UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA FUNDAMENTAL
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

PEDRO GABRIEL CORREIA DE LUCENA

USO DE TÉCNICAS ESPECTROSCÓPICAS E QUIMIOMÉTRICAS PARA RECONHECIMENTO DE PADRÕES E PREDIÇÃO DE PARÂMETROS DE SOLOS TROPICAIS SOB DIFERENTES CONDIÇÕES CLIMÁTICAS

Recife
2022

PEDRO GABRIEL CORREIA DE LUCENA

USO DE TÉCNICAS ESPECTROSCÓPICAS E QUIMIOMÉTRICAS PARA RECONHECIMENTO DE PADRÕES E PREDIÇÃO DE PARÂMETROS DE SOLOS TROPICAIS SOB DIFERENTES CONDIÇÕES CLIMÁTICAS

Dissertação apresentada ao Programa de Pós-Graduação em Química da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de mestre em Química.

Área de concentração: Química Analítica

Orientadora: Prof.ª Dra. Claudete Fernandes Pereira

Coorientadora: Dra. Juliet Emilia Santos de Sousa

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

L935u Lucena, Pedro Gabriel Correia de
Uso de técnicas espectroscópicas e quimiométricas para reconhecimento de padrões e predição de parâmetros de solos tropicais sob diferentes condições climáticas / Pedro Gabriel Correia de Lucena. – 2022.
90 f.: il., fig., tab.

Orientadora: Claudete Fernandes Pereira.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Química, Recife, 2022.

Inclui referências e apêndice.

1. Química analítica. 2. Mineralogia. I. Pereira, Claudete Fernandes (orientadora). II. Título.

543 CDD (23. ed.) UFPE - CCEN 2022-125

PEDRO GABRIEL CORREIA DE LUCENA

USO DE TÉCNICAS ESPECTROSCÓPICAS E QUIMIOMÉTRICAS PARA RECONHECIMENTO DE PADRÕES E PREDIÇÃO DE PARÂMETROS DE SOLOS TROPICAIS SOB DIFERENTES CONDIÇÕES CLIMÁTICAS.

Dissertação apresentada ao Programa de Pós-Graduação no Departamento de Química Fundamental da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Química.

Aprovada em: 09/06/2022

BANCA EXAMINADORA

_____
Profa. Claudete Fernandes Pereira (Orientadora)
Universidade Federal de Pernambuco

_____
Profa. Isabelle Moraes Amorim Viegas
Universidade Técnica da Dinamarca

_____
Prof. Yuri Jacques Agra Bezerra da Silva
Universidade Federal do Piauí

AGRADECIMENTOS

Agradeço à minha família, principalmente aos meus pais, Jurandir e Alice, por me apoiarem e sempre estarem ao meu lado, priorizando minha educação, não poupando sacrifícios para que aqui eu chegasse. Agradeço também a minha irmã, Jordana, por todo o carinho e ajuda no preparo das amostras.

À Profa. Dra. Claudete Pereira, minha orientadora, muito obrigado pela acolhida, gentileza, empatia e todos os ensinamentos nesse período.

À Dra. Juliet Sousa, obrigado por aceitar esse desafio de coorientação e por acreditar em mim. Obrigado pelas ideias, discussões e palavras que me acalmavam.

À Universidade Federal Rural de Pernambuco e Pós-Graduação em Ciência do Solo. Agradeço especialmente aos professores Dr. Valdomiro Souza Júnior, Dr. Ygor Silva e Dr. Mateus Ribeiro Filho, por todo o apoio e confiança na disponibilização de equipamentos e de amostras de solo para condução deste trabalho.

Ao Laboratório de Gênese e Mineralogia do Solo pelo suporte nas análises no MIR.

Ao Laboratório de Análise de Combustíveis pelo suporte nas análises no NIR.

Ao Laboratório de Refino e Tecnologias Limpas pelo suporte nas análises de EDXRF.

À Dra. Paula Araújo a frente da Coleção Mateus Rosas Ribeiro, sempre prestativa na seleção de amostras e compilação dos dados de solo.

Aos meus amigos, Mauro, Ramon, Everton, Luana, Caio, Marcela e Hanna pelos momentos compartilhados, reduziram o peso do mestrado no meio de uma pandemia.

Agradeço a todos os professores da área de solos e aqueles com quem convivi nesses 6 anos trabalhando no departamento de Agronomia pelos valiosos ensinamentos.

RESUMO

A granulometria e mineralogia dos solos interferem na sua capacidade de armazenamento de carbono orgânico, retenção de água e nutrientes. Esses parâmetros variam bastante em solos tropicais ao longo da paisagem. Dessa forma, métodos analíticos adequados são necessários para obter informações sobre os solos e, consequentemente, melhorar o manejo e preservação deste recurso. As técnicas espectroscópicas e quimiométricas surgem como excelentes alternativas para a análise de solos. Considerando-se esses aspectos, o objetivo deste trabalho foi desenvolver métodos espectroanalíticos para avaliação qualitativa da granulometria e mineralogia, e determinação quantitativa dos teores de areia, argila e carbono orgânico do solo (COS). Foram utilizadas 176 amostras superficiais e subsuperficiais de solos provenientes da Zona da Mata, Agreste e Sertão do Estado de Pernambuco. As amostras foram analisadas por espectroscopias na região do infravermelho médio com transformada de Fourier e acessório de refletância total atenuada (ATR-FTIR), no infravermelho próximo (NIR) com instrumentos portáteis (NanoNIR e MicroNIR) e fluorescência de raios-X por energia dispersiva (EDXRF). A Análise de Dimensões Comuns (ComDim) integrou informações ATR-FTIR e EDXRF, evidenciando semelhanças e diferenças entre as amostras de solo, agrupando-as em cinco categorias associadas às diferenças na composição granulométrica e mineral. Os resultados da discriminação não-supervisionada foram usados como ponto de partida para o desenvolvimento de modelos de classificação empregando espectros medidos em instrumentos NIR portáteis. O modelo desenvolvido com espectros medidos no NanoNIR e Máquina de Vetores de Suporte (SVM) alcançou 92,5% de exatidão. Já o modelo empregando espectros medidos no MicroNIR e seleção de variáveis combinada a SVM, apresentou 92,5% de exatidão. Os modelos de calibração desenvolvidos para predição dos teores de areia e argila utilizando-se espectros ATR-FTIR apresentaram resultados semelhantes ao método de regressão dos mínimos quadrados parciais (PLS) e seleção de variáveis por algoritmos genéticos (GA-PLS) e SVM. Por outro lado, os melhores resultados para predição de COS foram obtidos utilizando-se SVM. Por fim, os métodos desenvolvidos e validados são uma alternativa rápida e de baixo custo para caracterização, monitoramento e mapeamento de solos.

Palavras-chave: ATR-FTIR; EDXRF; NIR; mineralogia; ComDim; quimiometria.

ABSTRACT

Soil granulometry and mineralogy interfere with its capacity to store organic carbon and retention of water and nutrient. These parameters vary widely in tropical soils along the landscape. Thus, adequate analytical methods are necessary to obtain information about soils and, consequently, improve the management and preservation of this resource. Spectroscopic and chemometric techniques emerge as excellent alternatives for soil analysis. Considering these aspects, the objective of this work was to develop spectroanalytical methods for qualitative evaluation of particle size and mineralogy, and quantitative determination of sand, clay, and soil organic carbon (COS) contents. A total of 176 soil samples of surface and subsurface from zona da Mata, Agreste and Sertão in the state of Pernambuco were used. The samples were analyzed by spectroscopy in the mid-infrared region with Fourier transform and attenuated total reflectance accessory (ATR-FTIR), near-infrared spectroscopy (NIR) with portable instruments (NanoNIR and MicroNIR), and energy-dispersive X-ray fluorescence spectrometry (EDXRF). The Common Dimension Analysis (ComDim) integrated ATR-FTIR and EDXRF information, evidencing similarities and differences between soil samples, grouping them into five categories associated with differences in particle size and mineral composition. The results of unsupervised discrimination were used as a starting point for the development of classification models using spectra measured in portable NIR instruments. The model developed with spectra measured with NanoNIR and Support Vector Machine (SVM) achieved 92.5% accuracy. The model using spectra measured with MicroNIR and a selection of variables combined with SVM presented 92.5% accuracy. The calibration models developed to predict the sand and clay contents using ATR-FTIR spectra showed similar results with the partial least squares regression method (PLS) and selection of variables by genetic algorithms (GA-PLS), and SVM. On the other hand, the best results for predicting COS were obtained using SVM. Finally, the methods developed and validated are a fast and low-cost alternative for soil characterization, monitoring, and mapping.

Keywords: ATR-FTIR; EDXRF; NIR; mineralogy; ComDim; chemometrics.

LISTA DE FIGURAS

Figura 1 –	Mapa que demonstra os pontos de coleta de amostras de solo.	34
Figura 2 –	Espectrofotômetro FTIR Spectrum Two (Perkin-Elmer) com acessório de refletância total atenuada universal.	36
Figura 3 –	Instrumentos NIR portáteis, em que: NanoNIR (A) e MicroNIR (B).	37
Figura 4 –	Espectrômetro de fluorescência de raios-X por energia dispersiva NEX DE VS.	38
Figura 5 –	Espectros no MIR das amostras de solo obtidos com acessório de refletância total atenuada.	43
Figura 6 –	Espectros brutos no MIR de 400-1200 cm^{-1} (A) e pré-processados por SNV (B), MSC (C), primeira derivada com janela de 7 pontos (D), primeira derivada com janela de 11 pontos (E) e segunda derivada com janela de 7 pontos (F).	44
Figura 7 –	Espectros EDXRF sem pré-processamento nas condições Low-Z (A) e Mid-Z (B).	45
Figura 8 –	Espectros EDXRF após transformação logarítmica e suavização nas condições Low-Z (A) e Mid-Z (B).	45
Figura 9 –	Espectros EDXRF após transformação logarítmica, suavização e correção da linha de base nas condições Low-Z (A), Mid-Z (B) e concatenado (C)	47
Figura 10 –	Espectros Low-Z e Mid-Z autoescalados e concatenados.	47
Figura 11 –	Espectros MIR (A) e EDXRF (B) após normalização e centragem na média.	48
Figura 12 –	Métricas para decisão do número ideal de agrupamentos no PAM: índice de silhueta média (A) e estatística gap (B).	49

Figura 13 –	Gráfico de escores da ComDim, em que: agrupamento 1 (●), agrupamento 2 (●), agrupamento 3 (●), agrupamento 4 (●) e agrupamento 5 (●).	50
Figura 14 –	Gráfico de pesos da ComDim dos espectros no MIR (A, C, E e G) e EDXRF (B, D, F e H).	51
Figura 15 –	Distribuição dos teores de argila, silte e areia, onde letras diferentes indicam diferença significativa ($p<0.05$) de acordo com o teste de Wilcoxon.	53
Figura 16 –	Espectros no NIR sem pré-processamento dos instrumentos NanoNIR (A) e MicroNIR (B).	56
Figura 17 –	Espectros dos instrumentos NanoNIR (A) após correção da linha de base ALS (mínimos quadrados assimétricos) e MicroNIR (B) após primeira derivada.	58
Figura 18 –	Variáveis selecionadas pelo modelo RFE-SVM demonstrado nos espectros médios do conjunto de treinamento usando MicroNIR.	61
Figura 19 –	Mapa de *outliers* da PCA robusta baseado nos espectros MIR.	63
Figura 20 –	Variáveis selecionadas pelo modelo GA-PLS da argila (A), areia (B) e COS (C) demonstrado nos espectros médios do conjunto de calibração.	65
Figura 21 –	Gráficos de dispersão referência vs. preditos para argila, areia, e COS usando PLS, GA-PLS e SVM, em que: calibração (●) e validação externa (●).	67

LISTA DE TABELAS

Tabela 1 – Variação dos teores de argila, silte, areia e COS nas amostras de solos. 35

Tabela 2 – Linhas de emissão identificadas nos espectros EDXRF das amostras de solo. 46

Tabela 3 – Variância explicada e *saliences* da análise ComDim. 49

Tabela 4 – Mineralogia (Min.) das frações argila, silte e areia dos medoides dos agrupamentos, em que: caulinita (Ct), goethita (Gt), hematita (Hm), quartzo (Qz), ilita (Il), esmectita (Es), vermiculita (V), hornblenda (Hb), feldspatos (Fd), micas (Mi) e maghemita (Mh). 51

Tabela 5 – Matriz de separação dos agrupamentos usando o algoritmo PAM, onde em negrito foram destacadas as menores separações entre pares de agrupamentos. 53

Tabela 6 – Matrizes de confusão dos conjuntos de treinamento e teste dos instrumentos NIR portáteis usando PLS-DA. 57

Tabela 7 – Matrizes de confusão dos conjuntos de treinamento e teste dos instrumentos NIR portáteis usando SVM. 59

Tabela 8 – Parâmetros de classificação dos conjuntos de treinamento e teste usando os espectros NIR e SVM. 60

Tabela 9 – Parâmetros de classificação dos conjuntos de treinamento e teste usando instrumento MicroNIR e modelo RFE-SVM. 62

Tabela 10 – Parâmetros de regressão para argila, areia e COS usando espectros MIR e modelos PLS, em que: variáveis latentes (VL), coeficiente de determinação da calibração (R_C^2), coeficiente de determinação da validação-cruzada (R_{CV}^2), coeficiente de determinação da predição (R_P^2), erro quadrático médio da validação cruzada ($RMSE_{CV}$), erro quadrático médio da validação predição ($RMSE_P$), razão 64

entre o desempenho e a distância interquartil da predição ($RPIQ_P$) e bias da predição ($Bias_P$).

Tabela 11 - Parâmetros de regressão para argila, areia e COS usando espectros ATR-FTIR e modelos GA-PLS, em que: variáveis selecionadas (Var. sel), variáveis latentes (VL), coeficiente de determinação da calibração (R_C^2), coeficiente de determinação da validação-cruzada (R_{CV}^2), coeficiente de determinação da predição (R_P^2), erro quadrático médio da validação cruzada ($RMSE_{CV}$), erro quadrático médio da validação predição ($RMSE_P$), razão entre o desempenho e a distância interquartil da predição ($RPIQ_P$) e bias da predição ($Bias_P$). 66

Tabela 12 - Parâmetros de regressão para argila, areia e COS usando espectros ATR-FTIR e modelos SVM, em que: coeficiente de determinação da calibração (R_C^2), coeficiente de determinação da validação-cruzada (R_{CV}^2), coeficiente de determinação da predição (R_P^2), erro quadrático médio da validação cruzada ($RMSE_{CV}$), erro quadrático médio da validação predição ($RMSE_P$), razão entre o desempenho e a distância interquartil da predição ($RPIQ_P$) e bias da predição ($Bias_P$). 66

Tabela 13 - Resultados do teste F (95% de confiança) obtidos na comparação dos modelos PLS como GA-PLS e SVM, em que: não-significativo (ns) e significativo (*). 68

LISTA DE ABREVIATURAS E SIGLAS

ALS	Mínimos quadrados assimétricos
ATR	Refletância total atenuada
ATR-FTIR	Espectroscopia no infravermelho médio por transformada de Fourier com acessório de refletância total atenuada
Au	Augita
C	Hiperparâmetro *cost*
ComDim	Análise de dimensões comuns
COS	Carbono orgânico do solo
Ct	Caulinita
EDXRF	Fluorescência de raios-x por energia dispersiva
Es	Esmectita
Fd	Feldspatos
FN	Falso-negativo
FP	Falso-positivo
FTIR	Espectroscopia de infravermelho por transformada de Fourier
GA	Algoritmos genéticos
Gb	Gibbsita
Gt	Goethita
Hb	Hornblenda
Hm	Hematita
Il	Ilita
Mh	Maghemita
Mi	Micas
MIR	Infravermelho médio
MLR	Regressão linear múltipla
MSC	Correção multiplicativa de sinal
NIR	Infravermelho próximo
OSC	Correção ortogonal de sinal
PAM	Partição ao redor de medoides
PCA	Análise de componentes principais
PLS	Análise de mínimos quadrados parciais

PLS-DA	Análise de mínimos quadrados parciais associada à análise de discriminante
PCR	Regressão por componentes principais
Qz	Quartzo
R^2_C	Coeficiente de determinação da calibração
R^2_{CV}	Coeficiente de determinação da validação cruzada
R^2_P	Coeficiente de determinação da predição
RFE-SVM	Eliminação recursiva de variáveis combinada à máquina de vetores de suporte
$RMSE_C$	Raiz do erro quadrático médio da calibração
$RMSE_{CV}$	Raiz do erro quadrático médio da validação cruzada
$RMSE_P$	Raiz do erro quadrático médio da predição
RPIQ	Razão entre o desempenho e a distância interquartil
SIMCA	Modelagem independente flexível por analogia de classe
SNV	Variação normal padrão
SV	Vetores de suporte
SVM	Máquina de vetores de suporte
TFSA	Terra fina seca ao ar
V	Vermiculita
VL	Variáveis latentes
VN	Verdadeiro-negativo
VP	Verdadeiro-positivo
XRF	Fluorescência de raios-X

SUMÁRIO

1	**INTRODUÇÃO**	**16**
2	**FUNDAMENTAÇÃO TEÓRICA**	**18**
2.1	O SOLO	18
2.1.1	**Composição granulométrica e mineral do solo**	**19**
2.1.2	**Matéria orgânica do solo**	**20**
2.2	TÉCNICAS ESPECTROSCÓPICAS PARA ANÁLISE DIRETA DOS SÓLIDOS DO SOLO	20
2.2.1	**Fluorescência de raios-X por dispersão de energia (EDXRF)**	**21**
2.2.2	**Espectroscopia no infravermelho**	**21**
2.2.2.1	MIR	22
2.2.2.2	NIR	23
2.3	ANÁLISE QUIMIOMÉTRICA DE DADOS ESPECTROSCÓPICOS	24
2.3.1	**Pré-processamento de dados**	**24**
2.3.2	**Análise não-supervisionada**	**26**
2.3.3	**Calibração multivariada**	**28**
2.3.4	**Classificação**	**30**
2.3.5	**Seleção de amostras**	**31**
2.3.6	**Seleção de variáveis**	**32**
2.4	TÉCNICAS ESPECTROSCÓPICAS E QUIMIOMÉTRICAS NA ANÁLISE MINERALÓGICA E DOS TEORES DE ARGILA, AREIA E CARBONO ORGÂNICO DO SOLO	32
3	**MATERIAIS E MÉTODOS**	**34**
3.1	ÁREA DE ESTUDO E SELEÇÃO DE AMOSTRAS	34

3.2 ANÁLISES FÍSICAS E QUÍMICAS 35

3.3 AQUISIÇÃO DOS ESPECTROS 35

3.3.1 MIR 36

3.3.2 NIR 36

3.3.3 EDXRF 37

3.4 ANÁLISE QUIMIOMÉTRICA 38

3.4.1 Pré-processamento dos espectros 38

3.4.2 ComDim 38

3.4.3 Análise de agrupamento 39

3.4.4 Classificação 39

3.4.5 Calibração multivariada 40

4 RESULTADOS E DISCUSSÃO 42

4.1 CARACTERIZAÇÃO ESPECTROSCÓPICA E IDENTIFICAÇÃO DE PADRÃO NAS AMOSTRAS DE SOLO BASEADO NA INTEGRAÇÃO DE INFORMAÇÕES DE ATR-FTIR E EDXRF 42

4.1.1 Espectros no ATR-FTIR 42

4.1.2 Espectros EDXRF 44

4.1.3 Análise ComDim usando espectros no ATR-FTIR e de EDXRF 48

4.1.4 Perfil dos agrupamentos formados 52

4.2 CLASSIFICAÇÃO DE AMOSTRAS ASSOCIADA À COMPOSIÇÃO GRANULOMÉTRICA E MINERAL DE SOLOS, USANDO INSTRUMENTOS NIR PORTÁTEIS 55

4.2.1 Espectros NIR 56

4.2.2 PLS-DA 56

4.2.3 SVM 58

4.3 PREDIÇÃO DOS TEORES DE ARGILA, AREIA E COS USANDO ESPECTROSCOPIA ATR-FTIR 63

4.3.1 Mapa de *outliers* da PCA robusta 63

4.3.2 Modelos de calibração 63

4.3.3 Comparação dos modelos de calibração 67

5 CONCLUSÃO 69

6 PERSPECTIVAS FUTURAS 70

REFERÊNCIAS 71

APÊNDICE A – INFORMAÇÕES SOBRE AS AMOSTRAS 83

1 INTRODUÇÃO

O solo apresenta diversas funções, como auxiliar na sustentação e desenvolvimento de plantas, atuar como um filtro natural, influenciando na qualidade da água e controle da poluição; ademais, trata-se de um importante compartimento para o sequestro de carbono, atenuando a emissão de gases do efeito estufa (JOBBÁGY et al. 2000; LAL, 2002; FAO, 2019).

Os fatores de formação do solo (material de origem, clima, relevo, ação de organismos e tempo) atuam de forma sinérgica ditando o grau de desenvolvimento do solo, impactando em características como a sua composição granulométrica e mineral. Desta forma, os fatores de formação e o uso do solo possuem papel fundamental no armazenamento de carbono orgânico (NAIMAN et al., 2005; ZINN et al., 2007; MAURYA et al., 2020). A composição granulométrica e a mineralogia do solo são indicadores promissores do armazenamento de carbono orgânico do solo (COS), uma vez que influenciam na estabilização da matéria orgânica do solo (MOS) (LÜTZOW et al., 2006; ZINN et al., 2007; WIESMEIER et al., 2019; SANTOS et al., 2020).

Solos tropicais apresentam variações ao longo da paisagem associadas a mineralogia, composição granulométrica e COS (PINCUS et al., 2017). Entretanto, a falta de informações do solo pode levar ao uso e manejo inadequado, causando redução da produtividade e depleção do COS (LAL, 2002; ZINN et al., 2007, AJIBOYE et al., 2019). Logo, é importante obter informações qualitativas e quantitativas dessas variáveis para seu manejo e preservação de forma sustentável.

Os métodos convencionais de laboratório para análises mineralógicas, de composição granulométrica e COS possuem altos custos, requerem muito esforço e podem requerer reagentes perigosos à saúde humana e ambiental. A análise granulométrica consiste na separação e quantificação das proporções das frações areia, silte e argila. Esta análise é comumente realizada em laboratório e envolve as etapas de pré-tratamentos (remoção de sais solúveis e matéria orgânica em excesso), dispersão química e física das amostras de solo (com duração de 16 horas), seguida da separação e quantificação das frações granulométricas (EMBRAPA, 2017).Já a determinação do COS, devido a facilidade de implementação e não requerer uma ótima infraestrutura, muitos laboratórios, em especial de países em desenvolvimento,

produzem resultados para estimar o carbono do solo com base na oxidação de dicromato (McCarthy et al. (2010).

Silva et al. (2019) afirmam que a demanda global por informações detalhadas de solos e avanços tecnológicos estimulam o interesse no desenvolvimento de ferramentas que promovam a rápida obtenção de informações sobre o solo, tanto para fins agronômicos quanto ambientais. Diante deste cenário, é de fundamental importância dispor de metodologias e técnicas analíticas capazes de analisar solos de forma rápida, precisa e segura. Portanto, métodos espectroanalíticos tornam-se bastante úteis, com diversas aplicações e resultados promissores (TÜMSAVAŞ et al., 2019; COBLINSKI et al., 2020; TEIXEIRA et al., (2020); HELFER et al., 2020; SANTOS et al., 2021).

Diante do exposto, o objetivo deste trabalho foi utilizar técnicas espectroscópicas e quimiométricas para avaliação qualitativa e quantitativa de solos tropicais relacionada a mineralogia, composição granulométrica e teor de COS, visando gerar informações para subsídio do seu uso e manejo sustentável.

Para atender a esse objetivo geral foram definidos os seguintes objetivos específicos:

- Extrair informação qualitativa a respeito da química mineral de amostras de solo pela integração das técnicas espectroscópicas ATR-FTIR e EDXRF usando Análise de Dimensões Comuns (ComDim);
- Identificar agrupamentos naturais nas amostras de solo baseado nas espectroscopias ATR-FTIR e EDXRF usando análise de agrupamentos;
- Interpretar a formação dos agrupamentos baseado nos espectros (ATR-FTIR e EDXRF) e composição granulométrica do solo;
- Comparar o desempenho analítico de instrumentos NIR portáteis (NanoNIR e MicroNIR) para discriminação supervisionada de amostras de solo usando PLS-DA e SVM;
- Desenvolver modelos PLS e SVM usando espectroscopia ATR-FTIR para predição dos teores de argila, areia e COS em solos tropicais sob diferentes condições climáticas.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 O SOLO

O solo é um componente vital da Terra, uma vez que ele provê serviços ecossistêmicos, tais como: atuação como filtro de água, fornecimento de nutrientes às plantas, abastecimento dos seres humanos com alimento, fibra e energia. Além disso, o solo age como reservatório e/ou fonte de carbono, regulando a emissão de gases do efeito estufa (ROSSEL et al., 2016).

Sposito (2008) define solos como sistemas biogeoquímicos abertos, multicomponentes que contêm sólidos, líquidos e gases. Já a definição conforme o Sistema Brasileiro de Classificação de Solos (SiBCS) (EMPRAPA, 2018):

> Coleção de corpos naturais, constituídos por partes sólidas, líquidas e gasosas, tridimensionais, dinâmicos, formados por materiais minerais e orgânicos que ocupa a maior parte do manto superficial das extensões continentais do nosso planeta, contém matéria viva e pode ser vegetado na natureza onde ocorre e, eventualmente, ter sido modificado por interferências antrópicas.

Os solos são o resultado da ação sinérgica dos fatores de formação, sendo eles: material de origem, clima, relevo, ação de organismos e tempo (ARAÚJO FILHO et al., 2014). Ademais, os solos apresentam estratificações formando seções aproximadamente paralelas à sua superfície, sendo, portanto, resultante dos processos de formação, chamadas de horizontes (SANTOS et al., 2015). Os horizontes são distinguidos entre si devido às propriedades que são geradas pelos processos de formação de solo (OWENS; RUTLEDGE, 2005). Segundo o SiBCS (EMBRAPA, 2018), “os horizontes refletem os processos de formação do solo a partir do intemperismo do substrato rochoso ou de sedimentos de natureza diversa”.

Os solos brasileiros são classificados por meio de um sistema taxonômico de aperfeiçoamento permanente, o SiBCS, que leva em consideração os atributos diagnósticos do solo para classificá-los. Dentre eles estão a matéria orgânica, o material mineral, a atividade de argila, a saturação por bases, a cor e a mudança textural abrupta. Outros atributos como a cerosidade e a superfície de compressão auxiliam na diferenciação das classes de solo. Os solos são classificados em seis níveis categóricos: ordens, subordens, grandes grupos, subgrupos, famílias e séries. No Brasil, há ocorrência de 13 ordens de solos: Argissolos, Cambissolos,

Chernossolos, Espodossolos, Gleissolos, Latossolos, Luvissolos, Neossolos, Nitossolos, Organossolos, Planossolos, Plintossolos e Vertissolos (EMBRAPA, 2018).

No nordeste do Brasil, é encontrada a maior variedade de tipos de solos, de modo que, em Pernambuco, estão presentes as 13 ordens de solo do Brasil. Desta forma, variações na mineralogia, composição granulométrica e COS são observadas (BRASIL, 1973; JACOMINE, 1996).

2.1.1 Composição granulométrica e mineral do solo

Idealmente, 50% do volume do solo é composto por sólidos. Eles podem ser inorgânicos ou orgânicos, sendo a fração inorgânica representada pelos minerais do solo, e a orgânica, pela matéria orgânica do solo. O mineral se trata de um sólido inorgânico de ocorrência natural, possuindo arranjo atômico ordenado e composição química definida, embora não seja fixa. Também é possível a ocorrência de sólidos inorgânicos amorfos (STRAWN; HINRICH; CONNOR, 2015; SCHULZE, 1989).

Os minerais do solo podem ser divididos em minerais primários e secundários. Segundo Schulze (1989), os minerais primários são aqueles formados sob altas temperaturas e pelo resfriamento do magma. Já os minerais secundários são formados em reações de baixa temperatura quando comparados à de formação dos primários, e podem ser herdados de rochas sedimentares ou formados no solo por ação do intemperismo. Allen e Hajek (1989) citam como exemplos de minerais primários, quartzo (Qz), feldspatos (Fd) (ortoclásio, plagioclásio sódico, plagioclásio cálcico), micas (Mi), hornblenda (Hb), augita (Au); e de minerais secundários, hematita (Hm), goethita (Gt), maghemita (Mh), caulinita (Ct), esmectita (Es), vermiculita (V) e ilita (Il).

Os sólidos inorgânicos do solo são compostos por uma mistura de rochas, cascalhos, areia, silte e argila. No entanto, as rochas e cascalhos não são considerados na avaliação da textura do solo devido ao grande tamanho e por isso pequena área superficial (YU, 1997). Desta forma, a textura do solo é composta pela proporção dos componentes de natureza inorgânica no solo: areia, silte e argila. De forma geral, a textura influencia no suprimento de água e nutrientes às plantas. Segundo EMBRAPA (2017), a areia grossa é composta por partículas de 0,2 a 2 mm de diâmetro, a areia fina de 0,05 a 0,2 mm, o silte de 0,002 a 0,05 mm e a argila são partículas menores que 0,002 mm.

A análise granulométrica consiste na separação e quantificação das proporções das frações areia, silte e argila. Esta análise é comumente realizada em laboratório e envolve as etapas de pré-tratamentos (remoção de sais solúveis e matéria orgânica em excesso), dispersão química e física das amostras de solo, seguida da separação e quantificação das frações granulométricas (EMBRAPA, 2017). Em relação à mineralogia do solo, a areia e o silte tendem a ser compostos, majoritariamente, por minerais primários herdados do material de origem, enquanto a argila é composta majoritariamente por minerais secundários (EMBRAPA, 2018; SCHULZE, 1989). Entretanto, minerais secundários podem ocorrer nas frações de areia e silte, assim como minerais primários podem ocorrer em frações de argila (ALLEN; HAJEK, 1989).

2.1.2 Matéria orgânica do solo

A matéria orgânica é um componente muito importante do solo devido às suas funções biológicas, químicas e físicas. Ela é fonte de energia e de nutrientes, melhora a biodiversidade, atua na sorção de inorgânicos e orgânicos, prevenindo também sua lixiviação e melhorando a capacidade de troca de cátions no solo. Ela também melhora a retenção de água e a estabilidade de agregados, previne a compactação e erosão, além de alterar as propriedades térmicas do solo. Do ponto de vista econômico, a matéria orgânica reduz a adição de fertilizantes e melhora a produtividade e qualidade vegetal (ONDRASEK et al., 2019).

O teor de carbono orgânico total é utilizado para estimar a matéria orgânica do solo (SILVA e MENDONÇA, 2007). Sua determinação por via húmida, muito comum, pelo método Walkley-Black (WALKLEY e BLACK, 1934) gera mais de 1,0 L de resíduo a cada quatro amostras analisadas, estimam Souza et al. (2016). Tendo em vista a importância dos minerais e da matéria orgânica do solo, é de fundamental importância que se possa dispor de metodologias e técnicas analíticas capazes de analisar diretamente a fração sólida dos solos. Neste contexto, os métodos espectroscópicos são uma alternativa bastante viável. Sendo assim, é importante compreender o funcionamento das técnicas espectroscópicas para essa finalidade.

2.2 TÉCNICAS ESPECTROSCÓPICAS PARA ANÁLISE DIRETA DOS SÓLIDOS DO SOLO

De acordo com Gillespie et al. (2015), a compreensão das propriedades físicas, químicas e biológicas do solo só pode ser completamente obtida por meio de um conjunto de métodos analíticos. As técnicas espectroscópicas podem fornecer informações importantes para essa compreensão. Por exemplo, a fluorescência de raios-X (XRF, do inglês *X-ray fluorescence*) fornece a composição elementar do solo, já a espectroscopia no infravermelho gera informações a respeito dos componentes orgânicos e inorgânicos do solo (CONKLIN, 2013).

2.2.1 Fluorescência de raios-X por dispersão de energia (EDXRF)

A XRF é um método popular para análise química que dispensa pré-tratamento complexo das amostras. O método é baseado na interação de radiação raios-X com átomos, de modo que a transição eletrônica em camadas mais internas (K, L e M, por exemplo) produz uma emissão com energia característica (keV) de um elemento particular com alta seletividade (PANCHUK et al., 2018). Os sinais de XRF podem ser registrados por fluorescência de raios-X por dispersão de comprimento de onda e fluorescência de raios-X por energia dispersiva (EDXRF). Neste trabalho, utilizou-se um equipamento EDXRF. Estes equipamentos são mais compactos e rápidos na análise quando comparados aos equipamentos por dispersão de comprimento de onda. No entanto, possuem menor eficiência na detecção de elementos leves (RIBEIRO et al., 2017).

A análise qualitativa em XRF é baseada na identificação de linhas espectrais cuja energia é característica do elemento e, em geral, é uma tarefa simples. Por outro lado, a análise quantitativa dos elementos pode ser bastante complexa devido aos efeitos de matriz da amostra. No solo, a umidade, o tamanho de partícula e o teor de matéria orgânica podem interferir (BASTOS et al., 2012). A determinação dos elementos na amostra pode ser feita de forma semiquantitativa, usando os parâmetros fundamentais sem necessidade de padrões, ou, de forma quantitativa usando padrões (RIBEIRO et al., 2017). O uso de ferramentas quimiométricas combinadas aos espectros XRF não é comum, avaliou Panchuk et al. (2018). No entanto, essa combinação é muito benéfica, uma vez que torna possível avaliações mais complexas, como a respeito da procedência ou autenticidade de amostras.

2.2.2 Espectroscopia no infravermelho

A espectroscopia na região do infravermelho está localizada entre as regiões do visível e das micro-ondas, estando relacionada às transições vibracionais e rotacionais e é subdividida em três regiões: infravermelho próximo (NIR, do inglês *Near-infrared*) que vai de 12800 cm^{-1} a 4000 cm^{-1}; médio (MIR, do inglês *Mid-infrared*) de 4000 cm^{-1} a 200 cm^{-1}; e distante (200 cm^{-1} a 10 cm^{-1}). Quase todos os compostos que possuem ligação covalente, sejam eles orgânicos ou inorgânicos, podem absorver radiação eletromagnética na região do infravermelho, mas é necessário que a frequência radiação absorvida seja a mesma da frequência natural de vibração e que haja variação no momento de dipolo da ligação em função do tempo (ROBINSON; FRAME; FRAME II, 2014). Segundo Silverstein; Webster; Kiemle (2005), os espectros gerados se apresentam como bandas de absorção ao invés de linhas. Isso ocorre porque a mudança de cada nível de energia vibracional corresponde a uma série de mudanças de energia rotacional.

2.2.2.1 MIR

No MIR, os espectros de absorção apresentam picos mais bem definidos que no NIR e estão relacionados às vibrações moleculares fundamentais (MIRABELLA, et al. 2006). Entretanto, os espectrofotômetros no MIR possuem óptica mais complexa e sensível que instrumentos no NIR (PASQUINI, 2003).

Os instrumentos que vêm sendo mais empregados na aquisição de espectros no infravermelho são aqueles que usam a transformada de Fourier em detrimento aos instrumentos dispersivos. A espectroscopia no MIR é versátil e pode ser usada na análise de amostras líquidas, sólidas e gasosas. Além disso, a interação entre a radiação e as amostras pode ser medida por técnicas de transmissão/absorção e por refletância (ROBINSON; FRAME; FRAME II, 2014).

Na análise do solo, as medidas de refletância, como a refletância total atenuada (ATR, do inglês *attenuated total reflectance*) e refletância difusa, além da fotoacústica são preferidas em detrimento às medidas de transmissão (RAPHAEL, 2011). Entretanto, a refletância difusa tem sido mais utilizada nos estudos de solo do que a ATR (GE et al., 2014).

Neste trabalho, utilizou-se medidas de ATR. A ATR é uma forma de medida da espectroscopia de infravermelho que pode ser aplicada em uma grande variedade de

amostras, requerendo pouca ou nenhuma preparação de amostra, além de não ser destrutiva (MILOSEVIC, 2012). Segundo Mirabella (2006), a radiação eletromagnética ao se propagar de um meio de alto índice de refração, n_1, para um meio de baixo índice, n_2, sofre reflexão total interna quando o ângulo incidente é maior que o ângulo crítico (θ_c), definido pela Equação 1.

$$\theta_c = sen^{-1}(\frac{n_2}{n_1}) \qquad \text{(Equação 1)}$$

Na interface entre os dois meios, a reflexão interna cria a onda evanescente que penetra poucos milímetros na amostra (menor índice de refração). Se a amostra absorver a radiação, ela será atenuada nas frequências onde houver absorção. Sendo assim, a profundidade de penetração da onda evanescente é inversamente proporcional ao aumento do comprimento de onda (MIRABELLA, 2006).

2.2.2.2 NIR

Os espectros no NIR são caracterizados pelos sobretons e bandas de combinação entre as vibrações. Os sobretons são múltiplos da frequência fundamental de absorção. Já as bandas de combinação surgem quando duas bandas fundamentais absorvem energia simultaneamente. Essas bandas estão relacionadas a transições proibidas de acordo com o modelo aproximado do oscilador harmônico e estão associadas a grupos funcionais, como OH, CH e NH. (STUART, 2004, MIRABELLA, 2006).

As bandas de absorção no NIR apresentam sobreposição, reduzindo sua aplicabilidade para análises qualitativas. Desta forma, o uso de ferramentas quimiométricas para análise no NIR é bastante comum para o reconhecimento de padrões e determinação de parâmetros/propriedades de interesse (MIRABELLA, 2006; OZAKI et al., 2021).

A espectroscopia na região do NIR é caracterizada pela sua versatilidade, sendo utilizada para análise com pouco ou nenhum pré-tratamento de amostras, e empregada na análise de amostras sólidas, líquidas e gasosas. A técnica também permite a aplicação em laboratório e em campo. Quanto à instrumentação, os

espectrômetros podem ser do tipo dispersivo ou com transformada de Fourier (OZAKI et al., 2021).

2.3 ANÁLISE QUIMIOMÉTRICA DE DADOS ESPECTROSCÓPICOS

O grande volume de dados que podem ser obtidos das técnicas espectroscópicas demanda ferramentas estatísticas eficazes para extração de conhecimento químico relevante (FERREIRA, 2015). Neste contexto, a quimiometria se apresenta como uma ferramenta importante baseada em técnicas matemáticas e estatísticas aplicadas à química (FERREIRA, 2015). Kumar e Sharma (2018) acrescentam que a quimiometria extrai o máximo de informação útil de um conjunto de dados.

Os métodos quimiométricos combinados às técnicas espectroscópicas são ferramentas poderosas para reconhecimento de padrões ou para prever uma determinada propriedade em diferentes amostras (LUNA; GOIS, 2018). Tem sido amplamente demonstrado que os métodos quimiométricos são necessários para extrair informações das técnicas espectroscópicas no geral (ESTEBAN; ARIÑO-BLASCO; DÍAZ-CRUZ, 2020). Para Esbensen e Julius (2009), para que a abordagem quimiométrica seja bem-sucedida é necessário que os dados sejam obtidos por amostragem representativa, com qualidade analítica apropriada, e os modelos quimiométricos devem gerar resultados de validação confiáveis (COCCHI, 2019). Simeonov et al. (2003), destacam a validade da quimiometria na resolução de questões ambientais, já que essa ferramenta evita erros de interpretação no monitoramento de dados ambientais volumosos e complexos.

Na quimiometria, podem ser obtidas informações relevantes de dados espectroscópicos por meio de análise exploratória (reconhecimento de padrões não-supervisionado), classificação (reconhecimento de padrões supervisionado) e calibração multivariada.

2.3.1 Pré-processamento dos dados

O pré-processamento de dados espectrais visa remover ou reduzir variações irrelevantes que podem ser específicas da técnica espectroscópica utilizada. Desta forma, é essencial para cumprir melhor o objetivo da análise de dados (ENGEL et al.,

2013). Neste trabalho foram avaliadas técnicas de pré-processamento para correção da linha de base, espalhamento da radiação, ruídos e escala.

A correção de linha de base é utilizada para remover deslocamento vertical e/ou inclinação. Para isso, pode ser utilizada a derivada dos espectros. A primeira derivada eliminará o deslocamento vertical constante, enquanto a segunda derivada eliminará o deslocamento vertical e a inclinação da linha de base (FERREIRA, 2015). A correção por mínimos quadrados assimétricos (ALS, do inglês *asymmetric least squares*) usa regressão de segunda derivada ponderada (ENGEL et al., 2013).

As técnicas de variação normal padrão (SNV, do inglês *standard normal variate*) e correção multiplicativa de sinal (MSC, do inglês *multiplicative signal correction*) são comumente usadas para correção do espalhamento da radiação. Na SNV, a média dos espectros ($\bar{x}$) é subtraída de cada espectro (x_i) e, posteriormente, ele é normalizado pelo desvio-padrão (s_i), conforme Equação 2.

$$SNV_i = \frac{x_i - \bar{x}}{s_i} \qquad \text{(Equação 2)}$$

Enquanto na MSC, é estimado o coeficiente que descreve a dispersão usando o espectro médio ou um espectro de referência (x_r) (FERREIRA, 2015; ENGEL et al., 2013), conforme as Equações 3 e 4:

$$x_i = m_i x_r + ai \qquad \text{(Equação 3)}$$

$$MSC_i = \frac{a_i - x_i}{m_i} \qquad \text{(Equação 4)}$$

Em que, a_i e m_i são as estimativas dos efeitos aditivos e multiplicativos.

Os métodos de suavização são utilizados para melhorar a razão sinal/ruído. Entre esses métodos, destaca-se o filtro de Savitzky-Golay (GORRY, 1990). O algoritmo é baseado no ajuste de um polinômio de baixa ordem em janelas de dados pequenas usando o método de mínimos quadrados (FERREIRA, 2015; ENGEL et al., 2013).

No caso das variáveis, quando estas são expressas em diferentes escalas, faz-se necessário o escalonamento. Desta forma, todas as variáveis terão peso igual. No entanto, os ruídos terão a mesma influência nos modelos que as variáveis com informação útil (TAYLOR; FRANCIS, 2006). O auto-escalonamento (Equação 5) é

bastante utilizado para isso. Tal método consiste na subtração da média, seguida da divisão pelo desvio-padrão de cada coluna (FERREIRA, 2015; ENGEL et al., 2013).

$$Xs_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j} \qquad \text{(Equação 5)}$$

Em que, Xs é a matriz auto-escalada, x_{ij} é o elemento i da variável j, $\bar{x}_j$ e s_j são a média e o desvio padrão da variável j.

De modo alternativo, podem ser utilizadas transformações dos dados originais, como a transformação logarítmica. Essa transformação reduz a importância de valores altos e gera um “pseudo-escalonamento” (ENGEL et al., 2013).

$$Xlog_{ij} = log_a(x_{ij}) \qquad \text{(Equação 6)}$$

Onde, Xlog é a matriz após transformação logarítmica, *a* é a base, x_{ij} é o elemento i da variável j.

A correção ortogonal de sinal (OSC, do inglês *orthogonal signal correction*) é um método supervisionado de pré-tratamento de dados proposto por Wold et al. (1998). Esse método visa a remoção de variação nos espectros que não são relacionadas à variável dependente usando decomposição baseada em PLS, desta forma, podem ser tratados simultaneamente a linha de base e o espalhamento (ENGEL et al., 2013).

2.3.2 Análise não-supervisionada

Os métodos não-supervisionados não necessitam de nenhuma informação adicional sobre as amostras para realização da inspeção dos dados. Desta forma, visando extrair informação relevante dos espectros, duas abordagens não-supervisionadas comuns são: redução de dimensão e análise de agrupamentos (OLIVERI et al., 2021).

O método mais comum de redução de dimensão é a Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*). A ideia principal da PCA é aproximar a matriz de dados originais do produto de duas matrizes menores, a matriz de escores e a matriz de pesos. As variáveis originais são reduzidas por meio de

combinações lineares (das componentes principais) de forma a agrupar aquelas variáveis originais que apresentam informações semelhantes. As componentes principais são determinadas com base no critério de variância máxima e são ortogonais entre si. A maior parte da variância está contida na primeira componente seguida da segunda e assim sucessivamente. Portanto, em geral, as primeiras componentes são as mais úteis nos estudos exploratórios (FERREIRA, 2015). A Equação 7 demonstra o cálculo da PCA:

$$X = TP^t + E \qquad \text{(Equação 7)}$$

Conforme a Equação 7, a matriz de dados originais (**X**) pode ser decomposta em uma matriz de escores (**T**) e na matriz transposta de pesos ($\mathbf{P^t}$). Tendo em vista que a PCA fornece uma aproximação dos dados originais, os resíduos estão contidos na matriz **E**. Desta forma, os escores representam um novo conjunto de variáveis sintéticas das amostras, ao passo que, a matriz de pesos representa os pesos de cada variável (BIANCOLILLO e MARINI, 2018).

A PCA pode ser utilizada para identificação de possíveis amostras anômalas no conjunto de dados (SIRVEN et al., 2007, GENISHEVA et al., 2018, MOHAMMADI et al., 2020). No entanto, os resultados da PCA são sensíveis a essas amostras anômalas. Para superar este problema e identificar esse tipo de amostras, Hubert, Rousseeuw, Vanden Branden (2005) propuseram a PCA robusta. Este método é resistente a amostras anômalas nos dados e gera um mapa baseado nas distâncias de escore e ortogonal calculadas para cada observação, sendo útil para identificação de amostras anômalas.

A análise de agrupamentos é usada para identificar a presença de estrutura no conjunto de dados. Desta forma, não é possível obter relações entre os agrupamentos ou interpretação química detalhada dos agrupamentos usando somente esse tipo de análise. Esses métodos são divididos em hierárquicos e não-hierárquicos.

A análise de agrupamentos hierárquicos é uma técnica que busca representar graficamente o agrupamento hierárquico de dados multidimensionais. Essa técnica reúne amostras semelhantes de acordo com as variáveis utilizadas e complementa a PCA na busca por diferenças e semelhanças entre as amostras. Pode ser efetuada a PCA nos dados originais de modo a reduzi-los a um menor número de variáveis e utilizar os escores extraídos como variáveis que serão usadas na análise de

agrupamentos hierárquicos. Segundo Kumar e Sharma (2018), esse procedimento promove os melhores resultados quando os dados originais contêm muitas variáveis (FERREIRA, 2015; ADAMS, 2004).

Por outro lado, os métodos não-hierárquicos dividem o conjunto de dados em agrupamentos de forma simultânea sem a formação de estruturas hierárquicas (XIAO; YU, 2012). Estão incluídas as técnicas: k-means, partição ao redor de medoides (PAM, do inglês *partitioning around medoids*), OPTICS e DBSCAN. Esses métodos requerem a definição do número de agrupamentos inicial. Isso pode ser obtido com auxílio de métricas como índice de silhueta e estatística gap. A redução de variáveis pode ser útil para melhoria do desempenho das técnicas de agrupamento (OLIVERI et al., 2021).

Neste trabalho, utilizou-se o algoritmo PAM para identificação de agrupamentos. Este algoritmo baseia-se na busca por k objetos representativos (medoides) que minimizem a soma das diferenças das observações em relação ao seu medoide mais próximo entre as observações do conjunto de dados. Depois de identificar um conjunto de k medoides, os k agrupamentos são construídos associando cada observação ao medoide mais próximo (KAUFMAN e ROUSSEEUW, 2009).

Para análise exploratória quando há mais de uma técnica analítica (blocos) que descrevem o mesmo conjunto amostral, podem ser utilizados os métodos multibloco, com destaque para a análise ComDim. A ComDim permite a extração de componentes globais e os pesos específicos, para cada bloco, para se ter uma percepção da importância de cada uma por meio das *saliences*. Para isso, os componentes globais são extraídos sequencialmente, maximizando a variância da soma ponderada dos produtos cruzados das matrizes. Em seguida, os blocos individuais são decompostos nos componentes globais. Por fim, todo o procedimento é repetido nos blocos decompostos (MISHRA et al., 2021). Portanto, a análise não-supervisionada é adequada para a exploração dos dados e não para predição. Para predição de variáveis dependentes, é mais adequado utilizar métodos de calibração multivariada e classificação.

2.3.3 Calibração multivariada

Os métodos de regressão têm como objetivo relacionar variáveis independentes medidas a uma variável dependente quantitativa, de modo a

desenvolver um modelo matemático adequado para predição. No entanto, nem sempre modelos lineares e univariados são adequados para este fim. Desta forma, são comuns casos de regressão multivariada em que muitas variáveis independentes contribuem para a predição da variável dependente (ADAMS, 2004).

Diversos métodos de regressão multivariada podem ser utilizados para encontrar uma relação entre os espectros e as variáveis dependentes quantitativas, como a concentração de analitos ou propriedades da amostra (FERREIRA, 2015). Esse grupo de métodos pode ser dividido em métodos lineares e não-lineares. Entre os métodos lineares, encontram-se a Regressão Linear Múltipla (MLR, do inglês *Multiple Linear Regression*), Regressão por Componentes Principais (PCR, do inglês *Principal Component Regression*) e Regressão por Mínimos Quadrados Parciais (PLS, do inglês *Partial Least Squares Regression*). Por outro lado, exemplos de métodos não-lineares são: Máquina de Vetores de Suporte (SVM, do inglês *Support Vector Machine*), Floresta Aleatória e as Redes Neurais Artificiais.

O método PLS é um método que busca por relações lineares entre as variáveis independentes (**X**) e o vetor ou matriz de variáveis dependentes (Y), que contém a propriedade de interesse. Para isso, a manipulação dos dados é baseada em PCA. Tanto a matriz **X** quanto o vetor/matriz **Y** são utilizados para decomposição de dados nas matrizes de escores e pesos, conforme representado nas Equações 8 e 9 (ADAMS, 2004).

$$X = TP^t + E \qquad \text{(Equação 8)}$$

$$Y = UQ^t + F \qquad \text{(Equação 9)}$$

Em que, **T** e **U** correspondem aos escores das matrizes **X** e **Y**, respectivamente. Bem como, **P** e **Q** são os pesos das matrizes **X** e **Y**, respectivamente. Por fim, **E** é o resíduo da matriz **X** e **F** é o resíduo da matriz **Y**.

As variáveis latentes são calculadas a partir de combinações lineares das variáveis originais para predição da propriedade de interesse. Dessa forma, cada variável latente relaciona a matriz **X** à propriedade de interesse **y**. Ela é obtida maximizando a covariância entre os escores das matrizes **X** e **Y** (**T** e **U**) (FERREIRA, 2015).

O SVM é um método de aprendizado de máquina utilizado em tarefas de predição e classificação. Ele se baseia na construção de planos que definem os limites

de decisão. SVM cria um hiperplano usando um modelo linear para implementar limites de classe não lineares usando uma função *kernel* para mapear o conjunto de dados de entrada em um espaço multidimensional. Os vetores de suporte (SV, do inglês *support vectors*) são amostras do conjunto de calibração que se situam próximo ou no limite do hiperplano. Desta forma, os SVs influenciam na posição e orientação do hiperplano (DERIS, ZAIN, SALLEHUDDIN, 2011; GHOLAMI e FAKHARI, 2017).

Seja qual for o método de regressão utilizado é importante que os modelos estejam bem ajustados e que sejam validados. Para isso, utiliza-se a validação interna (ou cruzada) e externa.

A validação cruzada é um procedimento de reamostragem, que se baseia no particionamento dos dados de treino, seguido do reajuste dos modelos para cada subamostra. Este procedimento é essencial para a prevenção do sobreajuste, já que é utilizado para seleção do número de variáveis latentes no PLS e hiperparâmetros no SVM (BERRAR, 2019; FERREIRA, 2015). Entre os métodos de validação cruzada mais utilizados estão *Leave-one-out* e *k-fold*. Neste trabalho, utilizou-se o k-fold. O método divide o conjunto de calibração em K partes iguais, em seguida, o modelo é ajustado usando as K−1 partes, e a parte não usada na calibração é destinada à validação. Esse processo é repetido K vezes (em cada momento uma parcela diferente será validada), por fim, a média dos erros é calculada (BERRAR, 2019).

A validação externa é utilizada para avaliar as estimativas da variável dependente em amostras que não participaram da construção do modelo. Logo, esse procedimento é importante para análise da capacidade preditiva do modelo em amostras desconhecidas (FERREIRA, 2005). É importante que a faixa de variação da propriedade de interesse na validação externa não extrapole a faixa de variação do conjunto de calibração.

2.3.4 Classificação

Métodos de classificação são utilizados para atribuir uma classe pré-definida a uma amostra. Duas abordagens podem ser utilizadas para classificação de amostras: métodos discriminantes e de modelagem de classes. Os métodos discriminantes dividem o espaço multivariado definido pelas variáveis preditoras em tantos subespaços quantas foram as classes pré-definidas. Dessa forma, pelo menos duas classes devem ser pré-definidas para estabelecer a função discriminante. Por outro

lado, os métodos de modelagem de classes visam caracterizar uma classe de interesse, estabelecendo um espaço definido pela faixa de variabilidade das variáveis preditoras ao redor da classe de interesse em um determinado nível de confiança. Desta forma, fazem parte dos métodos discriminatórios a Análise Discriminante Linear, Análise de Mínimos Quadrados Parciais associada à Análise Discriminante (PLS-DA, do inglês *partial least squares with discriminant analysis*), K-ésimo vizinho mais próximo e SVM. Além desses, tem-se a Modelagem Independente Flexível por Analogia de Classe (SIMCA, do inglês *Soft Independent Modelling by Class Analogy*), que é o método de modelagem de classe mais comumente utilizado (OLIVIERI et al., 2021).

PLS-DA é um método linear de classificação caracterizado pela sua flexibilidade, lidando bem em situações em que o número de variáveis supera em muito as amostras. Este método usa o algoritmo PLS, mas as variáveis dependentes SÃO um rótulo para a classe à qual um objeto pertence (BRERETON, 2009). Para classificação, o método SVM é treinado para separar as classes no conjunto de treinamento por meio de uma superfície que maximize a margem entre elas (CERVANTES et al., 2020).

Os modelos de classificação podem ser validados usando validação cruzada e os outros métodos empregados na calibração multivariada. No entanto, ao invés do cálculo do erro é calculada a média da exatidão. Também é importante a validação externa para avaliação da qualidade dos modelos de classificação em amostras que não participaram da etapa de treinamento.

2.3.5 Seleção de amostras

A análise supervisionada dos dados conta com as etapas de calibração e validação. Para calibração, as amostras são divididas nos conjuntos de calibração e validação, de modo que o conjunto de validação não exerça nenhuma influência na calibração do modelo. A seleção de amostras para composição desses conjuntos pode ser feita de forma aleatória ou empregando algum algoritmo de seleção.

Entre os algoritmos mais comuns de seleção, tem-se o de Kennard-Stone. Este é um algoritmo que se baseia na distância euclidiana dos espectros (KENNARD; STONE, 1969). Por outro lado, a seleção de variáveis proposta por Næs (1987) combina PCA e k-means para seleção de amostras em conjunto de dados

multivariados. Este algoritmo consiste em reduzir a dimensionalidade dos dados usando PCA e decidir quantas componentes reter; depois aplicar k-means aos escores da PCA; em seguida, selecionar amostras que estejam mais próximas ou distantes do centro do agrupamento de acordo com o número de amostras desejadas.

2.3.6 Seleção de variáveis

Tanto nos métodos de reconhecimento de padrões quanto na calibração multivariada, a seleção de variáveis é uma etapa importante para otimização dos modelos, devido à remoção de variáveis espectrais que apresentam baixa contribuição para validação/predição.

Neste trabalho, utilizou-se a eliminação recursiva de variáveis (RFE, do inglês *recursive feature elimination*) para otimização dos modelos de classificação SVM. Este algoritmo realiza uma seleção por retro-eliminação (do inglês, *backwards elimination*). Para isso, inicialmente, o modelo SVM é treinado com todas as variáveis espectrais, então, as variáveis são ranqueadas de acordo com a sua importância para o modelo. Em seguida, valores candidatos ao número de preditores são avaliados e de acordo com seu desempenho são retidos para o modelo final (KUHN, 2019).

Para otimização dos modelos de regressão PLS, avaliou-se a seleção de variáveis por algoritmos genéticos (GA-PLS) neste trabalho. Este algoritmo busca um subconjunto de variáveis preditoras e é inspirado na teoria da evolução biológica e na seleção natural. O ajuste e avaliação do conjunto de variáveis selecionadas é feito usando modelo PLS e validação cruzada respectivamente. Ao completar o algoritmo de busca, o melhor conjunto de variáveis em termos de desempenho é selecionado (MEHMOOD et al., 2020).

2.4 TÉCNICAS ESPECTROSCÓPICAS E QUIMIOMÉTRICAS NA ANÁLISE MINERALÓGICA E DOS TEORES DE ARGILA, AREIA E CARBONO ORGÂNICO DO SOLO

O uso de técnicas espectroscópicas para análises qualitativas é menos comum que para predição de parâmetros de avaliação de solos. Terra et al. (2018) verificaram que espectros na região do visível e do NIR permitem a discriminação do solo com base no intemperismo. A análise qualitativa de Guimarães et al. (2021) indicou que os

espectros no MIR e de XRF carregam informação a respeito de minerais e química do solo, refletindo o seu intemperismo. As matrizes de dados individuais e concatenadas foram analisadas por PCA.

As espectroscopias MIR e NIR são ferramentas úteis para caracterização do solo. Zhang et al. (2021) observaram que características como matéria orgânica e minerais nos espectros MIR foram fundamentais para caracterizar e classificar o solo. Tendo em vista que a espectroscopia NIR possui informações úteis sobre a química dos minerais (MEER, 2018), Coblinski et al. (2021) observaram os efeitos da textura do solo na identificação de minerais.

A biblioteca espectral em escala nacional de Mendes et al. (2022) demonstrou a aplicabilidade para avaliações qualitativas e quantitativas do solo usando espectros MIR obtidos por refletância difusa, apresentando resultados consistentes para predição dos teores de argila e COS. Tavares et al. (2020) verificaram que a determinação do teor de argila, individualmente, foi melhor predita utilizando modelo com espectros Visível-NIR que aqueles medidos XRF.

Terra et al. (2015) utilizaram PLS combinado a espectros de refletância difusa NIR e MIR e observaram que os modelos MIR apresentaram melhor performance para a maioria dos atributos preditos, entre eles argila, areia e COS. De forma semelhante, no trabalho de Santos et al. (2020) com solos do nordeste do Brasil, os autores observaram que os modelos desenvolvidos usando SVM apresentaram desempenho inferiores aos modelos PLS para determinação do COS, com destaque para os espectros no MIR. A razão do desvio do desempenho e coeficiente de determinação da predição (R^2_P) foram melhores no modelo utilizando os dados espectrais no Visível-NIR (R^2_P = 0,86; razão = 2,61) que os dados espectrais de XRF (R^2_P = 0,83; razão = 1,82). Apesar do desempenho inferior observado pelos últimos autores, a espectroscopia XRF tem se apresentado como uma ferramenta promissora para determinação multivariada do COS. Tal determinação, pode ser realizada em até três minutos em comparação às duas horas do método de referência por via úmida, estimam Morona et. al (2017).

Verifica-se por meio dos trabalhos citados, que há maior interesse no uso de técnicas espectroscópicas no NIR, MIR e XRF para predição de parâmetros em comparação a análise qualitativa. Ademais, o uso de métodos multibloco para extração de informação global de técnicas espectroscópicas ainda é incipiente, demonstrando assim, a relevância do presente trabalho.

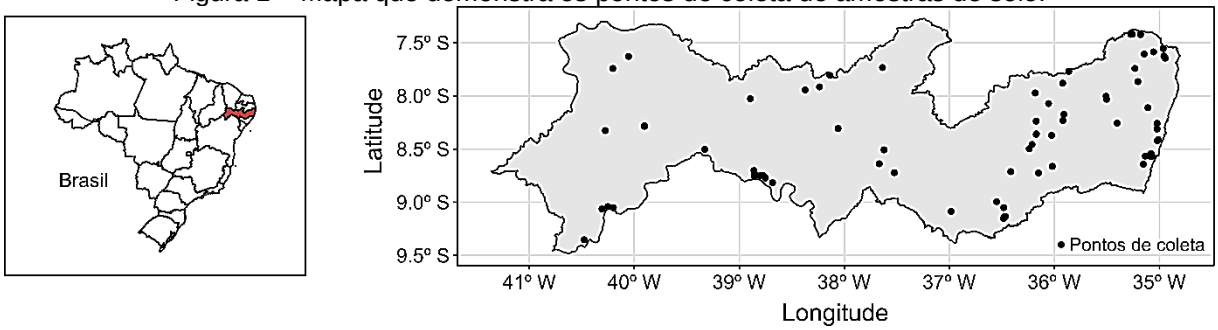
3 MATERIAIS E MÉTODOS

3.1 ÁREA DE ESTUDO E SELEÇÃO DAS AMOSTRAS

O estudo foi realizado com amostras de solo do Estado de Pernambuco, nordeste do Brasil. Pernambuco tem cerca de 98.281 km^2 e apresenta uma ampla variedade climática e geomorfológica (SILVA, 2004; ARAÚJO FILHO et al., 2014; Pernambuco, 2021), sendo dividido em três regiões fisiográficas: Zona da Mata, região mais úmida e com área de aproximadamente 11.776 km^2; Agreste, região de transição entre a Zona da Mata e o Sertão, que possui cerca de 17.970 km^2; e o Sertão, região mais seca e de maior extensão com cerca de 68.535 km^2 (BRASIL, 1973; NASCIMENTO et al., 2006). As ordens de solos que se destacam no estado são: Latossolos, Argissolos, Planossolos, Luvissolos e Neossolos. São encontrados em baixas proporções: Nitossolos, Chernossolos, Cambissolos, Espodossolos, Vertissolos, Gleissolos e Plintossolos (ARAÚJO FILHO et al., 2014).

Foram selecionadas amostras de horizontes superficiais e subsuperficiais de 88 sítios georreferenciados usando um dispositivo de Sistema de Posicionamento Global (GPS) (Figura 1), totalizando 176 amostras de solo. Este estudo compreendeu 12 das 13 ordens de solos do Sistema Brasileiro de Classificação de Solos, excetuando a ordem Organossolo (EMBRAPA, 2018). Informações auxiliares como a classificação dos solos até o segundo nível categórico, município e coordenadas do perfil de solo podem ser consultadas no Apêndice A.

Figura 1 – Mapa que demonstra os pontos de coleta de amostras de solo.



Fonte: o autor (2022).

As amostras selecionadas fazem parte do banco de solos do Departamento de Agronomia da Universidade Federal Rural de Pernambuco e foram coletadas

conforme Manual de Descrição e Coleta de Solo no Campo (LEMOS e SANTOS, 1996; SANTOS et al., 2015). É importante ressaltar que todas as amostras foram fornecidas na forma de terra fina seca ao ar (TFSA), isto é, foram secas ao ar, destorroadas e tamisadas em peneira cuja abertura foi de 2 mm.

3.2 ANÁLISES FÍSICAS E QUÍMICAS

Os resultados da composição granulométrica da TFSA foram fornecidos por pesquisas realizadas anteriormente (RIBEIRO, 1999; COSTA, 2012; LEMOS, 2013; ARAÚJO, 2014; BRILHANTE, 2014; LIMA, 2014; NEVES, 2014; SOUSA, 2015; CÂMARA, 2016; SILVA, 2016; HLAVANGUANE, 2017; BRILHANTE, 2018; NEVES, 2018; SÁ, 2018; SILVA, 2018; SOUSA, 2018; SANTANA, 2019; SOUSA, 2019; NASCIMENTO, 2020). Foi utilizado o método do hidrômetro com solução de calgon 0,1 mol L^{-1} como dispersante (EMBRAPA, 1997; EMBRAPA, 2009).

O teor de COS foi determinado baseado no método de Walkley-Black (Walkley e Black, 1934) descrito por EMBRAPA, 2017.

Ampla variação dos teores de argila, silte, areia e COS foi encontrada, conforme a Tabela 1.

Tabela 1 – Variação dos teores de argila, silte, areia e COS nas amostras de solos.

Parâmetro	**Argila (%)**	**Silte (%)**	**Areia (%)**	**COS (gC kg^{-1})**
Mínimo	4,0	0,4	4,7	0,57
Máximo	77,9	57,1	92,4	40,87
Mediana	45,75	16,95	34,90	6,81
Desvio-padrão	19,44	11,39	23,01	7,93

Fonte: Ribeiro (1999); Costa (2012); Lemos (2013); Araújo (2014); Brilhante (2014); Lima (2014); Neves (2014); Sousa (2015); Câmara (2016); Silva (2016); Hlavanguane (2017); Brilhante (2018); Neves (2018); Sá (2018); Silva (2018); Sousa (2018); Santana (2019); Sousa (2019); Nascimento (2020), o autor (2022).

3.3 AQUISIÇÃO DOS ESPECTROS

Para obtenção dos espectros, uma fração da TFSA foi macerada em almofariz de ágata e, em seguida, tamisadas em peneira de 65 mesh (0,212 mm).

3.3.1 MIR

Os espectros no MIR das amostras de solo foram obtidos utilizando espectrômetro de infravermelho com transformada de Fourier modelo Spectrum Two (Perkin Elmer), no laboratório de Mineralogia do Solo do Departamento de Agronomia da Universidade Federal Rural de Pernambuco. Para tanto, utilizou-se o acessório de Refletância Total Atenuada (ATR, do inglês *Attenuated Total Reflectance*) com reflexão única e um cristal de diamante (Figura 2). Tal acessório foi utilizado para aquisição dos espectros (ângulo de incidência de 45°) e a faixa espectral selecionada foi de 4000 cm^{-1} a 400 cm^{-1}. Foram feitas 32 varreduras para cada espectro com resolução de 4 cm^{-1}. O branco foi feito utilizando a atmosfera sem amostra pelo próprio equipamento. A amostra macerada e peneirada foi pressionada contra o cristal a uma força constante de 135 N. As amostras foram analisadas em triplicata e a média dos três espectros foi utilizada na análise quimiométrica. A limpeza do cristal do ATR foi feita com acetona P.A (marca: Química Moderna) entre análises.

Figura 2 – Espectrofotômetro FTIR Spectrum Two (Perkin-Elmer) com acessório de refletância total atenuada universal.

Fonte: Perkin-Elmer (2022).

3.3.2 NIR

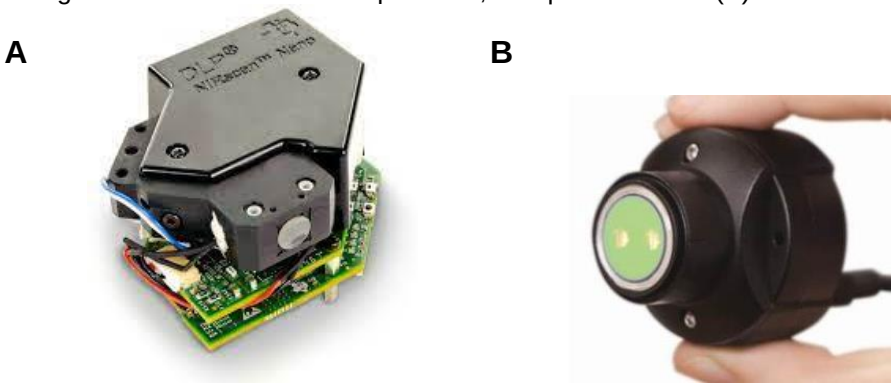
Dois equipamentos portáteis diferentes foram usados para aquisição dos espectros NIR das amostras de solo. As amostras foram dispostas em placas de petri de vidro para análise. Cada espectro foi obtido em triplicata, assim, a média das aquisições foi utilizada para análise quimiométrica.

O primeiro instrumento NIR portátil foi o DLP NIRscan Nano EVM (Texas Instruments), denominado NanoNIR (Figura 3A), que possui arquitetura pós-

dispersiva e opera na faixa espectral de 900-1700 nm. O equipamento possui ainda duas lâmpadas de filamento de tungstênio, grade de difração e detector de arranjo de diodos de InGaAs. Os espectros neste equipamento foram obtidos no modo Hadamard com média de 15 varreduras e resolução de 7 nm.

O segundo equipamento NIR portátil utilizado foi o MicroNIR 1700 (VIAVI Solutions Inc.), denominado MicroNIR (Figura 3B). Este instrumento opera na faixa espectral de 908 -1676 nm, conta com duas lâmpadas integradas de tungstênio sob vácuo, filtro linear variável (LVF, do inglês linear *variable filter*) e detector de InGaAs. Os espectros foram adquiridos com resolução < 1,25% do centro de comprimento de onda (cerca de 12,5 nm) e média de 50 varreduras, com tempo de integração de 10 ms.

Figura 3 – Instrumentos NIR portáteis, em que: NanoNIR (A) e MicroNIR (B).



Fonte: Texas Intruments (2022) e VIAVI Solutions (2022).

3.3.3 EDXRF

Os espectros de fluorescência de raios-X das amostras de solo foram obtidos utilizando EDXRF modelo NEX DE (Rigaku), no LabRefino-Lateclim localizado no LITPEG, Universidade Federal de Pernambuco. O equipamento conta com um tubo cuja máxima voltagem é 60 kV, corrente máxima de 1 mA e potência máxima de 12 W, com carrosel auto-amostrador para 10 amostras (Figura 4).

As análises foram realizadas sob ar atmosférico, em duas condições: Low-Z e Mid-Z. Sob condição Low-Z, a voltagem foi de 6,5 kV, corrente do tubo automática e tempo de medida de 100 s. Já sob condição Mid-Z, a voltagem foi de 35 kV, corrente do tubo automática e tempo de medida de 100 s.

Figura 4 – Espectrômetro de fluorescência de raio-X por energia dispersiva NEX DE VS.

Fonte: Rigaku (2022).

3.4 ANÁLISE QUIMIOMÉTRICA

3.4.1 Pré-processamento dos espectros

O pré-processamento, bem como demais tratamentos dos dados, foi realizado utilizando-se o *software* R. Foram avaliadas correções para o espalhamento de radiação, tais como SNV e MSC, usando os pacotes “prospectr” (STEVENS e RAMIREZ-LOPEZ, 2014) e “pls” (WEHRENS e MEVIK, 2007), respectivamente. Para correção da linha de base, avaliou-se o algoritmo de correção ALS e a primeira e segunda derivadas usando filtro de Savitzky-Golay (com diferentes tamanhos de janelas), usando o pacote “mdatools” (KUCHERYAVSKIY, 2020). A melhoria da relação sinal/ruído com a suavização usando filtro de Savitzky-Golay foi obtida também com o pacote “mdatools”. Avaliou-se também a transformação logarítmica para os espectros EDXRF e OSC, usando o pacote “mt” (LIN, 2022), para o desenvolvimento de modelos supervisionados.

3.4.2 ComDim

A análise ComDim é um método de reconhecimento de padrão não-supervisionado para avaliação global de matrizes de dados provenientes de diferentes técnicas analíticas para o mesmo conjunto de amostras. Para isso, utilizou-se o pacote “MBAnalysis” (MANGAMANA et al., 2021). O pré-processamento dos espectros ATR-

FTIR e EDXRF escolhido foi aquele que melhorou a formação de agrupamentos de acordo com a Estatística de Hopkins, usando o pacote “factoextra” (KASSAMBARA e MUNDT, 2020).

3.4.3 Análise de agrupamento

Para identificação de agrupamentos naturais, utilizou-se os escores da ComDim e o algoritmo de Partição ao Redor de Medoides (PAM, do inglês P*artitioning Around Medoids*) usando o pacote “cluster” (MAECHLER et al., 2022). O algoritmo PAM foi escolhido devido à sua flexibilidade e robustez (REYNOLDS et al., 2006; XIAO e YU, 2012). A decisão do número de agrupamentos ideais foi obtida baseada no índice médio da silhueta e estatística *gap*, usando o pacote “factoextra” (KASSAMBARA e MUNDT, 2020). Em caso de divergência entre as métricas, o índice de Dunn será utilizado com base no pacote “fpc” (HENNING, 2020). Para interpretação dos agrupamentos, avaliou-se os espectros ATR-FTIR e EDXRF, além da variação teores de areia, silte e argila.

3.4.4 Classificação

Para discriminação supervisionada das amostras de solo, as categorias de solos discriminadas pela análise de agrupamentos foram utilizadas como classes para o desenvolvimento de modelos de classificação. Em seguida, os espectros NIR sem pré-processamento são divididos nos conjuntos de treinamento (70% das amostras) e teste (30% das amostras) com auxílio do algoritmo de amostragem naes, proposto por Næs (1987) no pacote “prospectr” (STEVENS e RAMIREZ-LOPEZ, 2014). Foram avaliados os métodos PLS-DA e SVM usando o pacote “caret” (KUHN et al., 2022). O procedimento de validação cruzada usando o método *k-fold* (k = 10) foi utilizado para seleção do melhor número de variáveis latentes (VLs) no PLS-DA, seleção do *kernel* (linear, radial, polinomial ou sigmoide) e otimização de hiperparâmetro no SVM. A seleção do melhor pré-processamento deu-se pela análise da exatidão da validação cruzada e do conjunto de teste. Avaliou-se também a eliminação recursiva de variáveis combinada ao SVM (RFE-SVM) para otimização dos modelos SVM.

A eficiência da classificação foi avaliada com auxílio dos parâmetros de classificação sensibilidade, especificidade, precisão e acurácia, como demonstrado nas Equações 10-13:

$$Sensibilidade = \frac{VP}{(VP+FN)} \quad \text{(Equação 10)}$$

$$Especificidade = \frac{VN}{(VN+FP)} \quad \text{(Equação 11)}$$

$$Precisão = \frac{VP}{(VP+FP)} \quad \text{(Equação 12)}$$

$$Acurácia = \frac{VP+VN}{(VP+VN+FP+FN)} \quad \text{(Equação 13)}$$

Onde, VP, VN, FP e FN são o número de amostras verdadeiro-positivas, verdadeiro-negativas, falso-positivas e falso-negativas, respectivamente.

3.4.5 Calibração multivariada

Inicialmente, foi avaliada a ocorrência de amostras anômalas nos espectros ATR-FTIR usando o mapa de *outliers* da PCA robusta do pacote “rospca” (REYNKENS et al., 2018). Para determinação multivariada das medidas de argila, areia e COS, as 176 amostras de solo foram divididas usando os espectros sem pré-processamento nos conjuntos de calibração (70% das amostras) e validação externa (30% das amostras) com auxílio do algoritmo de amostragem naes. Em seguida, foram avaliados os métodos PLS, GA-PLS e SVM usando o pacote “caret” e “plsVarSel” (LILAND et al., 2022).

A avaliação dos modelos de calibração deu-se pela análise da raiz do erro quadrático médio da validação cruzada ($RMSE_{CV}$) e da predição ($RMSE_P$); coeficiente de determinação da calibração (R^2_C), validação cruzada (R^2_{CV}) e predição (R^2_P); além do *bias* da predição ($Bias_P$) e a razão entre o desempenho e a distância interquartil da predição ($RPIQ_P$). O RMSE, *bias* e RPIQ foram calculados de acordo com as Equações 14-16:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\ Referência_i - Predito_i)^2}{N}} \quad \text{(Equação 14)}$$

$$Bias = \frac{\sum_{i=1}^{N}(Referência_i - Predito_i)}{N} \quad \text{(Equação 15)}$$

$$RPIQ = \frac{Q3-Q1}{RMSE} \quad \text{(Equação 16)}$$

Onde, $Predito_i$ e $Referência_i$ é o valor predito e o medido pelo método de referência, respectivamente, para a amostra i. N é o número de amostras do conjunto. Q3 e Q1 são o terceiro e primeiro quartil, respectivamente.

Para o modelo PLS, as amostras anômalas foram avaliadas no conjunto de calibração baseados no *leverage* e em caso de amostras cujo resíduo exceda 3 x erro quadrático médio da calibração ($RMSE_C$). Após a remoção dos outliers, foi feita a seleção de variáveis usando GA para obtenção de modelos GA-PLS. Para seleção de variáveis, utilizou-se 20 iterações e tamanho da população de 50 cromossomos. Para os modelos SVM, avaliou-se amostras no conjunto de calibração cujo resíduo excederam 3 x $RMSE_C$.

Semelhante ao procedimento de classificação, a validação cruzada (k-fold, em que: k = 10) foi utilizada para seleção do melhor número de VLs no PLS, seleção do kernel e otimização de hiperparâmetro no SVM. A seleção do melhor pré-processamento deu-se pela análise do $RMSE_{CV}$ e $RMSE_P$.

4 RESULTADOS E DISCUSSÃO

Esta seção foi dividida em três partes. A primeira trata da integração dos espectros ATR-FTIR e EDXRF usando análise multibloco (ComDim) e de agrupamento (PAM) para reconhecimento de padrões não-supervisionado de amostras de solo. Já a segunda parte, trata de métodos de classificação (PLS-DA/SVM/RFE-SVM) usando instrumentos portáteis no NIR para discriminação supervisionada de amostras de solo. A terceira parte, traz o desenvolvimento de modelos de regressão (PLS/GA-PLS/SVM) para determinação multivariada dos teores de argila, areia e COS usando espectros ATR-FTIR.

4.1 CARACTERIZAÇÃO ESPECTROSCÓPICA E IDENTIFICAÇÃO DE PADRÕES NAS AMOSTRAS DE SOLO BASEADA NA INTEGRAÇÃO DE INFORMAÇÕES DE ATR-FTIR E EDXRF

4.1.1 Espectros ATR-FTIR

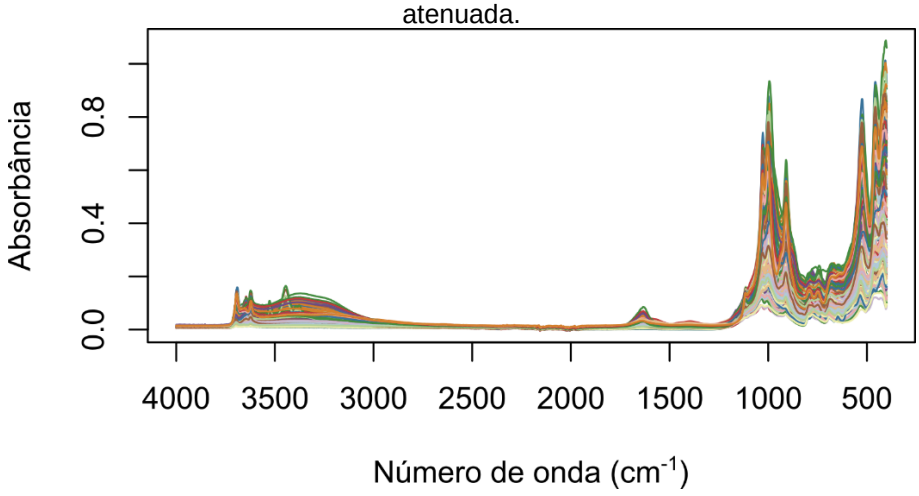
Os espectros MIR apresentam duas regiões mais informativas (Figura 5), de 1800 a 400 cm^{-1} e de 3750 a 2800 cm^{-1}. As associações entre as bandas de absorção e os minerais foram feitas baseadas na literatura (CHUKANOV, 2014; KUMAR; RAJKUMAR, 2014, NGUYEN et al.,1991; CHAUHAN; KUMAR; SHARMA, 2018; PARIKH et al., 2014; TINTI et al., 2015). A análise visual (Figura 5) revelou que as bandas de absorção são principalmente relacionadas ao quartzo e filossilicatos, semelhante a Guimarães et al. (2021).

O quartzo apresentou bandas associadas aos estiramentos e dobramentos de Si-O aproximadamente em 465, 692, 777, 796 e 1090 cm^{-1}. Ademais, as características espectrais do quartzo e sua abundância no solo tornam desafiadora a identificação de feldspatos, como também observou Müller et al. (2014)

Os filossilicatos 1:1, foram evidenciados por bandas em 427 cm^{-1} devido à deformação da ligação Si-O, 530 cm^{-1} associada ao dobramento Al-O-Si, e em 910 cm^{-1} por conta da deformação Al-OH. Foram também observados os estiramentos da ligação Si-O em 1000 cm^{-1} e 1030 cm^{-1}, enquanto os estiramentos da hidroxila em 3622 cm^{-1} e 3696 cm^{-1}. A diferenciação entre filossilicatos 1:1 e 2:1 deu-se pela absorção em filossilicatos 2:1, gerando uma banda alargada por volta de 3400 cm^{-1},

associada ao estiramento de -OH na água. Além desses, alguns minerais do tipo 2:1 apresentam banda relacionada ao dobramento de -OH na água que coordena os cátions entre camadas, por volta de 1640 cm^{-1}. Estiramentos -OH associados à gibbsita foram identificados em quatro amostras de solo em 3393, 3466 e 3527 cm^{-1}.

Figura 5 – Espectros no MIR das amostras de solo obtidos com acessório de refletância total atenuada.



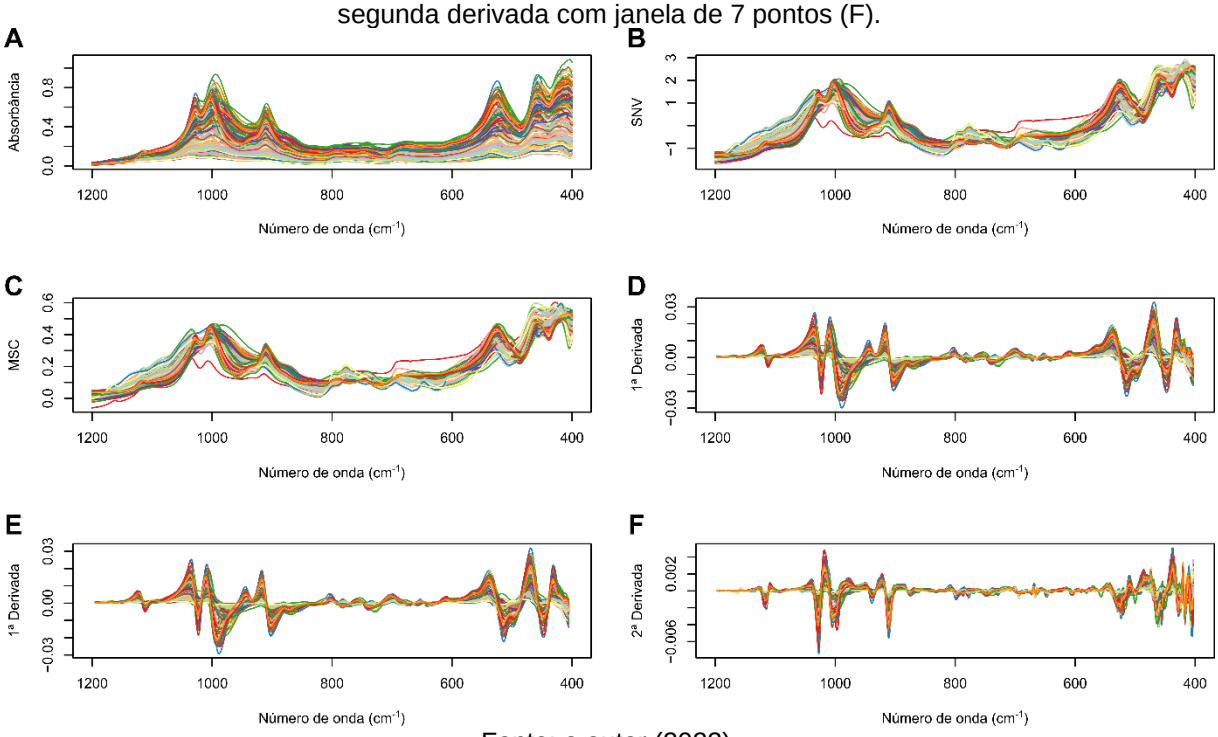
Fonte: o autor (2022).

Bandas de absorção associadas à matéria orgânica foram difíceis de serem identificadas devido à absorbância dominante da fração mineral do solo em relação às fracas absorbâncias de grupos funcionais orgânicos (MARGENOT et al., 2017). Ge et al. (2014), observaram bandas relacionadas ao estiramento da ligação C-H de compostos orgânicos da matéria orgânica (2919 e 2851 cm^{-1}) em solo com COS maior que 50 gC kg^{-1}. No entanto, o COS máximo encontrado neste trabalho foi de 40,87 gC kg^{-1}.

A região entre 1200 e 400 cm^{-1} foi caracterizada pelas absorbâncias mais intensas nos espectros. Segundo Larkin (2018), nos espectros ATR-FTIR a profundidade de penetração da onda evanescente é inversamente proporcional ao número de onda, em consequência, absorções em números de onda menores são mais intensas, portanto, ela foi escolhida para a análise exploratória. Nesta região espectral (Figura 6A), foi observado um espalhamento da radiação evidenciada pelas variações nas linhas de base dos dados espectrais, provavelmente relacionado às variações na composição granulométrica das amostras, como observado por Guimarães et al. (2021). Para correção do espalhamento da radiação foram utilizados os pré-processamentos SNV e MSC (Figuras 6B e 6C respectivamente). Já para a

correção da linha de base, foi empregada a 1ª derivada (polinômio de 2ª ordem e janelas de 7 e 11 pontos) e 2ª derivada (polinômio de 2ª ordem e janela de 7) com filtro de Savitzky-Golay (Figura 6D, 6E e 6F).

Figura 6 – Espectros brutos no MIR de 400-1200 cm^{-1} (A) e pré-processados por SNV (B), MSC (C), primeira derivada com janela de 7 pontos (D), primeira derivada com janela de 11 pontos (E) e segunda derivada com janela de 7 pontos (F).



Fonte: o autor (2022).

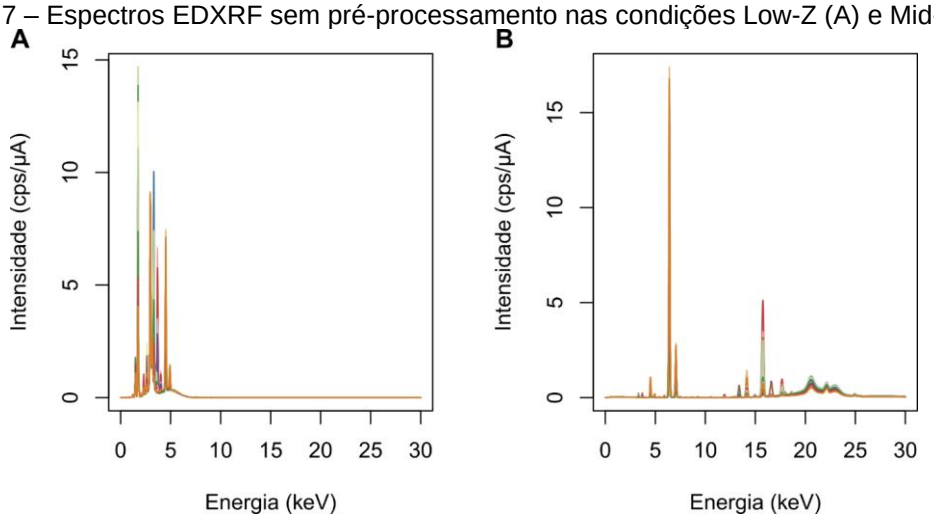
Os espectros pré-processados por SNV e MSC foram muito semelhantes. A primeira derivada (janela de 11 pontos e polinômio de 2º grau) foi suficiente para correção da linha de base sem perda de informação. A Estatística de Hopkins indicou que o pré-processamento usando SNV melhorou a tendência de formação de grupos, em relação à 1ª derivada, conforme apresentado na sequência deste trabalho.

4.1.2 Espectros EDXRF

Os resultados das análises de fluorescência de Raios-X nas condições Low-Z e Mid-Z das amostras de solo estão apresentados nas Figuras 7A e 7B. Os elementos apresentaram diferentes intensidades sob as distintas voltagens de aquisição, sendo elementos de energia na faixa do Mg até ao Ca favorecidos pela condição Low-Z, enquanto a Mid-Z favorece elementos do Ti ao Nb (GRIEKEN e MARKOWICZ, 2002). Os espectros na condição Mid-Z apresentam espalhamentos (> 18 keV) inelástico

(Compton) e elástico (Rayleigh) associados, principalmente, a elementos de baixo número atômico como H, C, O e N (GRIEKEN e MARKOWICZ, 2002; HODOROABA e RACKWITZ, 2014). Logo, a região dos espalhamentos foi removida dos espectros Mid-Z.

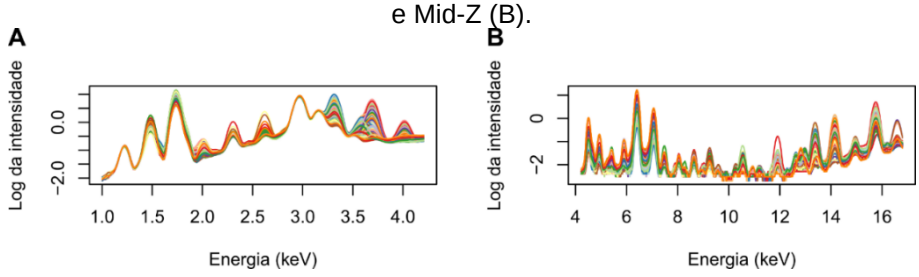
Figura 7 – Espectros EDXRF sem pré-processamento nas condições Low-Z (A) e Mid-Z (B).



Fonte: o autor (2022).

A transformação logarítmica dos espectros para as regiões selecionadas melhorou a visualização das linhas de emissão e identificação dos elementos (Figura 8).

Figura 8 – Espectros EDXRF após transformação logarítmica e suavização nas condições Low-Z (A) e Mid-Z (B).



Fonte: o autor (2022).

A análise visual dos espectros de fluorescência de raios-X em escala logarítmica indicou a presença de 23 elementos (Tabela 2). Na condição Low-Z, observou-se: magnésio (Mg), alumínio (Al), silício (Si), fósforo (P), enxofre (S), cloro (Cl), potássio (K) e cálcio (Ca). A linha de emissão referente ao argônio presente no ar foi observada, uma vez que as amostras foram irradiadas sob o ar atmosférico. No entanto, essa linha é irrelevante para a assinatura química das amostras. Na condição

Mid-Z, observou-se: titânio (Ti), cromo (Cr), manganês (Mn), ferro (Fe), níquel (Ni), cobre (Cu), zinco (Zn), gálio (Ga), arsênio (As), bromo (Br), rubídio (Rb), estrôncio (Sr), zircônio (Zr) e nióbio (Nb).

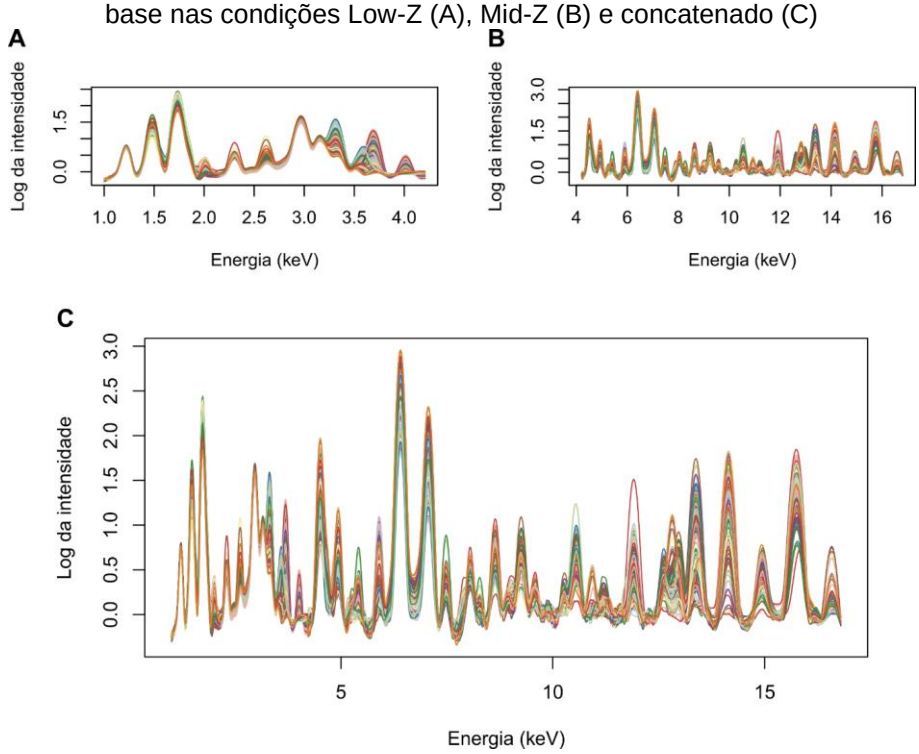
Tabela 2 – Linhas de emissão identificadas nos espectros EDXRF das amostras de solo.

Elemento	**Linhas identificadas**	**Elemento**	**Linhas identificadas**
Mg	K_α = 1,25	Fe	K_α = 6,40; K_β = 7,06
Al	K_α = 1,48	Ni	K_α = 7,48
Si	K_α = 1,74	Cu	K_α = 8,05
P	K_α = 2,01	Zn	K_α = 8,64
S	K_α = 2,31	Ga	K_α = 9,25
Cl	K_α = 2,62	As	K_α = 10,54
K	K_α = 3,31; K_β = 3,59	Br	K_α = 11,92
Ca	K_α = 3,69; K_β = 4,01	Rb	K_α = 13,40; K_β = 14,93
Ti	K_α = 4,51; K_β = 4,95	Sr	K_α = 14,17
Cr	K_α = 5,41	Zr	K_α = 50,78
Mn	K_α = 5,90	Nb	K_α = 16,62

Fonte: Grieken e Markowicz (2002).

Os espectros Low-Z e Mid-Z após transformação logarítmica apresentaram desvio da linha de base e ruído, principalmente nos espectros Mid-Z. Desta forma, os espectros tiveram suas linhas de base corrigidas usando o algoritmo ALS. Em seguida, foram suavizados usando filtro de Savitzky-Golay (polinômio de segunda ordem, com janela de 7 pontos para os espectros Low-Z e 11 pontos para os Mid-Z). Por fim, os espectros Low-Z e Mid-Z pré-processados (Figuras 9A e 9B) foram concatenados conforme apresentado na Figura 9.

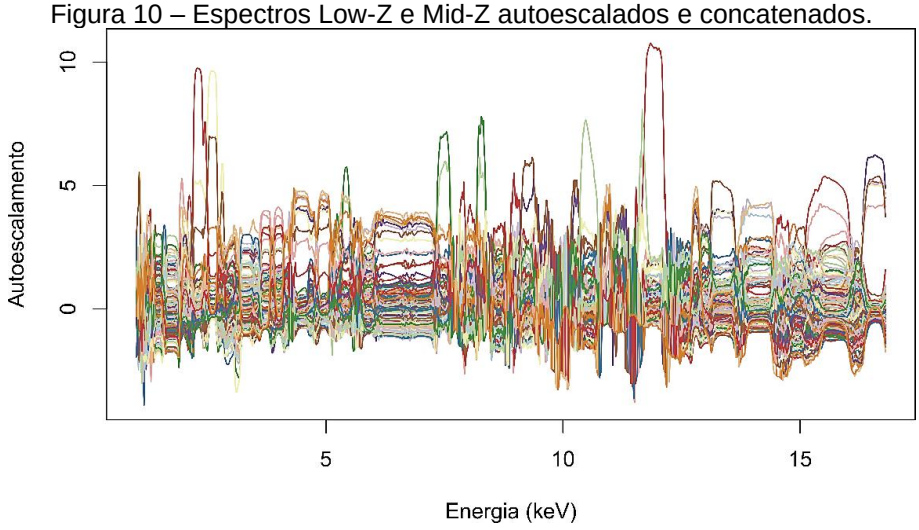
Figura 9 – Espectros EDXRF após transformação logarítmica, suavização e correção da linha de base nas condições Low-Z (A), Mid-Z (B) e concatenado (C)



Fonte: o autor (2022).

Também foi avaliado o auto-escalamento dos espectros brutos, seguida de concatenação para composição da matriz de dados EDXRF (Figura 10). Entretanto, variáveis irrelevantes passam a ter o mesmo peso que as relevantes, podendo gerar dificuldade na interpretação dos resultados (FERREIRA, 2015).

Figura 10 – Espectros Low-Z e Mid-Z autoescalados e concatenados.



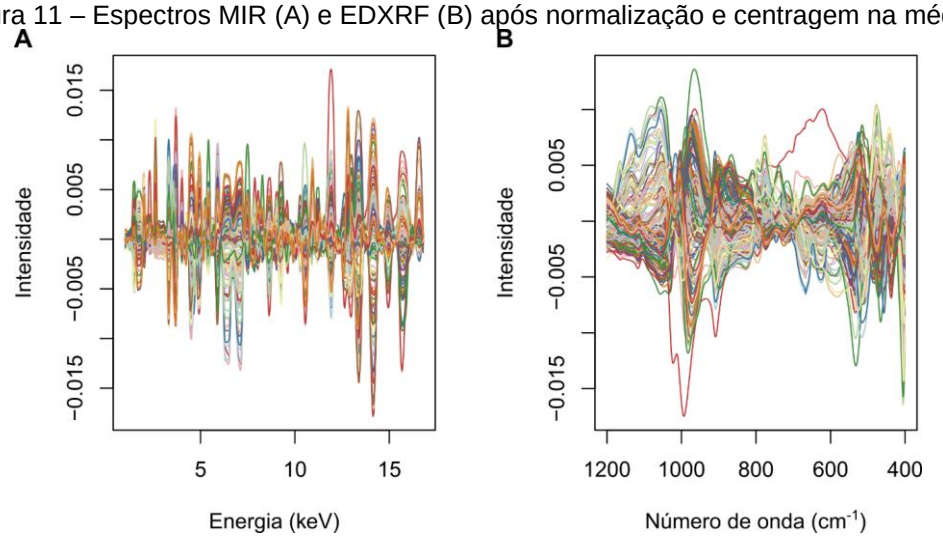
Fonte: o autor (2022).

A escolha do pré-processamento mais adequado ao conjunto de dados espectrais deu-se pela Estatística de Hopkins, semelhante ao ATR-FTIR. Logo, o pré-processamento formado pela combinação da transformação logarítmica, suavização e correção da linha de base foi o escolhido.

4.1.3 Análise ComDim usando espectros ATR-FTIR e EDXRF

Tendo em vista que os blocos de dados (ATR-FTIR e EDXRF) possuem unidades distintas, os espectros foram normalizados utilizando a norma de Frobenius e, posteriormente, centrados na média conforme a Figura 11.

Figura 11 – Espectros MIR (A) e EDXRF (B) após normalização e centragem na média.



Fonte: o autor (2022).

Na análise ComDim, foram extraídas 4 componentes comuns (CC), explicando 81,6% da variância dos dados. A decisão de reter 4 CCs deu-se uma vez que as fontes de variação mais importantes foram extraídas de acordo com a variância explicada e *saliences* de cada componente comum (Tabela 3). Posteriormente, as quatro primeiras CCs foram submetidas à análise de agrupamentos usando PAM para identificação de padrões naturais.

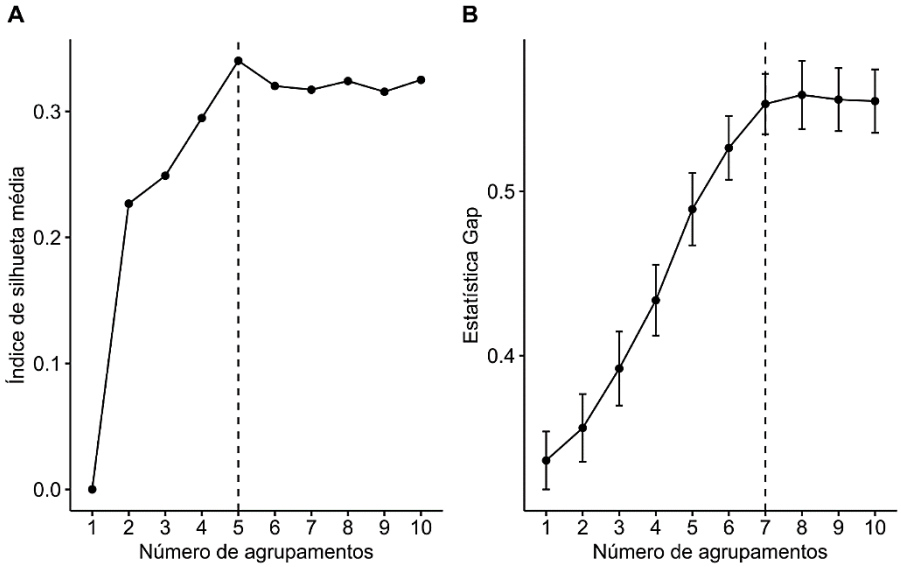
Tabela 3 – Variância explicada e *saliences* da análise ComDim.

Dimensão	Variância (%)	*Saliences*	
		EDXRF	MIR
CC1	45,9	0,211	0,440
CC2	29,8	0,273	0,282
CC3	3,2	0,083	0,098
CC4	2,7	0,119	0,014

Fonte: o autor (2022).

O número ideal de agrupamentos foi identificado usando o índice de silhueta média (Figura 12A) e a estatística *gap* (Figura 12B).

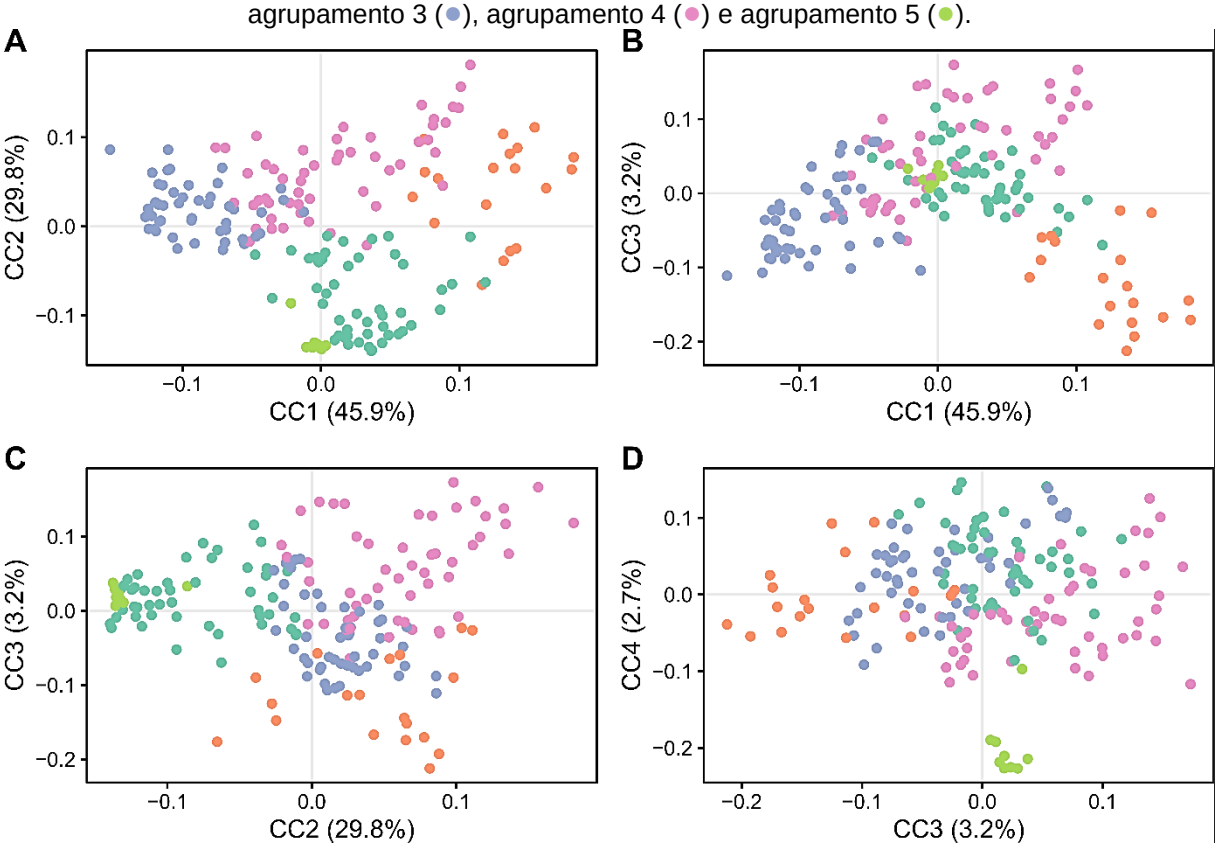
Figura 12 – Métricas para decisão do número ideal de agrupamentos no PAM: índice de silhueta média (A) e estatística gap (B).



Fonte: o autor (2022).

As duas métricas indicaram números de agrupamentos distintos. Para a decisão do número de agrupamentos ideais, utilizou-se o índice de Dunn. A métrica indicou o número ideal de 5 agrupamentos em concordância com o Índice de Silhueta. As amostras nos gráficos de escores (Figura 13) foram coloridas de acordo com o agrupamento ao qual pertencem. Os agrupamentos de cada amostra podem ser consultados no Apêndice A.

Figura 13 – Gráfico de escores da ComDim, em que: agrupamento 1 (●), agrupamento 2 (●), agrupamento 3 (●), agrupamento 4 (●) e agrupamento 5 (●).

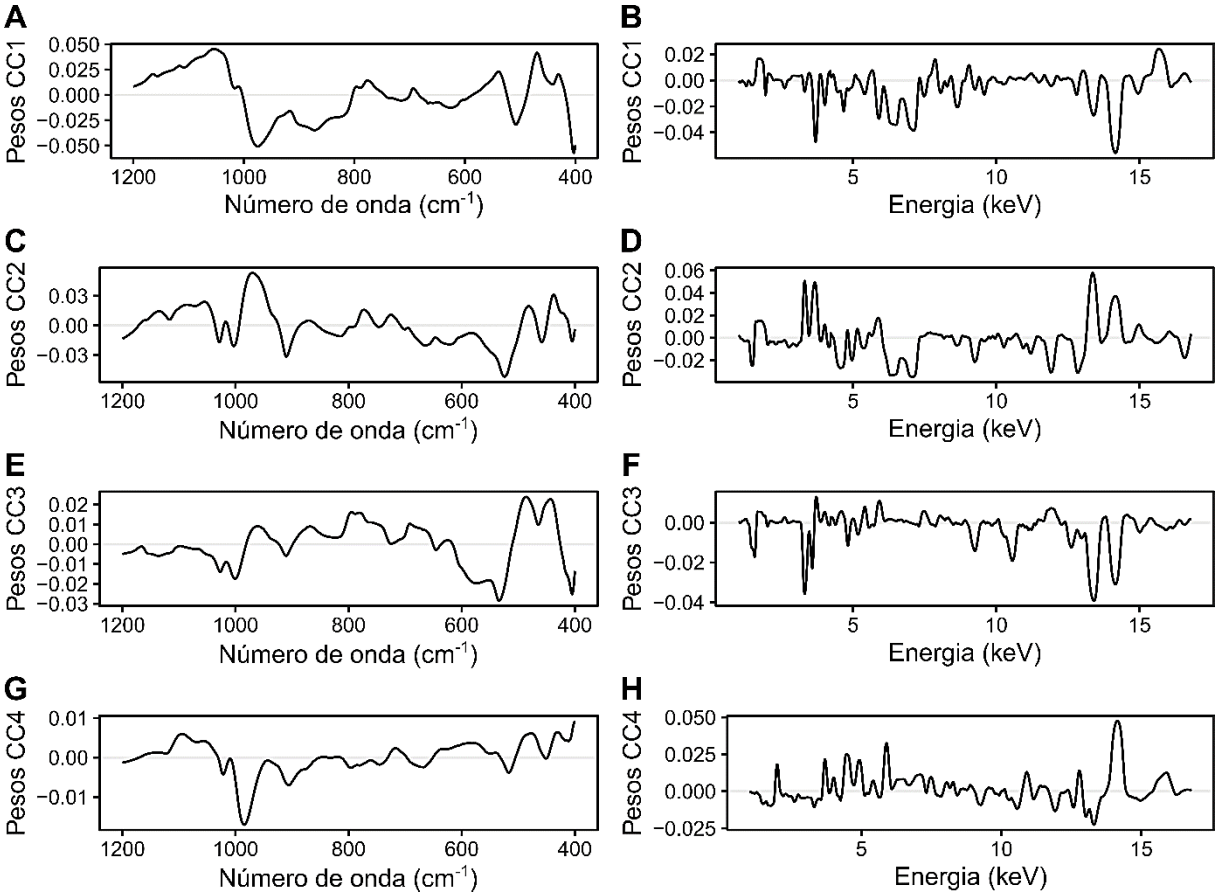


Fonte: o autor (2022).

A análise do gráfico de escores (Figura 13) indicou que o espaço CC1 x CC2 (Figura 13A) discrimina parcialmente os agrupamentos obtidos pelo algoritmo PAM, de modo que, as outras CCs foram importantes na diferenciação. No espaço CC1 x CC3 (Figura 13B) os agrupamentos 2 e 4 foram mais bem diferenciados, enquanto, os agrupamentos 1 e 2, e 1 e 4 foram mais bem discriminados no CC2 x CC3 (Figura 13C). No CC3 x CC4 (Figura 13D), os agrupamentos 1 e 5 foram separados.

A interpretação dos agrupamentos e identificação das variáveis espectrais mais importantes para a discriminação foi realizada com base nos pesos (Figura 14) e *saliences* (Tabela 2) para cada CC.

Figura 14 – Gráfico de pesos da ComDim dos espectros no MIR (A, C, E e G) e EDXRF (B, D, F e H).



Fonte: o autor (2022).

As *saliences* (Tabela 2) indicaram que o padrão das amostras na CC1 é devido, predominantemente, aos espectros no MIR (Figura 14A). A CC1 teve contribuições positivas dos espectros no MIR em 431, 470, 541, 1050, 1116 e 1164 cm^{-1}. As contribuições negativas estão associadas às bandas em 403, 507 871, 973 cm^{-1}. Apesar da menor contribuição do EDXRF para a CC1, os pesos positivos tiveram, principalmente, a contribuição do Si e Zr. Já Ca, Fe, Zn, Rb e Sr para os pesos negativos. Desta forma, com base nos resultados, a CC1 separou amostras com com caracerísticas de quartzo daquelas com presença de filossilicatos 2:1 nos espectros MIR.

Já para a CC2, houve contribuição semelhante dos dois blocos de dados. Houve contribuições positivas de bandas em 439, 480, 727, 775, 970, 1055, 1088 e 1136 cm^{-1}. Nos espectros EDXRF, houve influência de linhas de emissão do Si, K, Ca, Mn, Rb, e Sr. Logo, as contribuições positivas foram associadas ao quartzo e feldspatos. Os pesos negativos foram identificados em 525, 911, 1002 e 1028 cm^{-1}. No EDXRF, linhas de Al, Ti, Fe, Ga e Br tiveram pesos negativos. Logo, a CC2

discriminou amostras caracterizadas pela presença de quartzo e feldspatos daquelas com forte presença de filossilicatos 1:1 nos espectros ATR-FTIR e EDXRF.

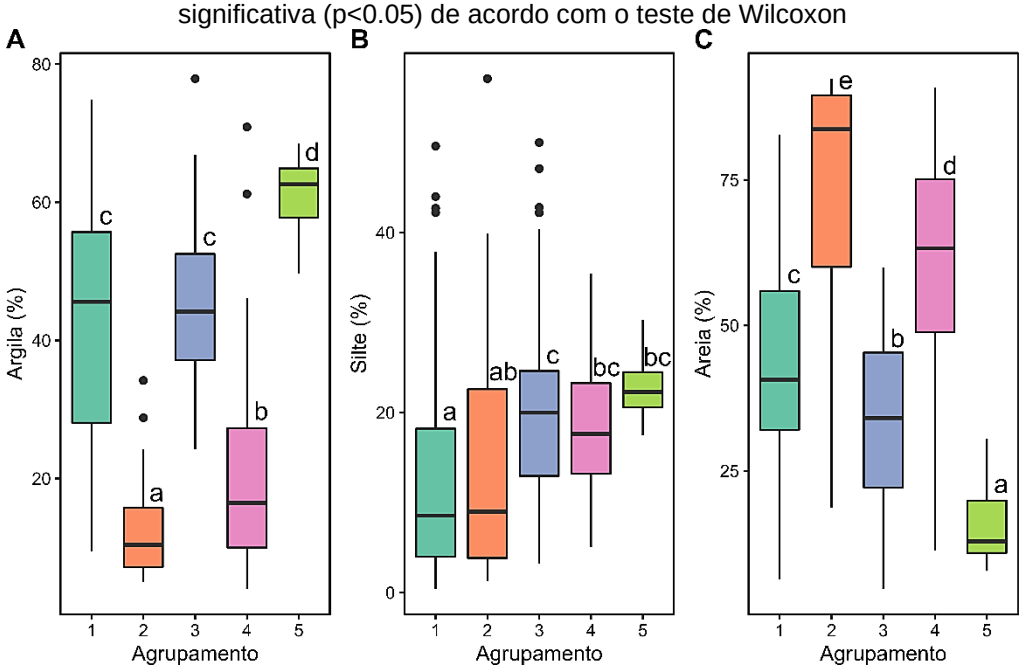
Devido as *saliences* próximas na CC3, também se avaliou os pesos dos dois blocos de dados. Pesos positivos associados ao bloco de dados do ATR-FTIR foram identificados em 444, 486, 692, 780, 795, 868 e 962 cm^{-1}. Os pesos negativos foram observados em 405, 532, 577, 1002 e 1027 cm^{-1}, além das linhas de emissão de Al, K, Ga, Rb e Sr. Assim, a CC3 foi importante para diferenciar amostras com predomínio de quartzo daquelas com predomínio de feldspatos.

A informação no EDXRF foi predominante no padrão de distribuição das amostras na CC4. Contribuições para os pesos positivos foram identificadas nas linhas de P, Ti, Mn e Sr. Já as contribuições negativas foram associadas às linhas de Br, Rb, Ga e Zn. Logo, foi importante para a discriminação de amostras dos agrupamentos 1 e 5.

4.1.4 Perfil dos agrupamentos formados

Os agrupamentos demonstraram diferenças na mineralogia, com isso, foram avaliadas se estão relacionados também com os teores de areia, silte e argila, já que alguns minerais podem ocorrer, preferencialmente, em algum dessas frações granulométricas. Os teores de argila, silte e areia não apresentaram distribuição normal, de acordo com o teste de Shapiro-Wilk ($p<0.05$), logo, para comparar os agrupamentos foram utilizados testes não-paramétricos (Kruskal-Wallis e Wilcoxon). O teste de Kruskal-Wallis indicou que houve diferença significativa ($p<0,05$) nos agrupamentos em relação aos parâmetros. Em seguida, o teste de Wilcoxon revelou os pares de agrupamentos com diferença significativa. Assim, a Figura 15 demonstra a variação nos valores desses parâmetros. Em geral, os teores de argila e areia foram mais importantes para a diferenciação dos agrupamentos.

Figura 15 – Distribuição dos teores de argila, silte e areia, onde letras diferentes indicam diferença significativa (p<0.05) de acordo com o teste de Wilcoxon



Fonte: O autor (2022).

Foi obtida de pesquisas realizadas anteriormente a mineralogia dos medoidoes dos agrupamentos (Tabela 4) para auxiliar a interpretação dos agrupamentos.

Tabela 4 – Mineralogia (Min.) das frações argila, silte e areia dos medoides dos agrupamentos, em que: caulinita (Ct), goethita (Gt), hematita (Hm), quartzo (Qz), ilita (Il), esmectita (Es), vermiculita (V), hornblenda (Hb), feldspatos (Fd), micas (Mi), gibbsita (Gb) e maghemita (Mh).

Agrupamento	Ordem	Horizonte	Min. da argila	Min. do silte	Min. da areia
1	Argissolo vermelho-amarelo	A1	Ct, Gt, Hm	Qz	Qz
2	Argissolo amarelo	A	Ct, Gt, Il	Qz	Qz
3	Planossolo nátrico	Btn	Ct, Es, V, Il, Hb, Fd	Qz, Fd, V	Qz, Fd
4	Planossolo háplico	A1	Ct, Es, Mi, Qz, Fd	-	Qz, Fd
5	Nitossolo háplico	Ap	Ct, Gt, Hm	Ct, Gb, Qz, Gt, Au, Mh,Hm	-

Fonte: Corrêa (2010), Sousa (2019), Câmara (2016) e Neves (2014).

Com base nos gráficos de pesos (Figura 14) e análise boxplot (Figura 15),,o agrupamento 1 foi composto por 48 amostras, argilosas com características cauliníticas devido à argilominerais 1:1, conforme análise ComDim e mineralogia da argila (Tabela 4).

O agrupamento 2 foi composto por 19 amostras com características arenosas e quartzosas. Conforme a Tabela 5, o agrupamento mais próximo a ele foi o agrupamento 1, no entanto, diferenças significativas nos teores de argila e areia são evidentes (Figura 15).

O agrupamento 3, contou com 48 amostras, apresenta semelhança com o agrupamento 1 quanto ao teor de argila, sendo que o agrupamento 3 é caracterizado pela presença de filossilicatos do tipo 2:1. As intensas linhas de emissão de K e Ca observadas nos pesos (Figura 14C) podem indicar possíveis reservas de K e Ca.

O agrupamento 4 foi formado por 52 amostras, com semelhanças ao agrupamento 3 (Tabela 5), característica também observada na mineralogia de areia e argila (Tabela 4). No entanto, houve diferenças significativas ($p<0,05$) nos teores de argila e areia. As amostras do agrupamento 4 apresentam menores teores de argila que as amostras do agrupamento 3. No entanto, também apresentam possíveis reservas de K e Ca (Figura 14C).

O agrupamento 5 foi constituído de 9 amostras, semelhantes ao agrupamento 1 (Tabela 5), caracterizados por alto teor de argila e presença marcante de filossilicatos 1:1 nos espectros ATR-FTIR. No entanto, apresentou diferenças consideráveis na intensidade das linhas do Fe, Ti e Mn nos espectros EDXRF.

Tabela 5 – Matriz de separação dos agrupamentos usando o algoritmo PAM, onde em negrito foram destacadas as menores separações entre pares de agrupamentos.

Agrupamentos	**Agrupamentos**				
	1	**2**	**3**	**4**	**5**
1	-	**0,03613**	0,04227	0,04079	**0,05931**
2	**0,03612**	-	0,08975	0,05774	0,19186
3	0,04227	0,08975	-	**0,01899**	0,12408
4	0,04071	0,05774	**0,01899**	-	0,09943
5	0,05931	0,19186	0,12408	0,09943	-

Fonte: O autor (2022).

O método de reconhecimento de padrões não-supervisionado proposto foi capaz de relacionar informações vibracionais e elementares, e diferenciar amostras de solo baseada em sua composição granulométrica e mineral. De fato, os espectros ATR-FTIR e EDXRF possuem abundância de informação da composição granulométrica e química do solo. Conforme observado por Silva et al. (2018), as informações XRF possuíram correspondência com a mineralogia do Cambissolo analisado. No entanto, esta não foi uma tarefa simples neste trabalho devido ao alto número de amostras e ocorrência de um mesmo elemento em múltiplos minerais. Müller et al. (2014) verificaram que espectros ATR-FTIR podem ser utilizados para análise qualitativa e quantitativa de minerais. Os espectros ATR-FTIR possuem correspondência mais simples com a mineralogia do solo que os espectros EDXRF. Já Heil et al. (2019), usando espectros MIR obtidos por refletância difusa também observaram que os dois agrupamentos obtidos apresentaram diferenças significativas quanto à granulometria. De forma semelhante, os espectros ATR-FTIR neste trabalho também foram influenciados pela composição granulométrica das amostras de solo. Conforme demonstrado no presente trabalho, a integração das duas técnicas permitiu uma avaliação mais abrangente do solo que as técnicas isoladas, sendo possível inferir a presença de feldspatos e óxidos de ferro.

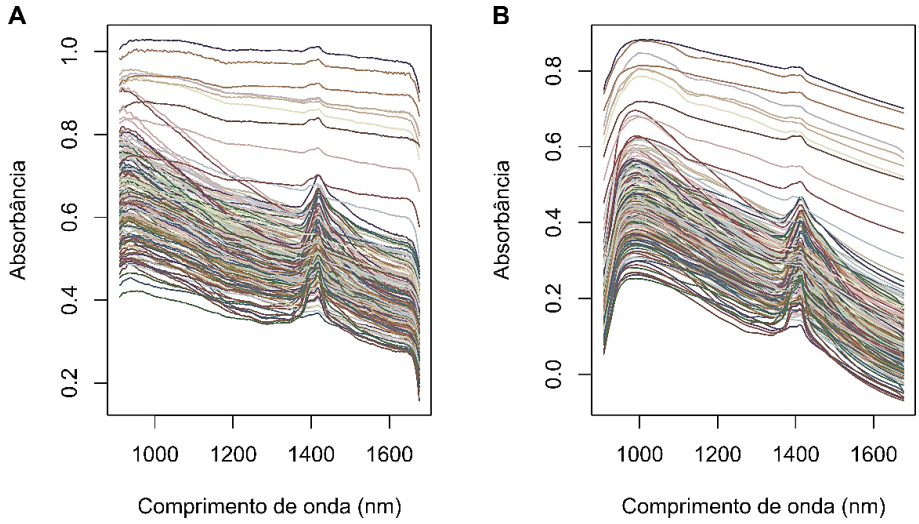
4.2 CLASSIFICAÇÃO DE AMOSTRAS ASSOCIADA À COMPOSIÇÃO GRANULOMÉTRICA E MINERAL DE SOLOS, USANDO INSTRUMENTOS NIR PORTÁTEIS

Diferenciar amostras de solo de acordo com suas características de composição granulométrica e mineral é útil para a tomada de decisões agrícolas e ambientais do solo (AJIBOYE et al., 2019; COBLINSKI et al., 2021; ZINN et al., 2007). No entanto, a análise não-supervisionada, como ComDim e PAM, não é adequada para predição de categoria em novas amostras desconhecidas. Logo, as categorias de solos discriminadas utilizando a abordagem ComDim-PAM foram utilizadas como ponto de partida para modelos PLS-DA e SVM. Para isso, utilizou-se espectros NIR obtidos utilizando-se instrumentos portáteis para classificar amostras em 5 classes associadas à composição granulométrica e mineral, tendo em vista que os equipamentos NIR requerem componentes mais simples e baratos que no MIR ou EDXRF (PASQUINI, 2003).

4.2.1 Espectros NIR

Os espectros NIR das amostras de solo nos instrumentos portáteis NanoNIR e MicroNIR são apresentados na Figura 16. Para comparação do desempenho analítico, as faixas espectrais dos dois instrumentos foram limitadas à faixa espectral comum a ambos: 908 - 1676 nm. Após esse procedimento, a matriz de dados dos espectros do NanoNIR correspondeu a 218 variáveis, enquanto a matriz do MicroNIR foi composta por 125 variáveis.

Figura 16 – Espectros no NIR sem pré-processamento dos instrumentos NanoNIR (A) e MicroNIR (B).



Fonte: O autor (2022)

Os espectros no NIR obtidos pelos instrumentos portáteis apresentaram somente a banda de absorção atribuída ao primeiro sobretom de OH associado a filossilicatos aproximadamente em 1400 nm (FANG et al., 2018). Espectros semelhantes foram obtidos por Mura et al. (2019).

4.2.2 PLS-DA

Na avaliação de modelos PLS-DA, empregando espectros do NanoNIR, utilizou-se suavização (janela de 9 pontos e polinômio do 2º grau) seguida de OSC (3 PCs). Já para espectros obtidos no MicroNIR, utilizou-se OSC (3 componentes). Os modelos OSC (3 componentes) foram desenvolvidos usando o conjunto de treinamento e, posteriormente, foram aplicados aos espectros do conjunto de teste.

Ambos os modelos requereram um número elevado de VL. O modelo PLS-DA usando espectros do NanoNIR requereu menos variáveis latentes e obteve melhor exatidão no conjunto de teste (12 VL e 88,7% de exatidão) que o modelo usando o MicroNIR (14 VL e 86,8% de exatidão). No entanto, conforme as matrizes de confusão (Tabela 6), ambos os modelos PLS-DA não foram eficientes para modelar amostras da classe 2.

Tabela 6 – Matrizes de confusão dos conjuntos de treinamento e teste dos instrumentos NIR portáteis usando PLS-DA.

Treinamento

	NanoNIR							**MicroNIR**					
		Referência							Referência				
		1	2	3	4	5			1	2	3	4	5
	1	33	1	2	0	0		1	30	0	0	1	0
Predição	2	0	8	1	0	0	Predição	2	0	3	0	1	0
	3	0	0	27	2	0		3	1	0	31	2	0
	4	0	6	3	34	0		4	1	12	2	32	0
	5	0	0	0	0	6		5	1	0	0	0	6

Teste

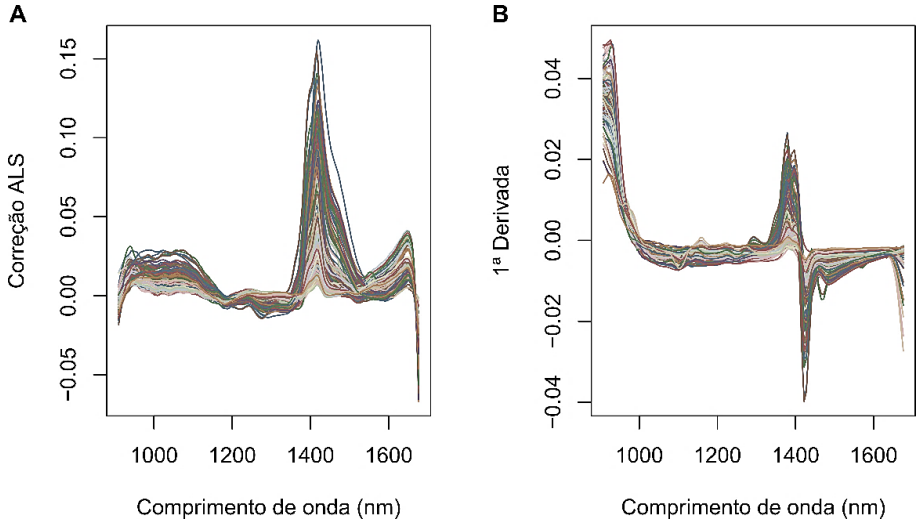
	NanoNIR							**MicroNIR**					
		Referência							Referência				
		1	2	3	4	5			1	2	3	4	5
	1	14	0	0	1	0		1	15	0	0	0	0
Predição	2	0	1	1	0	0	Predição	2	0	0	0	0	0
	3	0	0	14	0	0		3	0	0	12	0	0
	4	1	3	0	15	0		4	0	4	3	16	0
	5	0	0	0	0	3		5	0	0	0	0	3

Fonte: O autor (2022)

4.2.3 SVM

Tendo em vista que os modelos PLS-DA falharam na classificação de amostras da classe 2, avaliou-se modelos SVM. As melhores performances foram obtidas usando *kernel* linear. Foi utilizada a suavização com filtro de Savitzky-Golay (9 pontos) e seguida da correção de linha de base usando algoritmo ALS para os espectros do NanoNIR e primeira derivada usando filtro de Savitzky-Golay (polinômio de 2ª ordem e 5 pontos) para os espectros do MicroNIR. Os espectros pré-processados podem ser observados na Figura 17.

Figura 17 – Espectros dos instrumentos NanoNIR (A) após correção da linha de base ALS (mínimos quadrados assimétricos) e MicroNIR (B) após primeira derivada.



Fonte: O autor (2022).

O *cost* (C) é o único hiperparâmetro otimizável por validação cruzada usando *kernel* linear, assim, foram obtidos valores de C = $10^{-0.1}$ para o NanoNIR e C = $10^{0.5}$ para o MicroNIR. Os modelos SVM tiveram melhor desempenho quando comparados aos modelos PLS-DA (Tabela 7).

Tabela 7 – Matrizes de confusão dos conjuntos de treinamento e teste dos instrumentos NIR portáteis usando SVM.

Treinamento

NanoNIR

		Referência				
		1	2	3	4	5
Predição	1	33	1	0	0	0
	2	0	9	0	1	0
	3	0	0	31	1	0
	4	0	5	2	34	0
	5	0	0	0	0	6

MicroNIR

		Referência				
		1	2	3	4	5
Predição	1	33	1	0	0	0
	2	0	12	0	1	0
	3	0	0	33	0	0
	4	0	2	0	35	0
	5	0	0	0	0	6

Teste

NanoNIR

		Referência				
		1	2	3	4	5
Predição	1	13	0	0	0	0
	2	1	3	1	0	0
	3	0	0	14	0	0
	4	1	1	0	16	0
	5	0	0	0	0	3

MicroNIR

		Referência				
		1	2	3	4	5
Predição	1	12	0	0	0	0
	2	2	4	0	0	0
	3	0	0	14	1	0
	4	1	0	1	15	0
	5	0	0	0	0	3

Fonte: O autor (2022).

Os dados das matrizes de confusão (Tabela 7) foram utilizados para calcular os parâmetros de classificação para os conjuntos de treinamento e teste dos equipamentos NanoNIR e MicroNIR (Tabela 8). Similar aos modelos PLS-DA, os espectros obtidos usando o NanoNIR produziram exatidão ligeiramente melhor que os espectros obtidos usando o MicroNIR.

Tabela 8 – Parâmetros de classificação dos conjuntos de treinamento e teste usando os espectros NIR e SVM

NanoNIR					
Classe	**Conjunto**	**Sensibilidade**	**Especificidade**	**Precisão**	**Exatidão**
1		100%	98,9%	97,1%	
2		60,0%	99,1%	90,0%	
3	Treinamento	93,9%	98,9%	96,9%	91,9%
4		94,4%	92,0%	82,9%	
5		100%	100%	100%	
1		86,7%	100%	100%	
2		75,0%	95,9%	60,0%	
3	Teste	93,3%	100%	100%	92,5%
4		100%	94,6%	88,9%	
5		100%	100%	100%	
MicroNIR					
Classe	**Conjunto**	**Sensibilidade**	**Especificidade**	**Precisão**	**Exatidão**
1		100%	98,9%	97,1%	
2		80,0%	99,1%	92,3%	
3	Treinamento	100%	100%	100%	96,8%
4		97,2%	97,7%	94,6%	
5		100%	100%	100%	
1		80,0%	100%	100%	
2		100%	95,9%	66,7%	
3	Teste	93,3%	97,4%	93,3%	90,6%
4		93,8%	94,6%	88,2%	
5		100%	100%	100%	

Fonte: O autor (2022).

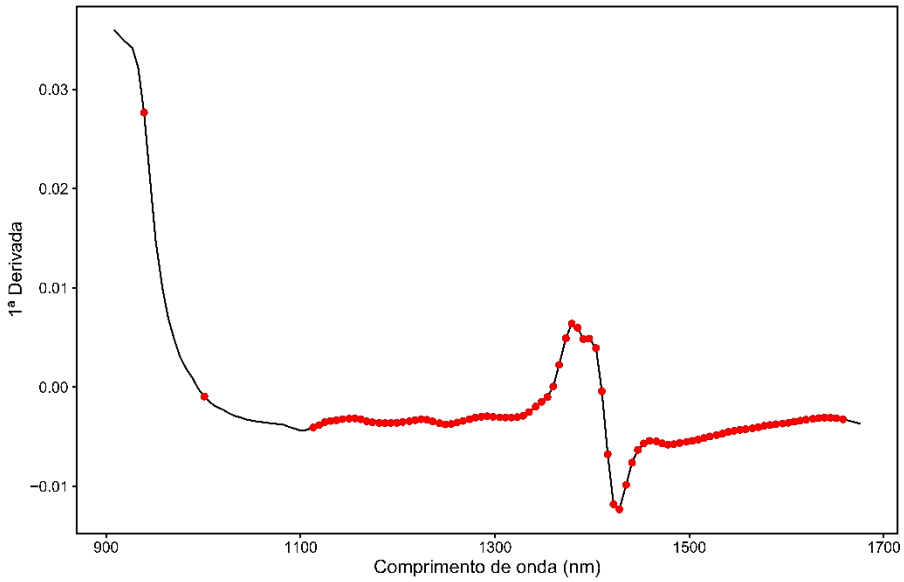
Os espectros no NIR possuem informação limitada na região de aquisição de dados, a qual não está diretamente relacionada ao quartzo, mineral muito importante na composição das amostras do agrupamento 2. Dessa forma, os melhores resultados

obtidos nos modelos SVM podem estar relacionados à capacidade do algoritmo de lidar com informações não-lineares, diferente do PLS-DA.

Avaliou-se, em seguida, a seleção de variáveis para otimização dos modelos de classificação SVM. O modelo RFE-SVM usando os espectros do NanoNIR apresentou um desempenho de predição inferior quando comparado ao modelo SVM usando todas as variáveis. Houve redução da exatidão do conjunto de teste de 92,5% para 86,8%.

A maioria das variáveis selecionadas dos espectros obtidos no MicroNIR possuem comprimento de onda maior que 1100 nm (Figura 18).

Figura 18 – Variáveis selecionadas pelo modelo RFE-SVM demonstrado nos espectros médios do conjunto de treinamento usando MicroNIR.



Fonte: O autor (2022).

O modelo selecionou 91 variáveis (cerca de 72,8% das variáveis originais) e gerou modelo com ligeira melhora na exatidão do conjunto de teste (Tabela 9), atingindo 92,5%.

Tabela 9 – Parâmetros de classificação dos conjuntos de treinamento e teste usando instrumento MicroNIR e modelo RFE-SVM.

Classe	Conjunto	Sensibilidade	Especificidade	Precisão	Exatidão
1		97,0%	98,9%	97,0%	
2		80,0%	98,1%	85,7%	
3	Treinamento	97,0%	100%	100%	95,1%
4		97,2%	96,6%	92,1%	
5		100%	100%	100%	
1		86,7%	100%	100%	
2		100%	95,9%	66,7%	
3	Teste	93,35	97,4%	93,3%	92,5%
4		93,8%	97,3%	93,8%	
5		100%	100%	100%	

Fonte: O autor (2022).

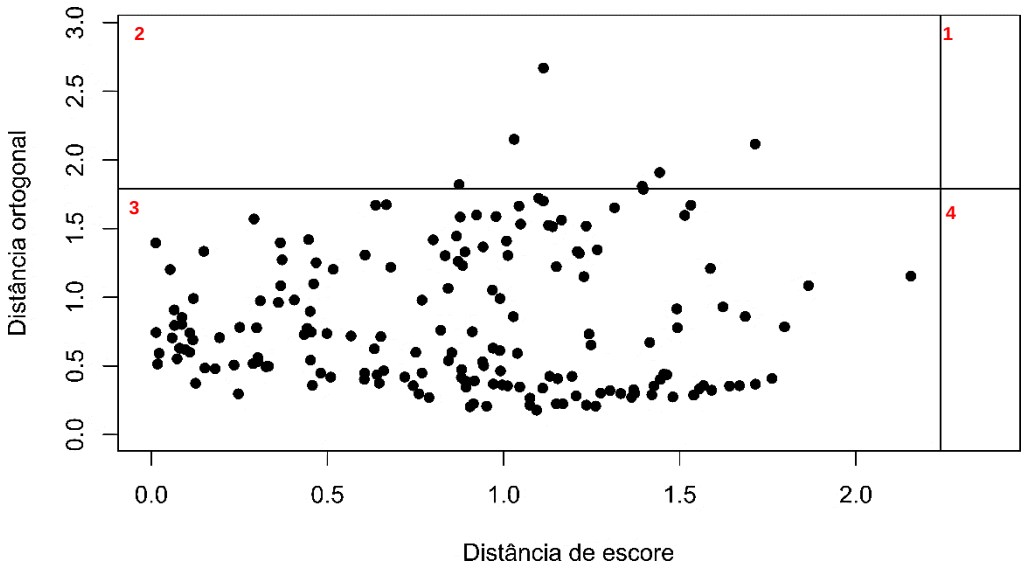
Os resultados demonstraram a aplicabilidade do método desenvolvido usando instrumentos NIR portáteis para triagem rápida do solo associada à composição granulométrica e mineral.

4.3 PREDIÇÃO DOS TEORES DE ARGILA, AREIA E COS USANDO ESPECTROSCOPIA ATR-FTIR

4.3.1 Mapa de outliers da PCA robusta

Os espectros MIR sem pré-processamento foram avaliados para identificação de anomalias espectrais usando o mapa de *outliers* da PCA robusta conforme apresentado na Figura 19. Não foram identificadas amostras anômalas com altas distâncias de escore e ortogonal (primeiro quadrante).

Figura 19 – Mapa de *outliers* da PCA robusta baseado nos espectros MIR.



Fonte: O Autor

4.3.2 Modelos de calibração

Modelos PLS foram inicialmente desenvolvidos para predição dos teores de argila, areia e COS, usando todas as variáveis espectrais. O melhor desempenho para predição desses parâmetros do solo foi obtido empregando o pré-processamento OSC (1 componente). Desta forma, foram desenvolvidos modelos OSC usando o conjunto de calibração e, posteriormente, foram aplicados aos espectros do conjunto de validação externa.

No conjunto de calibração, as amostras anômalas foram identificadas usando o *leverage* e análise dos resíduos (resíduos > 3 x $RMSE_C$). Os parâmetros dos modelos após remoção de outliers encontram-se na Tabela 10.

Tabela 10 – Parâmetros de regressão para argila, areia e COS usando espectros MIR e modelos PLS, em que: variáveis latentes (VL), coeficiente de determinação da calibração (R_C^2), coeficiente de determinação da validação-cruzada (R_{CV}^2), coeficiente de determinação da predição (R_P^2), erro quadrático médio da validação cruzada ($RMSE_{CV}$), erro quadrático médio da validação predição ($RMSE_P$), razão entre o desempenho e a distância interquartil da predição ($RPIQ_P$) e bias da predição ($Bias_P$)

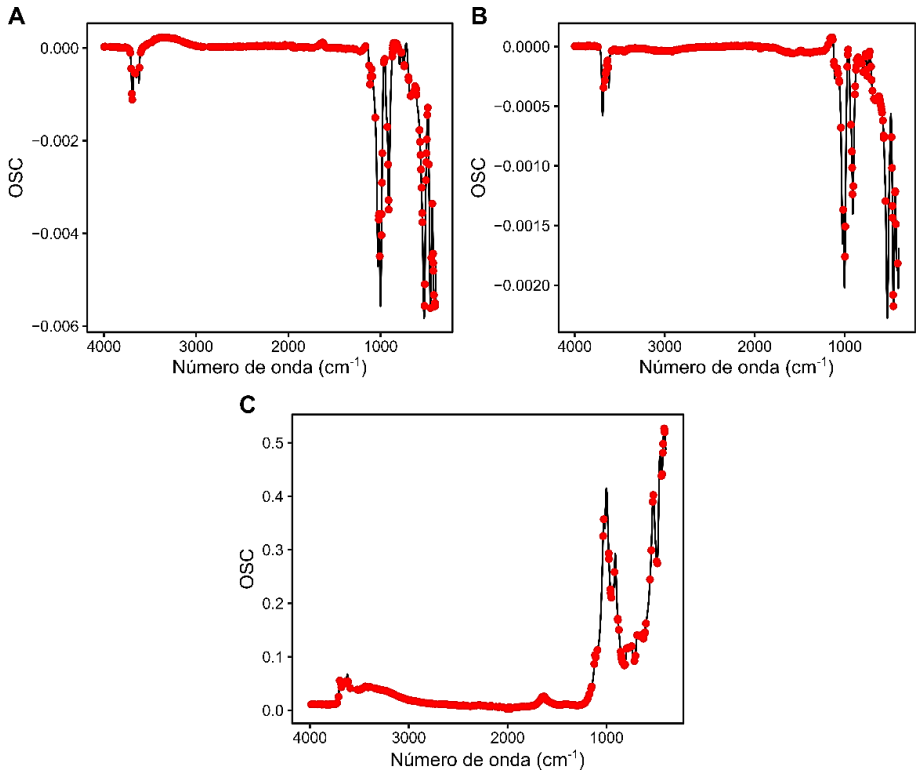
Parâmetro	**Calibração**					**Validação externa**			
	VL	**R_C^2**	**Outlier**	**R_{CV}^2**	**$RMSE_{CV}$**	**R_P^2**	**$RMSE_P$**	**$RPIQ_P$**	**$Bias_P$**
Argila (%)	3	0,84	4	0,84	7,54	0,81	10,12	4,12	3,42
Areia (%)	14	0,84	1	0,72	12,40	0,83	9,71	3,17	0,92
COS (gC kg^{-1})	12	0.88	4	0.83	3.47	0.83	4.11	1.21	-2.72

Fonte: O autor (2022).

O uso de espectros ATR-FTIR apresenta resultados variados para predição dos teores de argila, areia e COS. Ge et al. (2014) descreveram modelos PLS para predição de argila e areia que apresentaram R^2_P = 0,90 para argila e R^2_P = 0,88 para areia. Já o modelo PLS desenvolvido por Haghi et al., 2021, apresentou R^2_P = 0,62 para argila inferior ao descrito o por Ge et al. (2014). Em contrapartida, Haghi et al., 2021 tiveram excelentes resultados para o teor de areia (R^2_P = 0,90). Rial et al. (2016) obtiveram excelentes resultados (R^2_P = 0,88 e $RMSE_P$ = 2,14) para predição do COS em solos ácidos da Espanha. Já a validação-cruzada do modelo para COS de Rial et al (2017) usando solos da ilha de Santa Cruz, Galápagos teve baixo desempenho (R^2_{CV} = 0,58 e $RMSE_{CV}$ = 2,56). Desta forma, os resultados obtidos no presente trabalho com os modelos PLS e todas as variáveis espectrais para predição dos teores de argila, areia e COS foram satisfatórios e compatíveis com aqueles descritos na literatura.

A seleção de variáveis usando GA-PLS reduziu consideravelmente o número de variáveis espectrais para desenvolvimento de modelos, de modo que foram usados menos de 10,5% das 3601 variáveis originais (Figura 20).

Figura 20 – Variáveis selecionadas pelo modelo GA-PLS da argila (A), areia (B) e COS (C) demonstrado nos espectros médios do conjunto de calibração.



Fonte: O autor (2022).

As regiões com maior número de onda (> 1200 cm^{-1}) mesmo com menor absorção, aparentemente pouco informativas, foram importantes e por isso selecionadas nos modelos GA-PLS. No entanto, apesar da redução da considerável redução de variáveis, a redução no número de VLs requeridas foi possível somente no modelo para COS (Tabela 11).

Tabela 11 – Parâmetros de regressão para argila, areia e COS usando espectros ATR-FTIR e modelos GA-PLS, em que: variáveis selecionadas (Var. sel), variáveis latentes (VL), coeficiente de determinação da calibração (R_C^2), coeficiente de determinação da validação-cruzada (R_{CV}^2), coeficiente de determinação da predição (R_P^2), erro quadrático médio da validação cruzada ($RMSE_{CV}$), erro quadrático médio da validação predição ($RMSE_P$), razão entre o desempenho e a distância interquartil da predição ($RPIQ_P$) e bias da predição ($Bias_P$)

Parâmetro	Var. sel	Calibração				Validação externa			
		VL	R_C^2	R_{CV}^2	$RMSE_{CV}$	R_P^2	$RMSE_P$	$RPIQ_P$	$Bias_P$
Argila (%)	373	4	0,85	0,86	7,38	0,82	9,99	4,17	3,41
Areia (%)	348	14	0,84	0,73	12,48	0,84	9,30	3,31	0,93
COS (gC kg^{-1})	322	9	0.85	0.84	3.39	0.85	4.10	1.22	-2.78

Fonte: O autor (2022).

Os modelos SVM foram desenvolvidos usando *kernel* linear e todas as variáveis espectrais. Os espectros sem pré-tratamento apresentaram melhores resultados, os quais são apresentados na Tabela 12. Foram removidas amostras cujo resíduo excedeu 3 x $RMSE_C$. O hiperparâmetro C foi otimizado por validação cruzada para os modelos de argila (C = 0,0007943282), areia (C = 0,010000000) e COS (C = 0,005011872).

Tabela 12 – Parâmetros de regressão para argila, areia e COS usando espectros ATR-FTIR e modelos SVM, em que: coeficiente de determinação da calibração (R_C^2), coeficiente de determinação da validação-cruzada (R_{CV}^2), coeficiente de determinação da predição (R_P^2), erro quadrático médio da validação cruzada ($RMSE_{CV}$), erro quadrático médio da validação predição ($RMSE_P$), razão entre o desempenho e a distância interquartil da predição ($RPIQ_P$) e bias da predição ($Bias_P$)

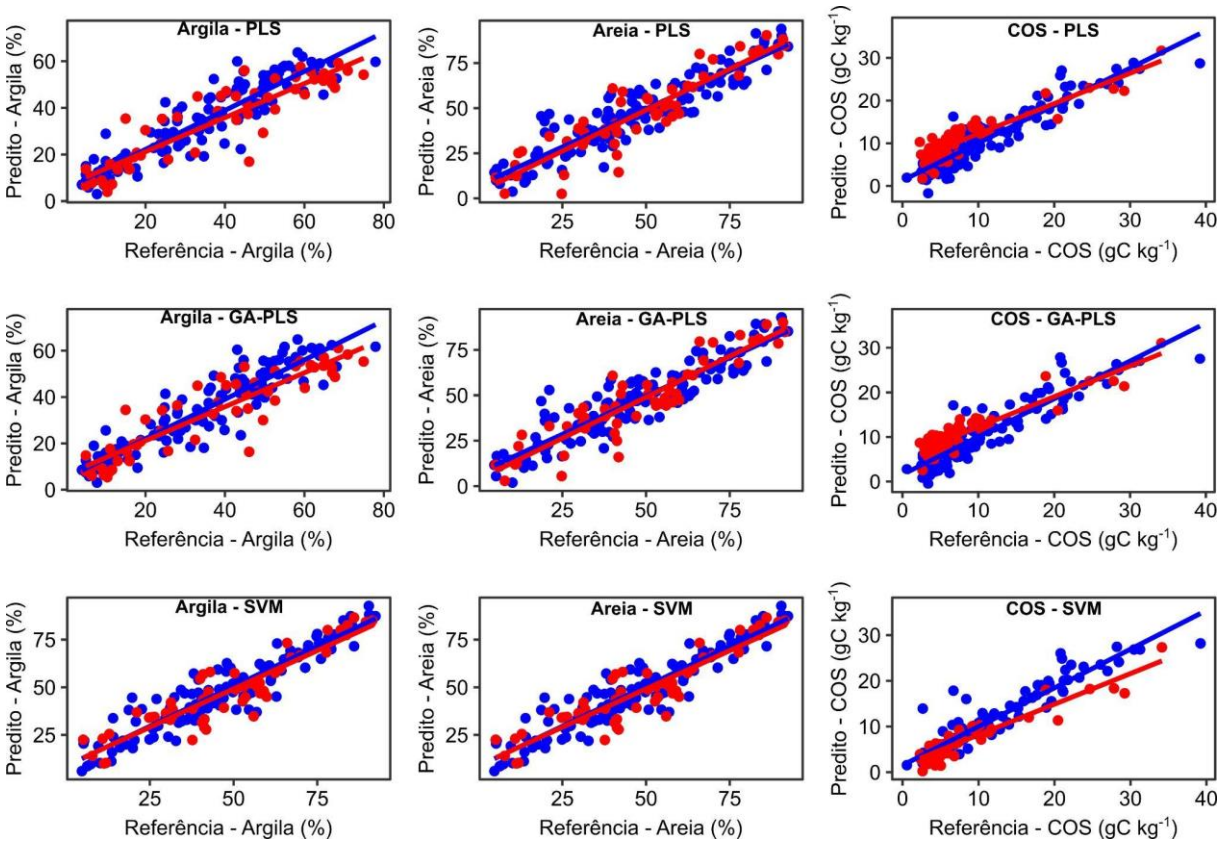
Parâmetro	Calibração				Validação externa			
	R_C^2	Outlier	R_{CV}^2	$RMSE_{CV}$	R_P^2	$RMSE_P$	$RPIQ_P$	$Bias_P$
Argila (%)	0,81	3	0,80	8,51	0,80	10,21	4,09	2,98
Areia (%)	0,88	2	0,72	12,33	0,84	9,17	3,36	0,57
COS (gC kg^{-1})	0,89	3	0,84	3,16	0,88	3,17	1,57	1,24

Fonte: O autor (2022).

4.3.3 Comparação dos modelos de calibração

A Figura 21 demonstra os resultados dos valores de referência em função dos valores preditos pelos modelos PLS, GA-PLS e SVM. A análise visual dos modelos PLS, GA-PLS e SVM indica a tendência linear em concordância com os coeficientes de determinação.

Figura 21 – Gráficos de dispersão referência vs. preditos para argila, areia, e COS usando PLS, GA-PLS e SVM, em que: calibração (●) e validação externa (●).



Fonte: O autor (2022).

Os modelos PLS foram comparados aos modelos GA-PLS e SVM em relação aos resultados de predição usando teste F (95% de confiança) (Tabela 13).

Os resultados do teste F (95% de confiança) apresentados na Tabela 13 indicam que os modelos GA-PLS não reduziram os erros de predição de forma significativa. No entanto, é importante destacar que o modelo GA-PLS manteve o desempenho analítico dos modelos PLS, usando menos variáveis. Quanto aos modelos SVM, somente para o COS houve redução significativa dos erros de

predição. Dessa forma, recomenda-se os modelos PLS para predição dos teores de areia e argila, tendo em vista que este algoritmo fornece modelos mais fáceis de serem interpretados. Por outro lado, para a predição do COS, recomenda-se o modelo SVM por ele apresentar melhor desempenho.

Tabela 13 – Resultados do teste F (95% de confiança) obtidos na comparação dos modelos PLS como GA-PLS e SVM, em que: não-significativo (^{ns}) e significativo (*).

	PLS - Argila	**PLS - Areia**	**PLS - COS**
	$F_{Calculado}$	**$F_{Calculado}$**	**$F_{Crítico}$**
GA-PLS	$1,03^{ns}$	$1,09^{ns}$	$1,00^{ns}$
SVM	$1,04^{ns}$	$1,12^{ns}$	1,67*

Fonte: O autor (2022)

Terra et al. (2015) obtiveram modelos utilizando-se espectros no MIR obtidos por refletância difusa com R^2_P e RPIQ superiores ao deste trabalho para determinação da argila (R^2_P = 0,88 e RPIQ = 4,38) e areia (R^2_P = 0,90 e RPIQ = 5,62). No entanto, o modelo para predição do COS teve R^2_P = 0,77, que foi inferior ao obtido neste trabalho. Deiss et al. (2020) observaram que a otimização dos hiperparâmetros de modelos SVM (*kernel* radial) gerou melhores resultados para determinação dos teores de argila, areia e COS quando comparados aos modelos PLS, ambos utilizando espectros no MIR. Os modelos desenvolvidos usando amostras dos Estados Unidos, apresentaram R^2_P superiores aos deste trabalho para predição de argila (0,91), areia (0,87) e COS (0,93). Por outro lado, utilizando-se amostras de solo da Tanzânia, o resultado de R^2_P foi compatível com este trabalho em relação à predição da argila (0,81) e inferior em relação ao COS (0,57).

5 CONCLUSÃO

Neste trabalho foram desenvolvidas abordagens quimiométricas utilizando informações espectroscópicas para avaliação qualitativa e quantitativa de solos.

A análise ComDim dos espectros na região MIR e de EDXRF permitiu a integração de informações vibracionais e atômicas, extração de informações sobre a relação entre as amostras, a importância das matrizes para a distribuição das amostras e a relação entre elas, e quais variáveis são responsáveis pelo padrão de distribuição das amostras.

A combinação das análises ComDim e PAM discriminou as amostras de solo em cinco categorias com diferenças na composição mineral e nos teores de argila e areia, os quais estão condizentes com a interpretação dos espectros ATR-FTIR e EDXRF, e distribuição dos teores de argila e areia nos agrupamentos.

A classificação de amostras de solo associada à composição granulométrica e mineral usando instrumentos portáteis indicou que os modelos SVM usando kernel linear apresentaram melhor desempenho que os modelos PLS-DA para classificação das amostras nas cinco classes.

A seleção de variáveis por RFE-SVM foi capaz de melhorar a exatidão somente do modelo usando espectros obtidos no MicroNIR, desta forma, atingindo a mesma exatidão do modelo SVM usando NanoNIR com todas as variáveis.

A seleção de variáveis por GA-PLS não melhorou a capacidade preditiva dos modelos PLS, conforme teste F (95% de confiança). Os modelos GA-PLS utilizaram menos de 10,5% das variáveis originais, dessa forma, foram eficientes para redução do esforço computacional. Por outro lado, o modelo SVM usando kernel linear e todas as variáveis espectrais gerou resultado significativamente melhor para COS que o PLS e GA-PLS.

Os métodos desenvolvidos foram úteis para avaliar solos tropicais sob diferentes condições climáticas quanto à mineralogia, composição granulométrica e teor de COS de forma rápida. Sendo possível a obtenção de informações para caracterização, monitoramento e mapeamento de solos

6 PERSPECTIVAS FUTURAS

O presente trabalho possui oportunidades de melhoria e aprofundamento, visando a continuidade do trabalho, como:

- Uso da espectroscopia de refletância difusa na região do ultravioleta-visível para obtenção de mais informações a respeito dos óxidos de ferro;
- Realizar a análise exploratória usando métodos multibloco que permitam a extração de informações comuns e distintas dos blocos de dados espectroscópicos;
- Comparar o desempenho analítico de instrumentos NIR portáteis com instrumento de bancada com maior resolução;
- Avaliar a aplicabilidade da transferência de modelos de classificação para outros instrumentos NIR portáteis;
- Avaliar modelos de calibração usando fusão de dados (EDXRF + ATR-FTIR, NIR + ATR-FTIR e NIR + ATR-FTIR).

REFERÊNCIAS

ADAMS, M. J. **Chemometrics in analytical spectroscopy**. 2. ed. Cambridge: Royal Society of Chemistry, 2004.

AJIBOYE, G. A.; OYETUNJI, C. A.; MESELE, S. A.; TALBOT, J. The Role of Soil Mineralogical Characteristics in Sustainable Soil Fertility Management: A Case Study of Some Tropical Alfisols in Nigeria. **Communications in Soil Science and Plant Analysis**, v. 50, n. 3, p. 333–349, 2019.

ALLEN, B. L.; HAJEK, B. F. Mineral; in Soil Environments. In: DIXON, J. B.; WEED, S. B.; DINAUER, R. C. **Minerals in soil environments**. 2. ed. Madison: Soil Science Society of America, cap. 7, p. 199-278, 1989.

ARAÚJO, J. K. S. **Latossolos com horizonte A húmico na Província Borborema: caracterização de solos e da matéria orgânica**. Tese de Doutorado. Recife, PE: UFRPE, 2014.

ARAÚJO FILHO, J. C.; ARAÚJO, M. S. B.; MARQUES, F. A.; LOPES, H. L. **Solos. In: Geodiversidade do estado de Pernambuco, Programa Geologia do Brasil**. Recife: CPRM. Cap. 7, p. 111-135, 2014.

BASTOS, R. O.; MELQUIADES, F. L.; BIASI, G. E. V. Correction for the effect of soil moisture on in situ XRF analysis using low-energy background: Correction for moisture on XRF analysis using low-energy background. **X-Ray Spectrometry**, v. 41, n. 5, p. 304–307, 2012.

BERRAR, Daniel. Cross-Validation. **Encyclopedia of Bioinformatics and Computational Biology**. Elsevier, 2019. p. 542–545.

BIANCOLILLO, A.; MARINI, F. Chemometric Methods for Spectroscopy-Based Pharmaceutical Analysis. **Frontiers in Chemistry**, v. 6, p. 576, 2018.

BRASIL. Ministério da Agricultura, **Levantamento exploratório – Reconhecimento de solos do Estado de Pernambuco**. Boletim Técnico n.° 26, Volume I. 1973.

BRERETON, Richard G. **Chemometrics for pattern recognition**. Chichester, U.K: Wiley, 2009.

BRILHANTE, S. A. **Gênese e caracterização de solos desenvolvidos de riolito em ambiente de Mata Atlântica, Pernambuco, Brasil**. Dissertação de Mestrado. Recife, PE: UFRPE, 2014.

BRILHANTE, S. A. **Geoquímica de elementos terras raras em solos da bacia Cabo, Nordeste do Brasil**. Tese de Doutorado. Recife, PE: UFRPE, 2018.

CÂMARA, E. R. G. **Caracterização e pedogênese de Luvissolos e Planossolos no núcleo de desertificação de Cabrobó**, Pernambuco. Tese de Doutorado. Recife, PE: UFRPE, 2016.

CERVANTES, Jair; GARCIA-LAMONT, Farid; RODRÍGUEZ-MAZAHUA, Lisbeth; LOPEZ, Asdrubal. A comprehensive survey on support vector machine classification: Applications, challenges and trends. **Neurocomputing**, vol. 408, p. 189–215, 2020.

CHAUHAN, R.; KUMAR, R.; SHARMA, V. Soil forensics: A spectroscopic examination of trace evidence. **Microchemical Journal**, v. 139, p. 74-84, 2018.

CHUKANOV, N. V. **Infrared spectra of mineral species: extended library**. 1. ed. Dordrecht: Springer, 2014.

COBLINSKI, João Augusto; GIASSON, Élvio; DEMATTÊ, José A.M.; et al. Prediction of soil texture classes through different wavelength regions of reflectance spectroscopy at various soil depths. **CATENA**, v. 189, p. 104485, 2020.

COBLINSKI, J. A.; INDA, A. V.; DEMATTÊ, J. A. M.; DOTTO, A. C.; GHOLIZADEH, A.; GIASSON, É. Identification of minerals in subtropical soils with different textural classes by VIS–NIR–SWIR reflectance spectroscopy. **CATENA**, v. 203, p. 105334, 2021.

COCCHI, Marina. Introduction. **Data Handling in Science And Technology**, p. 1-26, 2019. Elsevier.

CONKLIN, A. R. **Introduction to soil chemistry: analysis and instrumentation**. 2. ed. Hoboken: Wiley, 2013.

CORRÊA, M. **Mineralogia dos solos de referência do Estado de Pernambuco**, Recife, 2010 (dados não publicados).

COSTA, E. U. C. **Caracterização e gênese de Argissolos e Nitossolos na Bacia Cabo, Pernambuco**. Dissertação de Mestrado. Recife, PE: UFRPE, 2012.

DEISS, L.; MARGENOT, A. J.; CULMAN, S. W.; DEMYAN, M. S. Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. **Geoderma**, v. 365, p. 114227, 2020.

DERIS, Ashanira Mat; ZAIN, Azlan Mohd; SALLEHUDDIN, Roselina. Overview of Support Vector Machine in Modeling Machining Performances. **Procedia Engineering**, vol. 24, p. 308–312, 2011.

EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária. Centro Nacional de Pesquisa de Solos. **Manual de métodos de análise de solo**. Centro Nacional de Pesquisa de Solos. 2. Ed. Ver. Atual. Rio de Janeiro, 1997. 212p.

EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária. Centro Nacional de Pesquisa de Solos. **Manual de análises químicas de solos, plantas e fertilizantes**. Centro Nacional de Pesquisa de Solos. 2. Ed. Rev. Ampl. Brasília: DF: Embrapa Informação Tecnológica. 2009.

EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária. **Manual de métodos de análise de solo**. 3. Ed. Rev. Ampl. Brasília: DF: Embrapa Informação Tecnológica. 2017.
EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária. **Sistema brasileiro de classificação de solos**. 5. Ed. Brasília: EMBRAPA, 2018.

ENGEL, J.; GERRETZEN, J.; SZYMAŃSKA, E.; JANSEN, J. J.; DOWNEY, G.; BLANCHET, L.; BUYDENS, L. M. C. Breaking with trends in pre-processing?. **TrAC Trends in Analytical Chemistry**, v. 50, p. 96–106, 2013

ESTEBAN, M.; ARIÑO-BLASCO, M.C.; DÍAZ-CRUZ, J. M. **Chemometrics in Electrochemistry**. Comprehensive Chemometrics, p. 1-31, 2020. Elsevier.

ESBENSEN, K.H.; JULIUS, L.P. **Representative Sampling, Data Quality, Validation – A Necessary Trinity in Chemometrics**. Comprehensive Chemometrics, p. 1-20, 2009. Elsevier.

FANG, Q.; HONG, H.; ZHAO, L.; KUKOLICH, S.; YIN, K.; WANG, C. Visible and Near-Infrared Reflectance Spectroscopy for Investigating Soil Mineralogy: A Review. **Journal of Spectroscopy**, v. 2018, p. 1–14, 2018.

FAO. **Why Soil Matters?**. Disponível em: < http://www.fao.org/soils-portal/about/en/ >;. Acesso em: 19 jul. 2019.

FERREIRA, M. M. C. **Quimiometria - Conceitos, métodos e aplicações**. 1. ed. Campinas: Editora UNICAMP, 2015.

GE, Y.; THOMASSON, J. A.; MORGAN, C. L. S. Mid-infrared attenuated total reflectance spectroscopy for soil carbon and particle size determination. **Geoderma**, v. 213, p.57-63, 2014.

GENISHEVA, Z.; QUINTELAS, C.; MESQUITA, D.P.; FERREIRA, E.C.; OLIVEIRA, J.M.; AMARAL, A.L. New PLS analysis approach to wine volatile compounds characterization by near infrared spectroscopy (NIR). **Food Chemistry**, vol. 246, p. 172–178, 2018.

GHOLAMI, Raoof; FAKHARI, Nikoo. Support Vector Machine: Principles, Parameters, and Applications. **Handbook of Neural Computation.** Elsevier, 2017.

GILLESPIE, A; GREGORICH, E.; HELGASON, B.; PEAK, D. Soil Instrumental Methods. In: MEYERS, R. A. **Encyclopedia of Analytical Chemistry**, 1. ed. Chichester: John Wiley & Sons, 2015. p.1-27.

GORRY, P. A. General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method. **Analytical Chemistry**, v. 62, n. 6, p. 570–573, 1990.

GRIEKEN, R. V.; MARKOWICZ, A. (Orgs.). **Handbook of X-ray spectrometry**. 2. ed., rev.expanded. New York: Marcel Dekker, 2002. (Practical spectroscopy, v. 29).

GUIMARÃES, Clécia Cristina Barbosa; DEMATTÊ, José A. M.; AZEVEDO, Antônio Carlos de; DALMOLIN, Ricardo Simão Diniz; CATEN, Alexandre Ten; SAYÃO, Veridiana Maria; SILVA, Rafael Cipriano da; POPPIEL, Raul Roberto; MENDES, Wanderson de Sousa; SALAZAR, Diego Fernando Urbina. Soil weathering behavior assessed by combined spectral ranges: insights into aggregate analysis. **Geoderma**, v. 402, p. 115154, 2021.

HAGHI, R. K.; PÉREZ-FERNÁNDEZ, E.; ROBERTSON, A. H. J. Prediction of various soil properties for a national spatial dataset of Scottish soils based on four different chemometric approaches: A comparison of near infrared and mid-infrared spectroscopy. **Geoderma**, v. 396, p. 115071, 2021.

HEIL, J.; HÄRING, V.; MARSCHNER, B.; STUMPE, B. Advantages of fuzzy k-means over k-means clustering in the classification of diffuse reflectance soil spectra: A case study with West African soils. **Geoderma**, v. 337, p. 11–21, 2019.

HELFER, Gilson Augusto; VICTÓRIA BARBOSA, Jorge Luis; SANTOS, Ronaldo dos; COSTA, A. B. A computational model for soil fertility prediction in ubiquitous agriculture. **Computers and Electronics in Agriculture**, v. 175, p. 105602, 2020.

HENNING, C. **fpc: Flexible Procedures for Clustering**. R Package Vignette, Report No.: R Package Version 2.2-9, 2020.

HLAVANGUANE, A. C, J. O. **Intemperismo de anortositos e formação de solos na região semiárida, Nordeste do Brasil**. Dissertação de Mestrado. Recife, PE: UFRPE, 2017.

HODOROABA, Vasile-Dan; RACKWITZ, Vanessa. Gaining Improved Chemical Composition by Exploitation of Compton-to-Rayleigh Intensity Ratio in XRF Analysis. **Analytical Chemistry**, v. 86, n. 14, p. 6858–6864, 2014.

HUBERT, Mia; ROUSSEEUW, Peter J; VANDEN BRANDEN, Karlien. ROBPCA: A New Approach to Robust Principal Component Analysis. **Technometrics**, vol. 47, no. 1, p. 64–79, 2005.

JACOMINE, Paulo Klinger Tito. Solos sob caatingas: características e uso agrícola. **O solo nos grandes domínios morfoclimáticos do Brasil e o desenvolvimento sustentado**. Viçosa: SBCS/UFV, p. 95-111, 1996.

JOBBÁGY, Esteban G.; JACKSON, Robert B. The vertical distribution of soil organic carbon and its relation to climate and vegetation. **Ecological Applications**, v. 10, n. 2, p. 423–436, 2000.

KASSAMBARA, A.; MUNDT, F. **factoextra: Extract and Visualize the Results of Multivariate Data Analyses**. R Package Vignette, Report No.: R Package Version 1.0.7, 2020.

KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Finding groups in data: an introduction to cluster analysis**. John Wiley & Sons, 2009.

KENNARD, Ronald W.; STONE, Larry A. Computer aided design of experiments. **Technometrics**, v. 11, n. 1, p. 137-148, 1969.

KUCHERYAVSKIY, Sergey. mdatools–R package for chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 198, p. 103937, 2020.

KUHN, Max. **Recursive Feature Elimination**. 2019. Disponível em: http://topepo.github.io/caret/recursive-feature-elimination.html#backwards-selection. Acesso em: 15 jan. 2022.

KUHN, T.; WING, J.; WESTON, S.; WILLIAMS, A.; KEEFER, C.; ENGELHARDT, A.; COOPER, T.; MAYER, Z.; KENKEL., BENESTY, M.; LESCARBEAU, R.; ZIEM, A; SCRUCCA, L.; TANG, Y.; CANDAN, C.; HUNT, T. **caret: Classification and Regression Training**. R Package Vignette, Report No.: R Package Version 6.0-91, 2022.

KUMAR, R. S.; RAJKUMAR, P. Characterization of minerals in air dust particles in the state of Tamilnadu, India through FTIR, XRD and SEM analyses. **Infrared Physics & Technology**, v. 67, p.30-41, 2014.

KUMAR, R.; SHARMA, V. Chemometrics in forensic science. **Trac Trends in Analytical Chemistry**, v. 105, p.191-201, 2018.

LAL, R. The potential of soils of the tropics to sequester carbon and mitigate the greenhouse effect. In: **Advances in Agronomy**: Elsevier, 2002.

LARKIN, Peter J. Instrumentation and Sampling Methods. **Infrared and Raman Spectroscopy**, p. 29-61, 2018.

LEMOS, J. O. **Caracterização de gleissolos com e sem tiomorfismo da região litorânea do Estado de Pernambuco**. Dissertação de Mestrado. Recife, PE: UFRPE, 2013.

LEMOS, R. C.; SANTOS, R. D. **Manual de descrição e coleta de solo no campo**. 3. ed. Viçosa: Sociedade Brasileira de Ciência do Solo, 1996.

LILAND, K. G.; MEHMOOD, T.; SÆBØ, S. **plsVarSel: Variable Selection in Partial Least Squares**. R Package Vignette, Report No.: R Package Version 2.0-1.19, 2022.

LIMA, G. K**. Caracterização de Vertissolos do Nordeste brasileiro**. Dissertação de Mestrado. Recife, PE: UFRPE, 2014.

LIN, Wanchang. **mt: Metabolomics Data Analysis Toolbox**. R Package Vignette, Report No.: R Package Version 2.0-1.19, 2022.

LUNA, Aderval S.; GOIS, Jefferson S. de. **Application of Chemometric Methods Coupled with Vibrational Spectroscopy for the Discrimination of Plant Cultivars and to Predict Physicochemical Properties Using R**. Vibrational Spectroscopy for Plant Varieties And Cultivars Characterization, p. 165-194, 2018. Elsevier.

LÜTZOW, M. V.; KÖGEL-KNABNER, I.; EKSCHMITT, K.; MATZNER, E.; GUGGENBERGER, G.; MARSCHNER, B.; FLESSA, H. Stabilization of organic matter in temperate soils: mechanisms and their relevance under different soil conditions - a review: Mechanisms for organic matter stabilization in soils. **European Journal of Soil Science**, v. 57, n. 4, p. 426–445, 2006.

MCCARTHY, G . W.; REEVES III, J. B.; YOST, R.; DORAISWAMY, P. C.; DOUMBIA, M. Evaluation of methods for measuring soil organic carbon in West African soils. **African Journal of Agricultural Research**. V. 5, n. 16, p. 2169-2177, 2010.

MAECHLER, M.; ROUSSEEUW, P.; STRUYF, A.; HUBERT, M.; HORNIK, K.; STUDER, M.; ROUDIER, P.; GONZALEZ, J.; KOZLOWSKI, K.; SCHUBERT, E.; MURPHY, K. **cluster: "Finding Groups in Data"**. R Package Vignette, Report No.: R Package Version 2.1.3, 2022.

MANGAMANA, E. T.; CARIOU, V.; VIGNEAU, E. **MBAnalysis: Multiblock Exploratory and Predictive Data Analysis**. R Package Vignette, Report No.: R Package Version 0.1.0, 2021.

MARGENOT, A. J.; CALDERÓN, F. J.; GOYNE, K. W.; MUKOME, F. N. D.; PARIKH, S. J. IR Spectroscopy, Soil Analysis Applications, In: LINDON, J. C.; TRANTER, G. E.; KOPPENAAL, D. W. **Encyclopedia of Spectroscopy and Spectrometry**. 3. ed., Academic Press, p.448-454, 2017.

MAURYA, S.; ABRAHAM, J. S.; SOMASUNDARAM, S.; TOTEJA, R.; GUPTA, R.; MAKHIJA, S. Indicators for assessment of soil quality: a mini-review. **Environmental Monitoring and Assessment**, v. 192, n. 9, p. 604, 2020.

MEER, F. van der. Near-infrared laboratory spectroscopy of mineral chemistry: A review. **International Journal of Applied Earth Observation and Geoinformation**, v. 65, p. 71–78, 2018.

MEHMOOD, T.; SÆBØ, S.; LILAND, K. H. Comparison of variable selection methods in partial least squares regression. **Journal of Chemometrics**, v. 34, n. 6, 2020.

MENDES, W. de S.; DEMATTÊ, J. A. M.; ROSIN, N. A.; TERRA, F. da S.; POPPIEL, R. R.; URBINA-SALAZAR, D. F.; BOECHAT, C. L.; SILVA, E. B.; CURI, N.; SILVA, S. H. G.; JOSÉ DOS SANTOS, U.; SOUZA VALLADARES, G. The Brazilian soil Mid-infrared Spectral Library: The Power of the Fundamental Range. **Geoderma**, v. 415, p. 115776, 2022.

MILOSEVIC, M. **Internal reflection and ATR spectroscopy**. 1. ed. Hoboken: John Wiley & Sons, 2012.

MIRABELLA, F. M. Principles, Theory and Practice of Internal Reflection Spectroscopy. In: GRIFFITHS, P. R. **Handbook of Vibrational Spectroscopy**. John Wiley & Sons, 2006.

MISHRA, P.; ROGER, J.-M.; JOUAN-RIMBAUD-BOUVERESSE, D.; BIANCOLILLO, A.; MARINI, F.; NORDON, A.; RUTLEDGE, D. N. Recent trends in multi-block data

analysis in chemometrics for multi-source data integration. **TrAC Trends in Analytical Chemistry**, v. 137, p. 116206, 2021.

MOHAMMADI, M.; KHANMOHAMMADI KHORRAMI, M.; VATANI, A.; GHASEMZADEH, H.; VATANPARAST, H.; BAHRAMIAN, A.; FALLAH, A. Rapid determination and classification of crude oils by ATR-FTIR spectroscopy and chemometric methods. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, vol. 232, p. 118157, 2020.

MORONA, Franslley; SANTOS, Felipe R. dos; BRINATTI, André M.; MELQUIADES, Fábio L. Quick analysis of organic matter in soil by energy-dispersive X-ray fluorescence and multivariate analysis. **Applied Radiation and Isotopes**, v. 130, p. 13-20, dez. 2017.

MÜLLER, C. M.; PEJCIC, B.; ESTEBAN, L.; PIANE, C. D.; RAVEN, M.; MIZAIKOFF, B. Infrared Attenuated Total Reflectance Spectroscopy: An Innovative Strategy for Analyzing Mineral Components in Energy Relevant Systems. **Scientific Reports**, v. 4, n. 1, p. 6764, 2015.

MURA, S.; CAPPAI, C.; GREPPI, G. F.; BARZAGHI, S.; STELLARI, A.; CATTANEO, T. M. P. Vibrational spectroscopy and Aquaphotomics holistic approach to determine chemical compounds related to sustainability in soil profiles. **Computers and Electronics in Agriculture**, v. 159, p. 92–96, 2019.

NASCIMENTO, I. G. **Caracterização e gênese de solos originados de rochas metamórficas e carbonatadas na região semiárida de Pernambuco**. Dissertação de Mestrado. Recife, PE: UFRPE, 2020.

NÆS, Tormod. The design of calibration in near infra-red reflectance analysis by clustering. **Journal Of Chemometrics**, v. 1, n. 2, p. 121-134, abr. 1987. Wiley.

NAIMAN, R. J.; DÉCAMPS, H.; MCCLAIN, M. E.; LIKENS, G. E. **Structural Patterns**. In: RIPARIA: Elsevier, 2005. p. 79–123.

NASCIMENTO, Clístenes Williams Araújo do; OLIVEIRA, Abel Batista de; RIBEIRO, Mateus Rosas; et al. Distribution and Availability of Zinc and Copper in Benchmark Soils of Pernambuco State, Brazil. **Communications in Soil Science and Plant Analysis**, v. 37, n. 1–2, p. 109–125, 2006.

NEVES, L. V. M. **Caracterização de Nitossolos Háplicos-Argissolo Amarelo em uma topossequência na bacia Cabo, Sirinhaém-PE**. Dissertação de Mestrado. Recife, PE: UFRPE, 2014.

NEVES, L. V. M. **Pedogênses e diversidade microbiana em Planossolos sob processo de desertificação na região semiárida de Pernambuco**. Tese de Doutorado. Recife, PE: UFRPE, 2018.

NGUYEN, T. T.; JANIK, L. J.; RAUPACH, M. Diffuse reflectance infrared Fourier transform (DRIFT) spectroscopy in soil studies. **Soil Research**, v. 29, n. 1, p.49-67, 1991.

OLIVERI, P.; MALEGORI, C.; MUSTORGI, E.; CASALE, M. Qualitative pattern recognition in chemistry: Theoretical background and practical guidelines. **Microchemical Journal**, v. 162, p. 105725, 2021.

ONDRASEK, Gabrijel; BEGIć, Helena Bakić; ZOVKO, Monika; FILIPOVIć, Lana; MERIÑO-GERGICHEVICH, Cristian; SAVIć, Radovan; RENGEL, Zed. Biogeochemistry of soil organic matter in agroecosystems & environmental implications. **Science of The Total Environment**, v. 658, p. 1559-1573, 2019.

OWENS, P. R.; RUTLEDGE, E. M. Morphology. In: HILLEL, D.; HATFIELD, J. L. **Encyclopedia of Soils in The Environment**. 1. ed. Oxford: Elsevier/Academic Press, p.511-520, 2005.

OZAKI, Yukihiro; HUCK, Christian; TSUCHIKAWA, Satoru; BALLING ENGELSEN, Søren (Eds.). **Near-infrared spectroscopy: theory, spectral analysis, instrumentation, and applications**. Singapore: Springer Nature, 2021.

PANCHUK, V.; YAROSHENKO, I.; LEGIN, A.; SEMENOV, V.; KIRSANOV, D. Application of chemometric methods to XRF-data – A tutorial review. **Analytica Chimica Acta**, v. 1040, p. 19–32, 2018.

PARIKH, S. J.; GOYNE, K. W.; MARGENOT, A. J.; MUKOME, F. N. D.; CALDERÓN, F. J. **Soil Chemical Insights Provided through Vibrational Spectroscopy**. In: Advances Agronomy, 2014, p.1-148.

PASQUINI, C. Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications. **Journal of the Brazilian Chemical Society**, v. 14, n. 2, p. 198–219, 2003.

PERKIN-ELMER. **Universal ATR Sampling Accessory User's Guide**. Disponível em:<https://www.s4science.at/wordpress/wp-content/uploads/2019/04/L105023.pdf>;. Acesso em: 15 jan. 2022.

PERNAMBUCO. CONDEPE/FIDEM. (Comp.). **Base de Dados do Estado**. Disponível em: <http://www.bde.pe.gov.br>. Acesso em: 20 dez. 2021.

PINCUS, L.N.; RYAN, P.C.; HUERTAS, F.J.; et al. The influence of soil age and regional climate on clay mineralogy and cation exchange capacity of moist tropical soils: A case study from Late Quaternary chronosequences in Costa Rica. **Geoderma**, v. 308, p. 130–148, 2017.

RAPHAEL, L. **Application of FTIR Spectroscopy to Agricultural Soils Analysis**. In: NIKOLIC, G. Fourier Transforms - New Analytical Approaches and FTIR Strategies. Janeza Trdine: InTech, 2011.

REYNKENS, T.; TODOROV, V.; HUBERT, M.; SCHMITT, E.; VERDONCK, T. **rospca: Robust Sparse PCA using the ROSPCA Algorithm**. R Package Vignette, Report No.: R Package Version 1.0.4, 2018.

REYNOLDS, A. P.; RICHARDS, G.; DE LA IGLESIA, B.; RAYWARD-SMITH, V. J. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. **Journal of Mathematical Modelling and Algorithms**, v. 5, n. 4, p. 475–504, 2006

RIAL, M.; CORTIZAS, A. M.; RODRÍGUEZ-LADO, L. Mapping soil organic carbon content using spectroscopic and environmental data: A case study in acidic soils from NW Spain. **Science of The Total Environment**, v. 539, p. 26–35, 2016.

RIAL, M.; MARTÍNEZ CORTIZAS, A.; TABOADA, T.; RODRÍGUEZ-LADO, L. Soil organic carbon stocks in Santa Cruz Island, Galapagos, under different climate change scenarios. **CATENA**, v. 156, p. 74–81, 2017.

RIBEIRO, M. R. **Caracterização e classificação dos solos de referência do Estado de Pernambuco**. Recife: Universidade Federal Rural de Pernambuco, 1999.

RIBEIRO, B. T.; SILVA, S. H. G.; SILVA, E. A.; GUILHERME, L. R. G. Portable X-ray fluorescence (pXRF) applications in tropical Soil Science. **Ciência e Agrotecnologia**, v. 41, n. 3, p. 245–254, 2017.

RIGAKU. **Rigaku Nex DE**. Disponível em: < https://www.addspex.com/rigaku-nex-de.html>; Acesso em: 15 jan. 2022.

ROBINSON, J. W.; FRAME, E. M. S.; FRAME II, G. M. **Undergraduate Instrumental Analysis**. 7. ed. New York: M. Dekker, 2014.

ROSSEL, R. A. V.; BEHRENS, T.; BEN-DOR, E.; BROWN, D. J.; DEMATTÊ, J. A.; SHEPERD, K. D.; SHI, Z.; STENBERG, B. STEVENS, A.; ADAMCHUK, V.; AÏCHI, H.; BARTHÈS, B. G.; BARTHOLOMEUS, H. M.; BAYER, A. D.; BERNOUX, M.; BÖTTCHER, K.; BRODSKÝ, L.; DU, C. W.; CHAPPELL, A.; FOUAD, Y.; GENOT, V.; GOMEZ, C.; GRUNWALD, S.; GUBLER, A.; GUERRERO, C.; HEDLEY, C. B.; KNADEL, M.; MORRÁS, H. J. M.; NOCITA, M.; RAMIREZ-LOPEZ, L.; ROUDIER, P.; CAMPOS, E. M. R.; SANBORN, P.; SELLITO, V. M.; SUDDUTH, K. A.; RAWLINS, B. G.; WALTER, C.; WINOWIECKI, L. A.; HONG, S. Y.; JI, W . A global spectral library to characterize the world's soil. **Earth-science Reviews**, v. 155, p. 198-230, 2016.

SÁ, E. S. S. C. G. **Caracterização de solos e pedogênse em lagoa temporária no semiárido brasileiro**. Dissertação de Mestrado. Recife, PE: UFRPE, 2018.

SANTANA, M. B. **Atributos do solo relacionados com o potencial enológico de uvas viníferas no Vale do Submédio São Francisco**. Tese de Doutorado. Recife, PE: UFRPE, 2019.

SANTOS, R. D.; LEMOS, R. C.; SANTOS, H. G.; KER, J. C.; ANJOS, L. H. C.; SHIMIZU, S. H. **Manual de descrição e coleta de solo no campo**. 7. ed. Viçosa: Sociedade Brasileira de Ciência do Solo, 2015.

SANTOS, Uemeson José dos; DEMATTÊ, José Alexandre de Melo; MENEZES, Rômulo Simões Cezar; et al. Predicting carbon and nitrogen by visible near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy in soils of Northeast Brazil. **Geoderma Regional**, v. 23, p. e00333, 2020.

SANTOS, Felipe Rodrigues dos; DE OLIVEIRA, José Francirlei; BONA, Evandro; et al. Evaluation of pre-processing and variable selection on energy dispersive X-ray fluorescence spectral data with partial least square regression: A case of study for soil organic carbon prediction. **Spectrochimica Acta Part B: Atomic Spectroscopy**, v. 175, p. 106016, 2021.

SCHULZE, D. G. An Introduction to Soil Mineralogy. In: DIXON, J. B.; WEED, S. B.; DINAUER, R. C. **Minerals in soil environments**. 2. ed. Madison: Soil Science Society of America, 1989. cap. 1, p. 1-33.

SILVA, Vicente de Paulo Rodrigues da. On climate variability in Northeast of Brazil. **Journal of Arid Environments**, v. 58, n. 4, p. 575–596, 2004.

SILVA I. R., MENDONÇA E. S. Matéria orgânica do solo. In: Novais R. F., Alvarez V. V. H., Barros N. F., Fontes R. L. F., Cantarutti R. B., Neves J. C.L. (editores). **Fertilidade do solo**. Viçosa, MG: Sociedade Brasileira de Ciência do Solo; 2007. p.275-374.

SILVA, Y. J. A. B. **Influence of geochemical signature and mineralogy of granites on the pedogenesis and geochemistry of soils across a climosequence**. Tese de Doutorado. Recife, PE: UFRPE, 2016.

SILVA, V. R. F. **Gênese de luvissolos ao longo de uma climossequência sobre anfibolito no semiárido de Pernambuco**. Dissertação de Mestrado. Recife, PE: UFRPE, 2018.

SILVA, S. H. G.; HARTEMINK, A. E.; TEIXEIRA, A. F. dos S.; INDA, A. V.; GUILHERME, L. R. G.; CURI, N. Soil weathering analysis using a portable X-ray fluorescence (PXRF) spectrometer in an Inceptisol from the Brazilian Cerrado. **Applied Clay Science**, v. 162, p. 27–37, 2018.

SILVA, E. A; WEIDORF, D. C.; SILVA, S. H. G.; RIBEIRO, B. T.; POGGERE, G. C.; CARVALHO, T. S.; GONÇALVES, M. G. M.; GUILHERME, L. R. G.; CURI, N. Advances in Tropical Soil Characterization via Portable X-Ray Fluorescence Spectrometry. **Pedosphere**, v. 29, n. 4, p.468-482, 2019.

SILVERSTEIN, R. M.; WEBSTER, F. X.; KIEMLE, D. J. **Spectrometric identification of organic compounds**. 7. ed. Hoboken: John Wiley & Sons, 2005.

SIMEONOV, V; STRATIS, J. A; SAMARA, C.; ZACHARIADIS, G.; VOUTSA, D.; ANTHEMIDIS, A.; SOFONIOU, M.; KOUMITZIS, T. Assessment of the surface water quality in Northern Greece. **Water Research**, v. 37, n. 17, p.4119-4124, 2003.

SIRVEN, Jean-Baptiste; SALLÉ, Béatrice; MAUCHIEN, Patrick; LACOUR, Jean-Luc; MAURICE, Sylvestre; MANHÈS, Gérard. Feasibility study of rock identification at the surface of Mars by remote laser-induced breakdown spectroscopy and three chemometric methods. **Journal of Analytical Atomic Spectrometry**, vol. 22, no. 12, p. 1471, 2007.

SOUSA, J. E. S. **Gênese e mineralogia de planossolos sobre o maciço da Borborema, região semiárida de Pernambuco**. Dissertação de Mestrado. Recife, PE: UFRPE, 2015.

SOUSA, M. G. F. **Alterações nos atributos de Luvissolos sob irrigação na região do submédio do rio São Francisco, Pernambuco**. Dissertação de Mestrado. Recife, PE: UFRPE, 2018.

SOUSA, J. E. S. **Pedogênese e sequestro de carbon em Planossolos ao longo de um gradiente pluviométrico na Província da Borborema, Brasil**. Tese de Doutorado. Recife, PE: UFRPE, 2019.

SOUZA, Diego Mendes de; MORAIS, Pedro Augusto de Oliveira; MATSUSHIGE, Ivã; ROSA, Luismar Alves. Development of Alternative Methods for Determining Soil Organic Matter. **Revista Brasileira de Ciência do Solo**, v. 40, p. 1-17, 2016

SPOSITO, G. **Chemistry of Soils**. 2. ed. Nova Iorque: Oxford University Press, 2008.

STEVENS, Antoine; RAMIREZ-LOPEZ, Leonardo. **An introduction to the prospectr package**. R Package Vignette, Report No.: R Package Version 0.1, v. 3, 2014.

STRAWN, D. G.; HINRICH, L. B.; CONNOR, G. A. **Soil Chemistry**. 4. ed. Hoboken: Wiley, 2015.

STUART, Barbara. **Infrared spectroscopy: fundamentals and applications**. Chichester, West Sussex, England ; Hoboken, NJ: J. Wiley, 2004.

TAVARES, Tiago Rodrigues; MOLIN, José Paulo; JAVADI, S. Hamed; CARVALHO, Hudson Wallace Pereira de; MOUAZEN, Abdul Mounem. Combined Use of Vis-NIR and XRF Sensors for Tropical Soil Fertility Analysis: assessing different data fusion approaches. **Sensors**, v. 21, n. 1, p. 1-23, 2020.

TAYLOR; FRANCIS. **Pratical guide to chemometrics**. 2. ed. Nova York: CRC, 2006.

TEIXEIRA, Anita Fernanda dos Santos; HENRIQUE PROCÓPIO PELEGRINO, Marcelo; MISSINA FARIA, Wilson; et al. Tropical soil pH and sorption complex prediction via portable X-ray fluorescence spectrometry. **Geoderma**, v. 361, p. 114132, 2020.

TERRA, F. S.; DEMATTÊ, J. A. M.; ROSSEL, R. A. V. Spectral Libraries for Quantitative Analyses of Tropical Brazilian Soils: Comparing Vis–NIR and Mid-IR Reflectance Data. **Geoderma**, nº 255-256,81-93, 2015.

TERRA, Fabrício S.; DEMATTÊ, José A.M.; VISCARRA ROSSEL, Raphael A. Proximal spectral sensing in pedological assessments: vis–NIR spectra for soil classification based on weathering and pedogenesis. **Geoderma**, vol. 318, p. 123–136, 2018.

TEXAS INTRUMENTS. **DLP® NIRscan™ Nano EVM User's Guide**. Disponível em: < https://www.ti.com/lit/ug/dlpu030g/dlpu030g.pdf?ts=1649259084115>; Acesso em: 15 jan. 2022.

TINTI, A.; TUGNOLI, V.; BONORA, S.; FRANCIOSO, O. Recent applications of vibrational mid-Infrared (IR) spectroscopy for studying soil components: a review. **Journal of Central European Agriculture**, v. 16, n. 1, p.1-22, 2015.

TÜMSAVAŞ, Zeynal; TEKIN, Yücel; ULUSOY, Yahya; et al. Prediction and mapping of soil clay and sand contents using visible and near-infrared spectroscopy. **Biosystems Engineering**, v. 177, p. 90–100, 2019.

VIAVI SOLUTIONS. **VIAVI MicroNIR™ 1700 ES**. Disponível em: < https://www.viavisolutions.com/en-us/literature/micronir-1700es-data-sheets-en.pdf>; Acesso em: 15 jan. 2022.

WALKLEY, Aldous; BLACK, I. Armstrong. An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. **Soil science**, v. 37, n. 1, p. 29-38, 1934.

WEHRENS, Ron; MEVIK, B. H. **The pls package: principal component and partial least squares regression in R**. 2007.

WIESMEIER, M.; URBANSKI, L.; HOBLEY, E.; LANG, B.; VON LÜTZOW, M.; MARIN-SPIOTTA, E.; VAN WESEMAEL, B.; RABOT, E.; LIESS, M.; GARCIA-FRANCO, N.; WOLLSCHLÄGER, U.; VOGEL, H. J.; KÖGEL-KNABNER, I. Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales. **Geoderma**, v. 333, p. 149–162, 2019.

WOLD, S.; ANTTI, H.; LINDGREN, F.; ÖHMAN, J. Orthogonal signal correction of near-infrared spectra. **Chemometrics and Intelligent Laboratory Systems**, v. 44, n. 1–2, p. 175–185, 1998.

XIAO, Y.; YU, J. Partitive clustering (K-means family**). WIREs Data Mining and Knowledge Discovery**, v. 2, n. 3, p. 209–225, 2012.

YU, T. R. **Chemistry of variable charge soils**. New York: Oxford University Press, 1997.

ZHANG, Y.; HARTEMINK, A. E.; HUANG, J. Spectral signatures of soil horizons and soil orders – An exploratory study of 270 soil profiles. **Geoderma**, v. 389, p. 114961, 2021.

ZINN, Y. L.; LAL, R.; BIGHAM, J. M.; RESCK, D. V. S. Edaphic Controls on Soil Organic Carbon Retention in the Brazilian Cerrado: Texture and Mineralogy. **Soil Science Society of America Journal**, v. 71, n. 4, p. 1204–1214, 2007.

APÊNDICE A – INFORMAÇÕES SOBRE AS AMOSTRAS

Solo	Município	Longitude	Latitude	Horizonte (cm)	Categoria
Argissolo vermelho-amarelo	Itambé	-35,1817	-7,42278	A1 (0-40)	1
				Bt1 (155-250)	1
Argissolo amarelo	Goiana	-34,9583	-7,625	A (0-14)	1
				Bt1 (70-115)	1
Argissolo amarelo	Moreno	-35,1117	-8,11167	A (0-13)	1
				Bt1 (46-88)	1
Argissolo vermelho	Camutanga	-35,2731	-7,42139	Ap (0-17)	4
				Bt2 (105-138)	4
Argissolo amarelo	Aliança	-35,1453	-7,60556	Ap (0-28)	1
				Btx (112-160)	1
Argissolo vermelho	Riacho das Almas	-35,9097	-8,1725	Ap (0-16)	4
				Bt1 (35-68)	1
Argissolo amarelo	Vertentes	-35,9217	-7,87972	Ap (0-9)	2
				Bt (45-55)	1
Argissolo vermelho-amarelo	Jupi	-36,4167	-8,7125	Ap (0-12)	2
				Bt1 (128-178)	1
Argissolo amarelo	Brejão	-36,4833	-9,05	A (0-15)	1
				Btx (135-200)	1
Argissolo amarelo	Santa Cruz	-40,2767	-8,32639	A (0-8)	2
				Bt (24-44)	1

Solo	Município	Longitude	Latitude	Horizonte (cm)	Categoria
Argissolo vermelho-amarelo	Lagoa do Ouro	-36,466	-9,13494	A1 (0-9)	4
				Bt (34-50)	1
Argissolo vermelho	Pombos	-35,4047	-8,25779	Ap (0-8)	1
				Bt1 (30-46)	1
Argissolo amarelo	Sirinhaém	-35,0779	-8,54278	Ap (0-18)	1
				Bt1 (45-70)	1
Argissolo amarelo	Cabo de Santo Agostinho	-35,0239	-8,25889	A (0-19)	4
				Bt (50-75)	3
Argissolo vermelho-amarelo	Sirinhaém	-35,1357	-8,56625	Ap (0-16)	2
				Bt1 (16-34)	1
Argissolo vermelho-amarelo	Sirinhaém	-35,0858	-8,56536	Ap (0-22)	1
				Bt (35-70)	1
Argissolo vermelho-amarelo	Lagoa Grande	-40,1989	-9,05092	Apf (0-30)	2
				Btf1 (30-75)	1
Argissolo amarelo	Lagoa Grande	-40,1995	-9,05058	Ap (0-20)	1
				Bt (37-80)	1
Argissolo amarelo	Lagoa Grande	-40,2006	-9,05022	Ap (0-20)	2
				Bt (70-120)	1
Argissolo amarelo	Petrolina	-40,3072	-9,06278	A (0-6)	2
				Bt1 (28-50)	1
Argissolo acinzentado	Cabo de Santo Agostinho	-35,0233	-8,25944	A (0-17)	4
				Bt1 (30-50)	3

Solo	Município	Longitude	Latitude	Horizonte (cm)	Categoria
Argissolo acinzentado	Cabo de Santo Agostinho	-35,0228	-8,26	A (0-25)	4
				Bt (70-115)	3
Cambissolo háplico	Santa Cruz da Baixa Verde	-38,1375	-7,80806	Ap (0-15)	4
				Bi1 (15-45)	4
Cambissolo háplico	Vertente do Lério	-35,8639	-7,76942	Apk (0-15)	3
				Bivnkz (32-80)	3
Cambissolo háplico	Passira	-35,5086	-8,00375	A (0-10)	4
				Bi (33-58)	3
Cambissolo háplico	Passira	-35,5041	-8,0315	A (0-10)	4
				Bi (22-48)	3
Cambissolo háplico	Sirinhaém	-35,0861	-8,56764	A (0-9)	2
				Bi1 (27-39)	1
Chernossolo argilúvico	Nazaré da Mata	-35,2344	-7,74222	Ap (0-30)	2
				Bt (57-80)	3
Espodossolo humilúvico	Goiana	-34,9456	-7,64472	Ap (0-15)	2
				Bh (90-112)	2
Gleissolo háplico	Ipojuca	-35,0206	-8,42278	Ap (0-20)	4
				Cg1 (42-70)	3
Gleissolo háplico	Parnamirim	-39,9011	-8,28464	A (0-5)	3
				Cg (20-40)	3
Gleissolo háplico	Ipojuca	-35,0123	-8,40997	Agn (5-15)	3
				Cgn1 (15-54)	3

Solo	Município	Longitude	Latitude	Horizonte (cm)	Categoria
Gleissolo sálico	Goiana	-34,9646	-7,55511	Ap (0-16)	3
				Cgnz2 (34-50)	3
Gleissolo háplico	Sirinhaém	-35,0698	-8,56683	Ap (0-10)	1
				Cg2 (66-100)	1
Latossolo amarelo	Rio Formoso	-35,1542	-8,64417	A (0-10)	4
				Bw2 (70-126)	1
Latossolo vermelho-amarelo	Caruaru	-36,1744	-8,23861	A1 (0-12)	1
				Bw1 (70-110)	1
Latossolo amarelo	Trindade	-40,2031	-7,7425	A (0-20)	1
				Bw1 (45-77)	1
Latossolo amarelo	Pombos	-35,4036	-8,25678	A1 (0-9)	1
				Bw1 (50-69)	1
Latossolo amarelo	Itambé	-35,181	-7,42311	A1 (0-20)	1
				Bw (155-200)	1
Latossolo amarelo	Paudalho	-35,2048	-7,86617	A1 (0-20)	1
				Bw (145-180)	1
Latossolo vermelho-amarelo	Caruaru	-36,0278	-8,37314	A1 (0-20)	1
				Bw (120-180)	1
Latossolo amarelo	Brejão	-36,5506	-8,99406	A1 (0-15)	1
				Bw (130-195)	1
Luvissolo crômico	Cabrobó	-39,3275	-8,50278	A (0-23)	4
				2Bt (23-46)	4

Solo	Município	Longitude	Latitude	Horizonte (cm)	Categoria
Luvissolo crômico	Belém de São Francisco	-38,8583	-8,75194	A (0-9)	3
				Btv (22-45)	3
Luvissolo crômico	Belém de São Francisco	-38,8583	-8,74058	A (0-12)	4
				Btvnz (12-45)	3
Luvissolo crômico	Itacuruba	-38,7742	-8,74683	A (0-9)	4
				Btvz (9-40)	3
Luvissolo crômico	Serra Talhada	-38,2363	-7,91575	A (0-12)	4
				Btv (25-48)	3
Luvissolo crômico	Afogados da Ingazeira	-37,6369	-7,73364	A (0-10)	4
				Bt (23-47)	3
Luvissolo crômico	Itacuruba	-38,8551	-8,73042	A (0-7)	4
				Btv (18-55)	3
Luvissolo crômico	Itacuruba	-38,8546	-8,73144	A (0-8)	4
				Btvz (35-48)	3
Luvissolo crômico	Itacuruba	-38,8559	-8,74228	A (0-9)	4
				Btv2 (30-48)	3
Neossolo litólico	Caruaru	-36,0561	-8,07194	A1 (0-8)	4
				A2 (8-16)	4
Neossolo litólico	Verdejante	-38,8942	-8,02639	A1 (0-8)	4
				AC (8-23)	3
Neossolo quartzarênico	Petrolina	-40,4769	-9,35361	A (0-15)	2
				C1 (28-70)	2

Solo	Município	Longitude	Latitude	Horizonte (cm)	Categoria
Neossolo quartzarênico	Ibimirim	-37,6678	-8,63861	A (0-11)	2
				C1 (11-45)	2
Neossolo flúvico	Ibimirim	-37,5242	-8,72167	Azn (0-11)	3
				2Czn2 (40-85)	3
Neossolo flúvico	Ibimirim	-37,6233	-8,50722	Ap (0-25)	4
				3Czn (38-79)	3
Neossolo regolítico	São Caetano	-36,1739	-8,36167	Ap (0-20)	4
				C (58-135)	4
Neossolo regolítico	Caruaru	-35,9208	-8,23167	A (0-15)	4
				AC (15-30)	4
Neossolo regolítico	Panelas	-36,0194	-8,66308	A (0-11)	4
				C1 (17-46)	4
Neossolo regolítico	Jurema	-36,1532	-8,72527	A (0-16)	2
				C1 (56-93)	4
Neossolo regolítico	Betânia	-38,0601	-8,30631	A (0-3)	4
				C1 (9-17)	4
Neossolo regolítico	Águas Belas	-36,9848	-9,08575	Ap (0-5)	4
				C1 (17-26)	4
Nitossolo háplico	Sirinhaém	-35,0805	-8,5425	Ap (0-8)	5
				Bt1 (42-95)	1
Nitossolo háplico	Sirinhaém	-35,0799	-8,54247	Ap (0-20)	5
				Bt1 (55-100)	5

Solo	Município	Longitude	Latitude	Horizonte (cm)	Categoria
Nitossolo háplico	Sirinhaém	-35,0793	-8,54247	Ap (0-20)	5
				Bt1 (50-90)	5
Nitossolo háplico	Cabo de Santo Agostinho	-35,0239	-8,31194	Ap (0-15)	5
				Bt (30-63)	5
Nitossolo háplico	Cabo de Santo Agostinho	-35,0236	-8,31194	A (0-14)	5
				Bt (38-80)	5
Planossolo háplico	São Caetano	-36,1722	-8,35778	Ap (0-15)	4
				2Bt (78-105)	4
Planossolo háplico	São Caetano	-36,2136	-8,45722	Ap (0-11)	4
				2Bt (11-45)	3
Planossolo nátrico	Brejo da Madre de Deus	-36,1839	-7,97278	A (0-14)	4
				2Btn (14-45)	3
Planossolo nátrico	Petrolina	-40,2519	-9,04167	AE (0-15)	2
				Btn1 (30-55)	1
Planossolo háplico	Parnamirim	-39,9028	-8,28267	A1 (0-9)	2
				Bt (20-50)	3
Planossolo háplico	Lagoa do Ouro	-36,4839	-9,15144	Ap (0-13)	4
				Bt (44-55)	3
Planossolo háplico	Lagoa do Ouro	-36,4774	-9,14161	A (0-15)	4
				Btg (58-74)	3
Planossolo háplico	Lagoa do Ouro	-36,4749	-9,13375	A (0-22)	4
				Btg1 (94-111)	4

Solo	Município	Longitude	Latitude	Horizonte (cm)	Categoria
Planossolo háplico	Serra Talhada	-38,3742	-7,94478	A (0-10)	4
				Bt (21-48)	3
Planossolo háplico	Itacuruba	-38,8034	-8,75081	A1 (0-10)	4
				Bt1 (20-40)	3
Planossolo háplico	Itacuruba	-38,6843	-8,81569	A (0-12)	4
				Btn (12-30)	3
Planossolo háplico	Itacuruba	-38,8607	-8,70208	A (0-9)	4
				2Btn (26-52)	3
Planossolo nátrico	Itacuruba	-38,8555	-8,74189	A (0-10)	4
				2Btn (30-60)	3
Planossolo nátrico	Belém de São Francisco	-38,8556	-8,74186	A (0-8)	4
				Btn (26-67)	3
Planossolo nátrico	Camutanga	-35,2593	-7,41903	Ap1 (0-10)	4
				2Btn1 (48-72)	3
Plintossolo argilúvico	Goiana	-35,057	-7,58635	Ap (0-15)	2
				2Bt (75-93)	1
Vertissolo háplico	Bodocó	-40,0539	-7,62889	Ap (0-25)	3
				Biv (25-65)	3
Vertissolo hidromórfico	Parnamirim	-39,9041	-8,28564	Ag (0-10)	3
				Bgv (10-40)	3
Vertissolo ebânico	Cachoeirinha	-36,2403	-8,49806	Ap (0-8)	3
				Biv (20-68)	3