Artigo

APLICAÇÃO DE TÉCNICAS MULTIVARIADAS E INTELIGÊNCIA ARTIFICIAL NA ANÁLISE DE ESPECTROS DE INFRAVERMELHO PARA DETERMINAÇÃO DE MATÉRIA ORGÂNICA EM AMOSTRAS DE SOLOS

Diego M. Souza
Centro Nacional de Pesquisa de Arroz e Feijão, EMBRAPA, GO 462, km 12, 75375-000 Santo Antônio de Goiás – GO / Instituto de Química, Universidade Federal de Goiás, Campus Samambaia, CP 131, 74001-970 Goiânia – GO, Brasil
Beata E. Madari
Centro Nacional de Pesquisa de Arroz e Feijão, EMBRAPA, GO 462, km 12, 75375-000 Santo Antônio de Goiás – GO, Brasil
Freddy F. Guimarães*
Instituto de Química, Universidade Federal de Goiás, Campus Samambaia, CP 131, 74001-970 Goiânia – GO, Brasil

Recebido em 1/12/11; aceito em 6/6/12; publicado na web em 24/8/12

APPLICATION OF MULTIVARIATE CALIBRATION AND ARTIFICIAL INTELLIGENCE IN THE ANALYSIS OF INFRARED SPECTRA TO QUANTIFY ORGANIC MATTER IN SOIL SAMPLES. In this paper studies based on Multilayer Perception Artificial Neural Network and Least Square Support Vector Machine (LS-SVM) techniques are applied to determine of the concentration of Soil Organic Matter (SOM). Performances of the techniques are compared. SOM concentrations and spectral data from Mid-Infrared are used as input parameters for both techniques. Multivariate regressions were performed for a set of 1117 spectra of soil samples, with concentrations ranging from 2 to 400 g kg^{-1}. The LS-SVM resulted in a Root Mean Square Error of Prediction of 3.26 g kg^{-1} that is comparable to the deviation of the Walkley-Black method (2.80 g kg^{-1}).

Keywords: artificial neural network; LS-SVM; soil organic matter.

INTRODUÇÃO

A matéria orgânica do solo (MOS) inclui todas as substâncias orgânicas e é composta por uma mescla de resíduos animais e vegetais, em diversos estágios de decomposição. Sua importância para diversos processos físicos, químicos e biológicos é amplamente reconhecida na literatura.[1] A MOS desempenha diversas funções no solo, estando ligada a processos como a ciclagem e retenção de água e nutrientes, agregação do solo e dinâmica da água, além de ser fonte básica de energia para a atividade biológica.[2] Em solos brasileiros a capacidade de troca de cátions está relacionada à matéria orgânica, pois esta pode representar até 80% das cargas negativas presentes.[3] Assim, a determinação de MOS é um componente importante na avaliação da fertilidade de solo e recomendação de adubação, com influência na qualidade em todo processo agrícola.[4]

O método de Walkley-Black é o mais utilizado para determinação de MOS, cujo princípio é a oxidação do carbono orgânico a CO_2 por ação de íons dicromato em meio ácido. Nesta reação de oxirredução há a formação de íons Cr(III), que podem ser determinados indiretamente pela titulação dos íons dicromato em excesso por íons Fe(II), ou por espectrofotometria, graças à coloração esverdeada característica daquele íon.[4] Como o aumento da temperatura se dá somente pela diluição do H_2SO_4 em água, a oxidação ocorre parcialmente, sendo necessária a utilização de um fator de correção de 1,33. Como se determina carbono orgânico, a conversão para MOS é feita pelo fator de van Bemmelen (1,724), com base no pressuposto de que 58% da matéria orgânica é carbono.[5]

A determinação de MOS pelo método de Walkley-Black ou espectrofotométrico tem as desvantagens de utilizar ácidos concentrados e gerar resíduo com cromo. Ele ocorre naturalmente no ambiente no estado trivalente (Cr^{3+}) e é considerado essencial aos seres vivos. Contudo, quando assume a forma hexavalente (Cr^{6+}) é considerado tóxico aos seres humanos, podendo causar ulcerações, irritação, inflamação e, ainda, está associado a risco de câncer.[6] No Laboratório de Análise Agroambiental (LAA) do Centro Nacional de Pesquisa de Arroz e Feijão da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), em Goiás, por exemplo, foram realizadas 5000 determinações de MOS em 2010. Isso corresponde ao consumo de 2,5 kg de dicromato de potássio ($K_2Cr_2O_7$, grau analítico para análise - P.A.), 100 L de ácido sulfúrico (H_2SO_4) concentrado e geração de aproximadamente 1400 L de resíduo sulfocrômico. Tais desvantagens justificam esforços no desenvolvimento de metodologias limpas, como análise por espectroscopia no infravermelho médio (MIR - *Mid Infrared*). Os resultados apresentados neste trabalho demonstram que a aplicação de MIR na determinação de MOS apresenta baixo custo das análises e otimização da operacionalidade.

Materiais que contêm substâncias orgânicas são incorporados superficialmente no solo. Tais materiais são oriundos principalmente de restos vegetais. Essas substâncias são transformadas no solo por dois processos principais: humificação e mineralização. O material orgânico não transformado corresponde a substâncias de baixo peso molecular como carboidratos, aminoácidos, resinas, ligninas, alcoóis, aldeídos e ácidos alifáticos e aromáticos, que são originários de restos animais e vegetais. Dentre as substâncias de baixo peso molecular existem substâncias oriundas de metabolismo microbiano e das raízes. Estes dois grupos correspondem de 10 a 15% da carga orgânica do solo. O restante, de 85 a 90%, são substâncias húmicas que têm peso molecular variável e são produtos da transformação dos outros dois grupos por processos biológicos, físicos e químicos. As substâncias húmicas se caracterizam por coloração escura e podem apresentar peso molecular elevado por rotas de condensação ou apresentar estrutura complexa pela associação de várias substâncias de menor peso molecular.[7,8] Grupos químicos amino, carboxílico, hidroxílico, carbonila, cetona, éter, éste, e ligações C-C e C-H presentes nessas substâncias orgânicas são responsáveis por grande parte das absorções de amostras de solo no MIR.[9]

A espectroscopia no infravermelho próximo (NIR – *Near Infrared*) e MIR vem sendo utilizada constantemente na análise de

*e-mail: freddy@quimica.ufg.br

produtos agrícolas. Trabalhos recentes demonstram correlação satisfatória, $r^2 > 0,91$, entre métodos de análise de referência e análises utilizando NIR ou MIR em conjunto com calibrações multivariadas. A determinação de carbono em solo,[10] nitrogênio em tecido vegetal,[11] adulterantes em leite,[12] açúcar[13] e taninos[14] em café são exemplos destas aplicações. As principais vantagens destas técnicas são: ser limpa, por não necessitar da abertura das amostras por ação de reagentes tóxicos e/ou a altas temperaturas; ágil e capaz de atender à demanda crescente das diferentes determinações.[4,5,10,13]

As técnicas analíticas de espectroscopia MIR e NIR não fornecem diretamente o teor de um componente particular, sendo necessário um conjunto de amostras com teor determinado por uma metodologia de referência. Com esses valores, constrói-se um modelo de calibração multivariada, no qual se correlacionam matematicamente os espectros com os respectivos teores.[13]

Os métodos para calibração multivariada mais utilizados são regressão linear múltipla (MLR - *Multiple Linear Regression*); regressão em componentes principais (PCR –*Principal Components Regression*) e mínimos quadrados parciais (PLS – *Partial Least Square*). Além destas técnicas, ferramentas da inteligência artificial, como redes neurais artificiais (ANN – *Artificial Neural Network*) e algoritmo genético (GA – *Genetic Algorithm*), também estão sendo aplicadas em quimiometria para calibração multivariada e seleção de variáveis, respectivamente.[15,16] Uma ferramenta que tem se mostrado promissora para problemas de calibração multivariada e classificação é a máquina de vetor de suporte com mínimos quadrados (LS-SVM – *Least Square Support Vector Machine*).[12]

Algumas ferramentas auxiliares são necessárias para a obtenção de bons modelos multivariados, seja para retirar efeitos de espalhamento de luz, fenômeno comum em espectroscopia por reflectância difusa; reduzir a dimensionalidade dos dados por compressão ou seleção de variáveis, visando melhorar a generalização dos modelos.

As principais técnicas utilizadas no pré-tratamento dos dados espectrocópicos são 1ª derivada, 2ª derivada, alisamento Savitsky-Golay, transformação padrão normal de variação (SNV – *Standard Normal Variate*) e correção do espalhamento multiplicativo (MSC – *Multiplicative Scatter Correction*). As derivadas são utilizadas visando eliminar os desvios lineares de linha de base e problemas de sobreposição, mas trazem o inconveniente de diminuir a relação sinal/ruído. Esta desvantagem é compensada pelo uso simultâneo do alisamento Savitsky-Golay, que é aplicado para eliminação de ruídos espectrais. SNV e MSC amenizam problemas de dispersão de luz, muito comuns em varreduras espectrais por refletância difusa, onde a radiação é incidida diretamente na amostra em pó.[17,18]

Quando há muitas variáveis independentes para construção das calibrações, pode existir multicolinearidade, ou seja, variáveis que tenham alta correlação, também denominada de informação redundante. A inclusão de variáveis multicolineares poderá diminuir a capacidade de generalização dos modelos e, no caso de uma ANN, impossibilitar a convergência do erro durante seu treinamento.[11,19] Técnicas de pré-seleção de variáveis reduzem a dimensionalidade das variáveis independentes, produzindo modelos mais simples e robustos. Para este fim, aplica-se, principalmente, PLS por intervalos (i-PLS),[20] eliminação de variáveis não informativas (UVE – *Uninformative variable elimination*) por PLS,[21] e GA.[22] Outra maneira de reduzir a dimensionalidade é por compressão dos dados. A análise de componentes principais (PCA – *Principal Component Analysis*), apesar de desenvolvida para realizar análise exploratória de dados multidimensionais por modelagem da estrutura de covariância, é a ferramenta mais utilizada para compressão dos dados. Geralmente consegue-se reter uma grande quantidade da variância total em número pequeno de componentes principais; se isso acontecer, há uma simplificação ou redução de dimensionalidade das variáveis originais.[23]

Calibração multivariada por ANN-MLP e LS-SVM

As redes neurais artificiais são algoritmos que imitam, mesmo de forma simplificada, o mecanismo de aprendizado do cérebro humano. De modo prático, a ANN é uma caixa de processamento, que é treinada a partir de dados de entrada (*input*) previamente conhecidos e é capaz de fornecer parâmetro ou parâmetros respostas (*output*) para a qual ela foi treinada. Uma rede neural do tipo multicamadas de neurônios (MLP – *Multilayer Perceptron*) é elaborada a partir de cinco componentes básicos: neurônios artificiais, pesos sinápticos, funções de transferência, arquitetura de redes neurais e treinamento. Os neurônios artificiais, Figura 1, são as unidades básicas de processamento, que simulam o comportamento de um neurônio biológico. Os sinais de entrada, *input* (x_i), são multiplicados por pesos sinápticos, (w_i), sendo o sinal total "Net" a somatória dos produtos dos sinais de entrada pelos respectivos pesos, Figura 1. O sinal de saída é obtido por uma função de transferência que atua sobre o sinal de entrada oriundo da camada anterior, f(Net), as funções utilizadas são sigmoidal,

$$\text{Saída} = f(\text{Net}) = \frac{1}{1+e^{-(net)}} \quad (1)$$

linear,

$$\text{Saída} = a(\text{Net}) + b \quad (2)$$

e *Heaviside* (esta última assumindo apenas valor "0" ou "1"),

$$\text{Saída} = f(\text{Net}) \begin{cases} 1 \text{ se Net} > \text{valor limite;} \\ 0 \text{ se net} \leq \text{valor limite.} \end{cases} \quad (3)$$

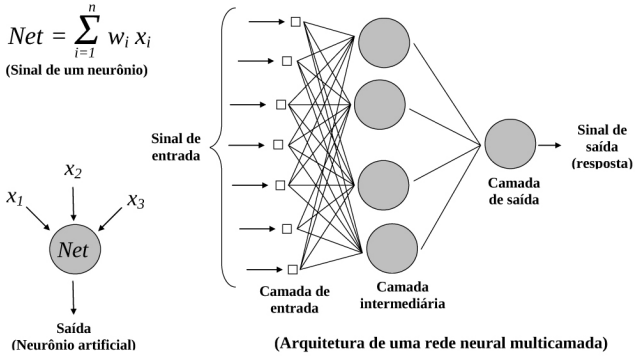

Figura 1. *Representação do neurônio artificial, sinais de entrada e saída (fórmula e esquema à esquerda)[11] e sua arquitetura (esquema da direita)*

A arquitetura das redes neurais corresponde à quantidade de neurônios nas camadas de entrada, intermediária e saída. O treinamento é a calibração da rede para que ela execute uma determinada tarefa, onde ocorrem os ajustes dos pesos sinápticos para diminuir o erro (diferença entre o valor real e o previsto). O método mais utilizado é a retropropagação de erros, onde o referido ajuste é realizado da última camada em direção à primeira, sendo cada ciclo de ajustes dos pesos sinápticos denominado de época de treinamento.[24,25]

O modelo utilizado para máquina de vetor de suporte é uma generalização do algoritmo *Generalized Portrait*, proposto nos anos 60 por Vapnik *et al.*.[26] A SVM é considerada em várias bibliografias um tópico de redes neurais artificiais.[25] Essa ferramenta apresenta as seguintes vantagens em relação à ANN-MLP: alta capacidade de generalização; robustez em espaços multidimensionais; teoria de

aprendizado bem estabelecida dentro da matemática e da estatística e, convexidade da função objetivo – convergindo para apenas um único mínimo na superfície de resposta. Contudo, a obtenção do modelo SVM depende de programação quadrática na resolução de equações não lineares, apresentando, muitas vezes, um alto custo computacional. Uma alternativa a esta desvantagem foi sugerida por Suykens *et al.*, que propuseram o uso de máquinas de vetores de suporte por mínimos quadrados (LS-SVM), baseando-se em uma estimativa linear ($y_i = w^T\phi(x_i) + b$), onde ϕ é uma função de linearização.

De forma análoga às redes neurais artificiais, o LS-SVM utiliza funções núcleos para processamento da informação, sendo a função Kernel de base radial (RBF - *Radial Basis Function*) a mais utilizada,

$$\phi = \exp\left(\frac{-\| x_i - x\|^2}{2\sigma^2}\right) \quad (4)$$

sendo σ^2 é um parâmetro ajustável, que controla a largura da gaussiana.

Atualmente, há algoritmos de treinamento disponíveis capazes de maximizar a capacidade de generalização de uma forma sistemática para LS-SVM. Para isso, é necessário minimizar a função custo,

$$C = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\gamma\sum_{i=1}^{N} e_i^2, \quad (5)$$

onde e é o erro de previsão, $\boldsymbol{w}$ são os pesos e γ o peso da função custo. Esta função penaliza o erro e pesos elevados, que geram excessiva variância comprometendo a capacidade de generalização do modelo. A regressão multivariada está representada na Equação 6:

$$y_i = \boldsymbol{w}^T\mathrm{f}(x_i) + b + e_i \quad (6)$$

onde x_i são as variáveis independentes e y_i, as variáveis dependentes. Através de ϕ, neste caso uma RBF, é possível mapear as variáveis não lineares para um espaço de maior dimensão linearmente separáveis, possibilitando a construção do modelo por mínimos quadrados. Assim, pode-se dizer que ϕ é uma função de linearização dos dados e a variância σ^2 (largura da função núcleo, Equação 4) será o indicador do esforço empregado ao linearizar os dados. Através do método dos multiplicadores de Lagrange resolve-se o problema de otimização convexa, imposto nas Equações 4 e 5 e, então, obtém-se os coeficientes de regressão ($\boldsymbol{w}$) como uma expansão dos multiplicadores de Lagrange.[12,27] A representação matemática da obtenção desses coeficientes é bem documentada e não será exposta aqui. Na literatura existem descrições e análises detalhadas sobre as ferramentas ANN-MLP e LS-SVM.[11,12,24-26]

A otimização do modelo pode ser realizada pela superfície de resposta para a raiz quadrada do erro médio quadrático da validação cruzada (RMSECV - *Root Mean Square Error of Cross-validation*) em função de σ^2 (largura da função núcleo, Equação 4) e γ (peso da função custo, Equação 5).[12]

A imensa variedade das amostras considerada neste trabalho, provenientes de todas as regiões do estado brasileiro, pode levar a desvios consideráveis da linearidade na correlação entre os espectros MIR e os teores de MOS. Isso justifica a proposição e avaliação do desempenho de modelos não lineares através das técnicas ANN-MLP e LS-SVM em relação ao método clássico linear PLS. Desvios da linearidade são esperados em espectroscopia de matrizes complexas, como alimentos, tecidos vegetais e solos, principalmente quando são considerados conjuntos amostrais diversificados.

Neste trabalho são propostas calibrações multivariadas aplicando ANN-MLP e LS-SVM aos espectros MIR de um conjunto de 1117 amostras de solo brasileiras para determinação de matéria orgânica. Além disto, foram aplicadas a MSC para amenizar efeitos de espalhamento de luz e a PCA para a redução da dimensionalidade. Avaliou-se a aplicação na rotina laboratorial dos modelos multivariados obtidos, comparando o erro de predição desses com o desvio laboratorial associado à metodologia de Walkley-Black.

PARTE EXPERIMENTAL

Amostras

Foram considerados 367 perfis de solo, de maneira a se obter um conjunto representativo do território brasileiro.[10] As amostras foram selecionadas de 1 a 4 horizontes diagnóstico por perfil de solo, resultando em 1117 amostras. O número dos horizontes diagnóstico depende do tipo de solo. As amostras foram fornecidas pelo Centro Nacional de Pesquisa de Solos (CNPS) da EMBRAPA e algumas amostras do Estado de São Paulo pelo Instituto Agronômico de Campinas (IAC). O pré-tratamento das amostras consistiu na sua secagem a 65 °C, seguida pela moagem e peneiração em malha de 80 mesh.

Método de referência

O método Walkley-Black foi usado como referência para determinação do teor de matéria orgânica nas amostras de solo. Este método é atualmente o mais utilizado para este fim, por demandar pouco investimento e ser bem estabelecido. Nesta metodologia a matéria orgânica do solo é oxidada pela ação de dicromato de potássio em meio ácido. Após a reação, o excesso de dicromato é titulado com solução de Fe^{2+} e a MOS é determinada indiretamente, usando-se ácido fosfórico e difenilamnia ou ferroína como indicador.[4,28]

Espectroscopia no infravermelho médio por refletância difusa

A varredura espectral na região do infravermelho médio foi realizada com um espectrômetro com transformada de Fourier Digilab FTS-7000 (Bio-Rad, Randolph, MA) equipado com amostrador automático e acessório de refletância. A região espectral utilizada foi 4000 a 400 cm^{-1} (MIR), com resolução de 4 cm^{-1}, com 64 varreduras/leitura. O KBr é utilizado como branco.

Análise estatística

Para o pré-tratamento de dados, remoção de *outliers*, agrupamentos de espectros, redução da dimensionalidade dos dados e calibrações multivariadas foram utilizados um microcomputador Processador Intel® Core™ 2 Quad 2.67 GHz, 4 Gbytes de memória RAM; e o software MATLAB™, versão 7.12.0.635 (*The MathWorks*, Natick, EUA), PLS Toolbox 6.2 (*Eigenvector Technologies*, Manson, EUA)[29] e LS-SVM Toolbox 1.5 (*World Scientific*, Singapore). http://www.esat.kuleuven.be/sista/lssvmlab/toolbox.html.[27]

Pré-tratamento espectral e remoção de *outliers*

Para se obter a correção do espalhamento multiplicativo de sinal (MSC) é necessário um espectro de referência. Neste trabalho, considerou-se como referência o espectro médio entre todas as amostras. Além da aplicação do MSC foi subtraído dos espectros originais o espectro de referência. Estes procedimentos corrigem desvios de linha de base não lineares (*drift*) e diminuem a dispersão dos espectros, provocados pelo espalhamento da luz que é intensificado pela heterogeneidade da granulometria do solo.[17,30]

A identificação de *outliers*, amostras anômalas, foi realizada por *leverage* (indicativo da distância entre uma amostra e a média de um conjunto de dados) extremo dos espectros (x_i) e por resíduos não

modelados na variável dependente, que é a MOS.

A identificação e remoção de *outliers* por *leverage h* seguiu a norma ASTME1655-05. Neste, amostras com h_i maiores que o limite

$$h_{limite} = 3\frac{A+1}{lc}, \quad (7)$$

foram removidas. Na Equação 7, *A* é o número de variáveis latentes e *lc* o número de amostras.

Após a calibração, foram identificadas e removidas como *outliers* as amostras que apresentarem erro absoluto (diferença entre o valor real e o previsto) maior que 3 vezes o valor da raiz quadrada do erro médio da calibração (RMSEC - *Root Mean Square Error of Calibration*). Este tipo de consideração, para uma distribuição normal, inclui nos ajustes 99% dos indivíduos, valor padrão utilizado neste tipo de análise.[31-33]

Seleção das amostras para validação

A separação dos grupos para calibração e validação dos modelos foi realizada pelo algoritmo de Kennard-Stone, que seleciona as amostras de calibração por distância euclidiana a partir dos espectros. O processo se inicia selecionando duas amostras, a mais próxima da média e a mais distante. Em seguida, ambas são removidas e o processo é repetido até se alcançar o número desejado para calibração.[34] Foi utilizado como critério 80% dos dados para calibração e 20% para validação.

A seguir, estão descritas quatro propostas para construção de modelos multivariados para predição de MOS, através dos quais é possível comparar as técnicas de ajuste ANN-MLP e LS-SVM. As combinações das ferramentas multivariadas em cada proposta avaliada estão resumidas na Tabela 1.

Tabela 1. Descrição dos modelos multivariados estudados nas quatro propostas

	Propostas			
	1	2	3	4
Pré-tratamento	MSC	MSC	MSC	MSC
Redução da dimensionalidade	PCA	PCA	-----	-----
Técnica de calibração	ANN-MLP	ANN-MLP	LS-SVM	LS-SVM
Amostras utilizadas	Todas	2 grupos	Todas	2 grupos

Proposta 1 – Regressão multivariada por ANN-MLP

Consiste em inicialmente realizar uma redução da dimensionalidade dos dados por PCA. Em seguida, avalia-se a raiz quadrada do erro médio quadrático de previsão (RMSEP - *Root Mean Square Error of Prediction*) em função do número de épocas de treinamento da rede neural multicamadas, onde o sinal de entrada são os escores de 15 componentes principais (PC - *Principal Components*), que explicam 99% da variância da matriz original dos espectros. A arquitetura inicial da ANN-MLP para esta avaliação foi: i) 1 camada intermediária com 6 neurônios com função de transferência sigmoidal; ii) 1 camada de saída com 1 neurônio com função de transferência linear e iii) treinamento por retropropagação de erros com 5 níveis para o número épocas ($5{,}0x10^3$; $1{,}0x10^4$; $1{,}5x10^4$; $2{,}0x10^4$; $2{,}5x10^4$);

Após a decisão do número de épocas de treinamento utilizadas, avalia-se o número de PC em 7 níveis (49, 42, 35, 28, 21, 14, 7) como sinal de entrada e o número de neurônios (N) da camada intermediária em 7 níveis (2, 4, 6, 8, 10, 12, 14). O treinamento em cada combinação de nível foi realizado em triplicata. Uma superfície de resposta (SR)[35] com os dados padronizados para identificar a região de mínimo para RMSEP é então construída.

Proposta 2 – Regressão multivariada por ANN-MLP em grupos espectrais

As amostras de solo são separadas em 4 grupos espectrais pela análise hierárquica de agrupamentos, usando distância euclidiana e ligação por média ponderada.[23,36] De forma análoga à proposta 1, é realizado o procedimento de otimização da arquitetura da ANN-MLP em função do RMSEP separadamente para os dois maiores grupos espectrais de solos.

Proposta 3 – Regressão multivariada por LS-SVM

Nesta proposta, não há redução de dimensionalidade, ou seja, todo espectro MIR é utilizado como sinal de entrada. Utiliza-se como técnica de ajuste multivariada a LS-SVM e como função núcleo, a RBF. A otimização do modelo é realizada pela superfície de resposta para RMSECV em função dos parâmetros σ^2 e γ, através de ferramenta fornecida pelo pacote LS-SVM.[27]

Proposta 4 – Regressão multivariada por LS-SVM em grupos espectrais

As amostras de solo são separadas em 4 grupos espectrais pela análise hierárquica de agrupamentos. Para os dois maiores grupos faz-se a construção e otimização do modelo por LS-SVM, ou seja, é aplicado procedimento análogo à proposta 3 para os grupos espectrais de solos.

RESULTADOS E DISCUSSÃO

Os espectros MIR das amostras de solo consideradas estão apresentados na Figura 2. A remoção de *outliers* por *leverage* extremo, através da norma ASTME1655-05, identificou 9,07% dos espectros como amostras anômalas. Considerando a diversidade da coleção de solos estudada, suspeita-se que os *leverages* extremos não estejam associados a *outliers* com características individuais, mas sim a um grupo ou grupos de solos com características físicas e químicas discrepantes da média global. Isso é corroborado ao se realizar a separação das amostras em grupos espectrais, propostas 2 e 4, reduzindo para 1,71% a remoção de *outliers*. Este resultado está de acordo com outros estudos que demonstram correlação entres os espectros no infravermelho e as características químicas e físicas do solo.[10,15,37] Os parâmetros texturais, argila e areia, e o teor de MOS dos 4 grupos espectrais estão descritos na Tabela 2. Os dados mostram uma alta dispersão dos parâmetros texturais e do teor de MOS. A formação dos grupos espectrais, em geral, está relacionada com parâmetros físico-químicos das amostras.

Propostas 1 e 2 - Regressão multivariada por ANN-MLP

A redução da dimensionalidade dos dados é necessária porque com um número muito grande de variáveis há risco do treinamento não convergir, subajuste, ou de necessitar de arquitetura complexa para modelagem dos dados, o que aumenta as chances de sobreajuste. Situação na qual a ANN-MLP se torna especialista nos dados do treinamento, o que não é conveniente para a aplicação desejada.[25,38,39] Neste trabalho, as arquiteturas testadas para a ANN-MLP não demonstraram capacidade de utilizar como sinal de entrada o espectro

Tabela 2. Características dos modelos MIR-ANN-MLP e MIR-LS-SVM para as quatro propostas estudadas, assim como os resultados do modelo MIR-PLS

Prop. - grupo - modelo	RMSEC (g kg^{-1})	%RMSEC	$r^2_{calibração}$	n_c	RMSEP (g kg^{-1})	%RMSEP	$r^2_{previsão}$	n_v
1 - geral - ANN-MLP	2,41	12,88	0,95	775	5,64	30,14	0,84	187
2 - grupo 1 - ANN-MLP	0,82	7,88	0,99	168	7,33	70,19	0,64	41
2 - grupo 2 - ANN-MLP	1,87	12,71	0,99	630	15,66	106,46	0,83	157
3 - geral - LS-SVM	1,60	8,55	0,97	766	3,61	19,29	0,93	187
3 - geral - LS-SVM*	1,26	7,91	0,98	398	3,26	17,42	0,96	99
4 - grupo 1 - LS-SVM	1,37	13,17	0,97	160	2,98	28,64	0,91	39
4 - grupo 2 - LS-SVM	0,21	1,43	0,99	602	4,36	29,64	0,98	144
Geral - PLS	3,90	20,84	0,78	781	5,60	29,93	0,76	195

*amostras com MOS>7 g kg^{-1}, Prop. = proposta, n_c = número de amostras na calibração, n_v = número de amostras na validação.

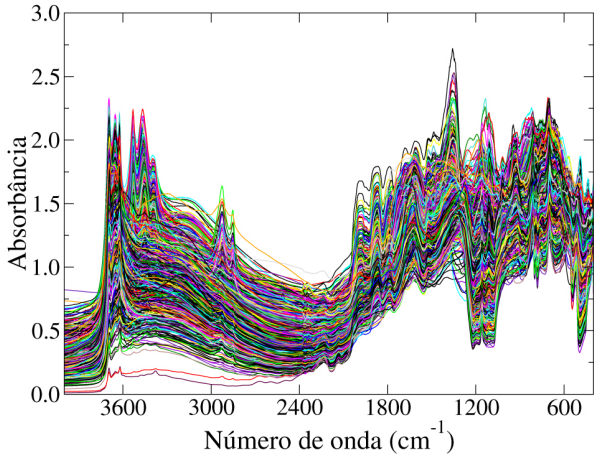

Figura 2. *Espectros MIR originais das 1117 amostras de solo*

completo. Isto ocorre devido ao subajuste que dificulta e acaba por impedir a convergência do erro de treinamento. Como solução alternativa utilizou-se a PCA, que foi capaz de reduzir a dimensionalidade dos dados sem perda significativa da informação. Apesar dos espectros originais possuírem 1868 dimensões, a redução por PCA mostrou que a utilização das 15 primeiras componentes principais é suficiente para reter 99% da variância total dos dados.

Testes prévios, utilizando as 15 PCs como sinal de entrada e 6 neurônios na camada intermediária da ANN-MLP, demonstraram que os modelos que apresentavam menor RMSEP eram obtidos em média entre 8500 e 13000 épocas de treinamento. Embora a rotina escrita para MATLAB™ armazene durante o treinamento a melhor rede para o conjunto de validação, este estudo prévio foi feito para definir o número mínimo de épocas que garante os mais baixos RMSEP, otimizando o custo computacional de obtenção dos modelos. Levando em conta uma margem de erro considerável, foram utilizadas 20000 épocas de treinamento como valor padrão nas simulações numéricas.

Para a proposta 1, o estudo para RMSEP(PC, N) com os dados normalizados resultou em um modelo quadrático não aditivo (RMSEP = -1,219 - 0,339 PC + 0,316 PC^2 + 0,573 N + 0,656 N^2), a superfície de resposta deste modelo está representada na Figura 3. Através de derivada parcial identificou-se o ponto de mínimo; a coordenada convertida para dimensão original, com menor RMESP foi PC = 26,45 e N = 8,84. Esses valores devem ser números inteiros; portanto, consideram-se 26 PCs como sinal de entrada e 9 neurônios para a camada intermediária da ANN-MLP como o melhor modelo. Trabalhos empregando PLS, PCR e ANN-MLP utilizam como sinal de entrada poucas PCs ou variáveis latentes, quase sempre um número menor que 10. Aqui, foram utilizadas 26, um número elevado em comparação com outros trabalhos, que em princípio pode incluir ruído espectral no modelo. Entretanto, justifica-se o uso deste número alto de PCs devido à complexidade das substâncias orgânicas presentes no solo e a diversidade das amostras estudadas. Assim, devem existir vários compostos orgânicos de concentrações independentes entre si e que vibram em diferentes regiões do espectro, gerando então variâncias em diferentes componentes principais. Esta proposição vai de encontro ao observado em trabalhos que caracterizam a matéria orgânica do solo.[7-9]

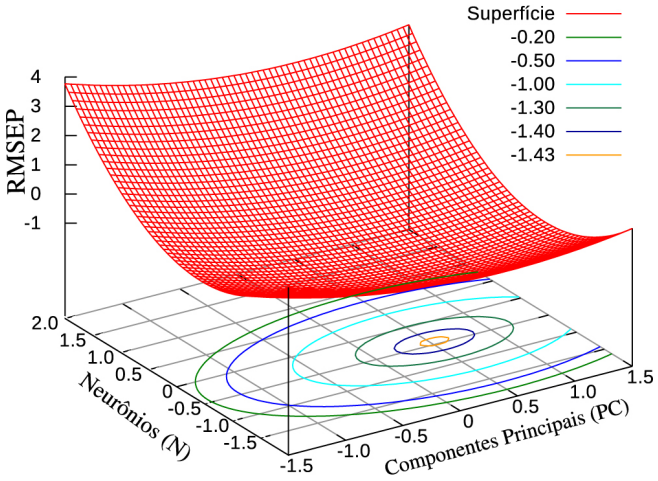

Figura 3. *Superfície de resposta normalizada obtida para a RMSEP da proposta 1*

Este modelo, proposta 1, apresentou RMSEP de 5,64 g kg^{-1} (30,14% relativo à média de todas as amostras) e RMSEC de 2,41 g kg^{-1} (12,88%). Estão condensadas as características das 4 propostas avaliadas na Tabela 2, juntamente com os resultados do modelo PLS considerando os dados de entrada do modelo da proposta 3. O valor encontrado para RMSEP, 5,64 g kg^{-1}, é alto quando comparado ao desvio de 2,8 g kg^{-1} associado ao método referência Walkley-Black; valor obtido em amostras de solo analisadas no LAA. Isto indica que a substituição do método referência por este modelo implicaria em uma diminuição na qualidade do resultado.

Para o grupo 1 na proposta 2, o estudo para RMSEP(PC, N) resultou em um modelo quadrático aditivo (RMSEP = -1,072 - 0,151 PC + 0,546 PC^2 + 0,396 N + 0,571 N^2 - 0.254 PC x N). Através desta equação identificou-se o ponto de mínimo da função como PC = 30 e N = 7. Este modelo apresentou RMSEP de 7,33 g kg^{-1} (70,19%) e RMSEC de 0,82 g kg^{-1} (7,88%), Tabela 2. Em relação à proposta 1, a previsão por RMSEP piorou e a calibração por RMSEC melhorou,

indicando sobreajuste do modelo. Para o grupo 2, o estudo para RMSEP(PC, N) resultou em um modelo quadrático não aditivo (RMSEP = -1,034 - 0,067 PC + 0,732 PC^2 + 0,333 N + 0,371 N^2). A condição ótima para a região estudada foi PC = 39 e N = 11; netas condições, a ANN-MLP apresentou RMSEP de 15,66 g kg^{-1} (106,46%) e RMSEC de 1,87 g kg^{-1} (12,71%), Tabela 2. Novamente os resultados indicam sobreajuste. O número elevado de variáveis na ANN, como as PCs, implica na necessidade de um número maior de neurônios , o que em princípio pode levar ao sobreajuste da rede neural artificial.[25,40]

Não foram realizadas otimização de arquitetura da ANN-MLP por superfície de resposta (RMSEC(PC,N)) para os 2 outros grupos da Tabela 3. Para o grupo 3, testes prévios na região mediana do planejamento experimental (PC = 28; N = 8) apresentaram RMSEP de 9,94 g kg^{-1} (121,03% relativo à média do teor de MOS do grupo 3, 7,99 g kg^{-1}) e r^2 = 0,25. Este erro inviabiliza a aplicação da técnica proposta para fins analíticos. O valor elevado do RMSEP obtido para o grupo 3 pode ser explicado pelo número reduzido de amostras, 69. Além disto, os solos do grupo 3 possuem baixos teores de MOS, onde existe uma baixa precisão da metodologia Walkley-Black na sua determinação. Para o grupo 4, não foi construído um modelo multivariado também devido ao número reduzido de amostras, 24. Este grupo é formado por organossolos que apresentam altos teores de MOS, encontrados geralmente na Região Amazônica.

Tabela 3. Grupos espectrais por análise hierárquica de agrupamentos e respectivos dados texturais e desvios-padrão associados

Grupos	n	MOS (g kg^{-1})	argila (g kg^{-1})	areia (g kg^{-1})
1	231	11,64 ± 9,59	131,67 ± 101,65	231,90 ± 152,64
2	846	16,18 ± 30,34	280,52 ± 186,47	509,19 ± 229,33
3	69	7,40 ± 8,34	508,33 ± 253,58	893,38 ± 109,45
4	24	199,93 ± 112,67	895,40 ± 38,59	945,20 ± 1,57

A calibração geral por ANN-MLP, onde são consideradas todas as amostras sem classificação por grupos, para determinação de MOS, proposta 1, é preferível frente à calibração intragrupo, proposta 2, por apresentar maior capacidade de predição de MOS. As duas propostas apresentaram RMSEP superiores ao desvio associado ao método referência Walkley-Black.

Propostas 3 e 4 - Regressão multivariada por LS-SVM

Através da otimização do modelo pela superfície de resposta para RMSECV, proposta 3, os melhores níveis testados para os parâmetros e foram 1077,9 e 8711,1, respectivamente, para o modelo LS-SVM. Este modelo apresentou RMSECV de 2,71 g kg^{-1} (25,72%), RMSEP de 3,61 g kg^{-1} (19,29%) e RMSEC de 1,60 g kg^{-1} (8,55%). As características do modelo estão descritas na Tabela 2. Não houve a necessidade da redução da dimensionalidade por PCA, o que já era previsto em outros trabalhos.[12,38] O resultado obtido para RMSEP, 3.61 g kg^{-1}, melhorou em relação à proposta 1, obtendo um desvio maior que 2,8 g kg^{-1} associado à metodologia de referência. Uma possível razão para isto são as limitações na determinação experimental em amostras com baixo teor de MOS.

As amostras com baixo teor de MOS apresentam erros relativos altos, devido à imprecisão do método Walkley-Black quando se tem baixas concentrações de MOS. Estes erros ocorrem, principalmente, devido a dois fatores: a dificuldade de se determinar o ponto de viragem da titulação e ao desvio associado à heterogeneidade granulométrica das amostras.[28,41] Estas deficiências, em princípio, comprometem a qualidade dos ajustes por LS-SVM e ANN-MLP aumentando o erro percentual, já que essas amostras contribuem para diminuição do teor médio de MOS.

A fim de verificar a interferência das amostras com baixo teor de MOS, otimizou-se um modelo por LS-SVM retirando as amostras com MOS < 7g kg^{-1}, reduzindo o número de amostras no ajuste multivariado para 497. Os melhores níveis testados para os parâmetros e foram 1180,3 e 41724, respectivamente. Nestas condições, na Figura 4 estão representados os valores previstos e os de referência. Este modelo apresentou RMSECV de 1,71 g kg^{-1} (9,14%), RMSEP de 3,26 g kg^{-1} (17,42%) e RMSEC de 1,29 g kg^{-1} (6,73%). Retirando-se as amostras de baixa concentração de MOS, obtém-se um modelo com maior capacidade de predição, ou seja, menor RMSEP. Além disso, ambos os ajustes da calibração e da previsão melhoraram, Tabela 2. Embora o RMSEP 3,26 g kg^{-1} deste modelo seja um pouco maior que o valor do desvio laboratorial 2,8 g kg^{-1}, podendo, em vários casos, ser utilizado na rotina laboratorial de análise, desonerando o trabalho técnico pela adoção de um erro relativo à média de MOS 2% acima do valor da técnica de referência. Além disso, em testes de proficiência (onde se considera a dispersão interlaboratorial) o desvio associado à metodologia Walkley-Black é maior que os apresentados por este método.

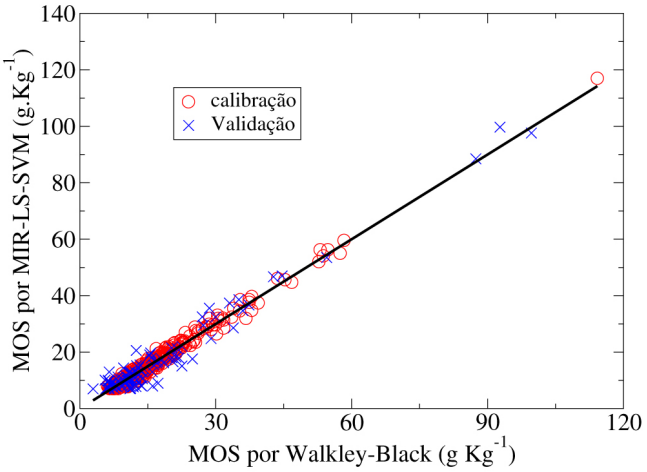

Figura 4. *Valor de referência, para as amostras com MOS > 7 g kg^{-1}, pelo método Walkley-Black versus o valor previsto pelo modelo de calibração MIR-LS-SVM, proposta 3*

Mesmo com número reduzido de amostras, o resultado melhorou em relação ao modelo LS-SVM anterior. Este resultado vem de encontro com as suspeitas da limitação do método Walkley-Black para determinar baixas concentrações de MOS. A retirada destas amostras não diminui a aplicação do modelo, pois para análise de fertilidade de solo é solicitada apenas a determinação de MOS na profundidade de 0 a 20 cm, região onde são encontrados teores acima de 10g kg^{-1}.[3,4]

Para o grupo 1, proposta 4, os melhores níveis testados para os parâmetros σ^2 e γ foram 1180,3 e 41724, respectivamente. Este modelo apresentou RMSEP de 2,98 g kg^{-1} (28,64%) e RMSEC de 1,37 g kg^{-1} (13,17%), Tabela 1. Para o grupo 2, os parâmetros otimizados foram σ^2 = 39,75 e g = 1180,3, resultando em RMSEP de 4,36 g kg^{-1} (29,64%) e RMSEC de 0,21 g kg^{-1} (1,43%), Tabela 1.

Comparando a capacidade de previsão dos modelos obtidos nas propostas 3 e 4, não se observa vantagem significativa que justifique realizar a calibração intragrupo. Contudo, optando pela aplicação da proposta 4, haveria a necessidade de um modelo classificador, ou seja, um modelo que indique a qual dos 4 grupos espectrais pertence uma amostra desconhecida. Para tanto, otimizou-se um modelo LS-SVM classificador, utilizando função RBF, os parâmetros otimizados são as matrizes σ^2 = [18,69; 5,15] e γ = [33,32; 10,40], já que para

classificar se utiliza linguagem binária, sendo necessários 2 sinais de saída para classificação em 4 grupos. Dos 1117 solos separados em 4 grupos espectrais, o modelo errou na classificação de apenas 3; 1 para o conjunto calibração e 2 para o conjunto validação. Isso corresponde um índice de acerto de mais de 99%.

Comparações de propostas e técnicas de ajuste multivariado

Independentemente da técnica de ajuste multivariado utilizada, ANN-MLP ou LS-SVM, as calibrações gerais, propostas 1 e 3, foram preferíveis por não exigir classificação prévia dos solos e apresentar RMSEP menor ou equiparável às calibrações intragrupo, propostas 2 e 4.

A ferramenta LS-SVM demonstrou maior capacidade de previsão de MOS que a ANN-MLP, resultado já apontado em um estudo recente que utilizou estas duas ferramentas em determinações químicas na madeira de eucalipto por cromatografia gasosa.[38] Isto se deve à existência de ferramentas mais eficazes para aperfeiçoar os modelos por LS-SVM, como otimização dos seus parâmetros por validação cruzada. Além disso, o modelo LS-SVM apresenta convexidade da função objetivo (RMSECV (σ^2,γ)) o que não ocorre com a ANN que pode convergir para mínimos locais.[12,38] Fidêncio *et al.* obtiveram RMSEP 2,5 g kg^{-1} e r^2 = 0,96 para determinação de MOS utilizando ANN, resultados compatíveis com os modelos LS-SVM aqui propostos, contudo consideraram um conjunto substancialmente mais restrito, 150 solos do Estado de São Paulo.[42]

Os resultados obtidos por LS-SVM foram superiores aos obtidos por PLS utilizando pré-processamento dos dados por MSC e 5 variáveis latentes, no qual obteve-se RMSEP de 5,60 g kg^{-1} e RMSEC de 3,90 g kg^{-1}, Tabela 2. O modelo LS-SVM também demonstrou melhor desempenho que os resultados obtidos por Madari *et al.*, que construíram modelos multivariados por PLS para determinação de MOS na mesma coleção de solos. Madari avaliou tanto a calibração geral como calibrações em grupos de 7 diferentes faixas de concentração de MOS; apenas em 1 dos grupos (4 g kg^{-1} < MOS < 30 g kg^{-1}) o desvio obtido de 2,70 g kg^{-1} é comparável aos RMSEPs das propostas 3 e 4. Todos os outros modelos por PLS propostos apresentaram RMSEP maior que 4,2 g kg^{-1}.[10]

Essa capacidade de ajuste equivalente e, em alguns casos, superiores do LS-SVM em relação à técnica mais difundida de calibração multivariada PLS já foi verificada por Ferrão *et al.*.[12] PLS é uma técnica de ajuste linear entre o parâmetro de interesse e as variáveis latentes obtidas a partir das variáveis independentes. No entanto, vários fenômenos podem causar o desvio dessa linearidade na região do infravermelho, dentre os quais destaca-se a dispersão por partículas ínfimas, interações moleculares, deslocamento de equilíbrio e alta concentração dos analitos, que é uma limitação da lei de Lambert-Beer.[17] Neste estudo, a diversidade de origem das amostras devido às dimensões continentais do Brasil, resultou em uma ampla faixa de teor de MOS e características físicas, o que intensificou desvios da linearidade, diminuindo em alguns casos a qualidade dos modelos por PLS. Estes fatores justificam a avaliação da utilização de técnicas capazes de modelar fenômenos não lineares, como ANN-MLP e LS-SVM, e explicam os resultados superiores obtidos por modelos LS-SVM em relação ao PLS.

Como já discutido anteriormente, o desvio associado à calibração geral por LS-SVM, proposta 3, tem resultado aplicável ao LAA. A implantação deste modelo em análises rotineiras implicaria, além de um ganho em operacionalidade, numa diminuição drástica no consumo de reagentes e geração de resíduos sulfocrômicos. Seria necessário esporadicamente analisar pelo método de referência amostras de solo para revalidação externa, como teste de controle de qualidade do modelo.

CONCLUSÃO

Dentre os modelos propostos por LS-SVM, a calibração geral para todas as amostras foi mais vantajosa e simples em relação às calibrações intragrupo. A ferramenta LS-SVM também mostrou uma grande capacidade na classificação dos solos, apresentando erros menores que 1%. Através da técnica LS-SVM foi possível ajustar modelos com exatidões próximas à metodologia de referência Walkley-Black. Assim, a associação da espectroscopia MIR com LS-SVM mostra ser uma alternativa limpa, operacional e de baixo custo para predição do teor de MOS. É importante destacar que os resultados de calibração são satisfatórios, mesmo quando os ajustes são aplicados a conjuntos de amostras de solos provenientes de todo território brasileiro. O que corresponde a uma grande diversidade da matriz de dados, resultando em modelos multivariados de aplicação abrangente.

AGRADECIMENTOS

Ao pesquisador Dr. R. M. Coelho, do Instituto Agronômico de Campinas (IAC), por disponibilizar amostras do estado de São Paulo para complementar a coleção de solos brasileiros utilizados neste estudo. Ao pesquisador Dr. J. B. Reeves III, do *United States Department of Agriculture* (USDA/ARS), pela colaboração técnica em calibração multivariada. O estudo teve apoio do CNPq processos 305031/2008-2 e 476287/2008-1 e Embrapa SEG 030605029.

REFERÊNCIAS

1. Novais, R. F.; Alvarez, V. V. H.; Barros, N. F.; Fontes, R. L. F.; Cantarutti, R. B.; Neves, J. C. L.; *Fertilidade do solo*, Sociedade Brasileira de Ciência do Solo: Viçosa, 2007.
2. Roscoe, R.; *Dinâmica da matéria orgânica do solo em sistemas conservacionistas: modelagem matemática e métodos auxiliares*, Embrapa Agropecuária Oeste: Dourados, 2006.
3. Madari, B. E.; Cunha, T. J. F.; Novotny, E. H.; Milori, D. M. B. P.; Martin Neto, L.; Benites, V. M.; Coelho, M. R.; Santos, G. A. Em *As terras pretas de* índio *da Amazônia: sua caracterização e uso deste conhecimento na criação de novas* áreas; Teixeira, W. G.; Kern, D. C.; Madari, B. E.; Lima, H. N.; Woods, W., eds.; Embrapa Amazônia Ocidental, 2009, cap. 13.
4. Silva, F. C.; *Manual de Análises Químicas de Solos, Plantas e Fertilizantes*, 1ª ed., Embrapa Comunicação para Transferência de Tecnologia: Brasília, 1999.
5. Raij, B. V.; Andrade, J. C.; Cantarella, H.; Quaggio, J. A.; *Análise química para avaliação da fertilidade de solos tropicais*, Instituto Agronômico: Campinas, 2001.
6. Nriagu, J. O.; Nieboer, E.; *Chromium in the natural and human environments*, Wiley Inter-Science: Ontario, 1988.
7. Silva Filho, A. V.; Silva, M. I. V.; *Resumos do Simpósio Nacional sobre as Culturas do Inhame e do Taro - II*, João Pessoa, Brasil, 2002.
8. Wander, M. Em *Advances in agroecology*; Magdov, F.; Weil, R., eds.; CRC: Boca Raton, 2004, chap. 3.
9. White, J. L.; Roth, C. B. Em *Methods of soil analysis. Part 1 - Physical and Mineralogical Methods*, 2nd ed., Agronomy: Madison, 1986, chap. 11.
10. Madari, B. E.; Reeves, J. B.; Coelho, M. R.; Machado, P. L. O. A.; De Polli, H.; *Spectrosc. Lett*. **2005**, *38*, 721.
11. Cerqueira, E. O.; De Andrade, J. C.; Poppi, R. J.; Mello, C.; *Quim. Nova* **2001**, *24*, 864.
12. Ferrão, M. F.; Mello, C.; Borin, A.; Maretto, D. A.; Poppi, R. J.; *Quim. Nova* **2007**, *30*, 852.
13. Morgano, M. A.; Faria, C. G.; Ferrão, M. F.; Ferreira, M. M. C.; *Quim. Nova* **2007**, *30*, 346.

14. Ferrão, M. F.; Furtado, J. C.; Neumann, L. G.; Konzen, P. H. A.; Morgano, M. A.; Bragagnolo, N.; Ferreira, M. M. C.; *Tecno-lóg.* **2003***, 7*, 9.
15. Sena, M. M.; Poppi, R. J.; Frighetto, R. T. S.; Valarini, P. J.; *Quim. Nova* **2000**, *23*, 547.
16. Barros Neto, B.; Scarminio, I. S.; Bruns, R. E.; *Quim. Nova* **2006**, *29*, 1401.
17. Rinnan, A.; Norgaard, L.; Ber, F. V. D.; Thygesen, J.; Bro, R.; Engelsen, S. B. Em *Infrared spectroscopy for food quality analysis and control;* Sun, D., ed.; 1st ed.; Elsevier: Burlington, 2009, chap. 1.
18. Dhanoa, M. S.; Lister, S. J.; Sanderson, R.; Barnes, R. J.; *J. Near Infrared Spectrosc.* **1994***, 2*, 43.
19. Hair, J. F.; Tatham, R. L.; Anderson, R. E.; Black, W.; *Análise multivariada de dados*, 5ª ed., Bookman: Porto Alegre, 2005.
20. NØrgaard, L.; Saudland, A.; Wagner, J. P.; Nielsen, L. M.; Engelsen, S. B.; *Appl. Spectrosc.* **2000**, *54*, 413.
21. Centner, V.; Massart. D.; de Noord, O. E.; Jong, S.; Vanfegniste, B. M.; Sterna, C.; *Anal. Chem.* **1996**, *68*, 3851.
22. Leardi, R.; Seasholtz, M. B.; Pell, R. J.; *Anal. Chim. Acta* **2002**, *461*, 189.
23. Ferreira, D. F.; *Estatística multivariada*, 2ª ed., Ed. UFLA: Lavras, 2011.
24. Schimidt, F.; Bueno, M. I. M. S.; Poppi, R. J.; *Quim. Nova* **2002**, *25*, 949.
25. Hu, Y. H.; Hwang, J. Em *Handbook of Neural Networks signal processing*; Poularikas, A., ed.; CRC Press LLC: Boca Raton, 2000, chap. 1.
26. Vapnik, V.; Lerner, A.; *Automat. Remote Control* **1963**, *24*, 774.
27. Suykens, J. A. K.; van Gestel, T.; de Brabanter, J.; de Moor, B.; Vandewalle, J.; *Least-Squares Support Vector Machines*, World Scientific: Singapore, 2002.
28. Pansu, M.; Gautheryou, J.; *Handbook of soil analysis: Mineralogical, organic and inorganic methods*, Springer: Maur des Fossés, 2003.
29. Wise, B. M.; Gallagher, N. B.; Bro, R.; Shaver, J. M.; Windig, W.; Koch, R. S.; *PLS Toolbox 2.1 for use with MATLAB™*, Eigenvector Research Inc.: Manson, 2001.
30. Cogdill, R. P.; Anderson, C. A.; Delgado, M.; Chisholm, R.; Bolton, R.; Herkert, T.; Afnan, A. M.; Drennen III, J. K.; *AAPS PharmSciTech.* **2005**, *6*, 273.
31. Annual Book of ASTM Standards; *Standards practices for infrared, multivariate, quantitative analysis*, E1655, vol 03.06. ASTM International: West Conshohocken, 2000.
32. Valderrama, P.; Braga, J. W.; Poppi, R. J.; *J. Agric. Food Chem.* **2007**, *55*, 8331.
33. Martens, H.; Naes, T.; *Multivariate calibration,* Wiley: New York, 1996.
34. Kennard, R. W.; Stone, L. A.; *Technometrics* **1969**, *11,* 137.
35. Barros Neto, B.; Scarminio, I. S.; Bruns, R. E.; *Como Fazer Experimentos. Pesquisa e Desenvolvimento na Ciência e na Indústria*, Ed. UNICAMP: Campinas, 2003.
36. Goodall, C. R. Em *Computation using the QR decomposition. Handbook in Statistics*, *Statistical Computing*; Rao, C. R., ed.; Elsevier: Amsterdam, 1993, vol. 9.
37. Madari, B. E.; Reeves III, J. B.; Machado, P. L. O. A.; Guimarães, C. M.; Torres, E.; McCarty, G. W.; *Geoderma* **2006**, *136*, 245.
38. Nunes, C. A.; Lima, C. F; Barbosa, L. C. A.; Colodette, J. L.; Fidêncio, P. H.; *Quim. Nova* **2011**, *34*, 279.
39. Sabin, J. G.; Ferrão, M. F.; Furtado, J. C.; *Rev. Bras. Ciênc. Farm.* **2004**, *40*, 387.
40. Ritchie, M. D.; White, B. C.; Parker, J. S.; Hahn, L. W.; Moore, J. H.; *Bioinformatics* **2003**, *4*, 28.
41. Harris, D. C.; *Análise química quantitativa*, 6ª ed., LTC: Rio de Janeiro, 2005.
42. Fidêncio, P. H.; Poppi, R. J.; Andrade, J. C.; *Anal. Chim. Acta* **2001**, *453*, 125.