Quim. Nova, Vol. 30, No. 5, 1301-1312, 2007

Revisão

TRANSFERÊNCIA DE CALIBRAÇÃO EM MÉTODOS MULTIVARIADOS

Fernanda Araújo Honorato e Benício de Barros Neto
Departamento de Química Fundamental, Universidade Federal de Pernambuco, 50670-901 Recife – PE, Brasil
Marcelo Nascimento Martins e Roberto Kawakami Harrop Galvão
Divisão de Engenharia Eletrônica, Instituto Tecnológico de Aeronáutica, 12228-900 São José dos Campos – SP, Brasil
Maria Fernanda Pimentel*
Departamento de Engenharia Química, Universidade Federal de Pernambuco, Av. Prof. Arthur de Sá, s/n, 50740-521 Recife – PE, Brasil

Recebido em 25/7/06; aceito em 9/11/06; publicado na web em 2/7/07

CALIBRATION TRANSFER IN MULTIVARIATE METHODS. Calibration transfer has received considerable attention in the recent literature. Several standardization methods have been proposed for transferring calibration models between equipments. The goal of this paper is to present a general revision of calibration transfer techniques. Basic concepts will be reviewed, as well as the main advantages and drawbacks of each technique. A case study based on a set of 80 NIR spectra of maize samples recorded on two different instruments is used to illustrate the main calibration transfer techniques (direct standardization, piecewise direct standardization, orthogonal signal correction and robust variable selection).

Keywords: calibration transfer; NIR; standardization

INTRODUÇÃO

A construção e implementação de modelos de calibração multivariada[1] passa por várias etapas, cada uma das quais é determinante para as subseqüentes. A aplicação de um modelo desenvolvido em um determinado equipamento – que vamos chamar de *primário* – a espectros medidos em um outro equipamento (o *secundário*)[1], ou a sua utilização sob novas condições ambientais ou instrumentais, por exemplo, exige muito cuidado. É provável que haja diferenças entre as respostas instrumentais de diferentes equipamentos e, mesmo que o equipamento seja um só, sua resposta pode variar ao longo do tempo. Tudo isto pode levar o modelo original a fornecer previsões errôneas.

Um modelo pode tornar-se inválido caso variações não incluídas na etapa de calibração sejam introduzidas nos novos espectros. Existem basicamente três causas para isto[2]: mudanças na constituição química e/ou física das amostras (viscosidade, tamanho de partículas, textura da superfície etc.); mudanças no ambiente do equipamento (variações de temperatura e umidade, por exemplo, podem provocar deslocamentos das bandas de absorção e mudanças não lineares nas suas intensidades[3]); mudanças na função de resposta do equipamento (que farão com que o espectro de uma dada amostra apresente o mesmo perfil, mas intensidades diferentes). Mudanças podem ser causadas por substituição de partes do equipamento, envelhecimento da fonte, instabilidade do detector ou, é claro, pela aquisição dos espectros em um outro equipamento.

Se as fontes de variação forem identificadas, podem ser incorporadas ao modelo, acrescentando-se novas amostras durante a calibração. Pode-se também recorrer a técnicas de planejamento experimental como Convergência de Instrumentos ("Instrument Matching")[4] e Atualização de Modelos[5]. Para que essas soluções funcionem, porém, é necessário que o analista tenha condições de prever futuras fontes de variação e controlar os parâmetros experimentais.

*e-mail: mfp@ufpe.br

Quando não se consegue identificar as fontes de variação ou quando se deseja evitar a despesa adicional de um processo de recalibração, é preciso usar alguma outra estratégia de transferência. As duas mais empregadas têm sido a adaptação do modelo de calibração ou das respostas espectrais (padronização) e o aumento da robustez do modelo de calibração, utilizando-se algum pré-processamento dos dados para eliminar informações irrelevantes e aumentar a seletividade para os parâmetros de interesse. Um excelente tutorial sobre transferência de calibração é o trabalho de Feudale e colaboradores[2], que foi uma das principais fontes de referência para este texto.

ESTRATÉGIAS DE TRANSFERÊNCIA

Adaptação do modelo de calibração ou das respostas espectrais - métodos de padronização

Dois equipamentos podem fornecer diferentes respostas instrumentais para a mesma amostra por diversas razões. Se ambos forem semelhantes, existem dois problemas principais: deslocamentos de comprimentos de onda e mudanças nas respostas espectrais medidas. Se forem de tipos diferentes, as chances de modificações nos dados serão evidentemente maiores. Nos métodos de padronização, a função de resposta do equipamento secundário é modificada para se tornar parecida com a função de resposta do equipamento primário, onde o modelo de calibração foi desenvolvido. Isto pode ser feito através de manipulações matemáticas nas respostas espectrais ou nos valores preditos.

Padronização das respostas espectrais

Estes métodos utilizam medidas espectrais de várias amostras (ditas *de transferência*) registradas nos equipamentos envolvidos na transferência, e não precisam de seus valores analíticos de referência. Se $\mathbf{X}_P$ ($n_{trans} \times q$) e $\mathbf{X}_S$ ($n_{trans} \times p$) são as matrizes contendo os espectros de n_{trans} amostras de transferência registrados nos equipamentos primário e secundário, respectivamente, a padronização

das respostas espectrais procura relacioná-las por meio de uma transformação linear,

$$\mathbf{X}_P = \mathbf{X}_S\mathbf{F} \quad (1)$$

onde **F** é uma matriz de dimensão ($p \times q$), sendo q e p o número de comprimentos de onda para os espectros registrados nos equipamentos primário e secundário, respectivamente. **F** é estimada com o objetivo de produzir, a partir de um espectro $\mathbf{x}_S$ registrado para uma nova amostra no equipamento secundário, um espectro ajustado, $\mathbf{x}^a$, que se assemelhe ao que seria obtido se a mesma amostra tivesse sido analisada no equipamento primário:

$$\mathbf{x}^a = \mathbf{x}_S\hat{\mathbf{F}} \quad (2)$$

onde o acento circunflexo indica que $\hat{\mathbf{F}}$ é uma estimativa de **F**. Se os dois equipamentos forem iguais, a matriz **F** será quadrada, porque os espectros serão registrados no mesmo número de comprimentos de onda. Se, além disto, ela também for diagonal, só o elemento $[\mathbf{x}_S]_j$ contribuirá para o cálculo do elemento $[\mathbf{x}^a]_j$. Cada elemento do novo espectro terá, portanto, uma dependência linear simples em relação ao elemento correspondente do espectro original, ou seja, a matriz **F** ajustará separadamente as medidas em cada comprimento de onda[6]. Se os valores não-nulos de **F** estiverem localizados em uma diagonal secundária abaixo da diagonal principal, cada elemento ajustado será afetado pelo elemento correspondente ao comprimento de onda seguinte. Isto corrigirá diferenças de deslocamento de comprimentos de onda se as escalas dos equipamentos estiverem deslocadas por uma mesma distância entre pontos sucessivos. Caso não estejam, o deslocamento será modelado por uma matriz **F** com um número maior de elementos não nulos, em diagonais paralelas à diagonal principal. Dependendo de sua forma, portanto, a matriz **F** tanto pode corrigir diferenças de intensidade quanto diferenças de deslocamentos de comprimento de onda. O ajuste pode ser feito por padronização direta ("Direct Standardization, DS"), padronização direta por partes ("Piecewise Direct Standardization, PDS"), padronização reversa ou padronização univariada.

Padronização direta (DS)

Neste método, proposto por Wang *et al.*[7] para fazer com que o espectro registrado no equipamento secundário se torne parecido com o que teria sido medido no equipamento primário, as respectivas matrizes das respostas obtidas para as amostras de transferência relacionam-se através da Equação 1.

Se $\mathbf{X}_S$ for inversível, a matrix **F** pode ser obtida diretamente através da multiplicação da Equação 1 por $(\mathbf{X}_S)^{-1}$:

$$\mathbf{F} = (\mathbf{X}_S)^{-1} \cdot \mathbf{X}_P \quad (3)$$

Contudo, isto raramente ocorre porque, como costuma haver menos amostras que variáveis, $\mathbf{X}_S$ é retangular. Nesses casos, **F** pode ser calculada utilizando-se métodos multivariados, tais como regressão em componentes principais (PCR) ou por mínimos quadrados parciais (PLS). Nestes métodos, as duas matrizes espectrais são projetadas sobre o espaço coluna reduzido de $\mathbf{X}_P$, normalmente calculado por decomposição em valores singulares. Os escores obtidos nessas projeções são empregados ao invés das matrizes originais, tanto na padronização como na posterior calibração.

A padronização do novo espectro é realizada pela matriz **F** calculada através da Equação 3. A pós-multiplicação por **F** faz com que o vetor $\mathbf{x}_s$ contendo a resposta espectral de uma amostra desconhecida registrada no equipamento secundário seja padronizado para o vetor resposta esperado para a mesma amostra no equipamento primário, $\mathbf{x}_a$.

A padronização direta apresenta duas desvantagens. Em primeiro lugar, as amostras de transferência precisam ser analisadas nos dois equipamentos exatamente nas mesmas condições, pois é preciso garantir que a matriz **F** de fato represente apenas diferenças instrumentais e não mudanças na composição química das amostras. A segunda desvantagem é que o número de amostras de transferência é sempre muito menor que o número de constantes necessárias para ajustar as medidas espectrais, o que pode levar a um sobreajuste ("overfitting") na estimativa de **F**.

Padronização direta por partes (PDS)

Este método, também proposto por Wang *et al.*[7], admite que as correlações espectrais podem ser limitadas a pequenas regiões. A diferença é que na padronização DS todo o espectro do equipamento secundário é, em princípio, utilizado para reproduzir cada comprimento de onda do equipamento primário. Na padronização por partes, cada comprimento de onda no primário é relacionado somente aos comprimentos de onda de uma pequena região do espectro obtido no equipamento secundário, utilizando-se uma janela espectral móvel em torno do comprimento de onda em questão. Isto minimiza o risco de sobreajuste, já que o número de termos calculados para estimar **F** é reduzido.

No método PDS, as matrizes espectrais das amostras de transferência também se relacionam pela Equação 1, porém para cada janela calcula-se um vetor de regressão **f**, por PCR ou PLS. Estes vetores são, então, dispostos em uma matriz **F**, cuja estrutura é dada por:

$$\mathbf{F} = [\mathbf{f}_1 \ \mathbf{f}_2 \ \ldots \ \mathbf{f}_j \ \ldots \ \mathbf{f}_q] \quad (4)$$

Nesta matriz cada coluna é um vetor da forma

$$\mathbf{f}_j = [0 \ \ldots \ f_{j-k,j} \ \ldots \ f_{jj} \ \ldots \ f_{j+l,j} \ \ldots \ 0]^T \quad (4a)$$

Neste vetor, a faixa de elementos não-nulos que vai de f_{j-k} até f_{j+l} define a janela usada para a padronização. O índice j indica a que comprimento de onda do vetor $\mathbf{x}^a$ o vetor $\mathbf{f}_j$ se refere. A absorbância nesse comprimento de onda está sendo padronizada empregando-se uma janela de largura $k + l + 1$. Desta forma, o espectro de qualquer nova amostra pode ser padronizado através da Equação 2, de modo a torná-lo similar a um espectro medido no equipamento primário.

A largura da janela móvel deve ser otimizada, o que pode ser feito comparando-se as posições dos picos e larguras de banda dos espectros obtidos nos dois equipamentos. A janela pode ser simétrica (caso em que $k = l$ nas expressões acima), o que é desejável quando os deslocamentos de comprimentos de onda são desprezíveis, ou assimétrica, quando há deslocamentos significativos de comprimentos de onda em torno de cada variável. O uso da janela móvel pode provocar efeitos de borda, que podem ser corrigidos truncando-se as extremidades do espectro ou estimando-as por extrapolação. Em muitos casos relatados na literatura recente, uma janela de 3 pontos já se mostra suficiente[8-10], mas às vezes janelas maiores podem ser necessárias. Zhang *et al.*[11], por ex., empregaram uma janela de 21 pontos para a determinação de níveis fisiológicos de glicose em uma matriz complexa.

O método PDS é empregado como referência para avaliação de novas propostas de transferência de calibração, por permitir correção simultânea de diferenças de intensidade, deslocamentos de comprimentos de onda e alteração na largura das bandas. Algumas modificações do algoritmo original já foram propostas, dentre elas a correção de diferenças aditivas entre espectros[12], que é semelhante a

incluir um termo de desvio sistemático ("offset") na equação do PDS, e a correção de variações contínuas de temperatura[3]. No entanto, o PDS apresenta problemas para estimar os postos locais dos modelos PCR/PLS, pois o número de autovalores retidos é pré-determinado. A dificuldade para estimar os postos locais é atribuída a ruídos de alta freqüência nos dados que, no entanto, podem ser amenizados por um pré-processamento dos dados espectrais antes da transferência. Vários pré-processamentos foram empregados com esta finalidade. Wang *et al.*[13] utilizaram derivadas e Chen *et al.*[14] transformadas de Fourier, ao passo que Gemperline *et al.*[15] recorreram a uma PCA para decompor e reconstituir o espectro apenas com fatores relevantes. Tan e Brown[16] combinaram decomposição em coeficientes "wavelet"[2] (coeficientes de aproximação – baixa freqüência, e de detalhes – alta freqüência), com DS e PDS para transferência.

Lima e Borges[17] avaliaram a eficiência dos métodos de padronização reversa (RS e PRS, onde a letra R corresponde a "reverse", e as demais às mesmas palavras da terminologia referente à padronização direta), obtendo resultados mais favoráveis que os dos métodos DS e PDS. Os métodos de padronização reversa são semelhantes aos de padronização direta e utilizam as mesmas equações. A diferença é que na padronização reversa a matriz de transformação é utilizada para ajustar os espectros de todas as amostras do conjunto de calibração do equipamento *primário*, como se elas tivessem sido medidas no equipamento secundário. Com este conjunto padronizado calcula-se um novo modelo, que será utilizado para prever as propriedades das amostras cujos espectros serão obtidos no instrumento secundário. Apesar de consumirem um pouco mais de tempo, os métodos reversos são mais flexíveis, por permitirem o uso de técnicas quimiométricas para, por exemplo, identificar e eliminar valores anômalos ("outliers"), selecionar o melhor conjunto de amostras de calibração e de variáveis etc.

Padronização univariada

Este método, proposto por Shenk e Westerhaus[18], destinava-se inicialmente a ajustar diferenças de intensidades e admitia que não houvesse deslocamento de comprimentos de onda entre os equipamentos. Posteriormente, acrescentou-se uma correção para essa possível diferença, e o método modificado foi patenteado[19]. Inicialmente, cada comprimento de onda *i* do equipamento primário é correlacionado aos comprimentos de onda contidos em uma janela em torno de *i* no equipamento secundário. Uma função quadrática é ajustada a esses dados, sendo o seu máximo, *j*, equivalente ao comprimento de onda correspondente do equipamento primário. Finalmente, a intensidade espectral é ajustada através de uma equação linear:

$$\mathbf{x}_{iP} = a_i + b_i \,.\, \mathbf{x}_{jS} \qquad (5)$$

onde $\mathbf{x}_{iP}$ é o vetor resposta no comprimento de onda *i* das amostras medidas no instrumento primário; $\mathbf{x}_{jS}$ o vetor resposta das mesmas amostras no comprimento de onda do equipamento secundário equivalente ao do primário e a_i e b_i são, respectivamente, os coeficientes linear ("offset") e angular da reta de regressão determinada a partir das amostras de transferência.

Bouveresse *et al.*[20] propuseram uma modificação na correção das intensidades, levando em conta a possibilidade de as amostras de transferência não serem da mesma natureza que as amostras a serem analisadas. Já no método proposto por Forina *et al.*[21], as diferenças de comprimentos de onda são corrigidas ajustando-se uma função quadrática às distâncias Euclidianas e não aos coeficientes de correlação, que são mais sujeitos a ruídos.

Com qualquer um desses métodos, a padronização univariada pode corrigir diferenças de comprimentos de onda e de intensidades, mas não variações nas larguras das bandas e outros efeitos, que exigem correção por métodos multivariados.

Padronização dos valores preditos

O ajuste dos valores preditos é comumente feito por uma transformação linear e univariada conhecida pelo acrônimo SBC[22] ("slope/bias correction"). As amostras de transferência têm suas propriedades **y** medidas e preditas em ambos os equipamentos utilizando-se o modelo de calibração desenvolvido no equipamento primário, e os resultados são usados para ajustar a equação de regressão

$$\mathbf{y}_P = a + b \,.\, \mathbf{y}_S \qquad (6)$$

cujos coeficientes linear e angular corrigem, respectivamente, deslocamentos da linha de base e de inclinação devidos à transferência. Esta correção, porém, só é apropriada se os dois equipamentos forem idênticos. Caso contrário, as diferenças entre as respostas instrumentais serão mais complexas e será necessário recorrer a outros métodos de padronização, como o PDS. Bouveresse *et al.*[22] desenvolveram um procedimento para determinar quando a correção SBC pode ser empregada, usando um teste F para comparar as variâncias residuais obtidas nos equipamentos.

Aumento da robustez do modelo

Um modelo de calibração é considerado robusto quando a resposta de novas amostras não é significativamente afetada por variações não calibradas. Essa robustez pode ser influenciada pela escolha dos equipamentos envolvidos na transferência, pelo pré-processamento dos espectros e pela técnica de calibração empregada.

Modelos globais

Para se obter um modelo global, é necessário incluir na modelagem todas as variações esperadas (por exemplo, na preparação das amostras, nas condições ambientais e nos parâmetros instrumentais) em uma larga faixa de condições experimentais, a fim de minimizar ou até mesmo evitar variações não calibradas. Para tanto, o analista deve ser capaz de prever as possíveis fontes de variações, bem como a extensão da influência dessas fontes sobre os dados. Vale ressaltar que tal caracterização pode não ser uma tarefa simples, pois é necessário registrar um conjunto de dados de calibração suficientemente amplo para englobar os efeitos de todas as fontes consideradas. Adicionalmente, para representar tais efeitos, os modelos globais tendem a ser de grandes dimensões, tornando-se, portanto, mais complexos que os modelos locais. Deve-se salientar que, se um determinado conjunto de amostras não foi substancialmente afetado pelas fontes de variações em questão, um modelo local tenderia a gerar melhores predições que o modelo global[5,23]. Contudo, o modelo global é mais robusto no sentido de que suas predições continuariam confiáveis em uma gama mais ampla de situações.

Wulfred *et al.*[24] estudaram a influência dos efeitos da temperatura nos espectros e nos modelos de calibração multivariados, e concluíram que esses efeitos conduzem a relações não-lineares entre as concentrações e as absorbâncias. Tais efeitos podem ser minimizados com modelos PLS ou PCR, ou por técnicas de modelagem não-linear[25].

Pré-processamentos

Os métodos de padronização são úteis quando as amostras de transferência podem ser medidas em todos os equipamentos envolvidos na transferência do modelo, o que em geral requer que os equipamentos estejam próximos e que as amostras sejam química

e fisicamente estáveis. Quando isto não é possível, pode-se aplicar métodos de pré-processamento para tentar remover variabilidades indesejáveis no espectro, causadas por efeitos de espalhamento ou variações entre equipamentos, e ao mesmo tempo preservar a informação relevante.

Um dos pré-processamentos mais utilizados é a derivação[26,27]. A primeira derivada remove deslocamentos sistemáticos da linha de base ("offsets"); com a segunda derivada, é possível eliminar uma variação linear da linha de base, normalmente devida a efeitos de espalhamento. Como o cálculo das derivadas é feito a partir de diferenças entre valores de pontos adjacentes, a relação sinal-ruído torna-se pior com esse tipo de pré-processamento. Por isto, antes da diferenciação é comum aplicar-se aos dados algum tipo de suavização. O algoritmo mais utilizado para este fim é o de Savitzky-Golay (SG)[28].

Outro pré-processamento muito utilizado é a correção de sinal multiplicativa[27] (MSC – "Multiplicative Signal Correction"), onde se faz uma regressão de cada espectro contra um espectro de referência, que normalmente é o espectro médio do conjunto de calibração. Os espectros são então corrigidos subtraindo-se dos valores originais os coeficientes lineares das respectivas retas de regressão e dividindo-se o resultado pelos coeficientes angulares dessas mesmas retas. Para uma dada calibração, a MSC produz dois efeitos importantes: simplifica o modelo, por reduzir o número de componentes, e melhora a sua linearidade. Entretanto, por admitir que as diferenças entre linhas de base e inclinações não se alteram ao longo de toda a região espectral empregada, esta técnica não traz muitos benefícios para resultados de transferências de modelos de calibração.

Os filtros de resposta a impulso finita[29] (filtros FIR – "Finite Impulse Response") foram propostos como método de transferência de calibração para o caso de as diferenças sistemáticas de linha de base e de inclinação variarem com o comprimento de onda. Neste método, análogo a uma MSC por partes, escolhe-se um espectro de referência, que pode ser o espectro de uma determinada amostra ou o espectro médio de todas elas, e ajustam-se os demais espectros de forma a se assemelharem ao espectro de referência escolhido. O ajuste é feito varrendo-se todo o espectro com uma janela móvel ajustando-se o comprimento de onda no centro da janela, da mesma forma que na MSC.

O método FIR tem sua robustez limitada pela introdução de variações espúrias no espectro transferido. Tais variações surgem do cálculo dos modelos de regressão local, onde as correlações entre o espectro da amostra e o espectro de referência são quase linearmente independentes. Tan *et al.*[30] propuseram uma melhoria no método FIR que torna a transferência mais robusta e produz espectros livres de variações espúrias. Tanto o FIR quanto a MSC têm aplicação limitada, devido à perda de informação química que ocorre quando os espectros das amostras são forçados a se tornarem parecidos com um espectro de referência.

A correção de sinal ortogonal[31-34] (OSC – "Orthogonal Signal Correction"), que foi desenvolvida para reduzir o número de fatores em regressões multivariadas, passou a ser também utilizada para transferir modelos de calibração entre instrumentos, por evitar perda de informação durante o pré-processamento dos espectros[35]. O objetivo da OSC é reduzir a variação em **X** (matriz das respostas instrumentais) que não esteja relacionada (isto é, seja ortogonal) a **Y** (matriz dos parâmetros a serem preditos pelo modelo), e assim simplificar o modelo. Na transferência de calibração, os vetores ortogonais a **Y** e comuns aos equipamentos envolvidos são removidos, o que torna o modelo mais transferível. Quando utilizada com calibração multivariada, a OSC é calculada apenas com amostras de calibração. Para fins de transferência, a correção deve ser calculada para um conjunto de amostras de transferência medidas nos equipamentos envolvidos ($\mathbf{X}_t$), a fim de permitir que todas as fontes de variação sejam consideradas[10,36].

O primeiro passo do algoritmo comumente utilizado para cálculo das componentes OSC é ortogonalizar a primeira componente principal (PC1) de $\mathbf{X}_t$ em relação a **y**, através da Equação[31]

$$\mathbf{t}_{ort} = \mathbf{t} - \mathbf{y}(\mathbf{y}^T\mathbf{y})^{-1}\mathbf{y}^T\mathbf{t} \quad (7)$$

em que **t** representa o vetor dos escores das amostras em PC1. Esta operação tem o objetivo de obter a maior variância possível nos dados que não esteja correlacionada com a informação química. Em seguida, com uma regressão PLS de $\mathbf{t}_{ort}$ contra $\mathbf{X}_t$, obtém-se o vetor dos pesos $\mathbf{w}_{ort}$ que minimiza a covariância entre $\mathbf{X}_t$ e **y**, uma vez que este vetor contém pouca informação química. Esses pesos são utilizados para atualizar o vetor ortogonal dos escores ($\mathbf{t}_{ort}$), que será aproveitado para calcular um vetor correspondente de "loadings" ($\mathbf{p}_{ort}$).

$$\mathbf{t}_{ort} = \mathbf{X}_t\mathbf{w}_{ort} \quad (8)$$

$$\mathbf{p}_{ort} = \mathbf{X}_t^T\mathbf{t}_{ort}/(\mathbf{t}_{ort}^T\mathbf{t}_{ort}) \quad (9)$$

A primeira componente OSC, que representa a parte do espectro que é irrelevante para prever **Y**, é uma combinação de $\mathbf{t}_{ort}$ e $\mathbf{p}_{ort}$. Esta componente OSC é então subtraída de $\mathbf{X}_t$, fornecendo a matriz de resíduos $\mathbf{E}_t$. A matriz $\mathbf{E}_t$, que possui uma melhor relação sinal/ruído, é usada para substituir $\mathbf{X}_t$ e o procedimento é repetido até que o número desejável de componentes OSC seja removido.

Como já comentado, o modelo OSC para transferência é construído apenas com as amostras de transferência, cujos espectros foram obtidos em ambos os instrumentos. Estas amostras não precisam ser idênticas, mas devem ser representativas das variações instrumentais. As amostras restantes (**X**), que não foram usadas para calcular os vetores $\mathbf{w}_{ort}$ e $\mathbf{p}_{ort}$, são corrigidas empregando-se a Equação 10

$$\mathbf{X}^a = \mathbf{X} - \mathbf{X}\mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{P}^T \quad (10)$$

em que **W** é a matriz que contém os vetores $\mathbf{w}_{ort}$ dispostos na ordem em que foram obtidos e **P**, a matriz contendo os respectivos "loadings" $\mathbf{p}_{ort.}$.

A Transformada Wavelet[37] (WT) tem sido usada como método de processamento alternativo para transferência de calibração[38,39]. Utilizando-se a WT os sinais originais são decompostos em coeficientes "wavelet" com diferentes freqüências, o que permite a remoção de certas freqüências. Em geral, os coeficientes de baixa freqüência estão relacionados a diferenças instrumentais e os de alta freqüência, a ruídos de medida, que correspondem à informação irrelevante para a análise. Além de ser mais versátil e dispensar o uso de amostras de transferência, a Transformada Wavelet é útil na remoção de variações de linhas de base entre os conjuntos de calibração e teste, e contribui para melhorar a relação sinal/ruído[40].

Andrew e Fearn[41] propuseram recentemente um método chamado de Transferência por Projeções Ortogonais (TOP – "Transfer by Orthogonal Projections"), que utiliza informação sobre a variação espectral entre instrumentos para fazer uma calibração robusta e transferível. Os espectros de um pequeno conjunto de amostras são registrados em *n* equipamentos, obtendo-se daí uma matriz contendo os espectros médios para cada equipamento. As *n*–1 componentes principais dessa matriz representam as principais direções de variação entre os equipamentos no espaço espectral. A matriz **X** dos dados espectrais para o conjunto de calibração é en-

tão ortogonalizada contra algumas dessas componentes, removendo-se as *d* dimensões que melhor descrevem a variabilidade entre equipamentos. O valor de *d* é uma questão ainda em aberto. Pode-se utilizar todas as componentes ($d = n-1$), ou selecionar as que forneçam os melhores resultados, a julgar pelo desempenho da calibração. Andrew e Fearn utilizaram o número de componentes principais necessário para descrever 98-99% da variação entre equipamentos[41].

Ao remover dimensões dos dados espectrais antes da calibração, o método TOP assemelha-se à OSC. A diferença é que o TOP utiliza informações do conjunto de transferência para remover dimensões que interferem na transferência, enquanto a OSC só remove dimensões ortogonais ao vetor das respostas **y**.

Cada pré-processamento opera eliminando uma fonte específica de variação, podendo levar a modelos mais parcimoniosos[23]. Em geral, os métodos descritos nesta seção são eficientes quando o problema se deve a variações nas intensidades das bandas espectrais, os equipamentos envolvidos são similares e as diferenças instrumentais são pequenas.

Seleção de variáveis

A seleção de variáveis insensíveis ou menos sensíveis a variações nas condições experimentais melhora a robustez dos modelos de calibração, além de aumentar a sua capacidade preditiva[42]. Existem várias razões práticas para isto. Por ex., algumas regiões espectrais podem não mostrar uma grande variação para o parâmetro de interesse, os espectros podem conter ruídos heteroscedásticos, pode haver comprimentos de onda em que a intensidade do sinal não seja linearmente correlacionada com o parâmetro de interesse, bem como pode haver comprimentos de onda que tenham uma correlação indireta com esse parâmetro. Nos métodos de seleção de variáveis, procura-se identificar as variáveis contendo informações relevantes para o problema de interesse, e faz-se a modelagem utilizando-se apenas essas variáveis. Os modelos assim obtidos são mais simples e fáceis de interpretar.

Para um modelo de calibração, várias técnicas de seleção de variáveis têm sido propostas[26]. Por ex., a seleção do melhor subconjunto baseada na comparação dos erros de predição de todos os modelos possíveis; a seleção progressiva ("forward selection"), que seleciona as melhores variáveis com base nos sucessivos modelos obtidos pela ampliação do número de variáveis; a eliminação regressiva ("backward elimination"), que começa com um modelo baseado em todas as variáveis e, de acordo com algum critério pré-estabelecido, vai descartando as menos importantes; o procedimento de seleção passo a passo ("stepwise procedure"), que é equivalente à seleção progressiva, diferenciando-se apenas pela eliminação prévia de variáveis que não contribuem para o modelo. Estes procedimentos têm como desvantagem a grande probabilidade de que o resultado da seleção seja apenas um ótimo local, e não o melhor conjunto de variáveis.

O Algoritmo Genético (AG) seleciona um conjunto de variáveis usando uma heurística de busca aleatória, sendo menos propenso a convergência para ótimos locais[43]. O algoritmo é baseado na teoria da evolução de Darwin, selecionando modelos construídos com combinações de variáveis que fornecem os melhores valores de predição. Cada combinação ("indivíduo") é representada por um "cromossomo", cujos genes são dígitos binários que codificam cada variável ("1" se a variável foi selecionada e "0" se foi descartada). Desta forma, cada combinação será representada por uma seqüência diferente de genes (diferentes posições de "uns" e "zeros"). Uma população é formada por um determinado número de seqüências diferentes, que são avaliadas a partir da capacidade preditiva do modelo. As melhores seqüências (aquelas cujos modelos forneceram os menores valores de erro para o conjunto de teste) sobrevivem, sofrem mutação e recombinam-se de acordo com critérios pré-definidos para formar os descendentes. A mutação é utilizada para incluir variáveis que não estão presentes nos modelos da população inicial (pais). A principal dificuldade é encontrar os valores adequados para os parâmetros que podem afetar o resultado final (tamanho da população, grau de mutação, regra de combinação, regra de seleção e número de gerações).

O algoritmo "jack-knife"[44] também pode ser usado para selecionar as variáveis que serão mantidas nos modelos de regressão PLS. Os erros padrão das estimativas dos coeficientes de regressão obtidos por PLS são avaliados de forma empírica, a partir de uma série de subconjuntos de dados. A partir daí, podem-se calcular intervalos de confiança para os coeficientes de regressão e avaliar a significância de cada variável utilizada no modelo. O mesmo procedimento pode ser utilizado para regressão em componentes principais (PCR) ou qualquer outro método de regressão.

Finalmente, pode-se selecionar variáveis por meio do Algoritmo das Projeções Sucessivas (APS), recentemente proposto[45-50], que utiliza operações simples para reduzir a colinearidade do conjunto de dados espectrais, de forma a melhorar o condicionamento numérico e reduzir a propagação de ruídos na construção de um modelo por Regressão Linear Múltipla ("Multiple Linear Regression, MLR").

A redução da colinearidade dos dados espectrais permite a utilização de modelos MLR, que são mais simples e parcimoniosos que os modelos PLS/PCR. Tal ganho de parcimônia tende a melhorar a capacidade de generalização do modelo, permitindo que seja aplicado com mais sucesso em novas situações, o que é importante em uma transferência de calibração. Além disto, estudos comparativos mostraram que o APS pode produzir modelos mais generalizáveis que os obtidos usando-se o AG para a seleção de variáveis[47,48].

O APS inicia a seleção por uma determinada variável e segue adicionando outras, até que uma condição de parada seja satisfeita. Três aspectos devem ser considerados: a escolha da variável inicial, o critério de seleção para as variáveis restantes, e a condição de parada (número *N* de variáveis a serem selecionadas).

O APS trabalha com base em conjuntos de calibração e teste, ambos contendo dados de respostas instrumentais (**X**) e valores dos parâmetros medidos por um método de referência (**y**). A principal operação do APS consiste em manipulações algébricas executadas na matriz $\mathbf{X}_{\mathbf{CAL}}$ (K_C x J), onde linhas e colunas correspondem a K_C amostras de calibração e J variáveis espectrais, respectivamente. A partir de uma coluna $\mathbf{x}_0$, correspondente à variável inicial da seleção, o APS determina qual das colunas restantes tem a maior projeção no subespaço S_0 ortogonal a $\mathbf{x}_0$. Esta coluna, chamada de $\mathbf{x}_1$, pode ser considerada como a que contém a maior quantidade de informação não incluída em $\mathbf{x}_0$. Na próxima iteração, o APS restringe a análise ao subespaço S_0, tomando $\mathbf{x}_1$ como a nova coluna de referência, e procede com os passos descritos anteriormente. Desta forma, o critério de seleção do APS favorece a minimização da colinearidade entre as variáveis. Vale ressaltar que no máximo K_C variáveis podem ser selecionadas desta maneira, porque, depois de cada operação de projeção, a dimensão do espaço coluna de $\mathbf{X}_{\mathbf{CAL}}$ é reduzida de um fator, ou seja, um grau de liberdade é removido. Assim, depois de K_C operações de projeção todos os vetores coluna de $\mathbf{X}_{\mathbf{CAL}}$ terão sido projetados sobre a origem do espaço, isto é, $\mathbf{X}_{\mathbf{CAL}}$ se tornará uma matriz de posto nulo.

Para determinar a melhor variável inicial (coluna de $\mathbf{X}_{\mathbf{CAL}}$) e o número ótimo de variáveis *N*, o APS emprega o seguinte procedimento: se *N* é fixo, *J* subconjuntos de *N* variáveis podem ser selecionados utilizando cada uma das *J* variáveis disponíveis como ponto de partida. Para cada um desses subconjuntos de variáveis,

um modelo MLR é construído e o erro quadrático médio de predição no conjunto de teste (*RMSEV*) é calculado de acordo com a Equação:

$$RMSEV = \sqrt{\frac{1}{Kv}\sum_{k=1}^{Kv}(y_v^k - \hat{y}_v^k)^2} \quad (11)$$

onde y_v^k e $\hat{y}_v^k$ são, respectivamente, os valores de referência e predito para o parâmetro de interesse na *k*-ésima amostra de teste, e *Kv* é o número de amostras de teste. O menor *RMSEV* assim obtido é denotado por $RMSEV^*(N)$, onde o asterisco é usado para indicar o melhor resultado para subconjuntos de *N* variáveis. Repetindo-se este procedimento para $N = 1, 2, ..., K_C$, o *N* ótimo pode ser obtido do mínimo da curva de $RMSEV^*(N)$. Para economizar tempo computacional, o procedimento pode ser interrompido antes de *N* alcançar K_C, se o analista julgar que o ponto mínimo da curva do $RMSEV^*(N)$ já está bem estabelecido.

Honorato *et al.*[9] propuseram uma melhora na robustez do conjunto de variáveis selecionadas pelo APS. Nessa variante, denotada pelo acrônimo APSV, o valor do *RMSEV* é substituído por um critério de seleção que também leva em conta o erro de predição para o conjunto de dados de transferência. Tal critério foi originalmente proposto por Swierenga *et al.*[51], que testaram a capacidade de transferência de modelos PLS para determinar o teor de água em tabletes. Para isso, utilizou-se recozimento simulado (SA – "Simulated Anneling") para selecionar um subconjunto das variáveis originais que preservasse a capacidade de predição do modelo quando este fosse transferido para outro equipamento. Nessas condições, foram obtidos resultados melhores que os produzidos com o método DS.

O novo critério é a minimização do valor do erro *E* dado por:

$$E = 1/2(RMSEV + RMSET) \quad (12)$$

em que *RMSET* é a raiz do erro médio quadrático de predição para as amostras do conjunto de transferência medidas no equipamento secundário, definido como:

$$RMSET = \sqrt{\frac{1}{Kt}\sum_{k=1}^{Kt}(y_t^k - \hat{y}_t^k)^2} \quad (13)$$

sendo y_t^k e $\hat{y}_t^k$ os valores de referência e predito para o parâmetro de interesse na *k*-ésima amostra de transferência, e *Kt* o número de amostras de transferência. Desta forma, a seleção de variáveis considerará não apenas o poder de predição do modelo (medido por *RMSEV*), mas também sua robustez (medida por *RMSET*), atribuindo a mesma importância aos dois critérios, independentemente do número de amostras de teste e de transferência utilizados.

Seleção das amostras de transferência

A escolha do conjunto de amostras que serão lidas nos equipamentos envolvidos na transferência de calibração (amostras de transferência ou de padronização) é obviamente importante para o sucesso da transferência. Para evitar variações espúrias no espectro transferido, as amostras devem ser representativas do espaço experimental, bem como varrer as diferenças instrumentais. Se isto não acontecer, os parâmetros de transferência calculados levarão a padronizações duvidosas.

Existem várias idéias, princípios e estratégias para a seleção de amostras de transferência[7,20,38,52]. Um algoritmo clássico é o Kennard-Stone (KS), que começa selecionando as duas amostras com a maior distância Euclidiana entre si no espaço **X**. Para cada uma das amostras restantes, calcula-se a distância mínima com respeito às amostras já selecionadas. Feito isso, a amostra com a maior distância mínima é retida, e o procedimento é repetido até que um determinado número de amostras seja selecionado[53].

Wang *et al*.[7] selecionaram amostras de acordo com seus valores de alavancagem ("leverage"). A alavancagem é uma medida da distância de uma determinada amostra em relação à média de todas as amostras. A seleção inicia-se com a amostra de maior alavancagem, que é removida do conjunto de dados e ortogonalizada contra as demais amostras. A seqüência é repetida até que o número desejado de amostras seja extraído.

Bouveresse *et al.*[52] compararam a seleção do conjunto de amostras de transferência usando o algoritmo KS, baseado nos valores de alavancagem, e um método baseado na maximização da menor distância entre pontos (MSID). Os melhores resultados foram obtidos utilizando-se os algoritmos KS e MSID.

Dantas Filho *et al.*[54] propuseram o algoritmo APSA (*APS* para amostras) para seleção de amostras de calibração. Trata-se do próprio APS, só que modificado para selecionar amostras ao invés de variáveis. A matriz das respostas instrumentais é transposta (isto é, suas colunas passam a ser as amostras de calibração) e o APS é usado para selecionar um subconjunto de amostras minimamente redundante. Comparado ao algoritmo clássico KS, o APSA forneceu resultados comparáveis ou ligeiramente melhores[9]. Com certas adaptações, o APSA também pode ser utilizado para transferência[9].

ESTUDO DE CASO

Para avaliar as diferentes estratégias de transferência de calibração, apresentaremos um estudo de caso baseado em dados espectrais de refletância difusa (R) para 80 amostras de milho registrados na região NIR (1100-1498 nm), disponíveis no endereço http://www.eigenvector.corn/Data/Corn/. Neste estudo, os equipamentos 1 e 2 são, respectivamente, considerados como principal e secundário, sendo os teores de umidade, óleo, proteína e amido das amostras os parâmetros de interesse. Os aspectos mais importantes de cada técnica, como número de amostras de transferência, tamanho da janela do PDS e número de componentes OSC serão discutidos, bem como suas respectivas vantagens e desvantagens.

Notação

As seguintes abreviaturas serão utilizadas:

RMSEP Erro quadrático médio de predição.

$RMSEP^P_P$ RMSEP para o conjunto de predição do equipamento primário utilizando o modelo calibrado nesse equipamento.

$RMSEP^S_S$ Idêntico ao anterior, só que no equipamento secundário.

$RMSEP^S_P$ RMSEP para o conjunto de predição do equipamento secundário utilizando o modelo calibrado no equipamento primário.

$RMSEP^S_{P\text{-}T}$ RMSEP para o conjunto de predição do equipamento secundário utilizando o modelo calibrado no primário, com a letra T representando a técnica de transferência utilizada (PLS-DS, PLS-PDS, PLS-OSC, APSV-MLR).

Softwares empregados

Os cálculos para o método PDS foram feitos com programas implementados no ambiente Matlab 6.5 R13[55], com base no pacote de ferramentas PLS Toolbox 3.0[56]. Empregou-se uma janela simétrica para o PDS, variando-se seu tamanho de 3 até 15 pontos. A

matriz **F** de padronização foi determinada através de uma PCR, retendo-se as componentes principais necessárias para explicar 99% da variância. A primeira derivada dos espectros com suavização por filtros Savitzky-Golay, os modelos MLR, a seleção de variáveis pelo APS e a seleção de amostras pelo KS também foram determinadas com o software Matlab, usando-se programas desenvolvidos pelos autores. A construção dos modelos PLS e a implementação do método OSC foram realizadas com o software Unscrambler 9.2[57].

Pré-processamento dos dados e modelagem PLS

Para evitar problemas devidos a variações sistemáticas na linha de base, utilizou-se o espectro derivado depois de uma suavização por filtros Savitzky-Golay com polinômio de 2ª ordem e janela de 21 pontos. A Figura 1 apresenta os perfis dos espectros das 80 amostras nos equipamentos primário e secundário. A Figura 2 mostra os perfis desses mesmos espectros depois dos pré-processamentos. Os espectros são representados por 680 variáveis.

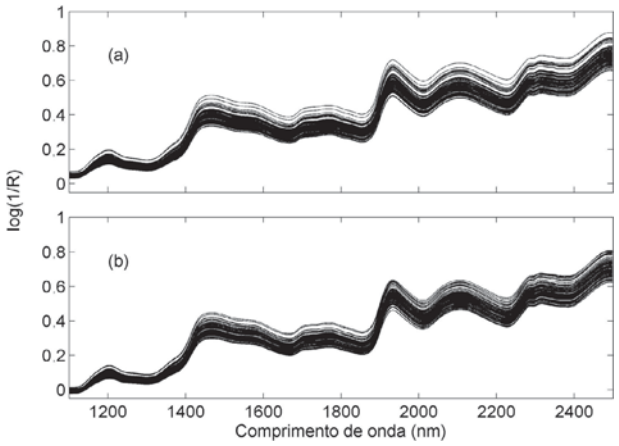

Figura 1. *Espectros de partida: (a) primário e (b) secundário*

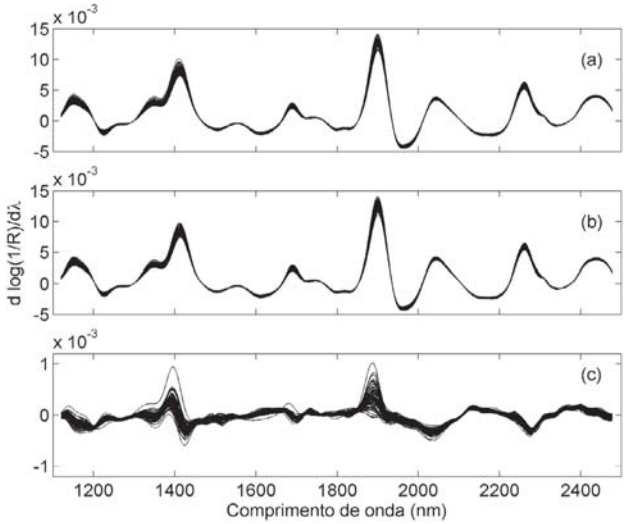

Figura 2. *Espectros derivados: (a) primário, (b) secundário e (c) diferenças entre o primário e secundário*

Uma análise PLS preliminar de todo o conjunto de dados espectrais foi realizada para se escolher os conjuntos de calibração, teste e predição. A partir de uma análise visual dos escores, o conjunto foi dividido em 40 amostras de calibração, 20 de teste e 20 de predição. As amostras de teste e de predição foram selecionadas de forma a cobrir a variabilidade em **X** e **Y**, evitando-se os pontos extremos desses espaços.

Um novo modelo PLS foi então desenvolvido para o conjunto de calibração, empregando-se as amostras de teste para definir o número apropriado de fatores para cada propriedade em estudo. Os resultados da aplicação do modelo PLS baseado nos dados do instrumento primário aos conjuntos de predição no primário e no secundário são apresentados na Tabela 1. Pode-se notar que o erro no equipamento secundário é uma ordem de grandeza maior que o erro no primário, o que justifica o emprego de técnicas de transferência de calibração.

Tabela 1. RMSEP (%m/m) obtidos com uma calibração PLS

	Umidade (4)	Óleo (13)	Proteína (7)	Amido (12)
$\mathbf{RMSEP^P_P}$	0,0675	0,0268	0,0859	0,1485
$\mathbf{RMSEP^S_P}$	1,4995	0,7917	0,5289	2,0537

O número de variáveis latentes empregadas na modelagem de cada parâmetro encontra-se indicado entre parênteses

Seleção das amostras de transferência: algoritmo Kennard-Stone

A Figura 3 mostra o gráfico dos escores das duas primeiras componentes principais das amostras do conjunto de calibração, destacando-se as 10 primeiras amostras selecionadas para transferência pelo algoritmo Kennard-Stone. A numeração das amostras segue a ordem resultante da aplicação do algoritmo. Pode-se observar que as amostras 1 a 4 correspondem aos extremos do conjunto de calibração. A partir da quinta amostra, o algoritmo procura fazer a seleção de forma uniforme dentro do espaço amostral, procurando evitar que sejam escolhidas amostras próximas entre si. Com isto, obtém-se um conjunto representativo do total de amostras.

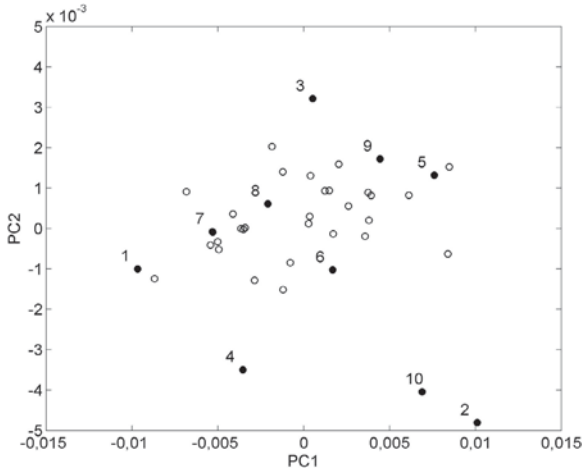

Figura 3. *Seleção de amostras de transferência pelo algoritmo Kennard-Stone. Apenas as 10 primeiras amostras estão indicadas. A numeração das amostras segue a ordem resultante da aplicação do algoritmo. As duas primeiras componentes principais explicam, respectivamente, 87 e 9% da variância dos dados*

Métodos de padronização: DS e PDS

A Figura 4 mostra o resultado da aplicação do método DS em função do número de amostras de transferência. Para cada propriedade, o gráfico apresenta o erro ($RMSEP^S_{P\text{-}DS}$) obtido empregando-

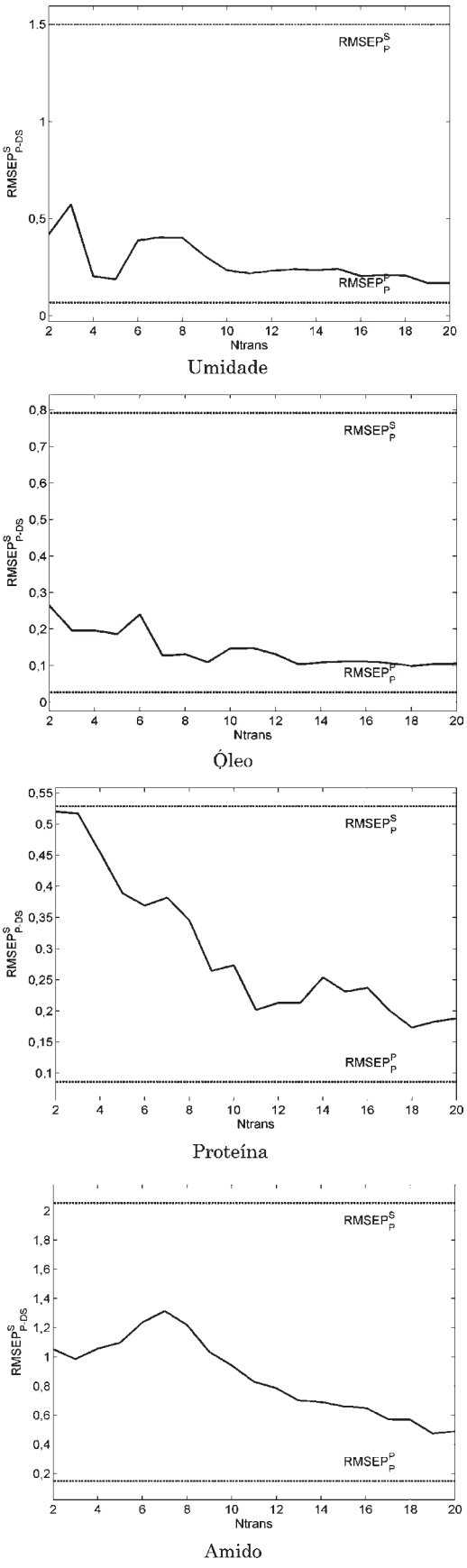

Figura 4. *RMSEP resultantes da aplicação do modelo desenvolvido para o equipamento primário aos espectros do secundário padronizados com DS. O número de amostras de transferência (Ntrans) utilizadas na padronização foi variado de 2 a 20*

se o modelo desenvolvido no equipamento primário e os espectros registrados no secundário (após a padronização pelo DS). Para fins de comparação são também apresentados os valores de $RMSEP^S_P$ (espectros obtidos no secundário usados como entrada para o modelo PLS desenvolvido no primário, sem utilização de técnicas de transferência) e $RMSEP^P_P$ (espectros obtidos no primário usados com o modelo desenvolvido no próprio primário). Observa-se que para todas as propriedades, e para qualquer número de amostras de transferência, o RMSEP quando se aplica o DS é menor que o valor obtido quando se aplica diretamente o modelo do primário aos espectros obtidos no secundário. O $RMSEP^S_{P\text{-}DS}$ tende a diminuir com o aumento do número de amostras de transferência, havendo um nivelamento do erro a partir de 13 amostras. De modo geral, os valores de $RMSEP^S_{P\text{-}DS}$ aproximam-se bastante dos valores do $RMSEP^P_P$, embora não cheguem a atingi-los. Os teores de proteína e de amido requerem um número maior de amostras de transferência que os de umidade e óleo, para que o erro se reduza a níveis próximos do $RMSEP^P_P$. Uma possível explicação pode ser dada com base na Figura 5, que apresenta a variação percentual de cada propriedade coberta pelo conjunto de transferência. Para cobrir a variação de proteína e amido, foi necessário usar mais amostras de transferência que para as outras duas propriedades.

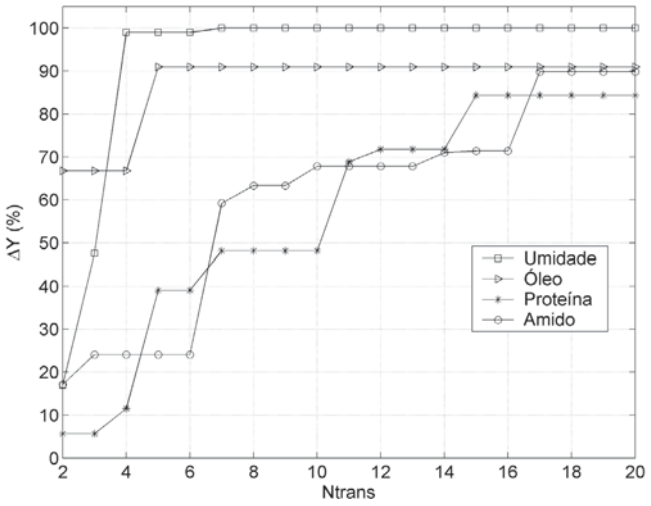

Figura 5. *Variação percentual dos valores de* y *no conjunto de calibração coberto pelas amostras de transferência para cada propriedade. Ntrans é o número de amostras de transferência*

O efeito da padronização DS no espectro derivado de uma amostra é ilustrado na Figura 6. Como se pode notar na ampliação, a padronização faz com que o espectro do secundário se aproxime do primário, o que explica a melhora no resultado da predição.

Os resultados da aplicação do método PDS são mostrados na Figura 7, em função do número de amostras de transferência e do tamanho da janela utilizada. Para cada propriedade, o gráfico apresenta o erro ($RMSEP^S_{P\text{-}PDS}$) obtido empregando-se o modelo desenvolvido no primário e os espectros registrados no secundário (após a padronização por PDS). Para fins de comparação são também apresentados os valores de $RMSEP^S_P$ e $RMSEP^P_P$. Para todas as propriedades, independentemente do número de amostras de transferência e do tamanho da janela, o RMSEP quando se aplica o PDS é menor que o obtido aplicando-se diretamente o modelo do primário aos espectros obtidos no secundário. O efeito do tamanho da janela varia com a propriedade, mas só é significativo quando o número de amostras de transferência é pequeno. A partir de 8 amostras, o tamanho da janela deixa de ser importante. Em comparação com o DS, o

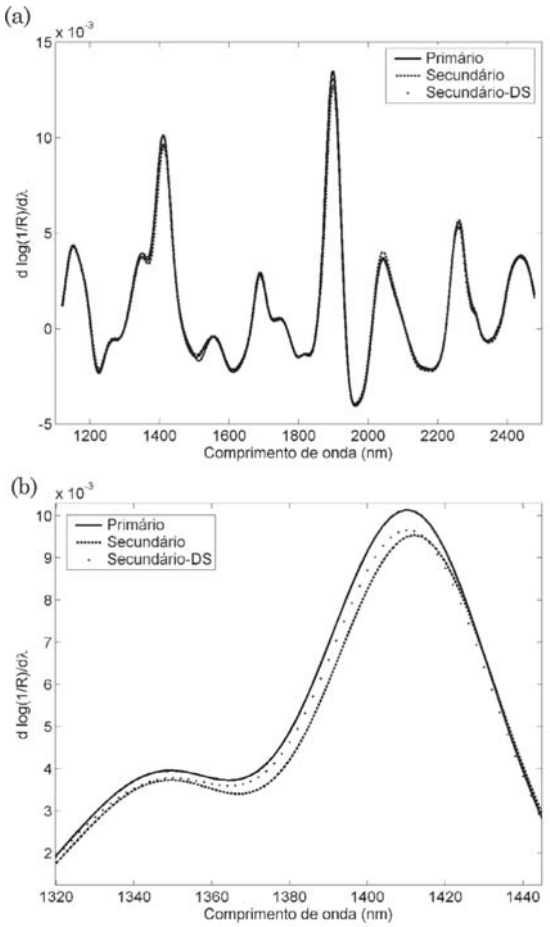

Figura 6. *(a) Comparação entre o espectro de uma mesma amostra no equipamento primário, no secundário e no secundário após a padronização por DS. (b) Ampliação da faixa 1320-1440 nm, para melhor visualização do efeito da padronização*

$RMSEP^{S}_{P\text{-}PDS}$ estabiliza-se com a utilização de um número menor de amostras de transferência, como se pode constatar na Figura 8, em que o $RMSEP^{S}_{P\text{-}PDS}$ é apresentado para uma janela de 3 pontos. A partir de 4 amostras já se obtém um nivelamento da curva do $RMSEP^{S}_{P\text{-}PDS}$ para os teores de óleo, proteína e amido. Para a umidade o erro ainda diminui até cerca de 12 amostras.

A Figura 9 mostra as diferenças entre os espectros derivados dos equipamentos primário e secundário, antes e após padronização por DS e PDS usando-se 13 amostras de transferência. Em geral, a padronização reduz as diferenças e tende a remover seu caráter sistemático. Além disto, a correção pelo PDS é mais eficaz que a correção pelo DS. Como se verá na análise comparativa final (Tabela 3), os erros de predição do método PDS são menores que os do DS.

Melhora da robustez do modelo

Pré-processamento com OSC

A variação dos resultados do método OSC com o número de componentes é ilustrada na Figura 10, que também mostra a influência da variação do número de amostras de transferência (5 ou 13). Para comparação, também são representados os valores de $RMSEP^{S}_{P}$ e $RMSEP^{P}_{P}$. Para todas as propriedades, os modelos OSC construídos no equipamento primário e aplicados aos espectros obtidos no próprio primário possuem valores de $RMSEP^{P}_{P\text{-}OSC}$ um pouco maiores que os modelos PLS sem OSC ($RMSEP^{P}_{P}$), independentemente do

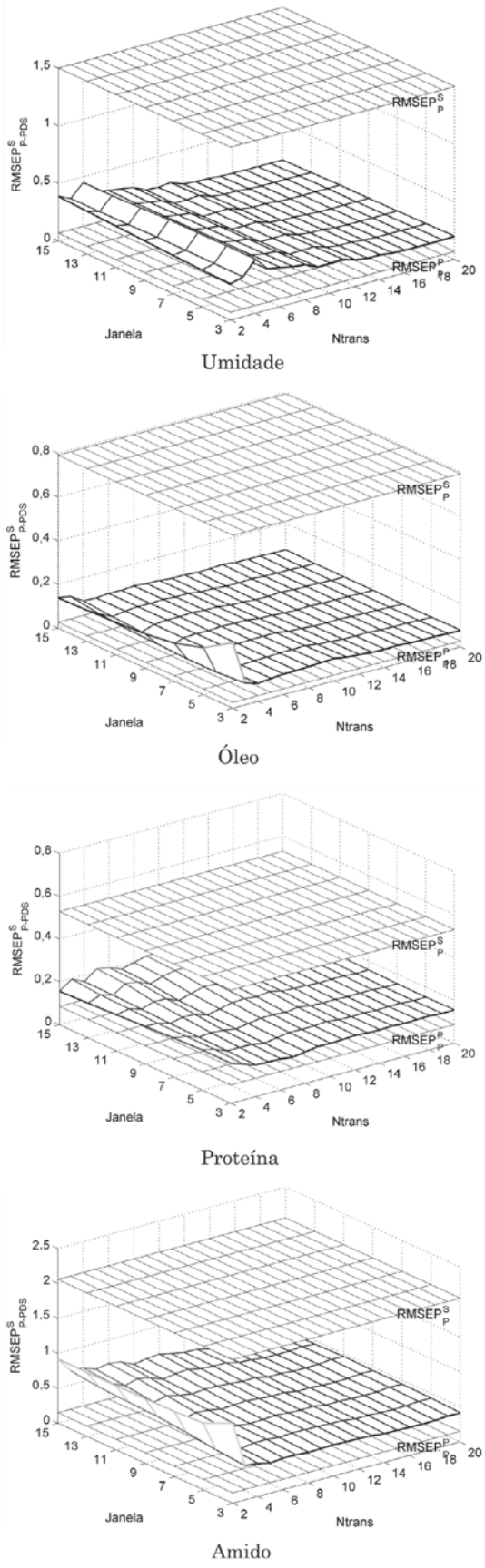

Figura 7. *RMSEP resultantes da aplicação do modelo desenvolvido para o equipamento primário aos espectros do secundário padronizados com PDS. O número de amostras de transferência utilizadas na padronização (Ntrans) foi variado de 2 a 20 e o tamanho da janela para o PDS, de 3 a 15*

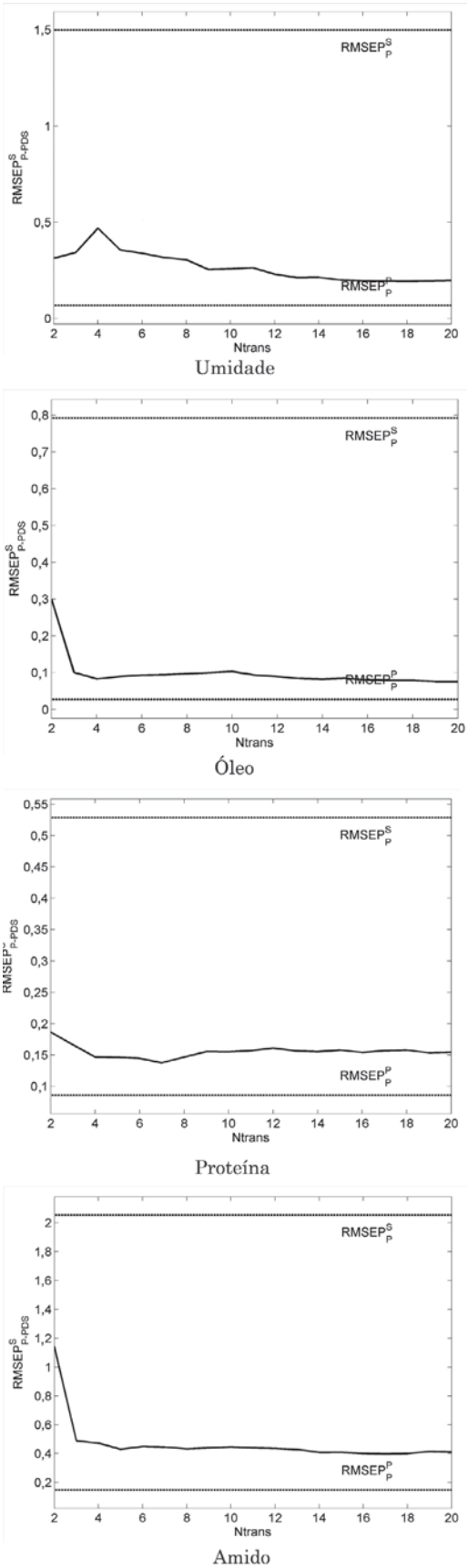

Figura 8. *RMSEP resultantes da aplicação do modelo do equipamento primário aos espectros do secundário padronizados com PDS. Foi empregada uma janela de 3 pontos, com o número de amostras de transferência (Ntrans) variando de 2 a 20*

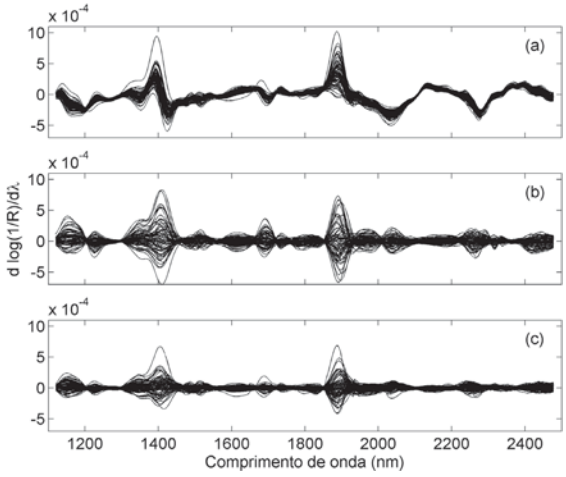

Figura 9. *Diferenças entre os espectros derivados dos equipamentos primário e secundário, (a) antes e após a padronização com 13 amostras de transferência por (b) DS e (c) PDS, com janela de 3 pontos*

número de amostras de transferência. Na transferência este pequeno acréscimo é compensado, uma vez que os RMSEP's dos espectros obtidos no secundário utilizando-se o modelo do primário são bem menores quando se usa OSC (RMSEP$^S_{P\text{-}OSC}$) que quando o PLS é aplicado sem essa técnica (RMSEP^S_P), particularmente quando o número de componentes OSC é igual ou maior que 3. O efeito do número de amostras de transferência torna-se mais significativo quando o número de componentes OSC diminui. A partir de 3 componentes OSC, os valores dos erros são semelhantes para todas as propriedades.

Seleção de variáveis robustas: algoritmo APSV

Os resultados da seleção de variáveis robustas em termos dos valores de RMSEP utilizando-se ou não amostras de transferência são apresentados na Tabela 2. Para os teores de umidade, óleo e proteína, os valores de RMSEP^S_P diminuem consideravelmente quando se utiliza o critério da Equação 12 (isto é, quando são empregadas amostras de transferência). Uma redução menos expressiva é observada para o amido. Em todos os casos, elevar o número de amostras de transferência de 5 para 13 não altera significativamente os resultados.

Tabela 2. RMSEP obtidos com calibração APSV-MLR com e sem o uso de amostras de transferência

RMSEP	*Ntrans*	Umidade	Óleo	Proteína	Amido
RMSEP^P_P	0	0,013(22)	0,029(25)	0,019(30)	0,124(24)
RMSEP^S_P		1,494	0,539	0,814	0,426
RMSEP^P_P	5	0,224 (6)	0,042 (22)	0,074 (21)	0,139 (24)
RMSEP^S_P		0,262	0,070	0,139	0,375
RMSEP^P_P	13	0,194 (5)	0,042 (22)	0,073 (10)	0,139 (24)
RMSEP^S_P		0,184	0,070	0,137	0,375

Os valores entre parênteses são os números de comprimentos de onda usados para cada parâmetro

Os métodos APSV-MLR e OSC não requerem que as mesmas amostras de transferência sejam medidas nos equipamentos primário e secundário. Isto pode ser uma vantagem decisiva se as medidas tiverem de ser feitas para conjuntos de dados obtidos separadamente, seja no espaço ou no tempo.

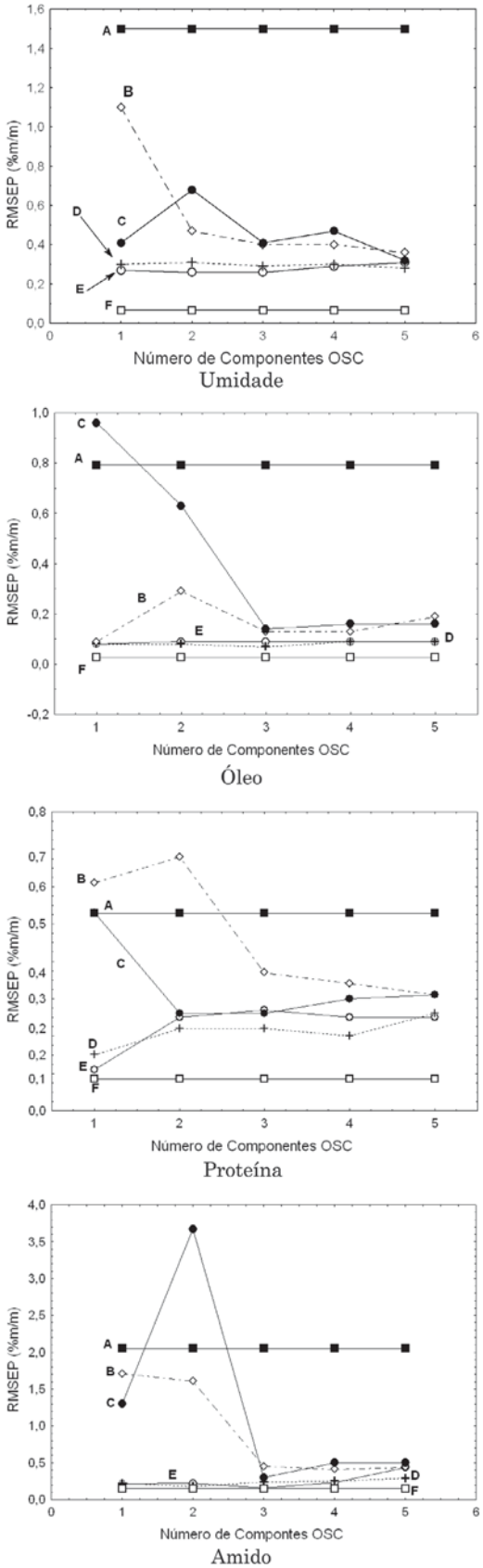

Figura 10. *RMSEP resultantes da aplicação do modelo do equipamento primário aos espectros do secundário, usando-se o método OSC, em função do número de componentes. A(■): $RMSEP^{S}_{P}$; B(◇):$RMSEP^{S}_{P\ OSC\text{-}5}$; C(●):$RMSEP^{S}_{P\ OSC\text{-}13}$; D(+):$RMSEP^{P}_{P\ OSC\text{-}13}$; E(○):$RMSEP^{P}_{P\ OSC\text{-}5}$; F(□):$RMSEP^{P}_{P}$*

Análise comparativa dos resultados

A Tabela 3 apresenta os valores de $RMSEP^{S}_{P}$ obtidos com as várias técnicas de transferência analisadas, usando-se 13 amostras de transferência, janela de 3 pontos para o PLS-PDS e 3 componentes OSC.

Tabela 3. Valores de $RMSEP^{S}_{P}$ para as técnicas de transferência utilizadas

	PLS	PLS-DS	PLS-PDS	OSC	APSV-MLR
Umidade	1,499	0,240	0,212	0,410	0,184
Óleo	0,792	0,104	0,084	0,140	0,070
Proteína	0,529	0,213	0,157	0,260	0,137
Amido	2,054	0,701	0,428	0,300	0,375

Como já se enfatizou, a aplicação das técnicas de transferência diminui os erros de predição para os espectros coletados no equipamento secundário, empregando-se o modelo desenvolvido no primário. Para todas as propriedades, o PDS apresenta resultados um pouco melhores que o DS. Os erros de predição empregando-se OSC foram um pouco menores para o teor de amido. Para os teores de umidade, óleo e proteína, a estratégia APSV-MLR mostrou-se levemente superior às demais.

CONCLUSÃO

Neste tutorial foram apresentados os principais conceitos e técnicas empregados para transferência de calibração. Em um estudo de caso envolvendo análise de amostras de milho por refletância difusa, foram comparados os resultados da transferência de modelos PLS com os métodos tradicionais (DS, PDS, OSC). Também foram apresentados resultados obtidos com uma técnica recente de seleção de variáveis robustas (APS), para o desenvolvimento de modelos MLR.

Todas as técnicas de transferência estudadas aqui diminuíram significativamente (em alguns casos por uma ordem de grandeza) os erros de predição resultantes do emprego de um modelo desenvolvido no equipamento primário com espectros obtidos no equipamento secundário. Entre as técnicas tradicionais, os melhores resultados foram obtidos pelo método PDS com uma janela de 3 pontos. Para uma única propriedade (o teor de amido), o método OSC mostrou-se superior. O método de seleção de variáveis mostrou-se uma alternativa viável às técnicas tradicionais, produzindo resultados mais satisfatórios para a predição de três das quatro propriedades estudadas.

AGRADECIMENTOS

Ao apoio recebido das agências financiadoras Finep/CT-PETRO, CAPES/PROCAD e CNPq (processo 475204/2004-2/Edital Universal, PRONEX 015/98, bolsas de mestrado, doutorado e produtividade em pesquisa).

REFERÊNCIAS

1. Swierenga, H.; Weijer, A. P.; Wijk, R. J.; Buydens, L. M. C.; *Chemom. Intell. Lab. Syst*. **1999**, *49*, 1.
2. Feudale, R. N.; Woody, N. A.; Tan, H.; Myles, A. J.; Brown, S. D.; Ferré, J.; *Chemom. Intell. Lab. Syst.* **2002**, *64*, 181.
3. Wulferd, F.; Kok, W. T.; Noord, O. E.; Smilde, A. K.; *Anal. Chem.* **2000**, *72*, 1639.
4. Adhihetty, I. S.; Mcguire, J. A.; Wangmaneerat, B.; Niemczyk, T. M.; Haaland, D. M.; *Anal. Chem.* **1991**, *63*, 2329.
5. Despagne, F.; Massart, D. L.; Chabot, P.; *Anal. Chem.* **2000**, *72*, 1657.
6. Fearn, T.; *J. Near Infrared Spectrosc.* **2001**, *9*, 229.

7. Wang, Y.; Veltkamp, D. J.; Kowalski, B. R.; *Anal. Chem.* **1991**, *63*, 2750.
8. Valverde, R. S.; Garcia, M. D. G.; Galera, M. M.; Goicoechea, H. C.; *Anal. Chim. Acta* **2006**, *562*, 85.
9. Honorato, F. A.; Galvão, R. K. H.; Pimentel, M. F.; Neto, B. B.; Araújo, M. C. U.; Carvalho, F. R.; *Chemom. Intell. Lab. Syst.* **2005**, *76*, 65.
10. Barboza, F. D.; Poppi, R. J.; *Anal. Bioanal. Chem.* **2003**, *377*, 695.
11. Zhang, L.; Small, G. W.; Arnold, M. A.; *Anal. Chem.* **2003**, *75*, 5905.
12. Wang, Z. Y.; Dean, T.; Kowalsky, B. R.; *Anal. Chem.* **1995**, *67*, 2379.
13. Wang, Y.; Lysaght, M. J.; Kowalsky, B. R.; *Anal. Chem.* **1992**, *64*, 562.
14. Chen, C. S.; Brown, C. W.; Lo, S. C.; *Appl. Spectrosc.* **1997**, *51*, 744.
15. Gemperline, P. J.; Cho, J. H.; Aldridge, P. K.; Sekulic, S. S.; *Anal. Chem.* **1996**, *68*, 997.
16. Tan, H. W.; Brown, S. D.; *J. Chemom.* **2001**, *15*, 647.
17. Lima, F. S. G.; Borges, L. E. P.; *J. Near Infrared Spectrosc.* **2002**, *10*, 269.
18. Shenk, J. S.; Westerhaus, M. O.; Templeton Jr, W. C.; *Crop Sci.* **1985**, *25*, 159.
19. Shenk, J. S., Westerhaus M.O.; *US Pat. 4,866,644* **1989**.
20. Bouveresse, E.; Massart, D. L.; Dardenne, P.; *Anal. Chem.* **1995**, *67*, 1381.
21. Forina, M.; Casolino, C.; *Quim. Anal.* **1999**, *19*, 49.
22. Bouveresse, E.; Hartmann, C.; Massart, D. L.; Last, I. R.; Prebble, K. A.; *Anal. Chem.* **1996**, *68*, 982.
23. Noord, O. E.; *Chemom. Intell. Lab. Syst.* **1994**, *23*, 65.
24. Wulfert, F.; Kok, W. T.; Smilde, A. K.; *Anal. Chem.* **1998**, *70*, 1761.
25. Centner, V.; Verdu-Andres, J.; Walczak, B.; Jouan, R. D.; Despagne, F.; Pasti, L.; Poppi, R.; Massart, D. L.; Noord, O. E.; *Appl. Spectrosc.* **2000**, *54*, 608.
26. Beebe, K. R.; Pell, R. J.; Seasholtz, M. B.; *Chemometrics – A practical guide,* Wiley: New York, 1988.
27. Naes, T.: Isaksson, T.: Fearn, T.: Davies, T.; *A user-friendly guide to Multivariate Calibration and Classification*, NIR Publications: Chichester - UK, 2002.
28. Savitzky, A.; Golay, M. J. E.; *Anal. Chem.* **1964**, *36,* 1627.
29. Blank, T. B.; Sum, S. T.; Brown S. D.; Monfre, S. L.; *Anal. Chem.* **1996**, *68*, 2987.
30. Tan, H.; Sum, S. T.; Brown, S. D.; *Appl. Spectrosc.* **2002**, *56*, 1098.
31. Wold, S.; Antti, H.; Lindgren, F.; Ohman, J.; *Chemom. Intell. Lab. Syst.* **1998**, *44*, 175.
32. Fearn, T.; *Chemom. Intell. Lab. Syst.* **2000**, *50*, 47.
33. Westerhuis, J. A.; Jong, S.; Smilde, A. K.; *Chemom. Intell. Lab. Syst.* **2001**, *56*, 13.
34. Pierna, J. A. F.; Massart, D. L.; Noord, O. E.; Ricoux, Ph.; *Chemom. Intell. Lab. Syst.* **2001**, *55*, 101.
35. Sjoblöm, J.; Svensson, O.; Josefson, M.; Kullberg, H.; Wold, S.; *Chemom. Intell. Lab. Syst.* **1998**, *44*, 229.
36. Woody, N. A.; Feudale, N. R.; Myles, A. J.; Brown, S. D.; *Anal. Chem.* **2004**, *76*, 2595.
37. Leung, A. K.; Chau, F.; Gao, J.; *Chemom. Intell. Lab. Syst.* **1998**, *43*, 165.
38. Park, K. S.; Ko, Y. H.; Lee, H.; Jun, C. H.; Chung, H.; Ku, M. S.; *Chemom. Intell. Lab. Syst.* **2001**, *55*, 53.
39. Yoon, J.; Lee, B.; Han, C.; *Chemom. Intell. Lab. Syst.* **2002**, *64*, 1.
40. Tan, H. W.; Brown, S. D.; *J. Chemom.* **2002**, *16*, 228.
41. Andrew, A.; Fearn, T.; *Chemom. Intell. Lab. Syst.* **2004**, *72*, 51.
42. Swierenga, H.; Haanstra, W. G.; Weijer, A. P.; Buydens, L. M. C.; *Appl. Spectrosc.* **1998**, *52*, 7.
43. Hibbert, D. B.; *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277.
44. Efron, B.; *An Introduction to the Bootstrap (Hardcover),* Chapman & Hall/ CRC, 1994.
45. Araújo, M. C. U.; Saldanha, T. C. B.; Galvão, R. K. H.; Yoneyama, T.; Chame, H. C.; Visani, V.; *Chemom. Intell. Lab. Syst.* **2001**, *57*, 65.
46. Galvão, R. K. H.; Pimentel, M. F.; Yoneyama, T.; Visani, V.; *Anal. Chim. Acta* **2001**, *443*, 107.
47. Breitkreitz, M. C.; Raimundo Jr, I.; Rohwedder, J. J. R.; Pasquini, C.; Dantas Filho, H. A.; José, G. E.; Araújo, M. C. U.; *Analyst* **2003**, *128*, 1204.
48. Coelho, C. J.; Galvão, R. K. H.; Araújo, M. C. U.; Pimentel, M. F.; Silva, E. C.; *Chemom. Intell. Lab. Syst.* **2003**, *66*, 205.
49. Coelho, C. J.; Galvão, R. K. H.; Araújo, M. C. U.; Pimentel, M. F.; Silva, E. C.; *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 928.
50. Galvão, R. K. H.; José, G. E.; Dantas Filho, H. A.; Araújo, M. C. U.; Silva, E. C.; PaivA, H. M.; Saldanha, T. C. B.; Souza, E. S. O. N.; *Chemom. Intell. Lab. Syst.* **2004**, *70*, 1.
51. Swierenga, H.; Groot, P. J.; Weijer, A. P.; Derksen, M. W. J.; Buydens, L. M. C.; *Chemom. Intell. Lab. Syst.* **1998**, *41*, 237.
52. Bouveresse, E.; Massart, D. L.; *Chemom. Intell. Lab. Syst.* **1996**, *32*, 201.
53. Kennard, R. W.; Stone, L. A.; *Technometrics* **1969***,* 11.
54. Dantas Filho, H. A.; Galvão, R. K. H.; Araújo, M. C. U.; Silva, E. C.; Saldanha, T. C. B.; José, G. E.; Pasquini, C.; Raimundo Jr, I.; Rohwedder, J. J. R.; *Chemom. Intell. Lab. Syst.* **2004**, *72*, 83.
55. *Matlab 6.5 R 13 User's Guide*, The Mathworks, Natick, MA, USA, 2003.
56. Wise, B. M.; Gallagher, N. B.; Bro, R.; Shaver, J. M.; *PLS Toolbox 3.0 - Program for use with Matlab version 6.5*, Eingenvector Research, Manson, 2003.
57. *The Unscrambler 9.2*, CAMO ASA Nedre Vollgate 8, N-0158 OSLO, Norway, 2005.