Análise de espectroscopia utilizando aprendizado de máquina para quantificação de fósforo e de potássio

Spectroscopy analysis using machine learning for phosphorus and potassium quantification

Vinicius Henrique Kieling[1], Giovanni Alfredo Guarneri[2],
Larissa Macedo dos Santos Tonial[3], Jefferson Tales Oliva[4]

RESUMO

Este trabalho tem o objetivo de quantificar, por meio de aprendizado de máquina, a incidência de fósforo e de potássio a partir de espectros de amostras de solo. Para isso, os dados analisados foram coletados por meio de um equipamento de infravermelho próximo (NIR). A metodologia consiste em tratar os dados do espectro utilizando os algoritmos de Savtisky-Golay, *Quasi Monte Carlo Discrepancy* (QMCD); fazer a separação dos conjuntos de treino e teste utilizando o algoritmo de Kennard-Stone; utilizar o algoritmo de Regressão *partial least squares* (PLS) para a construção do modelo. Como resultado, os modelos atingiram valores de R^2 entre 0,69 e 0,75 durante a calibração dos modelos. Durante a validação, valores de R^2 entre 0,73 e 0,74 foram obtidos. Dessa forma, neste trabalho foram atingidos resultados satisfatórios para cada um dos espectros. Conclui-se que uma abrangência maior de amostras é desejada para construir um modelo mais preciso, suprindo a maior variabilidade de amostras possível, e que os teores dos espectros indiretos são mais difíceis de serem estimados.
PALAVRAS-CHAVE: Aprendizado de Máquina; Espectroscopia; Espectroquímica; NIR; PLS.

ABSTRACT

This work aims to quantify the incidence of phosphorus and potassium from soil sample spectra using machine learning. To achieve this, the analyzed data were collected using near-infrared (NIR) equipment. The methodology involves processing the spectrum data using the Savitzky-Golay and Quasi Monte Carlo Discrepancy (QMCD) algorithms, separating the training and test sets using the Kennard-Stone algorithm, and using the Partial Least Squares (PLS) regression algorithm to build the model. As a result, the models achieved R2 values between 0.69 and 0.75 during model calibration. During validation, R2 values between 0.73 and 0.74 were obtained. Thus, this work achieved satisfactory results for each of the spectra. It is concluded that a broader range of samples is desired to build a more accurate model, covering as much sample variability as possible, and that the levels of indirect spectra are more challenging to estimate.
KEYWORDS: Machine Learning; Spectroscopy; Spectrochemistry; NIR; PLS.

INTRODUÇÃO

A espectroscopia de infravermelho próximo (NIR, do inglês *near infrared spectroscopy*) é comumente utilizada na análise de alimentos, medicamentos, plantas, compostos industriais, solo, entre outros. Na NIR ocorrem absorções devido a mudanças de estados rotacionais e vibracionais das moléculas. Na região do NIR de 2500 a 800 nm

[1] Bolsista da Fundação Araucária. Universidade Tecnológica Federal do Paraná, Pato Branco, Paraná, Brasil. E-mail: viniciush.kieling@gmail.com. ID Lattes: 7299665248676427.
[2] Docente no Departamento Acadêmico de Química. Universidade Tecnológica Federal do Paraná, Pato Branco, PR, Brasil.E-mail: larissasantos@utfpr.edu.br. ID Lattes: 9439814411927273.
[3] Docente no Programa de Pós-Graduação em Engenharia Elétrica. Universidade Tecnológica Federal do Paraná, Pato Branco, PR, Brasil. E-mail: giovanni@utfpr.edu.br. ID Lattes: 7436484622054922.
[4] Docente no Departamento Acadêmico de Informática. Universidade Tecnológica Federal do Paraná, Pato Branco, PR, Brasil. E-mail: jeffersonoliva@utfpr.edu.br. ID Lattes: 5086431818930800.

(4000 a 12500 cm-1) ocorre quase exclusivamente a absorção de vibrações relacionadas a sobretons e combinações. As bandas vibracionais de absorção de compostos orgânicos geradas por sobretons ocorrem na região de 800 a 2000 nm, e as bandas de combinação ocorrem na região de 1800 a 2500 nm, envolvendo estiramentos e deformações angulares, referentes às ligações C=O (1900-2000 nm), C-H (1100-1225 nm, 1300-1420 nm, 1620-1800 nm, 2200-2460 nm), C-O, N-H (1400-1600 nm, 2000-2200 nm), e O-H (1400-1600 nm, 1900-2000 nm, 2000-2200 nm) (WORKMAN e WEYER, 2008).

A problemática de espectros NIR remete à complexidade decorrente da sobreposição de bandas, motivo pelo qual esta não tem sido utilizada com propósito de interpretação de espectros, precisando do auxílio de ferramentas estatísticas. Deste modo, para a construção de modelos são necessários além dos espectros, dos teores determinados em laboratório por meio de metodologias de bancada (BEDIN et al., 2021).

O objetivo deste trabalho é construir modelos regressivos para quantificação de fósforo e potássio a partir de espectros obtidos em amostras do solo por meio do NIR.

MATERIAIS E MÉTODOS

AMOSTRAS E EQUIPAMENTO NIR

Para a obtenção dos espectros de NIR empregando o espectrômetro modelo MPAFT-NIR da marca Bruker (Bruker Optics Inc., Ettlingen, Alemanha), utilizou-se 20cm³ de amostra. A obtenção dos espectros foi realizada empregando reflectância difusa com a esfera integradora e a rotação da amostra. Para isso foi utilizado o software de aquisição e de processamento espectral OPUS 7.2[5]. Os dados foram registrados de 12.000 a 4.000 cm^{-1} com uma resolução de 16 cm^{-1}, e 64 varreduras por espectro. Dois espectros foram coletados para cada uma das 94 amostras.

MÉTODOS DE PRÉ PROCESSAMENTO

Nessa análise é utilizado o método de filtragem digital de Savitzky-Golay (SAVITZKY e GOLAY, 1964), que é utilizado para a suavização de dados. Esse método é uma generalização da média móvel que utiliza um polinômio de baixo grau e coeficientes gerados por mínimos quadrados. Esse filtro é aplicado nas seguintes etapas: É definido a origem do sinal, depois, são definidos a largura do intervalo e um ponto central, seguido da remoção desse ponto central do conjunto de pontos do intervalo. O polinômio é então ajustado e utilizado para estimar o valor do ponto central que fora removido previamente e então o intervalo é deslocado para o ponto seguinte, repetindo as etapas anteriores. O diferencial desse método é que utiliza um polinômio ao invés da média móvel para determinar pontos do intervalo, apresentando resultados superiores a outros filtros digitais (CERQUEIRA, 2000). Esse método foi implementado utilizando a linguagem Python sob o auxílio da biblioteca Scipy.

[5] https://www.bruker.com/en/products-and-solutions/infrared-and-raman/opus-spectroscopy-software.html

Depois de aplicar o filtro e reduzir os ruídos, é aplicado um método de seleção de intervalo de interesse (*Region of Interest* - ROI) na banda. Segundo Pellicia (2018), nem sempre é claro em qual região do espectro as informações mais importantes se encontram, muito menos quando se tenta automatizar esse processo. A priori, é considerado o espectro inteiro como estimativa inicial e depois são analisados quais tamanhos de onda contêm mais informações relevantes. As etapas do método são: Aplicar um PLS no espectro todo e avaliar sua qualidade; obter os coeficientes do melhor modelo encontrado onde cada coeficiente equivale a um pequeno intervalo de banda; descartar os intervalos com menores coeficientes, resultando em um espectro otimizado.

Em sequência, é necessário retirar as anomalias do conjunto de dados que desviam atipicamente do restante devido às falhas do equipamento ou falhas humanas no laboratório. Para isso, algoritmos de detecção de anomalias, também conhecidos como algoritmos de detecção de *outliers*, visam identificar essas instâncias. Inúmeros métodos de detecção foram desenvolvidos nas últimas décadas (HAN et al., 2022), nesse projeto é utilizado o método de remoção de *outliers* QMCD (*Quasi Monte Carlo Discrepancy*) por meio da biblioteca Python PYOD (*Python Outlier Detection*).

O QMCD é uma versão determinística da Simulação Monte Carlo tradicional. A QMCD não utiliza aleatoriedade e as sequências de números são mais uniformes e ordenadas, chamadas de sequências de baixa discrepância. É frequentemente mais rápida, especialmente em dimensões mais elevadas (LIN et al., 2023).

CONSTRUÇÃO DO MODELO

Depois do pré-processamento, foi aplicado o método Kennard-Stone para divisão das amostras entre conjuntos de treinamento e de teste. Segundo Daszykowski et al. (2002), o algoritmo Kennard-Stone seleciona um subconjunto de objetos *n*-dimensionais distribuídos uniformemente no espaço experimental, que é importante quando se desenvolve modelos de classificação ou de calibração. O conjunto selecionado deve abranger todos os tipos de dados e representar todas as variações possíveis das amostras. Para este projeto, escolhemos 30% das amostras pós-processamento para serem usadas como conjunto de teste/validação, deixando 70% das amostras como conjunto de treinamento/calibração. Assim, os conjuntos de treinamento foram utilizados para construir modelos regressores. Em seguida, esses modelos foram avaliados utilizando os conjuntos de teste. Para o código, foi escolhido a biblioteca *Astartes* ao invés da biblioteca *Scikit-Learn* para implementar o método Kennard-Stone por apresentar um desempenho 430 vezes mais rápido (PELLICIA, 2019).

Para a construção dos modelos de regressão é utilizado o método de *Partial Least Squares* (PLS). Como apresentado por Abdi (2003), PLS foi uma técnica desenvolvida para combinar características de *Principal Component Analysis* (PCA) e Regressão Múltipla para predizer um conjunto de variáveis dependentes de um outro conjunto maior de variáveis independentes. Em uma notação mais formal, dado **I** observações descritas

por **K** variáveis dependentes são armazenadas em uma matriz $\mathbf{Y}_{I \times K}$. Os valores de **J** preditores são coletados a partir de **I** e armazenadas em uma matriz $\mathbf{X}_{I \times J}$. Dizemos então que o objetivo é prever **Y** de **X** e descrever sua estrutura

AVALIAÇÃO DOS MODELOS

Para a avaliação dos regressores, foram utilizadas as métricas R^2 e erro médio quadrático (do inglês *root mean square error* - RMSE). A R^2 é o quadrado do coeficiente de correlação entre o conjunto do espectro (X) e o conjunto de teores conhecidos (Y). Nós comumente nos referimos a R^2 como abrangência de variabilidade nos dados explicados pelo modelo regressor (MONTGOMERY; RUNGER, 2011). O RMSE indica o erro do modelo ou a variabilidade de calibração. Quando validado externamente, pode resultar em maior valor para modelo de calibração. Quanto menor o valor, melhor é considerado o resultado (LOUW; THERON, 2010).

RESULTADOS

Ao final do processo, devemos avaliar nossos conjuntos de treino e teste. Os resultados das métricas de calibração são apresentados na Tabela 1.

Tabela 1 – Métricas da calibração

	R^2	RMSE
Fósforo (ppm)	0,69	203,11
Potássio (ppm)	0,75	2091,05

Fonte: Autoria Própria (2023).

Considerando a Tabela 1, tanto o Fósforo quanto o Potássio apresentam um R^2 que se encaixa na faixa de 66% a 81%, uma aproximação razoável da variabilidade. O RMSE do Potássio e Fósforo apresenta um valor alto, característico dos espectros NIR indiretos, semelhante aos resultados de Bedin (2021).

Os resultados das métricas de validação são apresentados na Tabela 2.

Tabela 2 – Métricas da validação

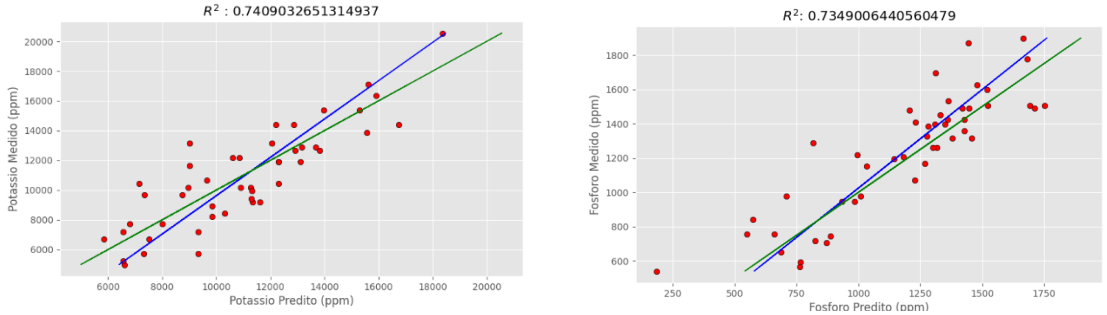
	R^2	RMSE
Fósforo (ppm)	0,73	180,21
Potássio (ppm)	0,74	1697,63

Fonte: Autoria Própria (2023).

Considerando a Tabela 2, obtemos resultados semelhantes a Tabela 1. Essa semelhança é desejada pois preserva a variabilidade dos dois conjuntos, conforme explicado na seção sobre a construção do modelo e separação dos conjuntos.

O processo de validação é caracterizado justamente por validar o conjunto de treinamento utilizando o conjunto de teste. Nessa validação, a variabilidade do modelo pode ser graficada conforme a Figura 1, onde a linha azul representa o valor médio dos teores medidos em bancada, enquanto a linha verde representa o valor médio dos teores estimados pelo modelo.

Figura 1 – PLS do Potássio e do Fósforo



Fonte: Autoria Própria

CONCLUSÃO

Por meio da análise de bancada das amostras de solo e após serem aplicados os métodos de pré-processamento apresentados neste resumo, foi possível treinar um modelo de regressão PLS que apresenta valores de R^2 considerados aceitáveis, conforme a Tabela 1 e Tabela 2. Isto significa que o modelo de regressão treinado pode predizer com grande eficácia e em poucos minutos os teores de Potássio e Fósforo. A mesma análise através das metodologias em bancada pode demorar diversos dias.

É de relevante importância considerar que a variabilidade e quantidade das amostras pode afetar significativamente o modelo treinado. Neste projeto, foram quantificados Potássio e Fósforo em amostras de solo, visto que as mesmas reagem fortemente à radiação emitidas em Carbono e Nitrogênio. Diz-se que todos os outros teores químicos são medidos indiretamente.

O código foi escrito visando a automatização dos parâmetros, tentando ao máximo criar uma análise não supervisionada que processa, avalia e escolhe o melhor modelo. Isso dará início a trabalhos futuros, analisando outros componentes químicos como o Carbono e Nitrogênio, visando aumentar o número de amostras e investigar outros métodos regressores como o *support vector machine* (SVM).

Agradecimentos

À Fundação Araucária pela fomentação de uma bolsa de iniciação científica.

Disponibilidade de código

O código fonte da avaliação experimental estará disponível publicamente após a tramitação do pedido de registro de software.

Conflito de interesse

Não há conflito de interesse.

REFERÊNCIAS

ABDI, H. Partial least square regression (PLS regression). **Encyclopedia for research methods for the social sciences,** 6.4: 792-795, 2003.

BEDIN, F. C. B., et al. NIR associated to PLS and SVM for fast and non-destructive determination of C, N, P, and K contents in poultry litter. **Spectrochimica acta part A: Molecular and biomolecular spectroscopy**, 245: 118834, 2021.

CERQUEIRA, Eduardo O., et al. Utilização de filtro de transformada de Fourier para a minimização de ruídos em sinais analíticos. **Química Nova**, 23: 690-698, 2000.

DASZYKOWSKI, M.; WALCZAK, B.; MASSART, D. L. Representative subset selection. **Analytica chimica acta**, 468.1: 91-103, 2002.

HAN, S., et al. Adbench: Anomaly detection benchmark. **Advances in Neural Information Processing Systems**, 35: 32142-32159, 2022.

LIN, G.; DENG, S.; WANG, X. An efficient quasi-Monte Carlo method with forced fixed detection for photon scatter simulation in CT. **Plos one**, 18.8: e0290266, 2023.

LOUW, E. D.; THERON, K. I. Robust prediction models for quality parameters in Japanese plums (Prunus salicina L.) using NIR spectroscopy. **Postharvest Biology and Technology,** 58.3: 176-184, 2010.

MONTGOMERY, D. C.; RUNGER, G. C. Applied statistics and probability for engineers. **John Wiley & Sons,** 2010.

PELLICCIA, D. **A variable selection method for PLS in Python.** 2018. Disponível em: https://nirpyresearch.com/variable-selection-method-pls-python/. Acesso em: 18 set. 2023.

PELLICCIA, D. **The Kennard-Stone algorithm.** 2022. Disponível em: https://nirpyresearch.com/kennard-stone-algorithm/. Acesso em: 18 set. 2023.

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. **Analytical Chemistry**, 36.8: 1627-1639, 1964.

WORKMAN, J.; WEYER, L. Practical Guide to Interpretive Near-Infrared Spectroscopy. **Angew. Chem. Int.** , v. 47, p. 4628-4629, 2008.