%) e o RMSE corresponde a um RPD de aproximadamente 1,8, valor que, pelos critérios de Chang et al. (2001), indica capacidade preditiva moderada, adequada para estimativas aproximadas e triagem, mas não para substituir a análise de referência em amostras individuais. A absorbância também melhorou o ajuste em relação ao espectro bruto (R^2 = 0,54; RMSE = 0,69%).

Os resultados encontrados reforçam a necessidade de aplicação de pré-processamentos, visto que os espectros brutos obtiveram o menor desempenho em ambos os algoritmos. A equivalência entre SNV e MSC é esperada, pois as duas transformações são algebricamente relacionadas e produzem espectros quase idênticos quando o espectro de referência da MSC é a média do conjunto (Dhanoa et al., 1994; Wadoux et al., 2021). O ganho obtido com essas técnicas indica que parte da variância espectral não relacionada ao COS decorre do espalhamento da luz por diferenças de tamanho e arranjo das partículas, efeito acentuado em um conjunto que reúne desde solos com mais de 900 g kg^{-1} de areia até solos siltosos e argilosos (Tabela 3). Dotto et al. (2018) também verificaram, em solos subtropicais, que as técnicas de correção de espalhamento superaram as derivadas na predição de COS por PLSR. Em uma biblioteca espectral de solos piauienses, Mendes et al. (2021) mostraram que o pré-processamento ótimo variou com o atributo e o algoritmo, o que reforça que a escolha deve ser testada em cada conjunto de dados, podendo ser influenciada por fatores como teores granulométricos, variação litológica ou quantidade de amostras utilizadas.

Os modelos PLSR com SPEC, SG, CRR e SMO apresentaram métricas idênticas (R^2 = 0,46; RMSE = 0,75%; RPIQ = 1,79). A inspeção da base de dados das amostras mostra que as predições desses quatro modelos coincidem em 51 das 52 amostras, o que não seria esperado para transformações tão distintas quanto a derivada de Savitzky-Golay e a remoção do contínuo. Esse resultado sugere que os pré-processamentos SG, CRR e SMO podem não ter sido efetivamente aplicados na rotina de PLSR, e não deve ser interpretado como insensibilidade do algoritmo a essas técnicas até que a rotina seja verificada pelos autores.

Conforme os resultados demonstrados na Tabela 4, os valores obtidos para as calibrações realizadas por RF foram relativamente baixos (R^2 de 0,28 a 0,58). O pré-tratamento que apresentou melhor ajuste foi CRR, que atingiu o maior R^2 (0,58), com RMSE de 0,66% e RPIQ de 2,02, enquanto o de menor precisão foi SPEC, sem aplicação de técnicas, como ocorreu no PLSR. A remoção do

contínuo normaliza as feições de absorção em relação a uma envoltória convexa e realça sua profundidade e forma (Clark; Roush, 1984), o que pode favorecer um algoritmo que seleciona comprimentos de onda isolados em cada divisão das árvores. No RF, MSC e SNV produziram métricas iguais ($R^2 = 0,49$), e SG apresentou R^2 de 0,52.

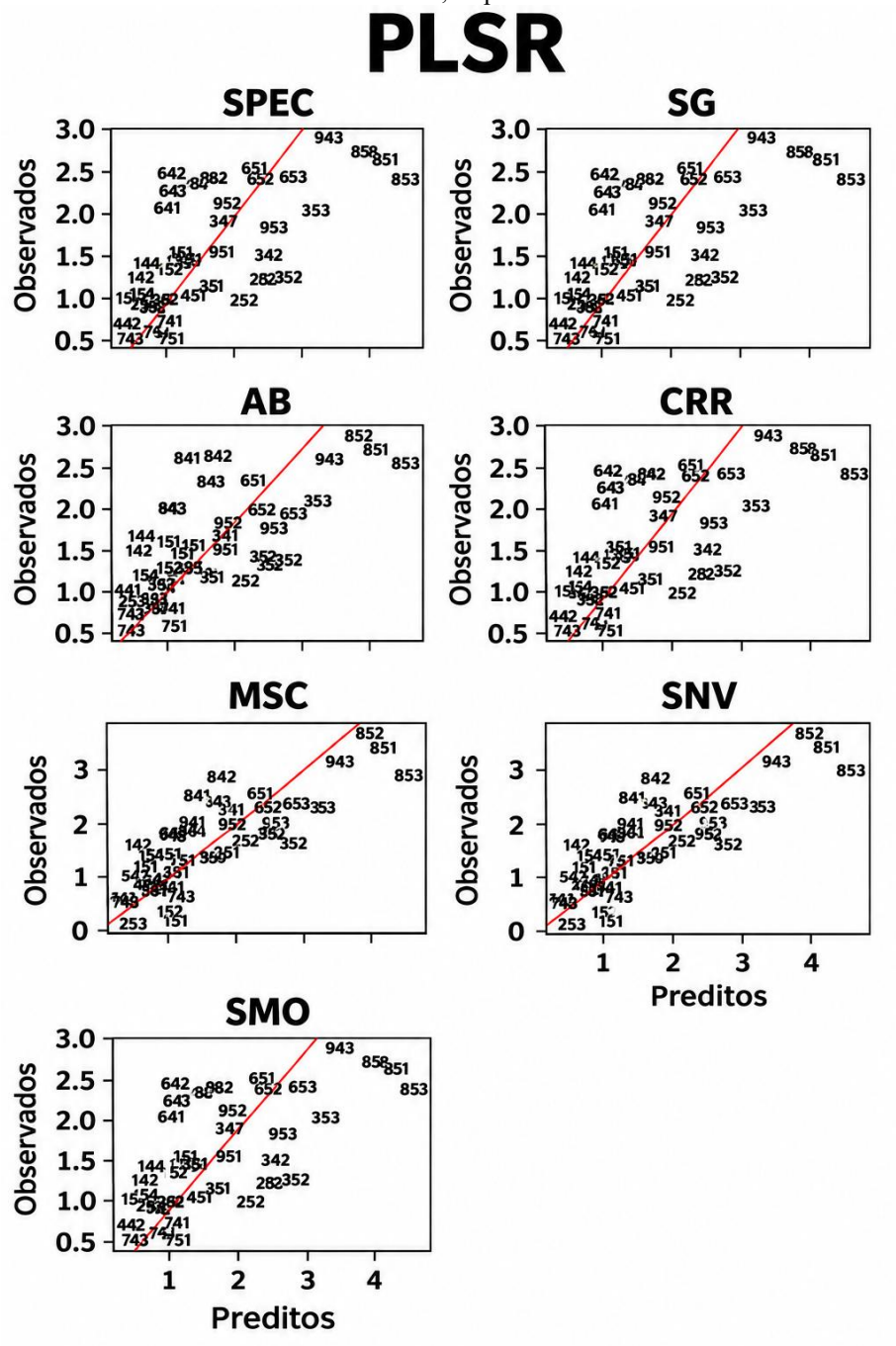
De modo geral, a calibração realizada por RF demonstrou baixa eficiência, que não ultrapassou R^2 de 0,58, o que pode ter relação com o número de amostras utilizado para calibração. A literatura mostra que o erro de predição diminui com o aumento do número de amostras de calibração e que conjuntos grandes, ou estratificados por similaridade, melhoram a predição de COS (Kuang; Mouazen, 2012; Moura-Bueno et al., 2019; Moura-Bueno et al., 2020). Isso ocorre principalmente para o modelo RF, pois este necessita de um conjunto amostral maior para obter um melhor ajuste, enquanto o PLSR, que tem como premissa a criação de componentes, tende a diminuir a variância. Horst-Heinen et al. (2021) também obtiveram desempenho mais consistente para PLSR do que para RF na predição de COS por Vis-NIR-SWIR em solos do Sul do Brasil, e comparações em bibliotecas espectrais não indicaram superioridade do RF sobre outros métodos de calibração, além de destacarem sua menor interpretabilidade em relação aos modelos lineares (Viscarra Rossel; Behrens, 2010).

É notável que os modelos de calibração construídos por PLSR apresentaram desempenho superior aos de RF para a maioria dos pré-processamentos. As exceções foram CRR e SG, únicos casos em que o RF superou o PLSR (R^2 de 0,58 e 0,52 contra 0,46), o que reforça a suspeita sobre a aplicação desses pré-processamentos no PLSR. Lalitha et al. (2023) também obtiveram baixo desempenho do RF para estimar COS em solos semiáridos cultivados da Índia ($R^2 = 0,22$), embora o algoritmo tenha sido efetivo para outros atributos, como pH e argila. Mahmoudzadeh et al. (2020), modelando o COS de 865 amostras do oeste do Irã com covariáveis ambientais, e não com espectros, também não obtiveram R^2 superior a 0,60 com RF. Assim, pode-se inferir que os baixos teores e a pequena amplitude de COS, característicos de áreas áridas e semiáridas, dificultam a predição por esse algoritmo. Ainda assim, o RPIQ de 2,02 do RF-CRR indica que algoritmos não lineares podem ser competitivos se treinados com bibliotecas maiores (Araújo et al., 2014).

4.3 PADRÕES DE SUBESTIMAÇÃO E SUPERESTIMAÇÃO

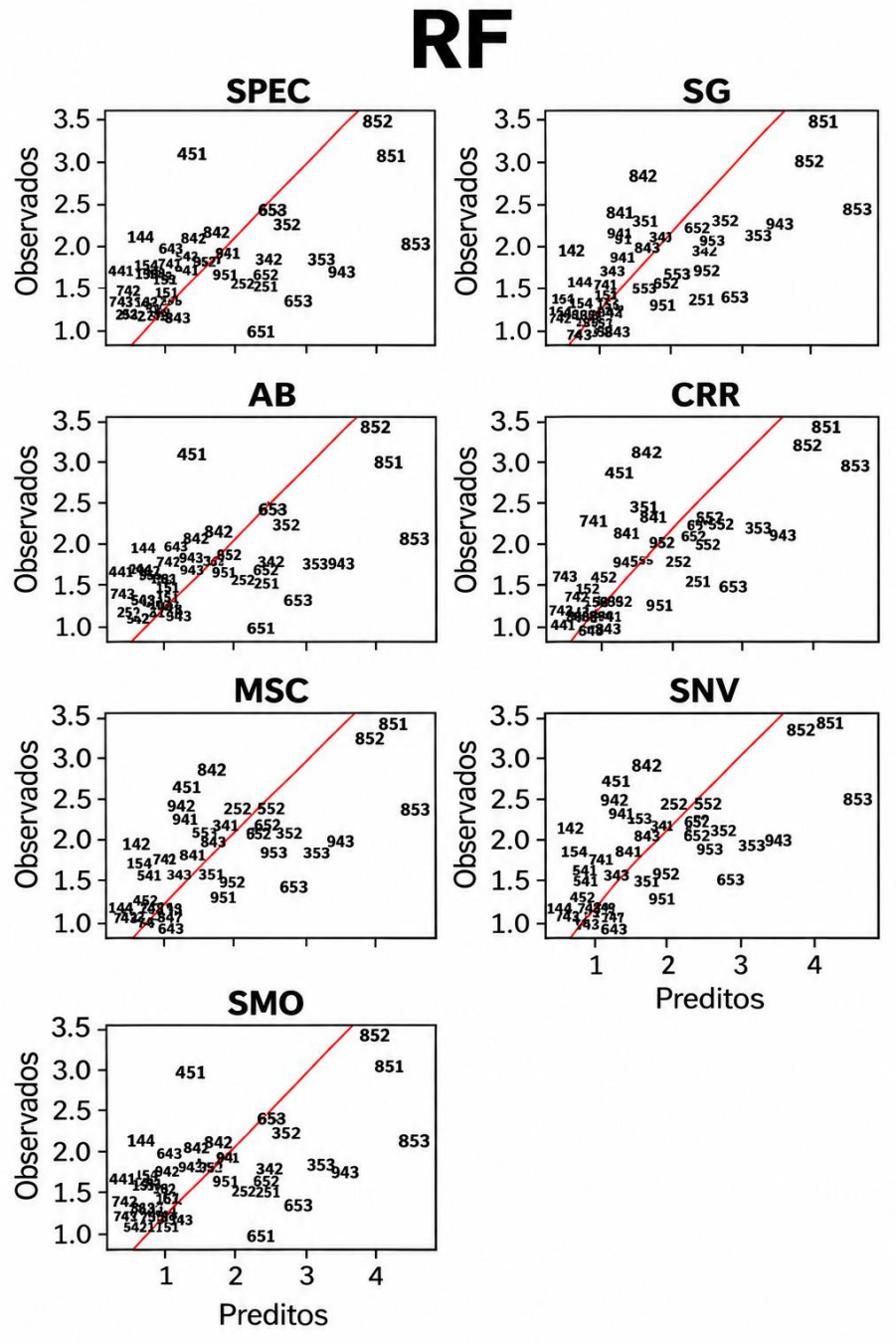
As relações entre os teores observados e preditos para os modelos PLSR (Figura 2) e RF (Figura 3) mostram tendência positiva em todos os modelos, mas com dispersão que varia conforme a área de coleta. Na comparação com a base de dados das amostras, verificou-se que o eixo horizontal desses gráficos contém os valores observados e o vertical, os preditos, ao contrário dos rótulos impressos nas figuras originais.

Figura 2 – Teores de COS (%) observados e preditos por validação cruzada leave-one-out nos modelos PLSR calibrados com o espectro em reflectância (SPEC) e com os pré-processamentos SG, AB, CRR, MSC, SNV e SMO. Os números identificam as amostras (município, condição e trincheira) e a linha vermelha é a linha de tendência traçada pelos autores. Nota: os rótulos dos eixos da figura original estão invertidos; o eixo horizontal contém os valores observados (até 4,64%) e o vertical, os preditos.



Fonte: Elaborada pelos autores (2023).

Figura 3 – Teores de COS (%) observados e preditos por validação cruzada leave-one-out nos modelos RF calibrados com o espectro em reflectância (SPEC) e com os pré-processamentos SG, AB, CRR, MSC, SNV e SMO. Os números identificam as amostras e a linha vermelha é a linha de tendência traçada pelos autores. Nota: como na Figura 2, o eixo horizontal contém os valores observados e o vertical, os preditos.

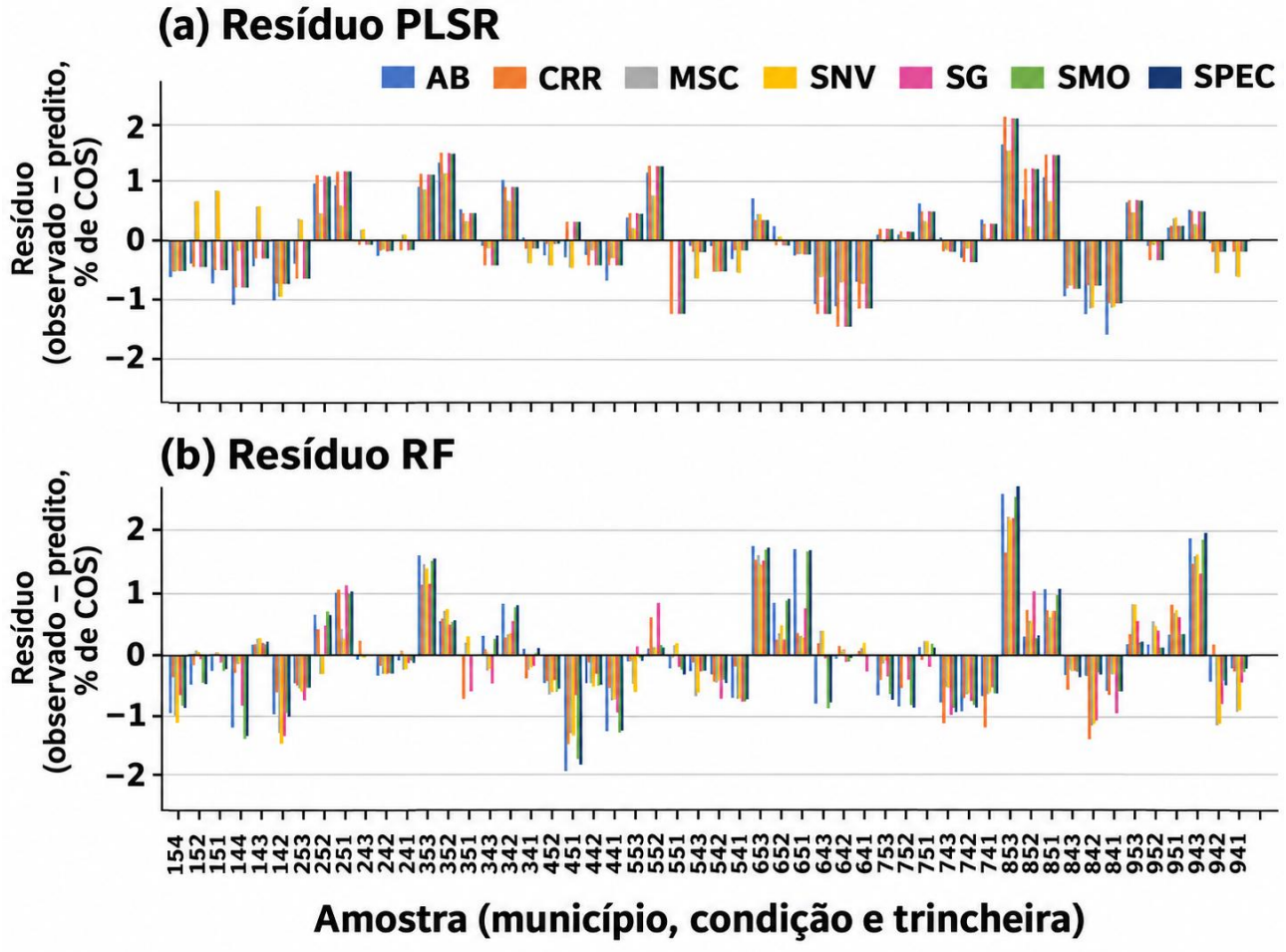


Fonte: Elaborada pelos autores (2023).

Os resíduos por amostra (Figura 4) evidenciam um padrão sistemático comum aos dois algoritmos: as amostras com teores elevados de COS, em sua maioria de áreas naturais, foram subestimadas, e as amostras com teores baixos, em sua maioria de áreas antropizadas, foram superestimadas. No PLSR-SNV, as dez amostras com COS acima de 2,5% tiveram resíduo médio de +0,70% (subestimação), e 12 das 16 amostras com COS abaixo de 1,0% foram superestimadas; no

conjunto, 20 das 26 amostras naturais foram subestimadas e 20 das 26 antropizadas foram superestimadas. Em Irecê-BA, as amostras naturais 851, 852 e 853 (4,03% a 4,64%) foram preditas entre 3,11% e 3,79% pelo PLSR-SNV, enquanto as antropizadas 841, 842 e 843 (1,44% a 1,78%) foram preditas entre 2,46% e 2,90% (Tabela 3). Esse mesmo comportamento foi frequente em outras áreas preservadas, como Sobral-CE e Araripina-PE, e nas áreas antropizadas de Feira Nova-SE e Campina Grande-PB. O texto original descreveu esse padrão de forma invertida (superestimação nas áreas naturais); a direção aqui relatada segue a definição de resíduo da Figura 4 e os valores observados e preditos da base de dados das amostras.

Figura 4 – Resíduos (observado − predito, % de COS) por amostra para todos os pré-processamentos e para o espectro em reflectância: (a) modelos calibrados por PLSR; (b) modelos calibrados por RF. Resíduos positivos indicam subestimação e negativos, superestimação; as cores identificam os pré-processamentos (AB, CRR, MSC, SNV, SG, SMO e SPEC).



Fonte: Elaborada pelos autores (2023).

Dois mecanismos explicam esse padrão. O primeiro é estatístico: modelos de regressão calibrados com poucas amostras e distribuição assimétrica tendem a contrair as predições em direção à média, porque os extremos são pouco representados na calibração e, na validação leave-one-out, a

retirada de uma das poucas amostras extremas reduz ainda mais essa representação (Bellon-Maurel et al., 2010; Kuang; Mouazen, 2012). No RF, o efeito é acentuado porque as predições são médias de valores observados nas folhas das árvores e, por isso, não ultrapassam a amplitude dos dados de treinamento (Breiman, 2001; Hengl et al., 2018); isso explica por que o RF subestimou mais os teores altos e superestimou mais os teores baixos do que o PLSR (Tabela 3). O segundo mecanismo é espectral: como o COS covaria com a granulometria e com a mineralogia (seção 4.1), os modelos podem ter aprendido parcialmente essa covariação, atribuindo teores elevados a solos argilosos ou siltosos, como as amostras antropizadas de Irecê-BA e Feira Nova-SE, e teores baixos a solos arenosos, independentemente do seu conteúdo real de matéria orgânica (Stenberg et al., 2010; Kuang; Mouazen, 2013). A sobreposição das feições de óxidos de ferro e argilominerais ao sinal orgânico reforça essa confusão em solos com baixos teores de COS (Aïchi et al., 2020).

Houve um comportamento bastante variável entre as amostras e as técnicas empregadas, e algumas amostras destoaram do padrão geral, como 451 (natural, 1,43%), fortemente superestimada pelo RF (predições de 2,07% a 3,31%), e 651 (natural, 2,39%, com 831 g kg^{-1} de argila), bem predita pelo PLSR, mas subestimada pelo RF com SPEC, AB e SMO. Considerando que as áreas antropizadas têm como característica o uso pecuário e, em alguns casos, agrícola, a maior frequência de superestimação nessas áreas pode estar associada também à variedade de características dos solos decorrente dos manejos realizados, que altera a relação entre o COS e as propriedades espectrais dominadas pela fração mineral.

4.4 LIMITAÇÕES E IMPLICAÇÕES

Os resultados devem ser interpretados à luz de algumas limitações. O conjunto de 52 amostras é pequeno para calibrações espectrais, especialmente para o RF, e a avaliação baseou-se apenas em validação cruzada leave-one-out, sem conjunto de teste independente; portanto, as métricas tendem a ser otimistas quanto à capacidade de generalização para outras áreas da Caatinga (Bellon-Maurel et al., 2010; Wadoux; Minasny; McBratney, 2020). As amostras foram coletadas em 2013 e restringem-se à camada de 0–5 cm, de modo que os modelos não representam o perfil do solo nem a condição atual das áreas. Os métodos de referência para COS e granulometria, a geometria de leitura, os parâmetros dos pré-processamentos (janela e ordem do Savitzky-Golay e da suavização) e os hiperparâmetros do RF não foram informados, o que limita a reprodutibilidade. Por fim, as predições idênticas do PLSR para SPEC, SG, CRR e SMO indicam a necessidade de verificar a rotina de pré-processamento antes de conclusões definitivas sobre esses tratamentos.

Apesar dessas restrições, o estudo oferece uma contribuição metodológica e regional. Em

termos metodológicos, demonstra que, em conjuntos pequenos e heterogêneos de solos semiáridos, o PLSR associado a pré-processamentos de correção de espalhamento é mais estável que o RF, e que o erro não é aleatório, mas concentrado nos extremos da distribuição de COS. Em termos regionais, fornece uma das primeiras avaliações de modelos espectrais com amostras de nove municípios da Caatinga, incluindo áreas naturais e de pastejo, o que é relevante para programas de monitoramento do carbono do solo em um bioma com elevada pressão de desmatamento (MapBiomas, 2023). Para avançar, recomenda-se ampliar a biblioteca espectral regional, com amostras estratificadas por classe de solo, textura e uso da terra e com maior representação de teores elevados de COS (Moura-Bueno et al., 2019; Demattê et al., 2019), reservar um conjunto de validação independente e testar a correção de efeitos externos e algoritmos baseados em similaridade espectral.

5 CONCLUSÃO

Os modelos de predição de COS calibrados com espectros Vis-NIR-SWIR de solos da Caatinga apresentaram capacidade preditiva moderada. O modelo de melhor ajuste foi PLSR, quando aplicadas as técnicas de pré-tratamento SNV e MSC, que obtiveram R^2 de 0,69 e uma pequena variação no RMSE com 0,56 e 0,57, respectivamente, bem como o RPIQ de 2,38 e 2,37. As calibrações com RF exibiram maior sensibilidade às técnicas, porém apresentaram métricas com menor precisão quando comparadas ao PLSR, e o pré-processamento CRR foi o mais satisfatório (R^2 de 0,58, RMSE de 0,66% e RPIQ de 2,02).

Os modelos evidenciaram a necessidade da aplicação de técnicas de pré-tratamento, visto que os espectros em reflectância não demonstraram bons ajustes, sendo necessárias correções para eliminação de ruídos, e/ou salientar feições e características específicas para um melhoramento dos resultados de predição. A equivalência entre SNV e MSC indica que a correção do espalhamento da luz é o principal ganho nesse conjunto de solos de texturas contrastantes.

Os teores de COS foram, em média, maiores nas áreas naturais do que nas antropizadas em todos os municípios, mas a ocorrência de áreas antropizadas com teores elevados indica que o manejo, a textura e a classe de solo também condicionam o armazenamento de carbono. Ambos os algoritmos subestimaram os teores elevados, predominantes nas áreas naturais, e superestimaram os teores baixos, padrão atribuído à contração das predições em direção à média e à covariação entre COS, textura e mineralogia.

O estudo mostra que a espectrorradiometria associada à quimiometria é uma ferramenta promissora para o monitoramento do COS no semiárido, principalmente para a Caatinga, que demonstra alto grau de degradação e suscetibilidade. As conclusões limitam-se, contudo, a 52 amostras

superficiais coletadas em 2013 e a métricas de validação cruzada. O baixo quantitativo de amostras pode ser uma possível influência para os valores de desempenho do modelo, fazendo-se necessárias pesquisas posteriores com um conjunto amostral mais robusto, validação independente e otimização dos processamentos para reduzir o viés das predições de COS.

AGRADECIMENTOS

Agradecemos a todos os envolvidos na elaboração da presente pesquisa, em especial aos financiadores Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e INCT IN-TREE (Instituto Nacional de Ciência e Tecnologia em Estudos Inter- e Transdisciplinares em Ecologia e Evolução), bem como ao Programa de Pós-Graduação em Modelagem em Ciências da Terra e do Ambiente (PPGM), sediado na Universidade Estadual de Feira de Santana; ao Laboratório de Sensoriamento Remoto e Espectrorradiometria pelo aparato técnico e científico; e à Embrapa Semiárido, em Petrolina-PE, pela parceria na aquisição de dados.

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e pelo Ministério da Ciência, Tecnologia e Inovação (MCTI), por meio do edital CNPq/MCTI nº 441271/2023-5 – Presente, Passado e Futuro da Biodiversidade do Semiárido: Inventários, Monitoramento, Impactos das Mudanças Climáticas e Implicações para o Uso e Conservação da Flora, Fauna e Fungos. W.J.F.R. foi apoiado por bolsa de produtividade em pesquisa do CNPq (processo nº 314954/2021-0), Projeto Prospecta 4.0 – CNPq (processo nº 407907/2022-0) e pelo projeto 30202010010121001 – CTBIO – FINEP, associada ao Centro Temático de Bioeconomia do Semiárido (UEFS – CTBio-CIMATEC SERTÃO). Os autores agradecem ao Instituto Nacional de Ciência e Tecnologia em Estudos Interdisciplinares e Transdisciplinares em Ecologia e Evolução (INCT IN-TREE), financiado pelo CNPq (408930/2024-1), pela CAPES (88887.195651/2025-00) e pela FAPESP. Financiamento adicional foi fornecido pela Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB), por meio do programa Bioeconomia, Recursos Hídricos e Sustentabilidade Ambiental na Bahia, no âmbito da Chamada nº 38/2022, Acordo de Cooperação Técnica nº 294/2023 e Termo de Outorga nº PPF0003/2023. Os autores agradecem ao Programa de Pós-Graduação em Modelagem em Ciências da Terra e do Ambiente (PPGM/UEFS) pelo apoio institucional, bem como ao SENAI CIMATEC pela contribuição no desenvolvimento científico e tecnológico voltado à inovação aplicada. R.N.V. foi apoiado pelo INCT IN-TREE e agradece a bolsa SET-G 30202010010121001 – CTBIO – FINEP, associada ao Centro Temático de Bioeconomia do Semiárido (UEFS – CTBio-CIMATEC SERTÃO), vinculado a Projetos Especiais.

REFERÊNCIAS

AÏCHI, H.; FOUAD, Y.; CAUSEUR, D. et al. Organic carbon and total iron effect on soil Vis-SWNIR spectra and quantification of their contents using PLS R models. Communications in Soil Science and Plant Analysis, v. 51, n. 9, p. 1253-1267, 2020. DOI: https://doi.org/10.1080/00103624.2020.1751187.

ALVARES, Clayton Alcarde; STAPE, José Luiz; SENTELHAS, Paulo Cesar et al. Köppen's climate classification map for Brazil. Meteorologische Zeitschrift, v. 22, n. 6, p. 711-728, 2013. DOI: https://doi.org/10.1127/0941-2948/2013/0507.

ARAÚJO, S. R.; WETTERLIND, J.; DEMATTÊ, J. A. M. et al. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. European Journal of Soil Science, v. 65, n. 5, p. 718-729, 2014. DOI: https://doi.org/10.1111/ejss.12165.

BARNES, R. J.; DHANOA, M. S.; LISTER, Susan J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied Spectroscopy, v. 43, n. 5, p. 772-777, 1989. DOI: https://doi.org/10.1366/0003702894202201.

BELLON-MAUREL, Véronique; FERNANDEZ-AHUMADA, Elvira; PALAGOS, Bernard et al. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC Trends in Analytical Chemistry, v. 29, n. 9, p. 1073-1081, 2010. DOI: https://doi.org/10.1016/j.trac.2010.05.006.

BEN-DOR, E. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400-2500 nm) during a controlled decomposition process. Remote Sensing of Environment, v. 61, n. 1, p. 1-15, 1997. DOI: https://doi.org/10.1016/S0034-4257(96)00120-4.

BEN-DOR, Eyal; ONG, Cindy; LAU, Ian C. Reflectance measurements of soils in the laboratory: standards and protocols. Geoderma, v. 245-246, p. 112-124, 2015. DOI: https://doi.org/10.1016/j.geoderma.2015.01.002.

BREIMAN, Leo. Random forests. Machine Learning, v. 45, n. 1, p. 5-32, 2001. DOI: https://doi.org/10.1023/A:1010933404324.

CHANG, Cheng-Wen; LAIRD, David A.; MAUSBACH, Maurice J. et al. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Science Society of America Journal, v. 65, n. 2, p. 480-490, 2001. DOI: https://doi.org/10.2136/sssaj2001.652480x.

CLARK, Roger N.; ROUSH, Ted L. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. Journal of Geophysical Research: Solid Earth, v. 89, n. B7, p. 6329-6340, 1984. DOI: https://doi.org/10.1029/JB089iB07p06329.

COZZOLINO, D.; MORÓN, A. Potential of near-infrared reflectance spectroscopy and chemometrics to predict soil organic carbon fractions. Soil and Tillage Research, v. 85, n. 1-2, p. 78-85, 2006. DOI: https://doi.org/10.1016/j.still.2004.12.006.

DALMOLIN, Ricardo Simão Diniz; GONÇALVES, Cristiano Nunes; KLAMT, Egon et al. Relação

entre os constituintes do solo e seu comportamento espectral. Ciência Rural, v. 35, n. 2, p. 481-489, 2005. DOI: https://doi.org/10.1590/S0103-84782005000200042.

DAS, Bappa; CHAKRABORTY, Debashis; SINGH, Vinod Kumar et al. Partial least square regression based machine learning models for soil organic carbon prediction using visible–near infrared spectroscopy. Geoderma Regional, v. 33, e00628, 2023. DOI: https://doi.org/10.1016/j.geodrs.2023.e00628.

DEMATTÊ, José A. M.; DOTTO, André Carnieletto; PAIVA, Ariane F. S. et al. The Brazilian Soil Spectral Library (BSSL): a general view, application and challenges. Geoderma, v. 354, 113793, 2019. DOI: https://doi.org/10.1016/j.geoderma.2019.05.043.

DHANOA, M. S.; LISTER, S. J.; SANDERSON, R. et al. The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. Journal of Near Infrared Spectroscopy, v. 2, n. 1, p. 43-47, 1994. DOI: https://doi.org/10.1255/jnirs.30.

DIAS, F. P. M.; SILVA, F. T. S.; SILVEIRA JUNIOR, M. C. et al. Content and carbon stock in organic production systems in the Bahia semiarid. Scientific Electronic Archives, v. 10, n. 3, p. 19-23, 2017.

DOTTO, André Carnieletto; DALMOLIN, Ricardo Simão Diniz; TEN CATEN, Alexandre et al. A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. Geoderma, v. 314, p. 262-274, 2018. DOI: https://doi.org/10.1016/j.geoderma.2017.11.006.

GAVA, Carlos Alberto Tuão; GIONGO, Vanderlise; SIGNOR, Diana et al. Land-use change alters the stocks of carbon, nitrogen, and phosphorus in a Haplic Cambisol in the Brazilian semi-arid region. Soil Use and Management, v. 38, n. 1, p. 953-963, 2022. DOI: https://doi.org/10.1111/sum.12716.

GHOLIZADEH, Asa; SABERIOON, Mohammadmehdi; POULADI, Nastaran et al. Quantification and depth distribution analysis of carbon to nitrogen ratio in forest soils using reflectance spectroscopy. International Soil and Water Conservation Research, v. 11, n. 1, p. 112-124, 2023. DOI: https://doi.org/10.1016/j.iswcr.2022.06.004.

GIONGO, V.; GALVÃO, S. R. S.; MENDES, A. M. S. et al. Soil organic carbon in the Brazilian semi-arid tropics. Dynamic Soil, Dynamic Plant, v. 5, n. esp. 1, p. 12-20, 2011. Disponível em: https://www.alice.cnptia.embrapa.br/alice/handle/doc/915039. Acesso em: 14 nov. 2026.

GOMEZ, Cécile; VISCARRA ROSSEL, Raphael A.; MCBRATNEY, Alex B. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: an Australian case study. Geoderma, v. 146, n. 3-4, p. 403-411, 2008. DOI: https://doi.org/10.1016/j.geoderma.2008.06.011.

GUO, Long; ZHANG, Haitao; SHI, Tiezhu et al. Prediction of soil organic carbon stock by laboratory spectral data and airborne hyperspectral images. Geoderma, v. 337, p. 32-41, 2019. DOI: https://doi.org/10.1016/j.geoderma.2018.09.003.

HELLAND, Inge S.; NÆS, Tormod; ISAKSSON, Tomas. Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. Chemometrics and Intelligent

Laboratory Systems, v. 29, n. 2, p. 233-241, 1995. DOI: https://doi.org/10.1016/0169-7439(95)80098-T.

HENGL, Tomislav; NUSSBAUM, Madlene; WRIGHT, Marvin N. et al. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ, v. 6, e5518, 2018. DOI: https://doi.org/10.7717/peerj.5518.

HORST-HEINEN, Taciara Zborowski; DALMOLIN, Ricardo Simão Diniz; SAMUEL-ROSA, Alessandro et al. The interplay among analytical method, preprocessing, and modeling on soil organic carbon Vis-NIR-SWIR predictions. In: EGU GENERAL ASSEMBLY 2021, online, 19-30 abr. 2021. Resumo EGU21-7851. DOI: https://doi.org/10.5194/egusphere-egu21-7851.

HUNT, G. R. Spectral signatures of particulate minerals in the visible and near infrared. Geophysics, v. 42, n. 3, p. 501-513, 1977. DOI: https://doi.org/10.1190/1.1440721.

IBGE – Instituto Brasileiro de Geografia e Estatística. Biomas e sistema costeiro-marinho do Brasil: compatível com a escala 1:250 000. Rio de Janeiro: IBGE, 2019. (Relatórios Metodológicos, v. 45).

ISAKSSON, Tomas; NÆS, Tormod. The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. Applied Spectroscopy, v. 42, n. 7, p. 1273-1284, 1988. DOI: https://doi.org/10.1366/0003702884429869.

IUSS WORKING GROUP WRB. World reference base for soil resources 2014, update 2015: international soil classification system for naming soils and creating legends for soil maps. Rome: FAO, 2015. (World Soil Resources Reports, 106).

JI, W.; VISCARRA ROSSEL, R. A.; SHI, Z. Improved estimates of organic carbon using proximally sensed vis-NIR spectra corrected by piecewise direct standardization. European Journal of Soil Science, v. 66, n. 4, p. 670-678, 2015. DOI: https://doi.org/10.1111/ejss.12271.

KER, João Carlos; CURI, Nilton; SCHAEFER, Carlos Ernesto G. R. et al. (ed.). Pedologia: fundamentos. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, 2012. 343 p.

KUANG, B.; MOUAZEN, A. M. Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale. European Journal of Soil Science, v. 63, n. 3, p. 421-429, 2012. DOI: https://doi.org/10.1111/j.1365-2389.2012.01456.x.

KUANG, Boyan; MOUAZEN, Abdul M. Non-biased prediction of soil organic carbon and total nitrogen with vis–NIR spectroscopy, as affected by soil moisture content and texture. Biosystems Engineering, v. 114, n. 3, p. 249-258, 2013. DOI: https://doi.org/10.1016/j.biosystemseng.2013.01.005.

KUHN, Max. Building predictive models in R using the caret package. Journal of Statistical Software, v. 28, n. 5, 2008. DOI: https://doi.org/10.18637/jss.v028.i05.

LAL, R. Soil carbon sequestration impacts on global climate change and food security. Science, v. 304, n. 5677, p. 1623-1627, 2004. DOI: https://doi.org/10.1126/science.1097396.

LALITHA, M.; DHARUMARAJAN, S.; GOMEZ, C. et al. Spectral assessment of soil properties in

semi-arid tropical regions of southern Karnataka Plateau. Archives of Agronomy and Soil Science, v. 69, n. 11, p. 2065-2080, 2023. DOI: https://doi.org/10.1080/03650340.2022.2134565.

MAHMOUDZADEH, Hamid; MATINFAR, Hamid Reza; TAGHIZADEH-MEHRJARDI, Ruhollah et al. Spatial prediction of soil organic carbon using machine learning techniques in western Iran. Geoderma Regional, v. 21, e00260, 2020. DOI: https://doi.org/10.1016/j.geodrs.2020.e00260.

MAPBIOMAS. Relatório anual do desmatamento no Brasil 2022. São Paulo: MapBiomas, 2023. 125 p. Disponível em: https://alerta.mapbiomas.org. Acesso em: 14 nov. 2026.

MENDES, Wanderson de Sousa; BOECHAT, Cácio Luiz; GUALBERTO, Adriano Venicius Santana et al. Soil spectral library of Piauí State using machine learning for laboratory analysis in Northeastern Brazil. Revista Brasileira de Ciência do Solo, v. 45, e0200115, 2021. DOI: https://doi.org/10.36783/18069657rbcs20200115.

MENESES, Paulo Roberto; ALMEIDA, Tati de; BAPTISTA, Gustavo Macedo de Mello. Reflectância dos materiais terrestres: análise e interpretação. São Paulo: Oficina de Textos, 2019.

MENESES, Paulo Roberto; MADEIRA NETTO, José da Silva (org.). Sensoriamento remoto: reflectância dos alvos naturais. Brasília, DF: Editora Universidade de Brasília; Planaltina: Embrapa Cerrados, 2001.

MENEZES, R. S. C.; SAMPAIO, E. V. S. B.; GIONGO, V. et al. Biogeochemical cycling in terrestrial ecosystems of the Caatinga Biome. Brazilian Journal of Biology, v. 72, n. 3, supl., p. 643-653, 2012. DOI: https://doi.org/10.1590/S1519-69842012000400004.

MENEZES, Rômulo Simões Cezar; SALES, Aldo Torres; PRIMO, Dário Costa et al. Soil and vegetation carbon stocks after land-use changes in a seasonally dry tropical forest. Geoderma, v. 390, 114943, 2021. DOI: https://doi.org/10.1016/j.geoderma.2021.114943.

MOURA-BUENO, Jean Michel; DALMOLIN, Ricardo Simão Diniz; TEN CATEN, Alexandre et al. Stratification of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions. Geoderma, v. 337, p. 565-581, 2019. DOI: https://doi.org/10.1016/j.geoderma.2018.10.015.

MOURA-BUENO, Jean Michel; DALMOLIN, Ricardo Simão Diniz; HORST-HEINEN, Taciara Zborowski et al. When does stratification of a subtropical soil spectral library improve predictions of soil organic carbon content? Science of the Total Environment, v. 737, 139895, 2020. DOI: https://doi.org/10.1016/j.scitotenv.2020.139895.

NOCITA, Marco; STEVENS, Antoine; VAN WESEMAEL, Bas et al. Soil spectroscopy: an alternative to wet chemistry for soil monitoring. Advances in Agronomy, v. 132, p. 139-159, 2015. DOI: https://doi.org/10.1016/bs.agron.2015.02.002.

PINHEIRO, Érika; CEDDIA, Marcos; CLINGENSMITH, Christopher et al. Prediction of soil physical and chemical properties by visible and near-infrared diffuse reflectance spectroscopy in the central Amazon. Remote Sensing, v. 9, n. 4, 293, 2017. DOI: https://doi.org/10.3390/rs9040293.

RIBEIRO, Sharon Gomes; TEIXEIRA, Adunias dos Santos; OLIVEIRA, Marcio Regys Rabelo de et

al. Soil organic carbon content prediction using soil-reflected spectra: a comparison of two regression methods. Remote Sensing, v. 13, n. 23, 4752, 2021. DOI: https://doi.org/10.3390/rs13234752.

RINNAN, Åsmund; BERG, Frans van den; ENGELSEN, Søren Balling. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry, v. 28, n. 10, p. 1201-1222, 2009. DOI: https://doi.org/10.1016/j.trac.2009.07.007.

SACRAMENTO, José Augusto Amorim Silva do; ARAÚJO, Ana Caroline de Moraes; ESCOBAR, Maria Eugenia Ortiz et al. Soil carbon and nitrogen stocks in traditional agricultural and agroforestry systems in the semiarid region of Brazil. Revista Brasileira de Ciência do Solo, v. 37, n. 3, p. 784-795, 2013. DOI: https://doi.org/10.1590/S0100-06832013000300025.

SANTOS, Humberto Gonçalves dos; JACOMINE, Paulo Klinger Tito; ANJOS, Lúcia Helena Cunha dos et al. Sistema brasileiro de classificação de solos. 5. ed. Brasília, DF: Embrapa, 2018.

SAVITZKY, Abraham; GOLAY, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, v. 36, n. 8, p. 1627-1639, 1964. DOI: https://doi.org/10.1021/ac60214a047.

SCHARLEMANN, Jörn P. W.; TANNER, Edmund V. J.; HIEDERER, Roland et al. Global soil carbon: understanding and managing the largest terrestrial carbon pool. Carbon Management, v. 5, n. 1, p. 81-91, 2014. DOI: https://doi.org/10.4155/cmt.13.77.

SCHULZ, Katharina; VOIGT, Karsten; BEUSCH, Christine et al. Grazing deteriorates the soil carbon stocks of Caatinga forest ecosystems in Brazil. Forest Ecology and Management, v. 367, p. 62-70, 2016. DOI: https://doi.org/10.1016/j.foreco.2016.02.011.

SHI, Xiaoyan; SONG, Jianghui; WANG, Haijiang et al. Improving soil organic matter estimation accuracy by combining optimal spectral preprocessing and feature selection methods based on pXRF and vis-NIR data fusion. Geoderma, v. 430, 116301, 2023. DOI: https://doi.org/10.1016/j.geoderma.2022.116301.

SILVA, Kleber Andrade da; ARAÚJO, Elcida de Lima; FERRAZ, Elba Maria Nogueira. Estudo florístico do componente herbáceo e relação com solos em áreas de caatinga do embasamento cristalino e bacia sedimentar, Petrolândia, PE, Brasil. Acta Botanica Brasilica, v. 23, n. 1, p. 100-110, 2009. DOI: https://doi.org/10.1590/S0102-33062009000100013.

SOUZA, Diego M.; MADARI, Beata E.; GUIMARÃES, Freddy F. Aplicação de técnicas multivariadas e inteligência artificial na análise de espectros de infravermelho para determinação de matéria orgânica em amostras de solos. Química Nova, v. 35, n. 9, p. 1738-1745, 2012. DOI: https://doi.org/10.1590/S0100-40422012000900007.

STENBERG, Bo; VISCARRA ROSSEL, Raphael A.; MOUAZEN, Abdul Mounem et al. Visible and near infrared spectroscopy in soil science. Advances in Agronomy, v. 107, p. 163-215, 2010. DOI: https://doi.org/10.1016/S0065-2113(10)07005-7.

TEKIN, Yücel; TÜMSAVAS, Zeynal; MOUAZEN, Abdul Mounem. Comparing the artificial neural network with parcial least squares for prediction of soil organic carbon and pH at different moisture

content levels using visible and near-infrared spectroscopy. Revista Brasileira de Ciência do Solo, v. 38, n. 6, p. 1794-1804, 2014. DOI: https://doi.org/10.1590/S0100-06832014000600014.

TERRA, Fabrício S.; DEMATTÊ, José A. M.; VISCARRA ROSSEL, Raphael A. Spectral libraries for quantitative analyses of tropical Brazilian soils: comparing vis-NIR and mid-IR reflectance data. Geoderma, v. 255-256, p. 81-93, 2015. DOI: https://doi.org/10.1016/j.geoderma.2015.04.017.

TIESSEN, H.; FELLER, C.; SAMPAIO, E. V. S. B. et al. Carbon sequestration and turnover in semiarid savannas and dry forest. Climatic Change, v. 40, n. 1, p. 105-117, 1998. DOI: https://doi.org/10.1023/A:1005342932178.

VISCARRA ROSSEL, R. A.; BEHRENS, T. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, v. 158, n. 1-2, p. 46-54, 2010. DOI: https://doi.org/10.1016/j.geoderma.2009.12.025.

VISCARRA ROSSEL, R. A.; BEHRENS, T.; BEN-DOR, E. et al. A global spectral library to characterize the world's soil. Earth-Science Reviews, v. 155, p. 198-230, 2016. DOI: https://doi.org/10.1016/j.earscirev.2016.01.012.

VISCARRA ROSSEL, R. A.; WALVOORT, D. J. J.; MCBRATNEY, A. B. et al. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma, v. 131, n. 1-2, p. 59-75, 2006. DOI: https://doi.org/10.1016/j.geoderma.2005.03.007.

VISCARRA ROSSEL, Raphael A.; BEHRENS, Thorsten; BEN-DOR, Eyal et al. Diffuse reflectance spectroscopy for estimating soil properties: a technology for the 21st century. European Journal of Soil Science, v. 73, n. 4, e13271, 2022. DOI: https://doi.org/10.1111/ejss.13271.

WADOUX, Alexandre M. J.-C.; MALONE, Brendan; MINASNY, Budiman et al. Soil spectral inference with R: analysing digital soil spectra using the R programming environment. Cham: Springer, 2021. (Progress in Soil Science). DOI: https://doi.org/10.1007/978-3-030-64896-1.

WADOUX, Alexandre M. J.-C.; MINASNY, Budiman; MCBRATNEY, Alex B. Machine learning for digital soil mapping: applications, challenges and suggested solutions. Earth-Science Reviews, v. 210, 103359, 2020. DOI: https://doi.org/10.1016/j.earscirev.2020.103359.

WOLD, Svante; SJÖSTRÖM, Michael; ERIKSSON, Lennart. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, v. 58, n. 2, p. 109-130, 2001. DOI: https://doi.org/10.1016/S0169-7439(01)00155-1.

ZERAATPISHEH, Mojtaba; AYOUBI, Shamsollah; JAFARI, Azam et al. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. Geoderma, v. 338, p. 445-452, 2019. DOI: https://doi.org/10.1016/j.geoderma.2018.09.006.