PREDIÇÃO DO CARBONO ORGÂNICO DO SOLO UTILIZANDO ESPECTROSCOPIA Vis-NIR, PLSR E REGRESSÃO *STEPWISE*

Franciele Romagnoli[1], Marcos Rafael Nanni[2], Carlos Antonio da Silva Junior[1], Anderson Antonio da Silva[1], Everson Cezar[2], Aline de Carvalho Gasparotto[1]

[1]Doutorando(a) em Agronomia, Universidade Estadual de Maringá (UEM), Departamento de Agronomia, Maringá, PR, Brasil. E-mail: romagnoli.fran@gmail.com.
[2]Professor Doutor, Universidade Estadual de Maringá (UEM), Departamento de Agronomia, Maringá, PR, Brasil. E-mail: mrnanni@uem.br.

Recebido em: 03/01/2014 – Aprovado em: 04/04/2014 – Publicado em: 12/04/2014

RESUMO

Este estudo teve como objetivo avaliar dois métodos de análise multivariada: O métotodo PLSR (ParLeS) e a regressão linear múltipla com seleção de variáveis *Stepwise* (SAS), para predição do carbono orgânico do solo por meio de dados de reflectância utilizando um espetrorradiometro de laboratório. Foram coletadas 156 amostras no noroeste do Estado do Paraná e estas foram encaminhadas ao laboratório de solos da Universidade Estadual de Máringá e submetidas à análise química de rotina. As leituras espectrais foram realizadas em um espectrorradiômetro na faixa de 350-2500 nm (Vis-NIR), permitindo a construção de um modelo de previsão para o carbono orgânico do solo por meio de sua resposta espectral. Os dois métodos multivariados apresentaram resultados similares para a fase de calibração do modelo de previsão. Na fase de validação, a técnica PLSR superou a regressão linear com seleção de variáveis *Stepwise*.
PALAVRAS-CHAVE: Modelo, química do solo, sensoriamento remoto.

PREDICTION USING SOIL ORGANIC CARBON VIS-NIR SPECTROSCOPY, PLSR AND STEPWISE REGRESSION

ABSTRACT

This study aimed to evaluate two methods of multivariate analysis: The method PLSR (ParLeS) and multiple linear regression with stepwise selection (SAS) variables for predicting soil organic carbon using reflectance data using a spectroradiometer laboratory. 156 samples were collected in northwestern Paraná State and these were sent to the soil laboratory at the State University of Maringá and submitted for chemical analysis routine. The spectral measurements were performed on a spectroradiometer in the range of 350-2500 nm (Vis-NIR), allowing the construction of a prediction model for soil organic carbon through its spectral response. The two multivariate methods showed similar results for the calibration phase of the forecasting model. In the validation phase, the PLSR technique outperformed linear regression with stepwise regression
KEYWORDS: Remote sensing, soil chemistry, models.

INTRODUÇÃO

O teor de Carbono Orgânico (C.O.) está diretamente relacionado com a qualidade do solo, tornando-se um condicionante do mesmo por atuar em mecanismos que permitem a manutenção da sua capacidade produtiva e sua conservação. Porém, a

análise laboratorial do C.O. no solo é difícil e demorada, além de gerar resíduos químicos que podem causar sérios problemas para o meio ambiente. Por isso novas tecnologias devem ser estudadas e desenvolvidas a fim de auxiliar ou até mesmo substituir a análise convencional de laboratório.

Dentre as técnicas de Sensoriamento Remoto aplicadas na agricultura, a espectroscopia de reflectância do Visível e do Infravermelho Próximo (Vis-NIR), vem sendo muito utilizada para a predição dos atributos do solo. Esta ferramenta vem apresentando-se como uma técnica atrativa para determinados campos donde se requer análises rápidas com baixo custo. Além disso, outra característica interessante desta técnica é a sua multiplicidade, já que, uma vez que se obtém o espectro, é possível a partir deste, estimar vários parâmetros de uma só vez (BEN DOR & BANIN, 1995, SHEPHERD & WALSH, 2002).

Para a utilização dos dados espectrais na quantificação dos atributos dos solos, necessitam-se de técnicas estatísticas para discriminar a resposta desses atributos a partir de suas características espectrais (VISCARRA ROSSEL et al., 2006). Para isto, tem se utilizado de análises multivariadas como PLSR *(Partial Least Square Regression*), PCR (*Principal Component Regression*), MARS (*Multivariate Adaptive Regression Splines*), MLR (*Multiple Linear Regression*), dentre outras.

As técnicas PLSR e a regressão linear múltipla com seleção de variáveis *Stepwise* vem sendo muito utilizadas por diversos pesquisadores. A regressão PLSR vem se mostrando ser uma ótima ferramenta para predição dos atributos do solo por meio dos dados espectrais, onde as variáveis preditoras são altamente colineares (BELMONTE, 2006). A técnica de seleção de variáveis *Stepwise,* realizada por meio da regressão linear múltipla, seleciona as variáveis preditoras em detrimento do coeficiente de determinação entre as variáveis dependentes e independentes. Por serem duas das técnicas de análise multivariadas mais utilizadas para predição dos atributos do solo, torna-se importante a análise comparativa entre elas, já que não se encontra, na literatura, trabalhos comparando-as.

Diante disto, o objetivo deste trabalho, foi de estimar o C.O. do solo por meio de seus dados espectrais e modelagens PLSR e regressão linear múltipla juntamente com o método de seleção de variáveis *Stepwise e* verificar seus desempenhos no ajuste e validação dos modelos de predição, bem como saber qual método multivariado oferece melhores resultados, quando comparado com as análises convencionais conduzidas em laboratório.

MATERIAL E MÉTODOS

As coletas de solo foram realizadas na região noroeste do Estado do Paraná. Essas amostras foram coletadas com um trado tipo holandês, acondicionadas em sacos plásticos e posteriormente encaminhadas ao laboratório de solos da Universidade Estadual de Maringá para determinação do C.O., conforme EMBRAPA (1997).

As leituras espectrais foram realizadas no laboratório de Geoprocessamento e Sensoriamento Remoto Aplicado ao Meio Ambiente, do Departamento de Agronomia da Universidade Estadual de Maringá. O equipamento utilizado foi um espectrorradiômetro FieldSpec 3, com intervalo espectral de 350 – 2500 nm (Vis-NIR) e resolução espectral de 3nm até 700 e de 30 nm de 700 a 2500nm, presente em ambiente controlado de umidade, luminosidade e temperatura.

O sensor ótico presente na ponta da fibra foi colocado em posição vertical a 8 cm de distância da amostra, sendo medida a luz refletida numa área aproximada de 2 cm² no centro da amostra. A fonte de iluminação utilizada foi uma lâmpada halógena de 650 W, com feixe não colimado para o plano visado, sendo posicionada a 35 cm da amostra e com um ângulo zenital de 30º.

Foram realizadas, para cada amostra, três leituras com o sensor girando-se a placa de petri a 120° entre as leituras, obtendo melhor varr edura da amostra, como realizado por NANNI, (2000). Para cada uma das três leituras, o equipamento fez 50 leituras, gerando a curva espectral média de cada amostra.

Para ajustar um modelo de calibração e posterior validação para predição do C.O., foram utilizados os Softwares, ParLeS 3.1, proposto por VISCARRA ROSSEL, (2008) e o Statistical Analysis System (SAS) (SAS, 2001).

A primeira técnica multivariada utilizada para construção dos modelos de predição foi a PLSR, por meio da validação cruzada. A PLSR é uma técnica de análise multivariada, muito utilizada em análises de predição por meio dos dados espectrais, no qual as variáveis preditoras são altamente colineares (BELMONTE, 2006). Durante a etapa de calibração, a modelagem PLSR utiliza as informações da matriz de dados X e matriz de concentração Y, obtendo-se novas variáveis denominadas, variáveis latentes (fatores PLS).

O segundo método utilizado para estimativa dos atributos do solo por meio dos dados espectrais, foi a regressão linear múltipla juntamente com o método de seleção de variáveis *Stepwise,* realizado pelo programa estatístico SAS. Todos os comprimentos de onda do espectro Vis-NIR (350-2500 nm) foram inseridos na rotina para uma possível seleção *Stepwise.*

Para a fase de calibração dos modelos de predição por meio do PLSR, foram selecionadas aleatoriamente 2/3 das curvas espectrais junto com seus respectivos valores determinados em laboratório, restando, portanto, 1/3 das amostras para validação do modelo. Proporções semelhantes de divisão entre amostras de calibração e validação foram utilizadas por REEVES et al. (1999), MCCARTY et al (2002), SHEPHERD & WALSH (2002) e ISLAM et al. (2003).

De acordo com VISCARRA ROSSEL et al., (2006), ao ajustar um modelo utilizando-se PLSR, pretende-se encontrar o menor número possível de fatores PLS necessários para explicar a maior parte da variação entre as variáveis. Durante a validação cruzada (*cross validation*), *n-1* amostras foram utilizadas para a calibração do modelo de predição. A validação cruzada foi realizada a fim de se obter menores números de fatores PLS que gerassem modelos com maiores R^2 e RPD e menores valores de RMSE, parâmetros estes que são indicadores da qualidade dos modelos.

A calibração obtida por meio da regressão *Stepwise* realizada pelo SAS (SAS, 2001), estabelece, passo a passo, o modelo cujas variáveis dependentes tenham maior coeficiente de determinação com as variáveis independentes. O sistema apresenta equações para cada atributo do solo selecionado na análise, determinando a correlação deste em cada comprimento de onda. Assim, cada comprimento terá um fator na equação, de acordo com a contribuição do elemento nessa faixa. As faixas onde o elemento não obtiver correlação, não aparecerão na equação. A avaliação qualitativa dos modelos de predição obtidos pela regressão Stepwise foi realizada por meio do R^2 e RMSE dos modelos.

De acordo com SAYES et al. (2005), valores de R^2 entre 0,50 e 0,65 indicam a possibilidade de discriminação de altas e baixas concentrações no modelo, enquanto que valores de R^2 de 0,66 a 0,80 indicam modelos aceitáveis, de 0,81 a 0,90 indicam modelos bons, e por fim, valores maiores que 0,9 indicam excelentes modelos de predição.

De acordo com DUNN et al. (2002) e CHANG et al. (2001), valores de RPD acima de 2,0 já podem ser considerados como modelos excelentes, de 1,4 à 2,0 modelos aceitáveis e menor que 1,4 modelos não confiáveis. O uso do R^2 juntamente com o RPD, como relatado por WILLIAMS (2001), são os indicadores mais importantes para avaliação da qualidade das análises por meio da Vis-NIR.

RESULTADOS E DISCUSSÃO

Por meio da Tabela 1, pôde-se analisar que o modelo de calibração do C.O. indicou apenas a possibilidade de discriminação de altas e baixas concentrações e modelagem não confiável para predição (R^2 = 0,44 e RPD = 1,34) (SAYES et al., 2005).

TABELA 1. Fatores PLS, R^2, RMSE e RPD dos modelos para o C.O.

Atributo	Fatores PLS	R^2	RMSE	RPD
C.O. (g dm^{-3})	8	0,44	2,59	1,34

O modelo ajustado pelo SAS (2001) para predição do C.O., apresentou as variáveis independentes (comprimentos de onda) que foram selecionadas pelo método *Stepwise* para ajuste do modelo de predição (Tabela 2).

TABELA 2. Equação de regressão múltipla obtida pelo método *Stepwise* para predição do C.O.

Atributo	Equação de calibração[1]	R^2	RMSE
C.O. (g dm^{-3})	C.O = 15,345 + 153,935*c578 – 166,400*c648	0,28	2,99

[1] significativo à 5%, c = comprimento de onda.

Analisando a Tabela 2, para a calibração do modelo de predição do C.O., não foi possível ajustar um modelo satisfatório (R^2 = 0,28) (SAYES et al., 2005).

Para comparar a habilidade dos dois métodos utilizados na construção dos modelos de predição, foi realizada a análise do R^2 e RMSE dos mesmos. Para o C.O., o Parles ajustou um modelo, cujo R^2 e RMSE foi maior e menor respectivamente comparados com o SAS.

Mas, conforme a Tabela 3, o test t possibilitou a análise entre as médias das variâncias dos valores estimados pelos dois métodos, onde se pôde verificar que, mesmo que a regressão PLSR tenha apresentado coeficiente de determinação superior ao obtido pela regressão *Stepwise*, estes não foram suficientemente maiores para diferirem sobre as médias de variância dos valores estimados pelos dois métodos estatísticos, no processo de construção dos modelos de predição.

TABELA 3. Teste t[1] aplicado aos valores estimados pelos dois métodos estatísticos

Método multivariado	C.O.[1]
SAS	6,04 a
Parles	6,03 a

[1] Letras iguais significa que os dados não diferem entre si e letras diferentes significa a diferença entre as médias de variâncias (p<0,05).

A faixa espectral do visível (400-700 nm) foi muito importante para a obtenção do modelo de predição do C.O., sendo esta responsável pela leitura de cor do solo. A cor do solo é um atributo facilmente determinado, e sua importância se dá pelo fato de que a matéria orgânica e os óxidos de ferro estão associados a ela (POST et al., 1993).

KRISHMAN et al. (1980), ao estudarem a reflectância espectral de solos para identificar comprimentos de onda mais adequados para predizer o conteúdo de matéria orgânica do solo, concluíram que a região do visível proporcionou as melhores informações. Isso também ocorreu neste trabalho, no qual as bandas c578 e c648 foram as mais importantes para predição do C.O.

Depois de obtidos os modelos para predições dos atributos do solo, foram realizadas suas validações, a fim de testar a sua capacidade de previsão. Essa validação

foi realizada utilizando as amostras que não entraram na fase de construção do modelo. Para isto, foram utilizadas 52 amostras correspondendo, portanto, a um terço das amostras totais.

A validação do modelo para predição do C.O. realizado pelo método PLSR, está apresentado na Tabela 4 e a dispersão dos dados determinados em laboratório e preditos pelo modelo na fase de validação, estão apresentados na Figura 1.

TABELA 4. R^2, RMSE e RPD resultantes da validação do modelo de predição

Atributos	R^2	RMSE	RPD
C.O. (g dm^{-3})	0,64	2,41	1,67

Para a validação do modelo de predição do C.O., o R^2 e RPD obtidos (0,64 e 1,67), foram superiores aos encontrados por SUMMERS et al. (2011), que obtiveram, para o modelo, R^2 de 0,57 e RPD de 1,8. No entanto, eles foram inferiores, aos obtidos por ISLAM et al. (2003) que encontraram R^2 e RPD de 0,76.e.1,7.respectivamente.

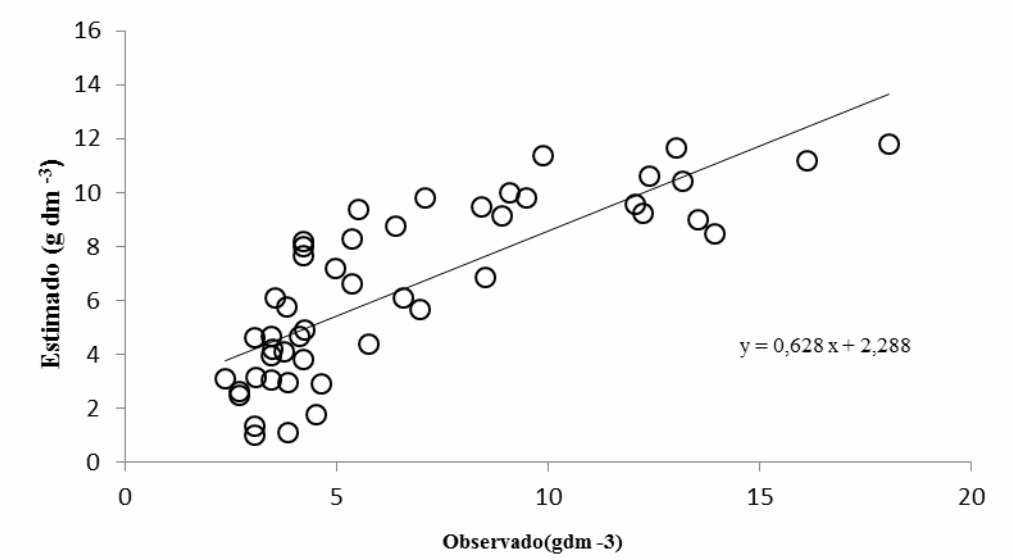

FIGURA 1. Dados de C.O. validados pelo modelo de predição x determinados em laboratório por meio do ParLeS.

Foi realizado o test t, entre os dados observados e determinados em laboratório para avaliar se existe ou não diferença entre as médias de variâncias dos dados, conforme apresenta na Tabela 5. Utilizando o procedimento t test, foi possível observar que a utilização das equações para predição dos atributos do solo possibilitou a estimativa de valores estatisticamente semelhantes aos determinados em laboratório para o C.O..

TABELA 5. Teste t[1] aplicado às médias dos atributos estimados e observados

Método multivariado	**C.O**
Observado	6,67 a
Estimado	6,47 a

[1] Letras iguais significa que os dados não diferem entre si e letras diferentes significa a diferença entre as médias de variâncias ($p<0,05$).

Analisando a Tabela 6, pôde-se verificar que o modelo de predição do C.O. (R^2 = 0,43), não apresentou bom desempenho nesta fase de validação, o que já era esperado, devido ao mau desempenho na calibração do modelo (R^2 = 0,28).

TABELA 6. R^2 referente a validação do modelo de calibração realizada pelo procedimento proc reg – stepwise do Sas

Atributos	**R^2 da validação**
C.O. (g dm^{-3})	0,43

A dispersão dos dados determinados em laboratório e validados pelo modelo de predição está apresentada na Figura 2.

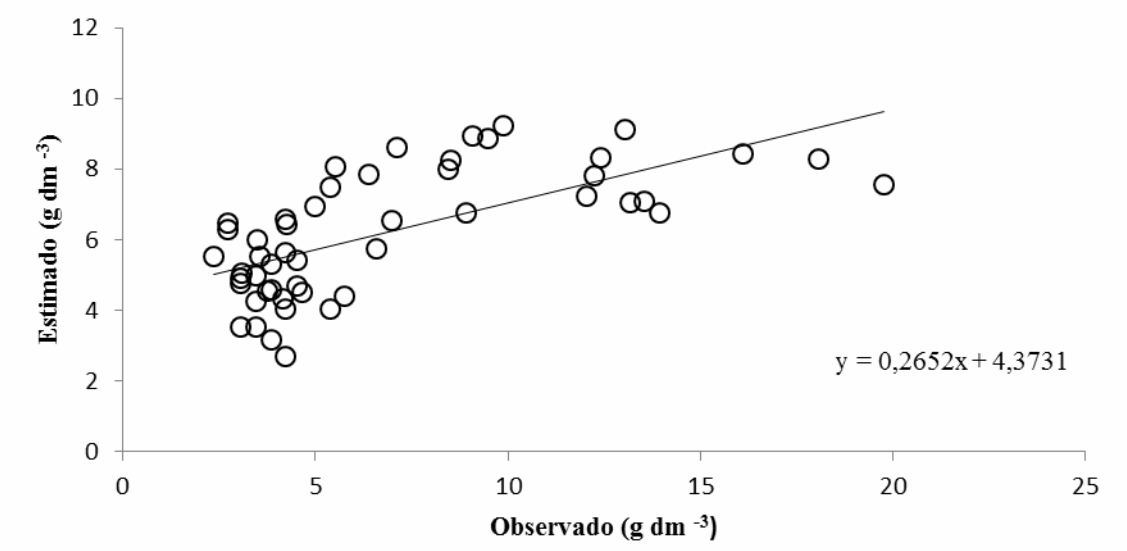

FIGURA 2. Dados de C.O. validados pelo modelo de predição x determinados em laboratório por meio do SAS.

A validação do modelo apresentou discrepância quando comparado com a fase de calibração (que obteve menor R^2). De acordo com MORGANO (2005), discrepâncias entre as correlações obtidas para a calibração e validação sugerem que a calibração pode ter sido ajustada com ruídos nos dados espectrais. Como não foi realizado nenhum tipo de tratamento nos espectros antes da construção dos modelos de predição, como no Parles, isso pode ter ocorrido com a modelagem obtida pelo SAS.

A validação do modelo realizado pelo Parles, em comparação com a validação realizada pelo SAS, (RMSE e o R^2) apresentou melhor desempenho.

Melhor validação foi obtida quando utilizado o programa ParLes (por meio da análise do RMSE e o R^2) . A técnica PLSR se mostrou melhor preditora na fase de validação quando comparada com o procedimento *proc reg* e função s*tepwise* realizada pelo SAS. De acordo com BILGILI et al. (2010), a técnica multivariada PLSR, é o método mais utilizado na calibração e validação da predição de atributos do solo, utilizando sua reflectância, devido sua superioridade sobre os métodos convencionais em lidar com a multicolinearidade dos dados.

Para verificar se as médias das variáveis respostas estimadas diferem das médias determinadas em laboratório, foi realizado o teste t. A Tabela 7 apresenta o teste t da média dos dados a um nível de 5% de significância.

TABELA 7. Teste t[1] aplicado à média do C.O. estimado e observado realizado pelo SAS

Método multivariado	**C.O.**
Observado	6,89 a
Estimado	6,20 b

[1] Letras iguais significa que os dados não diferem entre si e letras diferentes significa a diferença entre as médias de variâncias (p<0,05).

Analisando a Tabela 7, pode-se verificar que a utilização da equação para predição do C.O do solo, não possibilitou a estimativa de valores estatisticamente semelhantes aos determinados em laboratório, mostrando-se estatisticamente diferentes (a um nível de 5% de significância).

CONCLUSÕES

Foi possível determinar modelos de predição para o C.O. do solo por meio de sua resposta espectral para os dois métodos utilizados: Regressão por mínimos quadrados parciais (PLSR) e a regressão linear múltipla com seleção de variáveis *Stepwise.*

Na fase de obtenção do modelo de predição (calibração), os dois métodos analisados apresentaram desempenhos semelhantes. Na fase de validação dos modelos de predição, a técnica PLSR superou a regressão *Stepwise*.

REFERÊNCIAS

BELMONTE, R. Z. **Evaluación de la calidad ambiental en suelos de la provincia de Alicante desarrollo y aplicación de diferentes técnicas**. Tese (Doutorado). Elche. 2006.

BEN-DOR, E., BANIN, A., Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. **Soil Science Society of America Journal,** v.59, n.2, p.364–372, 1995.

BILGILI, A. V., Van ES, H. M., AKBAS, F., DURAK, A., HIVELY, W. D. Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey, **Journal of Arid Environments**, v.74, n.2, p.229-238, 2010.

CHANG, C., LAIRD, D. A., MAUSBACH, M. J., HURBURG JUNIOR, C. R. Near infrared reflectance spectroscopy – principal components regression analyses of soil properties. **Soil Science Society of American Journal**, v.25, n.2, p. 480-490, 2001.

DUNN, B.W., BEECHER, H.G., BATTEN, G.D., CIAVARELLA, S. The potential of near-infrared reflectance spectroscopy for soil analysis — a case study from the Riverine Plain of south-eastern Australia. **Australian Journal of Experimental Agriculture**, v.42, n.2, p.607-614, 2002.

EMBRAPA. **Manual de métodos de análises de solo**. 2.ed. Rio de Janeiro: Ministério da Agricultura e do Abastecimento, 1997, 212 p.

ISLAM, K., SINGH, B., MCBRATNEY, A.B., Simultaneous estimation of various soil properties by ultra-violet, visible and near-infrared reflectance spectroscopy. **Australian Journal of Soil Research,** v.41, n.6. p.1101– 1114, 2003.

KRISHNAN, P. et al. Reflectance technique for predicting soil organic matter. **Soil Science Society of America Journal**, v.44, n. 6. p.1282-1285, 1980.

MCCARTY, G.W., REEVES III, J.B., REEVES, V.B., FOLLETT, R.F., KIMBLE, J.M., Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurements. **Soil Science Society of America Journal,** v. 66, n. 2, p. 640– 646. 2002.

MORGANO, M. A. Aplicação do método Quimiométrico em análise de alimentos. **Tese** (Doutorado). Universidade de Campinas. Campinas. 2005.

REEVES, J.B., MCCARTY, G.W., MEISINGER, J.J., Near infrared reflectance spectroscopy for the analysis of agricultural soils. **Journal of Near Infrared Spectroscopy**, v.7, p.179– 193, 1999.

NANNI, M. R. Dados radiométricos obtidos em laboratório e no nível orbital na caracterização e mapeamento dos solos. **Tese (Doutorado)** – Escola Superior de Agricultura “Luiz de Queiroz”. Piracicaba, 2000.

POST, D.F. et al. **Soil Color**. Correlations between field and laboratory measurements of soil color., v. 31, p.35-49, 1993.

SAS - INSTITUTE. **SAS, software: user's guide, version 8.2**, Cary, 2001.

SAYES, W., MOUAZEN, A.M., RAMON, H. Potencial for onsite and online analysis of pig manire using visible and near infrared reflectance spectroscopy. **Biosystems Engineering,** v.91, n.4, p.393-402, 2005.

SUMMERS, D., LEWIS, M.; OSTENDORF, B., CHITTLEBOROUGH, D. Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties. **Ecological Indicators**. v.11, n.1, p. 123-131. 2011.

SHEPHERD, K.D., WALSH, M.G., Development of reflectance spectral libraries for characterization of soil properties. **Soil Science Society of America Journal,** v.66, n.3, p.988– 998, 2002.

VISCARRA ROSSEL, R. A. ParLeS: Software for chemometric analysis of spectroscopic data. Chemometrics Intelligent Laboratory. **Systems**, v.90, n.1, p.72-83, 2008.

VISCARRA ROSSEL, R., WALVOOT, D. J. J., McBRATNEY, A. B., JANIC, L. J., SKJEMSTAD, J. O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. **Geoderma**, v.131, n.1-2, p.59-75. 2006.

WILLIAMS, P.C. **Implementation of near infrared technology. In: Near infrared technology in the agricultural and food industries**. American Association of Cereal Chemist, Minnesota, 2001.