Artigo

# DETERMINAÇÃO DE AÇÚCAR TOTAL EM CAFÉ CRU POR ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO E REGRESSÃO POR MÍNIMOS QUADRADOS PARCIAIS

**Marcelo A. Morgano***
Centro de Química de Alimentos e Nutrição Aplicada, Instituto de Tecnologia de Alimentos, CP 139, 13073-001 Campinas – SP, Brasil
**Cristiano Gomes de Faria**
Centro Nacional de Pesquisa Gado de Leite, Empresa Brasileira de Pesquisa Agropecuária, 36038-330, Juiz de Fora – MG, Brasil
**Marco F. Ferrão**
Departamento de Química e Física, Universidade de Santa Cruz do Sul, 96815-000, Santa Cruz do Sul – RS, Brasil
**Márcia M.C. Ferreira**
Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13084-971, Campinas – SP, Brasil

Recebido em 16/1/06; aceito em 25/5/06; publicado na web em 31/10/06

DETERMINATION OF TOTAL SUGAR IN RAW COFFEE USING NEAR INFRARED SPECTROSCOPY AND PLS REGRESSION. In this work a fast method for the determination of the total sugar levels in samples of raw coffee was developed using the near infrared spectroscopy technique and multivariate regression. The sugar levels were initially obtained using gravimety as the reference method. Later on, the regression models were built from the near infrared spectra of the coffee samples. The original spectra were pre-treated according to the Kubelka-Munk transformation and multiplicative signal correction. The proposed analytical method made possible the direct determination of the total sugar levels in the samples with an error lower by 8% with respect to the conventional methodology.

Keywords: total sugar; raw coffee; near infrared spectroscopy.

## INTRODUÇÃO

As técnicas de espectroscopia no infravermelho próximo e médio têm sido freqüentemente utilizadas, como uma ferramenta analítica, na determinação de constituintes em alimentos. Diversos trabalhos demonstram o uso destas técnicas na determinação do conteúdo de açúcar em alimentos[1-5]. Os métodos empregando a espectroscopia no infravermelho são rápidos e exatos e têm ampla aplicação analítica em análises químicas e de controle de qualidade de produtos da áreas de agricultura e alimentos[6-10].

Na determinação de um componente em particular, a espectroscopia no infravermelho necessita de um conjunto de amostras de calibração, uma vez que a resposta instrumental não fornece diretamente a informação desejada. Normalmente a propriedade de interesse é determinada por meio de um método de referência. Neste estudo, o teor de açúcar total nas amostras de café cru foi determinado empregando-se como método de referência o gravimétrico da "Association of Official Analytical Chemists"[11]. A regressão por mínimos quadrados parciais (PLS) vem sendo usada para estabelecer uma relação matemática (modelo de calibração) entre os valores químicos de referência e os dados espectrais para uma determinada propriedade medida[12].

Pré-tratamentos matemáticos aplicados aos dados espectrais são, em geral, necessários para facilitar a interpretação dos espectros e, também, melhorar a qualidade de previsão dos modelos de calibração. Em medidas de reflectância na região do infravermelho é comum ocorrer o fenômeno conhecido como espalhamento de luz, que geralmente é provocado pela falta de homogeneidade ótica das amostras. Para compensar tal problema é utilizada a técnica de correção multiplicativa de sinal (MSC), que visa minimizar a interferência relativa ao espalhamento de luz com base no espalhamento médio de todos os espectros empregados na calibração multivariada[13].

*e-mail: morgano@ital.sp.gov.br

O objetivo deste trabalho foi desenvolver uma metodologia analítica não destrutiva para determinação do teor de açúcar total em amostras de café cru, utilizando-se dos espectros no infravermelho próximo (NIR) e da aplicação do método de regressão por mínimos quadrados parciais (PLS).

## PARTE EXPERIMENTAL

### Amostras

Na construção dos modelos de regressão foram usadas 49 amostras de café cru, da variedade arábica, procedentes de diferentes regiões produtoras de café do Brasil, dos estados de São Paulo, Minas Gerais, Bahia e Paraná. Todas as amostras foram inicialmente homogeneizadas em moinho de facas até a obtenção de partículas de tamanho reduzido e passaram por peneira de 0,84 mm de tamanho de abertura de poro.

### Método de referência

O método de referência utilizado na determinação do teor de açúcar total foi o de Munson-Walker[11], pela medida gravimétrica de óxido cuproso formado. As determinações do teor de açúcar total foram realizadas em duplicata nas amostras. Os teores de açúcar total nas amostras empregadas estão entre 6,46 e 11,24 g/100g.

### Espectroscopia no infravermelho próximo

Os espectros no infravermelho próximo foram coletados em um espectrômetro Bomem DA-08, equipado com um acessório de reflectância difusa (Jasco), sendo os sinais expressos em log (1/R). Foram realizadas três réplicas para cada amostra, com resolução de 4 $cm^{-1}$, 16 varreduras para cada espectro e a região espectral utilizada foi de 4.500 a 10.000 $cm^{-1}$.

**Análise dos dados**

O método de regressão multivariada utilizado no tratamento dos dados foi o método PLS e o software, o Matlab 5.1[14]. Este método é bem conhecido da comunidade científica[13] e, como qualquer outro método de regressão, tem como objetivo encontrar uma relação entre a matriz (**X**) contendo os espectros das amostras de café do conjunto de calibração e o vetor que armazena os respectivos teores de açúcar (**y**). O resultado é uma equação semelhante à Equação 1

$$\mathbf{y = Xb + e} \quad (1)$$

onde **b** é o vetor de regressão e **e** o vetor que representa os erros do modelo. O método PLS é especialmente indicado quando **X** contém variáveis altamente correlacionadas, como no presente trabalho (dados de espectroscopia). Outra vantagem desse método é que pode ser usado mesmo quando as amostras contêm interferentes (que devem estar presentes nas amostras do conjunto de calibração).

No modelo PLS, a matriz **X** é decomposta em escores, **t**, e pesos, **w**, i.e. **XW=T**, onde $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_k)$ é escolhido de maneira que $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_k)$ apresente covariância máxima com **y** (k é o número de variáveis latentes). Assim, as informações espectrais e as concentrações são usadas ao mesmo tempo na fase de calibração. Um fator de suma importância na construção de um modelo PLS é a escolha do número de variáveis latentes (VL), k, a serem incluídas no modelo.

**Estatística**

O conjunto total de espectros foi dividido em dois subconjuntos: um de calibração e outro de validação externa. Na determinação do número de VL utilizadas no modelo, é feita uma validação cruzada (validação interna) no conjunto de calibração: uma amostra do conjunto de calibração é excluída, o modelo é construído e, então, estimado o seu teor de açúcar. O processo é repetido até que todas as amostras sejam previstas para 1, 2, ... variáveis latentes. A habilidade do modelo de calibração para estimar (ou prever) o teor de açúcar baseado nos dados dos espectros NIR gerados foi avaliada usando os erros de previsão e os coeficientes de correlação entre os valores dos teores de açúcar estimados pelo modelo utilizando espectros NIR e os valores do método de referência das amostras do conjunto de calibração e de validação. Os parâmetros de erro empregados estão apresentados nas Equações 2 a 8:
Soma dos quadrados dos erros de previsão (PRESS):

$$\mathrm{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (2)$$

onde *n* representa o número de amostras do conjunto de calibração, $y_i$ é o valor de referência e $\hat{y}_i$ é o valor estimado pelo modelo para a iésima amostra.
Erro quadrático médio (MSE) ou a raiz quadrada do mesmo (RMSE):

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \quad (3)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (4)$$

Como estes erros podem ser medidos tanto para o conjunto de calibração quanto para o de validação externa é comum adicionar no final da sigla destes erros a letra C indicando serem estes relativos à calibração C (MSEC ou RMSEC) ou P quando forem relativos à previsão do conjunto de validação externa (MSEP ou RMSEP). Variância relativa (REV):

$$\mathrm{REV} = 1 - \mathrm{MSE}/s^2 \quad (5)$$

A estimativa da variância total dos dados ($s^2$) para *n* amostras, utilizada na expressão 5, é determinada segundo a expressão:

$$s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1) \quad (6)$$

onde: $\bar{y}$ é o valor médio.
Erro relativo percentual (ER (%)) entre o método de referência e o método desenvolvido (NIR-PLS):

$$\mathrm{ER}\ (\%) = \frac{(y_i - \hat{y}_i)}{y_i} 100 \quad (7)$$

Coeficiente de correlação entre os valores estimados e os valores experimentais do método de referência:

$$r = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\left[\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2\right]^{1/2} \left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right]^{1/2}} \quad (8)$$

**RESULTADOS E DISCUSSÃO**

Para a construção dos modelos de regressão foram utilizados os valores médios dos teores de açúcares totais obtidos na determinação empregando o método de referência e 147 espectros das amostras de café cru. Para formar o conjunto de calibração foram selecionados 81 espectros, correspondentes às 27 amostras, de forma a cobrir toda a faixa do teor de açúcar nas amostras. Os 66 espectros restantes, correspondentes às 22 amostras, foram utilizados para formar o conjunto de validação externa.

Devido aos efeitos de espalhamento de luz provocados pela não-homogeneidade das amostras, decorrentes principalmente de diferença de granulometria, foi necessário utilizar a técnica de correção de espalhamento de luz (MSC) nos espectros transformados pela Equação de Kubelka-Munk[15,16] antes de se proceder à construção dos modelos de regressão. Os resultados das transformações Kubelka-Munk e a correção multiplicativa de sinal (MSC) nos dados dos espectros originais podem ser visualizados na Figura 1.

A matriz de dados resultante, a partir dos espectros transformados, foi escalada pela variância, ou seja, cada coluna foi dividida pelo respectivo desvio padrão.

A seleção de variáveis foi obtida através das informações dadas pelo correlograma[4]. O correlograma foi gerado calculando-se o coeficiente de correlação linear (*r*) entre as variáveis independentes, ou seja, os resultados de log(1/R) (R = reflectância) (bloco-Y) e cada uma das variáveis dependentes, ou seja, os resultados do teor de açúcar obtidos pelo método de referência (bloco-X), de modo que no final do processo se tem para cada variável independente um valor associado de *r*, que ao serem dispostos em um gráfico geram o correlograma. O coeficiente de correlação linear ($r_i$) entre estes dois blocos foi utilizado para compor o correlograma em um gráfico no qual a ordenada é o coeficiente de correlação linear e a

abscissa, os comprimentos de onda lidos (os números representativos de cada variável).

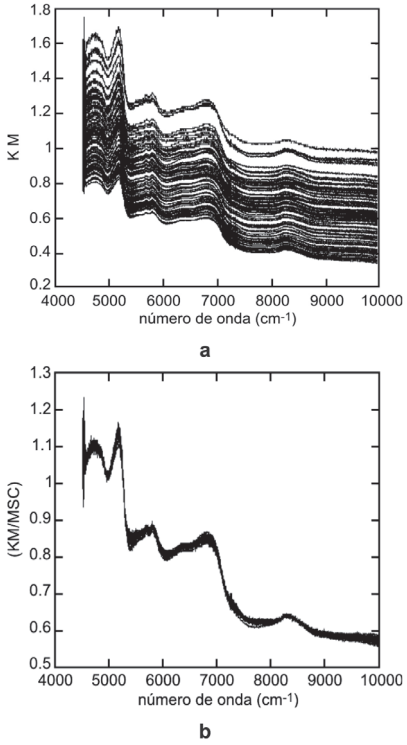


***Figura 1.*** *a. Espectros NIR após a transformação Kubelka-Munk (K M); b. Espectros NIR após K M e a correção multiplicativa de sinal (MSC)*

Para escolher qual o melhor valor de corte a ser empregado, acompanhou-se o valor do PRESS para o conjunto de calibração e de validação externa fazendo-se variar o valor de corte, para modelos PLS usando diferentes variáveis latentes. A Figura 2 apresenta estes resultados. Observa-se que o PRESS de calibração aumenta significativamente, quando se utilizam valores de corte superiores a 0,3. Nesta figura observa-se, também, que o PRESS da calibração diminui, à medida que o modelo se torna mais complexo, no entanto, esta tendência não é observada com relação ao PRESS da validação externa. Como o que se pretende é um modelo com uma boa capacidade de predição optou-se pelo modelo que apresentou o menor PRESS no conjunto de validação (assinalado com asterisco na Figura 2a), que corresponde ao modelo que utiliza 6 variáveis latentes e um valor de corte igual a 0,1.

O correlograma referente aos dados é mostrado na Figura 2b. Observa-se que a correlação linear entre o bloco-X e o bloco-Y não é muito pronunciada, apresentando valores inferiores a 0,6. Com um valor de corte de 0,1 foi possível selecionar 831 variáveis das 2250 variáveis originais. Os intervalos de números de onda selecionados estão relacionados principalmente com as bandas de combinação (estiramento + deformação angular) e sobretons das ligações C-H de grupos $CH_3$ (7353; 5900; 5865; 5797 $cm^{-1}$) e ligação O-H (6757- 6329 $cm^{-1}$) da glicose.

## Ajuste do modelo

O número de variáveis latentes, utilizado no modelo PLS, foi determinado a partir do PRESS exibido no conjunto de validação externa. A Figura 3 mostra o comportamento do PRESS para o conjunto de calibração e de validação externa em função do número de variáveis latentes utilizadas nos modelos. O valor mínimo do PRESS exibido pelo conjunto de validação externa é atingido quando se utiliza um

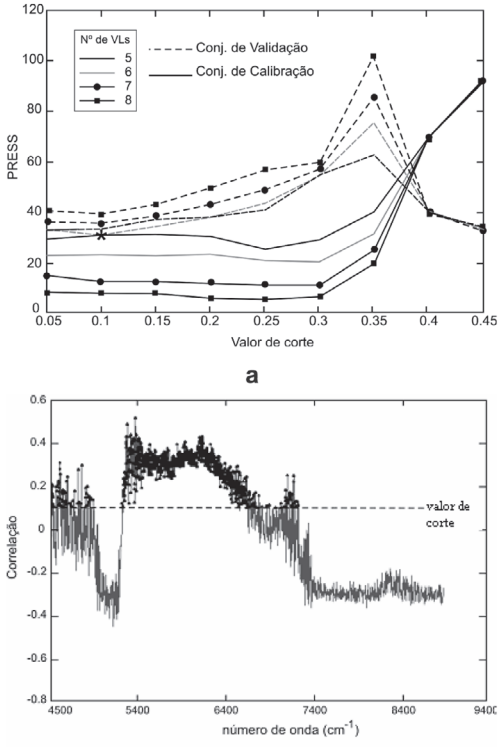


***Figura 2.*** *a. Gráfico do PRESS de validação interna e de validação externa em função do valor de corte do correlograma (selecionando apenas as variáveis positivamente correlacionadas) dos modelos PLS utilizando 5, 6, 7 e 8 variáveis latentes. b. Gráfico do correlograma apresentando o valor de corte utilizado (0,1) e as variáveis selecionadas*

número de variáveis latentes igual a 6. O comportamento do PRESS da validação indica que, se utilizando modelos com mais de 6 variáveis latentes, os dados sofrem um sobre-ajuste ("overfit"), que fica evidenciado também no decréscimo monotônico do PRESS do conjunto de calibração com o aumento da complexidade do modelo. Problemas de sub-ajuste ("underfit") ficam evidenciados no aumento dos valores do PRESS de ambos os conjuntos ao se utilizar modelos com um número de variáveis latentes inferiores a 6. Levando-se em conta os fatores abordados acima, concluiu-se que o uso de 6 variáveis latentes é o mais apropriado na construção do modelo de regressão PLS para determinar o teor de açúcar em amostras de café cru.

Com o modelo PLS utilizando 6 variáveis latentes foi possível descrever praticamente toda a variabilidade tanto do bloco – Y como

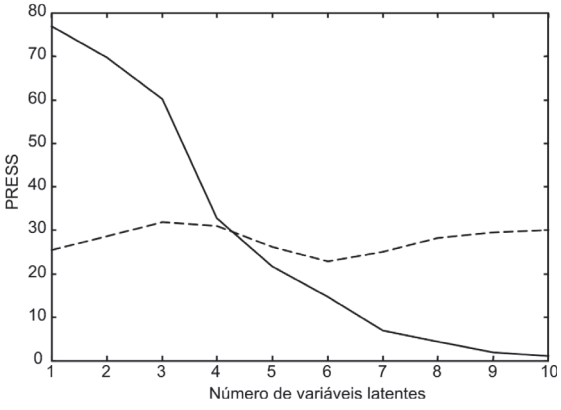


***Figura 3.*** *Gráfico do PRESS dos conjuntos de validação interna (— ) e externa (– –) em função do número de variáveis latentes utilizadas nos modelos*

do bloco – X, utilizados na calibração. A Tabela 1 traz os valores de variância capturados por cada VL; estes valores mostram que a estrutura dominante dos dados, tanto para o bloco-X quanto para o bloco-Y, é descrita basicamente pela primeira VL.

**Tabela 1.** Porcentagem de variância obtida para cada variável latente do modelo PLS

| | Bloco-X | | Bloco-Y | |
|---|---|---|---|---|
| VL* | Variância (%) | Variância Acumulada (%) | Variância (%) | Variância Acumulada (%) |
| 1 | 99,99 | 99,99 | 98,71 | 98,71 |
| 2 | 0,01 | 99,99 | 0,12 | 98,83 |
| 3 | 0,00 | 99,99 | 0,16 | 98,99 |
| 4 | 0,00 | 99,99 | 0,46 | 99,45 |
| 5 | 0,00 | 100 | 0,19 | 99,64 |
| 6 | 0,00 | 100 | 0,12 | 99,76 |

*VL = Variável latente.

Os resultados da calibração podem ser visualizados na Figura 4, que exibe os valores experimentais *versus* os valores preditos pelo modelo PLS com 6 variáveis latentes. Os resíduos da calibração obedecem a uma distribuição normal como se pode ver pelo histograma apresentado na Figura 5. Os parâmetros de erros (g/100 g) e ajuste relativos ao conjunto de calibração encontrados foram os seguintes: PRESS = 16,41; MSE = 0,197; RMSE = 0,444; REV = 0,826 e r = 0,907.

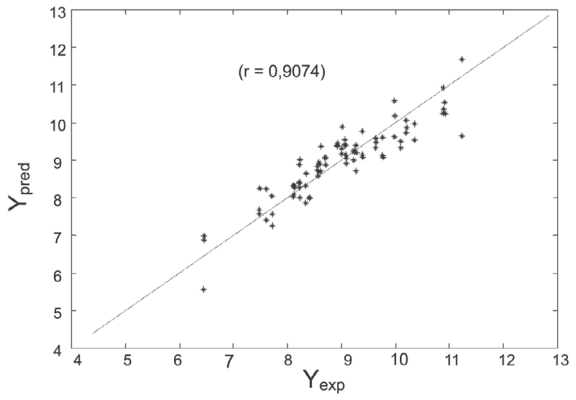


***Figura 4.*** *Valor experimental (Y exp)* versus *valor predito (Y pred) (em g /100 g de açúcares totais) para o conjunto de calibração do modelo de regressão usando 6 variáveis latentes*

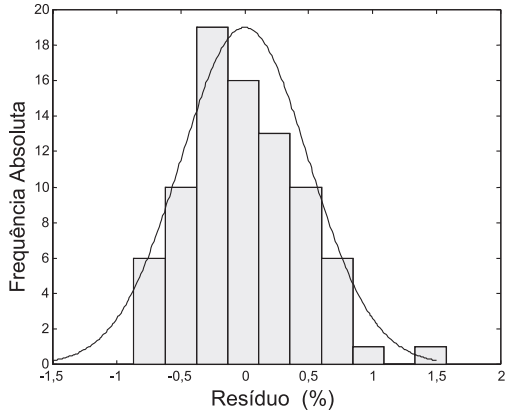


***Figura 5.*** *Histograma dos resíduos da calibração (%) do modelo PLS com 6 variáveis latentes*

**Validação do modelo**

O modelo de calibração para a predição do teor de açúcar total foi validado por validação externa, utilizando um conjunto de 66 espectros, relativos a 22 amostras em triplicata. A Tabela 2 apresenta os valores médios preditos pelo modelo, o resíduo e o erro relativo percentual para as amostras do conjunto de validação externa. O histograma para estes dados, mostrado na Figura 6, indica que o erro relativo percentual para este conjunto fica, na maioria dos casos, confinado entre os valores de ± 5%. O erro máximo obtido para o teor de açúcar total foi 7,7% para as 22 amostras do

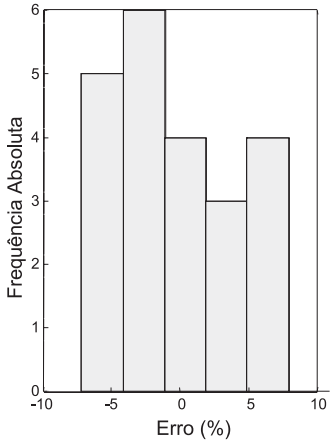


***Figura 6.*** *Representação gráfica da freqüência do erro relativo percentual (ER) para o conjunto de validação externa*

**Tabela 2.** Resultados de predição do teor de açúcares totais para as amostras do conjunto de validação

| Resultados obtidos pelo método referência (g/100 g) | Resultados obtidos pelo método NIR/PLS (g/100 g) | Resíduo (g/100 g) | ER (%) |
|---|---|---|---|
| 8,5 ± 0,1 | 8,6 ± 0,8 | -0,1 | -1,2 |
| 8,9 ± 0,2 | 8,8 ± 0,4 | 0,1 | 1,1 |
| 9,50 ± 0,09 | 9,0 ± 0,6 | 0,5 | 5,3 |
| 9,6 ± 0,4 | 9,0 ± 0,5 | 0,6 | 6,2 |
| 9,0 ± 0,2 | 9 ± 1 | 0 | 0 |
| 8,9 ± 0,2 | 9,3 ± 0,5 | -0,4 | - 4,5 |
| 9 ± 1 | 8,8 ± 0,5 | 0,2 | 2,2 |
| 9,2 ± 0,5 | 9,7 ± 0,2 | -0,5 | - 5,4 |
| 9,6 ± 0,1 | 9,4 ± 0,5 | 0,2 | 2,1 |
| 9,2 ± 0,3 | 9,0 ± 0,3 | 0,2 | 2,2 |
| 9,3 ± 0,3 | 8,9 ± 0,6 | 0,4 | 4,3 |
| 9,2 ± 0,1 | 9,7 ± 0,9 | -0,5 | -5,4 |
| 9 ± 1 | 9,2 ± 0,4 | -0,2 | -2,2 |
| 9 ± 2 | 9,1 ± 0,3 | -0,1 | -1,1 |
| 10,0 ± 0,1 | 9,5 ± 0,1 | 0,5 | 5,0 |
| 10 ± 1 | 9,8 ± 0,4 | 0,2 | 2,0 |
| 8,1 ± 0,2 | 8,7 ± 0,3 | -0,6 | -7,4 |
| 9,4 ± 0,1 | 8,9 ± 0,5 | 0,5 | 5,3 |
| 9,0 ± 0,4 | 9,1 ± 0,7 | -0,1 | -1,1 |
| 9,2 ± 0,7 | 9,3 ± 0,7 | -0,1 | -1,1 |
| 9,4 ± 0,2 | 9,1 ± 0,5 | 0,3 | 3,2 |
| 10,4 ± 0,2 | 9,6 ± 0,7 | 0,8 | 7,7 |

Resíduo: valor obtido pelo método de referência – valor obtido pelo método NIR/PLS; ER (%) = (resíduo/ valor obtido pelo método de referência) x 100.

conjunto de validação, o menor erro foi 0% e a média 3,4%. Os parâmetros de erro obtidos foram: PRESS = 3,354; MSE = 0,152; RMSE = 0,390; r = 0,645.

## CONCLUSÕES

A espectroscopia no infravermelho próximo combinada com o método de calibração multivariada (PLS) é uma técnica fácil e rápida na determinação do teor de açúcar total em amostras de café cru. Os modelos de regressão construídos apresentam melhor desempenho quando são empregadas as transformações Kubelka-Munk e a correção multiplicativa de sinal (MSC) nos dados escalados pela variância. As principais vantagens do método proposto quando comparado com o método tradicional gravimétrico são redução do tempo de análise, pouca manipulação das amostras, diminuição de resíduos químicos e redução do custo de análise.

## AGRADECIMENTOS

Ao Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café e à FUNAPE, pelo apoio financeiro para a realização deste trabalho.

## REFERÊNCIAS

1. Baker, D.; *Cereal Food. World* **1985**, *30*, 389.
2. Baker, D.; Norris, K. H.; *Appl. Spectrosc.* **1985**, *39*, 618.
3. Cadet, F.; Robert, C.; Offmann, B.; *Appl. Spectrosc.* **1997**, *51*, 369.
4. Bellon-Maurel, V.; Vallat, C.; Goffinet, D.; *Appl. Spectrosc.* **1995**, *49*, 556.
5. Kawano, S.; Fujiwara, T.; Iwamoto, M.; *J. Jpn. Soc. Hortic. Scienc.* **1993**, *62*, 465.
6. Osborne, B. G.; Fearn, T.; *Near-infrared spectroscopy in food analysis,* Longman Scientific & Technical: Harlow, 1986.
7. Barton, F. E.; *Near-infrared technology in the agricultural and food industries,* American Association Cereal Chemists: St. Paul, 1987.
8. Parreira, T. F.; Ferreira, M. M. C.; Sales, H. J. S.; Almeida, W. B.; *Appl. Spectrosc.* **2002**, *56*, 1607.
9. Ferrão, M. F.; Carvalho, C. W.; Muller, E. I.; Davanzo, C. U.; *Ciênc. Tecnol. Aliment.* **2004**, *24*, 333.
10. Ferrão, M. F.; Davanzo, C. U.; *Anal. Chim. Acta* **2005**, *540*, 411.
11. Horwitz, W., ed.; *Official Methods of Analysis of the Association of Official Analytical Chemists*, 2000, vol. 9, method 44.1.6.
12. Ferreira, M. M. C.; Antunes, A. M.; Melgo, M. S.; Volpe, P. L. O.; *Quim. Nova* **1999**, *22*, 724.
13. Martens, H.; Naes, T.; *Multivariate calibration***,** J. Wiley & Sons Ltd.: Chichester, 1989.
14. *MATLAB- The Language of Tecnical Computing*, version 5.1.0.421; The MathWorks Inc., 1997.
15. Olinger, J. M.; Griffiths, P. R.; *Appl. Spectrosc.* **1993a**, *47*, 687.
16. Olinger, J. M.; Griffiths, P. R.; *Appl. Spectrosc.* **1993b,** *47*, 695.