UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

TESE DE DOUTORADO

UMA NOVA TÉCNICA PARA SELEÇÃO DE VARIÁVEIS EM CALIBRAÇÃO MULTIVARIADA APLICADA ÀS ESPECTROMETRIAS UV-VIS E NIR

PEDRO GERMANO ANTONINO NUNES

João Pessoa – PB – Brasil

Março/2008

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

TESE DE DOUTORADO

UMA NOVA TÉCNICA PARA SELEÇÃO DE VARIÁVEIS EM CALIBRAÇÃO MULTIVARIADA APLICADA ÀS ESPECTROMETRIAS UV-VIS E NIR

PEDRO GERMANO ANTONINO NUNES

Tese apresentada como parte dos requisitos para obtenção do título de Doutor em Química pela Universidade Federal da Paraíba

Orientador: Prof. Dr. Edvan Cirino da Silva

2º Orientador: Dr. Wallace Duarte Fragoso

João Pessoa – PB – Brasil

Março/2008

N972u Nunes, Pedro Germano Antonino

Uma nova técnica para seleção de variáveis em calibração multivariada aplicada às espectrometrias UV-VIS e NIR/ Pedro Germano Antonino Nunes – João Pessoa, 2008.

106p.

Orientador: Prof. Dr. Edvan Cirino da Silva

Co-orientador: Dr. Wallace Duarte Fragoso

Tese (Doutorado) – UFPB/CCEN

1. Química Analítica. 2. Seleção de variáveis. 3. ASA. 4. Calibração multivariada. 5. Espectrometrias UV-VIS e NIR.

UFPB/BC 543(043)

A Deus

A minha família

Agradecimentos

- Aos meus orientadores, professores Edvan Cirino e Wallace Fragoso, pelas profícuas discussões e sugestões;

- Ao professor Mario Ugulino, por todo apoio, indispensável para a realização deste trabalho;

- Ao professor Roberto Kawakami (ITA) e ao colega Sófacles, pela colaboração imprescindível na programação do algoritmo;

- Aos professores e a Coordenação do Programa de Pós-Graduação em Química do DQ/CCEN/UFPB;

- Ao Departamento de Tecnologia Rural do Centro de Formação de Tecnólogos, pela liberação para realizar esta qualificação;

- Ao gerente de laboratório Silvio Rogério, da Empresa Guaraves Alimentos, pela presteza no fornecimento de amostras;

- Aos colegas Laqueanos, pela amizade, companheirismo, momentos de alegria: Elaine, Simone, Sérgio, Chicão, Márcio, Prof[a]. Teresa, Glédson, Ilanna, Alessandra, Glauciene, Wellington, Socorro, Sueny, Jô, Valdomiro, Urijatan, Germano, Luciano, Karina, Chicote, Cícero, Xande, Gaião,Rose, Osmundo, Amália, Aline, Mônica, Valmir, Renato, Fátima, Paulinho, Williames, Adamastor, Aline (IC), enfim, a todos que estiveram comigo nesta caminhada.

SUMÁRIO

CAPÍTULO 2. O ALGORITMO DE BUSCA ANGULAR

CAPÍTULO 3. EXPERIMENTAL

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ASA	Algoritmo de busca angular
MLR	Regressão linear múltipla
APS	Algoritmo das projeções sucessivas
AG	Algoritmo genético
SW	Stepwise
VIF	Fator de inflação da variância
UV-VIS	Ultravioleta e Visível
PLS	Regressão por mínimos quadrados parciais
PCR	Regressão em Componentes Principais
MLR-ASA-VIF	Modelo de calibração obtido com as variáveis selecionadas pelo algoritmo ASA-VIF
MLR-ASA	Modelo de calibração obtido com as variáveis selecionadas pelo algoritmo ASA
MLR-ASA-APS	Modelo de calibração obtido com as variáveis selecionadas pelo algoritmo APS
MLR-ASA-AG	Modelo de calibração obtido com as variáveis selecionadas pelo algoritmo AG
MLR-ASA-SW	Modelo de calibração obtido com as variáveis selecionadas pelo algoritmo SW
$RMSEP_{val}$	Raiz quadrada do erro médio quadrático de previsão para o conjunto de validação
$RMSEP_{prev}$	Raiz quadrada do erro médio quadrático de previsão para o conjunto de previsão
NIR	Infravermelho próximo
min-max	Procedimento mínimo-máximo
MON	Número de octanagem do motor
T90%	Temperatura correspondente a 90% de evaporados da gasolina

RESUMO

Este trabalho propõe o Algoritmo de Busca Angular (*Angular Search Algorithm*-ASA) como uma nova técnica para seleção de variáveis em calibração multivariada. Este algoritmo foi concebido para minimizar problemas de correlação e multicolinearidade em regressão linear múltipla (*Multiple Linear Regression*-MLR). Para isso, o ASA utiliza o conceito de produto interno para encontrar o cosseno dos ângulos intervetores definidos no espaço multidimensional das amostras. Um procedimento min-max é proposto e aplicado aos cossenos dos ângulos para gerar as cadeias contendo as variáveis com menor correlação par a par. Em seguida, o ASA utiliza a Análise de Inflação de Variância (*Variance Inflation Factors*-VIF) como critério para minimizar multicolinearidade entre as variáveis menos correlacionadas. O desempenho do ASA é avaliado por meio de quatro estudos de casos. O primeiro consiste de um conjunto de dados de misturas de quatro corantes alimentícios sintéticos (tartrazina, vermelho 40, amarelo crepúsculo e eritrosina), obtidos por espectrofotometria UV-VIS. Os outros três casos envolvem a utilização de dados de espectrometria NIR (*Near InfraRed*). Um dos conjuntos abrange amostras de trigo, cujas propriedades medidas foram proteína e umidade. O outro envolve amostras de milho onde foram determinados os teores de proteína, umidade, óleo e amido. No quarto, foram determinados MON (*Motor Octane Number*) e T90% (temperatura correspondente a 90% de evaporados) em amostras de gasolina. Os modelos MLR-ASA e MLR-ASA-VIF obtidos são comparados, em termos do menor erro médio quadrático de previsão para um conjunto independente de amostras, com os modelos MLR baseados em variáveis selecionadas pelos métodos: Stepwise, Algoritmo Genético e o Algoritmo de Projeções Sucessivas. Além disso, modelos PLS (*Partial Least Squares*), baseados nos espectros completos, também foram usados na comparação. Os desempenhos de predição dos modelos MLR-ASA e MLR-ASA-VIF foram geralmente similares ou ligeiramente melhores que os dos outros modelos de calibração. Contudo, os modelos MLR-ASA-VIF foram geralmente mais parcimoniosos (menor número de variáveis selecionadas). Portanto, o algoritmo proposto pode ser considerado uma ferramenta potencialmente útil para seleção de variáveis em calibração MLR, especialmente em análises espectrométricas UV-VIS e NIR.

Palavras-chave: Seleção de variáveis, ASA, calibração multivariada, Espectrometrias UV-VIS e NIR

ABSTRACT

The present work proposes the angular search algorithm (ASA) as a novel technique for variable selection in multivariate calibration based on multiple linear regression (MLR). The proposed algorithm was designed in order to minimize problems of correlation and multicollinearity in MLR calibration. For this purpose, ASA uses the concept of inner product to find the cosine of the angles between vectors defined in the multi-dimensional sample space. A min-max procedure is proposed and applied to the cosine of the angles to generate the chains containing the variables with the smallest pairwise correlation. Thereafter, ASA employs the Variance Inflation Factors (VIF) as criterion to minimize the multicollinearity problems between the variables less correlated. The efficiency of the proposed algorithm is illustrated in four case studies. The first consists of UV-VIS determination of four synthetic food colorants (tartrazine, allure red, sunset yellow and erythrosine) in aqueous solutions. The other cases concerning to NIR determinations of: protein and moisture in samples of wheat, protein, moisture, oil and starch in corn and MON (Motor Octane Number) and T90% (temperature at which 90% of the sample has evaporated) in gasoline. In all applications, the prediction abilities of MLR-ASA and MLR-ASA-VIF models are compared, in terms of the root-mean-square error obtained in an independent set, with those obtained by MLR models based on variables selected by: Stepwise, Genetic algorithm and Successive Projections Algorithm. Moreover, PLS (Partial Least Squares) models, based on full-spectrum, are also employed in comparison. The prediction performances of the MLR-ASA and MLR-ASA-VIF models were similar or slightly better than those obtained by other calibration models. However, MLR-ASA-VIF models are generally more parsimonious (a smaller number of selected variables). Therefore, proposed algorithm may be considered as a potentially useful tool for variable selection in MLR calibration, especially in UV-VIS and NIR spectrometric analyses.

Keywords: Variable selection, Multivariate calibration, ASA, UV-VIS and NIR spectrometric

CAPÍTULO 1

INTRODUÇÃO

1. INTRODUÇÃO

1.1 Caracterização geral da problemática

O sinal analítico fornecido por um instrumento de medida dificilmente poderá fornecer diretamente a informação quantitativa da espécie química de interesse (analito). A necessidade de tratar adequadamente os dados adquiridos para extrair a informação desejada tem levado ao desenvolvimento de novos procedimentos de um ramo da química analítica denominado Quimiometria. Esta pode ser definida como a área da Química que se propõe a fazer "Utilização de técnicas estatísticas e matemáticas para analisar dados químicos" (BEEBE et al, 1998). A quimiometria abrange vários tópicos, tais como, planejamento e otimização experimental, calibração univariada e calibração multivariada, processamento de sinais, reconhecimento de padrões, etc. A utilização dos métodos quimiométricos permite, entre outras coisas, a identificação de amostras, a análise de misturas complexas sem a necessidade de separações prévias, a possibilidade de determinar simultaneamente vários analitos, etc.

Em geral, o sinal analítico está relacionado a uma propriedade (óptica, elétrica, etc) do analito. Sendo assim, os métodos instrumentais são relativos, ou seja, para se determinar a quantidade de um analito presente em uma amostra é necessário comparar a propriedade medida com a de um conjunto de padrões de composição conhecida. Esse procedimento é conhecido como calibração.

A calibração é definida como o processo que permite estabelecer a relação entre a resposta instrumental (sinal analítico) e uma determinada propriedade da amostra (concentração do analito, por exemplo). A equação matemática que descreve a relação é denominada modelo de calibração e a representação gráfica é denominada curva analítica ou de calibração.

Nas últimas décadas, muita ênfase tem sido dada à calibração multivariada, na qual são realizadas medidas associadas a muitas variáveis simultaneamente ao se analisar uma dada amostra. Nesses sistemas, a conversão da resposta instrumental na informação química de interesse requer a utilização de técnicas de calibração multivariada. Essas técnicas se constituem no momento na melhor alternativa para a interpretação de dados e para a aquisição do máximo de informação sobre o sistema analítico.

A calibração multivariada tornou-se uma ferramenta analítica importante em diferentes campos de aplicação, especialmente nas análises de alimentos, farmacêutica, agricultura, ambiental e química industrial. É aplicada tanto para a determinação de espécies químicas quanto físicas. A razão do grande interesse na calibração multivariada é que o procedimento analítico é rápido e eficiente na abordagem de muitos problemas reais (FORINA, 2007). O processo de calibração pode ser realizado por intermédio de diversos métodos quimiométricos, utilizando toda a informação instrumental (métodos baseados em espectro completo) ou utilizando métodos de seleção de variáveis. Não obstante, o desempenho da calibração multivariada pode ser melhorado significativamente quando se efetua uma seleção de variáveis.

No contexto dos métodos espectroanalíticos, a seleção de variáveis envolve a escolha de uma determinada região do espectro, que podem ser comprimentos de ondas discretos ou faixas. Várias razões práticas corroboram para se optar por uma seleção de variáveis. Por exemplo, algumas regiões espectrais podem não apresentar relação com o parâmetro de interesse, os espectros podem conter ruídos heteroscedásticos, pode haver comprimentos de onda em que a intensidade do sinal não é linearmente correlacionada ao parâmetro de interesse, bem como pode haver comprimentos de onda correlacionados (colineares).

Os métodos de seleção de variáveis buscam encontrar as variáveis minimamente correlacionadas e colineares que contenham informações relacionadas ao parâmetro de interesse. Assim, a modelagem pode ser realizada com base nessas variáveis a fim de construir modelos mais simples, robustos, eficientes e fáceis de interpretar.

1.2 Apresentação e objetivos do trabalho

Neste trabalho, propõe-se uma nova técnica de seleção de variáveis para calibração multivariada baseada em regressão linear múltipla (MLR). Denominada algoritmo de busca angular (*Angular Search Algorithm* - ASA), esta técnica busca a minimização da correlação e da multicolinearidade entre as variáveis da matriz de respostas instrumentais. Para minimizar a correlação, o ASA calcula os ângulos (ou melhor, os cossenos dos ângulos) entre os vetores associados às variáveis definidas no espaço das amostras. Depois, um procedimento mínimo-máximo é então aplicado para construir cadeias de variáveis minimamente correlacionadas. Num

conjunto de dados centrados na média (comum nos métodos de regressão multivariada e também utilizado no ASA), o valor do cosseno é, matematicamente, igual a correlação entre duas variáveis (DRAPER e SMITH, 1998). No entanto, as variáveis minimamente correlacionadas podem ainda apresentar uma significativa multicolinearidade entre si. Para superar esse inconveniente, o ASA descarta as variáveis mais colineares por intermédio do *Fator de Inflação da Variância* (VIF, do inglês *Variance Inflation Factors*).

O VIF fornece uma medida da multicolinearidade entre as variáveis, selecionando as que apresentem valores abaixo de um determinado limiar (neste trabalho, adotou-se os limiares VIF < 5, 10, 30 e 50. Sendo assim, a combinação ASA-VIF possibilita a seleção um conjunto de variáveis minimamente correlacionadas e multicolineares. Além disso, essa associação permite selecionar as variáveis mais informativas com relação à propriedade de interesse (matriz **Y**) por intermédio do menor erro médio quadrático de previsão para um conjunto independe de amostras (validação externa).

A eficiência do ASA é ilustrada mediante quatro estudos de caso envolvendo duas diferentes técnicas espectroanalíticas. O primeiro caso se refere à determinação espectrofotométrica UV-VIS de uma mistura de corantes alimentícios sintéticos (tartrazina, vermelho 40, amarelo crepúsculo e eritrosina). Esse problema analítico enfatiza a habilidade do ASA ao lidar com forte sobreposição espectral causada pelas bandas largas que normalmente ocorrem nessa região. Efeitos de alta multicolinearidade estão presentes também nesses dados, proveniente da semelhança dos espectros dos corantes vermelho 40 e amarelo crepúsculo. A mistura dos corantes foi realizada através de um sistema automático de análise fluxo-batelada.

Os outros três casos envolvem a utilização de dados de espectroscopia de infravermelho próximo (*Near InfraRed* - NIR). Um conjunto de amostras de trigo, cujas propriedades medidas foram proteína e umidade, um conjunto de amostras de milho, onde se determinou os teores de proteína, umidade, óleo e amido, e um conjunto de gasolina, que envolve a determinação de duas propriedades: MON (*Motor Octane Number*) e T90% (temperatura correspondente a 90% de evaporados). Estes exemplos ilustram a aplicação do ASA à analise de amostras com matriz complexa.

Nas aplicações mencionadas, os modelos MLR-ASA são comparados com os modelos MLR obtidos a partir das variáveis selecionadas pelos outros três métodos de seleção de variáveis: Stepwise (SW), Algorítimo Genético (GA) e o Algoritmo de Projeções Sucessivas (APS). Além disso, os resultados são também comparados com os produzidos pelos modelos de Regressão por Mínimos Quadrados Parciais (PLS), baseados nos espectros completos.

1.3 CALIBRAÇÃO MULTIVARIADA

Os métodos de calibração podem ser classificados de acordo com os critérios apresentados na **Tabela 1.1** (SABOYA, 2002).

Tabela 1.1 – Critérios para classificação dos métodos de calibração

Critério	Método de calibração
Dependendo do número de variáveis	Univariado Multivariado
Dependendo do tipo de função matemática	Linear Não linear
Dependendo da obtenção dos parâmetros de calibração	Direta Indireta
Dependendo de qual é a variável independente	Clássica Inversa

Na calibração **univariada**, é estabelecida uma relação matemática entre uma única variável dependente e uma única variável independente. Quando a relação é entre mais de uma variável denomina-se calibração **multivariada**.

Os modelos de calibração também podem ser **lineares** quando relacionam as variáveis dependentes com funções lineares das variáveis independentes, ou seja, com funções polinomiais que são lineares nos coeficientes. Quando as funções não são deste tipo, os modelos são denominados **não lineares**. Quando os parâmetros de calibração são conhecidos diretamente a partir do sinal analítico de cada um dos analitos de forma individual, a calibração é **direta**. Quando os parâmetros são determinados a partir dos sinais analíticos de misturas de componentes, a calibração é **indireta**. Na calibração **clássica** a variável independente é a concentração e a variável dependente é o sinal analítico. No caso contrário, a calibração é inversa.

Também se distinguem entre métodos de **espectro completo** onde se utilizam todos os comprimentos de onda do espectro, ou de **seleção de variáveis**, cujas variáveis são selecionadas previamente. Dentro dos métodos de espectro

completo encontram-se os métodos de compressão de variáveis, baseados na decomposição dos dados em componentes principais (MARTENS e NAES, 1987).

A calibração multivariada é realizada, de forma geral, por intermédio das etapas mostradas abaixo:

- ***Preparação do conjunto de calibração***. Obtenção de um conjunto de amostras das quais se conheça a propriedade de interesse (obtida por um método de referência) e que seja representativo para realizar futuras previsões. Este conjunto deve ser representativo tanto nas fontes de variação química, quanto nas fontes de variação físicas (tamanho de amostra, granulometria, cristalização, etc.);

- ***Registro do sinal analítico***. A informação pode ser obtida de várias fontes. No caso dos métodos espectrométricos, o registro é chamado de espectro. A partir destes sinais são obtidas as informações químicas e/ou físicas desejadas;

- ***Pré-tratamento dos dados***. Nesta etapa, são minimizadas as possíveis contribuições não desejadas dos sinais, que diminuem a capacidade de previsão dos modelos. Um dos pré-processamentos mais utilizado é o método da derivada (PIZARRO et al., 2004). A primeira derivada remove os termos constantes (*offsets*) a todos os comprimentos de onda do espectro, ou seja, o deslocamento da linha de base. A segunda derivada elimina os termos que variam linearmente com a linha de base, normalmente devidos a efeitos de espalhamento. A maior limitação do uso dos métodos derivativos é a diminuição na relação sinal/ruído que se produz quando aumenta o grau da derivada, devido a alta sensibilidade do método a presença de ruído no espectro original. 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 (SAVITZKY e GOLAY, 1964).

- ***Construção do modelo***. Seleção do modelo que melhor estabelece a relação entre o sinal instrumental e a propriedade desejada;

- ***Validação do modelo***. Para assegurar a capacidade preditiva de um modelo é necessário realizar um processo de validação do mesmo, que consiste no estudo quantitativo dos resultados da aplicação do modelo em novas amostras (que não fizeram parte da etapa de calibração). Esta validação pode ser um processo externo, utilizando um conjunto de amostras independentes das utilizadas na calibração, mas que sejam representativas das mesmas e das futuras amostras a analisar, denominado conjunto de validação. A concentração deste conjunto de

validação deve ser conhecida para se verificar a relação entre estas e aquelas previstas pelo modelo.

A validação também pode ser feita por um processo interno ou validação cruzada (cross-validation). Por este método, divide-se o conjunto de calibração em vários segmentos. Um dos seguimentos é utilizado para validação e os outros para construir o modelo de calibração. São construídos modelos de acordo com o número de segmentos utilizados, de maneira que cada segmento seja excluído do modelo de calibração e utilizado na validação. O modelo final será aquele que apresentar o menor resíduo de concentração.

Para avaliar a capacidade de previsão utiliza-se o somatório do quadrado dos resíduos ($\Sigma(y_{prev} - y)^2$) denominado habitualmente de PRESS (*Predicted Residual Error Sum of Squares*) ou seu valor médio, MSE (*Mean Squared Error*), obtido dividindo-se o PRESS pelo número de amostras. Para se trabalhar nas mesmas unidades de concentração calcula-se a raiz quadrada do MSE, obtendo-se o RMSE (*Root Mean Squared Error*), o qual se calcula tanto para as amostras de calibração, RMSEC (*Root Mean Squared Error o Calibrationf*) quanto para as amostras de previsão, RMSEP (*Root Mean Squared Error of Prediction*).

- ***Previsão de novas amostras***. Utilização do modelo construído e validado para prever as propriedades de amostras novas.

Nas seções apresentadas a seguir, são descritos sucintamente os principais métodos para calibração multivariada reportados na literatura.

1.3.1 Métodos de Calibração Multivariada

1.3.1.1 Regressão em Componentes Principais

A Regressão em Componentes Principais (*Principal Component Regression – PCR*) decompõem os dados originais (matriz **X**) em Componentes Principais e realiza uma regressão múltipla inversa relacionando os escores desta matriz com a propriedade de interesse (JACKSON, 1991).

Supomos que cada amostra contenha *p* analitos. Desta forma, teremos *p* variáveis y_1, y_2, y_3,..., y_p referentes a concentração. Isto permite definir um vetor de concentrações **y** (**y** = y_1, y_2, y_3,..., y_p) que descreve a concentração de cada um dos *p* componentes da amostra. O espectro de cada amostra é dado pelas absorbâncias

nos diferentes comprimentos de onda, *k*. Por tanto, cada amostra é definida por um vetor **x** que contem k variáveis independentes (**x** = x_1, x_2, x_3,..., x_k). Quando se constrói um conjunto de calibração com *m* amostras, o sistema pode ser descrito por meio de uma matriz de dados: uma matriz **X**, que contem os dados espectrais (de dimensão *m* x *k*) e uma matriz **Y** contendo os valores das concentrações (de dimensão *m* x *p)*.

A primeira etapa deste processo é a decomposição da matriz **X** em Componentes Principais:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\ T} + \mathbf{E} \qquad (1)$$

Uma vez escolhido o número A de Componentes Principais ideal para descrever a matriz **X**, esta pode ser representada por sua matriz de escores **T**:

$$\mathbf{T} = \mathbf{X}\mathbf{P} \qquad (2)$$

A matriz de concentração Y é então relacionada com os escores de X:

$$\mathbf{Y} = \mathbf{T}\mathbf{B} + \mathbf{E} \qquad (3)$$

Onde a **B** é a matriz dos coeficientes da regressão e é calculada pela pseudo-inversa de **T** $[(\mathbf{T}^{\mathbf{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathbf{T}}]$, conhecendo-se os valores das concentrações do conjunto de calibração:

$$\mathbf{B} = (\mathbf{T}^{\mathbf{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathbf{T}}\mathbf{Y} \qquad (4)$$

Uma vez estabelecido o modelo da calibração, este é utilizado para prever as concentrações de novas amostras a partir dos espectros registrados. Este procedimento é realizado da seguinte forma:

A partir da matriz de pesos **P** calculada na etapa de calibração, para **A** componentes ótimos, calculam-se os escores ($\mathbf{T}^*$) das amostras de previsão ($\mathbf{X}^*$):

$$\mathbf{T}^* = \mathbf{X}^*\mathbf{P} \qquad (5)$$

E a partir da matriz dos coeficientes de regressão **B**, também calculada na etapa de calibração, calcula-se a propriedade a ser determinada das novas amostras:

$$\mathbf{Y}_{\mathbf{prev}} = \mathbf{T}^*\mathbf{B} \qquad (6)$$

Algumas vezes a variância relacionada a **Y** (concentrações) é uma parcela importante da variância global. Por usar apenas a matriz **X** na modelagem, PCR pode falhar ao encontrar as combinações lineares apropriadas das variáveis que modelam as concentrações. Além disso, se a escolha do número ótimo de PCs não

for realizada corretamente, o modelo PCR pode apresentar uma baixa capacidade de previsão.

1.3.1.2 Regressão por mínimos quadrados parciais

A técnica de regressão por mínimos quadrados parciais (*Partial Least Squares Regression – PLS*) foi introduzida por Wold (1975) e, geralmente, apresenta um poder de previsão melhor que o PCR. Durante a etapa de calibração, a modelagem PLS utiliza tanto a informação da matriz de dados **X** como da matriz de concentração **Y**, obtendo-se novas variáveis denominadas variáveis latentes, fatores ou componentes. Cada matriz é decomposta na soma de "A" variáveis latentes (VLs) da seguinte maneira:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{T} + \mathbf{E} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{T} + \mathbf{E} \qquad (7)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{T} + \mathbf{F} = \sum_{a=1}^{A} \mathbf{u}_a \mathbf{q}_a^{T} + \mathbf{F} \qquad (8)$$

onde **T** é a matriz dos escores, **P** dos loadings e **E** a matriz dos resíduos da matriz dos dados **X** e **U** é a matriz dos escores, **Q** dos loadings e **F** a matriz dos resíduos da matriz de concentração **Y**. Se tivermos M amostras, A fatores, k variáveis e P analitos, a dimensionalidade das matrizes é: **T** e **U** (M x A), $\mathbf{P}^{T}$ e $\mathbf{Q}^{T}$ (A x K) e (A x P), respectivamente.

A decomposição de ambas as matrizes não é independente, realiza-se de forma simultânea, estabelecendo-se uma relação interna entre os escores dos blocos **X** e **Y** de forma que, para cada fator a, a seguinte relação é obtida:

$$\mathbf{u}_a = \mathbf{b}_a \mathbf{t}_a \qquad (9)$$

onde $\mathbf{b}_a$ é o coeficiente de regressão para cada um dos fatores. Para as A VLs, a matriz dos coeficientes de regressão **B** (de dimensão A x A) é determinada pela seguinte relação:

$$\mathbf{Y} = \mathbf{T}\mathbf{B}\mathbf{Q}^{T} + \mathbf{F} \qquad (10)$$

No caso se ter a concentração de apenas um analito na matriz **Y**, o algoritmo é denominado de PLS1, e para mais de um analito, o algoritmo é chamado de PLS2.

Construído o modelo de calibração, este é usado para fazer a estimativa parâmetro em novas amostras. Todavia, os modelos PCR e PLS não permite uma interpretação físico-química direta dos resultados. A razão dessa dificuldade advém

do fato dessas técnicas realizarem a regressão no domínio dos dados transformados.

O problema acima pode ser contornado mediante a aplicação da regressão linear múltipla, pois esta técnica realiza a regressão no domínio original como descrito a seguir.

1.3.1.3 Regressão Linear Múltipla

A regressão linear múltipla (*Multiple Linear Regression–MLR*) foi introduzida por Stemberg et al. (1960). Essa técnica busca estabelecer uma relação linear entre sinal e concentração aplicando o método dos mínimos quadrados. Para isso, faz uso tanto na calibração clássica como na calibração inversa. O modelo MLR pode ser obtido a partir de uma matriz **X** de respostas instrumentais com dimensão (m x k), onde m representa o número de amostras e k o número de variáveis (no caso de espectros k = comprimentos de onda). Além disso, utiliza os dados de um vetor **y** de dimensão (m x 1) que contém as concentrações (ou outra propriedade) obtidas por um método de referência das amostras. Cada variável dependente de **y** é expressa como uma combinação linear das variáveis independentes da matriz **X** por intermédio da expressão:

$$\mathbf{y} = \mathbf{Xb} \qquad (11)$$

onde o vetor **b** contém os coeficientes da regressão e é calculado por mínimos quadrados a partir da pseudo-inversa de **X**:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad (12)$$

os índices sobrescritos -1 e T representam a inversão e transposição da matriz, respectivamente.

Com o modelo determinado, as concentrações de novas amostras podem ser estimadas a partir da seguinte equação:

$$\mathbf{y}_{prev} = \mathbf{X}^*\mathbf{b} \qquad (13)$$

$\mathbf{X}^*$ representa a matriz de dados para as novas amostras.

Não obstante seu grande potencial, o método MLR apresenta alguns problemas que limitam sua aplicação. Um deles é que o número de amostras deve ser igual ou superior ao número de variáveis. Uma vez que o modelo consiste na resolução de um sistema de equações lineares simultâneas, a essa condição necessita ser satisfeita. Caso contrário, o sistema torna-se indeterminado.

Um problema importante em calibração MLR é que a matriz ($\mathbf{X^TX}$) pode não ser invertida ou promover a propagação de erros quando existir forte correlação ou multicolinearidade entre as variáveis. A definição dos termos e medidas do grau de correlação e multicolinearidade são vistas a seguir.

1.3.2 Correlação e multicolinearidade

No uso estatístico geral, *correlação*, também chamada de *coeficiente de correlação*, indica a magnitude e a direção da relação linear entre duas variáveis aleatórias. Neste sentido geral, existem vários coeficientes medindo o grau de correlação, adaptados à natureza dos dados. O mais conhecido é o coeficiente de correlação de Pearson, o qual é obtido dividindo a covariância de duas variáveis pelo produto de seus desvios padrão. Segundo a definição da estatística, o valor da autocorrelação está situado entre 1 (correlação perfeita) e -1 (anti-correlação perfeita). O valor 0 significa ausência de correlação linear.

A *Multicolinearidade* ocorre quando qualquer variável independente é altamente correlacionada com um conjunto de outras variáveis independentes. Embora as estimativas dos coeficientes de regressão sejam muito imprecisas quando a multicolinearidade está presente, a equação do modelo ajustado pode ainda ser útil. A inclusão de variáveis multicolineares, ou irrelevantes, geralmente melhoram os modelos, mas ao preço de superajustar os dados e torná-los menos generalizáveis à população. Logo, a inclusão de variáveis redundantes pode ter vários efeitos potencialmente danosos, ainda que as variáveis adicionais não influenciem diretamente o desempenho do modelo (HAIR et al., 2005). Por exemplo, suponha que se deseje prever as novas propriedades de interesse. Se essas previsões forem interpolações na região original do espaço onde a multicolinearidade existe, então previsões satisfatórias seriam freqüentemente obtidas, pois a função $\sum_{j=1}^{k} \mathbf{b_j x_{ij}}$ pode ser bem estimada. No entanto, b_j individuais podem ser mal estimados. Por outro lado, se a previsão das novas observações requererem extrapolação além da região original do espaço *x* onde os dados foram coletados, então se espera obter resultados ruins. Extrapolação requer geralmente boas estimativas dos parâmetros individuais do modelo.

Um método para avaliar o grau de multicolinearidade entre um conjunto de variáveis é o *Variance Inflation Factors-VIF* (GALVÃO e ARAÚJO, 2007; DRAPER e SMITH, 1998).

O VIF para uma k^{th} variável pode ser expresso como:

$$\mathbf{VIF(k)} = \frac{\mathbf{1}}{\mathbf{1-\rho^2(k)}} \tag{14}$$

onde ρ(k) é o coeficiente de correlação entre a variável x_k e o x_{kestim} estimado pela regressão com as variáveis restantes, ou seja, realiza-se uma regressão entre x_k e as x-variáveis restantes. Encontra-se assim, o vetor de coeficientes, $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{x}_k$. Com os valores de **b**, encontra-se o $\mathbf{x}_k$ estimado, $\mathbf{x}_{kestim} = \mathbf{Xb}$. Então, o ρ(k) é dado pela expressão:

$$\boldsymbol{\rho(k)} = \frac{\mathbf{1}}{\mathbf{N-1}}\sum_{\mathbf{1}}^{\mathbf{N}}\left(\frac{\mathbf{x_{kestim}} - \bar{\mathbf{x}}_{\mathbf{kestim}}}{\mathbf{s_{xkestim}}}\right)\left(\frac{\mathbf{x_k} - \bar{\mathbf{x}}_{\mathbf{k}}}{\mathbf{s_{xk}}}\right) \tag{15}$$

onde $\bar{x}_{kestim}$, $\bar{x}_k$, $s_{xkestim}$ e s_{xk} são as médias e desvios-padrão de $\mathbf{x}_{kestim}$ e $\mathbf{x}_k$, respectivamente, para as **N** amostras.

O valor do VIF(k) indica o grau com que a variância de b_k estimado aumenta (inflaciona) com respeito à variância que resultaria se as x-variáveis fossem não correlacionadas. Para um conjunto de variáveis perfeitamente não-correlacionadas, o valor do VIF é igual a 1. Por outro lado, quando as variáveis são perfeitamente correlacionadas, o valor do VIF é infinito. Obviamente, um valor alto do VIF indica uma forte multicolinearidade da variável em relação às demais. Um valor considerado ideal do VIF é arbitrário, contudo, valores abaixo de 4, 5, 7, 10 são citados como indicadores de não-multicolinearidade (GIACOMELLI et al., 1998).

Outros métodos são citados na literatura com o objetivo de minimizar efeitos de multicolinearidade. Naes e Mevik (2001) utilizaram o método de Regressão por Componentes Principais (PCR) para esta finalidade, em problemas de regressão e também de análise discriminante.

Outro problema fundamental em MLR consiste na seleção das variáveis mais informativas e não redundantes para o modelo de calibração. Para alcançar esse objetivo, várias técnicas de seleção de variáveis têm sido propostas na literatura.

1.3.3 Técnicas de seleção de variáveis

A seleção de comprimentos de onda de um espectro que resulte em modelos de calibração multivariada com máxima precisão é ainda uma tarefa desafiadora (CANECA et al., 2006)**.** Diversos métodos têm sido propostos para essa finalidade entre os quais são descritos, a seguir, os mais utilizados no contexto da química analítica.

1.3.3.1 Regressão Stepwise (SW)

O método de regressão stepwise é um procedimento padrão para seleção de variáveis que combina dois outros métodos, o *forward selection* e *backward elimination* (MONTGOMERY et al. citado por CHONG e JUN, 2005). O algoritmo, progressivamente, adiciona novas variáveis ao modelo, iniciando daquela com maior correlação com a resposta, como no método *forward selection* e incorpora um mecanismo de eliminação de variáveis igual ao método de *backward elimination*. O *forward selection* é um método iterativo que começa com uma variável (x) e, progressivamente, adiciona mais variáveis ao modelo de regressão até que um critério de parada seja satisfeito. A variável inicial deve apresentar máxima correlação com a variável de resposta (y). A cada iteração, é construído um novo modelo e o efeito da variável incluída é avaliado por um teste-F. A variável com um valor de F maior que um F-crítico é incluída no modelo. No método *backward elimination*, inicia-se com a construção de um modelo de regressão com todas as variáveis disponíveis e, subseqüentemente, variáveis são retiradas e o efeito dessa eliminação é avaliado, da mesma forma que no método *forward selection*. As variáveis com valores de F menores que F-crítico são descartadas do modelo.

Uma recente modificação neste método é apresentada por Forina et al. (2007), onde é feita uma ortogonalização para eliminar o problema de multicolinearidade entre as variáveis selecionadas.

1.3.3.2 Algoritmo Genético (AG)

O Algoritmo Genético (AG) foi proposto por Holland (1975) e utiliza operadores matemáticos para simular o mecanismo de seleção natural inspirada na teoria da evolução de Charles Darwin. Esse algoritmo seleciona conjunto(s) de variáveis de forma mais aleatória e menos susceptível a soluções locais. O AG

permite a construção de baseado em combinações de variáveis que fornecem os melhores valores de predição (HIBBERT, 1993).

O funcionamento do AG baseia-se no processo evolutivo dos seres vivos, seguindo o princípio básico de que as gerações futuras (descendentes) serão mais “evoluídas” do que os seus antecedentes. Gerações melhores continuariam existindo, enquanto que gerações mais “frágeis” tenderiam a sucumbir.

O AG utiliza operadores genéticos, tais como, o cruzamento (*crossover)* e a mutação que manipulam indivíduos de uma população, por intermédio de gerações, para melhorar (aperfeiçoar) a adaptação (*fitness*) gradativamente. Os indivíduos numa população, também denominados de cromossomos, são representados por cadeias (*strings*) de números binários. A função de avaliação (*fitness*) estabelece a relação entre o AG e o problema de otimização. Em termos práticos, essa é avaliada mediante a construção de modelos MLR baseados nos comprimentos de onda indicados nos cromossomos. Em seguida, esses modelos são usados para estimar o parâmetro de interesse em um conjunto de amostras de validação e o valor de fitness é calculado como o inverso do RMSEP obtido.

Um procedimento que também pode ser utilizado no AG é o *elitismo*, onde cada indivíduo selecionado e cruzado com seu parceiro é colocado no lugar do pior indivíduo da população anterior. Para evitar que indivíduos que produziram bons resultados sejam perdidos, um determinado número desses indivíduos é mantido na nova geração.

1.3.3.3 Algoritmo das Projeções Sucessivas (APS)

O APS é uma técnica originalmente concebida para selecionar variáveis minimamente colineares em calibração multivariada baseada em MLR. Esse algoritmo foi aplicado inicialmente à seleção de comprimentos de onda em espectrometria UV-VIS, especialmente sob condições de forte sobreposição espectral (ARAÚJO et al., 2001). Os modelos MLR-APS apresentaram desempenho (medido em termos de habilidade de predição) melhor que os modelos PLS em muitas aplicações: UV-VIS (ARAÚJO et al., 2001), ICP-OES (GALVÃO et al., 2001) e espectroscopia NIR (BREITKREITZ et al., 2003). Além disso, o APS foi aplicado na construção de modelos MLR aplicado à análise de óleo lubrificante (CANECA et al., 2006).

O algoritmo APS compreende basicamente três etapas. Inicialmente, o algoritmo seleciona subconjuntos de variáveis com base no critério de minimização da multicolinearidade. Tais conjuntos são obtidos de acordo com uma seqüência de operações de projeções aplicadas nas colunas da matriz de calibração. Numa segunda fase, o melhor subconjunto é escolhido de acordo com um critério que avalia a habilidade de previsão de um modelo MLR, tal com o $PRESS_v$, se um conjunto independente é utilizado, ou $PRESS_{cv}$, se é empregado o método de validação cruzada. Numa terceira fase, o subconjunto escolhido é submetido a um procedimento de eliminação para determinar se alguma variável poderá ser removida sem perda significante da capacidade de previsão.

O APS também apresenta uma versão cuja validação é realizada pelo procedimento de validação cruzada (GALVÃO et al., 2007) e uma outra, que é usada para classificação de amostras (PONTES et al., 2005).

Uma limitação do APS é que o critério de seleção de variáveis não leva em consideração a correlação da variável da matriz de resposta instrumental (**X**) com a matriz da propriedade a ser determinada (**y**). Para contornar este problema, Kompany-Zareh e Akhlaghi (2007) apresentaram uma nova proposta onde cada vetor de projeção é multiplicado por um fator, dado pelo coeficiente de correlação entre a variável de **X** com **y**. Outra maneira de resolver este problema é apresenta por YE, et al.(2007), onde o APS é usado associado com outro método, denominado UVE (*Uninformative Variable Elimination*).

1.3.3.4 Generalized Simulated Annealing

É um método estocástico, assim como o AG, porém não se inspira em princípios biológicos, e sim, em princípios termodinâmicos simulando o “cozimento” de um sólido. É um algoritmo de busca global que visa a encontrar o mínimo global de uma determinada função de custo, evitando captura de mínimos locais. Para esse propósito, algum grau de “aleatoriedade” é imposto no procedimento de procura para permitir a exploração de regiões longe da tendência atual de minimização de custo (KALIVAS, et al., 1989; HIJRCHNER e KALIVAS, 1995).

1.3.3.5 Seleção de variável por “informação mútua”

Este método utiliza uma medida da informação mútua (*mutual information*) entre os dados espectrais (variáveis independentes) e a concentração do analito

(variável dependente) para selecionar a primeira variável, que tem a mais alta relação com o analito. Uma vez a primeira variável é selecionada, o procedimento Stepwise é usado para selecionar as próximas variáveis espectrais. É um método aplicado para calibração linear e não-linear (BENOUDJIT et al., 2004).

1.3.3.6 PLS-VIP (Variable Importance in the Projection)

Este método, proposto inicialmente por Wold et al. (citado por CHONG e JUN, 2005), realiza projeções nos escores obtidos pelo método PLS, e escolhe aqueles que realmente são relevantes em relação à resposta, ou seja, que têm coeficientes diferentes de zero.

1.3.3.7 UVE-PLS (*Uninformative Variable Elimination by PLS*)

É um método que se baseia na análise dos coeficientes de regressão (b_j). São adicionadas variáveis artificiais (ruídos) na matriz de dados e é obtida a relação $c_j = b_j/s(b_j)$ para $j = 1, ..., p$ *variáveis*, onde $s(b_j)$ são os desvios-padrão. São eliminadas aquelas variáveis cuja relação, $abs(c_j) < abs(Max(c_{artificial}))$ é satisfeita (CENTNER, et al. 1996).

1.3.3.8 Interval PLS (iPLS)

Este método foi proposto por Norgaard et al. (2000) para aplicações envolvendo dados espectrais. O procedimento de iPLS compreende dois passos. Primeiro, o espectro é dividido em intervalos de largura igual e modelos PLS locais são construídos para cada intervalo. Segundo, a posição do centro e a largura do intervalo que produziram o melhor modelo PLS (em termos de RMSEPCV, por exemplo) são ajustados para aperfeiçoar os resultados. Para esse propósito, o intervalo é trocado inicialmente à esquerda e à direita até um número de pontos máximo preestabelecido. Finalmente, depois que o intervalo é transladado para a melhor posição, a largura é otimizada testando o limite simétrico (dois lados) e assimétrico (unilateral) do intervalo.

Vale notar que o iPLS não testa todos os possíveis intervalos de uma maneira exaustiva e o resultado pode ser um mínimo local da função de custo adotada.

1.3.3.9 Seleção de variáveis usando Tikhonov Regularization

Stout et al. (2007) descrevem um método de seleção de variáveis para calibração multivariada, onde os coeficientes de regressão para as variáveis descartadas são zero, ou próximos de zero, usando o método "Tikhonov Regularization (TR)". Este método busca otimizar a seguinte relação:

$$\mathbf{min}(\|\mathbf{Xb}-\mathbf{y}\|_{\mathbf{a}}^{\mathbf{a}}+\boldsymbol{\lambda}\|\mathbf{Lb}\|_{\mathbf{b}}^{\mathbf{b}})$$

Os parâmetros a serem ajustados são a, λ, **L** e b. O termo da esquerda representa o erro da previsão (bias) e o temo da direita representa a variância. A minimização destes dois termos é o que torna este método mais preciso.

Nas próximas seções, são descritos os fundamentos das técnicas espectroanalíticas utilizadas para a obtenção dos modelos de calibração multivariada MLR baseados na seleção de variáveis.

1.4 ESPECTROMETRIA NO ULTRAVIOLETA E VISÍVEL (UV-VIS)

A espectroscopia de absorção molecular nas regiões do ultravioleta e visível (UV-VIS) utilizada radiação eletromagnética na faixa espectral compreendida entre 200 e 780 nm. Quando submetida a essa radiação, a molécula de um composto pode sofrer transições eletrônicas por ocasião da absorção de energia quantizada.

A radiação UV-VIS possui geralmente energia suficiente apenas para promover a excitação de elétrons de ligações π e de valência n (não ligantes). Isto requer que a molécula contenha pelo menos um grupo funcional insaturado (por exemplo, C=C, C=O). Esses grupos que absorvem radiações UV/VIS são chamados *cromóforos*, sendo responsável principalmente pelas transições $\pi \rightarrow \pi^*$ e $n \rightarrow \pi^*$ (CECCHI, 2003).

A Lei de Lambert-Beer estabelece uma relação matemática entre a transmitância (ou absorbância) medida, a espessura da amostra e a concentração das espécies absorventes (SKOOG et al., 2002). Segundo essa lei, a passagem de um feixe de radiação monocromática num número sucessivo de moléculas absorventes idênticas resulta na absorção de frações iguais de energia radiante que as atravessa. Assim podemos concluir que a absorbância de uma solução é diretamente proporcional à concentração da espécie absorvente quando se fixa o

comprimento do percurso; e diretamente proporcional ao comprimento do percurso quando se fixa a concentração.

Os espectros UV-VIS são usualmente obtidos com um espectrofotômetro e consistem de um gráfico de absorbância (ou transmitância) versus comprimentos de onda. As características principais de uma banda de absorção são a sua posição e intensidade. A posição de absorção corresponde ao comprimento de onda da radiação cuja energia é igual à necessária para que ocorra a transição eletrônica. Já a intensidade de absorção depende essencialmente de dois fatores: da probabilidade de transição e da energia dos orbitais moleculares.

Os espectros de absorção UV-VIS, quando obtidos em fase condensada, apresentam geralmente bandas largas e com baixa resolução, resultantes da sobreposição dos sinais provenientes de transições vibracionais e rotacionais ao sinal associado à transição eletrônica. Além disso, observam-se alterações na posição e na intensidade das bandas, originadas de interações entre suas moléculas e as do solvente. Um aumento na polaridade do solvente usualmente promove um deslocamento da banda para comprimentos de onda maiores (efeito batocrómico), se a transição associada é do tipo $\pi \rightarrow \pi^*$, e um deslocamento para menores comprimentos de onda (efeito hipsocrômico), quando a transição é do tipo $n \rightarrow \pi^*$. Efeitos envolvendo um aumento (efeito hipercrômico) ou uma diminuição (efeito hipocrômico) na intensidade da banda de absorção também podem ser verificados como resultado dessas interações.

Os espectros eletrônicos de absorção apresentam outra característica marcante que consiste da forte sobreposição de bandas associadas a duas ou mais substâncias presentes em uma amostra. Essas características decorrem da natureza alargada das bandas e da modesta correlação entre o espectro e a estrutura molecular. Com efeito, substâncias com estruturas moleculares muito diferentes, mas contendo o(s) mesmo(s) cromóforo(s), podem apresentar espectros UV-Vis com perfis similares e bandas localizadas nas mesmas regiões de comprimentos de onda (SKOOG, 2002).

É importante salientar que sobreposições espectrais pronunciadas ocasionam multicolinearidade entre as variáveis espectrais dificultando a implementação da calibração multivariada, sobretudo a baseada em MLR. Nesse contexto, a aplicação de técnicas de seleção de variáveis pode minimizar esses problemas de multicolinearidade.

1.5 ESPECTROMETRIA NO INFRAVERMELHO PRÓXIMO (NIR)

A radiação infravermelha compreende a faixa de 780 a 100000 nm. O espectro do infravermelho é dividido em infravermelho próximo (*Near InfraRed* – NIR), infravermelho médio (*Middle InfrareD* – MID) e infravermelho distante (*FAR Infrared* – FAR). A **Tabela 1.2** abaixo apresenta os limites aproximados para cada região (SKOOG et al., 2002).

Tabela 1.2 – Regiões espectrais do infravermelho

Região	Intervalo em número de ondas [cm^{-1}]	Intervalo em comprimento de onda, λ [nm]
Próximo (NIR)	12800 – 4000	780 – 2500
Médio (MID)	4000 – 200	2500 – 5000
Distante (FAR)	200 - 10	5000 - 100000

Na região NIR, as ocorrências espectrais correspondem aos sinais (de absorção, reflectância, etc) relacionados aos sobretons e combinações de transições fundamentais que ocorrem na região do MID. As ligações envolvidas nas transições vibracionais ativas no NIR são tipicamente C–H, N–H e O–H. Ao contrário da espectrometria MID, a técnica NIR é mais útil para análises quantitativas de compostos orgânicos contendo as referidas ligações devido às facilidades experimentais (p. ex., a menor interferência da banda de OH da água associada à umidade do ar).

A região do NIR foi a primeira faixa do espectro não-visível a ser descoberta, porém sua utilização como ferramenta analítica só ocorreu muito tempo depois. As primeiras aplicações analíticas apareceram na década de 50, com o aparecimento dos primeiros espectrofotômetros comerciais baseados em detectores fotoelétricos. O grande impulso dessa técnica ocorreu na década de 60 quando Karl Norris, chefe de um grupo de pesquisa da USDA (*United States Department of Agriculture*), passou a utilizá-la na análise de produtos agroalimentares. A partir de então, o interesse pela espectroscopia NIR cresceu notavelmente.

O desenvolvimento da microeletrônica, a partir do final da década de 70, permitiu o desenvolvimento de equipamentos bem melhores e mais acessíveis do que aqueles existentes até então. Assim, foram construídos os primeiros espectrofotômetros que possibilitava o registro de espectros de forma rápida e altamente reprodutível. Além disso, o desenvolvimento e o uso da Quimiometria

impulsionaram muito a expansão da técnica NIR, pois permitiu superar o problema da não especificidade das bandas de absorção.

O grande interesse que tem despertado a espectroscopia NIR nos setores industrial e acadêmico é conseqüência direta das vantagens que esta técnica oferece como ferramenta analítica para o controle de qualidade. Por um lado, a baixa absortividade molar das bandas de absorção permite se trabalhar no modo de reflectância, com a conseqüente vantagem de se poderem obter os espectros de amostras sólidas, sem necessidade de realizar qualquer tratamento nas mesmas. Por outro lado, a dependência do sinal com a natureza química e física da amostra permite tanto sua identificação como a quantificação de seus parâmetros químicos e físicos.

1.5.1 Origem da absorção no infravermelho próximo

Para que uma molécula possa absorver a radiação eletromagnética é necessário que duas condições sejam satisfeitas: primeiro, a radiação deve conter a energia exata para satisfazer os requerimentos energéticos do material; segundo, deve haver o acoplamento entre a radiação e a matéria. A radiação na região do infravermelho tem energia necessária para promover apenas transições vibracionais nas moléculas, e a primeira condição para absorção será satisfeita se uma determinada freqüência de radiação infravermelha corresponde exatamente a uma freqüência fundamental de vibração de uma determinada molécula. Para satisfazer a segunda condição de absorção, deve ocorrer uma variação no momento dipolar da molécula.

Para uma molécula diatômica a freqüência de vibração pode ser conhecida, aproximadamente, supondo o modelo do *oscilador harmônico*, em que um átomo se desloca de sua posição de equilíbrio com uma força proporcional ao deslocamento (lei de Hooke). Neste caso, a função de energia potencial seria uma parábola, centrada na distância de equilíbrio, com os níveis energéticos vibracionais igualmente espaçados (**Figura 1.1 (a)**) e só seriam permitidas transições entre níveis adjacentes, $\Delta E = \pm 1$ (PASQUINI, 2003).

Na prática, quando dois átomos se aproximam, a repulsão coulômbica entre os dois núcleos provoca um aumento mais rápido da energia potencial do que prediz a aproximação harmônica e quando a distância interatômica se aproxima da distância em que ocorre a dissociação, o nível de energia potencial se estabiliza.

Desta forma, podemos dizer que as moléculas reais apresentam um comportamento mais condizente com o modelo do *oscilador anarmônico* (**Figura 1.1 (b)**). As moléculas apresentam comportamento harmônico apenas quando a energia potencial é baixa, ou seja, em torno da posição de equilíbrio.

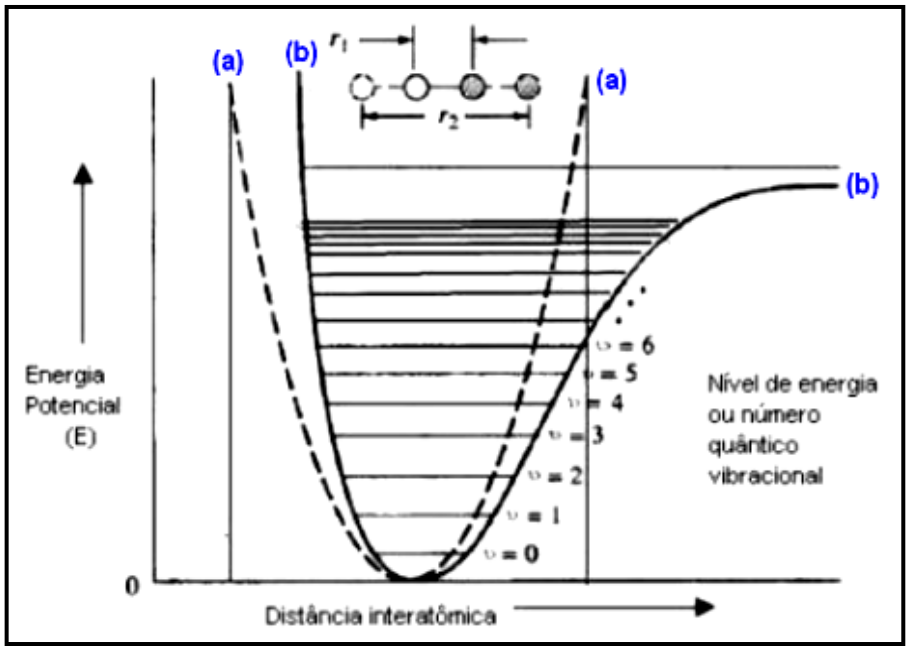

Figura 1.1. Curvas de energia potencial de modelos harmônico e anarmônico.

De acordo com o modelo do oscilador anarmônico, verifica-se que ocorrem transições tanto da banda fundamental ($\Delta\nu = \pm1$), como também de bandas correspondentes a outras transições ($\Delta\nu = \pm2, \pm3, \pm4, ...$), as quais são denominadas de *sobretons* (primeiro, segundo, terceiro sobreton, ..., respectivamente) e correspondem as bandas observadas no NIR. Outra conseqüência da anarmonicidade é que os níveis de energia não estão igualmente espaçados, com isso, os sobretons aparecem em freqüências ligeiramente menores que as correspondentes a múltiplos das freqüências fundamentais. As transições em que o número quântico vibracional ($\Delta\nu$) é maior que 1 são muito menos prováveis e, portanto, a intensidade dessas bandas é bem menor.

Além das bandas de sobretons, na região do NIR também aparecem as *bandas de combinação*, as quais são provenientes da mudança simultânea na energia de dois ou mais modos vibracionais e que se observa em freqüências dadas por $\nu = n_1\nu_1 + n_2\nu_2 + ...$, onde n_i são números inteiros e ν_i são as freqüências das transições que contribuem para a banda de combinação.

A informação contida num espectro NIR é um reflexo da contida num espectro de infravermelho médio, no entanto, as bandas de absorção no NIR estão normalmente muito sobrepostas e são de baixa intensidade. A intensidade das bandas de combinação e sobretons dependem do grau de anarmonicidade da ligação química. O átomo de hidrogênio, por ter uma massa menor, vibra com maior amplitude numa vibração de *estiramento*, fazendo com que este movimento se desvie apreciavelmente do modelo do oscilador harmônico. Como conseqüência, quase todas as bandas de absorção observadas na região do NIR provêm de sobretons das vibrações de estiramento de grupos AH_x, ou bandas de combinação destes grupos.

1.5.2 Modos de registro no NIR

Os registros das absorções na região do NIR podem ser obtidos nos modos de reflectância, transmitância ou transflectância. A diferença básica entre estes modos de medidas é a posição da amostra no instrumento, como mostra a **Figura 1.2**.

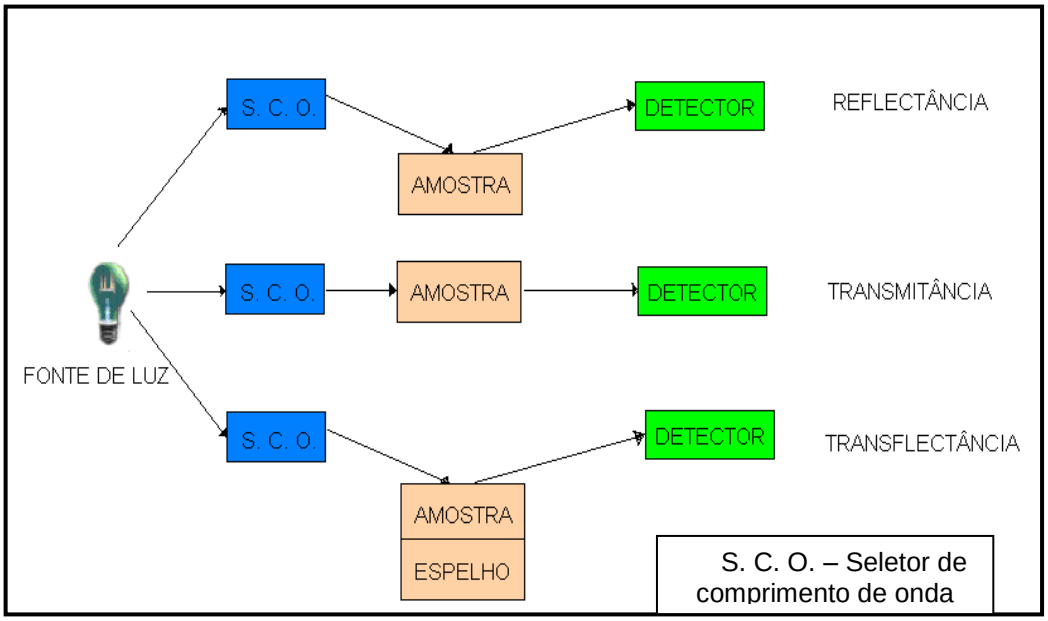

Figura 1.2. Modos de medida em espectroscopia NIR.

Em todos os casos, o sinal analítico que se obtém é uma função complexa que habitualmente se expressa como absorbância aparente (A = log(1/R)) ou unidade de Kubelka-Munk, quando as medidas se realizam do modo de reflectância,

ou como absorbância (A = log(1/T)) quando as medidas são feitas no modo de transmitância.

1.5.2.1 *Medidas obtidas por reflectância difusa*

A espectroscopia de reflectância estuda a radiação refletida por uma amostra, a qual pode ser especular ou difusa. A reflectância especular predomina quando o material sobre o qual se produz a reflexão tem valores altos dos coeficientes de absorção para o comprimento de onda incidente, quando a penetração da radiação é muito pequena e quando as dimensões da superfície reflectante é muito maior que o comprimento de onda. A reflectância difusa (**Figura 1.3**) ocorre em todas as direções da superfície como conseqüência dos processos de absorção e dispersão e predomina quando os materiais são fracamente absorventes e quando a penetração da radiação é grande em relação ao comprimento de onda incidente.

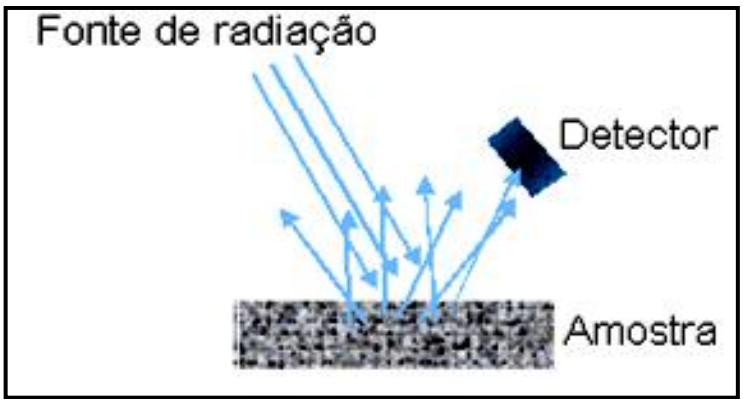

Figura 1.3. Ilustração da reflectância difusa

Normalmente, as medidas de reflectância contêm os dois componentes da reflexão. A componente especular praticamente não apresenta informação sobre a composição da amostra, e sua contribuição pode ser minimizada pela posição do detector em relação à amostra. Por outro lado, a componente difusa contém informação em relação à amostra e é, portanto, a base para as medidas que se realizam com esta técnica.

A reflectância difusa é explicada através da teoria de Kubelka-Munk (KUBELKA e MUNK, 1931). Esta teoria assume que a radiação que incide sobre um meio dispersante sofre simultaneamente um processo de absorção e dispersão, de forma que a radiação refletida pode ser descrita em função das constantes de

absorção (**k**) e dispersão (**s**). No caso de amostras opacas de espessura infinita (**y**), a função de Kubelka-Munk é descrita da seguinte maneira:

$$\mathbf{f(R_\infty) = \frac{(1-R_\infty)^2}{2R_\infty} = \frac{k}{s}} \qquad (16)$$

onde $\mathbf{R_\infty}$ é a reflectância absoluta da amostra.

Em análises quantitativas, a Equação 16 pode ser descrita em função da concentração do analito absorvente (**c**) da seguinte forma:

$$\mathbf{f(R_\infty) = \frac{(1-R_\infty)^2}{2R_\infty} = \frac{k}{s} = \frac{ac}{s}} \qquad (17)$$

onde (**a**) é a absortividade molar.

Na prática, ao invés da reflectância absoluta $\mathbf{R_\infty}$ se utiliza a reflectância relativa R, que é a relação entre as intensidades de luz refletidas pela amostra e por um padrão. Este padrão deve ser um material estável, com reflectância absoluta elevada e relativamente constante na região do NIR, tais como o brometo de potássio, teflon, sulfato de bário, óxido de magnésio, placas cerâmicas de alumina de alta pureza.

Reescrevendo a equação de Kubelka-Munk em termos de reflectância relativa, teremos:

$$\mathbf{f(R) = \frac{(1-R)^2}{2R} = \frac{ac}{s}} \qquad (18)$$

Para aquelas amostras que seguem a Equação 18, o gráfico de f(R) em função da concentração é uma linha reta com coeficiente angular igual a **a**/**s**. No entanto, se a matriz também absorve ou se o analito apresenta forte absorção, a reflectância difusa não obedece esta relação linear.

A equação de Kubelka-Munk, assim como a lei de Beer, é uma equação limitada que só pode ser aplicada para bandas de absorção de baixas intensidades. Esta limitação ocorre quando se aplica a técnica NIR, pois não se pode separar a absorção do analito da absorção da matriz (que freqüentemente absorve fortemente no mesmo comprimento de onda do analito), ocorrendo, portanto, desvios de linearidade.

Do ponto de vista prático, uma alternativa muito utilizada é a aplicação de uma relação entre a concentração e a reflectância relativa análoga à lei de Beer:

$$\log \frac{R_{padrão}}{R_{amostra}} = \log \frac{1}{R_{amostra}} + \log R_{padrão} \cong \frac{ac}{s} \quad (19)$$

Para radiação monocromática o $\log R_{padrão}$ é constante e a Equação 19 pode ser escrita como:

$$A = \log \frac{1}{R} = ac \quad (20)$$

onde **A** é a absorbância aparente, **R** a reflectância relativa, **c** a concentração e **a** é agora uma constante que engloba os termo 1/s e $\log R_{padrão}$. Embora esta expressão não tenha as bases teóricas da equação de Kubelka-Munk, apresenta resultados muitos satisfatórios nas condições usadas em muitas aplicações da espectroscopia por reflectância difusa.

1.5.2.2 *Medidas por transmitância*

A absorção da radiação NIR, assim como na região UV/VIS, segue a lei de Beer e, portanto, a absorbância é definida como:

$$A = \log \frac{I}{I_0} = -\log T \quad (21)$$

onde I_0 é a intensidade de energia incidente e **I** a intensidade da radiação detectada após passar pela amostra.

Desvios da lei de Beer também podem ocorrer aqui, em virtude de ligações por ponte de hidrogênio, complexação ou processos químicos.

Quando se analisa amostra sólida por transmitância, não se pode assumir que qualquer sistema siga a lei de Beer, já que por efeitos de dispersão parte da radiação pode sofrer reflectância difusa e, neste caso, **log1/T** não representa a atenuação da radiação por absorção. Na prática, análises de medidas por transmitância se procede do mesmo modo que nas medidas por reflectância, ou seja, usando a absorbância aparente. De todo modo, os instrumentos utilizados neste tipo de medida estão desenhados para minimizar os efeitos da dispersão e, portanto, o sinal analítico depende fundamentalmente da absorbância da amostra.

As medidas de sólidos por transmitância apresentam vantagens em relação à reflectância por ter maior sensibilidade e homogeneidade espectral (utiliza uma porção maior da amostra), mas apresenta desvantagem se na amostra contem componentes sensíveis à radiação, que podem ser afetados pela grande quantidade de radiação que atravessa a mesma (SABOYA, 2002).

Uma variação desta metodologia são as medidas feitas por transflectância. Neste caso, mede-se a transmitância depois que a radiação tenha atravessado duas vezes a amostra, a segunda depois que uma superfície reflectora (espelho) colocado abaixo da amostra faça com que a radiação retorne pela amostra antes de chegar ao detector.

1.5.3 Vantagens e desvantagens da espectroscopia NIR

As principais vantagens da espectroscopia NIR, como ferramenta de análise qualitativa e quantitativa, são:

- A técnica é não destrutiva e não invasiva;
- A manipulação de amostra é mínima. A possibilidade de realizar medidas tanto no estado líquido como sólido, permite a minimização da manipulação prévia da amostra e, com isso, aumenta a rapidez das análises;
- Baixo custo de análise. Como não usam reagentes ou outros tipos de materiais para preparo das amostras, o custo de análise é pequeno;
- A técnica permite a determinação de vários analitos da amostra sem a necessidade de um procedimento analítico para cada um deles separadamente;
- É possível determinar parâmetros não químicos de uma amostra;
- A resistência dos materiais utilizados e a ausência de partes móveis no sistema de detecção permitem que esta técnica seja usada no controle de processos em plantas industriais;
- Em muitos campos de aplicação, a exatidão da técnica NIR é comparável a outras técnicas analíticas.

Como qualquer outra técnica analítica, a espectroscopia NIR também apresenta seus inconvenientes:

- A complexidade do sinal NIR obriga a aplicação de métodos quimiométricos para modelar os dados e permitir a quantificação das propriedades de interesse;
- A etapa de calibração é geralmente exigente, pois é necessária uma grande quantidade de amostras para assegurar a variabilidade na complexidade da matriz das amostras;
- Torna-se difícil analisar uma amostra problema que apresente variabilidade (física ou química) não contemplada na etapa de calibração;

- A técnica é pouco sensível, especialmente em medidas de reflectância difusa, impossibilitando geralmente a determinação de componentes minoritários.

Apesar do grande potencial analítico da espectrometria NIR, os espectros de amostras com matrizes complexas (alimentos, combustíveis, etc) apresentam muitas sobreposições dificultando a implementação da análise quantitativa. Para resolver esse problema, pode-se recorrer aos métodos de calibração multivariada descritos na **Seção 1.3**.

1.6 SISTEMA AUTOMÁTICO FLUXO-BATELADA

O Sistema de análise em fluxo-batelada (do inglês: *Flow-Batch Analysis* – FBA) foi proposto por Honorato et. al (1999). Estes analisadores possuem como característica explorar estratégias inerentes a sistemas em fluxo e em batelada. A aspiração, bombeamento, transporte dos reagentes e amostra, bem como o monitoramento do sinal, ocorrem em fluxo, porém, a reação é desenvolvida em uma câmara aberta, utilizada para efetivar a mistura e a reação entre amostra e reagente, e é somente aspirada em direção ao detector, após a conclusão da mesma. Por apresentarem características importantes como, baixo custo por análise, alta velocidade analítica, boa precisão e exatidão, flexibilidade, versatilidade, robustez, os sistemas de análise em fluxo-batelada têm proporcionado um caminho promissor para automação em análises químicas (FORMIGA, 2003; LIMA, 2006; FREITAS, 2006).

Alguns aspectos importantes dos sistemas em fluxo-batelada**:**

- Usam válvulas solenóides de três vias (three way) ou de multidistribuição para direcionar os fluidos e uma câmara aberta para mistura, reação, preparação de soluções de calibração, adições de padrão, geração de gradientes de concentração, etc;
- A adição da amostra, reagentes, soluções padrão, tampão, diluentes, indicadores, etc, na câmara aberta são feitas em fluxo de forma simultânea, acionadas por microcomputador;
- A medida do sinal analítico é geralmente feita em fluxo usando ou não cela de fluxo, mas pode ser realizada na câmara;

- São sistemas bastante flexíveis porque permitem trabalhar em uma ampla faixa de concentração das amostras, mudando apenas os parâmetros de software;
- São muito versáteis porque, sem alterar as configurações físicas do sistema, permitem a implementação de diferentes processos analíticos (titulação, adições de padrão, preparação de soluções de calibração, screening analysis);
- São sistemas bastante robustos, simples, baratos e de baixo custo de manutenção;
- Por consumirem baixa quantidade de amostras e reagentes liberam pouco resíduo para o meio ambiente;
- Como nos sistemas em batelada, as medidas podem ser realizadas com a máxima sensibilidade, pois o equilíbrio físico e químico do processo analítico pode ser obtido;
- Permite explorar gradiente linear de concentração das amostras e/ou dos reagentes;
- As amostras podem residir no analisador por longo tempo, sendo adequado para metodologias analíticas envolvendo reações cineticamente lentas;
- A possibilidade de usar um único sistema em fluxo-batelada para realizar diferentes processos analíticos sem a necessidade de modificação física deste sistema pode enquadrá-lo dentro dos sistemas analíticos automáticos com caráter de universalidade.

CAPÍTULO 2

O ALGORITMO DE BUSCA ANGULAR

2. O ALGORITMO DE BUSCA ANGULAR

Neste capítulo serão descritos os fundamentos do algoritmo de busca angular (ASA) e os fluxogramas do programa ASA, bem como algumas ilustrações de como os resultados são apresentados.

2.1 Base Teórica

O ASA foi desenvolvido com o intuito de minimizar problemas de correlação e multicolinearidade em calibração multivariada usando MLR. Para isso, esse algoritmo calcula os cossenos dos ângulos entre as variáveis da matriz de respostas instrumentais das amostras de calibração ($\mathbf{X_{cal}}$), após a centralização na média. Em seguida, o algoritmo seleciona aquelas variáveis que apresentam os menores cossenos (ou maiores ângulos) entre si. A **Figura 2.1** ilustra o exemplo para o caso de uma matriz contendo três amostras (N_{cal} = 3) e cinco variáveis (J = 5).

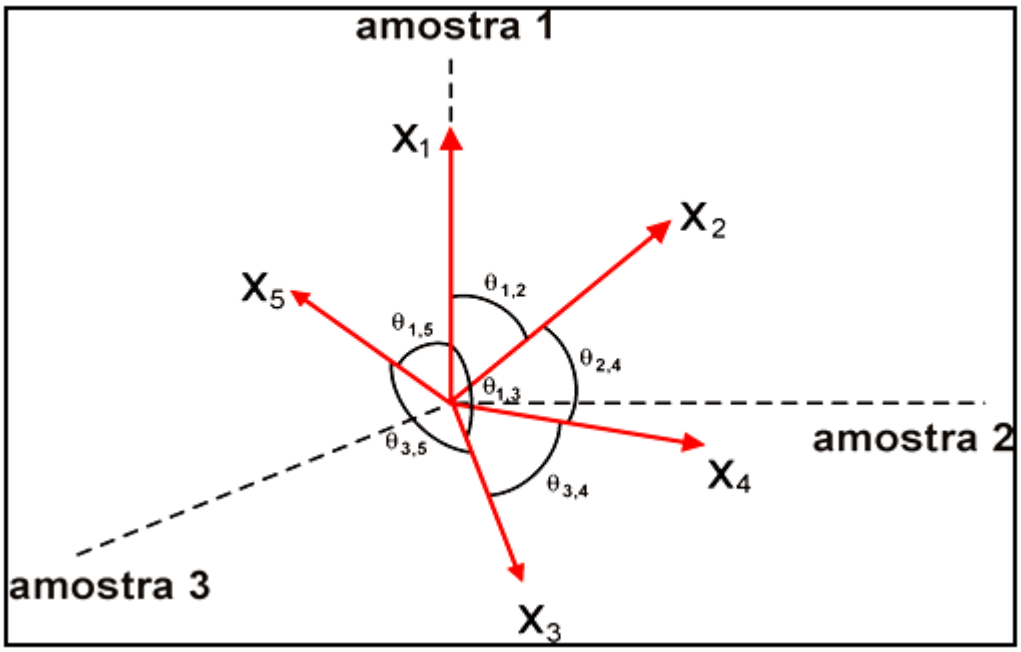

Figura 2.1. Representação geométrica dos ângulos entre os vetores colunas de uma matriz de calibração (J = 5 e N_{cal} = 3), não centralizada na média.

De acordo com a **Figura 2.1**, a *i*-ésima variável corresponde ao vetor $\mathbf{x}_i$ no espaço das amostras R^N. O cosseno do ângulo $\theta_{i,j}$ entre os vetores $\mathbf{x}_i$ e $\mathbf{x}_j$ é dado por:

$$\mathbf{C}_{i,j} = \cos\boldsymbol{\theta}_{i,j} = \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\|\mathbf{x}_i\| * \|\mathbf{x}_j\|} \qquad (22)$$

para *i* e *j* variando de 1 a J e as normas de $\|\mathbf{x}_i\|$ e $\|\mathbf{x}_j\|$ diferente de zero.

O produto interno $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ entre os vetores $\mathbf{x}_i$ e $\mathbf{x}_j$ é definido como

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = x_i^t . x_j \qquad (23)$$

o sobrescrito *t* significa o transposto do vetor.

Neste método, o número de cossenos que necessita ser calculado é J(J – 1)/2, pois $C_{i,i} = \cos 0° = 1$ e $C_{i,j} = C_{j,i}$.

O valor absoluto $|C_{i,j}|$ pode ser usado para calcular a correlação entre os vetores $\mathbf{x}_i$ e $\mathbf{x}_j$. Os vetores são perfeitamente correlacionados se $|C_{i,j}| = 1$ e serão mutuamente ortogonais de $|C_{i,j}| = 0$.

Vale salientar que, como as matrizes de dados $\mathbf{X_{cal}}$ e $\mathbf{Y_{cal}}$ são centradas na média antes de tudo, os valores dos cossenos correspondem aos coeficientes de correlação entre as variáveis $\mathbf{x_i}$ e $\mathbf{x_j}$. A demonstração matemática desse fato é apresentada no ANEXO 7.1.

Após o cálculo dos cossenos, o ASA seleciona subconjuntos contendo de uma a *nmax* = min{N_{cal} -1, *J*} variáveis usando o procedimento mínimo-máximo, iniciando a partir de cada *J* variável, conforme mostrado a seguir. A notação $V_{j,n}$ será usada para representar o subconjunto de *n* variáveis iniciado com x_j. O seguinte pseudo-código é usado para obter os sub-conjuntos $V_{1,j}$, $V_{2,j}$, ..., $V_{j,Ncal}$:

$V_{j,1} = \{ x_j \}$

$n = 1$

enquanto $n < N_{cal}$

Encontrar o indice *k* da próxima variável de acordo com o seguinte critério:

$$\min_{k \notin V_{j,n}} \max_{i \in V_{j,n}} | C_{i,k} |$$

$V_{j,n+1} = \{ V_{j,n} , x_k \}$

$n = n + 1$

fim

Cada variável obtida de acordo com o critério min-máx, é a menos colinear com respeito às variáveis selecionadas previamente. Como j varia de 1 a *J*, um total de (J x N_{max}) subconjuntos são gerados.

Para exemplificar, vamos considerar o exemplo da **Figura 2.1** e supor que os valores dos ângulos e respectivos cossenos sejam os mostrados na **Tabela 2.1**.

Tabela 2.1. Ângulos $\theta_{i,j}$ e correspondentes cossenos $C_{i,j}$ (em parênteses) para ilustrar o exemplo da **Figura 2.1**.

$j \downarrow$ $i \rightarrow$	1	2	3	4	5
1	0^o (1)	30^o (0.87)	85^o (0.09)	70^o (0.34)	20^o (0.94)
2	30^o (0.87)	0^o (1)	60^o (0.50)	40^o (0.77)	55^o (0.57)
3	85^o (0.09)	60^o (0.50)	0^o (1)	35^o (0.82)	80^o (0.17)
4	70^o (0.34)	40^o (0.77)	35^o (0.82)	0^o (1)	75^o (0.26)
5	20^o (0.94)	55^o (0.57)	80^o (0.17)	75^o (0.26)	0^o (1)

Neste caso, *nmax* = min{3, 5} = 3, assim, o procedimento min-max é aplicado para selecionar de uma até três variáveis. Usando o pseudo-código apresentado acima, os subconjuntos iniciando com x_1 são obtidos da seguinte forma: inicialmente $V_{1,1} = \{x_1\}$. De acordo com a coluna na **Tabela 2.1** correspondente a x_1 (i = 1), a segunda variável a ser selecionada é x_3, porque $C_{1,3}$ é o menor valor de cosseno desta coluna. Desta forma, $V_{1,2} = \{ V_{1,1} , x_3 \} = \{ x_1 , x_3 \}$. A seleção da terceira variável feita entre as variáveis restantes (x_2, x_4, x_5) é baseada no valor do cosseno nas colunas correspondentes a x_1 (i = 1) e x_3 (i = 3). Considerando estas duas colunas, os cossenos máximos para x_2, x_4, x_5 são 0,87, 0,82, 0,94, respectivamente. Assim, de acordo com o critério min-max, x_4 é selecionada e $V_{1,3} = \{ V_{1,2} , x_4 \} = \{ x_1 , x_3 , x_4 \}$. Um procedimento similar é usado para gerar os subconjuntos iniciando com x_2, x_3, x_4, x_5. Vale notar que todos os cossenos deste exemplo são positivos, caso contrário, o valor absoluto pode ser usado. Ao final deste processo, gera-se uma matriz contendo os índices das variáveis. Para esse exemplo, a matriz seria:

$$M = \begin{vmatrix} 1 & 3 & 4 \\ 2 & 3 & 5 \\ 3 & 1 & 4 \\ 4 & 5 & 2 \\ 5 & 3 & 2 \end{vmatrix}$$

As variáveis de cada sub-conjunto, apesar de serem menos correlacionadas, podem apresentar uma multicolinearidade espressiva dependendo do conjunto de dados. Para assegurar que as variáveis finais selecionadas pelo ASA tenham multicolinearidade desprezível, utiliza-se a Análise de Inflação de Variância (Variance Inflation Factors-VIF) como critério para minimizar a multicolinearidade entre as variáveis menos correlacionadas. Descartam-se, então, as variáveis que apresentam um valor de VIF a partir de um determinado limiar. No exemplo da **Figura.2.1** e **Tabela 2.1**, o VIF seria calculado para as variáveis nos subconjuntos: {1, 3}, {2, 3}, {3, 1}, {4, 5}, {5, 3}, {1, 3, 4}, {2, 3, 5}, {3, 1, 4}, {4, 5, 2}, {5, 3, 2}.

Após este procedimento, os subconjuntos com as variáveis mantidas, além das variáveis individuais (de 1 até 5, para este exemplo) serão empregados para construir os modelos MLR-VIF. O resultado dos modelos é comparado em termos do erro médio quadrático obtido em um conjunto de validação independente (RMSEV). Para as N_{val} amostras, o RMSEV é definido como:

$$\mathbf{RMSEV} = \sqrt{\frac{\mathbf{1}}{\mathbf{N_{val}}} \sum_{n=1}^{N_{val}} (\mathbf{y_n} - \hat{\mathbf{y}}_\mathbf{n})^2} \qquad (24)$$

onde y_n e $\hat{y}_n$ são os valores de referência e previstos do parâmetro determinado na n^{th} amostra de validação. O subconjunto que produz o menor valor de RMSEV é então escolhido.

Finalmente, um procedimento é realizado para reduzir o número de variáveis do subconjunto escolhido na fase anterior, como sugerido por Galvão et al. (2007). Esta etapa final é empregada para a obtenção de modelos mais parcimoniosos. O procedimento de eliminação consiste em descartar variáveis até que ocorra um aumento significativo no RMSEV de acordo com um teste-F. Esse critério é similar ao método de Haaland e Thomas (1998) empregado no contexto da modelagem PLS. No presente trabalho, um nível de significância α = 0,25 para o teste-*F* é adotado, conforme sugerido por Li et al (2005).

Resumidamente, a implementação do ASA compreende cinco etapas básicas:

Etapa 1: Cálculo dos cossenos $C_{i,j}$ para j = 1, …, J ; i = (j + 1),…, J na matriz $\mathbf{X_{cal}}$ centrada na média;

Etapa 2: Geração dos subconjuntos de variáveis $V_{j,n}$ para j = 1, …, J; n = 1,…, *nmax* com menor correlação par a par;

Etapa 3: Cálculo do VIF e minimização da multicolinearidade eliminando as variáveis com VIF maior que o limiar adotado;

Etapa 4: Seleção do melhor subconjunto (variáveis correlacionadas com o parâmetro de interesse, matriz **Y**) baseado no menor valor de RMSEV;

Etapa 5: Procedimento final de eliminação de variáveis.

2.2 Programa ASA

O programa ASA foi escrito em linguagem MATLAB® 6.5 e possui o código-fonte apresentado no ANEXO 7.2. Ele consiste de três rotinas: uma principal (**Rotina 1**) onde é realizada a seleção das variáveis, construído o modelo MLR e validado com um conjunto independente; a rotina (**Rotina 2**) que é usada para previsão de novas amostras; uma terceira (**Rotina 3**) onde as variáveis escolhidas são transformadas na unidade de interesse (por ex., comprimento de onda em nm) e são plotadas no espectro médio do conjunto de calibração e também apresenta o gráfico com os valores do VIF. Além dessas rotinas, o programa ASA utiliza uma sub-rotina para o cálculo do VIF. Essas rotinas são ilustradas nos fluxogramas a seguir.

A **Figura 2.2** mostra o fluxograma da **Rotina 1** do programa ASA. Esta rotina inicia com a entrada das matrizes dos espectros dos conjuntos de calibração e validação ($\mathbf{X}_{cal}$, $\mathbf{X}_{val}$), assim como das respectivas concentrações ($\mathbf{Y}_{cal}$, $\mathbf{Y}_{val}$), além do número máximo de variáveis a serem selecionadas (N) e o limite de VIF utilizado.

Para evitar variáveis com norma igual a zero, realiza-se uma eliminação destas possíveis variáveis. As variáveis mantidas são centradas na média.

Após a centralização da média, calculam-se os ângulos entre todas as variáveis da matriz $\mathbf{X}_{cal}$. Estes ângulos são armazenados em uma matriz onde a primeira linha contém os ângulos entre a primeira variável e todas as demais, na segunda linha, os ângulos entre a segunda variável e as outras, e assim por diante, formando, portanto, uma matriz diagonal, conforme exemplo mostrado na **Tabela 2.1**.

A próxima etapa é montar as cadeias com as variáveis minimamente correlacionadas, utilizando o procedimento mínimo-máximo.

Para os subconjuntos das cadeias formadas serão calculados os valores do VIF de cada variável, eliminando aquelas que apresentarem VIF > limiar adotado.

Cada cadeia de variáveis não-colineares é utilizada para construir modelos MLR, que são validados com a matriz $\mathbf{X_{val}}$, e são armazenados os valores de $RMSEP_{val}$. A cadeia que produz o menor valor de $RMSEP_{val}$ é então selecionada.

Com as variáveis selecionadas, é novamente construído e validado um modelo MLR.

Após esta seleção, é utilizado o critério para eliminar possíveis variáveis, buscando um modelo mais parcimonioso, conforme descrito na Seção 2.1. É calculado um critério de relevância de cada variável, multiplicando-se o desvio-padrão pela norma do coeficiente do modelo, relativo a cada variável. As variáveis são então colocadas em ordem decrescente de relevância e são construídos modelos MLR, partindo da variável de maior relevância e, seqüencialmente, adicionadas as outras variáveis. Para cada modelo construído é feita a validação e calculado o $RMSEP_{val}$. O teste-F é realizado, caso não haja diferença significativa entre os valores de $RMSEP_{val}$, a variável adicionada é eliminada, caso contrário, a variável é selecionada.

Após este procedimento, as variáveis selecionadas são usadas para construir novamente o modelo de calibração, que é validado e armazenado numa variável de estrutura, contendo todas as informações, inclusive aquelas necessárias nas outras rotinas.

A **Rotina 2**, utilizada para fazer previsão de novas amostras, é vista na **Figura 2.3**. Os dados de entrada são o *Modelo* construído na etapa anterior e a matriz de previsão, $\mathbf{X_{prev}}$ e de concentração, $\mathbf{Y_{prev}}$, se disponível.

Na matriz $\mathbf{X_{prev}}$ é feito o corte das variáveis que foram eliminadas na Rotina 1, usando o critério da norma zero.

Caso os valores da matriz $\mathbf{Y_{prev}}$ sejam conhecidos, é feita a centralização na média das duas matrizes ($\mathbf{X_{prev}}$ e $\mathbf{Y_{prev}}$), usando a média das matrizes de calibração ($\mathbf{X_{cal}}$ e $\mathbf{Y_{cal}}$). O modelo é então utilizado para realizar a previsão das amostras, e os parâmetros **Y** previsto, $RMSEP_{prev}$ e correlação, são obtidos. Estes dados são armazenados numa variável de estrutura.

Caso não se disponha da matriz com os valores de $\mathbf{Y_{prev}}$, a centralização na média é feita apenas na matriz $\mathbf{X_{prev}}$, e o modelo faz a previsão desta matriz, fornecendo os valores de **Y** previsto. Esta matriz é armazenada na variável de estrutura.

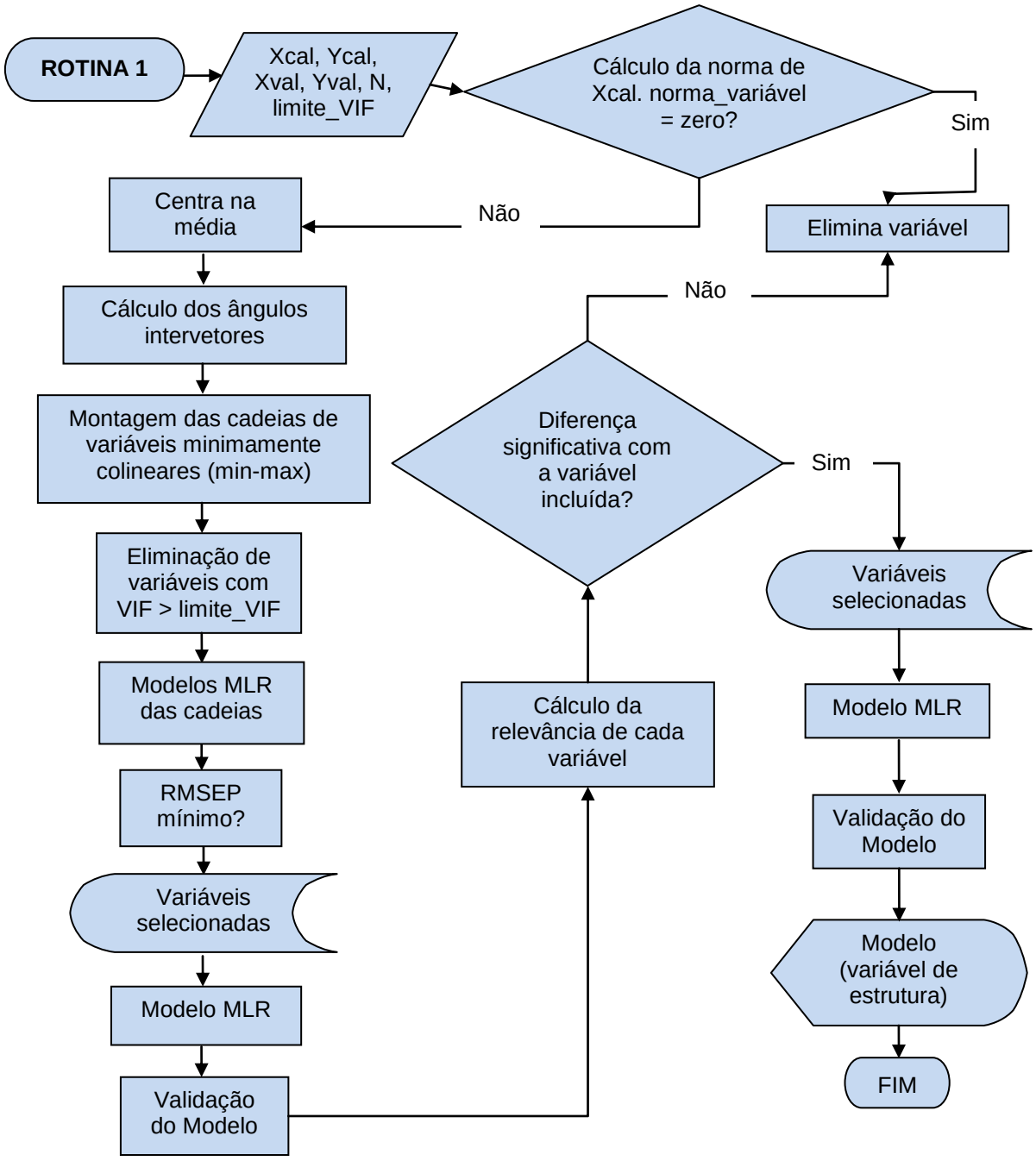

Figura 2.2. Fluxograma da Rotina 1 do programa ASA.

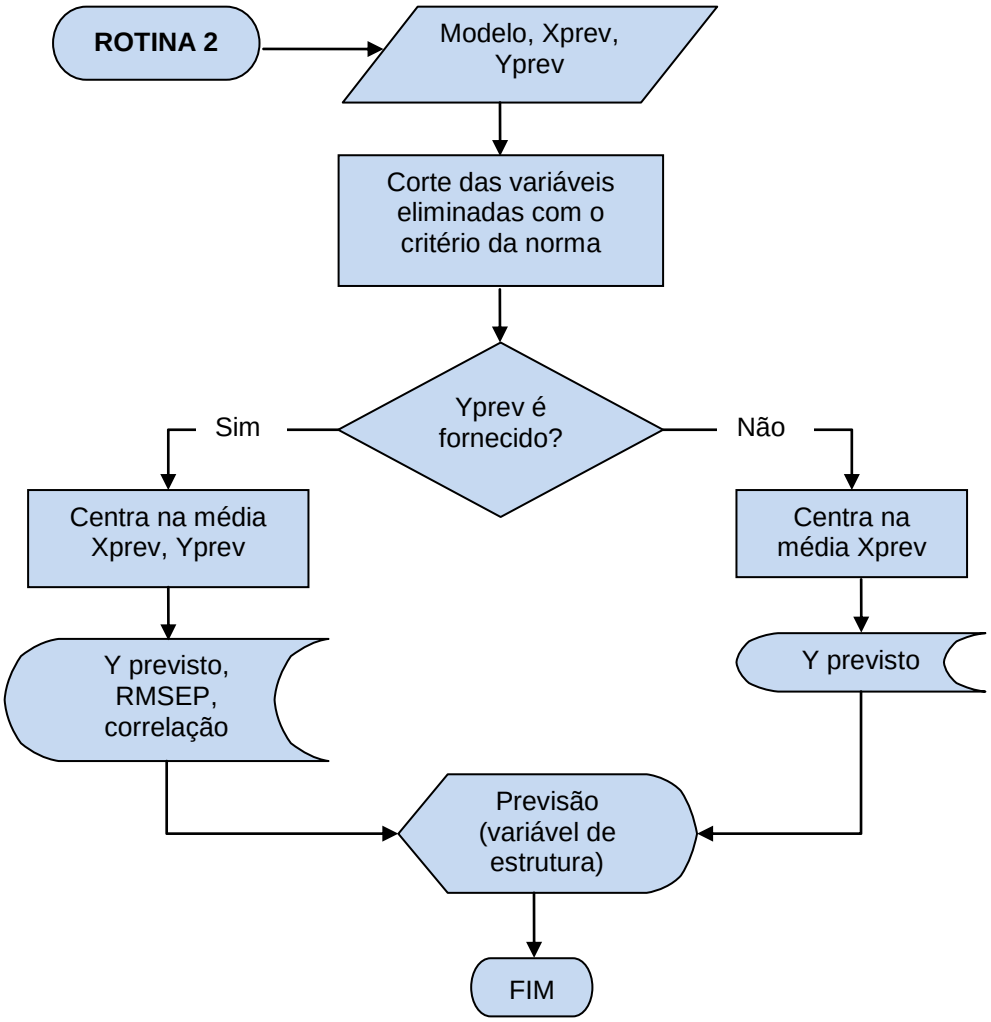

Figura 2.3. Fluxograma da Rotina 2 do programa ASA.

A **Rotina 3** é mostrada na **Figura 2.4**. Os dados de entrada são o *modelo*, *xaxis*, *lamb_inicial* e *res*. Do *modelo* é extraída a informação das variáveis selecionadas, porém em ordem numérica. O vetor *xaxis* contém os valores do eixo x na unidade de trabalho. Para fazer a transformação das variáveis numéricas para a unidade desejada, é fornecido o valor inicial desta unidade, contido na variável *lamb_inicial*. Outra informação necessária para esta transformação é a resolução das unidades, que é então armazenada na variável *res*.

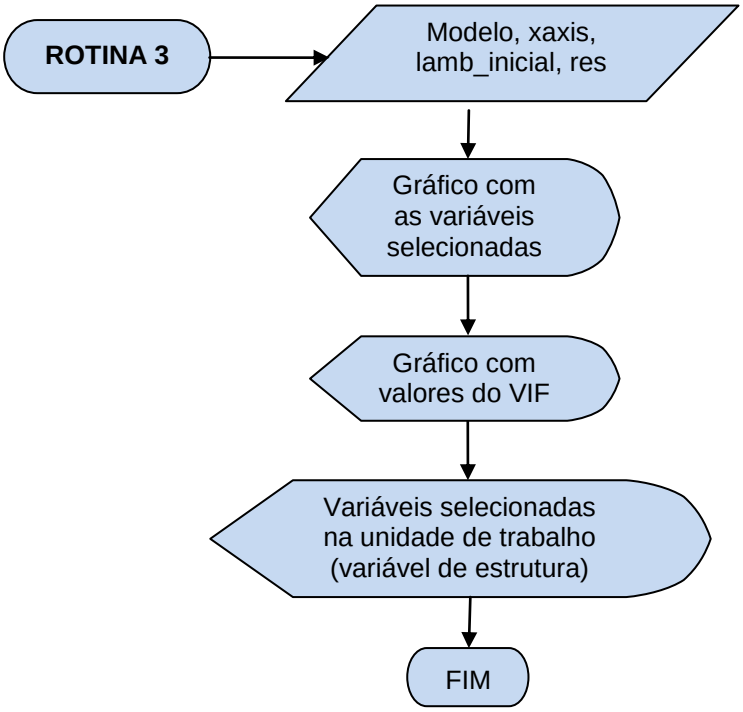

Figura 2.4. Fluxograma da Rotina 3 do programa ASA

2.3 Apresentação dos resultados e ferramentas de diagnóstico

A execução das rotinas do programa ASA apresenta vários resultados, que serão ilustrados a seguir, assim como algumas ferramentas de diagnóstico para avaliação do modelo de calibração e previsão:

2.3.1 Apresentação dos resultados da execução da Rotina 1

2.3.1.1 Modelo de calibração

O modelo de calibração é mostrado no quadro abaixo:

```
modelo =
            modelo: 'ASA_MLR'
             corte: [1x371 double]
   pre_processamento: 'centragem na media'
        media_Xcal: [1x371 double]
        media_Ycal: 5.7500
              coef: [4x1 double]
            No_var: 4
               var: [332 296 371 255]
             RMSEP: [0.0646 1.3379]
        correlacao: 0.9995
          valor_N: 20
              VIF: [4x1 double]
```

O modelo assim obtido apresenta as seguintes informações:

- *modelo* - nome do modelo (variável de estrutura contendo todas as variáveis do modelo);
- *corte*: variáveis mantidas (variáveis que permanecem após o corte de variáveis nulas);
- pré-processamento: as matrizes são centradas na média;
- *media_Xcal e media_Ycal* - média da matriz de espectros e respostas (usadas para centrar na média a matriz de previsão, Rotina 2);
- *coef* - coeficientes do modelo;
- *var* - variáveis selecionadas (em ordem numérica, e não em comprimento de onda);
- *RMSEP* - erro de previsão da matriz de validação (absoluto e relativo);
- *correlação* - coeficiente de correlação entre valores de referência e previstos da matriz de validação;
- *valor_N* - valor do número máximo de variáveis que podem ser selecionadas;
- VIF – Valores do VIF de cada variável.

Para se ter acesso a qualquer variável do modelo é necessário digitar o seguinte comando no Matlab: [nome do modelo].[variável]. Por exemplo, fazendo: modelo.coef, obtém os valores dos coeficientes do modelo.

2.3.1.2 Gráfico dos resíduos de concentração das amostras de calibração

Este gráfico mostra os resíduos dos valores de concentração de referência ($\mathbf{Y_{cal}}$) e os valores da concentração prevista pelo modelo ($\mathbf{Y_{estimado}}$) para as amostras de calibração. Com as variáveis selecionadas, o modelo é determinado pela expressão:

$$\mathbf{b} = (\mathbf{X_{cal}}'\mathbf{X_{cal}})^{-1}\mathbf{X_{cal}}'\mathbf{Y_{cal}}$$

e

$$\mathbf{Y_{estimado}} = \mathbf{X_{cal}}*\mathbf{b}$$

Portanto, o resíduo é calculado por: $\mathbf{Y_{cal}} - \mathbf{Y_{estimado}}$

Este resíduo é plotado versus o número de amostras, conforme ilustrado na **Figura 2.5**.

Através desta ferramenta de diagnóstico podemos verificar se alguma amostra do conjunto de calibração pode ser considerada anômala (outliers).

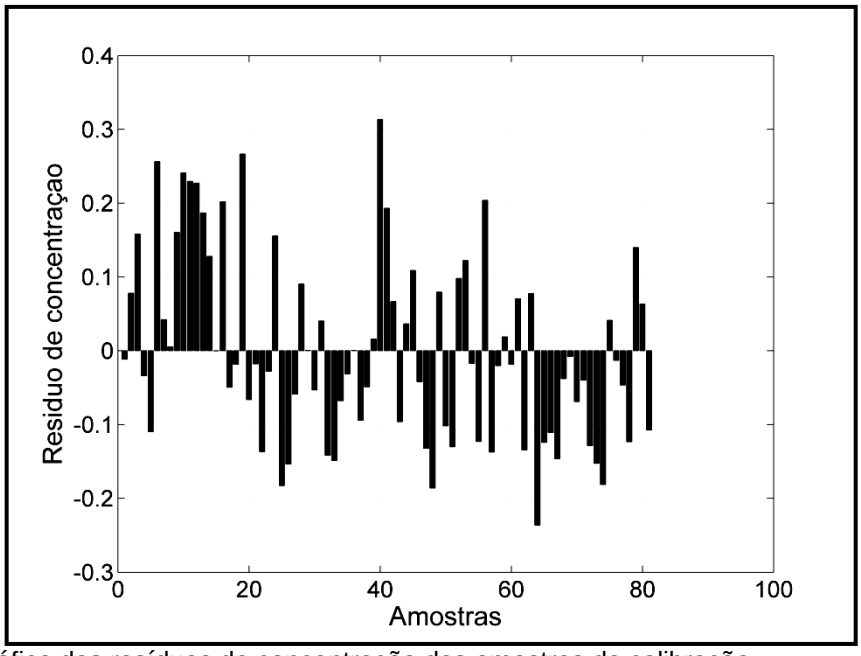

Figura 2.5. Gráfico dos resíduos de concentração das amostras de calibração.

2.3.1.3 Gráfico Scree plot

Este gráfico mostra o comportamento das variáveis selecionadas em função do RMSEV, no procedimento de eliminação de variáveis baseado no critério de Haaland e Thomas (1998).

Para este exemplo ilustrativo, o número de variáveis selecionadas seria cinco (**Figura 2.6**), porém, a diferença nos valores de RMSEV não é estatisticamente significante (no nível aqui adotado de 75%) após a quarta variável. Desta forma, o número de variáveis selecionadas foram quatro, produzindo um modelo mais parcimonioso.

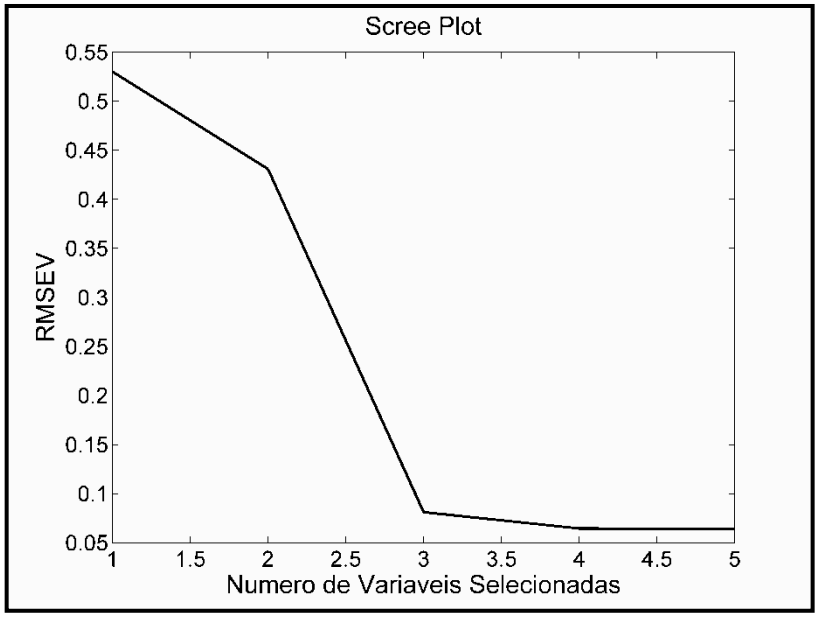

Figura 2.6. Gráfico Scree plot.

2.3.1.4 Gráfico de valores previstos versus valores de referência para as amostras de validação

Esta é outra ferramenta de diagnóstico, que mostra o comportamento dos valores de concentração previstos pelo modelo de calibração e os valores de referência para o conjunto de validação. Os valores devem se distribuir em torno da bissetriz, de forma aleatória, conforme a **Figura 2.7**. Esta figura também mostra os valores de RMSEV e correlação.

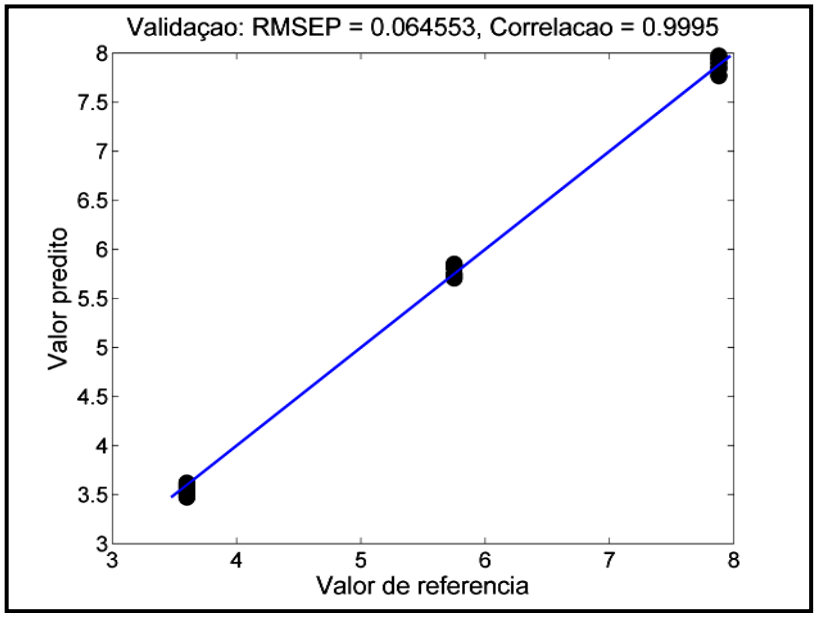

Figura 2.7. Valores previstos versus valores de referência para as amostras de validação.

2.3.2 Apresentação dos resultados da execução da Rotina 2

Esta Rotina é elaborada para fornecer os valores de concentração de novas amostras, após o modelo está definitivamente validado e pronto para ser usado em análise de rotina.

2.3.2.1 Previsão de novas amostras

A previsão de novas amostras, que não fizeram parte da obtenção do modelo, mas o qual dispõe dos valores de referência, aqui denominado de conjunto de previsão, é obtida com a execução da Rotina 2, mostrando um resultado da seguinte maneira:

```
previsão =
previsao: 'ASA_MLR'
    Yestimado: [30x1 double]
        RMSEP: [0.0782 1.9950]
   correlacao: 0.9995
```

- *Previsão* – variável de estrutura que armazena os resultados da previsão
- *Yestimado* - valores previstos pelo modelo para o parâmetro determinado;
- *RMSEP* - erro de previsão da matriz de previsão (absoluto e relativo);
- *correlação* - coeficiente de correlação entre valores de referência e previstos da matriz de previsão.

Este resultado pode ser considerado uma validação externa do modelo de calibração.

2.3.2.2 Gráfico de valores previstos versus valores de referência para as amostras de previsão

Esta ferramenta de diagnóstico (**Figura 2.8**), lembrando que só é fornecida se o conjunto dispuser dos valores de referência, mostra a relação entre os valores previstos e os valores de referência, com o mesmo significado da **Figura 2.7**, porém para o conjunto de previsão.

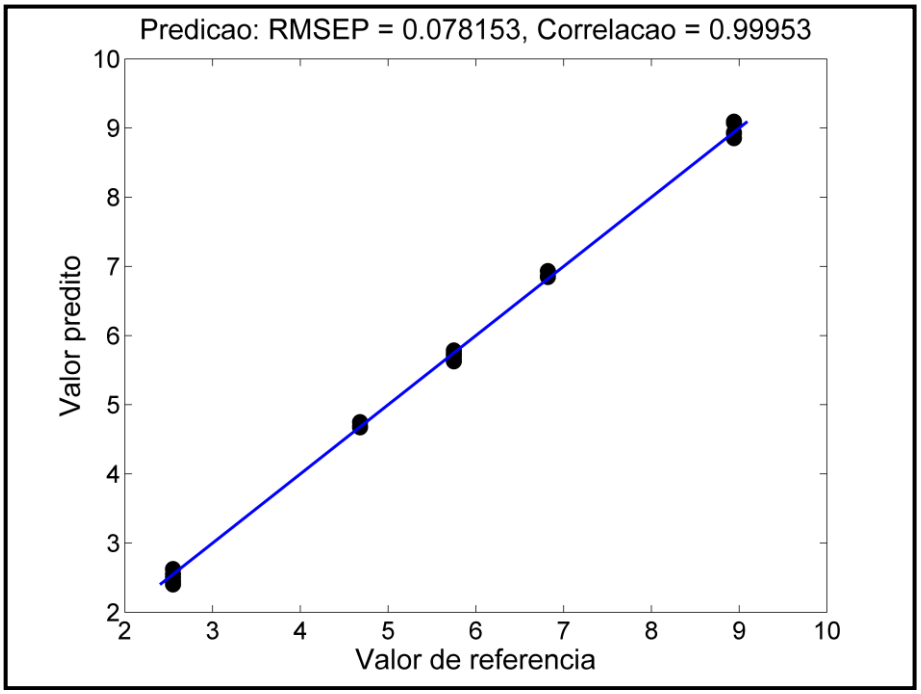

Figura 2.8. Valores previstos versus valores de referência para as amostras de previsão.

2.3.3 Apresentação dos resultados da execução da Rotina 3

Esta rotina fornece os valores das variáveis na unidade em que os dados foram obtidos. Para este exemplo, as variáveis são dadas em nanômetro.

```
var_sel =

   variaveis_selecionadas: [484 525 561 600]
```

Estas variáveis são mostradas também de forma gráfica, facilitando sua visualização no espectro (**Figura 2.9**). Nesta mesma figura, são mostrados os valores do VIF para as variáveis selecionadas.

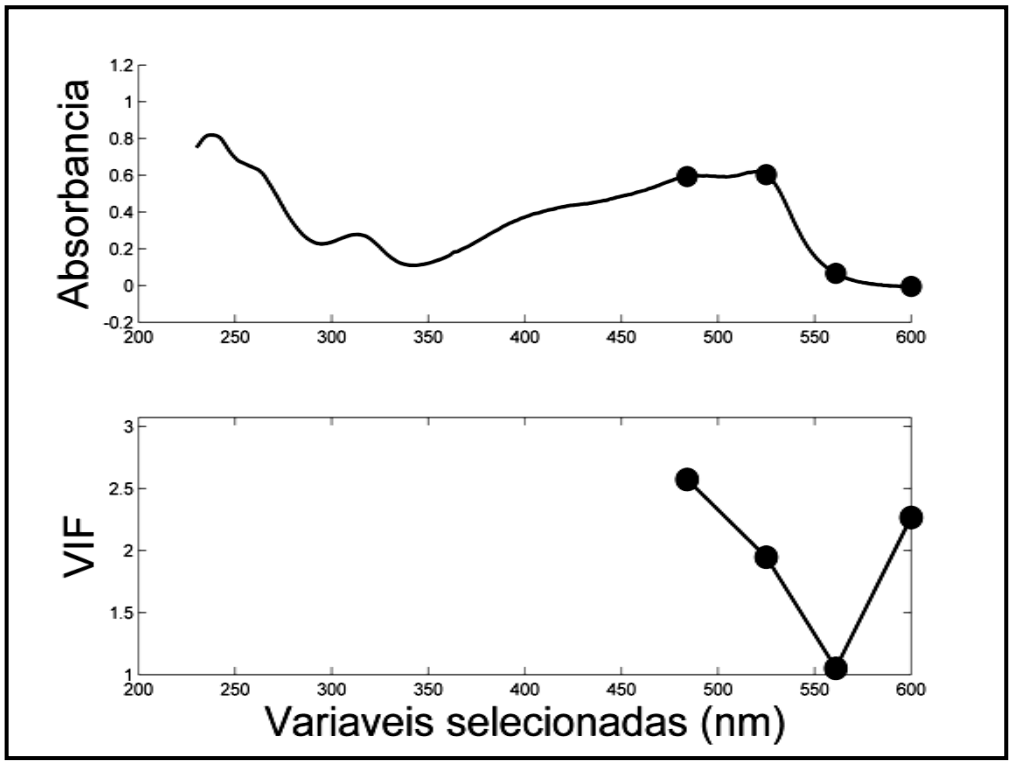

Figura 2.9. Variáveis selecionadas pelo ASA-VIF e respectivos valores de VIF.

CAPÍTULO 3

EXPERIMENTAL

3. EXPERIMENTAL

O desempenho do algoritmo proposto, usando o critério VIF (algoritmo ASA-VIF) e sem o uso do VIF (algoritmo ASA) foi avaliado por intermédio de quatro conjuntos de dados obtidos por duas técnicas espectroanalíticas. Um dos conjuntos consiste de dados de espectrometria absorção molecular UV-VIS de misturas de quatro corantes alimentícios sintéticos. Os outros três conjuntos foram obtidos por espectrometria NIR: um conjunto de dados de trigo, um de milho e outro de gasolina. Para fins de comparação, todos os resultados foram também avaliados por outras técnicas de seleção de variáveis e regressão linear múltipla: MLR-APS, MLR-AG e MLR-Stepwise (MLR-SW), implementados em MATLAB® 6.5. Também comparamos os resultados do MLR-ASA e MLR-ASA-VIF com o PLS, que é um método que usa todos os dados do espectro.

3.1 Dados espectrométricos UV-VIS de misturas de corantes

Neste trabalho, utilizamos um conjunto de dados resultante de espectros de absorção molecular UV-VIS de uma mistura de quatro corantes sintéticos (tartrazina, vermelho 40, amarelo crepúsculo, e eritrosina) usados em produtos alimentícios. Cada corante foi determinado individualmente pelos modelos de calibração.

Soluções estoques dos corantes com concentração 1000 mg L^{-1} foram preparadas em solução tampão fosfato de sódio monobásico anidro/hidróxido de sódio de pH 7,0. Essas soluções foram diluídas, manualmente, no mesmo tampão de modo atingir a concentração de 40 mg L^{-1}.

As amostras dos quatro corantes puros foram cedidas pela empresa Plury Química LTDA (Diadema – SP).

A água utilizada na preparação das soluções-tampão era sempre recém destilada e deionizada.

Os experimentos foram realizados no Laboratório de Automação e Instrumentação em Química Analítica/Quimiometria (LAQA) – UFPB.

Foram definidos três conjuntos de dados, um para calibração, outro para validação (validação interna) e um terceiro para previsão (validação externa).

A cor dos produtos alimentícios é uma de suas principais características, pois está associada à imediata percepção e avaliação. A aparência, segurança, aceitabilidade e características sensoriais dos alimentos são todas afetadas pela cor.

Embora esses efeitos sejam puramente psicológicos, eles interferem na escolha dos produtos.

Alguns alimentos industrializados não apresentam cor naturalmente, e outros, têm sua cor destruída durante o processamento e/ou estocagem (PRADO E GODOY, 2004). Com isso o uso de corantes sintéticos ou naturais é um recurso amplamente utilizado em diversos produtos, tais como, biscoitos, cereais, sorvetes, bebidas, queijos, produtos de confeitaria, entre outros, apesar da polêmica em torno de seu uso, por trazer danos à saúde de consumidores sensíveis a estes produtos. Os principais objetivos do uso de corantes são:

- compensar a perda de cor decorrente da exposição à luz, ar, temperaturas extremas, umidade e armazenamento do produto;
- corrigir variações naturais da cor (*alimentos sem cor são associados, geralmente, e de forma incorreta, à baixa qualidade)*;
- intensificar as cores, o que ocorre naturalmente, porém em níveis mais baixos; estabelecer uma identidade de cor entre o produto final, que a princípio seria descolorido, e o alimento que o origina *(exemplo: a adição de corantes vermelhos a sorvetes de sabor morango)*;
- proteger o sabor e o valor nutritivo de alimentos que poderiam ser afetados pela incidência de luz durante o armazenamento.

3.1.1 Conjunto de calibração

O conjunto de calibração foi definido de acordo com um planejamento fatorial completo de 3 níveis e 4 fatores (3^4), totalizando 81 misturas (CALADO et al., 2003). Os níveis de concentração considerados encontram-se na **Tabela 3.1**.

Tabela 3.1. Níveis de concentração (mg L^{-1}) dos corantes nas misturas de calibração

Corante	Planejamento Fatorial completo 3^4		
	Níveis		
	-1	0	1
Tartrazina	2,0	6,0	10,0
Vermelho 40	1,5	5,8	10,0
Amarelo crepúsculo	1,5	5,8	10,0
Eritrosina	0,5	3,8	7,0

3.1.2 Conjunto de validação

O conjunto de validação foi obtido conforme um planejamento fatorial fracionário 3^{4-1}, contendo 27 misturas. Na **Tabela 3.2** são apresentados os valores da concentração de cada corante.

Tabela 3.2. Níveis de concentração (mg L^{-1}) dos corantes nas misturas do conjunto de validação.

Corante	Planejamento Fatorial fracionário 3^{4-1}		
	Níveis		
	-1	0	1
Tartrazina	4,0	6,0	8,0
Vermelho 40	3,6	5,8	7,9
Amarelo crepúsculo	3,6	5,8	7,9
Eritrosina	2,1	3,8	5,4

3.1.3 Conjunto de previsão

Um terceiro conjunto, contendo 30 amostras, foi obtido para verificar a capacidade preditiva do modelo de calibração. Este conjunto foi preparado com concentrações, dentro da faixa de calibração, distribuídas aleatoriamente (sorteio). Os valores encontram-se na **Tabela 3.3**.

Tabela 3.3. Concentração dos corantes nas misturas do conjunto de previsão.

Corante	Concentração [mg/L]				
Tartrazina	3,0	5,0	6,0	7,0	9,0
Vermelho 40	2,6	4,7	5,8	6,8	8,9
Amarelo crepúsculo	2,6	4,7	5,8	6,8	8,9
Eritrosina	1,3	2,9	3,8	4,6	6,2

Como pode ser observado, o número de experimentos para obtenção dos conjuntos dos corantes é extremamente elevado. 81 amostras de calibração, mais 27 de validação e mais 30 de previsão. Como realizamos em duplicata, o total de misturas é de 276. Isto se torna, obviamente, impraticável de ser realizado da forma clássica (manual). Por este motivo, utilizamos um sistema automático em fluxo-batelada, em virtude das vantagens inerentes deste método, descritas na **Seção 1.6**.

3.1.4 Descrição do sistema de Análise em Fluxo-Batelada

O sistema em fluxo-batelada utilizado é composto de seis válvulas solenóides de três vias e uma câmara de mistura. As válvulas V_1, V_2, V_3 e V_4 (uma para cada

corante) e V_{TP} permitem a introdução das soluções dos corantes e do tampão TP (utilizado como diluente e fluido de limpeza) na câmara de mistura. A válvula V_M permite conduzir a mistura, após homogeneização, da câmara para uma cela de fluxo no espectrofotômetro onde são registrados os espectros UV-VIS. Um esquema simplificado do sistema é mostrado na **Figura 3.1**.

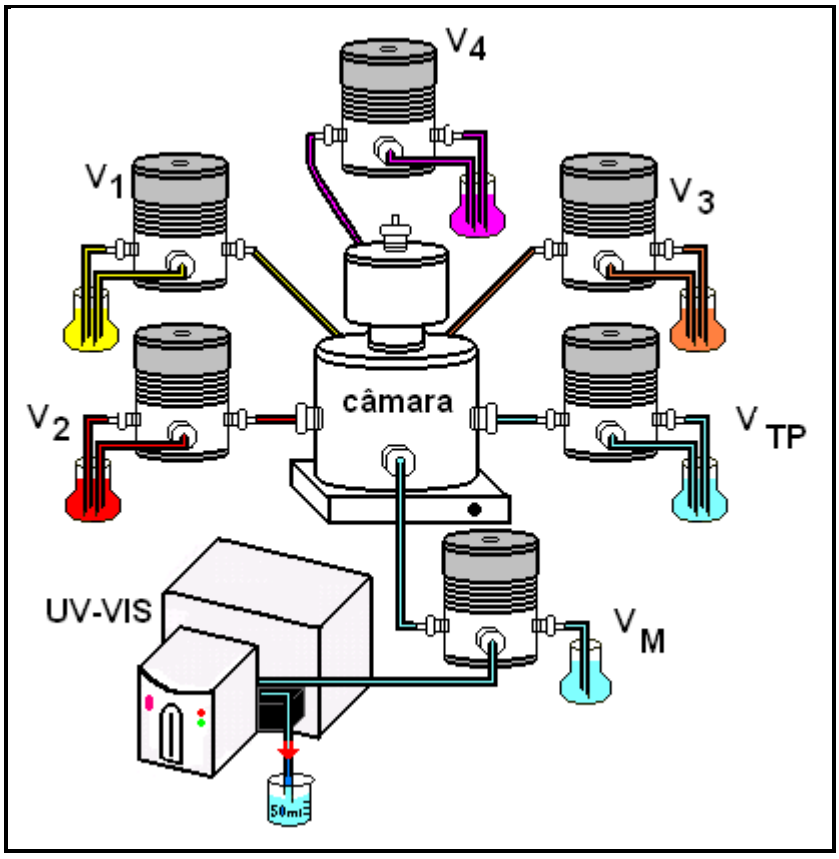

Figura 3.1. Esquema simplificado do sistema em fluxo-batelada.

O fluxo é obtido por intermédio de uma bomba peristáltica (não mostrada no esquema) que fica localizada entre os recipientes (amostras, diluentes) e as válvulas. Para os canais associados às válvulas V_1, V_2, V_3, V_4 e V_{TP}, o fluxo é aspirado, ou seja, é na direção do recipiente para a câmara. No canal da válvula V_{TP} que transposta mistura para o espectrofotômetro, o fluxo é bombeado em sentido contrário. A vazão desses fluxos depende do diâmetro dos tubos utilizados e da rotação da bomba peristáltica. Como algumas variáveis podem afetar a vazão em cada canal do sistema de maneira diferente (por exemplo, diferenças nos tamanhos e diâmetros dos tubos e nos tempos de abertura das válvulas), é necessário realizar, inicialmente, uma calibração para se obter a vazão correta em cada canal. Neste

caso, a calibração foi realizada com fluxo de água destilada, sendo cada canal calibrado individualmente.

A etapa de calibração é realizada da seguinte maneira: tomam-se diferentes tempos (t) de acionamento da válvula e o volume de água transportada é recolhida em um béquer, colocado em uma balança analítica, obtendo-se o peso desta água. Este procedimento é feito em triplicata. Com a massa média e a densidade da água, determina-se o volume (V). Plota-se em um gráfico o volume versus o tempo de acionamento da válvula, e a vazão (q) é obtida pelo coeficiente da reta de regressão: V = q t.

O tempo de acionamento das válvulas é controlado por um programa escrito em linguagem visual *LabView 5.1*. Para este experimento, o programa apresenta a tela principal mostrada na **Figura 3.2** (o diagrama de comandos não é aqui mostrado).

Figura 3.2. Controle de tempo do sistema em fluxo-batelada

Os tempos de cada corante e tampão (usado como diluente) são digitados para cada mistura (amostra) na primeira linha da tela (**Figura 3.2**), de acordo com seus valores, que são calculados da seguinte maneira: calcula-se o volume inicial de cada corante na mistura através da equação: Vi = (Cf * Vf)/Ci, onde Vi = volume inicial, Cf = concentração final, Vf = volume final e Ci = concentração inicial. De

posse da vazão(q) de cada canal, obtido na etapa de calibração, o tempo de acionamento de cada válvula é dado então pela expressão: t = Vi/q. O volume final (Vf) da câmara foi mantido constante em 1,6 mL.

Após o final da mistura, esta fica na câmara durante 10s para homogeneização, depois teste tempo, a válvula DC é acionada durante 20s, e a bomba é desligada. Esse tempo é suficiente para a amostra estar presente na cela de fluxo. Neste momento é registrado o espectro. Como ainda resta amostra na câmara, a válvula DC é acionada novamente durante 15s para descarregar toda a amostra (estes tempos são digitados na segunda linha da tela da **Figura 3.2**).

A etapa de limpeza, realizada após o final de cada amostra, é feita pelo acionamento da válvula TP durante 32s, depois, a válvula DC é acionada por 35s para esvaziar a câmara, que estará pronta para a próxima amostra (estes tempos são digitados na terceira linha da tela da **Figura 3.2**).

O sistema completo é visto na **Figura 3.3**.

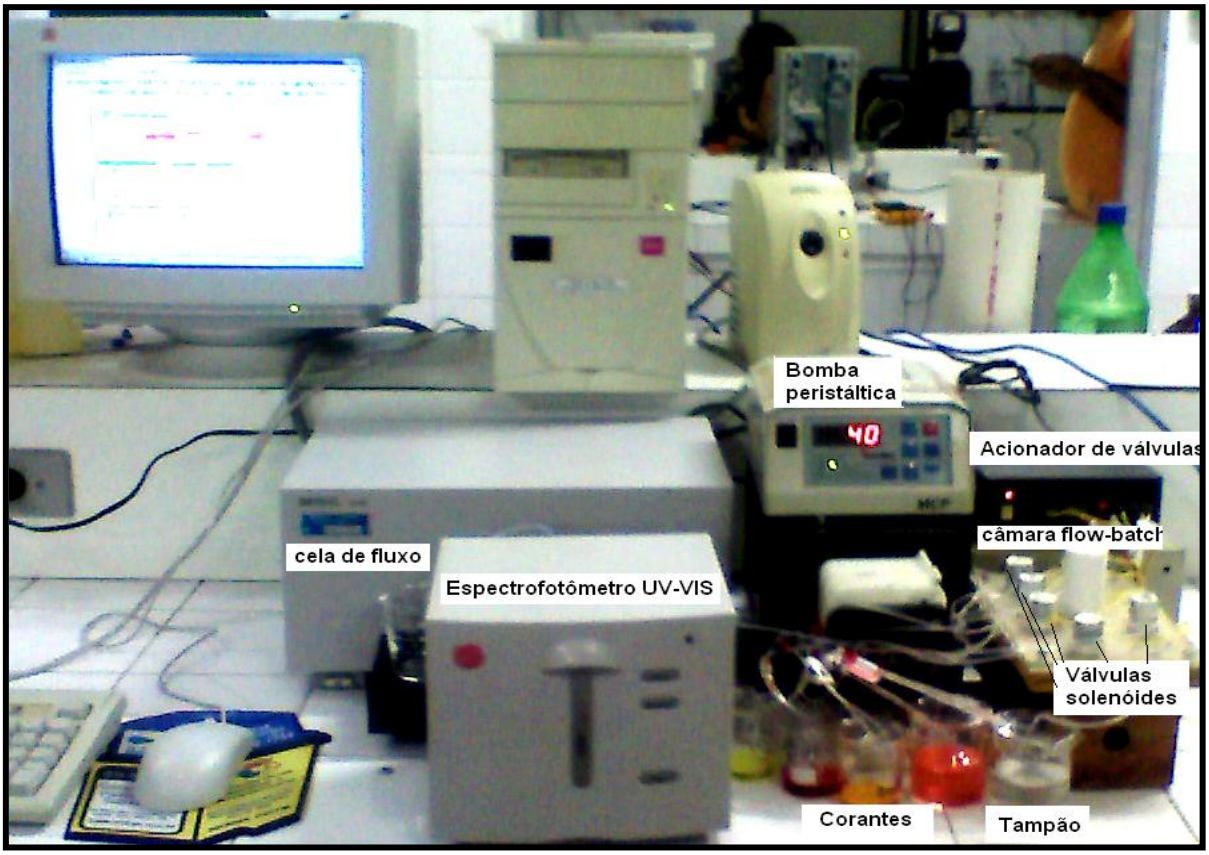

Figura 3.3. Sistema completo de análise em fluxo-batelada.

Todos os componentes do sistema de análise em fluxo-batelada são descritos a seguir:

- ***Válvulas Solenóides***

  Válvulas solenóides *three-way*, da Cole-Parmer.

- *Acionador de Válvulas*

  A abertura das válvulas solenóides foi controlada por um acionador de válvulas *lab-made*. O acionador é controlado via microcomputador através de um *software* de gerenciamento, escrito em linguagem visual *LabView 5.1*.

- ***Câmara de Mistura***

  Essa câmara foi usada para promover a mistura dos corantes com o diluente. Trata-se de um cilindro de acrílico *lab-made,* contendo em seu interior uma barra magnética. A rotação da barra magnética se dá com o auxílio de um agitador magnético localizado abaixo da câmara.

- ***Agitador Magnético***

  Para auxiliar na homogeneização dos fluidos na câmara de mistura foi usado um agitador magnético IKA, modelo 8068.

- ***Cela de Fluxo***

  Foi utilizada neste sistema uma cela de fluxo de quartzo comercial da HELLMA com caminho óptico de 1,0 cm e um volume morto de 90μl.

- ***Bomba Peristáltica***

  Os fluidos foram propulsionados utilizando uma bomba peristáltica Ismatec MCP, modelo 78002-00, da Cole-Parmer Instrument Company, com rotação de 40 rpm.

- ***Espectrofotômetro de Absorção Molecular***

  Os espectros das amostras foram registrados utilizando um Espectrofotômetro de Absorção Molecular UV-VIS, com arranjo de fotodiodos, marca Hewllet Packard, modelo 8453.

- ***Microcomputador***

Foi utilizado um microcomputador Pentium Intel 233 MHZ para controle e aquisição dos dados.

3.1.5 Calibração dos canais do sistema em fluxo-batelada

Como descrito na **Seção 3.1.4**, a obtenção dos conjuntos de dados dos corantes, usando o sistema automático em fluxo-batelada, requer uma calibração dos canais de fluxo. O resultado dessa calibração é mostrado na **Figura 3.4**.

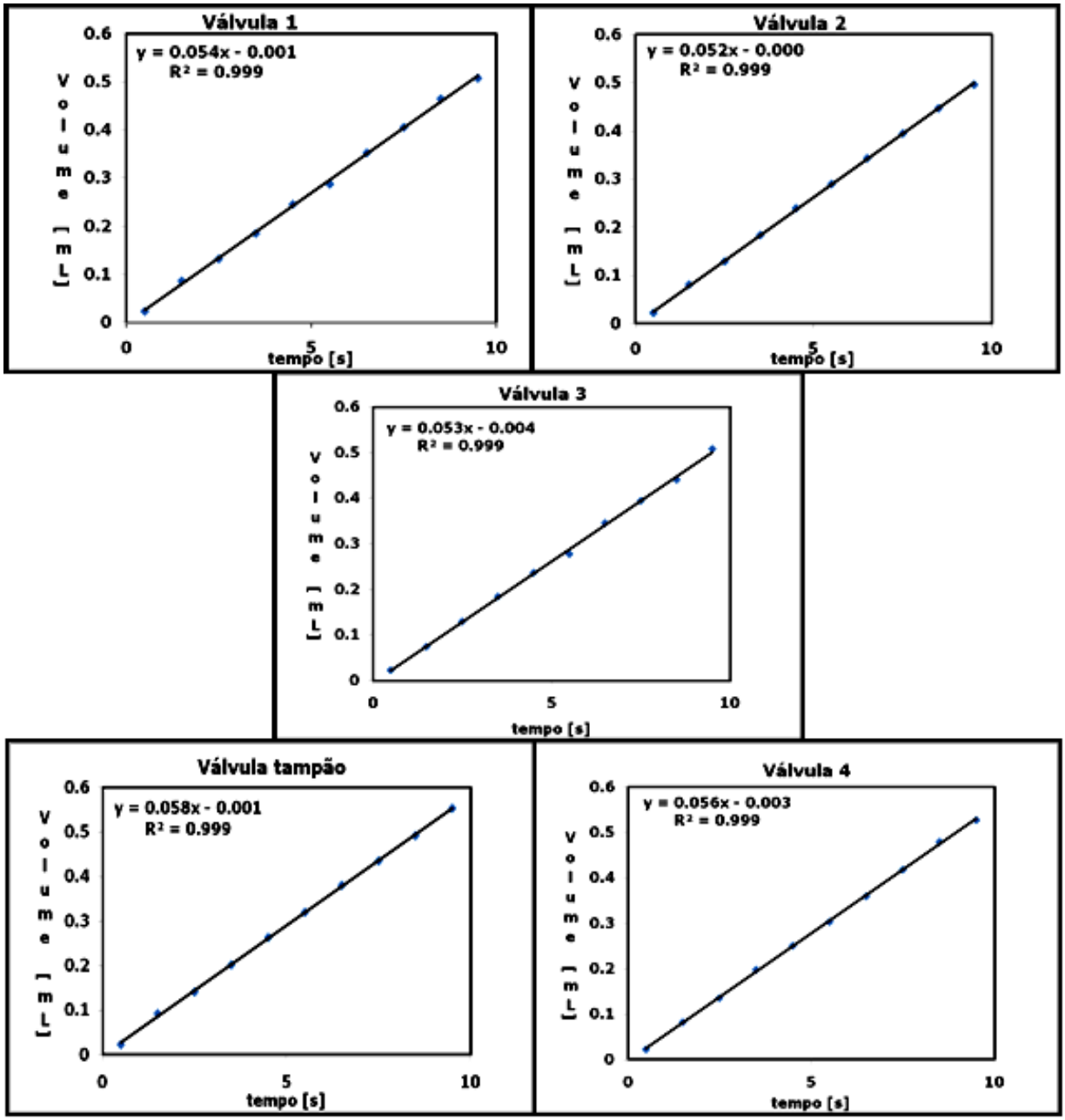

Figura 3.4. Calibração dos canais de fluxo.

Na **Figura 3.4** pode-se observar uma linearidade satisfatória dos volumes dos canais em função do tempo de abertura das válvulas, demonstrando uma boa eficiência das válvulas mesmo quando se utilizam tempos abaixo de 1s. O coeficiente angular de cada reta fornece a vazão em cada canal, cujos valores são mostrados na **Tabela 3.4**:

Tabela 3.4. Vazão dos canais de fluxo do sistema em fluxo-batelada.

	Canais de fluxo associados às válvulas				
	V_1	V_2	V_3	V_4	V_{TP}
Vazão (mL s^{-1})	0,054	0,052	0,053	0,056	0,058

A **Tabela 3.4** revela que os valores de vazão diferem pouco entre si, porém a correção dos tempos usados nas misturas, em cada canal, é fundamental para garantir exatidão na concentração desejada de cada corante na mistura.

3.2 Dados espectrométricos NIR de trigo

Esse conjunto engloba os espectros NIR de reflectância difusa de 100 amostras de farinha de trigo, obtidos na faixa de 1101 a 2502 nm com resolução de 2 nm. As propriedades medidas são proteína e umidade. Este conjunto de dados encontra-se disponível no site ftp://ftp.clarkson.edu/pub/hopkepk/Chemdata/Kalivas/ (KALIVAS, citado por FORINA et al., 2007).

O trigo contém proteínas de importância nutricional (albuminas e globulinas), pois contêm todos os aminoácidos essenciais, e aquelas de importância tecnológicas, que são as prolaminas e gluteninas, responsáveis pelas características de viscoelasticidade (glútem), adequadas na produção do pão.

O teor de umidade presentes no trigo, assim como no milho, tem grande importância na manutenção da qualidade destes cereais. Um teor de umidade elevado favorece a proliferação de diversos microrganismos (bactérias, fungos), responsáveis pela sua deterioração. A quantificação da umidade é importante também do ponto de vista econômico, visto que, a comercialização destes produtos é feita em unidades de peso.

Usamos aqui o mesmo pré-processamento descrito por Forina et al.(2007), que consiste em realizar a segunda derivada de Savitzky-Galay com polinômio de 3ª

ordem e janela de 11 pontos. Este procedimento permite a comparação dos resultados.

O algoritmo SPXY- *Sample set Partitioning based on joint X-Y distances* (GALVÃO et al., 2005) foi empregado para dividir as amostras em conjuntos de calibração (50 amostras), validação (25 amostras) e previsão (25 amostras).

As faixas de concentração para cada conjunto são mostradas na **Tabela 3.5**.

Tabela 3.5. Faixas de concentração dos conjuntos de dados NIR de trigo.

Conjunto	Propriedade	
	Proteina [%]	Umidade [%]
calibração	7,75 – 14,28	12,45 – 17,36
validação	10,36 – 12,96	12,69 – 16,59
previsão	10,59 – 14,02	12,70 – 16,94

3.3 Dados espectrométricos NIR de milho

Esses dados resultam dos espectros NIR de reflectância difusa de 80 amostras registrados na faixa de 1100 a 2498 nm com resolução de 2 nm. Esse conjunto de dados também se encontra disponível na internet por intermédio do site http://software.eigenvector.com/Data/Corn/index.html (WISE e GALLAGHER citado por FORINA et al., 2007). As propriedades estudadas neste caso foram proteína, umidade, óleo e amido.

A proteína presente no milho, embora em quantidade significante, possui qualidade inferior a de outras fontes vegetais e animais, pois é deficiente em alguns aminoácidos essenciais, tais como, metionina, lisina e triptofano.

O óleo (lipídeos) do milho é rico em ácidos graxos insaturados. Possui uma composição favorável em termos de ácidos essenciais, sendo considerado um óleo de alta qualidade. O óleo extraído do milho é uma fonte de fitosteróis, fitostanóis, ferulato éster de sitostanol e campesterol, e são utilizados como produtos redutores de colesterol (http://www.setor1.com.br/oleos/mi_lho.htm, acessado em 21/01/2008).

O amido é um polissacarídeo heterogêneo, composto de dois polímeros de glicose: a amilose e a amilopectina. O amido de milho é muito utilizado na alimentação como espessante em molhos, cremes, sopas, macarrão, biscoitos, cremes e na panificação. É utilizado ainda nas indústrias de papel e papelão, artesanato e em colas (http://www.abmamidos.com.br/, acessado em 21/01/2008).

O pré-processamento usado para estes dados foi a primeira derivada de Savitzky-Galay com polinômio de 2ª ordem e janela de 21 pontos.

Usamos o mesmo procedimento dos dados anteriores (algoritmo SPXY) para obter os conjuntos de calibração (40 amostras), validação (20 amostras) e previsão (20 amostras). As faixas de concentração são mostradas na **Tabela 3.6**.

Tabela 3.6. Faixas de concentração dos conjuntos de dados NIR de milho.

Conjunto	Propriedade			
	Proteína [%]	Umidade [%]	Óleo [%]	Amido [%]
calibração	7,654 – 9,711	9,377 – 10,993	3,088 – 3,832	62.826 - 66.472
validação	7,908 – 9,694	9,673 – 10,936	3,264 – 3,822	63.021 - 65.841
previsão	7,759 – 9,410	9,430 – 10,882	3,212 – 3,787	63.784 - 65.720

3.4 Dados espectrométricos NIR de gasolina

Este conjunto de dados contém 169 amostras de gasolina coletadas na cidade de São Paulo e analisadas no Laboratório de Química Analítica da UNICAMP. A gasolina apresenta aproximadamente 25% (v/v) de etanol, em conformidade com os padrões estabelecidos pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). As amostras de gasolina foram estocadas em frascos de vidro âmbar sob refrigeração a 5°C.

Duas propriedades da gasolina foram analisadas, MON (*Motor Octane Number*), e T90% (temperatura em que 90% da amostra são evaporadas). MON é uma importante propriedade fundamental para controle de qualidade da gasolina, pois fornece uma medida da eficiência antidetonante. T90% é uma propriedade relevante para caracterizar os componentes com alto ponto de ebulição. Altas temperaturas de ebulição para tais componentes melhoram as características antidetonantes. Entretanto, valores altos de T90% podem produzir depósitos na câmara de combustão, bem como aparecimento de goma no sistema de admissão de combustível.

Os valores de referências para MON e T90% foram obtidos de acordo com as normas da ASTM (*American Society for Testing and Materials*) D 2700 e D 86, respectivamente.

Os espectros de infravermelho próximo (NIR), na faixa de 850 a 1800 nm foram adquiridos usando um espectrômetro FT-NIR, da marca Bomen, modelo MB

160 D, com uma resolução espectral de 2 nm, com um caminho ótico de 1,0 cm (476 variáveis correspondentes aos comprimentos de onda). Para eliminar feições de linha de base nos espectros NIR, foi utilizado o método de Savitzky-Golay com janela de 21 pontos e 1ª derivada com polinômio de 2ª ordem.

O algoritmo SPXY foi empregado para dividir as amostras em conjuntos de calibração (84 amostras), validação (42 amostras) e previsão (43 amostras).

As faixas de concentração para cada conjunto são mostradas na **Tabela 3.7**.

Tabela 3.7. Faixas de concentração dos conjuntos de dados NIR de gasolina.

Conjunto	Propriedade	
	MON	T90% [°C]
calibração	82,48 – 86,98	168,2 – 189,1
validação	83,28 – 86,18	170,0 – 188,3
previsão	83,48 – 86,38	198,7 – 186,3

Veremos a seguir a descrição dos outros algoritmos utilizados neste trabalho.

3.5 Algoritmo Genético

A formulação do AG, adotada neste trabalho, codifica os subconjuntos de variáveis em cromossomos binários com comprimento igual a *J* comprimentos de onda no espectro (um "1" gene indica um comprimento de onda selecionado). A função de aptidão foi avaliada construindo um modelo de MLR com os dados de calibração restritos aos comprimentos de onda indicados no cromossomo. Cada modelo foi aplicado ao conjunto de validação e o valor de aptidão foi calculado como o inverso do RMSEV resultante. Utilizou-se também o método da roleta, no qual a probabilidade de um determinado subconjunto de variáveis seja selecionado é proporcional a sua aptidão.

Foram empregados operadores cruzamento e mutação com probabilidades de 60% e 5%, respectivamente. Cada geração foi completamente substituída pelas dos seus descendentes, tal que o tamanho de população foi mantido constante. Para evitar a perda de soluções boas foi retido sempre o melhor indivíduo (elitismo). O tamanho da população foi fixado em 100 cromossomos e o número de gerações fixado em 80. O algoritmo foi aplicado, separadamente, para cada analito em consideração. Em cada caso, o AG foi executado cinco vezes e o melhor resultado, em termos de RMSEV, foi escolhido.

3.6 Regressão Stepwise (SW)

O algoritmo começa da variável x, que tem a maior correlação com a variável dependente y. Cada repetição subseqüente do procedimento de seleção inclui uma fase de inclusão seguida por uma fase de exclusão que é guiada através de testes-F parciais para as x-variáveis.

Quatro diferentes níveis de significância para o teste-F (α = 0.01, 0.05, 0.10, 0.20) foi testado para cada propriedade y em consideração. Em cada caso, o melhor valor de α foi selecionado de acordo com o critério do RMSEV.

3.7 Algoritmo das Projeções Sucessivas (APS)

O algoritmo APS foi utilizado na sua versão GUI - *Graphical User Interface* (GALVÃO et al., 2007). Nesta versão, a calibração e a previsão são realizadas em duas etapas. O modelo de calibração é construído usando a tela apresentada na **Figura 3.5**, onde são introduzidos os dados das matrizes de calibração (**Xcal**) e de validação (**Xval**), com as respectivas concentrações (**ycal** e **yval**). Além disso, são informados os valores do número mínimo e máximo de variáveis que podem ser selecionadas. Neste trabalho, não foi utilizada a opção de autoescalonamento dos dados, apenas a centralização na média, que é "*default*" do programa.

A etapa de previsão é realizada usando a tela seguinte (**Figura 3.6**). Novamente são inseridos os dados da matriz de calibração e de previsão (**Xpred**) e suas concentrações (**ypred**), bem como as variáveis selecionadas na etapa anterior.

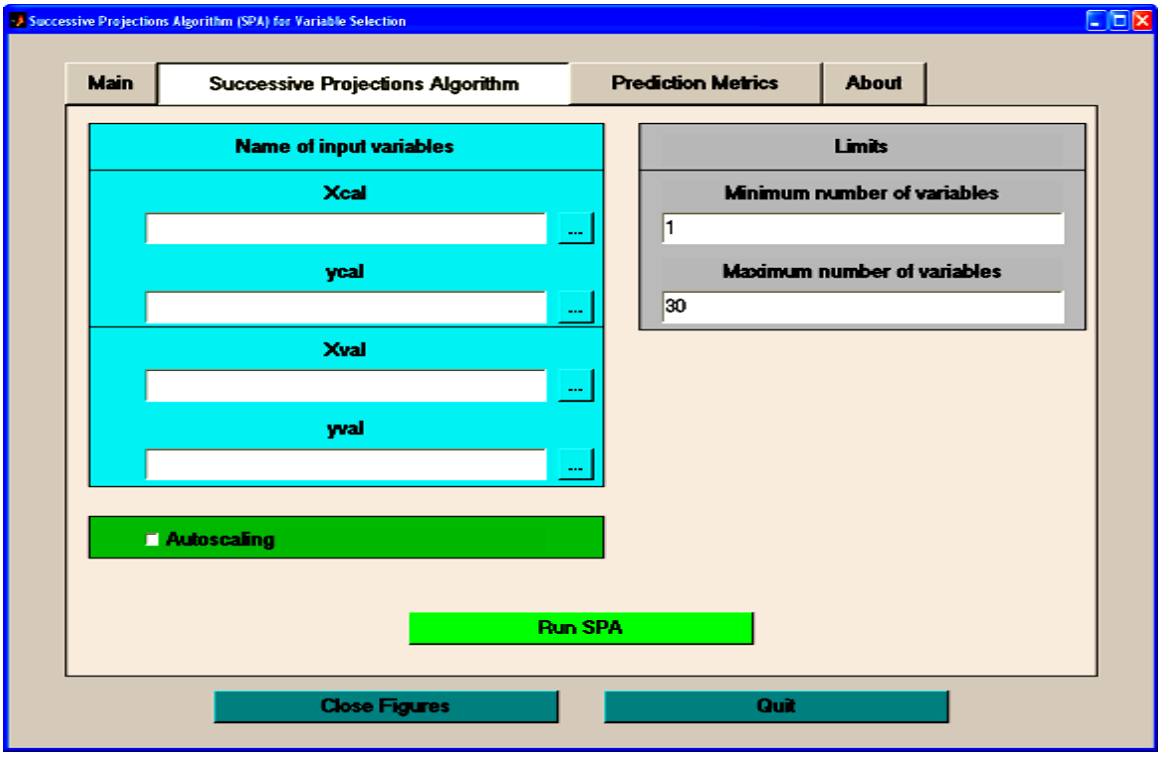

Figura 3.5. Tela do APS GUI para construção do modelo de calibração.

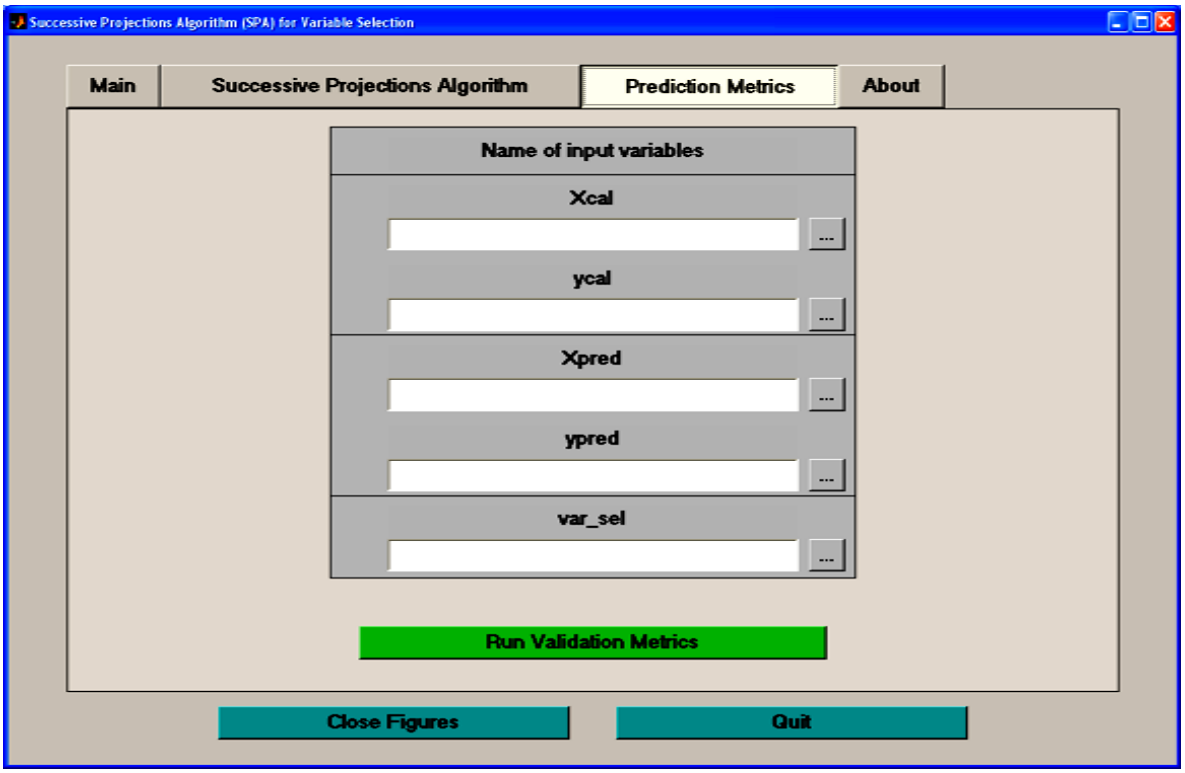

Figura 3.6. Tela do APS GUI para realizar a previsão.

3.8 Calibração PLS

Os modelos de calibração usando a técnica de regressão por mínimos quadrados parciais (PLS) foram construídos empregando o Unscrambler® 9.6

(CAMO COMO, Oslo, Noruega). O número de variáveis latentes para o PLS foi determinado baseado no erro de validação usando o default do software.

CAPÍTULO 4

RESULTADOS E DISCUSSÃO

4. RESULTADOS E DISCUSSÃO

Na determinação de todos os parâmetros, os modelos MLR baseados nas variáveis selecionadas pelo algoritmo ASA-VIF são identificados por MLR-ASA-VIF, cujo número entre parênteses indica o valor do limiar de corte adotado para o VIF. Quando o critério VIF não foi usado no processo de seleção, os modelos são identificados simplesmente como MLR-ASA.

Vale salientar que foram calculados os valores de VIF para as variáveis finais selecionadas pelo ASA (sem VIF), bem como pelas técnicas de seleção de variáveis usadas para fins de comparação (AG, APS e SW). Os valores assim obtidos são apresentados graficamente.

Os gráficos resultantes da utilização das ferramentas de diagnósticos (gráficos de resíduo, Scree plot, valores estimados versus valores de referência), descritos na **Seção 2.6.3**, não são mostrados para evitar excesso de informações. Não obstante, essas ferramentas foram usadas em todas as aplicações do algoritmo proposto e nenhum problema de modelagem ou amostra anômala foi observado.

4.1 Análise de misturas de corantes por espectrometria UV-VIS

Antes de apresentar os resultados dessa aplicação, discute-se a seguir as características de absorção dos quatro corantes e suas relações com a estrutura molecular.

A **Figura 4.1** mostra que os perfis espectrais dos quatro corantes se superpõem ao longo de toda a faixa de trabalho. Além disso, observa-se uma notável similaridade entre as estruturas e os espectros dos corantes vermelho 40 e amarelo crepúsculo. De fato, a diferença estrutural se deve à presença de dois grupos (CH_3 e OCH_3) ligados a um anel aromático na molécula do vermelho 40. O grupo OCH_3 é o provável responsável pelo deslocamento da banda da direita para maiores comprimentos de onda no espectro desse corante (**Figura 4.1**). Essas características podem contribuir para acentuar os problemas de correlação e multicolinearidade entre as variáveis nos espectros das misturas desses corantes.

Os problemas enfatizados acima realçam a exigência quanto ao desempenho requerido na aplicação dos algoritmos utilizados para a seleção das variáveis (comprimentos de onda) menos redundantes e mais informativas para os modelos de calibração MLR.

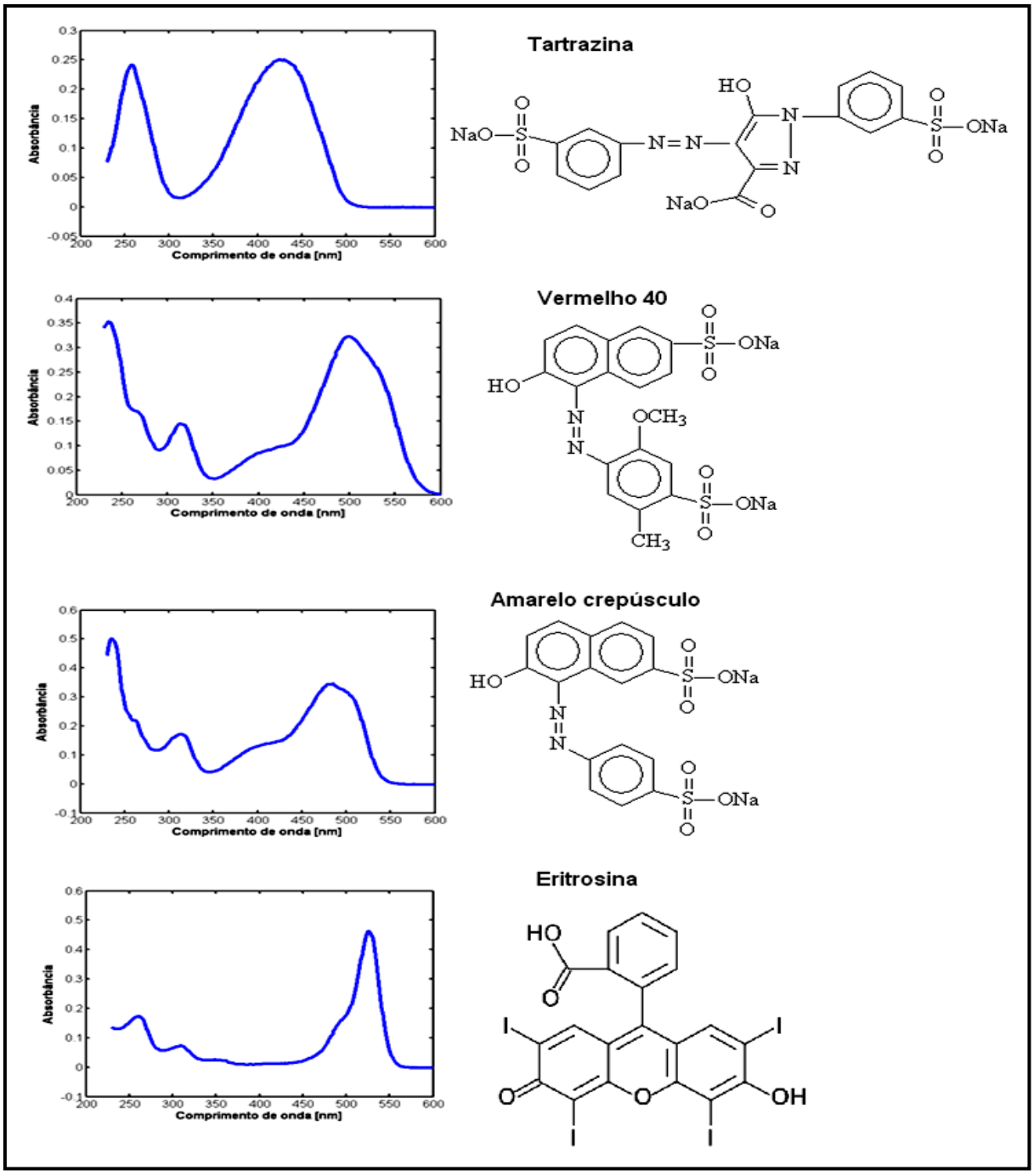

Figura 4.1. Espectros de absorção UV-VIS e estruturas moleculares dos corantes puros.

4.1.1 Determinação do corante tartrazina

A **Tabela 4.1** apresenta os resultados da aplicação dos modelos MLR (baseados nas variáveis selecionadas pelos algoritmos) e PLS (usando os espectros completos) usados na estimativa da concentração de tartrazina nas misturas. Os resultados são expressos em termos de $RMSEP_{val}$ e $RMSEP_{prev}$ obtidos, respectivamente, para as amostras dos conjuntos de validação e previsão. Para o modelo PLS, o número de variáveis selecionadas refere-se a “variáveis latentes”.

Tabela 4.1. Valores de RMSEP [mg L^{-1}] para o corante tartrazina.

Modelo	Nº de variáveis selecionadas	$RMSEP_{val}$	$RMSEP_{prev}$
MLR-ASA-VIF (5)	4	0,1	0,1
MLR-ASA-VIF (10)	4	0,1	0,1
MLR-ASA-VIF (30)	4	0,1	0,1
MLR-ASA-VIF (50)	4	0,1	0,1
MLR-ASA	7	0,1	0,1
MLR-APS	16	0,1	0,1
MLR-AG	23	0,1	0,1
MLR-SW	5	0,4	0,3
PLS	4	0,1	0,1

Com exceção do MLR-SW, todos os modelos apresentaram os menores valores de RMSEP (0,1 mg L^{-1}), os quais correspondem à precisão da concentração das soluções dos corantes usadas na preparação das misturas. Todavia, o algoritmo ASA-VIF produziu modelos MLR consideravelmente mais parcimoniosos (menor número de variáveis selecionadas) baseados em variáveis com menor multicolinearidade (**Figura 4.2**). O mesmo resultado foi obtido usando todos os limiares de cortes adotados para o VIF (5, 10, 30 e 50).

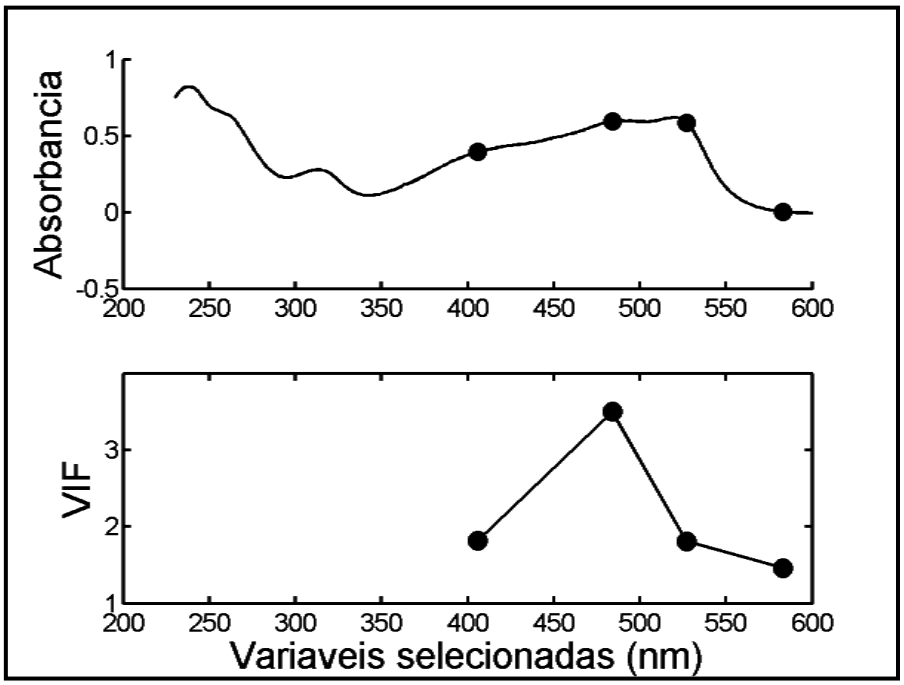

Figura 4.2. Variáveis selecionadas (e respectivos valores de VIF) pelo algoritmo ASA-VIF (com limiares de 5, 10, 30 e 50) para a tartrazina.

Um fato a ser ressaltado diz respeito ao número de variáveis selecionadas pelo algoritmo ASA-VIF. Como esse conjunto de dados resulta dos espectros das misturas dos quatro corantes, obtidas mediante um planejamento fatorial ortogonal, espera-se que apenas quatro variáveis sejam necessárias para modelá-los. Resultado similar também foi obtido pelo modelo PLS, porém empregando toda a faixa espectral.

A **Figura 4.3** mostra que o valor máximo do VIF para as variáveis selecionadas pelo ASA (sem o VIF) encontra-se em torno de 10000, sendo bem menor que os valores obtidos para o APS e AG (valores de VIF em torno de 10^4). Quanto ao número de variáveis, o ASA selecionou um número menor que os algoritmos AG e APS, porém maior que o do SW. As variáveis selecionadas pelo SW apresentaram valor máximo de VIF comparável ao das variáveis do ASA, mas produziram modelos MLR com menor capacidade preditiva (**Tabela 4.1**).

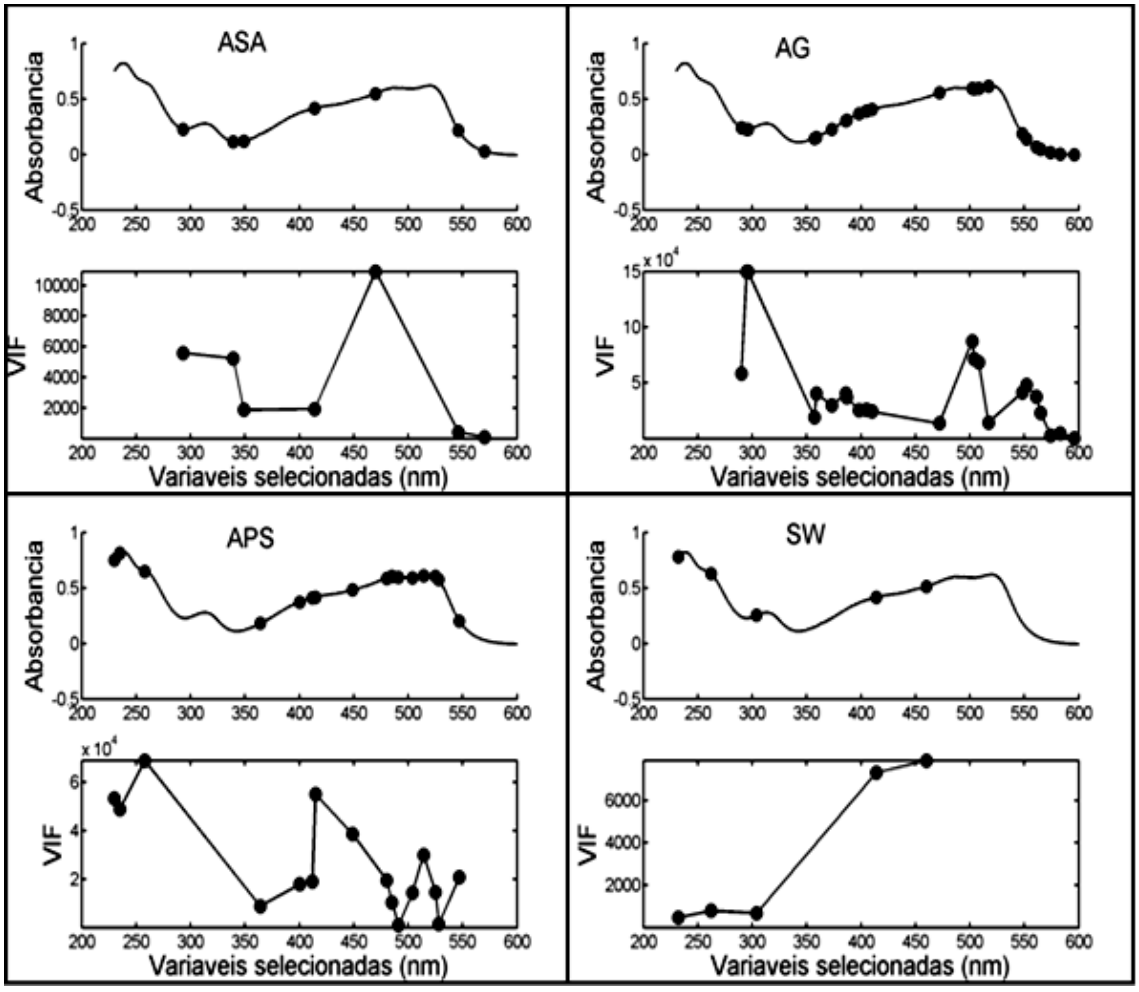

Figura 4.3. Variáveis selecionadas (e respectivos valores de VIF) pelos algoritmos ASA (sem o VIF), APS, AG e SW para o corante tartrazina.

4.1.2 Determinação do corante vermelho 40

A **Tabela 4.2** revela que os valores de RMSEP obtidos pelos modelos MLR-ASA-VIF para o vermelho 40 são geralmente similares aos dos outros modelos. Em relação às variáveis selecionadas, apesar de serem em mesmo número, não são exatamente as mesmas (**Figura 4.4**). Podemos observar também que, mesmo usando um limiar de corte mais ampliado (VIF < 50), as variáveis não apresentaram uma multicolinearidade acentuada como revelado pelos valores de VIF das variáveis selecionadas no final.

O resultado apresentado pelo PLS não é muito satisfatório, pois o modelo utiliza 6 variáveis latentes e produz um RMSEP ligeiramente maior que os outros para o conjunto das amostras de predição (**Tabela 4.2**). Provavelmente, isso se deve à forte sobreposição espectral decorrente da semelhança do perfil espectral entre o corante vermelho 40 e o amarelo crepúsculo.

Tabela 4.2. Valores de RMSEP [mgL^{-1}] para o corante vermelho 40.

Modelo	Nº de variáveis selecionadas	$RMSEP_{val}$	$RMSEP_{prev}$
MLR-ASA-VIF (5)	4	0,1	0,1
MLR-ASA-VIF (10)	4	0,1	0,1
MLR-ASA-VIF (30)	4	0,1	0,1
MLR-ASA-VIF (50)	4	0,1	0,1
MLR-ASA	9	0,1	0,1
MLR-APS	7	0,1	0,1
MLR-AG	20	0,03	0,1
MLR-SW	3	0,1	0,2
PLS	6	0,1	0,2

A **Figura 4.5** mostra que o algoritmo ASA selecionou um número de variáveis maior que os algoritmos APS e SW, porém as variáveis apresentam valores de VIF comparáveis aos do SW e significativamente menores que os das variáveis as selecionadas pelo APS. Em comparação com o AG, o algoritmo ASA selecionou um número muito menor de variáveis com valores de VIF consideravelmente menores.

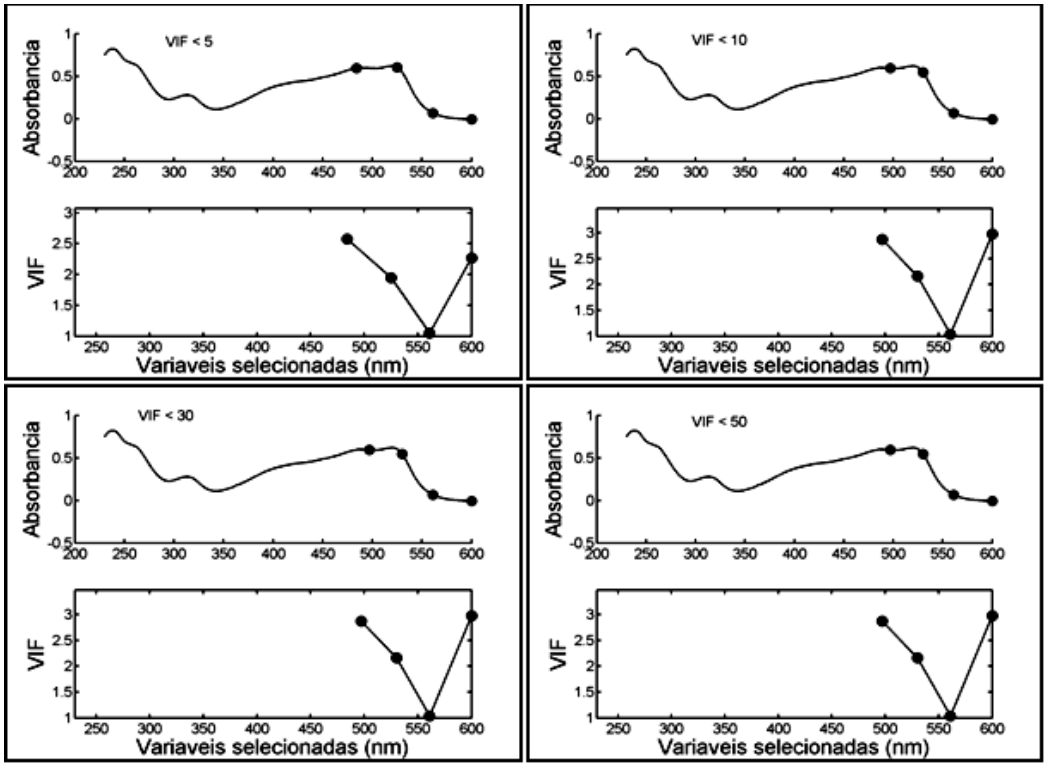

Figura 4.4. Variáveis selecionadas (e respectivos valores de VIF) pelo algoritmo ASA-VIF (com limiares de 5, 10, 30 e 50) para o corante vermelho 40.

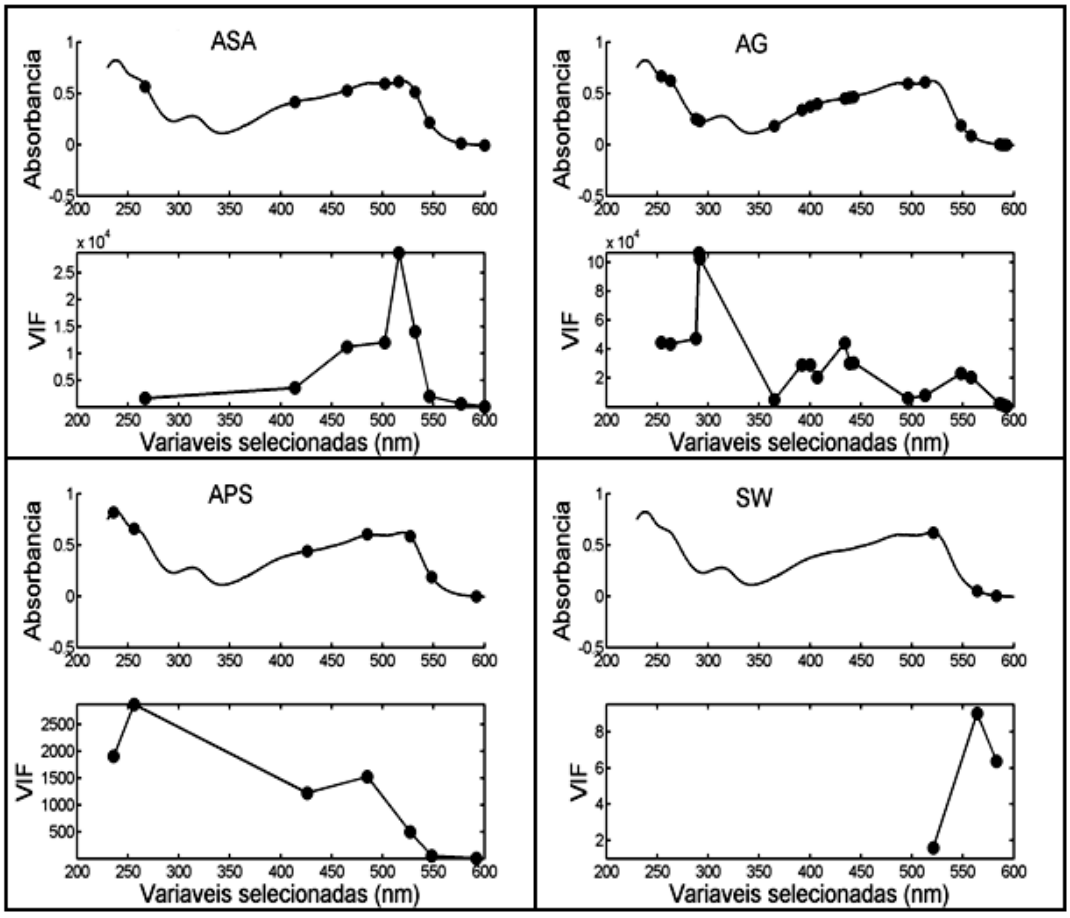

Figura 4.5. Variáveis selecionadas (e respectivos valores de VIF) pelos algoritmos ASA (sem o VIF), APS, AG e SW para o corante vermelho 40.

4.1.3 Determinação do corante amarelo crepúsculo

Como ressaltado antes, este corante apresenta um perfil espectral muito parecido com o do corante vermelho 40 (**Figura 4.1**), o que pode justificar os resultados dos modelos MLR-ASA-VIF exibidos na **Tabela 4.3**. De fato, observa-se que o algoritmo ASA-VIF selecionou 5 variáveis para todos os limites de VIF usados. Neste caso, selecionou-se uma variável a mais que para os outros corantes, produzindo modelos com RMSEP de validação maiores (para VIF menores que 5 e 10) e RMSEP de previsão maiores (para os valores de VIF menores que 30 e 50).

A **Tabela 4.3** revela também que os valores de RMSEP do modelo MLR-ASA são satisfatórios, embora apresentem da mesma ordem de grandeza que valores de produzidos pelo AG (**Figura 4.7**). Os resultados do modelo PLS parecem corroborar com o argumento de que a sobreposição espectral dificulta a calibração, principalmente para os métodos que usam todo o espectro. De fato, a qualidade da predição do modelo PLS foi muito mais afetada que a dos modelos MLR.

Tabela 4.3. Valores de RMSEP [mgL^{-1}] para o corante amarelo crepúsculo.

Modelo	Nº de variáveis selecionadas	$RMSEP_{val}$	$RMSEP_{prev}$
MLR-ASA-VIF (5)	5	0,2	0,1
MLR-ASA-VIF (10)	5	0,2	0,1
MLR-ASA-VIF (30)	5	0,1	0,2
MLR-ASA-VIF (50)	5	0,1	0,2
MLR-ASA	9	0,1	0,1
MLR-APS	5	0,1	0,1
MLR-AG	19	0,1	0,1
MLR-SW	5	0,2	0,1
PLS	6	0,4	0,5

A **Figura 4.6** mostra que as variáveis selecionadas pelo ASA-VIF não são as mesmas para os diferentes limiares adotados para o VIF. Além do mais, elas apresentam valores de VIF menores que 10, indicando uma pequena ou nenhuma multicolinearidade. Por outro lado, a **Figura 4.7** revela que os algoritmos APS e SW selecionaram apenas 5 variáveis enquanto o ASA selecionou 9 e o AG 19. No entanto, as variáveis selecionadas pelos algoritmos APS e SW apresentam ainda multicolinearidade bastante acima do limite máximo proposto na literatura (GIACOMELLI et al., 1998), que é de valores de VIF menores que 10.

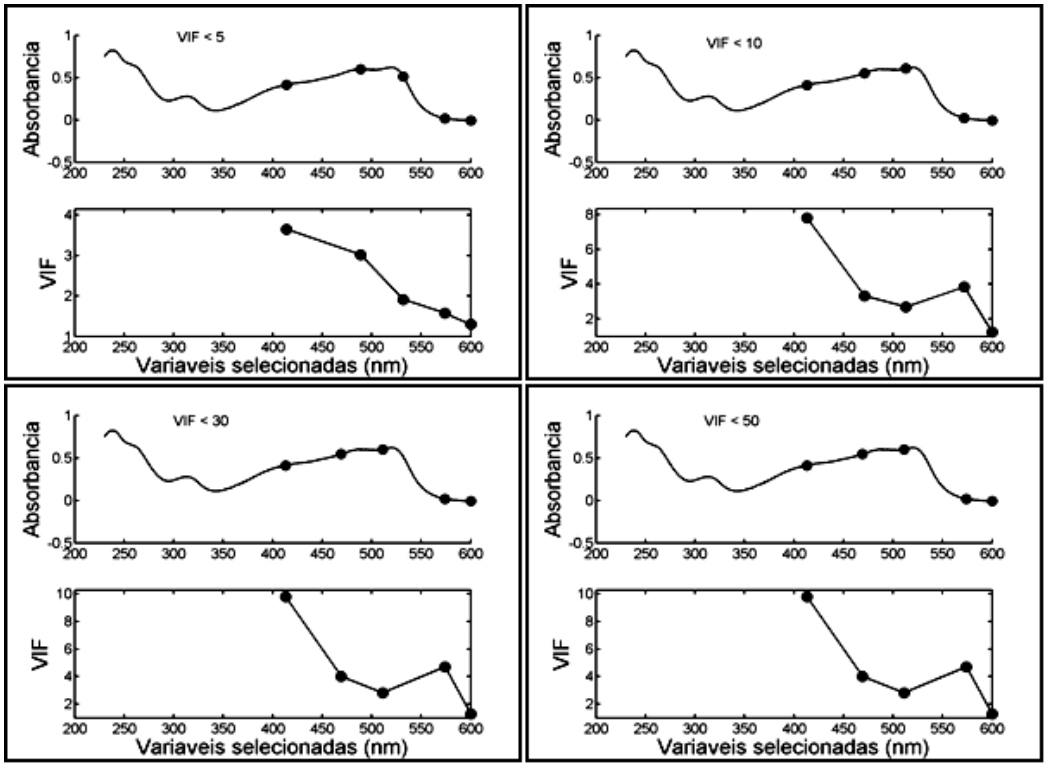

Figura 4.6. Variáveis selecionadas (e respectivos valores de VIF) pelo algoritmo ASA-VIF (com limiares de 5, 10, 30 e 50) para o corante amarelo crepúsculo.

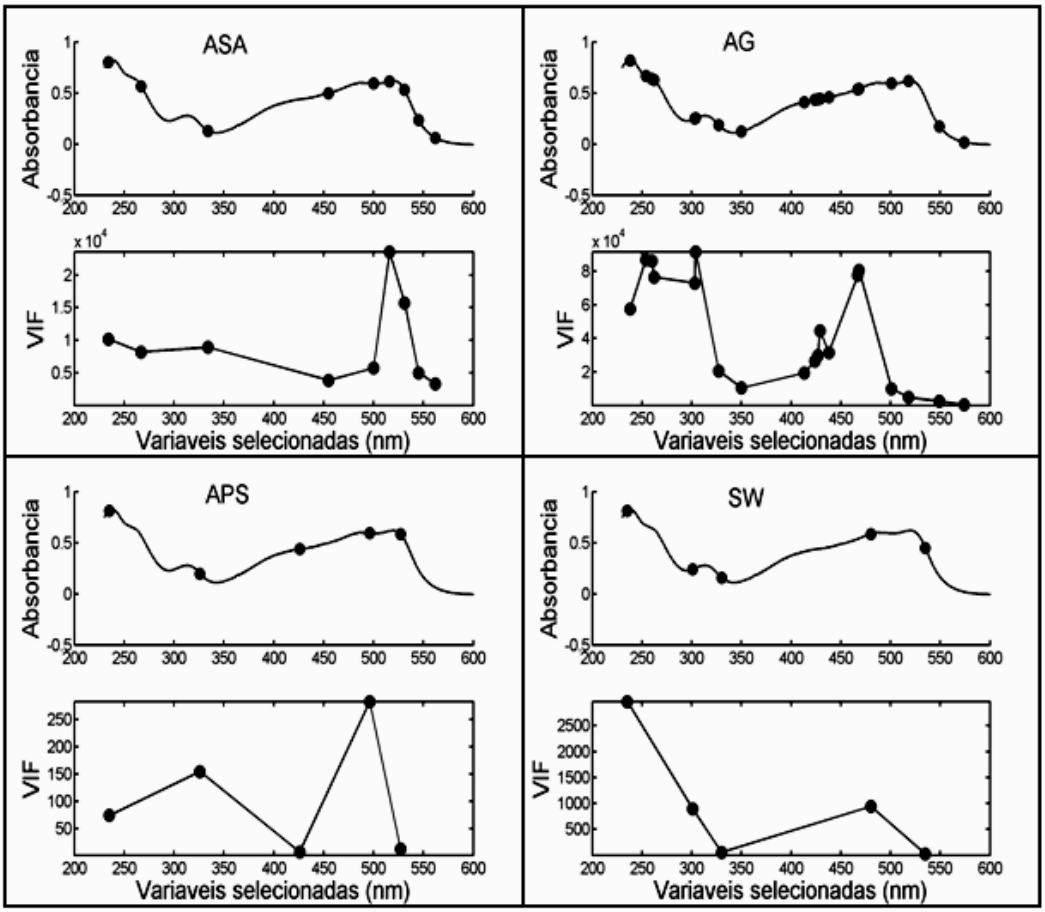

Figura 4.7. Variáveis selecionadas (e respectivos valores de VIF) pelos algoritmos ASA (sem o VIF), APS, AG e SW para o corante amarelo crepúsculo.

4.1.4 Determinação do corante eritrosina

Os algoritmos ASA-VIF e APS apresentaram desempenho similar, pois produziram modelos MLR com os mesmos valores de RMSEP (**Tabela 4.4**) e selecionaram variáveis com valores de VIF menores que 4 (**Figuras 4.8** e **4.9**). Isso indica que, mesmo diante de uma considerável sobreposição espectral (porém menos acentuada que no caso do amarelo crepúsculo), conseguiu-se efetuar uma calibração para os modelos menos afetada por problemas de multicolinearidade.

A **Tabela 4.4** revela também que o desempenho do modelo PLS foi similar da determinação dos corantes vermelho 40 e amarelo crepúsculo, ou seja, possui um número relativamente alto de variáveis latentes e baixa capacidade de predição.

Tabela 4.4. Valores de RMSEP [mgL^{-1}] para o corante eritrosina.

Modelo	Nº de variáveis selecionadas	$RMSEP_{val}$	$RMSEP_{prev}$
MLR-ASA-VIF (5)	4	0,1	0,1
MLR-ASA-VIF (10)	4	0,1	0,1
MLR-ASA-VIF (30)	4	0,1	0,1
MLR-ASA-VIF (50	4	0,1	0,1
MLR-ASA	15	0,1	0,1
MLR-APS	4	0,1	0,1
MLR-AG	19	0,04	0,1
MLR-SW	7	0,2	0,1
PLS	7	0,4	0,4

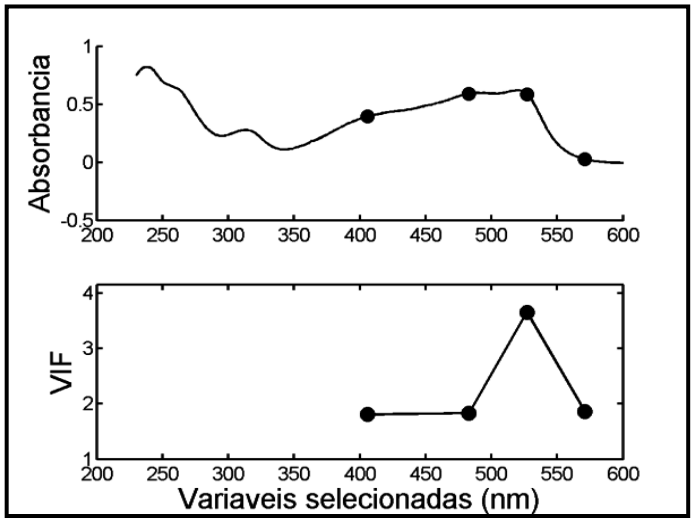

Figura 4.8. Variáveis selecionadas (e respectivos valores de VIF) pelo algoritmo ASA-VIF (com limiares de 5, 10, 30 e 50) para o corante eritrosina.

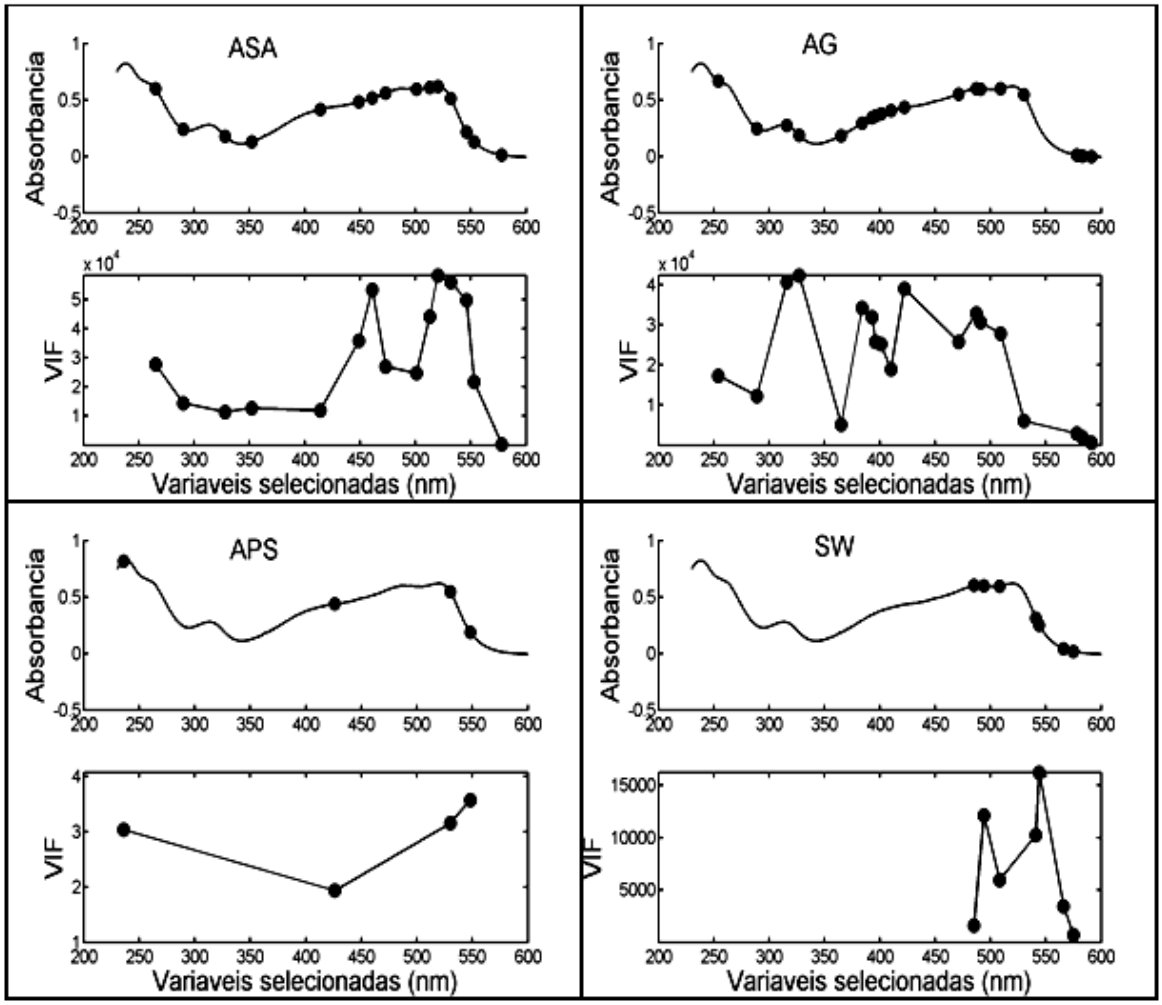

Figura 4.9. Variáveis selecionadas (e respectivos valores de VIF) pelos algoritmos ASA (sem o VIF), APS, AG e SW para o corante eritrosina.

4.2 Análise de amostras de trigo por espectrometria NIR

4.2.1 Determinação de proteína no trigo

A **Tabela 4.5** revela que na determinação do teor de proteína no trigo o uso do critério VIF não afetou a escolha das variáveis selecionadas pelo algoritmo ASA. De fato, as **Figuras 4.10** e **4.11** mostram que as variáveis selecionadas pelos algoritmos ASA e ASA-VIF foram exatamente as mesmas, independente do limiar de corte utilizado para o VIF. Estes resultados sugerem uma baixa multicolinearidade entre as variáveis espectrais, porém não descarta a possibilidade das correlações par a par serem expressivas.

Como pode ser visto na **Tabela 4.5**, os modelos MLR-ASA-VIF apresentaram valores de RMSEP comparáveis aos obtidos com os outros modelos de seleção de variáveis. Por outro lado, o algoritmo SW produziu um RMSEP de previsão

relativamente maior, porém selecionou apenas duas variáveis. Os valores de RMSEP para o modelo PLS são ligeiramente menores que os obtidos pelos métodos de seleção de variáveis, mas o número de variáveis latentes utilizadas é bastante elevado.

Tabela 4.5. Valores de RMSEP [%] para proteína no trigo.

Modelo	Nº de variáveis selecionadas	$RMSEP_{val}$	$RMSEP_{prev}$
MLR-ASA-VIF (5)	5	0,24	0,35
MLR-ASA-VIF (10)	5	0,24	0,35
MLR-ASA-VIF (30)	5	0,24	0,35
MLR-ASA-VIF (50)	5	0,24	0,35
MLR-ASA	5	0,24	0,35
MLR-APS	4	0,21	0,32
MLR-AG	18	0,14	0,32
MLR-SW	2	0,35	0,44
PLS	10	0,24	0,28

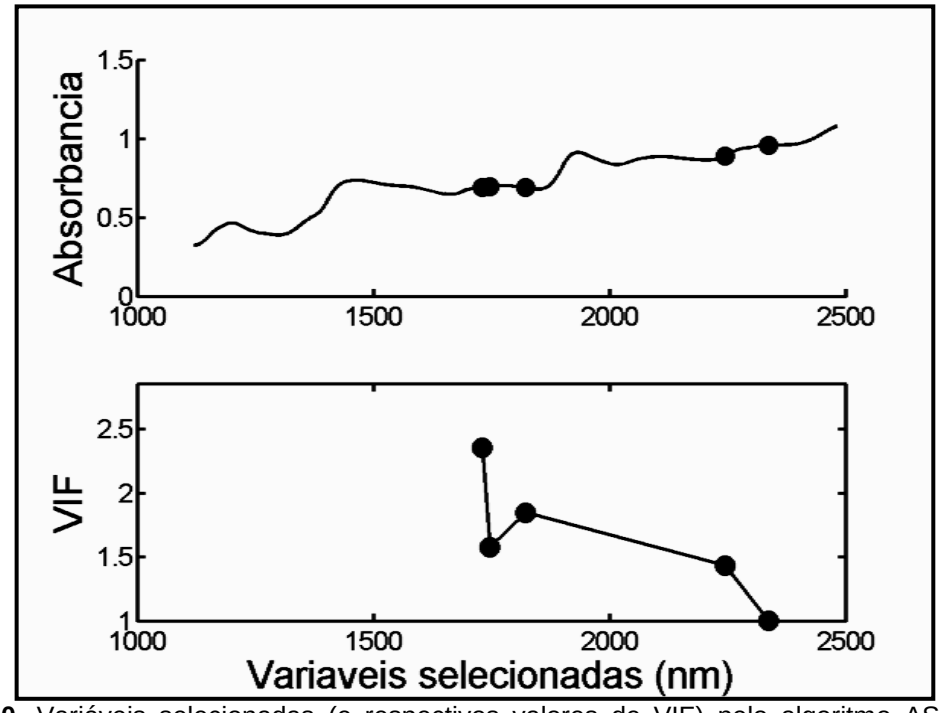

Figura 4.10. Variáveis selecionadas (e respectivos valores de VIF) pelo algoritmo ASA-VIF (com limiares de 5, 10, 30, 50) para a proteína do trigo.

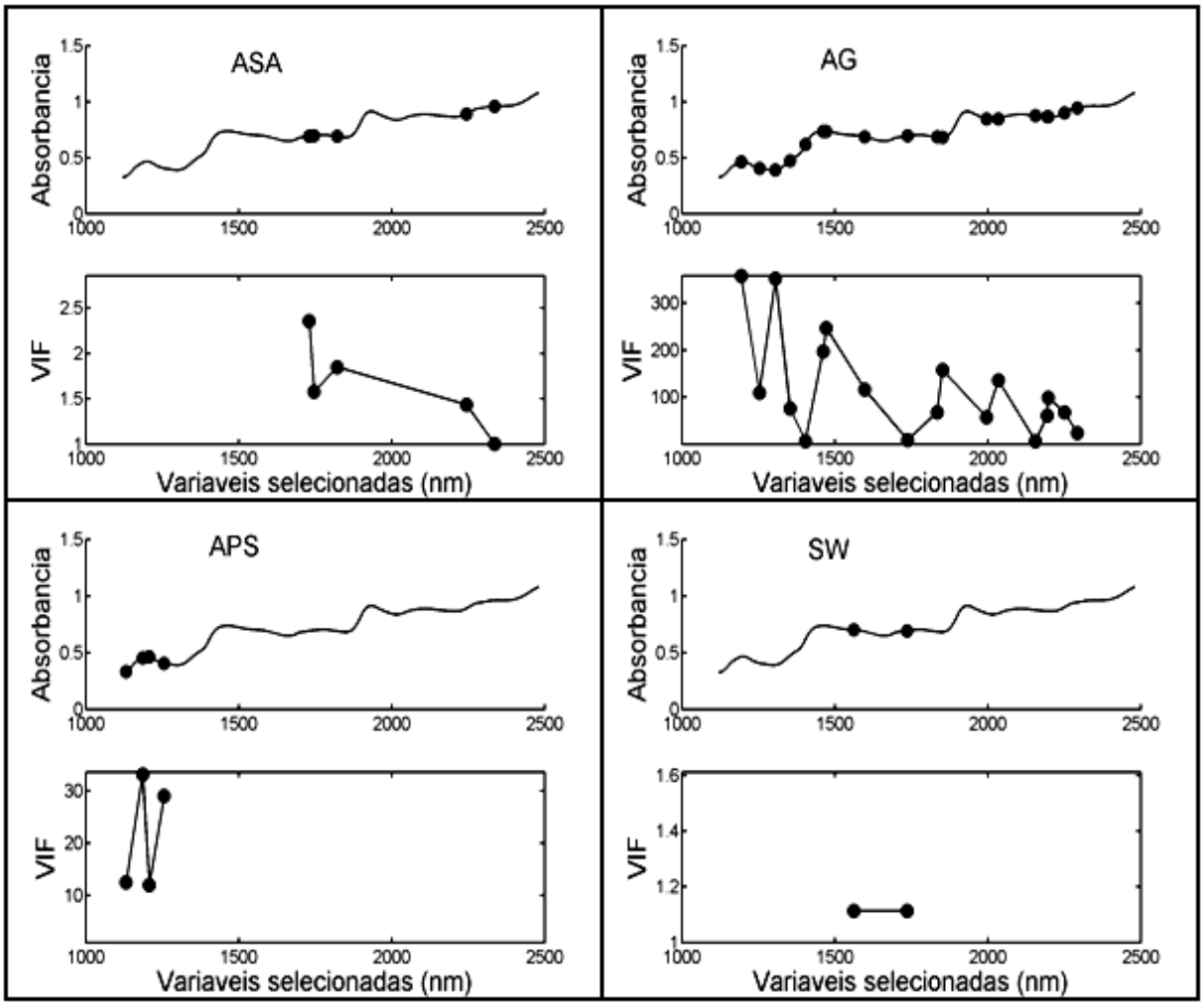

Figura 4.11. Variáveis selecionadas (e respectivos valores de VIF) pelos algoritmos ASA (sem o VIF), APS, AG e SW para a proteína do trigo.

4.2.2 Determinação de umidade no trigo

A **Tabela 4.6** apresenta os resultados da aplicação dos modelos de calibração à determinação da umidade no trigo. Novamente, pode-se constatar que o uso do critério VIF não afetou a escolha das variáveis selecionadas pelo algoritmo ASA. O número de variáveis selecionadas é consideravelmente menor que o dos outros algoritmos. Em termos de RMSEP, pode-se notar que os valores fornecidos por todos os modelos são praticamente iguais entre si.

A **Figura 4.12** mostra que os valores de VIF das variáveis selecionadas pelo algoritmo ASA-VIF foram exatamente os mesmos para todos os limiares de corte adotados, tal como ocorreu no caso da determinação da proteína no trigo. Estes resultados revelam uma maior robustez do método ASA-VIF, o que não ocorreu com as outras técnicas (**Figura 4.13**). De fato, estas técnicas produziram resultados

diferentes daqueles relativos à proteína, tanto em termos do número de variáveis selecionadas como no grau de multicolinearidade (neste caso, mais elevado).

Tabela 4.6. Valores de RMSEP [%] para umidade no trigo.

Modelo	Nº de variáveis selecionadas	$RMSEP_{val}$	$RMSEP_{prev}$
MLR-ASA-VIF (5)	4	0,20	0,27
MLR-ASA-VIF (10)	4	0,20	0,27
MLR-ASA-VIF (30)	4	0,20	0,27
MLR-ASA-VIF (50)	4	0,20	0,27
MLR-ASA	4	0,20	0,27
MLR-APS	14	0,21	0,28
MLR-AG	22	0,10	0,28
MLR-SW	13	0,26	0,27
PLS	4	0,27	0,27

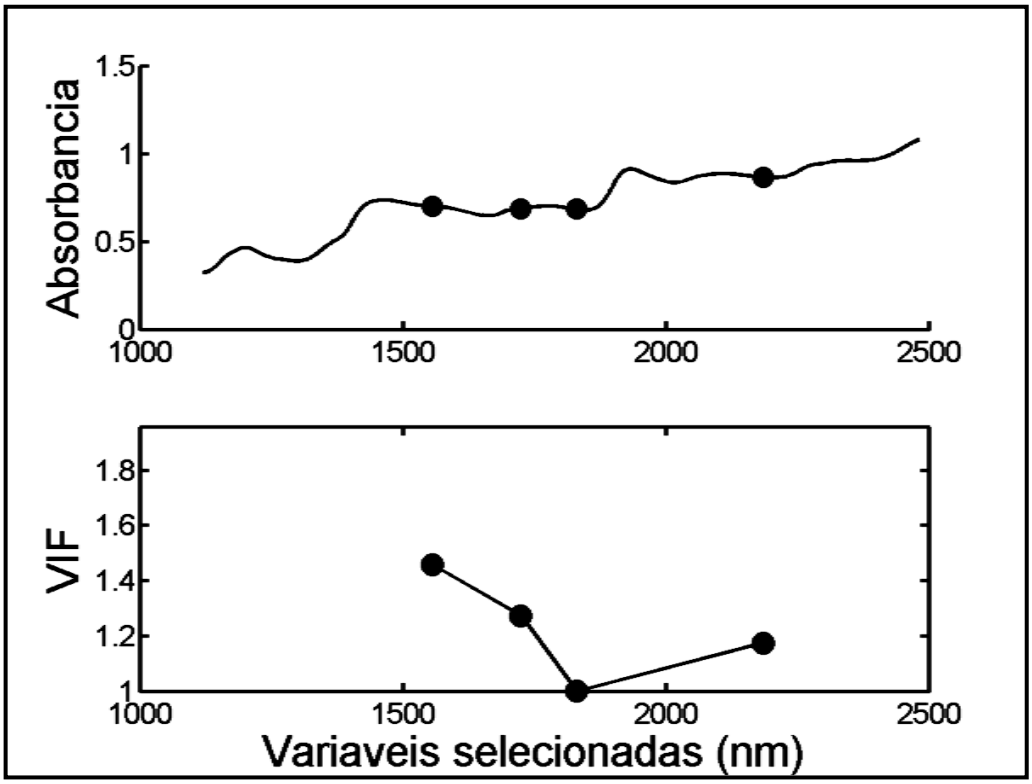

Figura 4.12. Variáveis selecionadas (e respectivos valores de VIF) pelo algoritmo ASA-VIF (com limiares de 5, 10, 30, 50) para a umidade no trigo.

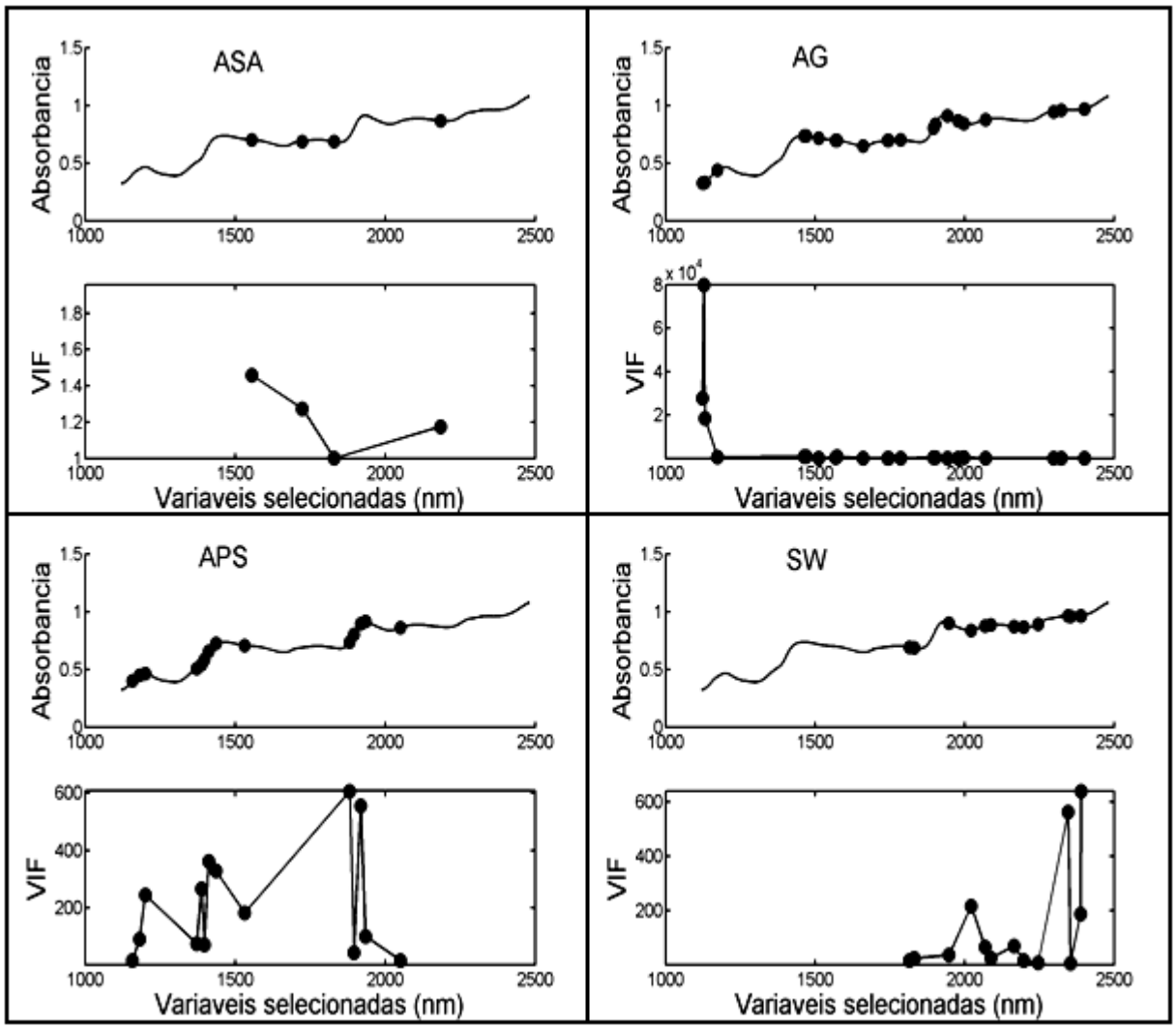

Figura 4.13. Variáveis selecionadas (e respectivos valores de VIF) pelos algoritmos ASA (sem o VIF), APS, AG e SW para a umidade no trigo.

4.3 Análise de amostras de milho por espectrometria NIR

4.3.1 Determinação de proteína no milho

A **Tabela 4.7** mostra que, ao contrário dos parâmetros anteriores, o comportamento do algoritmo ASA-VIF foi diferente (em termos de número de variáveis e valores de RMSEP) para os diversos limites de VIF adotados. Para os limites VIF < 5 e VIF < 10 (**Figura 4.14**), as variáveis selecionadas foram iguais. Estas variáveis encontram-se bem distribuídas, não apresentando sobreposições, porém produziram modelos com RMSEP bem maiores que os outros métodos (**Tabela 4.7**). Quando usamos o limite VIF < 30, o número de variáveis diminuiu e, principalmente, os valores de RMSEP. No entanto, algumas das variáveis apresentaram valores próximos (**Figura 4.14**). Para o limite VIF < 50, as variáveis

selecionadas também não são multicolineares, mas apresentam o mesmo problema de proximidade entre variáveis. Em compensação, os valores de RMSEP foram reduzidos ainda mais.

É importante notar ainda na **Tabela 4.7** que os RMSEP de previsão apresentam menores valores quando se utiliza um maior número de variáveis, mesmo sendo fortemente multicolineares. O compromisso que se deve assumir é que as variáveis sejam menos correlacionadas mesmo que se tenha um RMSEP um pouco maior, porém dentro de um erro permitido. Nesse sentido, pode-se constatar que mesmo no caso dos maiores valores de RMSEP obtidos pelos modelos MLR-ASA-VIF, o erro relativo é somente cerca de 2%. Sendo assim, o método proposto pode ser considerado adequado para a determinação deste parâmetro.

Tabela 4.7. Valores de RMSEP [%] para proteína no milho.

Modelo	Nº de variáveis selecionadas	$RMSEP_{val}$	$RMSEP_{prev}$
MLR-ASA-VIF (5)	7	0,103	0,177
MLR-ASA-VIF (10)	7	0,103	0,177
MLR-ASA-VIF (30)	5	0,069	0,141
MLR-ASA-VIF (50)	6	0,058	0,122
MLR-ASA	15	0,048	0,095
MLR-APS	13	0,023	0,030
MLR-AG	25	0,010	0,024
MLR-SW	18	0,010	0,0129
PLS	7	0,063	0,081

A **Figura 4.15** mostra que os modelos MLR-ASA, MLR-AG e MLR-SW, apesar de terem produzido baixos valores de RMSEP, utilizam um número elevado de variáveis com forte multicolinearidade, principalmente o MLR-AG e o MLR-SW. O modelo PLS também utiliza um número considerável de variáveis latentes (ver **Tabela 4.7**).

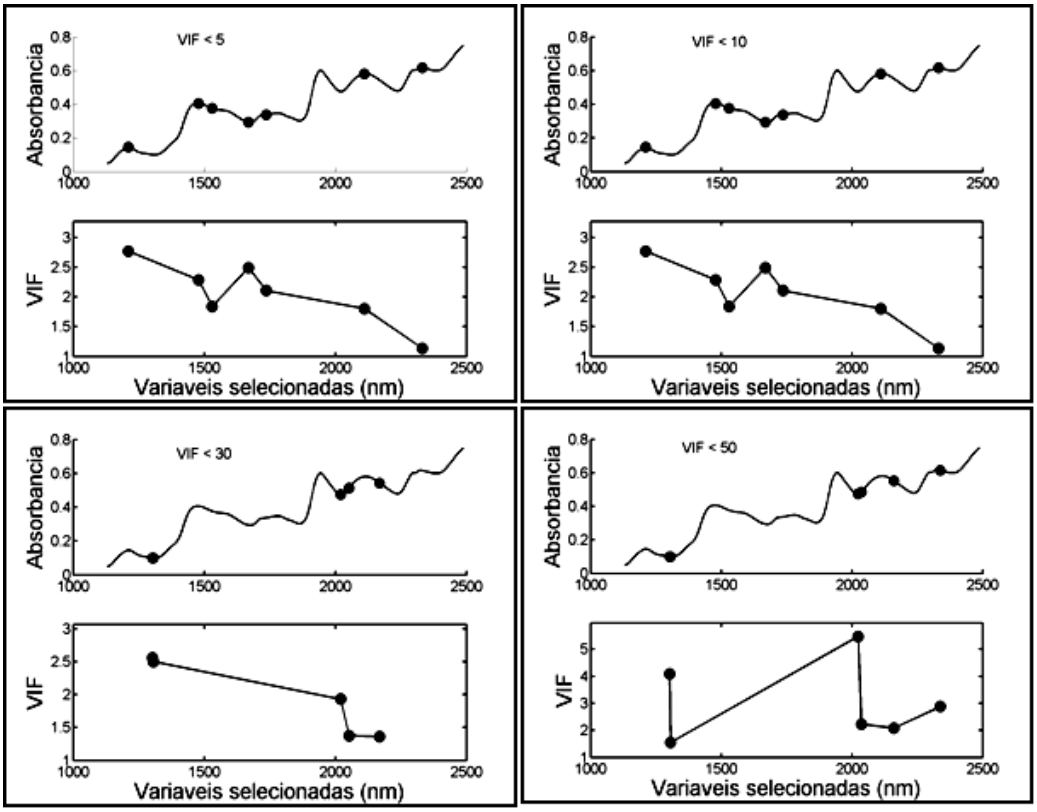

Figura 4.14. Variáveis selecionadas (e respectivos valores de VIF) pelo algoritmo ASA-VIF (com limiares de 5, 10, 30, 50) para a proteína no milho.

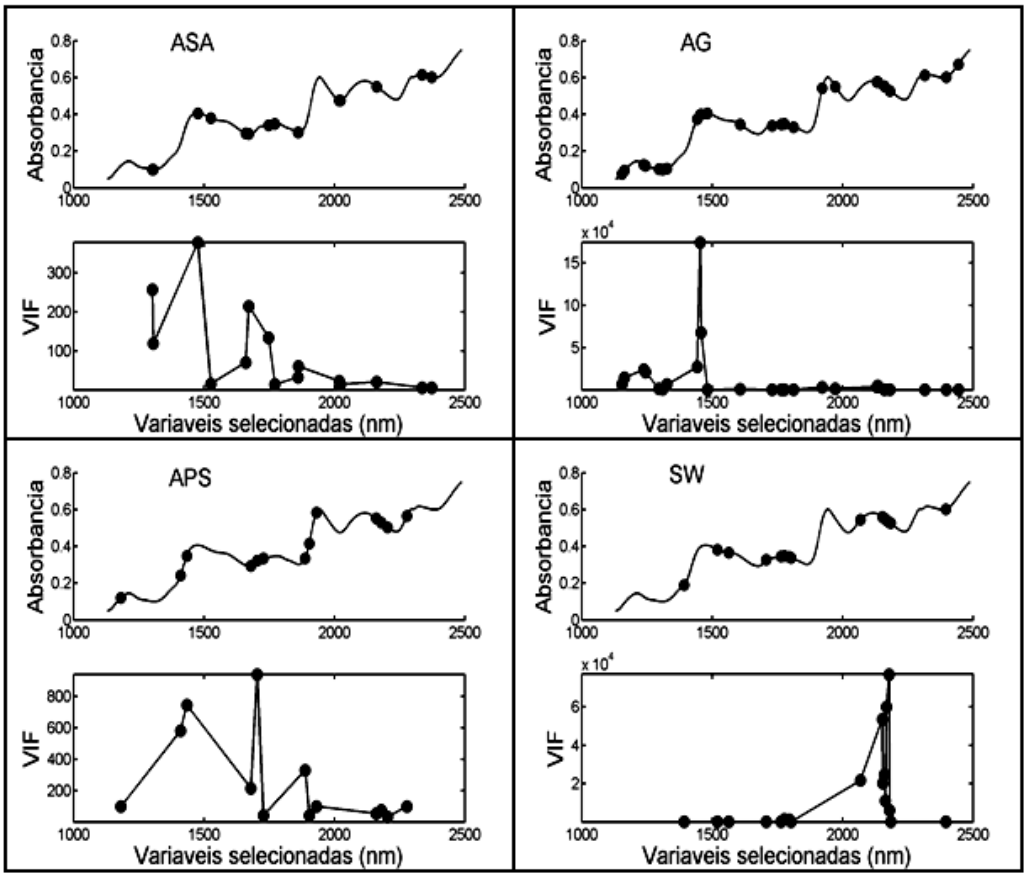

Figura 4.15. Variáveis selecionadas (e respectivos valores de VIF) pelos algoritmos ASA (sem o VIF), APS, AG e SW para a proteína no milho.

4.3.2 Determinação de umidade no milho

A **Tabela 4.8** revela que, nessa determinação, os diferentes limites de VIF usados pelo algoritmo proposto também produziram resultados diferentes, principalmente em termos de RMSEP. De fato, houve uma redução significativa dos valores de RMSEP de previsão com o aumento dos valores de VIF adotados.

Tabela 4.8. Valores de RMSEP [%] para umidade no milho.

Modelo	Nº de variáveis selecionadas	$RMSEP_{val}$	$RMSEP_{prev}$
MLR-ASA-VIF (5)	6	0,042	0,063
MLR-ASA-VIF (10)	7	0,025	0,048
MLR-ASA-VIF (30)	7	0,025	0,048
MLR-ASA-VIF (50)	6	0,025	0,028
MLR-ASA	14	0,017	0,022
MLR-APS	15	0,014	0,019
MLR-AG	28	0,009	0,018
MLR-SW	9	0,021	0,026
PLS	10	0,025	0,028

A **Figura 4.16** indica que mesmo utilizando o maior limite adotado para o VIF, as variáveis selecionadas não possuem multicolinearidade significativa. Além disso, as variáveis produziram modelos mais parcimoniosos com capacidade preditiva comparável a dos outros modelos. Dessa forma, o modelo MLR-ASA-VIF (VIF < 50) parece ser mais indicado para a determinação deste parâmetro.

Por outro lado, a **Figura 4.17** demonstra que as variáveis selecionadas pelo APS e AG apresentam forte multicolinearidade, bem como são muito numerosas (**Tabela 4.8**). O algoritmo SW selecionou um número considerável de variáveis, porém apresentam baixa multicolinearidade.

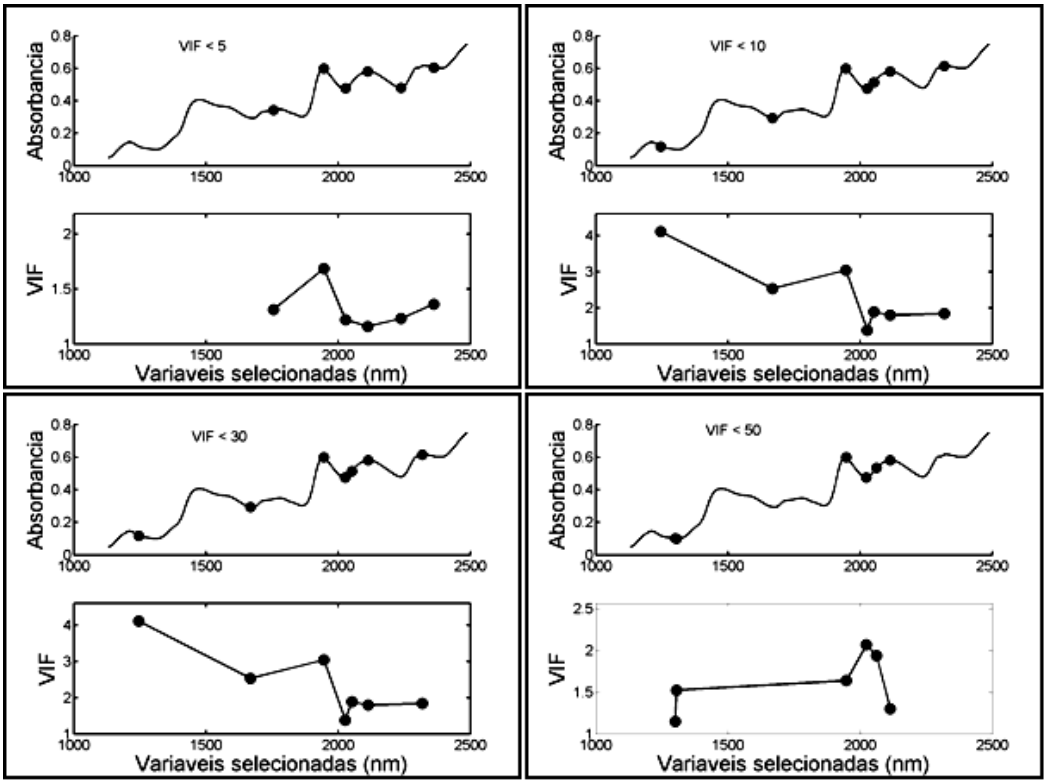

Figura 4.16. Variáveis selecionadas (e respectivos valores de VIF) pelo algoritmo ASA-VIF (com limiares de 5, 10, 30, 50) para a umidade no milho.

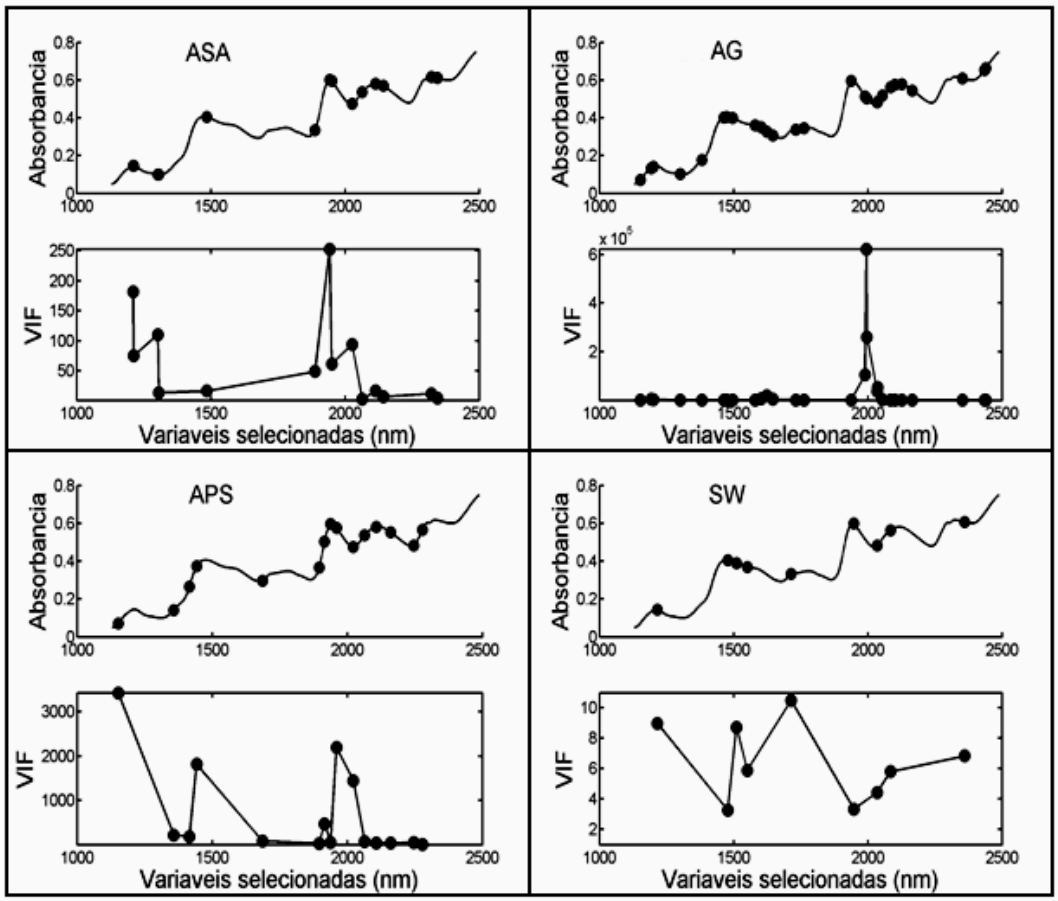

Figura 4.17. Variáveis selecionadas (e respectivos valores de VIF) pelos algoritmos ASA (sem o VIF), APS, AG e SW para a umidade no milho.

4.3.3 Determinação de óleo no milho

Como se pode observar na **Tabela 4.9**, os modelos MLR-ASA-VIF apresentam a mesma tendência de diminuição do RMSEP à medida que os valores adotados para o VIF aumentam. Todavia, o número de variáveis selecionadas pelo ASA-VIF tendeu a aumentar ainda mais neste caso. Além disso, ao adotar o limite de corte menos exigente (VIF < 50), os valores de VIF para as variáveis selecionadas ultrapassam o limite VIF = 10. Este limiar é recomendado na literatura para selecionar, na prática, as variáveis que não apresentam uma multicolinearidade significativa (**Figura 4.18**).

Os resultados sugerem que o modelo o MLR-ASA-VIF (30) é mais indicado para a determinação de óleo no milho, pois utiliza o número razoável de variáveis minimamente correlacionadas, bem como valores de RMSEP compatível com os outros modelos.

Tabela 4.9. Valores de RMSEP [%] para o óleo no milho

Modelo	Nº de variáveis selecionadas	$RMSEP_{val}$	$RMSEP_{prev}$
MLR-ASA-VIF (5)	3	0,050	0,054
MLR-ASA-VIF (10)	3	0,050	0,054
MLR-ASA-VIF (30)	5	0,032	0,035
MLR-ASA-VIF (50)	10	0,028	0,023
MLR-ASA	11	0,025	0,026
MLR-APS	18	0,025	0,030
MLR-AG	20	0,013	0,019
MLR-SW	13	0,052	0,039
PLS	6	0,052	0,045

Quanto à aplicação dos algoritmos APS, AG e SW, a **Figura 4.19** revela um menor desempenho, principalmente devido ao maior número de variáveis com uma multicolinearidade acentuada. De fato, os valores máximos de VIF ultrapassaram sempre o valor 1000.

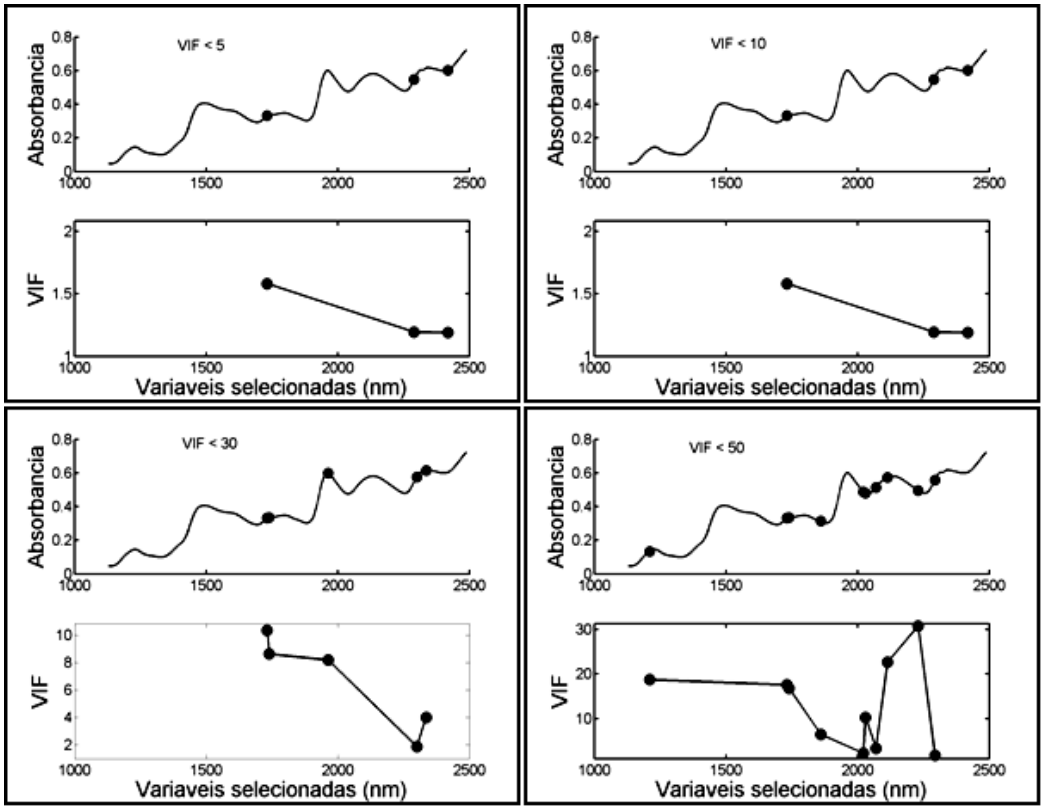

Figura 4.18. Variáveis selecionadas (e respectivos valores de VIF) pelo algoritmo ASA-VIF (com limiares de 5, 10, 30, 50) para o óleo no milho.

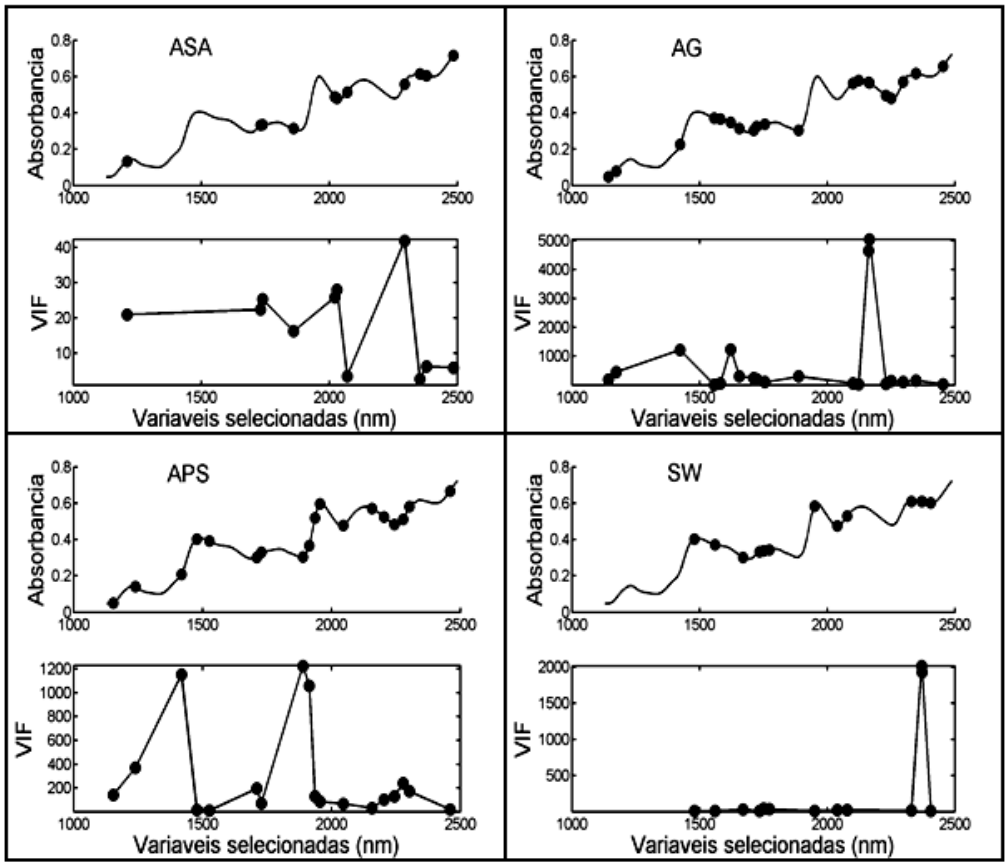

Figura 4.19. Variáveis selecionadas (e respectivos valores de VIF) pelos algoritmos ASA (sem o VIF), APS, AG e SW para o óleo no milho.

4.3.4 Determinação de amido no milho

A **Tabela 4.10** mostra que os quatro modelos MLR-ASA-VIF, construídos para estimar o conteúdo de amido no milho, apresentam um comportamento divergente daquele alcançado para os outros parâmetros. O valor do RMSEP (tanto de previsão como de validação) para o limiar (VIF < 5) é o menor, depois aumenta quando o VIF < 10, e volta a diminuir com o aumento do limiar adotado. As variáveis selecionadas com o limite inferior permanecem sendo escolhidas para os outros limites (**Figura 4.20**), porém a adição de novas variáveis não melhora a capacidade preditiva dos modelos. Desta forma, o modelo MLR-ASA-VIF (5) parece o mais indicado.

Tabela 4.10. Valores de RMSEP [%] para o amido no milho

Modelo	Nº de variáveis selecionadas	$RMSEP_{val}$	$RMSEP_{prev}$
MLR-ASA-VIF (5)	3	0,126	0,156
MLR-ASA-VIF (10)	7	0,168	0,194
MLR-ASA-VIF (30)	7	0,153	0,180
MLR-ASA-VIF (50)	7	0,118	0,162
MLR-ASA	7	0,118	0,162
MLR-APS	13	0,119	0,177
MLR-AG	25	0,037	0,074
MLR-SW	16	0,123	0,129
PLS	5	0,277	0,228

A **Figura 4.21** demonstra que os modelos MLR-APS, MLR-AG e MLR-SW utilizam um número de variáveis elevado com forte multicolinearidade. Não obstante, a previsão nem sempre é tão boa, exceto para o MLR-AG (**Tabela 4.10**). O modelo PLS utilizou poucas variáveis latentes, mas a capacidade preditiva ficou aquém dos outros métodos.

A correlação e multicolinearidade entre as variáveis deste conjunto de dados são bastante expressivas. Isto fica patente pelos valores do VIF das variáveis selecionadas pelos algoritmos APS, AG e SW. Os valores de RMSEP obtidos pelos modelos MLR-ASA-VIF foram algumas vezes maiores que os dos outros modelos. Contudo, um menor número de variáveis selecionadas com pouca multicolinearidade foi sempre obtido.

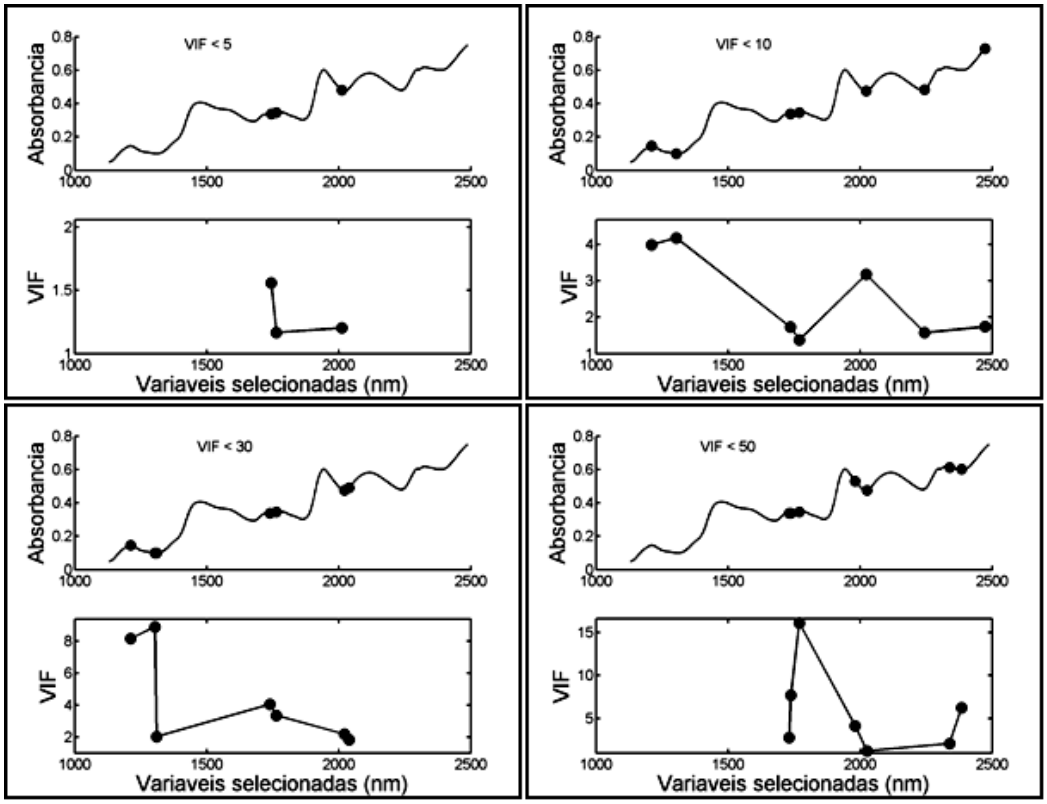

Figura 4.20. Variáveis selecionadas (e respectivos valores de VIF) pelo algoritmo ASA-VIF (com limiares de 5, 10, 30, 50) para o amido no milho.

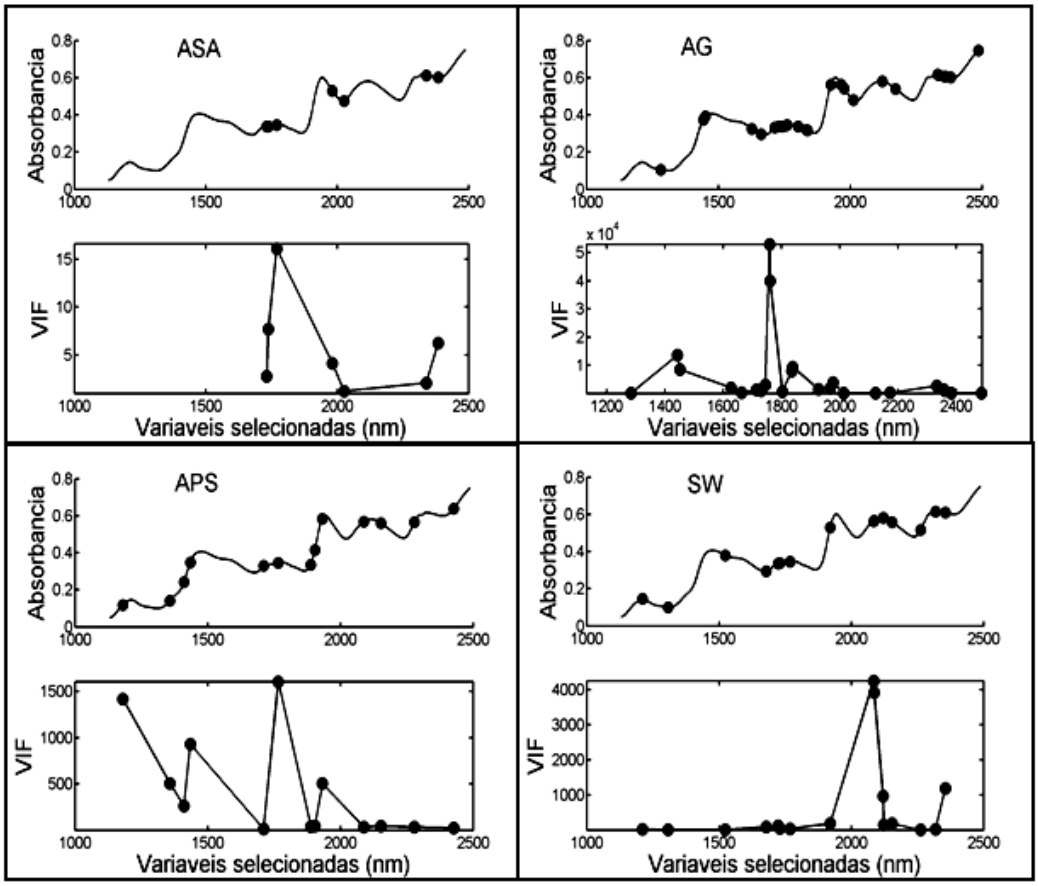

Figura 4.21. Variáveis selecionadas (e respectivos valores de VIF) pelos algoritmos ASA (sem o VIF), APS, AG e SW para o amido no milho.

4.4 Análise de amostras de gasolina por espectrometria NIR

4.4.1 Determinação de MON de gasolina

Na quantificação do parâmetro MON da gasolina, verifica-se na **Tabela 4.11** o mesmo resultado para os modelos MLR-ASA-VIF e MLR-ASA. Os valores de RMSEP foram ligeiramente menores que os dos outros modelos e com variáveis fortemente multicolineares (**Figuras 4.22** e **4.23**). Os métodos de seleção de variáveis produziram modelos MLR mais simples (poucas variáveis) que o PLS.

Tabela 4.11. Valores de RMSEP para o parâmetro MON de gasolina.

Modelo	Nº de variáveis selecionadas	$RMSEP_{val}$	$RMSEP_{prev}$
MLR-ASA-VIF (5)	5	0,31	0,37
MLR-ASA-VIF (10)	5	0,31	0,37
MLR-ASA-VIF (30)	5	0,31	0,37
MLR-ASA-VIF (50)	5	0,31	0,37
MLR-ASA	5	0,31	0,37
MLR-APS	6	0,29	0,39
MLR-AG	10	0,35	0,38
MLR-SW	3	0,34	0,41
PLS	10	0,37	0,40

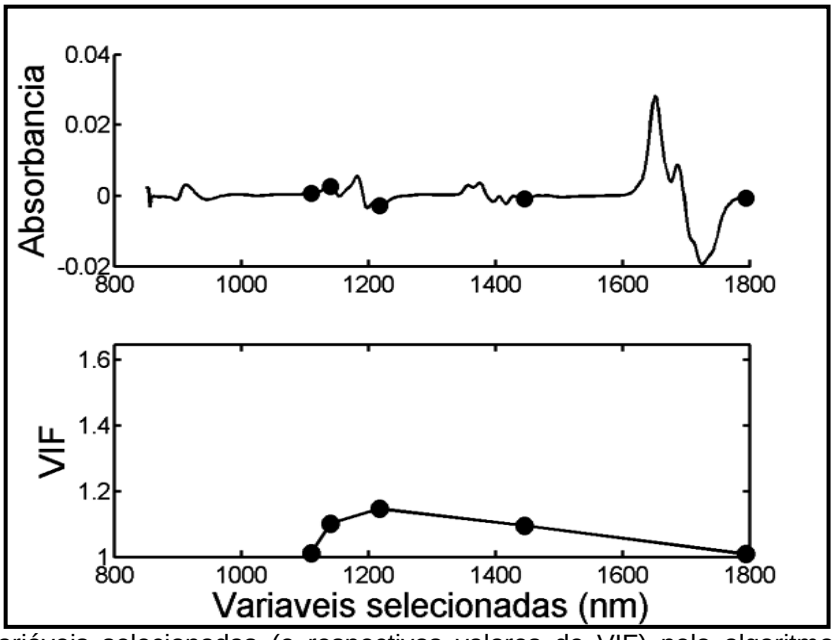

Figura 4.22. Variáveis selecionadas (e respectivos valores de VIF) pelo algoritmo ASA-VIF (com limiares de 5, 10, 30, 50) para o parâmetro MON de gasolina.

A **Figura 4.23** revela que, na determinação de MON de gasolina, o algoritmo AG apesar de ter selecionado variáveis com pouca multicolinearidade (VIF < 10), o número ainda significativamente maior que o dos outros algoritmos (**Tabela 4.11**).

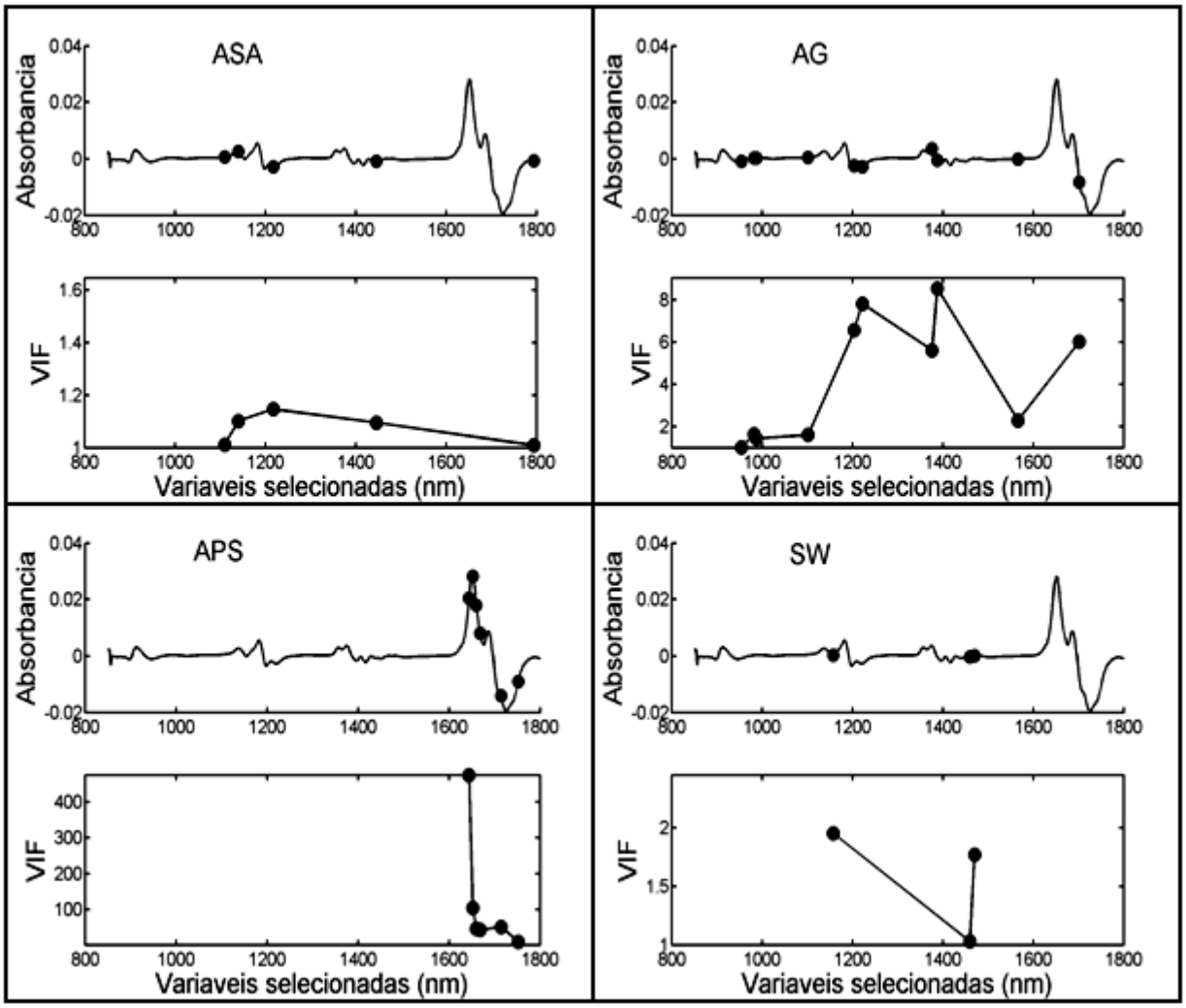

Figura 4.23. Variáveis selecionadas (e respectivos valores de VIF) pelos algoritmos ASA (sem o VIF), APS, AG e SW para o parâmetro MON de gasolina.

4.4.2 Determinação de T90% de gasolina

A **Tabela 4.12** revela que, novamente, os resultados dos modelos MLR-ASA-VIF são exatamente iguais aos resultados estimados pelo MLR-ASA, ou seja, o critério VIF não influenciou o processo de seleção. Os valores de RMSEP obtidos pelos modelos MLR-ASA-VIF são equivalentes aos obtidos pelos outros modelos.

As variáveis selecionadas pelos algoritmos ASA-VIF (**Figura 4.24**) e SW (**Figura 4.25**) são praticamente não multicolineares, sendo que o SW selecionou apenas 2 variáveis, produzindo neste caso um modelo MLR mais parcimonioso.

Os resultados da aplicação do AG, por outro lado, mudaram completamente no caso deste parâmetro. O número de variáveis e o grau de multicolinearidade das variáveis selecionadas são muito maiores que os obtidos na determinação do MON (**Figura 4.25**).

No que diz respeito ao resultado do PLS, o número de variáveis latentes pode ser considerado muito alto. Em compensação, o valor de RMSEP é similar aos dos outros modelos de calibração.

Tabela 4.12. Valores de RMSEP [ºC] para o parâmetro T90% de gasolina

Modelo	Nº de variáveis selecionadas	$RMSEP_{val}$	$RMSEP_{prev}$
MLR-ASA-VIF (5)	4	1,5	1,9
MLR-ASA-VIF (10)	4	1,5	1,9
MLR-ASA-VIF (30)	4	1,5	1,9
MLR-ASA-VIF (50)	4	1,5	1,9
MLR-ASA	4	1,5	1,9
MLR-APS	7	1,5	1,7
MLR-AG	20	2,0	2,1
MLR-SW	2	1,9	2,1
PLS	10	1,8	1,9

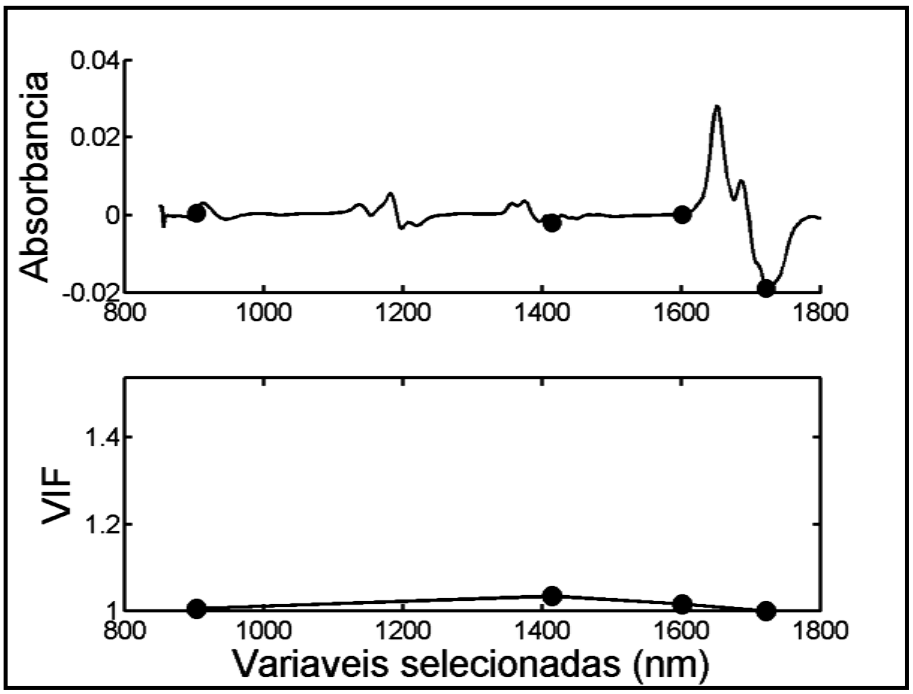

Figura 4.24. Variáveis selecionadas (e respectivos valores de VIF) pelo algoritmo ASA-VIF (com limiares de 5, 10, 30, 50) para o parâmetro T90% de gasolina.

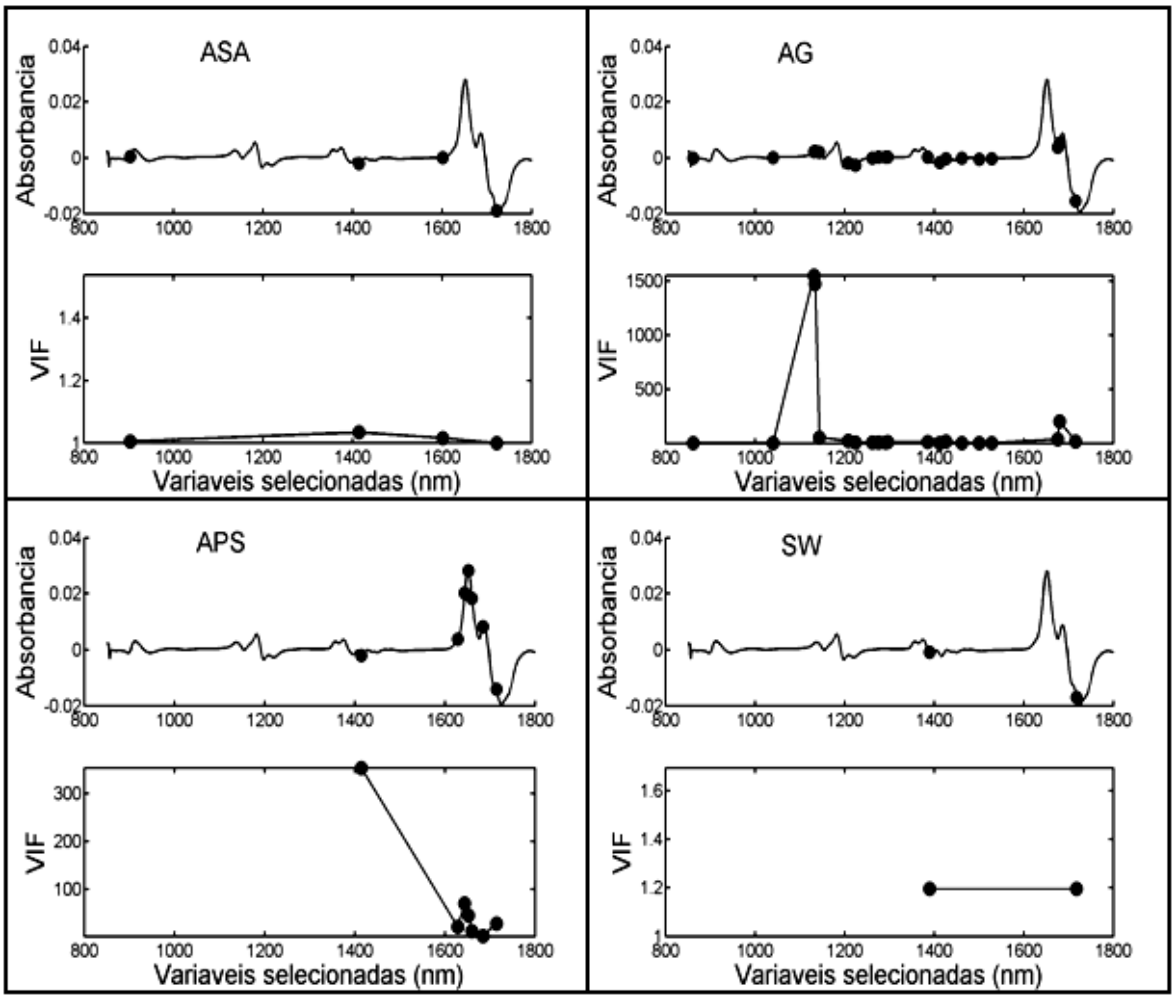

Figura 4.25. Variáveis selecionadas (e respectivos valores de VIF) pelos algoritmos ASA (sem o VIF), APS, AG e SW para o parâmetro T90% de gasolina.

Os resultados verificados para este conjunto de dados de gasolina mostram um comportamento semelhante, para o algoritmo proposto, aos dados do conjunto de trigo. Na determinação dos quatro parâmetros destes conjuntos, o ASA-VIF e o ASA produziram resultados iguais, em termos de RMSEP, utilizando variáveis selecionadas não-colineares. Isto comprova que o principio da busca angular pode, na prática, escolher também variáveis pouco multicolineares.

CAPÍTULO 5

CONCLUSÕES

5. CONCLUSÕES

Esta tese de doutorado apresentou uma nova técnica de seleção de variáveis (ASA-VIF) proposta para minimização de correlação e multicolinearidade em calibração multivariada baseada em Regressão Linear Múltipla (MLR). O algoritmo ASA-VIF permite selecionar as variáveis menos redundantes, melhorando o condicionamento da matriz de dados instrumentais. Para isso, o ASA-VIF incorporou no processo de busca o procedimento min-max para minimizar a correlação entre pares de variáveis e utiliza o critério VIF para abordar o problema da multicolinearidade.

O algoritmo proposto foi validado por meio de estudos de caso envolvendo quatro conjuntos de dados obtidos por duas técnicas instrumentais diferentes: um conjunto de absorção molecular UV/VIS (mistura de quatro corantes) e três conjuntos de dados NIR envolvendo a determinação de dois parâmetros de trigo, quatro de milho e dois parâmetros de qualidade da gasolina. Seu desempenho foi comparado com o de outras técnicas para seleção de variáveis (APS, AG e SW) em calibração MLR e com o método popular de calibração multivariada PLS.

Na determinação dos quatro corantes, o algoritmo ASA-VIF apresentou desempenho muito satisfatório selecionando quatro variáveis para três dos quatro corantes e cinco para um deles. Os resultados, em termos de número de variáveis e valores de RMSEP, permaneceram os mesmos independentemente do limiar adotado para o VIF. Além disso, as variáveis selecionadas usando os limites de VIF de 5, 10, 30 e 50 não apresentaram uma multicolinearidade significativa.

Na análise das amostras de trigo, assim como no caso da gasolina, o algoritmo ASA apresentou um desempenho similar usando ou não o critério VIF. Isso indica que o processo de busca angular, utilizado para selecionar variáveis minimamente correlacionadas, também pode reduzir a multicolinearidade. Todavia, nem sempre isso ocorrerá e, daí, a necessidade de implementar o critério VIF para minimizar definitivamente eventuais multicolinearidade dos dados.

No caso do milho, os modelos MLR, construídos a partir das variáveis selecionadas pelo ASA-VIF, produziram valores de RMSEP consideravelmente maiores que os outros métodos, especialmente para proteína e umidade. Mesmo assim, podemos considerar que os modelos são de utilidade prática, pois os erros médios relativos não ultrapassam 2% e as variáveis usadas não apresentam problemas de multicolinearidade.

O algoritmo ASA-VIF produziu modelos MLR com valores de RMSEP, para todos os parâmetros analisados, similares aos outros métodos usados para comparação, exceto para os parâmetros do milho. Neste caso, os valores de RMSEP foram maiores, porém o número de variáveis é consideravelmente menor, inclusive para os parâmetros dos outros conjuntos de dados. Como resultado, o ASA-VIF levou a obtenção de modelos MLR mais parcimoniosos sem comprometer significativamente a capacidade preditiva dos modelos.

Finalmente, podemos justificar o desempenho do algoritmo proposto, especialmente no tocante à maior parcimônia, como decorrentes dos seguintes fatos: i) incorporação do procedimento min-max para minimizar a correlação entre pares de variáveis e ii) aplicação do VIF nas cadeias de variáveis resultantes do procedimento min-max com o intuito de minimizar a multicolinearidade. Por conseguinte, o algoritmo proposto é uma ferramenta que tem potencial para seleção de variáveis em calibração multivariada via MLR como demonstrado nas análises espectrométricas UV-VIS e NIR.

5.1 Propostas Futuras

Na qualidade de propostas de trabalhos em continuação, podemos destacar as seguintes possibilidades:

- Estudar a robustez do algoritmo ASA-VIF. Isto pode ser feito variando os conjuntos de validação e verificando se ocorrem mudanças significativas nas variáveis selecionadas.

- Verificar o desempenho do algoritmo ASA-VIF em conjuntos de dados obtidos por outras técnicas analíticas instrumentais.

- Aplicar as variáveis selecionadas por esse algoritmo na escolha de sensores para construção de fotômetros dedicados.

- Modificar a estrutura básica do algoritmo a fim de realizar seleção de amostras minimamente redundantes, porém representativas do conjunto de dados.

CAPÍTULO 6

REFERÊNCIAS BIBLIOGRÁFICAS

6. REFERÊNCIAS BIBLIOGRÁFICAS

ARAÚJO, M. C. U.; SALDANHA, T. C. B.; GALVÃO, R. K. H.; YONEYAMA, T.; CHAME, H. C.; VISANI, V. The sucessive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems,* **57: 65, 2001.**

BEEBE, K.R,; PELL, R.J.; SEASHOLTZ, M.B. *Chemometrics: A Practical Guide*. John Wiley & Sons, INC. New York, 1998.

BENOUDJIT, N.; FRANÇOIS, D.; MEURENS,M.; VERLEYSEN, M. Spectrophotometric variable selection by mutual information. *Chemometrics and Intelligent Laboratory Systems*, **74: 243, 2004**.

BREITKREITZ, M. C.; RAIMUNDO, I. M., JR.; ROHWEDDER, J. J. R.; PASQUINI, C.; DANTAS FILHO, H. A.; JOSÉ, G. E.; ARAÚJO, M. C. U. Determination of total sulphur in diesel fuel employing nir spectroscopy and multivariate calibration. *Analyst,* **128: 1204, 2003.**

CALADO, V.; MONTGOMERY, D. C. Planejamento de experimentos usando *Statistica*. E-Papers Serviços Editoriais, Rio de Janeiro, 260 p., 2003.

CANECA, A. R.; PIMENTEL, M. F.; GALVÃO, R. K. H.; MATTA, C. E.; CARVALHO, F. R.; RAIMUNDO, JR., I. M.; PASQUINI, C.; ROHWEDDER, J. J. R. Assessment of infrared spectroscopy and multivariate techniques for monitoring the service condition of diesel-engine lubricating oils. *Talanta*, **70: 344, 2006**.

CECCHI, H. M. Fundamentos teóricos e práticos em análise de alimentos**,** 2.ed., p.189-207, Campinas, SP: Editora da Unicamp, 2003.

CENTNER, V.; MASSART, D. L.; NOORD, O. E.; JONG, S.; VANDEGINSTE, B. M.; STERNA, C. Elimination of uninformative variables for multivariate calibration. *Anal. Chem.* 68: **3851, 1996**.

CHONG, IL-GYO, JUN, CHI-HYUCK. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, **78: 103, 2005**.

DRAPER, N. R.; SMITH, H. *Applied Regression Analysis*, 3rd ed.; Wiley, New York, 1998.

FORINA, M.; LANTERI, S.; CASALE, M. Multivariate calibration: review. *Journal of Chromatography A*, **1158: 61, 2007.**

FORINA, M.; LANTERI, S.; CASALE, M.; OLIVEROS, M. C. C. Stepwise orthogonalization of predictors in classification and regression techniques: Na “old” technique revisited. *Chemometrics and Intelligent Laboratory Systems*, **87: 252, 2007.**

FORMIGA, F. M.; MEDEIROS, E. P.; NETO, J. G. V.; GAIAO, E. N.; SILVA, E. C.; ARAÚJO, M. C. U. - Um turbidímetro/nefelômetro de fluxo acoplado a um sistema *Flow-Batch* para "*Screening Analysis"* automática de cátions em medicamentos. *Controle & Instrumentação*, **83: 65, 2003**.

FREITAS, S. K. B. Uma Metodologia para *Screening Analysis* de Sucos Cítricos Utilizando um Analisador Automático em Fluxo-Batelada, Espectrometria UV-VIS e Técnicas Quimiométricas. Dissertação de Mestrado, João Pessoa, 2006.

GALVÃO, R. K. H.; ARAÚJO, M. C. U. (in press). Linear regression modelling: Variable selection. In B. Walczak, R. T. Ferré, S. Brown (Eds), Comprehensive chemometrics (2007).

GALVÃO, R. K. H.; ARAÚJO, M. C. U.; FRAGOSO, W. D.; SILVA, E. C.; JOSÉ, G. E.; SOARES, S. F. C.; PAIVA, H. M. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. *Chemometrics and Intelligent Laboratory Systems,* 2007, submetido.

GALVÃO, R. K. H.; PIMENTEL, M. F.; ARAÚJO, M. C. U.; YONEYAMA, T.; VISANI, V. Aspect of the sucessive projections algorithm for variable selection in multivariation calibration applied to plasma emission. *Analytica Chimica Acta*, **443: 107, 2001**.

GIACOMELLI, L.; BOGGETTI, H.; AGNELLI, H.; ANUNZIATA, J.; SILBER, J. J.; CATTANA, R. Relevant physicochemical factors in chromatographic separation of *Alternaria alternata* mycotoxins. *Analytica Chimica Acta*, **370: 79, 1998.**

GELADI, P.; KOWALSKI, B.R. *Anal. Chim. Acta,* **185: 1, 1985**.

HAALAND, D. M.; THOMAS, E.V. Partial Least-Squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem*., **60: 1193, 1988**.

HAIR, J. F.; TATHAM, R. L.; ANDERSON, R. E.; BLACK, W. *Análise Multivariada de Dados*. 5ª ed. Bookman, Porto Alegre, 2005.

HIBBERT, D. B. Genetic Algorithms in Chemistry. *Chemometrics and Intelligent Laboratory Systems*, **19: 277, 1993**.

HIJRCHNER, U.; KALIVAS, J. H. Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection. *Anal. Chim. Acta,* **311: 1, 1995**.

HOLLAND, J.H. *Adaptation in natural and artificial systems.* MIT Press, Ann Arbor, Michigan, 1975.

HONORATO, R. S.; ARAÚJO, M. C. U.; LIMA, R. A. C.; ZAGATTO, E. A. G.; LAPA, R. A. S.; LIMA, J. L. F. C., A flow-batch titrator exploiting an one-dimensional optimisation algorithm for end point search. *Anal. Chim. Acta*, **396: 91, 1999**.

JACKSON, J.E. *User´s Guide to Principal Componentes.* Wiley, New York, 1991.

KALIVAS, J.H.; ROBERTS, N.; SUTTER, J.M. Global optimization by simulated annealing with wavelength selection for ultraviolet-visible spectrophotometry. *Anal. Chem.* **61: 2024, 1989.**

KONZEN, P.H.A.; FURTADO, J.C.; CARVALHO, C.W.; FERRÃO, M. F.; MOLZ, R.F.; BASSANI, I.A.; HUNING, S.L. Otimização de métodos de controle de qualidade de fármacos usando algoritmo genetico e busca tabu. *Pesquisa Oeracional*, **23: 1, 2003**.

KOMPANY-ZAREH, M.; AKHLAGHI, Y. Correlation weighted successive projections algorithm as a novel method for variable selection in QSAR studies: investigation of anti-HIV activity of HEPT derivatives. *Journal of Chemometrics*, **21: 239, 2007**.

KUBELKA, P., MUNK, F. *Z. Tech. Physik***. 12: 593, 1931**.

LIMA, R. A. C. Um analisador fluxo-batelada multitarefa para a determinação de parâmetros físico-químicos de controle de qualidade de águas naturais - Tese de doutorado, Recife , 2006.

MARTENS, H., NAES, T. M*ultivariate calibration by data compression* in *near infrared technology in the agricultural and food industries*. Ed. P.Williams e K. Norris, American Society of Cereal Chemist, Inc. St. Paul, Minnesota, 1987.

NAES, T.; MEVIK, B.H. Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, **14: 413, 2001.**

MINDEL, B.D. *Process Control and Quality*, **9: 173, 1997**.

NORGAARD, L.; SAUDLAND, A.; WAGNER, J.; NIELSEN, J. P.; MUNCK, L.; ENGELSEN, S. B. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc.*, **54: 413, 2000**.

PASQUINI, C. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications*. J. Braz. Chem. Soc.,* **14: 198, 2003**.

PIMENTEL, M. F.; BARROS NETO, B. Calibração: uma revisão para químicos analíticos. *Química Nova*, **19: 268, 1996**.

PIZARRO, C.; ESTEBAN-DÍEZ, I.; NISTAL, A. J.; GONZÁLEZ-SÁIZ, J. M. Influence of data pre-processing on the quantitative determination of the ash content and lipids in roasted coffee by near infrared spectroscopy. *Analytica Chimica Acta*, **509: 217, 2004**.

PONTES, M. J. C.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; MOREIRA, P. N. T.; PESSOA NETO, O. D.; JOSÉ, G. E.; SALDANHA, T. C. B. The successive projections algorithm for spectral variable selection in classification problems. *Chemometrics and Intelligent Laboratory Systems*, **78: 11, 2005**.

PRADO, M. A.; GODOY, H. T. Determinação de corantes artificiais por cromatografia líquida de alta eficiência (CLAE) em pó para gelatina. *Química Nova*, **27: 22, 2004**.

SABOYA, N. P. *Análisis de control de preparados farmacêuticos mediante espectroscopia em el infrarrojo próximo*. Bellaterra, Universitat Autònoma de Barcelona, 2002, Tese de Doutorado.

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry,* **36: 1627, 1964.**

SCHENK, J.S.; WORKMAN JR. J.J.; WESTERHAUS, M.O. en *Handbook of Near-InfraredAnalysis*, ed. D.A. Burns, E.W. Ciurczak, Marcel Dekker Inc., New York, cap. 15, 1992.

SKOOG, D. A.; Holler, F. J.; Nieman, T. A.; Princípios de Análise Instrumental. 5ª Ed., Bookman, Porto Alegre, 2002.

STERNBERG, J.C.; STILLS, H.S.; SCHWENDEMAN, R.H. *Anal. Chem*., **32: 84, 1960**.

STOUT, F.; KALIVAS, J. H.; HÉBERGER, K. Wavelength selection for multivariate calibration using Tikhonov regularization. *Applied spectroscopy*, **61: 85, 2007**.

WOLD, H. *Soft Modeling by Latent Variables; the Non-linear Iterative Partial Least Squares Approach,* in *Perspectives in Probability and Statistics,* Ed. J. Gani, Academic Press, London, 1975.

WOLD, S.; ESBENSEN, K.; GELADI, P. *Chemometrics and Intelligent Laboratory Systems,* **2: 37, 1987**.

YE, S.; WANG, D.; MIN, S. Successive Projections Algorithm Combined with Uninformative Variable Elimination for Spectral Variable Selection. *Chemometrics and Intelligent Laboratory Systems,*(2007), DOI: 10.1016/j.chemolab.2007.11.005.

ANEXOS

7. ANEXOS

7.1 Cossenos de ângulos intervetores e coeficientes de correlação

Apresenta-se abaixo uma demonstração da correspondência que existe entre os cossenos dos ângulos intervetores, calculados pelo Algoritmo de Busca Angular, e os valores de correlação entre as variáveis associadas $\mathbf{x_i}$ e $\mathbf{x_j}$.

Por definição, o cosseno entre dois vetores x_i e x_j é dado pela razão entre o produto interno e produto das normas, de acordo com a expressão:

$$\cos(x_i, x_j) = \frac{<x_i, x_j>}{\|x_i\|\|x_j\|} \quad (1)$$

Se cada vetor é centralizado pela respectiva média, então o cosseno destes vetores é dado por:

$$\cos(x_i - \overline{x_i}, x_j - \overline{x_j}) = \frac{<x_i - \overline{x_i}, x_j - \overline{x_j}>}{\|x_i - \overline{x_i}\|\|x_j - \overline{x_j}\|} \quad (2)$$

Usando a definição de produto interno entre dois vetores e a definição da norma de um vetor, a Equação 2 fica:

$$\cos(x_i - \overline{x_i}, x_j - \overline{x_j}) = \frac{\sum(x_i - \overline{x_i})(x_j - \overline{x_j})}{\sqrt{\sum(x_i - \overline{x_i})^2}\sqrt{\sum(x_j - \overline{x_j})^2}} \quad (3)$$

Multiplicando e dividindo cada termo do denominador da Equação 3 por (N – 1), termos:

$$\cos(x_i - \overline{x_i}, x_j - \overline{x_j}) = \frac{\sum(x_i - \overline{x_i})(x_j - \overline{x_j})}{\sqrt{\frac{\sum(x_i - \overline{x_i})^2 (N-1)}{(N-1)}}\sqrt{\frac{\sum(x_j - \overline{x_j})^2 (N-1)}{(N-1)}}} \quad (4)$$

Como o produtos de duas raízes é a raiz do produto de seus termos, a Equação 4 torna-se:

$$\cos\left(x_i - \overline{x_\iota}, x_j - \overline{x_j}\right) = \frac{\sum\left(x_i - \overline{x_\iota}\right)\left(x_j - \overline{x_j}\right)}{\sqrt{\frac{\sum\left(x_i - \overline{x_\iota}\right)^2}{(N-1)}(N-1)\frac{\sum\left(x_j - \overline{x_j}\right)^2}{(N-1)}(N-1)}} \quad (5)$$

Multiplicando os temos (N – 1) do numerador da raiz, teremos:

$$\cos\left(x_i - \overline{x_\iota}, x_j - \overline{x_j}\right) = \frac{\sum\left(x_i - \overline{x_\iota}\right)\left(x_j - \overline{x_j}\right)}{\sqrt{\frac{\sum\left(x_i - \overline{x_\iota}\right)^2}{(N-1)}(N-1)^2\frac{\sum\left(x_j - \overline{x_j}\right)^2}{(N-1)}}} \quad (6)$$

Retirando o termo $(N - 1)^2$ da raiz, obteremos:

$$\cos\left(x_i - \overline{x_\iota}, x_j - \overline{x_j}\right) = \frac{1}{(N-1)}\frac{\sum\left(x_i - \overline{x_\iota}\right)\left(x_j - \overline{x_j}\right)}{\sqrt{\frac{\sum\left(x_i - \overline{x_\iota}\right)^2}{(N-1)}\frac{\sum\left(x_j - \overline{x_j}\right)^2}{(N-1)}}} \quad (7)$$

O desvio-padrão de cada vetor é dado pelas respectivas equações:

$$S_{x_i} = \sqrt{\frac{\sum(x_i - \overline{x_\iota})^2}{(N-1)}} \quad (8)$$

$$S_{x_j} = \sqrt{\frac{\sum\left(x_j - \overline{x_j}\right)^2}{(N-1)}} \quad (9)$$

Substituindo as Equações 8 e 9 na Equação 7, obteremos a Equação 10, que é a expressão da correlação entre os dois vetores x_i e x_j:

$$\cos\left(x_i - \overline{x_\iota}, x_j - \overline{x_j}\right) = \frac{1}{(N-1)}\frac{\sum\left(x_i - \overline{x_\iota}\right)\left(x_j - \overline{x_j}\right)}{S_{x_i}S_{x_j}} = corr(x_i, x_j) \quad (10)$$

Portanto, verifica-se que o cosseno e a correlação entre dois vetores x_i e x_j são matematicamente iguais.

7.2 Código-fonte do Programa ASA

O código-fonte do programa ASA, escrito em linguagem MATLAB 6.5, é apresentado a seguir:

Rotina 1:

```
function modelo = ASA_modelo(Xcal,Ycal,Xval,Yval,N,ind_VIF)


% - Algoritmo de Busca Angular(ASA-Angular Search Algorithm)para
Selecao de Variaveis em Calibracao
%      Multivariada por RLM Baseado nos Maiores Ângulos entre Vetores-
Colunas de Xcal -

%  - ROTINA DE CALIBRAÇAO - VALIDAÇAO SERIE DE TESTE

% LINHA DE COMANDO PARA EXECUÇAO DO PROGRAMA:
%  modelo = ASA_modelo(Xcal,Ycal,Xval,Yval,N)
% DADOS DE ENTRADA:
% Xcal   --> Respostas instrumentais das amostras do conjunto de
calibracao (#amostras x #variaveis)
% Ycal   --> Concentracoes de analitos ou parametros das amostras do
conjunto de calibracao (#amostras x #analitos)
% Xval --> Respostas instrumentais das amostras do conjunto (ou serie)
de validaçao ou teste (#amostras x #variaveis)
% Yval --> Concentracoes de analitos ou parametros das amostras da
serie de validaçao ou de teste (#amostras x #parametros)
% N  --> Numero de variaveis a selecionar (min= numero de analitos ou
parametros e max= numero de amostras de Xcal)

% SAIDA DO PROGRAMA:
% modelo --> modelo de calibraçao
   %  modelo.modelo='ASA_MLR';
   % modelo.corte = indices_mantidos;  -> variaveis apos o corte usando
o citerio de 0% da norma maxima (eliminar possiveis variaveis zero)
   % modelo.media_Xcal = mXc;  -> media de Xcal que ser usada na
centralização de Xprev (rotina ASA_prev)
   % modelo.media_Ycal = mYc;  -> media de Ycal que ser usada na
centralização de Yprev (rotina ASA_prev)
   % modelo.coef = B;   -> coeficientes do modelo, usado para
determinar Yestimado
   % modelo.No_var = length(msele);  -> numero de variaveis
selecionadas
   % modelo.var = msele;   -> variaveis selecionadas
   % modelo.RMSEP = rmsep;  -> RMSEP absoluto e relativo
   % modelo.correlacao = rp;  -> correlaçao entre valores estimados e
referencia
```

```
% Autores do Programa ASA
% Prof. Edvan Cirino da Silva - DQ/UFPB (orientador)
% Prof Roberto Kawakami (ITA)
% Pedro Germano - Doutorando (UFPB)
% Sofacles Figueredo - Iniciaçao Cientifica (UFPB)
% Versao 1.0 (Dezembro 2007)

% Eliminaçao das variaveis com normas menores que 0% da norma maxima

norma_Xcal = sqrt(sum(Xcal.^2));
Limiar_Corte = (0/100)*max(norma_Xcal);
indices_mantidos = find(norma_Xcal > Limiar_Corte);
Xcal = Xcal(:,indices_mantidos);
Xval = Xval(:,indices_mantidos);

 % CENTRALIZAÇÃO NA MEDIA

Nmist_cal = size(Xcal,1);
Nmist_val = size(Yval,1);
% Centralizaçao do X
mXc = mean(Xcal);
Xcal = Xcal - repmat(mXc,Nmist_cal,1);
Xval = Xval - repmat(mXc,Nmist_val,1);
% Centralizaçao do Y
mYc = mean(Ycal);
Ycal = Ycal - repmat(mYc,Nmist_cal,1);
Yval = Yval - repmat(mYc,Nmist_val,1);

% Definiçoes de veriaveis
    Nmist_cal = size(Xcal,1);
    M = size(Xcal,2);
    Nmist_val = size(Yval,1);
    Ncomp = size(Yval,2);


 % Calculo dos cossenos dos angulos intervetores da matriz Xcal

m = ones(M,N);
variaveis = 1:M;
C = ones(M,M);

for z = 1:M
        for j = z+1:M
            Produtonormas = norm(Xcal(:,z))*norm(Xcal(:,j));
            ProdutoInterno = sum(Xcal(:,z).*Xcal(:,j));
            C(z,j)=ProdutoInterno/Produtonormas;
            C(j,z) = C(z,j);
        end
end
C = abs(C);

% Montagem das cadeis com as variaveis menos correlacionadas

h = waitbar(0,'Montagem das cadeias...');
loopStart = now;

for variavel_inicial = 1:M
    loopEnd = loopStart + (now-loopStart)*M/variavel_inicial;
```

```
    waitbar(variavel_inicial/M,h,['Montando cadeias. Conclusao: '
datestr(loopEnd)]);
    V = [];
    V(1,1) = variavel_inicial;
    for i = 2:N
       pool = setdiff(variaveis,V);
       cmax = [];
       for j = 1:length(pool)
           indexa = pool(j);
           c = [];
           for k = 1:length(V)
               indexb =  V(k);
               c(k) = C(indexa,indexb);
           end
           cmax(j) = max(c);
       end

    [dummy,index] = min(cmax);
    V(1,i) = pool(index);
    end
    m(variavel_inicial,:) = V;
end
close(h);

% Determinação do RMSEP p/ todas as cadeias de variaveis

R = zeros(1,N);
Lopt = zeros(N,N);
rmsep = zeros(N,M);

h = waitbar(0,'Realizando regressoes...');
loopStart = now;

for n = 1:N
  loopEnd = loopStart + (now-loopStart)*N/n;
  waitbar(n/N,h,['Realizando regressoes. Conclusao: '
datestr(loopEnd)]);
  for i = 1:M
    lambdas = m(i,1:n);
     Xcal2 = Xcal(:,lambdas);

     % CALCULO DO VIF
     if size(Xcal2,2) > 1
       [VIF]= calculo_vif(Xcal2);
       VIF = VIF';
       ind = find(VIF < ind_VIF);
       lambdas = lambdas(:,ind);
   end

     Xcal2 = Xcal(:,lambdas);
     Xval2 = Xval(:,lambdas);
    B = inv(Xcal2'*Xcal2)*Xcal2'*Ycal;
     %B = Xcal2\Ycal;
    Yestimado = Xval2*B;
    rmsep(n,i)=sqrt(sumsqr(Yestimado-Yval)/(Nmist_val*Ncomp));
end
  [R(n) imin] = min(rmsep(n,:));
  Lopt(1:n,n) = m(imin,1:n)';
```

```
end
close(h);
[Rbest,Nbest] = min(R);
msele = (Lopt(1:Nbest,Nbest))';

%Modificacao 12 Fev 2008
Xcal2 = Xcal(:,msele);
if size(Xcal2,2) > 1
    [VIF]= calculo_vif(Xcal2);
    VIF = VIF'
    ind = find(VIF < ind_VIF);
    msele = msele(:,ind);
end

% Etapa de eliminacao de variaveis baseado no criterio de Haaland-
Thomas

% Passo 1: Restringir os conjuntos de calibracao e teste as variaveis
% selecionadas
Xcal2 = Xcal(:,msele);
Xval2 = Xval(:,msele);
% Passo 2: Determina o modelo via RLM + QR nas variaveis selecionadas
 B = inv(Xcal2'*Xcal2)*Xcal2'*Ycal;
%B = Xcal2\Ycal; % Jbar x Ncomp
% Passo 3: Calcula o coeficiente de relevancia
desvio = std(Xcal2); % 1 x Jbar
for i = 1:length(msele)
    magB(i) = norm(B(i,:));
end
relev = desvio.*magB;
% Passo 4: Dispoe as variaveis em ordem decrescente de relevancia
[dummy,sortrelev] = sort(relev); % Ordem crescente
sortrelev = fliplr(sortrelev); % Ordem decrescente
% Passo 5: Realiza regressoes e calcula RMSEV em funcao do numero de
% variaveis incluidas no modelo (seguindo a ordem decrescente de
relevancia)
for i=1:length(msele)
    Xcal3 = Xcal2(:,sortrelev(1:i));
    Xval3 = Xval2(:,sortrelev(1:i));
     B = inv(Xcal3'*Xcal3)*Xcal3'*Ycal;
   % B = Xcal3\Ycal;
    Yestimado = Xval3*B;
    rmsev(i)=sqrt(sumsqr(Yestimado-Yval)/(Nmist_val*Ncomp));
end
sortrelev;
figure
axes('FontSize',16)
plot(1:i,rmsev,'k-','LineWidth',2),grid
title('Scree Plot','FontSize',18)
xlabel('Numero de Variaveis Selecionadas','FontSize',18)
ylabel('RMSEV','FontSize',18)
print -dtiff -r600 scree.tif
% Passo 6: Encontra o ponto de RMSEV minimo
rmsevmin = min(rmsev);
% Passo 7: Determina o ponto de corte para RMSEV com base no teste F
com
% significancia alpha = 0.25
```

```
alpha = 0.25;
dof = Nmist_val*Ncomp; % Numero de graus de liberdade no numerador e no
denominador
fcrit = finv(1-alpha,dof,dof);
% Passo 8: Encontra o menor numero de variaveis que ainda conduz a um
RMSEV que nao eh
%          significativamente maior que o RMSEV minimo.
rmsevmax = rmsevmin*sqrt(fcrit);
indexopt = min(find(rmsev < rmsevmax));
% Passo 9: Monta o vetor de variaveis selecionadas
msele = msele(sortrelev(1:indexopt));


% Resultados de validaçao

Xcal2 = Xcal(:,msele);
Xval2 = Xval(:,msele);
 B = inv(Xcal2'*Xcal2)*Xcal2'*Ycal;
%B = Xcal2\Ycal;

% Grafico de residuos de concentraçao das amostras de calibraçao
Yest_cal = Xcal2*B;
resid_cal = Yest_cal - Ycal;
figure
axes('FontSize',16)
bar(resid_cal,'k'),grid
xlabel('Amostras','FontSize',18)
ylabel('Residuo de concentraçao','FontSize',18)
print -dtiff -r600 resid.tif

Yestimado = Xval2*B;
rmsep=sqrt(sumsqr(Yestimado-Yval)/(Nmist_val*Ncomp));
rmsepr=sqrt((1/Nmist_val)*sumsqr((Yestimado-Yval)./(Yval+mYc))); %
rmsep relativo
rmsep = [rmsep rmsepr*100];


% Correlacao

if size(Yval,2) == 1

x = Yval;
y = Yestimado;
n = length(y);
num = n*sum(x.*y) - sum(x)*sum(y);
den = sqrt( ( n*sum(x.^2) - (sum(x))^2 ) * ( n*sum(y.^2) - (sum(y))^2 )
);
rp = num/den;
end

%Construindo arquivo de saída
modelo.modelo='ASA_MLR';
modelo.corte = indices_mantidos;
modelo.pre_processamento = 'centralização na media';
modelo.media_Xcal = mXc;
modelo.media_Ycal = mYc;
modelo.coef = B;
modelo.No_var = length(msele);
```

```
modelo.var = msele;
modelo.RMSEP = rmsep;
modelo.correlacao = rp;
modelo.valor_N = N;

% Grafico de Valor Predito vs Valor de Referencia
 rmsep = rmsep(1);
 figure
 axes('FontSize',16)
 plot(Yval + mYc,Yestimado +
mYc,'ko','MarkerSize',12,'MarkerEdgeColor','k','MarkerFaceColor','k')
 title(['Validaçao: RMSEP = ' num2str(rmsep) ', Correlacao = '
num2str(rp)],'FontSize',18)
 xlabel('Valor de referencia','FontSize',18)
 ylabel('Valor predito','FontSize',18)

% % Reta bissetriz
 minimo = min([Yval + mYc;Yestimado + mYc]);
 maximo = max([Yval + mYc;Yestimado + mYc]);
 line([minimo,maximo],[minimo,maximo],'LineWidth',2)
print -dtiff -r600 prevrefval.tif

% Calculo do VIF das variavei selecionadas
 Xcal = Xcal + repmat(mXc,Nmist_cal,1);
 Xcal2 = Xcal(:,modelo.var);
[VIF] = calculo_vif(Xcal2);
modelo.VIF = VIF;
```

Rotina 2:

```
function previsao = ASA_prev(modelo,Xprev,Yprev)

% - Algoritmo de Busca Angular(ASA-Angular Search Algorithm)para
Selecao de Variaveis em Calibracao
%      Multivariada por RLM Baseado nos Maiores Ângulos entre Vetores-
Colunas de Xcal -

%  - ROTINA DE PREVISAO

% LINHA DE COMANDO PARA EXECUÇAO DO PROGRAMA:
%  previsao = ASA_prev(modelo,Xprev,Yprev)
%
% DADOS DE ENTRADA:
% modelo -> modelo de calibraçao obtido na rotina " ASA_modelo "
% Xprev  --> Respostas instrumentais das amostras do conjunto de
previsao (#amostras x #variaveis)
% Yprev  --> Concentracoes de analitos ou parametros das amostras de
previsao (#amostras x #parametros)

% SAIDA DO PROGRAMA:
% PREVISAO -> Dados da previsao
% previsao.previsao='ASA_MLR';
% previsao.Yestimado = Yestimado;  -> valores do Y (concentraçao)
estimdo
% previsao.RMSEP = rmsep; -> RMSEP absoluto e relativo (caso seja dado
os valores de Yprev)
% previsao.correlacao = rp; -> correlaçao (caso seja dado os valores de
Yprev)
```

```
Xprev = Xprev(:,modelo.corte);
Nmist_prev = size(Xprev,1);

% Centralizaçao do X
Xprev = Xprev - repmat(modelo.media_Xcal,Nmist_prev,1);

 % Previsao do conjunto Xprev

 Xpred2 = Xprev(:,modelo.var);
 Yestimado = Xpred2*modelo.coef;

 % Calculo do RMSEP associado a cadeia, caso seja dado os valores de
Yprev

 if nargin > 2

  Ncomp_prev = size(Yprev,1);
  Ncomp = size(Yprev,2);

% Centralizaçao do Y
Yprev = Yprev - repmat(modelo.media_Ycal,Ncomp_prev,1);

rmsep=sqrt(sumsqr(Yestimado-Yprev)/(Ncomp_prev*Ncomp));  % rmsep
absoluto
rmsepr=sqrt((1/Nmist_prev)*sumsqr((Yestimado-Yprev)./(Yprev +
modelo.media_Ycal))); % rmsep relativo
rmsep = [rmsep rmsepr*100];

% Correlacao e Grafico Predito vs Referencia --> Se y so tiver uma
coluna

 x = Yprev;
 y = Yestimado;
 n = length(y);
num = n*sum(x.*y) - sum(x)*sum(y);
den = sqrt( ( n*sum(x.^2) - (sum(x))^2 ) * ( n*sum(y.^2) - (sum(y))^2 )
);
rp = num/den;  % correlaçao

%Construindo arquivo de saída
previsao.previsao='ASA_MLR';
previsao.Yestimado = Yestimado + modelo.media_Ycal;
previsao.RMSEP = rmsep;
previsao.correlacao = rp;


% % Grafico de Valor Predito vs Valor de Referencia
rmsep = rmsep(1);
figure
axes('FontSize',16)
plot(Yprev + modelo.media_Ycal,Yestimado +
modelo.media_Ycal,'ko','MarkerSize',10,'MarkerEdgeColor','k','MarkerFac
eColor','k')
title(['Predicao: RMSEP = ' num2str(rmsep) ', Correlacao = '
num2str(rp)],'FontSize',18)
xlabel('Valor de referencia','FontSize',18)
ylabel('Valor predito','FontSize',18)
```

```
% Reta bissetriz
minimo = min([Yprev + modelo.media_Ycal;Yestimado +
modelo.media_Ycal]);
maximo = max([Yprev + modelo.media_Ycal;Yestimado +
modelo.media_Ycal]);
line([minimo,maximo],[minimo,maximo],'LineWidth',2)
print -dtiff -r600 prevrefprev.tif
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
else
  %Construindo arquivo de saída
previsao.previsao='ASA_MLR';
previsao.Yestimado = Yestimado + modelo.media_Ycal;
end
```

Rotina 3:

```
function var_sel = ASA_plot(modelo,Xcal_bruto,xaxis,lamb_inicial,res)

% - Algoritmo de Busca Angular(ASA-Angular Search Algorithm)para
Selecao de Variaveis em Calibracao
%      Multivariada por RLM Baseado nos Maiores Ângulos entre Vetores-
Colunas de Xcal -

%  - Rotina para fornecer as variaveis selecionadas pelo " ASA_modelo "
na unidade desejada e em forma grafica

% LINHA DE COMANDO PARA EXECUÇAO DO PROGRAMA:
%  var_sel = ASA_plot(modelo,xaxis,lamb_inicial,seq)
%
% DADOS DE ENTRADA:
% modelo -> modelo de calibraçao obtido na rotina "ASA_modelo"
% xaxis -> valores do eixo x na unidade desejada
% lamb_inicial -> valor do comprimento de onda inicial
% res -> resoluçao dos valores de xaxis (se de 1 em 1, 2 em 2, ...)


% % Grafico das variaveis selecionadas no espectro original

espectro = mean(Xcal_bruto);

a2 = modelo.corte(modelo.var);
a1 = [espectro(:,a2)];
 figure
 subplot(2,1,1)
 hold on
 plot(xaxis,espectro,'k-','LineWidth',2)
vasa = xaxis(:,a2);
plot(vasa,a1,'ok','MarkerSize',12,'MarkerEdgeColor','k','MarkerFaceColo
r','k')
ylabel('Absorbancia','FontSize', 24)
hold off
subplot(2,1,2)
var_sel = sort(res*a2 + (lamb_inicial - res));
plot (var_sel,modelo.VIF,'-
ok','LineWidth',2,'MarkerSize',12,'MarkerEdgeColor','k','MarkerFaceColo
r','k')
xlabel('Variaveis selecionadas (nm)','FontSize', 24)
ylabel('VIF','FontSize', 24)
axis([min(xaxis),max(xaxis),1,max(modelo.VIF)+0.5])
```

```
%print -dtiff -r600 varselvif.tif
var_sel = sort(res*a2 + (lamb_inicial - res));

var_sel.variaveis_selecionadas = var_sel; % vaiaveis selecionadas na
unidade desejada
```

Sub-rotina para cálculo do VIF

```
function [VIF] = calculo_vif(x)
% Rotina para determinar multicolinearidade
% Metodo VIF
% Pedro Germano
% 13/12/2007

M = size(x,2);
N = size(x,1);

for z = 1:M
    c = [[1:z-1] [z+1:M]];
    tc = size(c,2);
    xr = x(:,c);
    y = x(:,z);
    ry = repmat(mean(y),N,1);
     %Xones = [ones(N,1) xr];
    Xones = [xr];
    b = inv(Xones'*Xones)*Xones'*y;
    yp = Xones*b;
    ryp = repmat(mean(yp),N,1);
    cor(z,:) = (1/(N-1))*sum(((y - ry)/std(y)).*((yp - ryp)/std(yp)));

    VIF(z,:) = 1/(1 - cor(z)^2);
end
```