Avaliação da viabilidade do pré-processamento de espectros para predição de atributos do solo em tempo real

Ricardo Canal Filho[1]; Marcelo Chan Fu Wei[2]; José Paulo Molin[3]
[1]Engº Agrônomo, Mestrando - Depto. Engenharia de Biossistemas, ESALQ, USP, Piracicaba - SP. ricardocanal@usp.br / (16) 9 9262-8917; [2]Engº Agrônomo, Doutorando - Depto. Engenharia de Biossistemas, ESALQ, USP, Piracicaba - SP; [3]Engº Agrícola, Professor Doutor - Depto. Engenharia de Biossistemas, ESALQ, USP, Piracicaba - SP

Apresentado no
Congresso Brasileiro de Agricultura de Precisão- ConBAP 2022
Campinas, SP, 09 a 11 de agosto de 2022

RESUMO: Na ciência do solo, a região do infravermelho próximo, em inglês *near infrared* (NIR) é a mais estudada para predição de atributos de interesse agronômico na área da espectroscopia de reflectância difusa, em inglês *diffuse reflectance spectroscopy* (DRS). Os principais objetivos com o uso da técnica é de aumentar a densidade amostral para melhor caracterizar a variabilidade dos atributos do solo. Para isso a técnica precisa ser levada ao campo e tem que ser capaz de lidar com os desafios que a operação de campo apresenta. Para tratar as interferências na leitura, muitos autores têm utilizado técnicas de pré-processamento dos espectros buscando reduzir os ruídos e auxiliar os modelos multivariados a quantificar os atributos de interesse. Entretanto, é limitada a abordagem sobre a efetiva melhora dessa abordagem na predição, e o quanto a adição dessas etapas significa para o processamento dos dados. Desta forma, este estudo propôs proceder essa avaliação a partir da coleta em tempo real de 383 espectros NIR em uma lavoura experimental. Os dados espectrais foram organizados em três conjuntos de dados: dados brutos (DB), sequência de pré-processamento 1 (PP1) e sequência de pré-processamento 2 (PP2). Cada conjunto foi utilizado para predizer, via regressão parcial por mínimos quadrados, em inglês *partial least squares regression* (PLSR), seis atributos-chave de fertilidade do solo: argila, areia, matéria orgânica (MO), capacidade de troca catiônica (CTC), pH e potássio (K). Os modelos calibrados apresentaram bons parâmetros de acurácia para quase todos os atributos avaliados, com $R^2 < 0,60$ apenas para a predição de pH. Apesar disso, considerando a variação dos atributos em campo, os erros de predição de todos os modelos foram consideravelmente baixos. A predição a partir dos espectros de campo pré-processados não diferiu estatisticamente da predição com DB. Contudo, observou-se aumento médio de 200% no tempo demandado de processamento dos dados para predição. O uso de espectros brutos foi a mais eficiente estratégia para uso da DRS NIR para predição de atributos do solo em tempo real.

PALAVRAS-CHAVE: quimiometria; espectroscopia no infravermelho próximo; aprendizado de máquina

EVALUATION OF THE FEASIBILITY OF SPECTRA PRE-PROCESSING TECHNIQUES FOR ONLINE SOIL ATTRIBUTES PREDICTION

ABSTRACT: In soil science, the near-infrared region (NIR) is the most studied for predicting attributes of agronomic interest using diffuse reflectance spectroscopy (DRS). The main objectives with the use of the technique is to increase sampling density in order to successfully represent the spatial variability of soil attributes. For this, the technique needs to be taken to the field, being able to deal with the challenges that a field operation can offer. To deal with spectra measure interference, many authors have used spectra pre-processing techniques, in order to reduce noise and aid multivariate models to quantify the attributes of interest. However, it is limited an approach of evaluating if there is an effective improvement in the prediction, and how much the addition of these steps means for data processing. Thus, this study proposed to proceed this evaluation, using 383 online NIR spectra acquired in an experimental field. Spectral data were organized into three datasets: raw data (DB), pre-processing sequence 1 (PP1) and pre-processing sequence 2 (PP2). Each dataset was used to predict, via partial least squares regression (PLSR), six soil fertility key attributes: clay, sand, organic matter (MO), cation exchange capacity (CTC), pH and potassium (K). The calibrated models showed good parameters of accuracy for almost all attributes evaluated, with $R^2 < 0.60$ only for pH prediction. Despite this, considering the variation of each attribute in the field, the prediction errors of all models were considerably low. The prediction using the pre-processed field spectra did not differ statistically from the prediction with raw spectra. However, an average increase of 200% was observed in the time demanded of

processing data for prediction. The use of raw spectra was the most efficient strategy for the proposed use of DRS NIR of online prediction of soil attributes.

KEYWORDS: chemometrics; near-infrared spectroscopy; machine learning

INTRODUÇÃO: A espectroscopia de reflectância difusa, em inglês *diffuse reflectance spectroscopy* (DRS) na região do infravermelho próximo, em inglês *near infrared* (NIR) tem demonstrado potencial de aplicação na agricultura para predição de atributos do solo. Essa faixa do espectro eletromagnético exprime interações primárias e secundárias da energia com os atributos do solo (Stenberg et al., 2010). Tais interações ocorrem na forma de absorção, reflexão ou transmissão da energia, e podem ser relacionadas com os atributos do solo em quantidade e qualidade (Nocita et al., 2015).

Estabelecida no meio acadêmico, a DRS NIR tem seu uso documentado principalmente em laboratório. Contudo, um dos desafios para que a técnica contribua efetivamente é torna-la capaz de atuar coletando espectros em tempo real, no campo, a fim de aumentar a densidade amostral dos dados de solo. As densidades amostrais praticadas na maioria das lavouras são comprovadamente ineficazes para identificar adequadamente a variabilidade dos atributos (Wollenhaupt et al., 1994; Montanari et al., 2012; Cherubin et al., 2014; Cherubin et al., 2015), o que resulta em tomadas de decisão questionáveis e consequente gestão ineficiente dos recursos utilizados na produção agrícola.

Pesquisadores têm se dedicado ao uso dos espectros NIR combinados à aplicação de modelos estatísticos multivariados para predição dos atributos de solo (Pasquini et al., 2018). O avanço no uso de métodos de aprendizado de máquina, em inglês *machine learning* (ML) e o uso intensivo da inteligência artificial, vem impulsionando a aplicação de métodos estatísticos multivariados para modelos de predição. É comum a aplicação de técnicas de pré-processamento do espectro antes de inseri-los nas calibrações de ML (Franceschini et al., 2018; Munnaf et al., 2021a; Zhang et al., 2021). Essas técnicas possuem o objetivo de remover ruídos, enfatizar feições e extrair informações úteis para os modelos de predição (Dotto et al., 2018). Entretanto, seu uso representa maior demanda de processamento dos dados, fato relevante para aplicações em tempo real. Ainda, alguns dos principais modelos estatísticos utilizados nesta área, como é a regressão parcial por mínimos quadrados (PLSR), são técnicas de redução de dimensionalidade dos dados, auxiliando na remoção de dados ruidosos, redundantes e irrelevantes, semelhante ao objetivo das técnicas de pré-processamento (Velliangiri & Alagumuthukrishnan, 2019). Dessa forma, compreende-se que a viabilidade da aplicação de técnicas de pré-processamento é muitas vezes negligenciada. Com o objetivo de aumentar a densidade de informações acerca do solo, o banco de dados coletado em lavouras comerciais aumenta também e um processamento eficiente é necessário para permitir o uso da técnica em larga escala. Neste sentido, este trabalho se propõe a avaliar a viabilidade de diferentes formas de processamento na construção de modelos de ML para predição de atributos do solo com espectros NIR coletados diretamente em campo, aferindo o custo de processamento e os resultados obtidos pelas diferentes calibrações testadas.

MATERIAIS E MÉTODOS:

Foi utilizada uma área experimental do departamento de Engenharia de Biossistemas da ESALQ/USP, em Piracicaba, SP, com coordenadas centrais 22º43’03.51”S e 47º36’50.03”O, com aproximadamente 6,0 hectares. Nos últimos três anos houve o cultivo de soja na safra de verão e período de pousio durante o inverno. A textura do solo é classificada como franco-arenosa.

Em 2021 foi realizada a coleta de dados espectrais em tempo real utilizando uma estrutura montada no engate hidráulico de três pontos do trator. Nessa estrutura foi acoplado uma haste escarificadora atuando a 0,15 m de profundidade no solo e o fundo do sulco é suavizado pela ponteira da haste. Um compartimento metálico é acoplado na parte traseira da haste por um conjunto pantográfico, e dentro dele é alojado o espectrômetro MicroNIR OnSite-W (Viavi Solutions Inc., California, EUA) que coletou espectros de solo em tempo real na resolução espectral de 908,1-1676,2 nm, com leituras a cada 6,2 nm, resultando em 125 diferentes comprimentos de onda (Figura 1). Os espectros são coletados na base do compartimento, através de uma janela de safira, transportados por um cabo USB e convertidos para transmissão por meio de um cabo Ethernet, gravados em um computador portátil conectado ao conjunto. Uma placa de 99% reflectância foi usada como referência do branco (máxima reflectância), e o próprio equipamento possui medição interna de referência do

preto (mínima reflectância). Cada espectro coletado em campo foi associado às suas coordenadas geográficas com uso de um receptor de sistema de navegação global por satélite (GNSS) modelo Ag-Star (Novatel, Calgary, Canadá) com correção diferencial TerraStarC (Hexagon, São Paulo, Brasil).

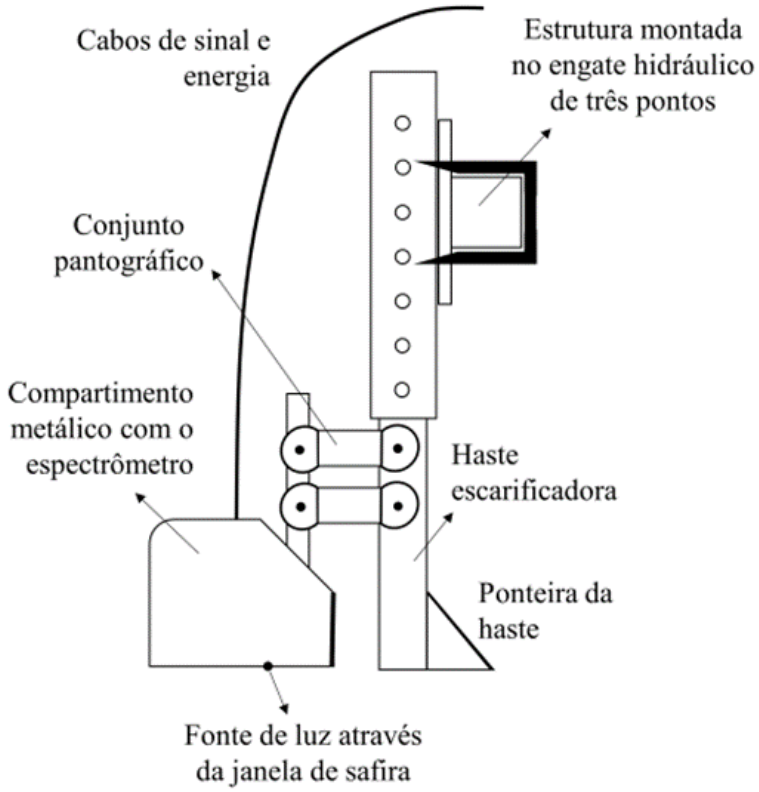

FIGURA 1. Estrutura montada no engate hidráulico de três pontos e haste subsoladora instrumentada com o espectrômetro MicroNIR. **Structure mounted on hydraulic three-point hitch and subsoiler shank with MicroNIR.**

O trator percorreu a área no sentido do tráfego de máquinas, limitado pela presença de terraços e com espaçamento de 12 m, o que resultou em 12 linhas de coleta de dados. Na coleta de espectros em tempo real, o equipamento agrupa amostras e faz uma análise de componentes principais, excluindo amostras que estão fora do limite de confiança estabelecido no software, e assim gera um espectro pela média. A densidade de coleta de espectros foi definida pelo tempo de coleta de 10 segundos à velocidade de 0,583 m s^{-1} (2,1 km h^{-1}), resultando em 383 espectros ao longo da área do experimento (Figura 2).

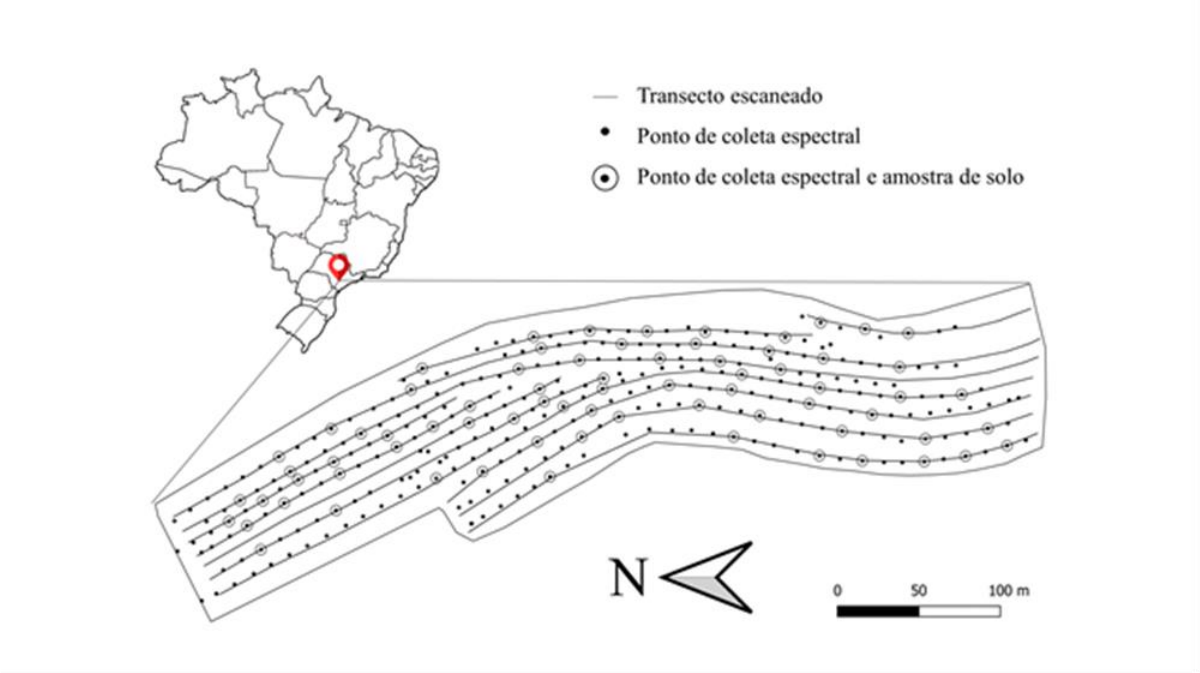

FIGURA 2. Localização da área experimental, em Piracicaba, SP, Brasil. O mapa mostra o contorno da área, os pontos pretos, associados a cada espectro, e os pontos circulados, demarcados para amostragem manual de solo. **Location of experimental area, Piracicaba, SP, Brazil. Plot show the area shape. Black dots alone represent each spectra acquired and black dots contoured represents a spectra associated with a soil sample.**

Um dia após a coleta dos espectros, 72 amostras de solo foram coletadas no fundo do sulco deixado pela haste sulcadora. Com o objetivo de criar calibrações com os espectros de campo, era necessário que cada amostra de solo coincidisse com o respectivo transecto escaneado pelo sensor. Para isso, demarcou-se o início de pontos amostrais apontado pelo software de aquisição de dados espectrais do solo. Após a coleta dos espectros, o produto entre a velocidade de deslocamento e tempo de coleta do software resultou no comprimento do transecto a ser coletado para cada amostra.

As análises físico-químicas do solo foram realizadas por um laboratório comercial. Os atributos utilizados neste estudo e os respectivos métodos de análise foram argila e areia – HMFS+NaOH, matéria orgânica (OM) - oxidação, capacidade de troca catiônica (CTC) – cálculo das bases mais acidez total, pH – $CaCl_2$ e potássio trocável (K) - resina.

O processamento dos dados de espectros ocorreu no ambiente de programação do software Jupyter Notebook (Kluyver et al., 2016; Python Software Foundation, 2022). Foi conduzida uma análise descritiva e observação espacial dos pontos, para filtrar apenas espectros onde houve um erro claro de medição por alguma condição durante a coleta, como manobras de cabeceira onde o sensor permaneceu ligado, ou perda de calibração do espectrômetro. Essa análise resultou na retirada de 77 espectros, permanecendo 306 no banco de dados.

A estes dados foram adicionados espectros coletados em bancada, de amostras provenientes de A junção de espectros de campo com espectros de bancada no mesmo banco de dados para calibrar modelos de ML é uma estratégia que pode auxiliar no desempenho preditivo do modelo (Munnaf et al., 2019). Ainda, é comprovado que aumentar o número de observações no banco de dados de amostras de solo na calibração de modelos de ML pode auxiliar a predição, sejam essas amostras provenientes da mesma região geográfica ou não (Guerrero et al., 2021).

A fim de comparar o desempenho preditivo com a aplicação de técnicas de pré-processamento, foram criados três bancos de dados: dados brutos (DB), pré-processamento 1 (PP1) e pré-processamento 2 (PP2). Diferentes pré-processamentos foram utilizados para testar o desempenho dessas técnicas, sendo avaliadas por meio do desempenho dos modelos de predição, como proposto por Munnaf et al. (2019).

Os métodos de pré-processamento compreenderam a média móvel (MM), para redução do efeito de bandas ruidosas no espectro; a normalização pelo máximo (NM), que é um método para conformar os espectros na mesma escala e criar uma distribuição uniforme das variações (Rinnan et al., 2009); dois métodos de derivação, a primeira e segunda derivada de Savitzky-Golay (SG), utilizadas para redução de ruído e realçar feições espectrais fracas e possíveis informações ocultas (Ben Dor et al., 1995); e o algoritmo SG de suavização. Esses métodos foram testados em duas sequências (Tabela 1).

TABELA 1. Bancos de dados criados para calibração dos modelos. **Datasets created for models calibration.**

Acrônimo	Ordem de aplicação das técnicas de pré-processamento						
DB	----------						
PP1	MM	+	1ª derivada SG	+	2ª derivada SG	+	algoritmo SG
PP2	MM	+	NM (0,1)	+	1ª derivada SG	+	algoritmo SG

DB: dados brutos; PP1: sequência de pré-processamento 1; PP2: sequência de pré-processamento 2; MM: média móvel, NM: normalização pelo máximo; SG: Savitzky-Golay

Por conseguinte, cada conjunto de dados empregado para o ML foi separado na proporção 70% para calibração e 30% para validação, utilizando a validação cruzada k-fold, definindo-se k = 10 Esta técnica é recomendada para avaliar modelos de ML, pois geralmente resulta em uma estimativa menos tendenciosa ou menos otimista da habilidade do modelo do que outros métodos (Jung et al., 2018), e consiste basicamente nas etapas: 1) embaralhar o banco de dados (linhas) aleatoriamente; 2) dividir o banco de dados em um número K de grupos; 3) para cada grupo: separar um grupo para validação, utilizar os grupos restantes como um banco de treinamento, ajustar um modelo no conjunto de treinamento e avaliá-lo no conjunto de validação, manter a avaliação e descartar o modelo; 4) resumir a habilidade do modelo pela média das avaliações realizadas. Dessa forma, cada amostra é utilizada uma vez no conjunto de validação e K-1 vezes no conjunto de treinamento, garantindo uma estimativa menos tendenciosa da habilidade de predição do modelo.

Utilizou-se o modelo matemático da regressão parcial por mínimos quadrados, em inglês *partial least squares regression* (PLSR). O PLSR é um método comumente aplicado às análises quantitativas via espectros, pois permite lidar com grande número de variáveis preditoras. Baseia-se em reduzir a dimensão dos dados espectrais gerando novas variáveis, as variáveis latentes (VL). As combinações entre as VL geram o modelo de regressão linear (Kuang et al., 2015).

As métricas utilizadas para avaliação dos modelos preditivos foram o coeficiente de determinação (R^2), a raiz do erro quadrático médio (RMSE), o erro médio absoluto (MAE) e a relação de desempenho para distância interquartil (RPIQ). Quanto maiores os valores de R^2 e RPIQ, e menores o RMSE e MAE, subentende-se que melhor foi o desempenho dos modelos. O custo de processamento considerado neste trabalho foi aferido pelo tempo necessário para a máquina realizar uma operação, pela função *datetime* da biblioteca *datetime* do software utilizado (Python Software Foundation, 2022). Os dados foram processados em um notebook com as especificações: SSD NVMe M.2 256 Gb, processador IntelCore i5, memória RAM 8 Gb.

A fim de identificar quais estratégias diferiam estatisticamente e qual seria apontada como melhor desempenho, os valores preditos para cada atributo foram submetidos ao teste de Kruskal-Wallis a 95% de confiança. O teste de Kruskall-Wallis é um teste não-paramétrico, escolhido pelas populações de cada atributo não apresentarem normalidade de distribuição pelo teste de Shapiro-Wilk. A estatística Kruskall-Wallis é utilizada para testar se k grupos são semelhantes. A hipótese nula é de que os grupos são semelhantes, e a hipótese alternativa é que pelo menos um grupo difere dos demais. O teste apresenta os graus de liberdade, a estatística calculada, neste caso o chi-quadrado (X^2) e o p-valor calculado.

RESULTADOS E DISCUSSÃO: As métricas de avaliação dos modelos de PLSR criados com os três conjuntos de dados para predição dos atributos de solo foram semelhantes (Tabela 2). Para os atributos de

física do solo, argila e areia, em valores absolutos, o modelo de DB apresentou os melhores parâmetros de predição.

TABELA 2. Resultados da validação dos modelos preditivos calibrados com os três conjuntos de dados utilizados neste estudo, testando o uso de espectros de solo brutos e as duas sequências de pré-processamento dos espectros. **Results presented in validation of prediction models calibrated using the three datasets of this study, testing the use of raw soil spectral data and two different spectra pre-processing sequences.**

	DB					PP1					PP2				
	VL	R^2	RMSE	MAE	RPIQ	VL	R^2	RMSE	MAE	RPIQ	VL	R^2	RMSE	MAE	RPIQ
argila	**7**	**0,94**	**29,22**	**23,82**	**8,61**	4	0,93	31,68	23,78	8,41	3	0,80	57,04	40,54	4,94
areia	**10**	**0,97**	**39,12**	**30,19**	**12,22**	16	0,96	44,76	36,98	10,97	7	0,96	48,38	36,42	9,97
MO	10	0,77	3,15	2,45	3,97	8	0,78	2,94	2,37	4,08	**9**	**0,80**	**2,87**	**2,17**	**4,53**
CTC	5	0,65	15,58	10,75	1,68	**2**	**0,69**	**10,88**	**7,80**	**2,02**	3	0,56	16,58	11,09	1,57
pH	1	0,29	0,45	0,33	1,33	1	0,29	0,47	0,35	1,17	**3**	**0,48**	**0,37**	**0,28**	**1,20**
K	3	0,63	1,72	1,16	1,51	**3**	**0,81**	**1,44**	**1,04**	**4,07**	3	0,68	1,72	1,28	2,42

DB: dados brutos; PP1: sequência de pré-processamento 1; PP2: sequência de pré-processamento 2; VL: variáveis latentes/número de componentes utilizados na calibração do modelo; R^2: coeficiente de determinação; RMSE: raiz do erro quadrático médio; MAE: erro médio absoluto; RPIQ: relação de desempenho para distância interquartil; MO: matéria orgânica; CTC: capacidade de troca catiônica; K: potássio. Destacados em negrito estão os modelos com melhores indicadores por avaliação de valores absolutos.

Destaca-se o modelo de argila de PP2, que apresentou erros de predição – RMSE e MAE - quase duas vezes maiores do que ambos os modelos de DB e PP1. Para a química do solo, a MO, atributo de resposta direta no NIR, através das vibrações fundamentais (Nocita et al., 2015), o melhor modelo foi o de PP2. Entretanto, os parâmetros de avaliação apresentaram valores semelhantes. A CTC, o pH e o K, são atributos de resposta secundária no NIR. Portanto não se espera que eles apresentem feições de absorção diretas nessa região espectral, e o sucesso de suas predições depende da covariância com atributos de resposta primária (Stenberg et al., 2010). Esse fato se concretiza pela análise de correlação de Pearson dos atributos do conjunto de dados deste trabalho (Tabela 3), e observam-se modelos de CTC e K, com $R^2 > 0,60$ para praticamente todos os modelos, com exceção da predição de CTC por PP2. Destacam-se os modelos de PP1 para a predição destes atributos. O pH foi o atributo que apresentou pior desempenho de predição. Apesar dos baixos erros numéricos quando comparados à variação possível deste atributo no campo (3,0 a 9,0 - Lopes, 1989; Lopes et al., 1990), os valores de R^2 e RPIQ ficaram distantes dos observados para os demais atributos.

TABELA 3. Correlação de Pearson entre os atributos do conjunto de dados deste trabalho. **Pearson's correlation observed on attributes of dataset used in this study.**

	Argila	Areia	MO	CTC	pH	K
Argila	1					
Areia	-0,67	1				
MO	0,25	-0,50	1			
CTC	0,16	-0,44	0,56	1		

pH	0,12	0,28	-0,46	-0,05	1	
K	0,12	-0,17	0,42	0,23	-0,19	1

MO: matéria orgânica; CTC: capacidade de troca catiônica; K: potássio

A avaliação dos valores absolutos permitiria apontar o melhor modelo avaliado. Pode-se, a partir disso, recomendar o uso de uma estratégia – DB, PP1 e PP2 – para cada atributo, como é visto na literatura (Guerrero et al., 2021; Munnaf et al., 2021a). Apontar somente um caminho como o ideal não se torna possível, visto que cada estratégia apresentou, igualmente, dois modelos com melhores parâmetros. Entretanto, é comumente desconsiderado se os melhores parâmetros são fruto de uma variação ao acaso, ou se diferem estatisticamente.

Para aferir se algum modelo apresentou estatisticamente melhor desempenho, os modelos calibrados a partir dos conjuntos de dados DB, PP1 e PP2 foram extrapolados para predição dos atributos em área total, com os espectros de campo coletados na área experimental. Os valores preditos de cada atributo foram submetidos ao teste de Kruskall-Wallis (Tabela 4). Nenhum dos grupos diferiu significativamente de outros a 95% de confiança. Esses resultados denotam que as melhores métricas que apresentaram os modelos preditivos – argila e areia de DB, CTC e K de PP1 e OM e pH de PP2 – não podem ser atribuídas ao tratamento dos dados, mas ao acaso. Dessa forma, a indicação de um modelo de tratamento dos dados para predição dos atributos utilizados neste estudo não é fundamentada.

TABELA 4. Resultados do teste de Kruskall-Wallis para a comparação dos atributos físico-químicos do solo preditos a partir do uso de espectros coletados em campo. **Kruskall-Wallis test results for comparison of physicochemical soil attributes predicted atributes using online NIR spectra.**

	G.L.	X^2	p-valor
Argila	2	0,086	0,9577
Areia		1,502	0,4719
MO		3,708	0,1566
CTC		1,301	0,5217
pH		5,075	0,0791
K		0,041	0,9795

G.L.: graus de liberdade; X^2: qui-quadrado; MO: matéria orgânica; CTC: capacidade de troca catiônica; K: potássio

As técnicas de pré-processamento dos espectros são comumente utilizadas para predição de atributos do solo com DRS. Entretanto, os autores tendem a definir um melhor tratamento dos dados com base somente na observação das métricas apresentadas (Benedet et al., 2020; Wang et al., 2020), e outras vezes, ainda, as métricas de avaliação dos modelos não são apresentadas, apenas o tratamento escolhido (Munnaf et al., 2021a; Munnaf et al., 2021b). Por óbvio, o objetivo de remover ruídos, enfatizar feições e extrair as melhores informações para os modelos de ML é factível. Entretanto, a adição de etapas no processamento dos dados deve ser fundamentada, pois representa inevitavelmente maior custo de processamento, exigindo ou maior tempo para predição ou o uso de máquinas mais sofisticadas, encarecendo a aplicação da técnica e dificultando a sua difusão. Assim, aferiu-se o tempo de processamento da máquina para realizar dez vezes a predição de cada atributo a partir dos modelos calibrados com os três bancos de dados. O tempo de cada repetição foi calculado pela média da predição dos seis atributos avaliados (Tabela 5). O tempo demandado para predição a partir dos espectros com pré-processamento foi, em média, 200% maior do que a partir do banco de dados calibrado com os dados espectrais brutos.

TABELA 5. Custo de processamento médio aferido pelo tempo demandado pela máquina para realizar dez predições dos seis atributos avaliados neste estudo. **Average processing cost measured by the time demanded by the machine to predict ten times the six attributes evaluated in this study.**

	DB	PP1	PP2
		segundos	
1	0,1346	0,2692	0,2692
2	0,0470	0,0997	0,1012
3	0,1037	0,2074	0,2074
4	0,1166	0,2356	0,2333
5	0,1306	0,2586	0,2639
6	0,1126	0,2422	0,2422
7	0,1346	0,2558	0,2558
8	0,1152	0,2315	0,2315
9	0,0985	0,1757	0,2010
10	0,1198	0,2397	0,2397
Média	0,1113	0,2216	0,2245
%	-	199%	202%

DB: dados brutos; PP1: sequência de pré-processamento 1; PP2: sequência de pré-processamento 2; %: porcentagem comparado ao modelo de menor tempo de processamento – DB.

Dessa forma, sugere-se que a adição das etapas demandadas pelo pré-processamento acarretou no dobro do custo de processamento, aqui aferido pelo tempo demandado para predição, sem ganho de acurácia preditiva nos modelos de ML. Portanto, não há sentido prático em indicar o uso de técnicas de pré-processamento do espectro para a predição dos atributos de solo a partir da DRS NIR em tempo real no cenário proposto neste estudo.

CONCLUSÃO: Os modelos calibrados apresentaram bons parâmetros de acurácia tanto para os atributos físicos – argila e areia - como para os atributos químicos do solo – MO, CTC, pH, K - na área de estudo. Os atributos secundários puderam ser preditos provavelmente pela correlação que apresentaram com atributos de resposta primária no NIR. Nenhum grupo de valores preditos, para nenhum dos seis atributos avaliados, diferiu estatisticamente dos demais, denotando que as pequenas variações observadas em cada predição podem ser atribuídas ao acaso. O uso de técnicas de pré-processamento não atingiu o objetivo esperado de auxiliar significativamente a predição de atributos do solo pelo tratamento dos espectros coletados em campo. Contudo, a aplicação dessas técnicas aumentou o tempo demandado pela máquina para predição dos espectros de campo da área de estudo. Sugere-se que o uso dos dados brutos foi o mais eficiente para o cenário avaliado neste trabalho.

REFERÊNCIAS

Ben-Dor, E., & Banin, A. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. **Soil Science Society of America Journal**, 59(2), 364-372. 1995.

Benedet, L., Faria, W. M., Silva, S. H. G., Mancini, M., Demattê, J. A. M., Guilherme, L. R. G., & Curi, N. Soil texture prediction using portable X-ray fluorescence spectrometry and visible near-infrared diffuse reflectance spectroscopy. **Geoderma**, *376*, 114553. 2020.

Cherubin, M. R., Santi, A. L., Eitelwein, M. T., Menegol, D. R., Ros, C. O. D., Pias, O. H. D. C., & Berghetti, J. Eficiência de malhas amostrais utilizadas na caracterização da variabilidade espacial de fósforo e potássio. **Ciência Rural**, 44, 425-432. 2014.

Cherubin, M. R., Santi, A. L., Eitelwein, M. T., Amado, T. J. C., Simon, D. H., & Damian, J. M. Dimensão da malha amostral para caracterização da variabilidade espacial de fósforo e potássio em Latossolo Vermelho. **Pesquisa Agropecuária Brasileira**, 50(2), 168-177. 2015.

Dotto, A. C., Dalmolin, R. S. D., ten Caten, A., & Grunwald, S. A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. **Geoderma**, 314, 262-274. 2018.

Franceschini, M. H. D., Demattê, J. A. M., Kooistra, L., Bartholomeus, H., Rizzo, R., Fongaro, C. T., & Molin, J. P. Effects of external factors on soil reflectance measured on-the-go and assessment of potential spectral correction through orthogonalisation and standardisation procedures. **Soil and Tillage Research**, 177, 19-36. 2018.

Guerrero, A., De Neve, S., & Mouazen, A. M. Data fusion approach for map-based variable-rate nitrogen fertilization in barley and wheat. **Soil and Tillage Research**, *205*, 104789. 2021.

Kuang, B., Tekin, Y., Mouazen, A.M., 2015. Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. **Soil Tillage Research**. 146, 243–252. 2015.

LOPES, A.S. Manual de fertilidade do solo. São Paulo: ANDA/POTAFOS. 153 p. 1989.

LOPES, A. S.; SILVA, M. C.; GUILHERME, L. R. Acidez do solo e calagem. 3a ed. Ver. ANDA. 22 p. **Boletim Técnico**. 1990.

Montanari, R., Souza, G. S. A., Pereira, G. T., Marques, J. U. N. I. O. R., Siqueira, D. S., & Siqueira, G. M. The use of scaled semivariograms to plan soil sampling in sugarcane fields. **Precision Agriculture**, 13(5), 542-552. 2012.

Munnaf, A. M., Nawar, S., & Mouazen, A. M. Estimation of Secondary Soil Properties by Fusion of Laboratory and On-Line Measured Vis–NIR Spectra. **Remote Sensing,** 11(23), 2819. 2019.

Munnaf, M. A., Guerrero, A., Nawar, S., Haesaert, G., Van Meirvenne, M., & Mouazen, A. M. A combined data mining approach for on-line prediction of key soil quality indicators by Vis-NIR spectroscopy. **Soil and Tillage Research**, 205, 104808. 2021a.

Munnaf, M. A., Haesaert, G., Van Meirvenne, M., & Mouazen, A. M. Multi-sensors data fusion approach for site-specific seeding of consumption and seed potato production. **Precision Agriculture**, *22*(6), 1890-1917. 2021.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., ... & Wetterlind, J. Soil spectroscopy: An alternative to wet chemistry for soil monitoring. In **Advances in agronomy** (Vol. 132, pp. 139-159). Academic Press. 2015.

Pasquini, C. (2018). Near infrared spectroscopy: A mature analytical technique with new perspectives–A review. **Analytica chimica acta**, 1026, 8-36.

Rinnan, Å., Van Den Berg, F., & Engelsen, S. B. Review of the most common pre-processing techniques for near-infrared spectra. **TrAC Trends in Analytical Chemistry**, 28(10), 1201-1222. 2009.

Stenberg B., Viscarra Rossel R.A., Mouazen A.M., & Wetterlind J. Visible and near-infrared spectroscopy in soil science. **Advances in Agronomy**, 107: 163-215. 2010.

Velliangiri, S., & Alagumuthukrishnan, S. (2019). A review of dimensionality reduction techniques for efficient computation. **Procedia Computer Science**, 165, 104-111.

Wang, Y. P., Lee, C. K., Dai, Y. H., & Shen, Y. Effect of wetting on the determination of soil organic matter content using visible and near-infrared spectrometer. **Geoderma**, *376*, 114528. 2020.

Wollenhaupt, N. C., Wolkowski, R. P., & Clayton, M. K. Mapping soil test phosphorus and potassium for variable?rate fertilizer application. **Journal of production agriculture**, 7(4), 441-448. 1994.

Zhang, J., Guerrero, A., & Mouazen, A. M. Map-based variable-rate manure application in wheat using a data fusion approach. **Soil and Tillage Research**, 207, 104846. 2021.