# MODELAGEM QUIMIOMÉTRICA BASEADA EM JANELA MÓVEL DE REGRESSÃO POR MÍNIMOS QUADRÁTICOS PARCIAIS ASSOCIADA COM OTIMIZAÇÃO COLÔNIA DE FORMIGAS

C. RANZAN[1;2], A. STROHM[2], L. F. TRIERWEILER[1], B. HITZMANN[2] e J. O. TRIERWEILER[!]

[1] Universidade Federal do Rio Grande do Sul, Departamento de Engenharia Química
[2] Universität Hohenheim, Stuttgart
E-mail para contato: (Cassiano, Jorge)@enq.ufrgs.br

RESUMO – O principal objetivo da otimização de processos é a obtenção de alta produtividade e lucro em processos, sejam químicos ou bioquímicos. Para isso, técnicas de controle, altamente correlacionadas com nossas capacidades de caracterizar processos, devem ser aplicadas. Sensores ópticos associados com modelagem quimiométrica são considerados uma escolha natural para medidas não intrusivas e de alta sensibilidade. Este trabalho esta focado na caracterização de farinhas (operação usual e obrigatória na indústria de alimentos) através de um método proposto para monitoração usando dados de infravermelho próximo reflexivo (NIR). Dados de espectroscopia são avaliados através da combinação dos métodos de regressão de mínimos quadráticos parciais em janelas móveis e otimização de colônia de formigas. Resultados de predição de proteína são comparados com modelos PLSR padrão, apresentando uma melhora na capacidade preditiva de cerca de 30% usando CSMWPLS-ACO, confirmando a eficiência da proposta de caracterização e da estratégia quimiométrica.

## 1. INTRODUÇÃO

Atualmente, *soft-sensors* são amplamente aceitos e aplicados nas mais variadas atividades, sendo vistos como alternativas viáveis para a predição e caracterização de variáveis de processo. Apesar da grande quantidade de sensores online disponíveis, aqueles baseados em técnicas de regressão PLS (*Partial Least Squares*) ainda são os preferencialmente adotados, principalmente devido a suas vantagens na forma de tratar ruídos e correlação de variáveis, características usualmente presentes em dados industriais (Du, Liang et al. 2004, Brown 2013, Cariou, Verdun et al. 2014, Chi, Fei et al. 2014).

Regressão por mínimos quadráticos parciais é um método de calibração multivariável conhecido mundialmente. Este método possui particular aplicabilidade em análise espectral multicomponentes, fato este, que fez deste a escolha majoritária para estudo e desenvolvimento de sensores baseados em espectroscopia vibracional, como de Infravermelho (*Infrared - IR*), infravermelho próximo (*Near Infrared – NIR*), espectroscopia Raman ou espectroscopia de fluorescência (Hasegawa 2001, Du, Liang et al. 2004).

Apesar da aceita capacidade do PLS em lidar com problemas de calibração envolvendo dados de espectroscopia completos, a seleção ou filtragem de regiões espectrais ainda é um assunto de suma importância, já que seu impacto na capacidade preditiva de modelos é significativo. Esta característica é advinda principalmente da alta sensibilidade de métodos de espectroscopia à variações das características dos meios reacionais, e suas influências nas caracterizações das variáveis de estado (Xu and Schechter 1996, Sato, Kiguchi et al. 2004, Sratthaphut and Ruangwises 2012).

Com relação à seleção de regiões espectrais, Jiang *et al*. (2002) faz uma discussão detalhada sobre métodos para este fim, propondo um método chamado pelos autores de: Regressão por Mínimos Quadráticos Parciais com Janela Móvel (*Moving Window Partial Least Squares Regression* - MWPLSR). Este método busca por regiões espectrais informativas utilizando uma série de modelos do tipo PLS, com número de entradas fixas, segmentando o espaço espectral e mapeando cada região em função da predição da variável de interesse. Cada janela do espectro é avaliada utilizando a estratégia de validação cruzada (*Cross-Validation*).

Baseado no trabalho de Jiang *et al*. (2002), Du *et al*. (2004) propôs uma evolução ao método MWPLSR, introduzindo duas variações deste, para seleção espectral, o método de PLS com Janela Móvel e Tamanho Variável (*Changeable Size Moving Window PLS* - CSMWPLS) e o método de Busca Combinatória de Modelos PLS com Janela Móvel (*Searching Combination Moving Window PLS* - SCMWPLS).

Ambas as técnicas propostas por Du e colaboradores são combinações da rotina MWPLSR. No método CSMWPLS, é feita a varredura MWPLSR não apenas variando a posição da janela no espaço espectral, mas também variando o tamanho da própria janela, adicionando um novo grau de liberdade na modelagem. Já no SCMWPLS é feita a combinação dos melhores resultados obtidos para diversos tamanhos de janela, em regiões espectrais não subsequentes, possibilitando a combinação de regiões espectrais não sequenciais.

Apesar das vantagens inerentes à aplicação destas técnicas (como a obtenção de modelos PLS mais precisos em função do numero de variáveis independentes dos modelos, em comparação a modelos PLS aplicados diretamente a dados espectrais não filtrados) estas técnicas apresentam algumas desvantagens na forma de segmentar regiões espectrais quando mais de um tamanho de janela é avaliado conjuntamente. Além disto, uma vez que os núcleos destes métodos são modelos PLS aplicados em intervalos sequenciais de espectro, a seleção de regiões não subsequentes pode ser prejudicada pela estratégia da janela móvel, mascarando a combinação de regiões espectrais em avaliações inapropriadas de elementos espectrais, dentro das janelas.

Neste trabalho é proposta a resolução desta limitação do método de CSMWPLS através da utilização conjunta deste, com o método de busca e seleção de componentes espectrais puros para a predição de variáveis espectrais, denominado Modelagem Quimiométrica com Elementos Espectrais Puros (*Pure Spectral Chemometrical Modeling* - PSCM) e o algoritmo de otimização global baseado em colônia de formigas (*Ant Colony Optimization* – ACO).

PSCM é a estratégia de modelagem quimiométrica apresentada previamente por nosso grupo

(Ranzan, Strohm et al. 2014), onde regiões espectrais são classificadas e elementos espectrais selecionados para modelagem e predição de variáveis de processo. ACO é utilizado como ferramenta de busca para seleção do grupo de variáveis espectrais ideal para predição de estados. A utilização de ACO para esta função permite a caracterização qualitativa dos dados espectrais em função da variável de interesse, permitindo a segmentação dos dados espectrais através de dados qualitativos chamados de feromônios (alusão ao fenômeno real depositado no solo por formigas durante o processo de busca por alimentos).

A combinação do CSMWPLS com PSCM/ACO permite a filtragem de dados espectrais, de forma não sequencial, associada à modelagem da variável de interesse através de modelos PLS com tamanho de modelo e amplitude de dados espectrais otimizados em função da variável de interesse.

Para avaliação da combinação de metodologias de analise de dados espectrais, proposta neste trabalho, dados de espectroscopia de infravermelho próximo oriundos de amostras de farinha de centeio e trigo serão utilizados como caso de estudo. A escolha destes dados foi feita devido ao fato de espectroscopia NIR ser considerada uma ferramenta padrão para caracterização de amostras de farinha (Technology 1986, Technology 1995), de forma que, resultados oriundos de modelagem PLS com dados NIR são aplicados industrialmente, podendo ser considerados um padrão aceitável para comparação de resultados. Além disto, incrementos na capacidade preditiva de modelos PLS apresentam impacto direto na indústria alimentícia.

## 2. MATERIAIS E MÉTODOS

### 2.1. Conjunto de dados Experimental

O conjunto de dados experimental utilizado neste trabalho é composto por 34 amostras de farinha, dos tipos centeio e trigo, oriundas de diferentes marcas, coletadas no comercio dos arredores de Stuttgart, Alemanha. Medidas NIR das amostras foram coletadas em triplicata, totalizando 102 espectros. As amostras foram divididas aleatoriamente em dois grupos, um para calibração de modelos (2/3 do conjunto) e um para testes de modelos (1/3 do conjunto).

As medidas NIR foram realizadas no equipamento Multi-Purpose NIR Analyzer (Bruker Optik GmbH - Ettlingen, Alemanha), variando o comprimento de onda de 800nm até 2800nm, com incremento variável, totalizando 1150 comprimentos de onda independentes por espectro. Os dados espectrais foram normalizados com SNV (*Standard Normal Variate*).

A caracterização da concentração de proteínas das amostras foi feita laboratorialmente através de análise farinográfica com o aparato Brabender GmbH & Co. KG, Duisburg, Alemanha, modelo FD0234H. A variabilidade da concentração de proteínas no grupo amostral foi de 6 até 14 gramas de proteína por 100 gramas de amostra.

Análises de componente principal (PCA) foram realizadas com o objetivo de avaliar qualitativamente o conjunto de dados espectral. Os resultados desta análise qualitativa mostraram que os dados apresentam semelhanças entre as amostras, além da segmentação destas nos referidos

grupos, indicando o agrupamento entre amostras de farinha de trigo e centeio, através dos primeiros cinco componentes principais.

## 2.2. Métodos Quimiométricos

Todas as rotinas de cálculo foram implementadas e desenvolvidas no software MATLAB (Ver. 5.3.0.10183 R11, The Mathworks, Inc., Natick, USA). A avaliação da capacidade preditiva dos modelos foi feita através do índice estatístico raiz quadrada do erro médio (*Root Mean Square Error* - RMSE) da etapa de teste dos modelos, de forma que os dados referentes à etapa de calibração não são levados em consideração para comparação entre os modelos.

A análise de dados espectrais para predição de variáveis de estado, proposta neste trabalho, é composta pela combinação das funcionalidades entre as metodologia CSMWPLS e PSCM/ACO. Como ambas estas metodologias apresentam a possibilidade de serem aplicadas para filtragem de dados ou modelagem de processos, neste trabalho é avaliada a utilização conjunta destas técnicas, onde a metodologia PSCM/ACO é utilizada para pré-filtragem de dados espectrais e a metodologia CSMWPLS para modelagem do sistema em questão.

CSMWPLS: O método quimiométrico PLS móvel com janela variável trata do método de varredura PLS com janela móvel expandido para avaliar um intervalo de tamanhos de janela específicos, de forma a selecionar qual o melhor modelo PLS obtido no espaço espectral, em função da posição e tamanho da janela.

O tamanho da janela referida no método diz respeito a quantidade de elementos espectrais agrupados em ordem sequencial de comprimentos de onda e levados em consideração na modelagem do referido sistema (Du, Liang et al. 2004). Cada janela é considerada um subgrupo do conjunto dos dados original e utilizada na predição da variável de estado. O número de vetores de carga (*load vector* - LV) dos modelos é avaliado por varredura, sendo utilizado como variável independente para comparação de metodologias.

Os modelos obtidos por CSMWPLS são, desta forma, caracterizados em função do número de LV's utilizado nos modelos, determinando o melhor tamanho de janela e a posição mais indicada da janela, para cada tamanho de modelo.

PSCM/ACO: Esta metodologia de análise quimiométrica trata da combinação de elementos espectrais puros para a predição de variáveis de estado (Ranzan, Strohm et al. 2014). O centro desta estratégia é a forma de seleção do grupo de elementos espectrais dentro do universo de possíveis grupos. Na referida metodologia o algoritmo de otimização de colônia de formigas (ACO) é utilizado para a seleção do grupo ótimo de elementos espectrais.

A modelagem PSCM pode ser dividida em três fases: seleção do grupo de elementos espectrais, calibração do modelo e por fim, teste do mesmo. A seleção do grupo de elementos espectrais busca o melhor grupo para a predição das variáveis de estado com dados unicamente espectrais, através de modelos do tipo MISO (*Multiple Input Single Output*) ou então MIMO (*Multiple Input Multiple*

*Output*) (Skoog, Holler *et a*l. 2007). Estes modelos são lineares em relação aos parâmetros, permitindo a estimação de parâmetros utilizando mínimos quadráticos ordinários (Joe Qin 1998).

ACO é baseado no comportamento real de formigas, mais especificamente, na forma com que estes indivíduos se comunicam utilizando secreção de feromônios (Dorigo and Blum 2005, Dorigo, Birattari *et al.* 2006, Mullen, Monekosso *et al.* 2009). A ideia principal deste algoritmo é que a convergência de trilha de menor distancia entre o ninho e a fonte de alimentos é caracterizada pela maior concentração de feromônios (Deneubourg, Aron et al. 1986). Detalhes mais aprofundados sobre esta metodologia e sua implementação podem ser obtidos em Mullen *et al*. (2009) e Ranzan *et al* (2014).

A grande vantagem associada à esta metodologia é a capacidade do método de caracterizar o conjunto de dados espectrais de forma qualitativa através do vetor da trilha de feromônios (criado no decorrer da otimização). Este vetor funciona como fonte de informação numérica para que as formigas possam construir a solução probabilística do problema, e também pode ser usado como fonte parâmetro para filtragem e seleção de regiões de espectroscopia, como demonstrado em Ranzan et al (Ranzan, Strohm et al. 2014).

## 3. RESULTADOS

Inicialmente, o conjunto de dados experimental completo, normalizado usando SNV, foi submetido à modelagem PLS, de forma a gerar a base de resultados padrão. A Figura 1 apresenta os resultados de variância explicada apresentada para os vinte primeiros LV's, além dos resultados de RMSE para a predição (RMSEP), em função do número de LV's utilizados nos modelos ajustados. Esta análise inicial foi realizada sem nenhum tipo de pré-tratamento dos dados espectrais, além da normalização SNV, de forma que os dados espectrais estão da forma que foram coletados, sem restrição de regiões de espectroscopia.

Resultados obtidos por analise PLS em dados NIR completos indicam que a melhor configuração de modelos, para predição de concentração de proteína em amostras de farinha, é de 11 a 13 LV's, uma vez que modelos desta dimensão apresentaram os menores valores de RMSEP, assim como variância explicada acumulada entorno de 95% do total apresentado pelos dados de espectroscopia. A análise de variância acumulada foi realizada com as 102 medidas NIR, enquanto que a modelagem PLS foi realizada com 2/3 deste total e o 1/3 restante foi utilizado para os testes.

Após a modelagem PLS padrão, foi iniciado o processo de filtragem e tratamento dos dados NIR através da metodologia PSCM/ACO. Nesta etapa, modelos MISO com número de variáveis de entrada variando de um a nove foram ajustados pela rotina de otimização. Os modelos gerados não são avaliados na predição do conjunto de dados de teste, uma vez que este resultado não é de interesse, sendo apenas a informação qualitativa armazenada na forma do vetor trilha de feromônios analisada.

No decorrer de cada rotina de otimização, um novo resultado de concentração de feromônios é coletado, associado ao referido número de variáveis de entrada utilizado nos modelos ajustados.

Como os vetores possuem ordens de grandeza distintas, eles são normalizados de acordo com o valor máximo em cada um, e é feita a média de feromônios depositadas em cada elemento espectral. O resultado da média de feromônio atribuída a cada elemento é considerado a assinatura da variável de estado no plano espectral avaliado, sendo esta informação utilizada para a filtragem dos dados espectrais.

A Figura 2 apresenta a assinatura de feromônios obtida para o conjunto de dados experimentais. Nesta figura, são salientadas as regiões espectrais que apresentam maior correlação com a concentração de proteína em amostras de farinha.

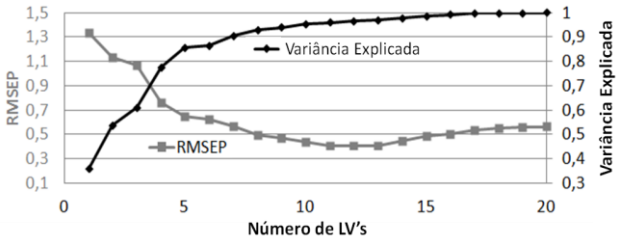


Figura 1 – Resultados de RMSEP e variância explicada para predição de proteína em amostras de farinha usando modelos PLS aplicados ao conjunto de dados NIR completos.

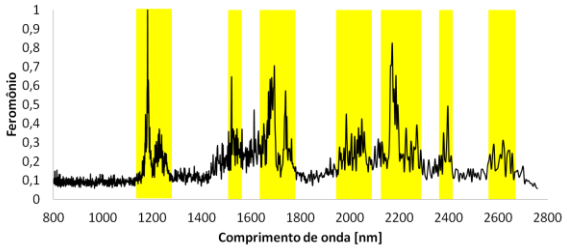


Figura 2 – Trilha de feromônios média associada à predição de dados de proteína utilizando medidas NIR em amostras de farinha de trigo e centeio.

As regiões de espectroscopia NIR salientadas pela metodologia PSCM/ACO, e graficadas na Figura 2, podem ser relacionadas com a teoria de sobretons de espectroscopia. De acordo com os trabalhos de Sun (2008) e Champe e Harvey (2005), algumas das regiões salientas pela metodologia PSCM/ACO são correlacionadas com ligações C-H (sobretons em 1200nm e 1800nm), ou bandas de amidos (2100nm), resultados esperados para medidas de NIR de moléculas de proteína. Resultado esperado para a matriz de compostos das amostras de farinha.

Os elementos espectrais contidos nas regiões salientadas são ordenados em ordem decrescente de feromônios associados e filtrados de forma a representarem 20% do total de elementos espectrais totais, selecionando assim, 230 elementos do total de 1150. A ordem de filtragem empregada foi escolhida baseada em estudos preliminares para este conjunto de dados (Ranzan, Strohm et al. 2014).

Partindo do conjunto de dados NIR, filtrados e otimizados por PSCM/ACO para a predição da concentração de proteína em amostras de farinha, é feita a modelagem do referido sistema utilizando a metodologia CSMWPLS. Nesta etapa, o método busca as melhores configurações (etapa de calibração) dos modelos PLS, fixando o tamanho dos modelos de um até 20 LV's, determinando a melhor combinação entre tamanho de janela e posição da mesma para predição da variável de estado. Com os modelos estruturados, é feito o teste dos mesmos (no conjunto de amostras de teste) e os respectivos resultados são apresentados na Figura 3.

Os resultados apresentados na Figura 3 mostram que para modelos com até 8 LV's, os dados NIR filtrados com PSCM/ACO e modelados com CSMWPLS apresentaram melhores resultados para

predição do grupo de amostras de testes, com incremento máximo alcançado de 30% para os modelos de 2 e 3 LV's (Figura 3b), em comparação aos modelos PLS com dados espectrais originais.

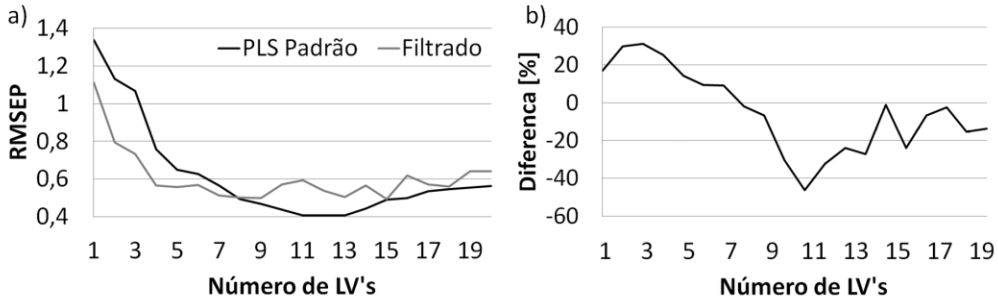


Figura 3 – (a) Resultados de RMSEP usando modelos PLS padrão e CSMWPLS com dados espectrais filtrados. (b) Diferença percentual entre os resultados dos modelos CSMWPLS/PSCM em relação aos modelos PLS padrão.

Para modelos com mais de 8LV's, os resultados de predição da variável de estado foram mais precisos, apresentando valores mínimos de RMSEP iguais a 0,4, para os modelos de 11 a 13 LV's, indicando que a informação útil presente na matriz de dados espectrais foi extraída e concentrada nos primeiros 8 LV's, entretanto, alguma informação significativa não foi captada e retirada desta matriz pelo processo de filtragem e modelagem propostos, fazendo com que modelos usando todos os elementos disponíveis apresentassem melhores predições.

## 4. CONCLUSÕES

Ambas as metodologias quimiométricas CSMWPLS e PSCM apresentam a funcionalidade de serem empregadas para a filtragem e/ou modelagem de dados de espectroscopia para a predição de variáveis de estado em processos. Desta forma, a utilização conjunta destas técnicas abrange um universo de possibilidades elevado na quimiometria, uma vez que ambas possuem diversos parâmetros de ajuste que devem ser determinados pelo usuário.

Neste universo de possibilidades, os resultados apresentados para caracterização de farinhas, utilizando dados de espectroscopia NIR, mostram se promissores em comparação com métodos quimiométricos padrões (PLS).

A modelagem quimiométrica testada é composta pela filtragem de dados utilizando o método PSCM/ACO (descartando 80% dos dados espectroscópicos), sendo o processo caracterizado através de modelos PLS ajustados por CSMWPLS. Para modelos com baixos números de LV's, os dados filtrados apresentaram melhora em relação aos modelos PLS padrão, atingindo incremento máximo de 30% para modelos com 3 LV's. Apesar disso, modelos com número de LV's superiores à 8 apresentaram redução na capacidade preditiva, indicando que o processo de filtragem descartou informação significativa do processo, e aprimoramentos na metodologia ainda devem ser feitos.

## 5. AGRADECIMENTOS

Os autores são gratos à CAPES, CNPQ, DAAD e FAPERGS pelo auxilio financeiro.

## 6. REFERÊNCIAS

Brown, S. D. (2013). Transfer of Multivariate Calibration Models. Reference Module in Chemistry, Molecular Sciences and Chemical Engineering, Elsevier.

Cariou, V., S. Verdun and E. M. Qannari (2014). "Quadratic PLS regression applied to external preference mapping." Food Quality and Preference **32, Part A**(0): 28-34.

Champe, P. C. and R. A. Harvey (2005). Biochemistry. Philadelphia, Lippincott/Williams & Wilkins.

Chi, Q., Z. Fei, Z. Zhao, L. Zhao and J. Liang (2014). "A model predictive control approach with relevant identification in dynamic PLS framework." Control Engineering Practice **22**(0): 181-193.

Deneubourg, J. L., S. Aron, S. Goss, J. M. Pasteels and G. Duerinck (1986). "Random behaviour, amplification processes and number of participants: How they contribute to the foraging properties of ants." Physica D: Nonlinear Phenomena **22**(1–3): 176-186.

Dorigo, M., M. Birattari and T. Stuetzle (2006). "Ant colony optimization - Artificial ants as a computational intelligence technique." Ieee Computational Intelligence Magazine **1**(4): 28-39.

Dorigo, M. and C. Blum (2005). "Ant colony optimization theory: A survey." Theoretical Computer Science **344**(2-3): 243-278.

Du, Y. P., Y. Z. Liang, J. H. Jiang, R. J. Berry and Y. Ozaki (2004). "Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares." Analytica Chimica Acta **501**(2): 183-191.

Hasegawa, T. (2001). Handbook of Vibrational Spectroscopy. J. Chalmers and P. R. Griffiths. Chichester, UK, Wiley**:** 2293.

Jiang, J.-H., R. J. Berry, H. W. Siesler and Y. Ozaki (2002). "Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window Partial Least-Squares Regression with Applications to Mid-Infrared and Near-Infrared Spectroscopic Data " Anal. Chem. **74**: 3555 - 3565.

Joe Qin, S. (1998). "Recursive PLS algorithms for adaptive data modeling." Computers & Chemical Engineering **22**(4–5): 503-514.

Mullen, R. J., D. Monekosso, S. Barman and P. Remagnino (2009). "A review of ant algorithms." Expert Systems with Applications **36**(6): 9608-9617.

Ranzan, C., A. Strohm, L. Ranzan, L. F. Trierweiler, B. Hitzmann and J. O. Trierweiler (2014). "Wheat flour characterization using NIR and spectral filter based on Ant Colony Optimization." Chemometrics and Intelligent Laboratory Systems **132**(0): 133-140.

Sato, H., M. Kiguchi, F. Kawaguchi and A. Maki (2004). "Practicality of wavelength selection to improve signal-to-noise ratio in near-infrared spectroscopy." NeuroImage **21**(4): 1554-1562.

Skoog, D. A., F. J. Holler and S. R. Crouch (2007). Principles of instrumental analysis. Belmont, CA, Thomson Brooks/Cole.

Sratthaphut, L. and N. Ruangwises (2012). "Genetic Algorithms-Based Approach for Wavelength Selection in Spectrophotometric Determination of Vitamin B12 in Pharmaceutical Tablets by Partial Least-Squares." Procedia Engineering **32**(0): 225-231.

Sun, D.-W. (2008). Infrared Spectroscopy for Food Quality Analysis and Control, Elsevier.

Technology, I.-I. A. f. C. S. a. (1986). Procedure for near infrared (NIR) reflectance analysis of ground wheat and milled wheat products. **202**.

Technology, I.-I. A. f. C. S. a. (1995). Determination of Protein by Near Infrared Reflectance (NIR) Spectroscopy. **159**.

**Xu, L.** and I. **Schechter** (1996). "Wavelength selection for simultaneous spectroscopic analysis. Experimental and theoretical study." **Anal Chem 68**: 2392–2400.