XLVI Congresso Brasileiro de Engenharia Agrícola - CONBEA 2017
Hotel Ritz Lagoa da Anta - Maceió - AL
30 de julho a 03 de agosto de 2017

CAPACIDADE DA ESPECTROSCOPIA VIS-NIR EM DIFERENCIAR NÍVEIS DE FÓSFORO NO SOLO

MICAEL FELIPE DE SOUZA[1], MARCOS A. N. COUTINHO[2], LUCAS R. DO AMARAL[3], HENRIQUE C. J. FRANCO[4]

[1] Eng. Agrícola, Acadêmico de mestrado, Faculdade de Engenharia Agrícola (FEAGRI), Universidade Estadual de Campinas (UNICAMP), Campinas – SP, Fone: (19) 998024302, micaelfelipe@gmail.com.br.
[2] Eng. Agrícola, Acadêmico de mestrado, FEAGRI/UNICAMP, Campinas – SP.
[3] Eng. Agrônomo, Professor Doutor, FEAGRI/UNICAMP, Campinas – SP
[4] Eng. Agrônomo, PhD, Coordenador Divisão Agrícola, CTBE/CNPEM, Campinas – SP.

Apresentado no
XLVI Congresso Brasileiro de Engenharia Agrícola - CONBEA 2017
30 de julho a 03 de agosto de 2017 - Maceió - AL, Brasil

RESUMO: A espectroscopia de refletância difusa na região do Visível e Infravermelho próximo (Vis-NIR) é uma ferramenta promissora na análise de parâmetros químicos do solo, devido à praticidade e não utilização de reagentes químicos. Assim, o objetivo do trabalho foi avaliar se é possível classificar os níveis de fósforo disponível no solo (P) por meio da espectroscopia. Foi utilizado o espectrorradiômetro FieldsSpec 4 para coleta dos espectros em amostras de solo submetidas a doses diferenciadas de P. Com base na análise química das amostras de solo, dividiu-se os teores de P em três classes e testou-se diferentes modelos de classificação obtidos com os dados espectrais. A eficácia dos modelos foi avaliada de acordo com a acurácia, estatística Kappa e precisão de cada classe. A qualidade geral dos modelos de classificação gerados foi razoável, mas a precisão em cada classe foi baixa. A ausência de resposta espectral direta do P justifica esses resultados. Dessa forma, a espectroscopia no Vis-NIR apresenta limitações nas análises de disponibilidade desse elemento no solo, pois sua predição é dependente de correlação com outras propriedades do solo.
PALAVRAS-CHAVE: Modelos de classificação; fertilidade do solo, mineração de dados.

VIS-NIR SPECTROSCOPY CAPABILITY TO DIFFERENTIATE SOIL PHOSPHORUS LEVELS

ABSTRACT: Visible and Near Infrared spectroscopy (Vis-NIR) is a promising approach to predict soil chemical properties due to its practicality and non-use of chemical reagents. This study aimed to evaluate whether is possible to classify levels of soil phosphorus availability (P) by spectroscopy. FieldSpec 4 Spectroradiometer was used to collect spectral measurements of soil samples that were submitted to P rates. Based on soil chemical P analysis, samples were divided in 3 availability levels. Several classification modelling procedures were tested to analyze spectral data. Effectiveness of the models was evaluated according to overall accuracy, Kappa statistics and accuracy of each class. The overall effectiveness of the classification models was reasonable, but the accuracy in each class was low. These results can be explained by absence of P direct influence on the spectra. Thus, Vis-NIR Spectroscopy has limitations for identify soil P availability, because its prediction is dependent of other soil properties.
KEYWORDS: classification models, soil fertility, data mining.

INTRODUÇÃO: O fósforo (P), após ser absorvido da solução do solo pelas raízes, participa de processos importante na planta, como fotossíntese, respiração, armazenamento e transferência de energia. Dessa forma, o conhecimento dos teores disponíveis para as plantas no solo é de extrema importância. As técnicas laboratoriais convencionais utilizadas na determinação do P-disponível são demoradas, além de utilizarem reagentes químicos e apresentarem custos elevados. Nesse sentido, a espectroscopia é uma ferramenta promissora para o mapeamento das propriedades do solo (WETTERLIND et al., 2008). O uso dessa técnica apresenta vantagens como praticidade, não utilização de reagentes químicos, baixo custo e a possibilidade de em uma única amostragem ser possível inferir sobre diversas propriedades do solo (VISCARRA ROSSEL et al., 2006). De posse dos dados espectrais, são necessárias ferramentas para tratá-los e extrair informações relevantes, favorecendo a produção de modelos de predição das variáveis de interesse (MENDES, 2014). Nesse contexto, a aplicação de técnicas de mineração de dados e análises multivariadas são alternativas promissoras, pela possibilidade de análise de grandes volumes de dados. Dentre as tarefas de mineração de dados é possível destacar a classificação, que consiste na busca por uma função que permita associar corretamente cada instância do banco de dados a uma classe. Assim, o objetivo desse trabalho foi avaliar se é possível criar modelos de classificação para níveis de fósforo no solo (P) por meio da espectroscopia e técnicas de mineração de dados.

MATERIAL E MÉTODOS: Foram analisadas 100 amostras de solo com textura argilosa provenientes de um experimento conduzidos em casa de vegetação em vasos com 1 dm^3. O solo para preenchimento dos vasos foi coletado em área de vegetação nativa na região de Campinas/SP (-22,804032-S, -47,050278-O) a 0,20 m de profundidade. Após o preenchimento dos vasos com solo, foram aplicadas doses de P na forma de Super Fosfato Triplo. As doses foram 0, 15, 40, 60 e 120 $mg.dm^{-3}$ de P_2O_5. Os vasos foram mantidos por 35 dias para que ocorresse a reação dos fertilizantes com o solo, sendo realizado revolvimento do solo 15 dias após a montagem do experimento. Em seguida, essas amostras foram encaminhadas ao laboratório para análise química do fósforo, pelo método da resina de troca iônica, com a determinação do fósforo disponível (método de referência). A partir desses resultados foi realizada a classificação das amostras de solo de acordo com os teores de P-resina em: baixo (0 – 12 $mg.dm^{-3}$), médio (13 – 30 $mg.dm^{-3}$) e alto (>31 $mg.dm^{-3}$). Para a coleta dos dados de espectroscopia as amostras foram secas em estufa por 24 horas a 45°C e peneiradas a 2,00 mm. Foram coletados 3 espectros por amostra, resultado da média de 10 leituras consecutivas (valores de absorbância) cada um. Para isso, utilizou-se o espectroradiômetro FieldSpec 4 (Analytical Spectral Devices, Boulder, Colorado, EUA). As leituras foram realizadas na faixa espectral de 450 a 2500 nm, com resolução de 1,4 nm. Em seguida, foi utilizado o programa computacional WEKA 3.6.3. (WITTEN et al., 2011) para a criação dos modelos de classificação. Foram testados alguns dos principais métodos classificadores: árvores de decisão (J48, Random Forest – RF), Máquinas de vetores suporte (Support Vector Machine – SVM), Meta (Regressão por discretização) – Boosting, Regressão logística (Simple Logistic). Para avaliar o desempenho dos modelos, utilizou-se a abordagem de divisão dos conjuntos de treinamento e validação nas proporções 2/3 e 1/3 respectivamente. Após a execução dos algoritmos, obteve-se as tabelas com os resultados de desempenho: Acurácia (%), Kappa e Precisão das Classes. Segundo Landis e Koch (1997), os modelos são classificados pelo índice Kappa como péssimo (Kappa<0); ruim (0,00<Kappa<0,20); razoável (0,21<Kappa<0,40); bom (0,41<Kappa<0,60); muito bom (0,61<Kappa<0,80); excelente (0,81<Kappa<1,00).

RESULTADOS E DISCUSSÃO: De forma geral, os modelos obtidos com os algoritmos obtiveram valores de acurácia maiores que 60% (TABELA 1). Esse parâmetro de avaliação representa a porcentagem de acerto do classificador. No entanto, os valores obtidos para a estatística Kappa, que indica a concordância entre as classes preditas e observadas e o número esperado de acerto do classificador foram razoáveis (LANDIS E KOCH, 1997) (TABELA 1). Os baixos valores de Kappa podem estar relacionados com o fato de o P não possuir resposta espectral direta na região analisada (Vis-NIR), sendo necessária sua correlação com outras propriedades do solo (TERRA et al., 2015). Como o P foi aplicado no mesmo solo, não houve correlação do teor de P com outra propriedade do solo, como frequentemente ocorre para argila ou matéria orgânica em amostragens espacializadas. Essa falta de resposta espectral nessa região ocorre porque as vibrações moleculares fundamentais de alguns elementos no solo ocorrem na região do infravermelho médio (MIR), enquanto que apenas seus harmônicos e combinações são detectados no NIR (VISCARRA ROSSEL et al., 2006). Dessa forma, alguns autores destacam que melhores modelos para previsão do P no solo podem ser obtidos na região do MIR (SORIANO-DISLA et al., 2014). Além disso, outro fator que pode ter limitado o desempenho dos modelos de classificação é o desbalanceamento nas classes, já que o conjunto de dados foi composto por 72% de amostras classificadas com teor alto, 10% médio e 18% baixo. Apesar de ser aplicado o balanceamento nas classes para a criação dos modelos, o efeito do desbalanceamento pode ter sido limitante no desempenho dos modelos, já que alguns classificadores tendem a ser oprimidos por classes maiores e ignorar classes menores (CHAWLA et al., 2004). Apesar do baixo desempenho dos modelos, o melhor modelo de classificação para os níveis de fósforo no solo (maiores valores de acurácia e Kappa) foi obtido com o algoritmo Random Forest (RF). Esse algoritmo é baseado na construção de um conjunto de árvores preditoras que são capazes de produzir separadamente uma resposta para o conjunto de dados preditores (MURSALIN et al., 2017). Essa estratégia revela-se muito boa em comparação com outros classificadores como SVM e Redes Neurais e é robusta contra overfitting (BREIMAN, 2001). As principais vantagens que tornam o RF uma escolha favorável para as previsões de propriedades do solo por espectroscopia são: ser robusto a ruídos, não precisar de ajustes finos dos parâmetros para produzir boas previsões e principalmente por ser adequado para conjunto de dados com muitas variáveis (comprimentos de onda) e poucas instâncias (amostras) o que resulta em uma matriz com número muito maior de colunas do que linhas (DÍAZ-URIARTE e DE ANDRÉS, 2006; VISCARRA ROSSEL e BEHRENS, 2010). Alguns autores já comprovaram essas vantagens, obtendo resultados satisfatórios com a utilização do RF para prever propriedades do solo como carbono orgânico (ZHANG et al., 2017) e textura (CHAGAS et al., 2016). O algoritmo Boosting apresentou resultados semelhantes ao RF, justificado pelo fato desse algoritmo ser apropriado para conjunto de dados com desequilíbrio de classes pois fornece mais oportunidades de aprendizado para amostras de classe minoritárias através do processo de aprendizagem iterativa (KIM et al., 2015).

TABELA 1. Resultados obtidos com os algoritmos de classificação

Algoritmo	**Acurácia [%]**	**Kappa**	**Precisão das classes**		
			Baixo	**Médio**	**Alto**
Random Forest	75	0.33	0.67	0.25	0.79
Boosting	75	0.31	0.70	0.25	0.78
Simple Logistic	72	0,25	0.67	0	0.71
J48	71	0.22	0.46	0.33	0.76
SVM	66	0.21	0.21	0.67	0.74

CONSIDERAÇÕES FINAIS: Foi possível criar modelos de classificação para níveis de fósforo no solo por meio da espectroscopia e técnicas de mineração de dados. No entanto, os resultados foram razoáveis. Resultados substancialmente melhores poderiam ser obtidos com conjunto de dados balanceado (números equivalentes de amostras em cada classe). A ausência de resposta espectral direta do P na região analisada (Vis-NIR) justifica esses resultados razoáveis. Dessa forma, a espectroscopia nessa região apresenta limitações nas análises de disponibilidade de P no solo, pois sua predição é dependente de correlação com outras propriedades do solo. O entendimento da relação do P com parâmetros que possuem resposta espectral direta na região do Vis-NIR bem como algoritmos mais robustos são necessários para as análises de espectroscopia nessa região.

REFERÊNCIAS

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

CHAGAS, C. et al. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. **Catena**, v. 139, p. 232-240, 2016.

CHAWLA, N.V.; JAPKOWICZ, N.; KOTCZ, A. Editorial: special issue on learning from imbalanced data sets. **ACM Sigkdd Explorations Newsletter**, v. 6, n. 1, p. 1-6, 2004.

DÍAZ-URIARTE, R.; DE ANDRES, S. A. Gene selection and classification of microarray data using random forest. **BMC bioinformatics**, v. 7, n. 1, p. 3, 2006.

KIM, M-J; KANG, D-K; KIM, H. B. Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. **Expert Systems with Applications**, v. 42, n. 3, p. 1074-1082, 2015.

LANDIS, J.R.; KOCH, G.G. The measurement of observer agreement for categorical data. **Biometrics**, v.33, n.1, p. 159-174, 1977.

MENDES, R. A. G. **Utilização da espectroscopia em reflectância no infravermelho próximo para discriminação de espécies da família Myrtaceae**. 2014. 85 f., il. Dissertação (Mestrado em Botânica)—Universidade de Brasília, Brasília, 2014.

MURSALIN, Md. et al. Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. **Neurocomputing**, v. 241, p. 204-214, 2017.

SORIANO-DISLA, J. M. et al. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. **Applied Spectroscopy Reviews**, v. 49, n. 2, p. 139–186, 2014.

TERRA, F.S.; DEMATTÊ, J.AM; ROSSEL, R.A.V. Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data. **Geoderma**, v. 255, p. 81-93, 2015.

VISCARRA ROSSEL, R.A.; BEHRENS, T. Using data mining to model and interpret soil diffuse reflectance spectra. **Geoderma**, v. 158, n. 1, p. 46-54, 2010.

VISCARRA ROSSEL, R.A., WALVOORT, D.J.J., MCBRATNEY, A.B., JANIK, L.J., SKJEMSTAD, J.O. Visible, near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. **Geoderma** 131, 59–75, 2006

WITTEN, I. E., FRANK, E., MARK, A. **Data Mining – Practical Machine Learning Tools and Techniques**. 3 ed. United States: Elsevier, 2011.

WETTERLIND J; STENBERG B; JONSSON A. Near infrared reflectance spectroscopy compared with soil clay and organic matter content for estimating within-field variation in N uptake in cereals. **Plant Soil**, v. 302, n. 317–327, 2008.

ZHANG, H. et al. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model. **Science of The Total Environment**, v. 592, p. 704-713, 2017.