Resumos Expandidos- VI Simpósio Brasileiro de Educação em Solos- 22 a 25 de Maio de 2012 em Sobral, CE, Brasil.
Promoção: Sociedade Brasileira de Ciência do Solo. Organização: UVA-PROEX/LAPPEGEO

Proposta de tutorial de Quimiometria utilizando técnicas modernas para a análise de solos

André Marcelo de Souza[(1)]; Maurício Rizzato Coelho[(2)], Paulo Figueiras[(3)], Thais Affonso Fernandes Cunha[(4)], Ricardo de Oliveira Dart[(5)], Jerônimo Guedes Parés[(4)], Priscila Luzia Simon[(6)], Bruno Gomes da Cruz[(4)], Ronei Jesus Poppi[(7)], Maria de Lourdes Mendonça Santos[(2)], Ricardo Luis Louro Berbara[(8)]

[(1)] Pesquisador; Embrapa Solos; Empresa Brasileira de Pesquisa Agropecuária - Embrapa, Rua Jardim Botânico, 1.024 - Jardim BotânicoRio de Janeiro, RJ - Brasil - CEP 22460-000, andre.souza@cnps.embrapa.br; [(2)] Pesquisador; Embrapa Solos; Empresa Brasileira de Pesquisa Agropecuária - Embrapa, Rua Jardim Botânico, 1.024 - Jardim Botânico Rio de Janeiro, RJ - Brasil - CEP 22460-000; [(3)] Estudante; Instituto de Química; Universidade estadual de Campinas - Unicamp, Universidade Estadual de Campinas, Caixa Postal 6154, Campinas, SP, CEP 13083-970, [(4)] Estagiário; Embrapa Solos; Empresa Brasileira de Pesquisa Agropecuária - Embrapa, Rua Jardim Botânico, 1.024 - Jardim Botânico Rio de Janeiro, RJ - Brasil - CEP 22460-000, [(5)] Analista; Embrapa Solos; Empresa Brasileira de Pesquisa Agropecuária - Embrapa, Rua Jardim Botânico, 1.024 - Jardim Botânico Rio de Janeiro, RJ - Brasil - CEP 22460-000; [(6)] Departamento de Solos e Engenharia Agrícola; Universidade Federal do Paraná, Rua dos Funcionários, 1540 – Juveve – Curitiba-PR CEP 80035-050; [(7)] Pesquisador; Instituto de Química; Universidade estadual de Campinas - Unicamp, Universidade Estadual de Campinas, Caixa Postal 6154, Campinas, SP, CEP 13083-970; [(8)] Professor, Curso de Agronomia – Universidade Federal Rural do Rio de janeiro - UFRRJ, BR 465, KM 7, Seropédica, RJ, CEP. 23890-000.

Abstract
The use of spectroscopic techniques applied to soil science is recent and promising. Its application is invariably associated with Chemometrics. This paper aims to introduce the basic knowledge of these two techniques for graduate students and researchers in soil science. To this end, we propose a tutorial based on a study of 52 soil samples collected in the municipality of Manga, north of Minas Gerais, collected from profiles of different classes of soils, contrasting with respect to their chemical properties, physical and mineralogical. Soil samples were analyzed by Near Infrared Spectroscopy (NIR) and the spectra obtained were evaluated by Principal Component Analysis (PCA). The presented results and their presentation possible, students and teachers at undergraduate and postgraduate students in the area of soil science, a practical introduction to the PCA in Matlab, though little explored in the literature.

Resumo
A utilização de técnicas espectroscópicas aplicadas à ciência do solo é recente e promissora. Sua aplicação invariavelmente está associada à Quimiometria. Este trabalho tem por objetivo introduzir os conhecimentos básicos destas duas técnicas para estudantes de pós-graduação e pesquisadores da área de ciência do solo. Para tal, é proposto um tutorial baseado no estudo de 52 amostras de solos obtidas no município de Manga, norte de Minas Gerais, coletadas em perfis de diferentes classes de solos, contrastantes em relação aos seus atributos químicos, físicos e mineralógicos. Amostras de solos foram analisadas por Espectroscopia no Infravermelho Próximo (NIR) e os espectros obtidos, avaliados por Análise por Componentes Principais (PCA). Os resultados apresentados e sua forma de apresentação possibilitam, aos alunos e professores em nível de graduação e pós-graduação da área de ciência do solo, uma introdução prática da PCA em Matlab, ainda pouco divulgada na literatura mundial.

Palavras-Chave: Análise por Componentes Principais, aula prática, tutorial.

INTRODUÇÃO

O espectro eletromagnético contém desde as ondas de rádio, as microondas, o infravermelho, a luz visível, os raios ultravioleta, os raios X, até à radiação gama. A região do infravermelho está localizada logo a após a região do visível e se inicia em uma faixa conhecida como infravermelho próximo -do inglês, *Near Infrared Spectroscopy*, NIR- na faixa de 12800 a 4000 cm^{-1} ou 780 a 2500 nm, médio - do inglês, *Mid Infrared Spectroscopy*, MIR - na faixa de 4000 a 200 cm^{-1} ou 2500 a 5000 nm, e distante – do inglês, *Far Infrared Spectroscoy*, FAR - na faixa de200 a 10 cm^{-1} ou 5000 a 10000 nm. A maior parte das aplicações em ciência de solos tem sido realizada no infravermelho médio e próximo, amplamente utilizadas para análises qualitativas e quantitativas (Pasquini, 2003). A radiação na região do infravermelho é energética o suficiente para causar transições vibracionais de energia, que basicamente resulta numa impressão digital química e física da amostra de solos (Poppi *et al*. 2002). As técnicas de Espectroscopia no Infravermelho Próximo e no Infravermelho Médio estão sendo cada vez mais utilizadas em ciência do solo para medir vários atributos, tanto aqueles relacionados à sua composição química {por exemplo, teores de carbono orgânico, capacidade de troca catiônica, pH} e mineralógica, como aos parâmetros físicos dos solos (por exemplo, granulometria) (Vohland *et al*. 2011; Cañasveras *et al*. 2010; Demattê *et al*. 2004; Poppi *et al*. 2002; Mouazen *et al*. 2010; Nocita *et al*. 2011; Rossel and Behrens 2010; Fontán *et al*. 2010). A efetividade de seu uso generalizado como técnica analítica de rotina nos laboratórios de solos ainda requer muito esforço da pesquisa. No entanto, Fuentes et al. (2012) afirmam que é crescente na literatura as calibrações de sucesso entre os espectros NIR e, por exemplo, os teores de um ou mais atributos do solo; calibrações essas já suficientes para serem utilizadas como substitutas das análises convencionais de solos.

A radiação eletromagnética na região do Infravermelho próximo (do inglês, *Near Infrared Spectroscopy*), localizada entre 13.300 a 400 cm^{-1} ou 750 a 2500 nm, foi descoberta por Frederick William Herschel em 1800. Entretanto, esta faixa do espectro eletromagnético, caracterizada por bandas de absorção sobrepostas e fracas, foi ignorada por químicos analíticos e espectroscopistas até a metade do século XX (Pasquini 2003). Apesar desse ostracismo precoce, as potencialidades analíticas dessa região espectral foram claramente demonstradas três décadas depois em um artigo escrito por Wetzel, cujo sugestivo título é "Near-Infraed Reflectance Analysis - Sleeper Among Spectroscopic Techniques" publicado em 1983 (Davies, 1998), sendo justamente a partir da década de 1980 que a espectroscopia NIR começa a se estabelecer como promissora técnica analítica. Uma década depois, diversos artigos começaram a ser publicados com títulos enfáticos ressaltando a nova realidade da técnica, como por exemplo, o artigo publicado em 1994 com o título "Near Infrared Spectroscopy – The Gian its Running Strongly" (Pasquini 2003) e "The history of Near Infrared Spectroscopy Analysis: past, present and future – from sleeping technique to the morning star of the spectrosocpy, publicado em 1998 (Pasquini, 2003; Davies, 1998). O desenvolvimento de equipamentos com componentes eletrônicos e ópticos mais sofisticados e o advento de computadores capazes de processar eficientemente as informações contidas nos espectros NIR facilitaram a expansão da técnica em um número crescente de aplicações em diferentes campos (Pasquini, 2003).

O crescente interesse em espectroscopia NIR como técnica instrumental alternativa para análise de solos pode ser justificado pelas inúmeras e marcantes vantagens que esta técnica apresenta em relação às análises convencionais. Fuentes et al. (2012) argumentam que o fato da tecnologia NIR ser não destrutiva e livre de indesejáveis resíduos e impactos ambientais, ser barata, rápida, requerendo pouco manuseio das amostras quando combinada à quimiometria (ou análise multivariada), estão dentre as suas principais vantagens. Entretanto, a interpretação dos espectros NIR não é imediata, pois na sua região espectral são observadas bandas de sobreton e de combinção, que são pouco intensas e sobrepostas. A fim de possibilitar o uso dos espectros NIR e a interpretação eficaz dos seus resultados é fundamental a utilização de métodos quimiométricos. (Pasquini, 2003).

A análise exploratória através da PCA é um dos métodos mais empregados em Quimiometria, tendo cada vez mais novos usuários do meio acadêmico (alunos de graduação, pós-graduação e pesquisadores) interessados no seu emprego para interpretação de dados de espectroscopia NIR de amostras de solos (Sarkhot *et al*. 2011). Contudo, esses novos usuários estão sujeitos a limitações iniciais no estudo da Quimiometria devido à falta de conhecimento teórico prévio e habilidade de operação de softwares e ambientes computacionais indispensáveis à sua aplicação. Em geral, os artigos científicos e os livros da área apresentam a teoria e aplicação da PCA. Entretanto, sua abordagem na utilização dos ambientes computacionais é muito pouco didática. Por outro lado, a importância do entendimento dos procedimentos realizados pelos softwares é fundamental para avaliação dos resultados obtidos, bem como para o questionamento da maneira pelo qual tais softwares os realizam. Diversos softwares quimiométricos e ambientes computacionais estão disponíveis no mercado, dentre eles, pode-se destacar o Matlab, o Minitab, o Pirouette, o SIMCA-P+ e o Unscrambler, além de ambientes baseados em software livre, como o Octave e o R (Souza e Poppi, 2012).

A PCA permite a redução da dimensionalidade através da representação do conjunto de dados em um novo sistema de eixos, denominados Componentes Principais (PC), possibilitando a visualização da natureza multivariada dos dados em poucas dimensões. No espaço original, as amostras são pontos localizados em um espaço n- dimensional. Com a redução de dimensionalidade proporcionada pela PCA, as amostras passam a ser localizadas em espaços reduzidos, por exemplo, bi ou tri dimensionais. Matematicamente, na Análise de Componentes Principais, a matriz **X** é decomposta em um produto de duas matrizes, denominadas escores (do inglês, *scores*, **T**) e pesos (do inglês, *loadings*, **P**), mais uma matriz de erros (**E**), como mostrado na Equação 1:

$$\boldsymbol{X = TP^T + E} \qquad \text{Equação 1}$$

Os escores representam as coordenadas das amostras no sistema de eixos formados pelos Componentes Principais. Cada PC é constituído pela combinação linear das variáveis originais e os coeficientes da combinação são denominados pesos. Matematicamente, os pesos são os cossenos dos ângulos entre as variáveis originais e os Componentes Principais (PC), representando, portanto, o quanto cada

variável original contribui para uma determinada PC. A Primeira Componente Principal (PC1) é traçado no sentido da maior variação no conjunto de dados; a segunda (PC2) é traçado perpendicularmente a primeira, com o intuito de descrever a maior porcentagem da variação não explicada pela PC1 e assim por diante. Enquanto os escores representam as relações de similaridade entre as amostras, os pesos indicam a contribuição de cada variável para a formação das PC. Através da análise conjunta do gráfico de escores e pesos, é possível verificar quais variáveis são responsáveis pelas diferenças observadas entre as amostras. O número de PC a ser utilizado no modelo PCA é determinado pela porcentagem de variância explicada pelas PC. Assim, seleciona-se um número de PC de tal maneira que a maior porcentagem da variação presente no conjunto de dados originais seja capturada (Beeb, 1998). Diversas rotinas estão disponíveis para a realização da PCA e quatro deles aparecem freqüentemente na literatura: *Nonlinear Iterative Vartial Least Squares* (NIPALS), *Singular Value Decomposition* (SVD), os quais utilizam a matriz de dados **X**, POWER e *Eigenvalue Decompostion* (EVD) que trabalham com a matriz de produto cruzado **X**'.**X**. (Souza e Poppi, 2012; Wold, 2002). Neste tutorial será utilizado SVD devido a sua facilidade de execução tanto no Matlab quanto no Octave.

A etapa de pré-processamento dos dados é fundamental para o sucesso da análise multivariada. Os principais objetivos da aplicação das técnicas de pré-processamento são eliminar informações não relevantes do ponto de vista químico e tornar a matriz de dados melhor condicionada para a análise, possibilitando a subsequente análise exploratória do conjunto de dados com eficiência. Existe uma vasta literatura disponível a respeito dos diversos métodos de processamento de dados em espectroscopia. Centrar os dados na média, derivar e suavizar utilizando o algoritmo de Savitzky-Golay e aplicação de correção de espalhamento multiplicativo (MSC, Multiplicative Scatter Correction) são alguns dos métodos mais aplicados (Beebe *et al*., 1998).

A centralização na média (Beebe) consiste em fazer com que, para cada variável, seus valores tenham média zero. Para centrar os dados na média, obtêm-se, para cada coluna, o valor médio e, em seguida, subtrai-se este valor de cada variável dessa mesma coluna. Desta forma, ocorre a mudança do sistema de coordenadas para o centro dos dados. A Equação e é utilizada para centrar os dados na média.

$$x_{(i,j)}cm = x_{(i,j)} - \bar{x}_j \qquad \text{Equação 2}$$

em que, amostra $x_{(i,j)}$; corresponde ao valor centrado na média para a variável j, $x_{(i,j)}$, é o valor da variável j na amostra i e x_j é a média das amostras na coluna calculada pela Equação 3

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{(i,j)} \qquad \text{Equação 3}$$

Onde n representa o número de amostras.

A aplicação da primeira ou segunda derivada sobre os dados espectrais brutos é um procedimento que pode destacar ombros espectrais, bem como minimizar o efeito de inclinações provocadas na linha de

base dos espectros, devido à morfologia das partículas. Ao aplicarmos as operações de derivação aos espectros, as informações contidas ao longo dos diferentes comprimentos de onda são geralmente acentuadas. Não só os sinais espectrais, mas também os ruídos tornam-se acentuados, portanto, deve-se ter cuidado com a qualidade dos espectros com os quais se deseja aplicar o cálculo das derivadas (Beebe *et al*., 1998).

O método de correção de espalhamento multiplicativo (MSC - do inglês, Multiple Scatering Correction) é comumente aplicado em espectroscopia para a correção de linha base, proveniente principalmente da não homogeneidade da distribuição de partículas na matriz. Este método assume que os comprimentos de onda da luz espalhada possuem uma dependência distinta entre a luz espalhada e a absorvida pelos constituintes da amostra. Portanto teoricamente, é possível separar estes dois sinais. Este método tenta remover o efeito do espalhamento pela linearização de cada espectro por um espectro ideal. Para efeito de cálculo, considera-se que o espectro ideal é o espectro médio do conjunto de dados para o qual se deseja realizar a correção da linha base. Em seguida, utiliza-se uma regressão linear para calcular o coeficiente angular e linear do gráfico entre o espectro ideal e o espectro que vai ser corrigido. O espectro corrigido é calculado subtraindo cada ponto do espectro pelo valor do coeficiente linear e dividindo este valor pelo coeficiente angular (Svensson *et al*., 2002).

Matematicamente, e resumindo, a correção é feita da seguinte forma:

1. A partir do conjunto total de espectros, calcula-se o espectro médio $\bar{x}_l$.
2. Faz-se a regressão linear para cada um dos k espectros $x_{(ik)}$; do conjunto total de espectros, contra o espectro médio, sobre todos os i comprimentos de onda (Equação 4):

$$x_{(ik)} = \upsilon_k \bar{x}_i + \mu_k \qquad \text{Equação 4}$$

Onde υ_k é o coeficiente linear e μ_k é o coeficiente angular (Equação 5).

3. Correção final

$$x_{ik}{}^{(corrigido)} = \frac{x_{ik}{}^{(corrigido)} - \mu_k}{\upsilon_k} \qquad \text{Equação 5}$$

O objetivo desta pesquisa foi desenvolver um tutorial para ser utilizado como guia didático para alunos de pós-graduação e professores que desejam iniciar trabalhos de espectroscopia NIR e quimiometria aplicada a solos. Para tal, será utilizado o programa comercial Matlab.

MATERIAL E MÉTODOS

Caracterização dos solos e da área estudados

As amostras estudadas, em número de 52, foram obtidas de 11 perfis de solos descritos e coletados para fins de mapeamento de solos do Parque Estadual da Mata Seca (PEMS), município de Manga (MG). Com o clima semi-árido, vegetação de Caatinga e litologia relacionada aos depósitos quarternários do Rio São Francisco e neoproterozóicos da Formação Bambuí (calcáreos) (Iglesias e Uhlein, 2009), os solos estudados apresentam grande varibilidade espacial de seus atributos, incluindo-se o teor de Ca, com valores que variam de menos de 0,3 a 18,1 $cmol_c\ kg^{-1}$. Latossolos, Cambissolos, Vertissolos, Plintossolos, Neossolos Flúvicos, Gleissolos e Chernossolos são os solos predominantes do PEMS (Dart et al., 2010).

Obtenção dos espectros

As análises de NIR foram realizadas em um espectrômetro Spectrum 100N, PerkinElmer, equipado com acessório de reflectância difusa NIRA, sendo os espectros obtidos em duplicata na faixa de 600 a 4000 cm^{-1} com resolução de 4 cm^{-1}, 64 varreduras, na faixa de 4000 a 7800 cm^{-1}.

Realização da PCA

A PCA foi executada em Matlab 7.0 e foi utilizado um toolbox gratuito, desenvolvido por Wen Wu and Sijmen de Jong, do FABI – Vrije Universiteit Brussel. Esta rotina permite o cálculo do PCA de uma matriz de dados, onde cada amostra é colocada numa linha e fornece os escores e pesos, assim como a porcentagem de variância descrita em cada componente principal e os escores de uma nova matriz teste.

RESULTADOS E DISCUSSÃO

O Matlab possui versões diferentes para diferentes ambientes operacionais e este tutorial está baseado em versões mais recentes para o ambiente Windows XP/Vista/7. Também as operações descritas a seguir podem ser realizadas em Octave. O Octave é um software livre e a sua utilização é bastante similar à do Matlab. O Octave pode ser encontrado no seguinte site: http://download.famouswhy.com/octave/free_download.html. Neste site está disponível a versão 3.2.4. Após a instalação do Octave, pode-se instalar o GUIOctave 1.0.14 (também disponível em http://www.soft82.com/get/download/windows/gui-octave). Este programa funciona como uma "máscara" para o Octave e permite a utilização de uma interface mais amigável. A seqüência típica de passos a ser executada com o objetivo de executar a PCA, a partir dos espectros NIR das amostras de solos é:

1. Carregar o conjunto de espectros

```
>>T011 = dlmread('T011.ASC','\t',56,1);
>>T012 = dlmread('T012.ASC','\t',56,1);
>>T021 = dlmread('T021.ASC','\t',56,1);
>>T022 = dlmread('T022.ASC','\t',56,1);
```

```
>>T031 = dlmread('T031.ASC','\t',56,1);
>>T032 = dlmread('T032.ASC','\t',56,1);
>>T041 = dlmread('T041.ASC','\t',56,1);
>>T042 = dlmread('T042.ASC','\t',56,1);
>>T051 = dlmread('T051.ASC','\t',56,1);
>>T052 = dlmread('T052.ASC','\t',56,1);
>>T061 = dlmread('T061.ASC','\t',56,1);
>>T062 = dlmread('T062.ASC','\t',56,1);
>>T071 = dlmread('T071.ASC','\t',56,1);
>>T072 = dlmread('T072.ASC','\t',56,1);
>>T081 = dlmread('T081.ASC','\t',56,1);
>>T082 = dlmread('T082.ASC','\t',56,1);
>>T091 = dlmread('T091.ASC','\t',56,1);
>>T092 = dlmread('T092.ASC','\t',56,1);
>>T101 = dlmread('T101.ASC','\t',56,1);
>>T102 = dlmread('T102.ASC','\t',56,1);
>>T111 = dlmread('T111.ASC','\t',56,1);
>>T112 = dlmread('T112.ASC','\t',56,1);
>>T121 = dlmread('T121.ASC','\t',56,1);
>>T122 = dlmread('T122.ASC','\t',56,1);
>>MANGA011 = dlmread('MANGA011.ASC','\t',56,1);
>>MANGA012 = dlmread('MANGA012.ASC','\t',56,1);
>>MANGA021 = dlmread('MANGA021.ASC','\t',56,1);
>>MANGA022 = dlmread('MANGA022.ASC','\t',56,1);
>>MANGA031 = dlmread('MANGA031.ASC','\t',56,1);
>>MANGA032 = dlmread('MANGA032.ASC','\t',56,1);
>>MANGA041 = dlmread('MANGA041.ASC','\t',56,1);
>>MANGA042 = dlmread('MANGA042.ASC','\t',56,1);
>>MANGA051 = dlmread('MANGA051.ASC','\t',56,1);
>>MANGA052 = dlmread('MANGA052.ASC','\t',56,1);
>>MANGA061 = dlmread('MANGA061.ASC','\t',56,1);
>>MANGA062 = dlmread('MANGA062.ASC','\t',56,1);
>>MANGA071 = dlmread('MANGA071.ASC','\t',56,1);
>>MANGA072 = dlmread('MANGA072.ASC','\t',56,1);
>>MANGA081 = dlmread('MANGA081.ASC','\t',56,1);
>>MANGA082 = dlmread('MANGA082.ASC','\t',56,1);
>>MANGA091 = dlmread('MANGA091.ASC','\t',56,1);
```

>>MANGA092 = dlmread('MANGA092.ASC','\t',56,1);

>>MANGA101 = dlmread('MANGA101.ASC','\t',56,1);

>>MANGA102 = dlmread('MANGA102.ASC','\t',56,1);

>>MANGA111 = dlmread('MANGA111.ASC','\t',56,1);

>>MANGA112 = dlmread('MANGA112.ASC','\t',56,1);

>>MANGA121 = dlmread('MANGA121.ASC','\t',56,1);

>>MANGA122 = dlmread('MANGA122.ASC','\t',56,1);

>>MANGA131 = dlmread('MANGA131.ASC','\t',56,1);

>>MANGA132 = dlmread('MANGA132.ASC','\t',56,1);

>>MANGA141 = dlmread('MANGA141.ASC','\t',56,1);

>>MANGA142 = dlmread('MANGA142.ASC','\t',56,1);

O sinal de maior (>>) indica que os comandos digitados acima estão organizados de forma a ser executado diretamente na janela *Command Window* do Matlab ou do Octave. Qualquer dúvida na execução de algum comando do Matlab pode ser sanada pela utilização do comando *help* como, por exemplo: help *load*.

2. Montar a matrix X

A matriz **X** é organizada da seguinte maneira: nas linhas são colocadas as amostras e nas colunas as variáveis. No caso dos espectros, nas linhas temos as amostras de solos e nas colunas os valores de log(1/R), onde R é a refletância que é proporcional a absorbância. O sinal de apostrofo (‘) representa transposição. Neste trabalho, foi necessário transpor os espectros porque estes são exportados em colunas, entretanto, na matriz de dados, estes devem estar presentes em linhas.

A matriz X contendo as 52 amostras de solo pode ser criada da seguinte forma:

>>**X**=[T011';T012';T021';T022';T031';T032';T041';T042';T051';T052';T061';T062';T071';T072';T081';T082';T091';T092';T101';T102';T111';T112';T121';T122';MANGA011';MANGA012';MANGA021';MANGA022';MANGA031';MANGA032';MANGA041';MANGA042';MANGA051';MANGA052';MANGA061';MANGA062';MANGA071';MANGA072';MANGA081';MANGA082';MANGA091';MANGA092';MANGA101';MANGA102';MANGA111';MANGA112';MANGA121';MANGA122';MANGA131';MANGA132';MANGA141';MANGA142'];

3. Criar um vetor correspondente aos valores de número de ondas utilizando da seguinte forma:

>>**abs**=[4000:1:7800];

4. Transformar os espectros dados em %R para log(1/R) e construir o gráfico com o conjunto de espectros (Figura 1):

>>**X**t=log10(1./(**X**./100));

>>abs=[4000:1:7800];

>>plot(abs,**X**t)

>>xlabel('cm-1');

>>ylabel('log 1/R');

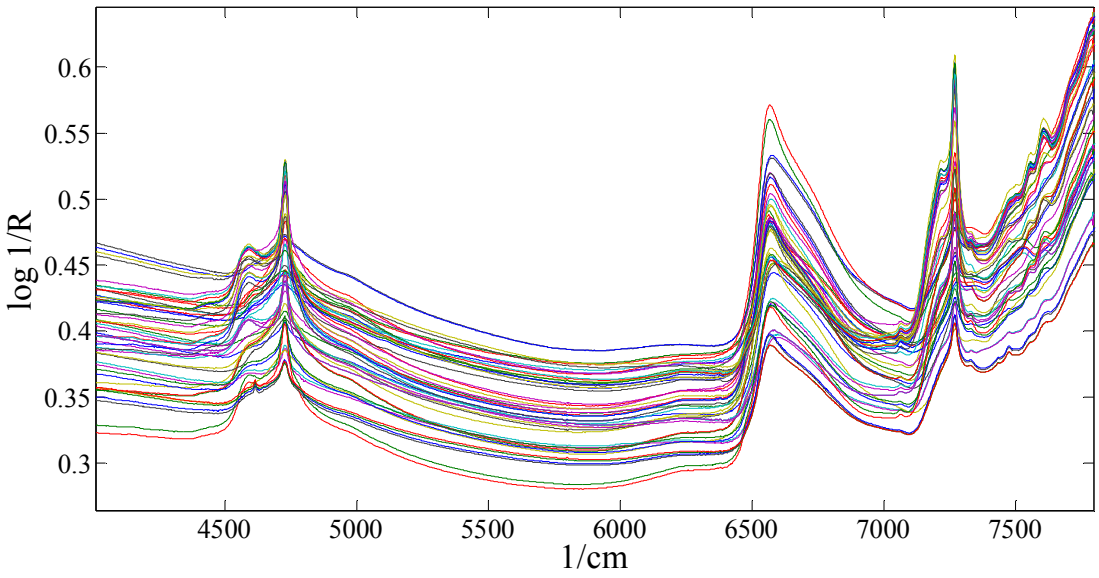

Figura 1: Faixa espectral da região do NIR.

5. Preprocessar os espectros por MSC (Figura 2)

>>[**X**tmsc]=msc(**X**t);

>>plot(abs,**X**tmsc)

>>xlabel('cm-1');

>>ylabel('log 1/R');

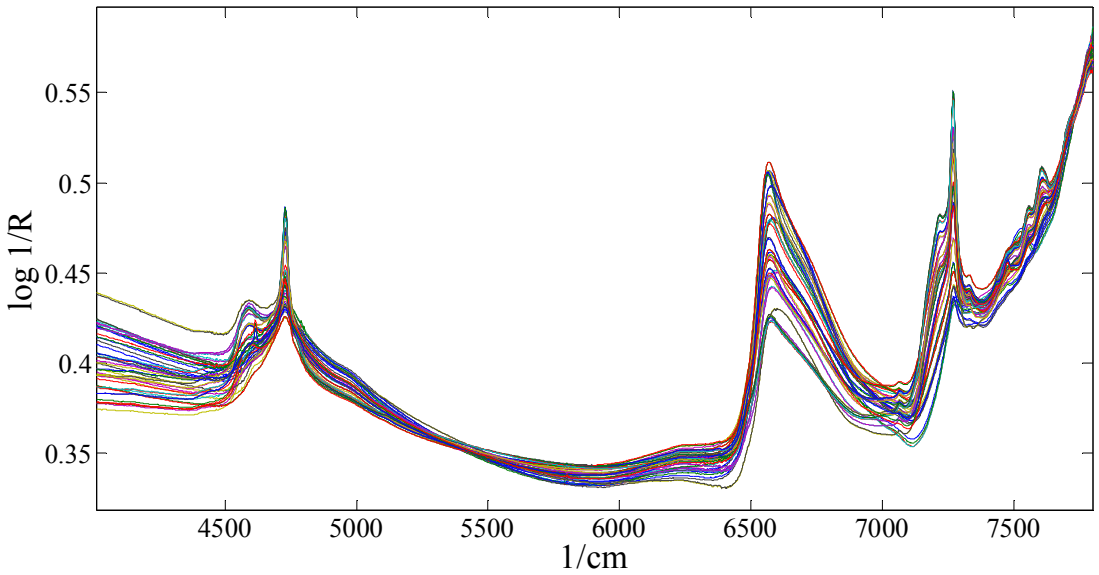

Figura 2: Espectros pré-processados por MSC.

6. Preprocessar os espectros pela 1ª Derivada (Figura 3)

>>[**X**td]=deriv(**X**t,1,15,2);

>>plot(abs,**X**td)

>>xlabel('cm-1');

>>ylabel('1ª Derivada');

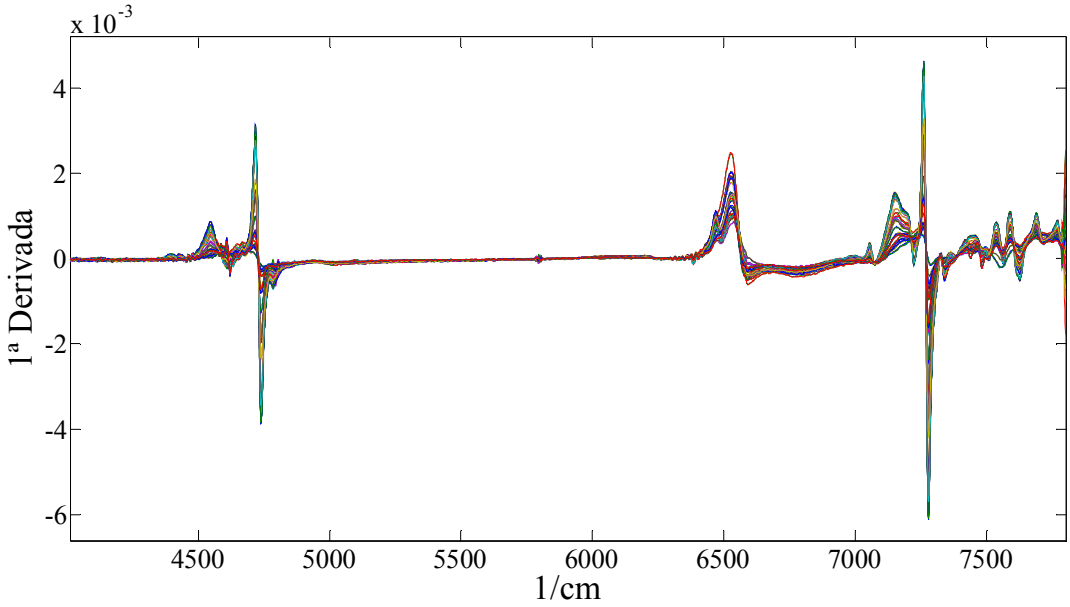

Figura 3: Espectros pré-processados pela 1ª derivada.

7. Preprocessar os espectros pela 2ª Derivada (Figura 4)

>>[**X**tdd]=deriv(**X**t,2,15,2);

>>plot(abs,**X**tdd)

>>xlabel('cm-1');

>>ylabel('2ª Derivada');

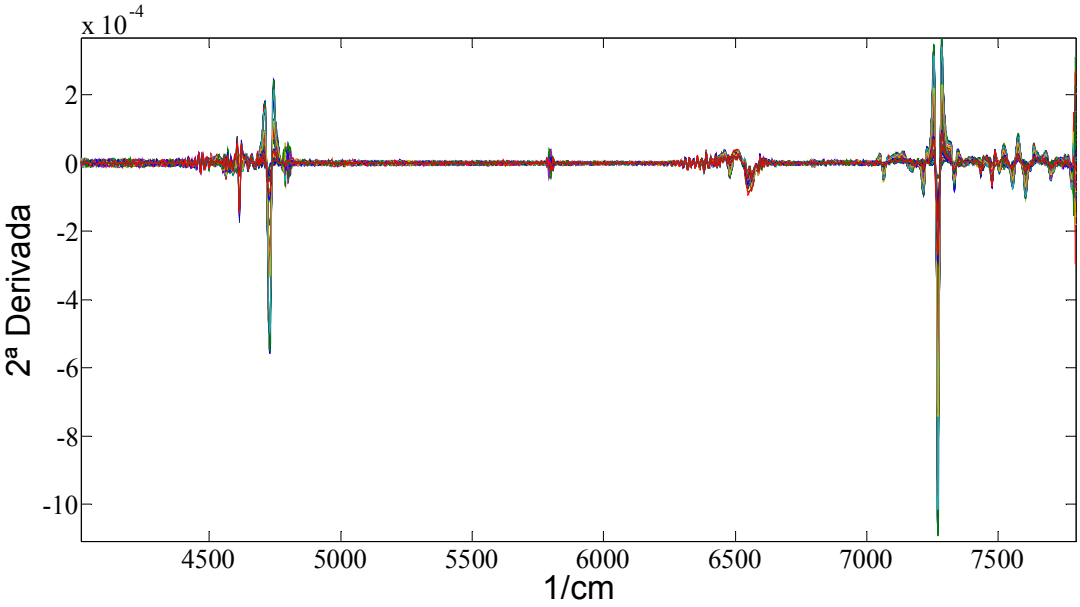

Figura 4: Espectros pré-processados pela 2ª derivada.

8. Centrar na média depois de aplicado o MSC

>>**X**tmscm=mean(**X**tmsc);

>>for i=1:52

>>**X**tmscmmsc(i,:)=**X**tmsc(i,:)-**X**tmscm;

>>end

9. PCA usando SVD - dados preprocessados por MSC e centrados na média

O SVD (Equação 6) é baseado no teorema da álgebra linear que afirma que uma matriz **X** (mxn), m colunas e n linhas, pode ser transformada em um produto de três matrizes **U**, **S**, $\mathbf{V}^T$ (T subscrito significa transposta) e que têm propriedades específicas: (1) as matrizes **U** e **V** são quadradas e ortonormais (2), a matriz **S** é uma matriz retangular diagonal contendo os valores singulares na diagonal e todos os elementos fora da diagonal iguais a zero (Geladi e Kowalski, 1986). Nesse caso, os pesos são dados pela matriz **V** e os escores por: **T**=**US**.

$$\boldsymbol{X = USV^T} \qquad \text{Equação 6}$$

>>[**u**,**s**,**v**]=svd(**X**tmscmmsc), onde o gráfico dos escores (PC1xPC2) é mostrado na Figura 5, o gráfico dos pesos na PC1 na Figura 6, e na Figura 7 é mostrados o gráfico dos pesos na PC2;

10. Cálculo da porcentagem de variância explicada

>>percent = 100*s/sum(s)

>>Gráfico dos scores

>>y=**u***s(:,1:10);

>>plot(y(1:24,1),y(1:24,2),'xg',y(25:48,1),y(25:48,2),'+r')

>>grid

>>for i=1:52;

>>text(y(i,1),y(i,2),num2str(i))

>>end

>>xlabel('PC1 (61.3813%)');

>>ylabel('PC2(34.1874%)');

>>title('MSC + Centrado na média + SVD')

>>legend('Alto teor de Ca','Baixo teor de Ca');

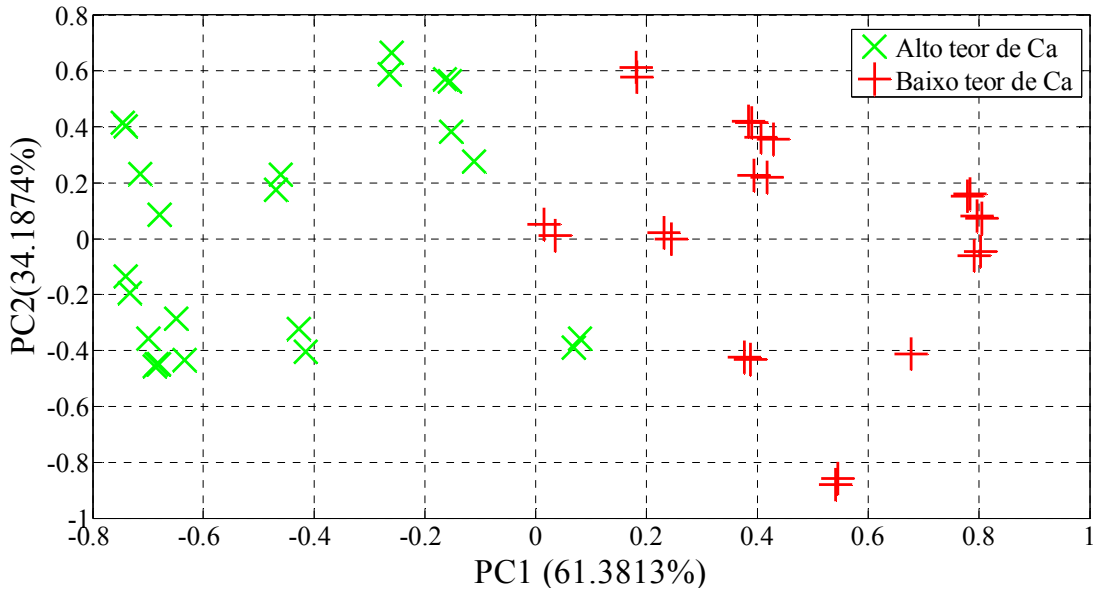

Figura 5: Gráfico de escores da PC1xPC2

>>Gráfico dos loadings

>>plot(abs,**v**(:,1))

>>xlabel('1/cm)');

>>ylabel('PC1(61.3813%)');

>>title('MSC + Centrado na média + SVD')

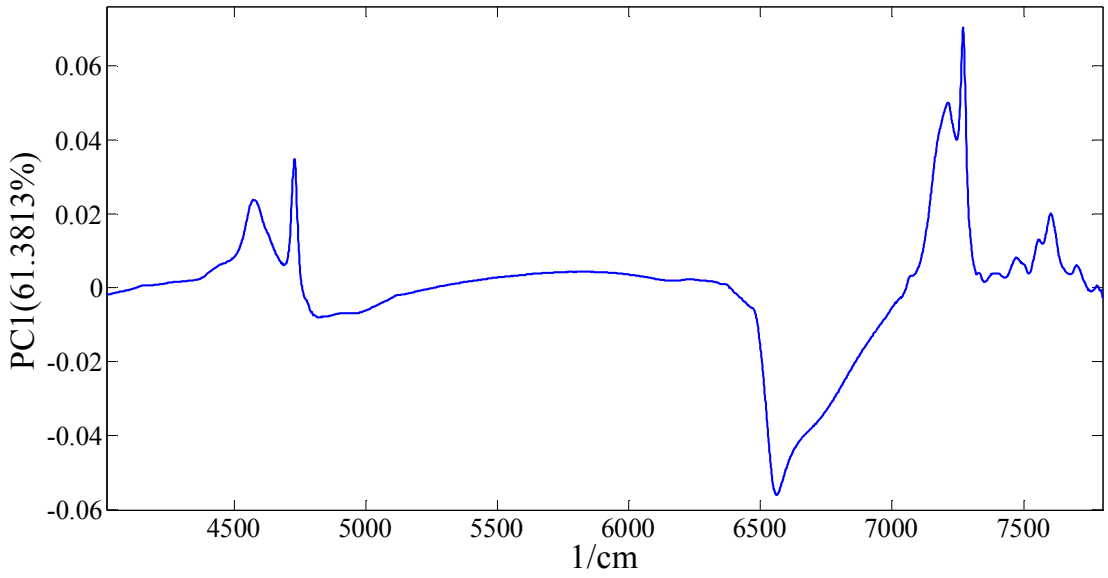

Figura 6: Gráfico de pesos da PC1.

>>figure(7)

```
>>plot(abs,v(:,2))
>>xlabel('1/cm)');
>>ylabel('PC2(34.1874%)');
>>title('MSC + Centrado na média + SVD')
```

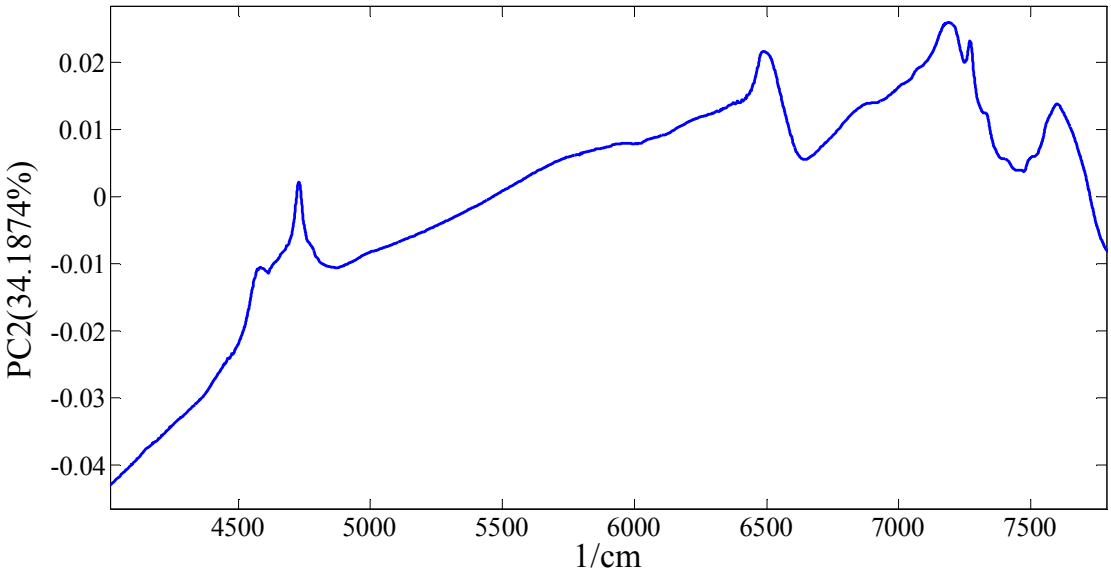

Figura 7: Gráfico de pesos da PC2.

A porcentagem de variância explicada pelas cinco primeiras PC é, respectivamente, 61,3813 %, 34,1874 %, 2,4979 %, 1,1737 % e 0,3222 %. O gráfico de escores da PC1 x PC2 (Figura 5) evidenciou a formação de dois agrupamentos distintos entre as amostras de solos analisadas. As duas primeiras PC, juntas, explicaram mais que 95 % da variância explicada, separando eficientemente as de maior teor de Ca^{2+} daquelas de menor. Por sua vez, os gráficos dos pesos mostrados nas Figuras 6 e 7, respectivamente, apontam quais são as variáveis, em número de ondas, mais importantes para o modelo de PCA descrito. Observando-se a Figura 6 verifica-se que as variáveis com número de onda próximo de 6500 cm-1 e 7300 são as que mais contribuem para a PC1. Esta informação é utilizada, por exemplo, para selecionar as variáveis mais relevantes a fim de melhorar a qualidade do modelo.

Estudos dessa natureza possibilitam, dentre outros usos, avaliar a eficiência dos sistemas taxonômicos de solos, como o Sistema Brasileiro de Classificação de Solos (SiBCS; Embrapa, 2006), que está em permanente atualização e ainda carece de validação de seus conceitos e critérios taxonômicos. Por exemplo, se amostras de solos forem agrupadas, tal como observado na Figura 5, e tais amostras reconhecidamente pertençam à mesma classe de solos segundo o SiBCS, há fortes evidências de que sejam similares e homogêneas. Portanto, a espectroscopia associada à quimiometria podem indicar que os solos foram adequadamente classificadas no SiBCS, validando-o.

CONCLUSÕES

Este tutorial pode ser utilizado como um guia de experimento porque possibilita aos alunos e professores manter um histórico de todas as operações executadas. Os autores recomendam, para aqueles que não dispõem do Matlab, o uso do software livre Octave, onde todos os comandos executados no Matlab podem ser aplicados diretamente ou adaptados com facilidade. Foram utilizadas rotinas disponíveis na internet para a execução da PCA e de alguns pré-processamentos como normalização e MSC. Entretanto, a partir dos conceitos básicos sobre PCA e a utilização do Matlab para esta finalidade é fortemente recomendado que o aluno ou pesquisador procure formular suas próprias rotinas de trabalho neste software. Nesse sentido, o uso da espectroscopia NIR associada a métodos quimiométricos, tanto para a quantificação de atributos dos solos, ou como ferramenta complementar para a sua classificação taxonômica, é um método promissor que pode ser desenvolvido a partir do aperfeiçoamento do tutorial aqui ensinado.

AGRADECIMENTOS

Os autores agradecem aos órgãos de fomento a pesquisa CAPES, CNPq, FAPERJ e FAPESP.

REFERÊNCIAS

BEEBE, K.R.; PELL, R.J.; SEASHOLTZ, M.B. Chemometrics: a pratical guide. John Wiley and Sons, New York, 1998.

CAÑASVERAS, J.C.; BARRÓN, V.; DEL CAMPILLO, M.C.; TORRENT, J. ; GÓMEZ, J.A. Estimation of aggregate stability indices in Mediterranean soils by diffuse reflectance spectroscopy. Geoderma, 158:78-84, 2010.

DART, R.O.; COELHO, M.R.; MENDONÇA-SANTOS, M.L.; PARES, J.G. ; BERBARA, R.L.L. Digital soil mapping at Parque Estadual da Mata Seca, Minas Gerais state, Brazil: applying Regression Tree to predict soil classes. In: INTERNATIONAL WORKSHOP ON DIGITAL SOIL MAPPING, 4, 2010, Rome. From Digital Soil Mapping to Digital Soil Assessment: identifying key gaps from fields to continents. Rome, **The INTERNATIONAL UNION OF SOIL SCIENCES The Working Group on Digital Soil Mapping (WG-DSM)**. Rome, JRC, 2010. CD-ROM.

DEMATTÊ, J.A.M.; CAMPOS, R.C.; ALVES, M.C.; FIORIO, P.R.; NANNI, M.R. Visible-NIR reflectance: a new approach on soil evaluation. Geoderma, 121: 95-112, 2004.

EMPRESA BRASILEIRA DE PESQUISA AGROCUPECUÁRIA (EMBRAPA). Centro Nacional de Pesquisa em Solos. Sistema Brasileiro de Classificação de Solos. 2. ed. Brasília: EMBRAPA, Produção de informação; Rio de Janeiro: EMBRAPA, Centro Nacional de Pesquisa em Solos, 2006. 306p.

FIDÊNCIO, P.H.; POPPI, R.J.; ANDRADE, J.C. ; CANTARELLA, H. Determination of organic matter in soil using near-infrared spectroscopy and partial least squares regression. Communications in Soil Science and Plant Analysis 33, 1607–1615.

FILGUEIRAS, P.R, SOUZA, A.M., POPPI, R.J.; COELHOS, M.R., PARÉS, J.G. CUNHA, T.A.F. ; DART, R.O. Avaliação de modelos de calibração PLS e SVM na determinação do carbono orgânico do solo por espectroscopia NIR. CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO, 33, 2011, Uberlândia. **Solos nos biomas brasileiros: sustentabilidade e mudanças climáticas.** Uberlândia, Sociedade Brasileira da Ciência do Solo, 2011. CD-ROM.

FONTÁN, J.M.; CALVACHE, S.; LÓPEZ-BELLIDO, R.J.; LÓPEZ-BELLIDO, L. Soil carbon measurement in clods and sieved samples in a Mediterranean Vertisol by visible and near-infrared reflectance spectroscopy. Geoderma, 156:93-98, 2010.

FUENTES, M.; HIDALGO, C.; GONZÁLEZ-MARTÍN, I.; HERNÁNDEZ-HIERRO, J.M.; GOVAERTS, B.; SAYRE, K.D.; ETCHEVERS, J. NIR Spectroscopy: an alternative for soil analysis, communications in soil Science and plant analysis, 43:1-2, 346-356.

IGLESIAS, M.; UHLEIN, A. Estratigrafia do Grupo Bambuí e coberturas fanerozóicas no vale do rio São Francisco, norte de Minas Gerais. Revista Brasileira de Geociências, 39(2): 256-266, 2009.

MOUAZEN, A.M.; KUANG, B.; DE BAERDEMAEKER, J.; RAMON, H. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. Geoderma, 158:23-31, 2010.

NOCITA, M.; KOOISTRA, L.; BACHMANN, M.; MÜLLER, A.; POWELL, M. ; WEEL, S. Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa. Geoderma, 167-168: 295-302, 2011.

PASQUINI, C. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. J. Braz. Chem. Soc., Vol.14, No. 2, 198-219, 2003.

ROSSEL, R.A.V. ; BEHRENS, T. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158:46-54, 2010.

SOUZA, A.M.; POPPI, R.J. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: um tutorial, parte I. Quim. Nova, Vol. 35, No 1, 223-229, 2012.

SOUZA, A.M.; COELHO, M.R.; NOVOTNY, E.H.; POPPI, R.J.; DART, R.O.; SANTOS, M.L.M.; BERBARA, R.L.L. Determinação do carbono orgânico do solo por espectroscopia de infravermelho próximo e regressão por quadrados mínimos parciais. In: CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO, 33, 2011, Uberlândia. **Solos nos biomas brasileiros: sustentabilidade e mudanças climáticas.** Uberlândia, Sociedade Brasileira da Ciência do Solo, 2011. CD-ROM.

VOHLAND, M.; BESOLD, J.; HILL, J.; FRÜND, H.C. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. Geoderma, 166: 198-205, 2011.