UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE QUÍMICA

UNICAMP

DISSERTAÇÃO DE MESTRADO

SELEÇÃO DE VARIÁVEIS EM CALIBRAÇÃO MULTIVARIADA A PARTIR DA HESSIANA DOS ERROS

SILVIO LUIS TOLEDO DE LIMA

ORIENTADOR: PROF. DR. RONEI JESUS POPPI

CAMPINAS
NOVEMBRO – 2000

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO INSTITUTO DE QUÍMICA UNICAMP

L628s

Lima, Silvio Luis Toledo de
Seleção de variáveis em calibração multivariada a partir da Hessiana dos erros / Silvio Luis Toledo de Lima. -- Campinas, SP: [s.n], 2000.

Orientador: Ronei Jesus Poppi.

Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Química.

1. Seleção de variáveis. 2. Cana de açúcar. 3. Mínimos quadrados parciais. 4. Infravermelho próximo. I. Poppi, Ronei Jesus. II. Universidade Estadual de Campinas, Instituto de Química. III.Título.

"... a loucura de Deus é mais sábia que a sabedoria do homem, e a fraqueza de Deus é mais forte que força do homem... Ninguém se engane a si mesmo; se alguém dentre vós se tem por sábio neste mundo, faça-se louco para se tornar sábio. Porque a sabedoria deste mundo é loucura diante de Deus... O Senhor conhece os pensamentos dos sábios, que são vãos."

PAULO, APÓSTOLO DE JESUS CRISTO,
I CARTA À IGREJA DE CORINTHO, 1:25, 3:18-19 E 3:20B[1]

[1] Texto Extraído de "A Bíblia Sagrada", traduzida para o português por João Ferreira de Almeida, Edição Revista e Corrigida, 17ª Impressão, Imprensa Bíblica Brasileira, Rio de Janeiro, 1984.

À minha esposa, Helen, pela paciência e incentivo; ao meu pai, Edison, pelos bons exemplos de um homem Culto e à minha Mãe, Maria José (in memoriam).

Aos meus Tios José, Tita e Apparecida, por todo o apoio, e aos meus irmãos, Billy e Heloísa, pelo incentivo e confiança em mim depositados.

AGRADECIMENTOS

- Primeiramente a Deus, pela minha vida e por ter me concedido inúmeras oportunidades, tanto acadêmicas quanto pessoais, que possibilitaram a realização e conclusão de mais este trabalho;

- Ao Prof. Ronei Jesus Poppi, pela amizade, paciência e pelo conhecimento transmitido, e ao Prof. César Alexandre Mello, pelas discussões que geraram a idéia central deste trabalho;

- Ao Centro de Tecnologia Copersucar de Piracicaba, pela gentil cessão das amostras de caldo e xarope de cana-de-açúcar;

- Ao Lídio K. Takayama, da FEMTO, pela cessão dos espectros adquiridos para estas amostras;

- Ao Magnus Norgaard, da *Technical University of Denmark*, por disponibilizar uma versão do *software* "*OBSPRUNE*", que serviu como referência durante o desenvolvimento desta dissertação;

- A todo o grupo do Laboratório de Quimiometria em Química Analítica, Paulo Fidêncio, Cleidiane, Rosângela, Jacqueline, Marcelo, Eduardo pelo convívio agradável e amizade, e em especial ao Paulo Augusto, pelas boas discussões acadêmicas e filosóficas;

- Ao Pablo Marcelo, Jefferson, André do Couto, Alessandro Adinolfi, Alexsandro Sunaga, Fábio e Francis Bozolan, pela amizade e paciente convivência;

- Aos funcionários do Instituto de Química, pela diligente execução de suas funções, em especial à Bel, da Coord. de Pós-Graduação, e ao Toninho, da Biblioteca do Instituto, por sempre fazerem muito além do que lhes era exigido.

- Ao Puka e ao Borys, pelos momentos de descontração;

- À CAPES pelo apoio financeiro;

- À todos aqueles que direta ou indiretamente tiveram sua parcela de participação durante a execução e conclusão deste trabalho.

SELEÇÃO DE VARIÁVEIS EM CALIBRAÇÃO MULTIVARIADA A PARTIR DA HESSIANA DOS ERROS

Silvio Luis Toledo de Lima*, Prof. Dr. Ronei Jesus Poppi†.

RESUMO

Neste trabalho foi desenvolvido um método para seleção de variáveis em determinações espectroscópicas multicomponente utilizando calibração multivariada, através da eliminação seletiva das variáveis não informativas.

Este método, baseia-se na análise do modelo PLS1 (Partial Least Squares), que estabelece uma relação linear entre as variáveis dependentes (quantidade de analito em uma amostra) e as variáveis independentes (absorbância a vários comprimentos de onda), através do cálculo de um vetor de coeficientes de regressão. A eliminação de uma determinada variável ocorre pela anulação do valor do coeficiente de regressão associada a ela. A escolha da variável a ser eliminada dá-se segundo a análise da Hessiana da matriz de erros, buscando minimizar o erro de previsão do modelo.

O método proposto foi aplicado em dados espectroscópicos no infravermelho próximo, cujo objetivo era determinar as quantidades de açúcares em amostras de caldo de cana e xarope de açúcares.

Os resultados obtidos foram bastante satisfatórios, pois o método foi capaz de reduzir o número de variáveis em até 96% sem, com isso, provocar uma diminuição na capacidade de previsão.

A partir destes resultados é possível simplificar os modelos construídos para determinação da quantidade de açúcares presentes no caldo de cana, possibilitando o emprego de métodos mais robustos como a regressão linear múltipla.

Palavras-Chave: *Seleção de Variáveis, Cana-de-açúcar, Mínimos Quadrados Parciais, Infravermelho Próximo*

---

* Autor
† Orientador

VARIABLE SELECTION IN MULTIVARIATE CALIBRATION FROM THE HESSIAN OF ERRORS

Silvio Luis Toledo de Lima*, Prof. Dr. Ronei Jesus Poppi†.

ABSTRACT

A method of variables selection was developed and applied to multicomponent spectroscopic determinations using multivariate calibration through selective elimination of uninformative variables.

This method is based on PLS (Partial Least Squares) model analysis, that establish a linear relationship between dependent variables (quantities of analyte in a sample) and independent variables (absorbance in several wavelengths), by the calculation of one regression coefficient vector. The elimination of a specific variable occurs when its associated regression coefficient is set to zero. The choosing of a variable that is going to be eliminated is given by the Hessian error's matrix analysis, in which is tried to minimize the model's predictive error.

The proposed method was applied to near infrared spectroscopic data, which objective was to determine sugar quantities in two kinds of samples, sugar cane juice and sugar syrup.

The results obtained were very satisfactory, since the method was able to reduce even up to 96% a set of data, without produce a decrease in its predictive ability.

From these results, it is possible simplify the models built to determine sugar quantities presents in sugar cane juice, or sugar syrup, allowing use of more robust methods such as the Multiple Linear Regression.

Keywords: Variable Selection, Sugar Cane, Partial Least Squares, Near Infrared

* Author
† Advisor

Cidade Universitária Zeferino Vaz s/n
Barão Geraldo – Campinas – SP
Cx. Postal: 6154 – CEP 13083-970
(19) 3788-3126
slima@iqm.unicamp.br

Sílvio Luis Toledo de Lima

Informações Pessoais

- Brasileiro, 25 anos, casado

Idiomas

- INGLÊS: Boa Leitura , Conversação e Escrita (Nível Intermediário)
- ESPANHOL: Boa Leitura (Técnico-Científico)

Informática

- APLICATIVOS: Windows98, Microsoft Office, Origin 5.0, CorelDraw 7.0, ChemWin, ISIS, InternetExplorer, FrontPage
- PROGRAMÁVEIS: Matlab 5.3, TPascal 7.0, Mathematica 3.0, Noções de C^+ e Java

Formação Acadêmica

1994 -1997 **Bacharel em Química pela Universidade Estadual de Campinas**

1997 - 2000 **Licenciado em Química pela Universidade Estadual de Campinas**

1998 - 2000 **Mestre na Área de Química Analítica pela Universidade Estadual de Campinas**

Doutoramento

"Detecção de Falhas em Métodos Espectro-Analíticos usando Redes Bayesianas" – Prof. Dr. Ronei Jesus Poppi – 11/00 a 10/04

Experiência Didática

2000 Instituto de Química - UNICAMP Campinas - SP
Palestrante
- Seminário: "Amostragem de Suspensões: Uma Técnica Aliada da Espectrometria Atômica"

2000 Instituto de Química - UNICAMP Campinas - SP
Palestrante
- Seminário: "Otimização de Sistemas de Várias Variáveis Restritas à Condições de Contorno, pelo Método dos Multiplicadores de Lagrange"

1999 Instituto de Química - UNICAMP Campinas - SP
Palestrante
- Seminário: "Eliminação de Variáveis em Determinações Espectroscópicas Multicomponente Utilizando Calibração Multivariada"

1998-2000 Aulas Particulares Campinas - SP
Professor de Química
- Aulas particulares teóricas para 2° Grau.

Simpósios e Congressos

Maio/96	19ª Reunião Anual da Soc. Brasileira de Química	Poços de Caldas - MG
Maio/97	20ª Reunião Anual da Soc. Brasileira de Química	Poços de Caldas - MG
Maio/98	21ª Reunião Anual da Soc. Brasileira de Química	Poços de Caldas - MG
Ago-Set/99	10º Encontro Nacional de Química Analítica	Santa Maria - RS

Interesses

Amizades, religião, cinema, computadores, surdez, cães.

SELEÇÃO DE VARIÁVEIS EM CALIBRAÇÃO MULTIVARIADA A PARTIR DA HESSIANA DOS ERROS

Índice

INTRODUÇÃO

Com os avanços alcançados na tecnologia digital eletrônica, novos equipamentos e técnicas instrumentais de análise em química surgiram. Aliado a isso, a possibilidade de interfaceamento destes instrumentos a computadores dotados de processadores de alta velocidade permitiu que uma enorme quantidade de dados pudesse ser adquirida em laboratório num curto intervalo de tempo. Porém, maior quantidade de dados não significa necessariamente maior quantidade de informações, pois apenas quando tais dados são devidamente interpretados e postos em uso, é que se pode dizer que estes se tornaram importantes para analistas de diversas áreas.

Por esta razão, métodos de análise capazes de manipular uma grande quantidade de dados tornaram-se uma necessidade para a prática em química, principalmente a química analítica. Nesse sentido surgiu a *Quimiometria*, uma área que propunha cumprir com eficiência estas novas exigências da química moderna.

Dentro da *Análise Multivariada*, que constitui o cerne da Quimiometria, encontra-se a *Calibração Multivariada* - métodos que correlacionam dados multivariados às propriedades características de determinada substância de interesse. Como exemplo pode-se correlacionar os espectros de misturas de açúcares, adquiridos na região do infravermelho próximo [1, 2], ou infravermelho médio [3,4], com as quantidades efetivas destas substâncias.

Determinações quantitativas através de técnicas espectroscópicas são muito apropriadas para análise de misturas, visto que são simples, rápidas, baratas e não destrutivas. Outro fator importante para o emprego destas técnicas reside no fato de algumas delas exibirem respostas lineares quanto às quantidades de analito em estudo. Por estes motivos surgiu, na indústria açucareira, o interesse de aliar as duas técnicas: de um lado a espectroscopia e de outro a calibração multivariada.

O exemplo desta tendência pode ser observado no Centro de Tecnologia da Copersucar (CTC) em Piracicaba, onde a rapidez das análises tem um papel fundamental durante todo o processamento industrial ao que a cana-de-açúcar é submetida, a começar pelo pagamento da matéria prima. O preço pago pela

tonelada de cana-de-açúcar depende do teor de açúcar que este possui; assim, quanto maiores os teores, melhores os preços pagos. Desta forma é importante uma análise rápida e confiável para determinar os índices de açúcares num dado lote.

O controle de qualidade durante o processamento da cana-de-açúcar é também uma etapa crucial à empresa e por esse motivo deve ser executado com eficiência e precisão.

A determinação dos teores de açúcares presentes nas amostras pode ser feita através da construção de um modelo de calibração, em que amostras de caldo de cana-de-açúcar são submetidas às análises padrões via úmida e, posteriormente, analisadas por espectroscopia na região do infravermelho. Métodos quimiométricos como a MLR (*Multiple Linear Regression*)[5], ou PLS (*Partial Least Squares*) [6] tentam estabelecer as possíveis relações lineares entre as quantidades de açúcares e os respectivos espectros adquiridos - valores de absorbância medidos em vários comprimentos de onda. Se o número de variáveis (comprimentos de onda) for reduzido é possível empregar o MLR que permitirá também a utilização de um número reduzido de amostras para construção do modelo de calibração. Caso se utilize o método PLS, ou um caso particular deste – PLS1 [7], para que um modelo seja capaz de determinar, com baixos erros, quantidade de açúcares em novas amostras, é necessário que este seja construído a partir de um grande número de amostras cujas quantidades de analito sejam conhecidas.

Assim, aparentemente é mais vantajosa a utilização do MLR em detrimento ao PLS ou PLS1, mas como os instrumentos modernos têm razoável resolução espectral, um grande número de variáveis é medido para cada espectro, o que implicaria na necessidade de um igual número de amostras, isto é, o método MLR só tem solução quando o número de variáveis for igual ao número de amostras. Assim, em termos práticos, os métodos PLS ou PLS1 são os mais amplamente empregados, pois nenhuma relação entre o número de varáveis e amostras, necessita ser obedecida.

Porém, visando simplificar o modelo, é comum o emprego de métodos de seleção de variáveis. Estes métodos são capazes de diminuir o número de comprimentos de onda possibilitando a utilização do método MLR.

O objetivo principal deste trabalho foi o desenvolvimento e aplicação de um método de seleção de variáveis, que visava a simplificação dos modelos de calibração aplicados a dados da indústria açucareira. Assim, a primeira etapa a ser realizada consistiu da adaptação de um *software* desenvolvido por Magnus Norgaard, da Universidade Técnica da Dinamarca, que ao ser aplicado em Redes Neurais Artificiais era capaz de simplificá-las. O *software* baseia-se no trabalho de Hassibi e Stork sobre o OBS (*Optimal Brain Surgeon*) [8].

Supôs-se que o desempenho deste método, aplicado às Redes Neurais Artificiais, também se observaria para modelos de calibração multivariada construídos através de métodos quimiométricos de ajustes lineares, como o PLS1.

A utilização de dados simulados foi de suma importância durante a verificação do desempenho do método de seleção, e aqui será também apresentado em caráter ilustrativo, a fim de facilitar o entendimento de toda a metodologia de análise empregada em dados espectroscópicos de amostras de cana-de-açúcar fornecidos pelo Centro de Tecnologia da Copersucar de Piracicaba.

A apresentação deste trabalho será dividida da seguinte forma: o primeiro capítulo exibe todos os conceitos teóricos necessários à execução e à compreensão da metodologia proposta. Será apresentada, no segundo capítulo, a análise de dados simulados, oportunidade em que se avalia o desempenho do método desenvolvido. Neste capítulo são feitas algumas considerações importantes tendo, ao final, uma sucinta discussão a respeito dos seus resultados. O ponto principal deste estudo é apresentado e discutido no terceiro capítulo, onde são analisados dados reais compreendidos por espectros de caldo de cana-de-açúcar. A finalização se dá pela apresentação de uma conclusão geral e o índice de referências bibliográficas consultadas. Há, ainda, um Apêndice contendo algumas deduções matemáticas um pouco mais formais.

Os códigos dos programas desenvolvidos para executar a seleção de variáveis não serão anexados, com o objetivo único e exclusivo de reduzir o volume final deste trabalho. Caso o leitor venha a se interessar sobre o assunto e quiser maiores informações, este poderá procurar o autor pessoalmente no Instituto de Química da Unicamp – na sala I-132 – ou através do correio eletrônico slima@iqm.unicamp.br, ou ainda contatar com o Prof. Ronei Jesus Poppi pessoalmente ou ainda no endereço ronei@iqm.unicamp.br.

CAPÍTULO 1

1. CONCEITOS

1.1. MODELOS DE CALIBRAÇÃO – DESCRIÇÃO FORMAL

Modelos de calibração aplicados às análises químicas multicomponentes consistem, em geral, de dois passos: calibração e previsão [6]. Primeiramente, as características de um método são investigadas a fim de se tentar encontrar um modelo fenomenológico que descreva seu comportamento. Um modelo, na verdade, é uma função matemática $Y = f(X)$ que relaciona dois grupos de variáveis, chamadas dependentes Y e independentes X. Esta etapa representa a ***calibração*** ou ***etapa de treinamento***; por isso o conjunto de dados empregados para essa finalidade é chamado ***conjunto de calibração*** ou ***de treinamento***. Os parâmetros do modelo (função) são chamados de coeficientes de regressão (b). Tais parâmetros são matematicamente determinados a partir de dados experimentais [7, 11].

A validação é o passo seguinte à calibração. Nesta etapa as variáveis independentes (X) obtidas, para um outro conjunto de amostras, são usadas, juntamente com os coeficientes de regressão, para calcular os valores previstos para as variáveis dependentes (Y_p). Utiliza-se como conjunto de validação, amostras cujas variáveis Y sejam conhecidas, pois a comparação entre os valores previstos pelo modelo e os conhecidos previamente, permite calcular o desvio entre eles, que serve como parâmetro de avaliação do desempenho deste modelo.

1° Passo

Função linear $Y = f(X)$:

$$Y_c = X_c * b$$

Conjunto de calibração: Y_c e X_c conhecidos inicialmente

Calcula-se matematicamente o valor de b.

2° Passo

Calculado b no 1° passo, determina-se Y_p (previsto)

$Y_p = X_p * b$

conjunto de previsão X_p, coeficientes de regressão b.

Há diversos métodos para a construção dos modelos de calibração. A função *f*, que ajusta as variáveis dependentes e independentes, pode ser linear ou não, dependendo da complexidade do sistema em estudo [12].

Verifica-se que a grande maioria dos métodos de calibração multivariada empregados em espectroscopia utiliza ajustes lineares entre as variáveis, uma vez que estes representam o modelo de mais fácil elaboração e interpretação.

1.2. Regressão Linear Múltipla (MLR – *Multiple Linear Regression*)

1.2.1. Princípios

A regressão linear múltipla estabelece, como o próprio nome diz, uma relação linear entre duas variáveis. A primeira variável pode ser um vetor com *m* medidas de absorbâncias, x_i ($i = 1 \rightarrow m$), que compõem um espectro de infravermelho, e a segunda é um escalar, *y* – quantidade total de açúcares em amostras de caldo de cana. Esta relação pode ser representada matematicamente por:

$$y = x_1b_1 + x_2b_2 + x_3b_3 + x_4b_4 + x_5b_5 + \ldots + x_mb_m + e \qquad (1a)$$

$$y = \sum_{i=1}^{m} b_i x_i + e \qquad (1b) \qquad \text{ou} \qquad y = \mathbf{x}^t\mathbf{b} + e \qquad (1c)$$

Nas equações 1a e 1b, os x_i são as variáveis independentes, *y* a variável dependente, os b_i são os coeficientes de regressão e *e* é o erro residual. Na Equação 1c, *y* é um valor escalar, $\mathbf{b}$ é um vetor coluna e $\mathbf{x}^t$ é um vetor linha.

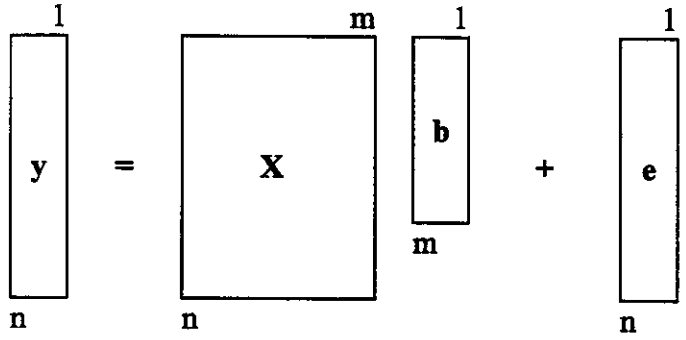
Assim, as equações 1a, 1b e 1c descrevem as dependências multilineares para apenas uma amostra. Tomando-se *n* amostras, os y_j ($j = 1 \rightarrow n$) podem ser escritos como um vetor coluna $\mathbf{y}$, $\mathbf{b}$ mantém-se o mesmo, e o vetor $\mathbf{x}^t_j$ corresponde ao *j*-ésimo espectro da matriz de dados $\mathbf{X}$. Podemos, representar esta situação genérica como segue:

$$\hat{y}_j = \boldsymbol{x}_j^t \quad \boldsymbol{b}_{mx1} + \boldsymbol{e}_j \qquad (2a)$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ . \\ . \\ . \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} x_1^1 b_1 + x_2^1 b_2 + x_3^1 b_3 + x_4^1 b_4 + ... + x_m^1 b_m \\ x_1^2 b_1 + x_2^2 b_2 + x_3^2 b_3 + x_4^2 b_4 + ... + x_m^2 b_m \\ . \\ . \\ . \\ x_1^n b_1 + x_2^n b_2 + x_3^n b_3 + x_4^n b_4 + ... + x_m^n b_m \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e} \qquad (2b)$$

que graficamente pode ser representada por:



Neste caso, *n* é o número de amostras de caldo de cana e *m* o número de variáveis independentes – comprimentos de onda.

O modelo é, então, obtido pela determinação do vetor ***b*** que apresente melhor capacidade de previsão.

Distingue-se três casos possíveis para determinação dos coeficientes de regressão das equações 2a e 2b:

1) Quando m > n, há mais variáveis que amostras.

   Neste caso, há um número infinito de soluções para ***b***.

2) Quando m = n, o número de variáveis é igual ao de amostras.

   Dá uma solução única para ***b***, permitindo-nos escrever:

   $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb} = \boldsymbol{0}$ (3)

   e é o vetor de resíduos, sendo neste caso um vetor de zeros (***0***)

   Esta situação em geral não é encontrada em termos práticos.

3) Quando m < n, há mais amostras que variáveis.

Aqui não é possível uma solução exata para ***b***, no entanto, pode-se chegar a uma solução minimizando-se o vetor de resíduos ***e***, na seguinte equação

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb} \qquad (4)$$

O método mais popular para resolução desta equação é chamado "mínimos quadrados" e é dado por:

$$\boldsymbol{b} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y} \qquad (5)$$

A Equação 5 remete ao problema mais freqüente em MLR: a matriz inversa de $\boldsymbol{X}^t\boldsymbol{X}$ não existirá se houver colinearidade entre os dados, isto é, se o determinante de $\boldsymbol{X}^t\boldsymbol{X}$ for muito próximo de zero.

Para contornar esse problema, em geral sugere-se que se tenha sempre um número de amostras *pelo menos* tão grande quanto o de variáveis. Caso se tenha m > n, é possível efetuar a eliminação ou seleção de algumas variáveis por métodos apropriados, a fim de que o número de amostras torne-se igual ao de variáveis (m = n).

1.3. Mínimos Quadrados Parciais (PLS – *Partial Least Squares*)

1.3.1. Princípios

No método PLS, o bloco ***X*** de variáveis independentes (neste caso os espectros digitalizados) é relacionado com um bloco ***Y*** de variáveis dependentes (quantidades totais de açúcares em amostras de caldo de cana).

Neste método tanto a matriz ***X*** quanto a ***Y*** são representadas pela análise de componentes principais.

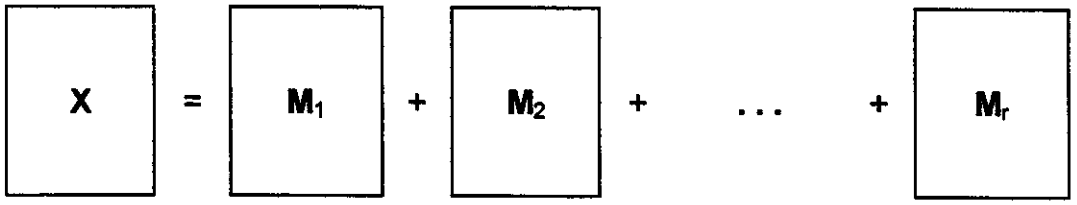
$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} \qquad (6)$$

$$\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F} \qquad (7)$$

A análise de componentes principais é um método na qual se representa uma matriz ***X*** de ordem *r* como a soma de *r* matrizes de ordem 1:

$$\mathbf{X} = \mathbf{M}_1 + \mathbf{M}_2 + \mathbf{M}_3 + \ldots + \mathbf{M}_r \qquad (8a)$$

que graficamente pode ser representado por:

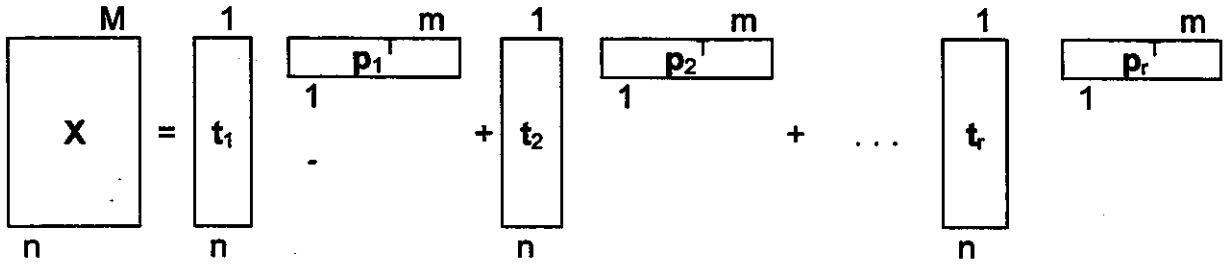


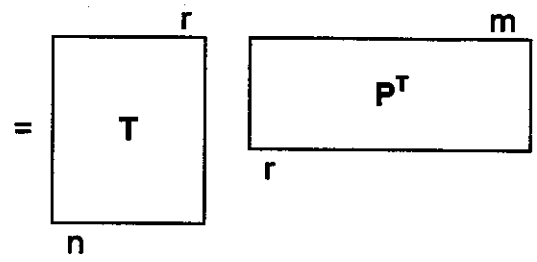
As matrizes $\mathbf{M}_i$ constituem os chamados Componentes Principais e podem ser escritos pelo produto de dois vetores, ***t*** (*scores*) e ***p*** (*loadings*), assim:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \mathbf{t}_3\mathbf{p}_3^T + \ldots + \mathbf{t}_r\mathbf{p}_r^T \qquad (8b)$$

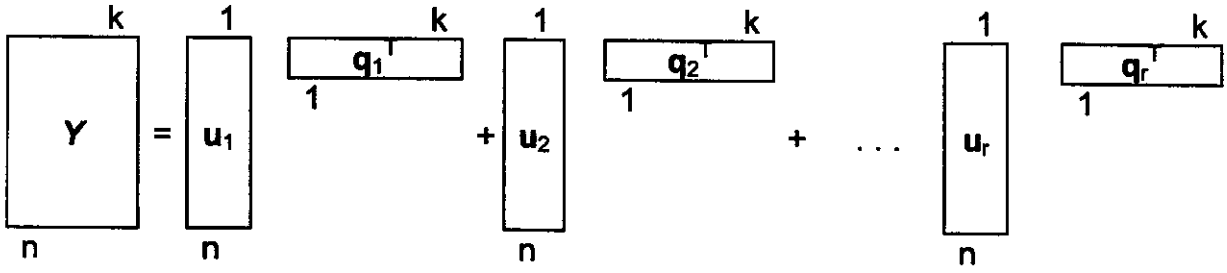
$$\mathbf{X} = \mathbf{T}\mathbf{P}^T \qquad (8c)$$

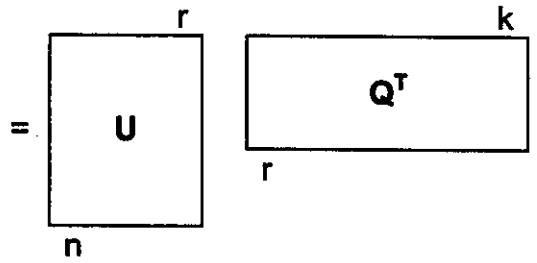
ou graficamente:





e





Assim, uma relação entre os dois blocos (***X*** e ***Y***) pode ser obtida correlacionando, linearmente, os *scores* de cada bloco:

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

onde:

$$\mathbf{b}_h = \mathbf{u}_h \mathbf{t}_h / \mathbf{t}_h^T \mathbf{t}_h \qquad (10)$$

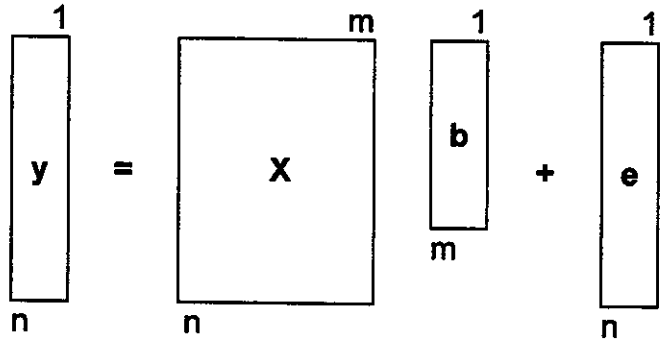
para cada h = 1, 2, ..., r, componentes principais.

O melhor modelo possível é aquele que consegue obter as menores matrizes de resíduo **E** e **F**, ao mesmo tempo que obtém a melhor relação linear entre **t** e **u**.

No PLS isto é alcançado através de uma ligeira mudança nos valores dos *scores*, de forma a produzir a melhor relação possível.

1.4. MÉTODO PLS1

Quando a rotina PLS é executada para um conjunto de amostras (variáveis independentes) e um vetor de variáveis dependentes, recebe o nome de PLS1, com a qual é possível obter o vetor de coeficientes de regressão ***b***, uma vez que o modelo pode ser representado pela Equação 2b, que graficamente é representado por:



O vetor coeficientes de regressão ***b*** é aproximadamente igual aos coeficientes de regressão obtidos pelo MLR (quando existir) [7], e pode ser obtido através do PLS1 do seguinte algoritmo.

Chamando A_{max} do número máximo de fatores a serem computados pelo PLS1, executar os passos de 1 a 6 para cada um dos fatores a = 1,2, ... , A_{max}.

1) Encontar o vetor peso $\hat{\mathbf{w}}_a$ pela maximização da covariância entre a combinação linear $\mathbf{X}_{a-1}\hat{\mathbf{w}}_a$ e **y**, com a condição de que $\hat{\mathbf{w}}_a^t\hat{\mathbf{w}}_a = 1$. Isto corresponde a encontrar o vetor unitário $\hat{\mathbf{w}}_a$ que maximiza $\hat{\mathbf{w}}_a^t\mathbf{X}_{a-1}^t\mathbf{y}_{a-1}$, ou seja, a covariância escalada entre $\mathbf{X}_{a-1}$ e $\mathbf{y}_{a-1}$;
2) Encontrar os *scores*, $\hat{\mathbf{t}}_a$ como a projeção de $\mathbf{X}_{a-1}$ em $\hat{\mathbf{w}}_a$, isto é,

$$\hat{\mathbf{t}}_a = \mathbf{X}_{a-1}\hat{\mathbf{w}}_a; \qquad (11)$$

3) Encontrar os *loadings* $\hat{\mathbf{p}}_a^t = \frac{\mathbf{X}_{a-1}^t \hat{\mathbf{t}}_a}{\hat{\mathbf{t}}_a^t \hat{\mathbf{t}}_a}$ (12)

4) Encontrar $\hat{\mathbf{q}}_a$, onde $q_a = \frac{\hat{\mathbf{y}}_{a-1}^t \hat{\mathbf{t}}_a}{\hat{\mathbf{t}}_a^t \hat{\mathbf{t}}_a}$ (13)

5) Subtrair $\hat{\mathbf{t}}_a \hat{\mathbf{p}}_a^t$ de $\mathbf{X}_{a-1}$ chamando-o de nova matriz $\mathbf{X}_a$;

6) Chamando $\begin{cases} \hat{\mathbf{W}} = \{\hat{\mathbf{w}}_a\} \quad (a = 1,2,...,A) \\ \hat{\mathbf{P}} = \{\hat{\mathbf{p}}_a\} \quad (a = 1,2,...,A) \end{cases}$,

onde A corresponde ao número de fatores ótimos da etapa de validação, calcula-se $\hat{\boldsymbol{b}}$ como $\hat{\boldsymbol{b}} = \hat{\mathbf{W}}(\hat{\mathbf{P}}^t\hat{\mathbf{W}})^{-1}\hat{\mathbf{q}}$, [7] (14)

A vantagem deste método sobre o MLR está no fato de não haver limitação quanto ao número de amostras necessárias para determinação de $\boldsymbol{b}$, uma vez que neste caso o número de variáveis independentes não mais importa, mas sim o número de componentes necessários para descrever satisfatoriamente a variância do sistema.

1.5. Seleção de Variáveis

A redução do número de variáveis (comprimentos de onda) utilizando apenas as que realmente contêm informação relacionada à propriedade de interesse, pode produzir um modelo mais robusto, simples de interpretar e com melhor precisão nas previsões. Assim, em análises espectroscópicas no infravermelho próximo, os comprimentos de onda que apenas induzem ruídos, informações irrelevantes ou não-linearidades, podem ser eliminados. Em algumas publicações têm-se indicado que pode ser possível aumentar o poder de previsão para algumas espécies pela seleção de conjuntos de comprimentos de onda que exibem boa sensibilidade e linearidade para o analito de interesse e não enfatizam características de outras espécies potencialmente interferentes [13, 14]. Neste sentido, têm surgido diferentes métodos para a seleção de comprimentos de onda em análises multivariadas [15-19].

A seleção de variáveis busca encontrar o conjunto de variáveis independentes mais restrito que produza o menor erro de previsão, por isso deve-se monitorar sempre, concomitantemente à seleção das variáveis, um parâmetro regulador representado pela função $E(\boldsymbol{b})$ – Equação (15), que descreve o desvio das previsões em relação aos valores esperados. A eliminação dos coeficientes fica, portanto, vinculada à minimização da função Erro ($E(\boldsymbol{b})$), ou seja, é necessário atingir o mínimo da função erro simultaneamente à eliminação do valor de um determinado coeficiente b_i. $E(\boldsymbol{b}) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ (eq. 15)

onde y_i e $\hat{y}_i$ são, respectivamente, os o valores conhecidos e previsto da i-ésima variável, N é o número de amostras.

1.6. Métodos de Simplificação de Redes Neurais Artificiais

O primeiro método de simplificação das Redes Neurais Artificiais, baseado na análise da Hessiana da matriz de erros foi apresentada por Le Cun, Denker e Solla [9] com o nome de OBD (*Optimal Brain Damage*), em seguida modificações foram propostas por Hassibi e Stork, o que deu origem ao OBS (*Optimal Brain Surgeon*) [8]. As modificações produziram resultados mais precisos, indicando que as Redes Neurais obtidas pelo novo método além de mais simples apresentavam-se mais genéricas, mas também mostrou-se computacionalmente mais dispendioso comparativamente ao método OBD. Apesar deste custo computacional, optou-se pela adaptação do método OBS à seleção de variáveis, aplicado a espectros de caldo de cana-de-açúcar, por ser mais preciso, embora se observe que o desenvolvimento tecnológico tem permitido o desenvolvimento de microprocessadores cada vez mais rápidos, apontando para uma tendência de minimização desse possível efeito negativo do método.

1.7. Rede Neural Artificial – Processamento Básico

O entendimento da adaptação feita no *software* de Norgaard será mais fácil após a compreensão de como se dá a simplificação das Redes Neurais Artificiais.

Em termos práticos podemos considerar uma Rede Neural Artificial como uma "caixa de processamento" que pode ser treinada a partir de um conjunto de dados de entrada (*inputs*) a fim de gerar uma ou mais saídas (*outputs*), conforme representado a Figura 1.

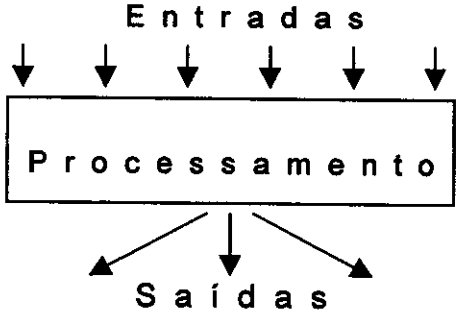


Figura 1: Representação esquemática de como se dá o processamento em Redes Neurais Artificiais

O processo de treinamento de uma rede é a etapa principal na criação de um modelo podendo, para isso, ser utilizado diferentes algoritmos, sendo os mais comuns o *Back Propagation* [20] e Marquardt-Levenberg [21].

Os chamados "Neurônios Artificiais" são as unidades básicas de processamento de informação, projetadas para simular o comportamento de um neurônio biológico. Muito simplificadamente, um neurônio biológico pode ser descrito como um corpo celular contendo dois tipos de ramificações: *dendritos* e *axônios*. Analogamente ao neurônio biológico, o "Neurônio Artificial" possui um corpo de processamento de informação, com duas classes de ramificações: *inputs* (dendritos) e *outputs* (axônios). A rede neural, portanto, consiste de uma coleção de neurônios artificiais que se interligam por suas "ramificações".

Assim, se vários neurônios dispostos em camadas estiverem conectados entre si, estará estabelecido o que se chama por arquitetura de rede. Esquematicamente ter-se-ia a representação esquematizada pela Figura 2.

Aqui uma certa coleção de valores de entrada representada por x_1 até x_6 corresponderá aos *inputs* de uma rede, estando conectados aos neurônios de processamento, N_1 até N_3 – correspondendo ao que se convencionou chamar camada escondida. Mais abaixo estaria a camada de saída, que neste caso é composta apenas pelo neurônio N_4.

O sinal total que entra no corpo de processamento de um neurônio artificial é comumente chamado ***Net***. O valor do *Net* é calculado pelo somatório dos *inputs* previamente multiplicados pelos pesos das "sinapses". O chamado processamento do neurônio consiste em aplicar uma função de transferência ao *Net* enviando o resultado deste processo como argumento do *output*. Genericamente representa-se a saída do neurônio como:

$$out = f(Net) \tag{16}$$

Há diversas funções possíveis, lineares ou não-lineares. Entre as funções de transferência não-lineares, pode-se destacar as duas mais amplamente empregadas: a função de transferência sigmoidal e a tangente hiperbólica. Para o caso da função de transferência linear o argumento do *output* será simplesmente o valor do *Net*, neste caso $out = f(Net) = Net$. Assim:

$$Net = \sum_{i=1}^{m} input_i * b_i \tag{17}$$

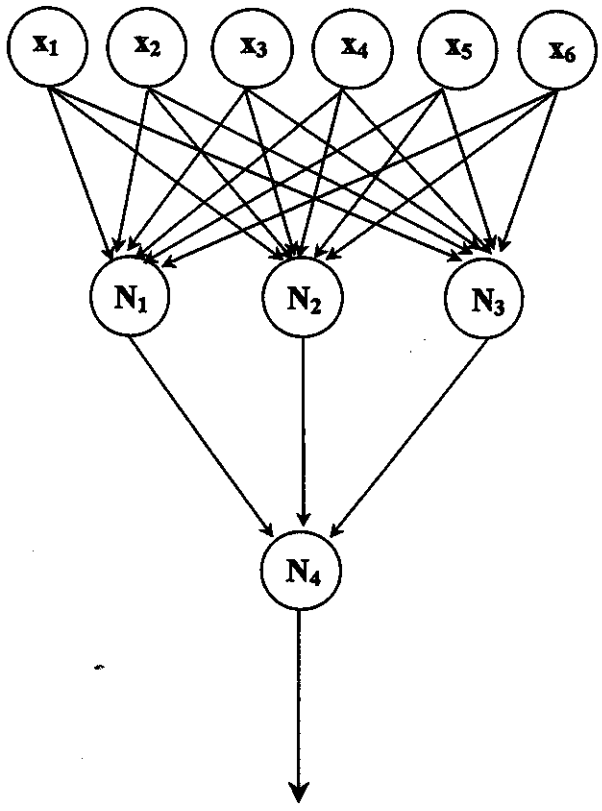


Figura 2: Representação esquemática da arquitetura de uma rede neural.

A Figura 3 ilustra a situação descrita:

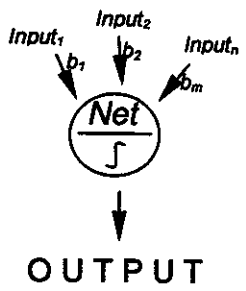


Figura 3: Representação esquemática do processamento de um neurônio artificial

1.8. MÉTODOS DE PODA

O método de simplificação das arquiteturas de Redes Neurais Artificiais se dá pela eliminação de algumas das conexões existentes, selecionando, a partir de uma arquitetura inicial, outra mais simples e talvez com maior capacidade de previsão.

Este método, por analogia direta à sua atuação, é conhecida por *poda*, do inglês "*pruning*". Não há apenas um único método de simplificação através das podas. As diferenças entre os métodos estão ligadas aos critérios de decisão de quais conexões devem ser eliminadas. Alguns se baseiam na Matriz Hessiana de erros [8, 9, 10] e outros se baseiam no *decaimento de pesos* [22] e na *eliminação de pesos* [23]. Porém, todos possuem um critério de eliminação seletiva e ordenada, buscando melhorar o poder de previsão do modelo.

Considera-se um poder de previsão alto, quando a resposta do modelo para um dado conjunto de amostras, aproxima-se muito dos valores reais. Operacionalmente isso ocorre quando se alcança a minimização da função erro, dada pela Equação 15.

A poda, neste caso o OBS, é um método de otimização que encontra o melhor conjunto de conexões, através da avaliação da função erro, neste caso sua matriz Hessiana (derivada segunda). O procedimento básico deste método consiste em:

1 – partindo-se de uma rede já treinada, e com *m* conexões ativas, computar o erro de previsão;

2 – eliminar uma única conexão segundo critérios matemáticos apropriados;

3 – estimar, usando as *m-1* conexões restantes qual o erro de previsão associado ao modelo;

4 – executar os passos 2 e 3, até que reste apenas uma conexão a ser eliminada.

Ao fim do procedimento têm-se *m* modelos, cujos números de conexões variam de 1 a *m*, e *m* diferentes erros de previsão (um valor de erro associado a cada um dos modelos). A arquitetura e, portanto, o conjunto de conexões que produzir o menor erro de previsão dentre todos é escolhida.

1.8.1. *Optimal Brain Surgeon* – Descrição Formal

A idéia básica deste método está no uso das informações contidas na derivada segunda da superfície de erros, de modo a estabelecer um compromisso entre a complexidade da rede e a eficiência do modelo. No caso particular em que um modelo de superfície de erros é construído para fazer previsões analíticas quanto aos efeitos da variação dos pesos sinápticos de uma rede, utiliza-se uma função erro, através do emprego da série de Taylor em torno de um ponto desejado:

$$E(\boldsymbol{b}+\delta\boldsymbol{b}) = E(\boldsymbol{b}) + \nabla E(\boldsymbol{b})\delta\boldsymbol{b} + \frac{HE(\boldsymbol{b})}{2!}(\delta\boldsymbol{b})^2 + \frac{\theta}{3!}(\delta\boldsymbol{b})^3 + \ldots \qquad (18)$$

onde $\boldsymbol{b}$ são os pesos das sinapses, $\delta\boldsymbol{b}$ é a perturbação aplicada, sendo dada por:

$$\delta\boldsymbol{b} = \boldsymbol{b}_{(0)} - \boldsymbol{b}_{(-1)} \qquad (19)$$

$\nabla E(\boldsymbol{b}_{(-1)})$ é o gradiente da função erro, dado por:

$$\nabla E(\boldsymbol{b}) = \frac{\partial E(\boldsymbol{b})}{\partial \boldsymbol{b}} \qquad (20)$$

$\mathbf{H}E(\boldsymbol{b}_{(-1)})$ é a matriz Hessiana:

$$\mathbf{H}E(\boldsymbol{b}) = \frac{\partial^2 E(\boldsymbol{b})}{\partial \boldsymbol{b}^2} \qquad (21)$$

Para se determinar o valor de $\delta\boldsymbol{b}$ que produz o menor aumento no valor da função custo, empregam-se duas aproximações:

1 – Próximo ao mínimo da função erro, podemos aproximá-lo à uma função quadrática, possibilitando ignorar os termos de ordem superior, da série de Taylor;

2 – A eliminação das conexões só se inicia após a convergência do processo de treinamento. A implicação desta suposição é que as conexões devem representar um conjunto de valores correspondentes a um mínimo local ou global da superfície de erro, então, neste caso, o termo gradiente deverá ser igual a zero.

Através destas suposições temos:

$$E(\boldsymbol{b}+\delta\boldsymbol{b}) = E(\boldsymbol{b}) + \frac{\boldsymbol{H}E(\boldsymbol{b})}{2!}(\delta\boldsymbol{b})^2 \qquad (22)$$

Admitindo

$$\Delta E(\boldsymbol{b}) = E(\boldsymbol{b} + \delta\boldsymbol{b}) - E(\mathbf{b}) \qquad (23)$$

então, pode-se escrever a Equação 18 como:

$$\Delta E(\boldsymbol{b}) = \frac{1}{2}\delta\boldsymbol{b}^t \boldsymbol{H}E(\boldsymbol{b})\delta\boldsymbol{b} \qquad (24)$$

A estratégia necessária para determinação de $\delta\boldsymbol{b}$ é apresentada a seguir.

1.8.1.1. CÁLCULO DA SALIÊNCIA

Na poda, deseja-se que uma das conexões (b_i) receba o valor zero, tanto quanto se minimize ΔE, na Equação 24. A eliminação do coeficiente b_i pode ser expressa como:

$$\delta b_i + b_i = 0 \tag{25a}$$

ou mais genericamente

$$\mathbf{1}_i^t \delta \mathbf{b} + b_i = 0 \tag{25b}$$

onde $\mathbf{1}_i^t$ é um vetor cujos elementos são todos zeros, exceto para o *i*-ésimo elemento, que é igual a unidade.

O método dos multiplicadores indeterminados de Lagrange pode ser usado para solucionar o problema da poda restrita à diminuição da função erro:

$$S(\mathbf{b}) = \Delta E(\mathbf{b}) + \lambda(\mathbf{1}_i^t \delta \mathbf{b} + b_i) \tag{26}$$

Tomando-se a derivada de $S(\mathbf{b})$ em relação a $\delta\mathbf{b}$, e fazendo-se o coeficiente b_i receber o valor zero, o $\delta\mathbf{b}$ ótimo é dado por:

$$\delta \mathbf{b} = -\frac{b_i}{[\mathbf{H}^{-1}]_{ii}} \mathbf{H}^{-1} \mathbf{e}_i \tag{27}$$

A chamada saliência, $S_i(\mathbf{b})$, corresponde à solução da Equação 26, otimizada em relação a $\delta\mathbf{b}$, sujeita à restrição de que o *i*-ésimo coeficiente seja eliminado.

$$S_i(\mathbf{b}) = \frac{1}{2} \frac{\tilde{b}_i^2}{[\mathbf{H}^{-1}]_{ii}} \tag{28}$$

onde $\mathbf{H}^{-1}$ é a inversa da matriz Hessiana $\mathbf{H}$, e $[\mathbf{H}^{-1}]_{ii}$ é o *ii*-ésimo elemento desta matriz inversa.

Na verdade, a Saliência S_i representa o aumento no erro quadrático médio (parâmetro que determina eficiência do modelo) resultante da eliminação de b_i. Nota-se que o valor de S_i é proporcional a b_i^2. Assim, em princípio, pequenos valores de b_i teriam pequeno efeito no erro quadrático médio. Porém, pela Equação 28, verifica-se que a Saliência é também inversamente proporcional aos elementos da diagonal da inversa da Matriz Hessiana. Com isso até mesmo pequenos pesos podem ter um efeito substancial no valor do erro quadrático médio.

A dificuldade do procedimento do OBS refere-se ao aspecto computacional e reside na inversão da Matriz Hessiana **H**, visto que, para que sua inversa seja computável, esta não pode ser singular, isto é, seu determinante não pode ser igual a zero. O cálculo da matriz Hessiana inversa $\mathbf{H}^{-1}$, utilizado por Hassibi em seu trabalho deu-se através de conhecimentos de álgebra matricial como a inversão de matriz *lemma*. A descrição formal para a inversão da matriz Hessiana é apresentada no Apêndice A.

1.8.2. Adaptação do OBS à Calibração Multivariada PLS1

A possibilidade de se representar o modelo PLS1 através de uma arquitetura particular de Rede Neural Artificial possibilitou a utilização dos princípio do OBS.

Em termos ilustrativos podemos representar uma Rede Neural Artificial extremamente simplificada, fazendo com que os valores de entrada – *inputs* sejam os valores de absorbâncias medidos em diversos comprimentos de onda, na região do infravermelho próximo, e a saída – *output* seja a quantidade total de açúcar de dada amostra. Esta rede é composta por um único neurônio na camada de saída, sem camadas intermediárias, e com função de transferência linear, portanto, neste modelo a saída da rede será dada simplesmente pela Equação 17.

Esquematicamente têm-se:

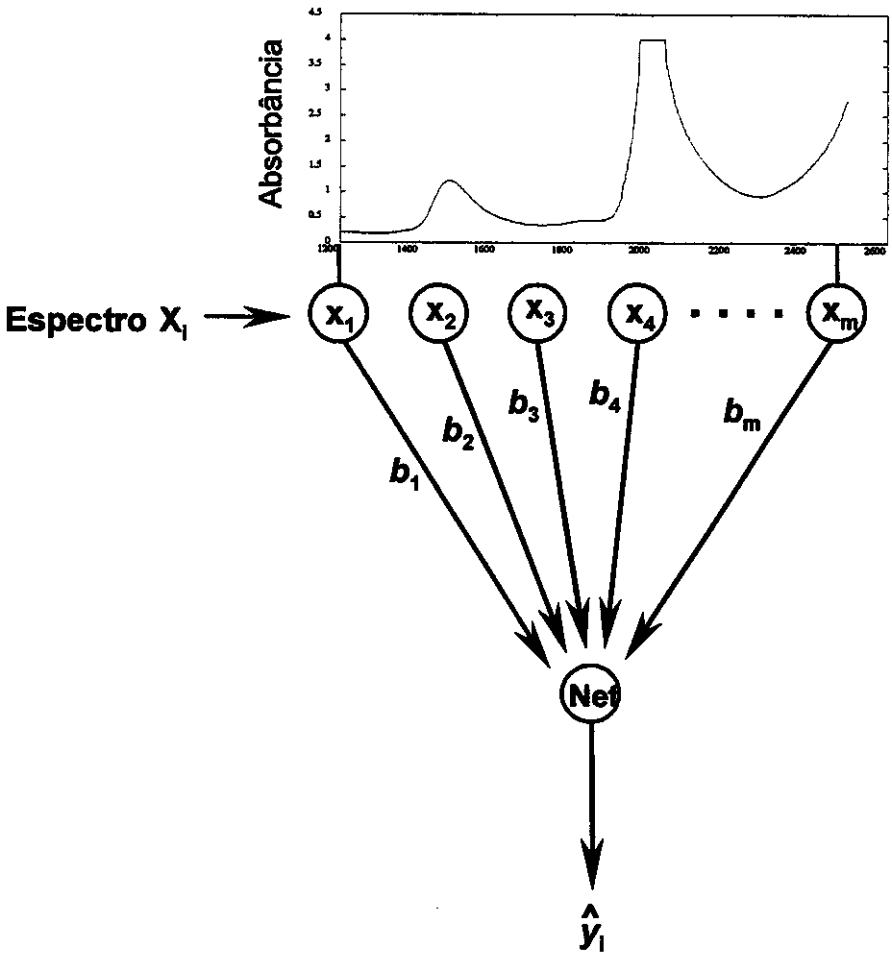


Figura 4: Representação de uma regressão linear multicomponente descrita em "arquitetura" de redes neurais

A inicialização do *software* OBS dá-se pela definição da arquitetura e dos valores dos pesos das sinapses correspondentes. Como o OBS pode ser aplicado à qualquer arquitetura possível, utilizou-se a arquitetura representada na Figura 4, sendo que as sinapses eram compostas pelo valores dos coeficientes de regressão calculados a partir do método PLS1.

A exigência imposta pelo método OBS é a de que a rede inicialmente empregada devesse estar totalmente treinada, isto é, teria que ter alcançado um ponto de mínimo na superfície de erros de previsão. Esta restrição foi observada durante a construção do modelo PLS1 que, através do número de variáveis latentes empregadas durante a execução deste método [24], garantia a obediência às exigências do método OBS, validando, assim, a adaptação proposta.

O ponto central da adaptação está no processamento executado após a eliminação de uma determinada conexão. O *software* de Norgaard originalmente utiliza o método de Marquardt-Levenberg para determinar novos valores para as conexões que são mantidas – esta etapa é conhecida por retreinamento da rede. O método proposto ao invés disso utiliza o método PLS1 para estabelecer os novos coeficientes de regressão, e estes são utilizados como sendo os novos pesos das sinapses. A Tabela I, abaixo, ajuda o entendimento desta diferença.

Ao fim do procedimento têm-se *m* modelos, cujos números de variáveis variam de 1 a *m*, e *m* diferentes erros de previsão (um valor de erro associado a cada modelo). Por isso, concomitantemente à eliminação das conexões são observados os valores dos erros de previsão, possibilitando, com isso, identificar o conjunto de coeficientes de regressão responsáveis pelo modelo mais preciso.

Tabela I: Comparação entre os algoritmos básicos dos métodos OBS e Poda-PLS1

Algoritmo básico do OBS	**Algoritmo básico da Poda-PLS1**
1 – a partir de uma rede totalmente treinada, e com *m* conexões ativas, computar o erro de previsão deste modelo.	1 - estimar, inicialmente, os coeficientes de regressão para *m* variáveis dependentes pelo método PLS1 e computar erro de previsão deste modelo
2 – eliminar uma única conexão segundo a análise da ***saliência***, que avalia a variação da função erro provocada pela exclusão deste parâmetro.	2 – eliminar um único coeficiente de regressão segundo a análise da ***saliência***, que avalia a variação da função erro provocada pela exclusão deste parâmetro.
3 – estimar, para as *m-1* variáveis restantes os novos valores dos pesos das sinapses pelo método de Marquardt-Levenberg e computar, também para este caso, qual o erro de previsão associado ao novo modelo construído.	3 – estimar, para as *m-1* variáveis restantes os novos coeficientes de regressão pelo método PLS1 e computar, também para este caso, qual o erro de previsão associado ao novo modelo construído.
4 – executar os passos 2 e 3, até que reste apenas uma conexões a ser eliminada.	4 – executar os passos 2 e 3, até que reste apenas um coeficiente de regressão a ser eliminado.

CAPÍTULO 2

2. Dados Simulados

Uma vez apresentada a modificação ao método OBS, utilizou-se dados simulados para validá-lo. Os dados simulados servirão, também, para ilustrar com maior clareza todos os passos executados até o resultado final das análises.

Como o método proposto visa selecionar as variáveis mais importantes para a calibração num conjunto de dados, decidiu-se simular um conjunto de espectros com as seguintes características:

- Número de Espectros Simulados: 1200;
- Número de Variáveis: 70;
- Número de Bandas em cada espectro: 3;
  - Banda I:
    Largura máxima: 12,5 unidades arbitrárias (U.A.);
    Variável com o Valor de Máximo: 23,5;
    Amplitude: Constante em 0,3 unidades arbitrárias (U.A.).
  - Banda II:
    Largura máxima: 10 unidades arbitrárias (U.A.);
    Variável com o Valor de Máximo: 33,5;
    Amplitude: 0,11– 0,18 unidades arbitrárias (U.A.) distribuída segundo uma variação aleatória
  - Banda III:
    Largura máxima: 15 unidades arbitrárias (U.A.)
    Variável com o Valor de Máximo: 50;
    Amplitude: 0,2 – 0,3 unidades arbitrárias (U.A.) distribuída segundo uma variação linear eqüidistante.

Baseando-se na teoria de análises espectrofotométricas, utilizou-se os princípios da Lei de Beer, que estabelece uma relação linear entre o valor de absorbância, medido para uma determinada espécie química, e sua concentração. Semelhantemente, decidiu-se estabelecer, para a terceira banda, uma relação

linear entre os valores correspondentes às absorbâncias e os valores correspondentes às concentrações, que neste caso passarão a ser chamadas de "quantidades de analito simuladas", ou simplesmente "quantidades simuladas".

As três bandas foram geradas a partir de funções gaussianas, tendo sido acrescentado erros aleatórios com distribuição normal (μ = 0,000 U.A., σ = 0,005 U.A.) simulando, simplificadamente, flutuações características em espectroscopia[25].

A Figura 5 mostra parte da coleção de espectros simulados.

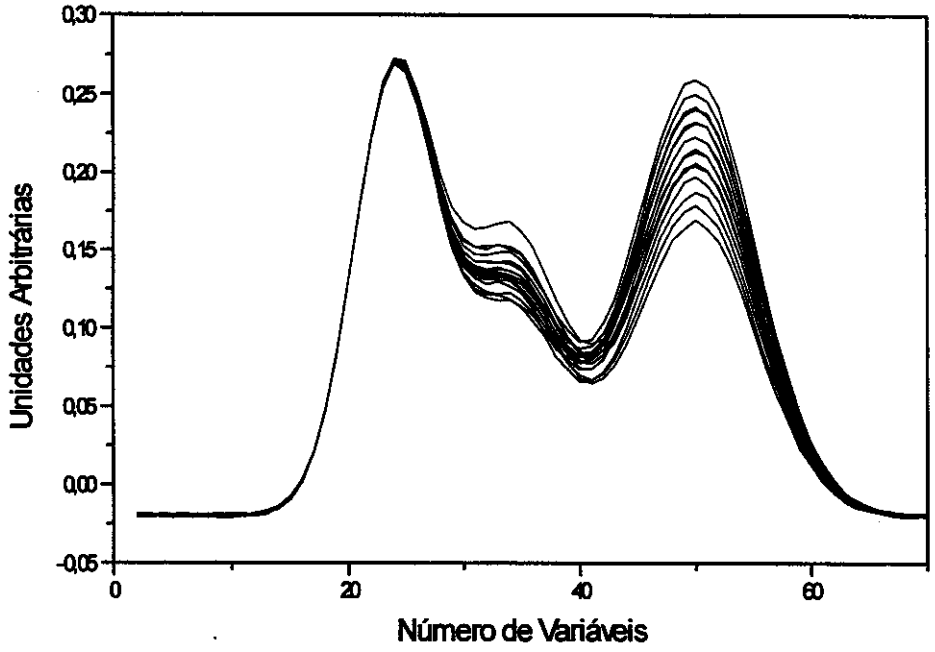


Figura 5: Representação dos perfis dos espectros simulados. Cada composto possui três bandas pouco resolvidas.

Verifica-se que a segunda banda não apresenta boa resolução, sobrepondo-se tanto à primeira (à esquerda) quanto à terceira (à direita).

A utilização de dados simulados com tais características permite validar o método de seleção de variáveis proposto, pois como este método seleciona as variáveis mais importantes à construção do modelo de calibração, espera-se que apenas as variáveis que apresentam uma dependência linear com as quantidades

simuladas de uma suposta substância devam ser selecionadas, que neste caso correspondem àquelas que compõe a terceira banda dos espectros.

As variáveis da primeira banda sendo constantes, e as variáveis da segunda variando aleatoriamente, não devem mostrar-se importantes para a construção do modelo de calibração e, por isso, não devem, a princípio, ser selecionadas.

2.1. Pré-Tratamento dos Dados

Os dados simulados não foram submetidos a nenhum pré-tratamento, visto que correspondem a modelos matemáticos pré-estabelecidos. O ruído adicionado aos dados foi mantido durante toda a análise. Ficou também descartada a possibilidade de haver alguma amostra anômala (*outlier*). Optou-se, também, não centrar os dados na média, pois verificou-se que os resultados centrando na média ou não, apresentavam baixíssimos desvios.

2.2. Distribuição das Amostras

Os 1200 espectros simulados foram divididos em três conjuntos: modelagem (600 espectros); previsão (300 espectros) e teste (300 espectros).

Devido ao grande número de amostras, a distribuição deu-se de forma aleatória e mesmo assim foi possível que todos os diferentes perfis de espectro estivessem presentes em cada um dos conjuntos.

A Figura 6 representa 100 dos espectros que compõem o conjunto de modelagem (linhas pretas), 11 dos espectros que compõem o conjunto de previsão (linhas vermelhas) e 11 dos espectros que compõem o conjunto teste (linhas azuis). Pode-se verificar que os perfis dos espectros do conjunto de modelagem abrangem uma faixa mais larga, fazendo com que os espectros dos conjuntos de previsão e teste estejam contidas nela. Decidiu-se não utilizar para

elaboração da Figura 6 todas as amostras de modelagem, previsão e teste, visto que a clareza na visualização ficaria prejudicada.

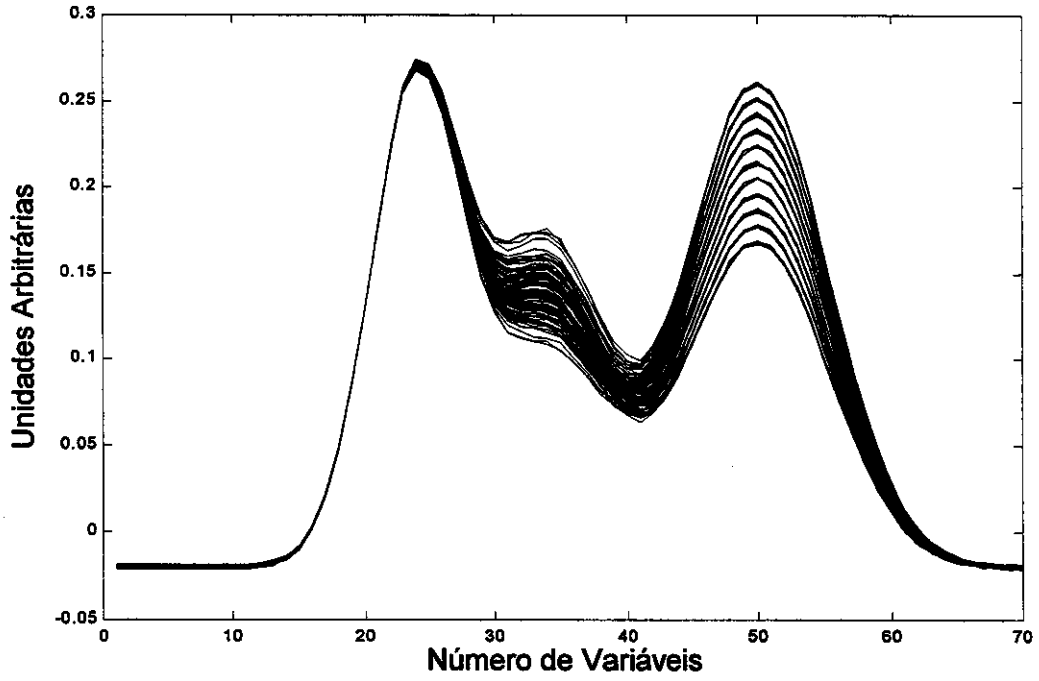


Figura 6: Distribuição dos espectros em três conjuntos: preto – modelagem; vermelho – previsão e azul - teste

A Figura 7 mostra as quantidades simuladas associadas a cada amostra. Os símbolos * representam as amostras do modelo, os △ indicam as amostras proveniente do conjunto de previsão e os □do conjunto teste. Pode-se verificar que as "quantidades" correspondentes às amostras de modelagem distribuem-se homogeneamente dentro da faixa estipulada na simulação. Observando-se as quantidades de analito simuladas, referentes às amostras de previsão e teste, percebe-se que para cada uma delas há um número suficiente de amostras de modelagem que assumem o mesmo valor, favorecendo, assim, a elaboração de um modelo de calibração mais preciso.

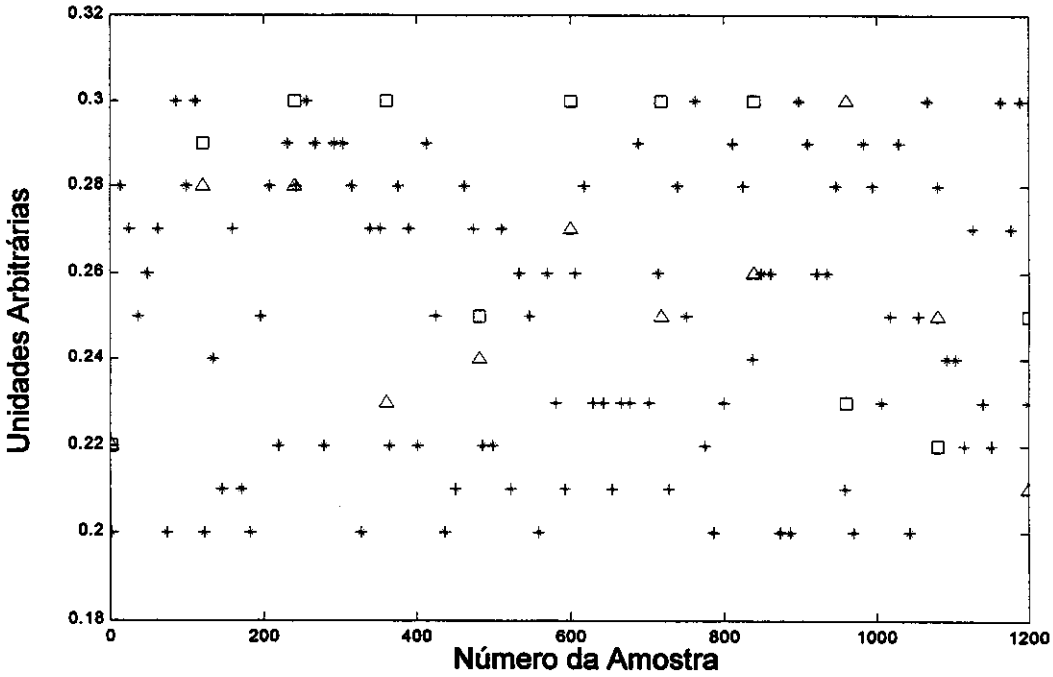


Figura 7: Distribuição das quantidades de analito simuladas associadas aos espectros simulados, sendo: ✳ – modelagem; △ – previsão e □ – teste.

2.3. Construção do Modelo

Após concluir a etapa de organização das amostras partiu-se para elaboração dos modelos de calibração através do PLS1. Vale lembrar que este método é capaz de encontrar um vetor de coeficientes ***b*** que relaciona o conjunto de variáveis independentes (espectros gerados) com um vetor de variáveis dependentes ("quantidades" simuladas). O conjunto de previsão foi utilizado para realizar a validação do modelo.

A Figura 8 mostra o gráfico da soma dos quadrados dos erros de previsão (*Predictor Error Sum Squares - PRESS*) contra o número de variáveis latentes obtido para este conjunto de dados. Observa-se que o erro de previsão praticamente não se altera quando são utilizados de duas a sete variáveis latentes para construção do modelo de calibração. Assim, para evitar sobreajuste do modelo a estes dados [24], utilizou-se três variáveis latentes, o que corresponderia a um erro de previsão igual a $3{,}00.10^{-4}$ U.A. (Figura 8). Os valores dos erros de

previsão são muito pequenos quando se trata de dados simulados, visto que as variáveis dependentes apresentam uma relação linear extremamente alta com as variáveis independentes, assim, qualquer método de ajuste linear é capaz de gerar modelos com baixíssimos erros de previsão e altíssima correlação.

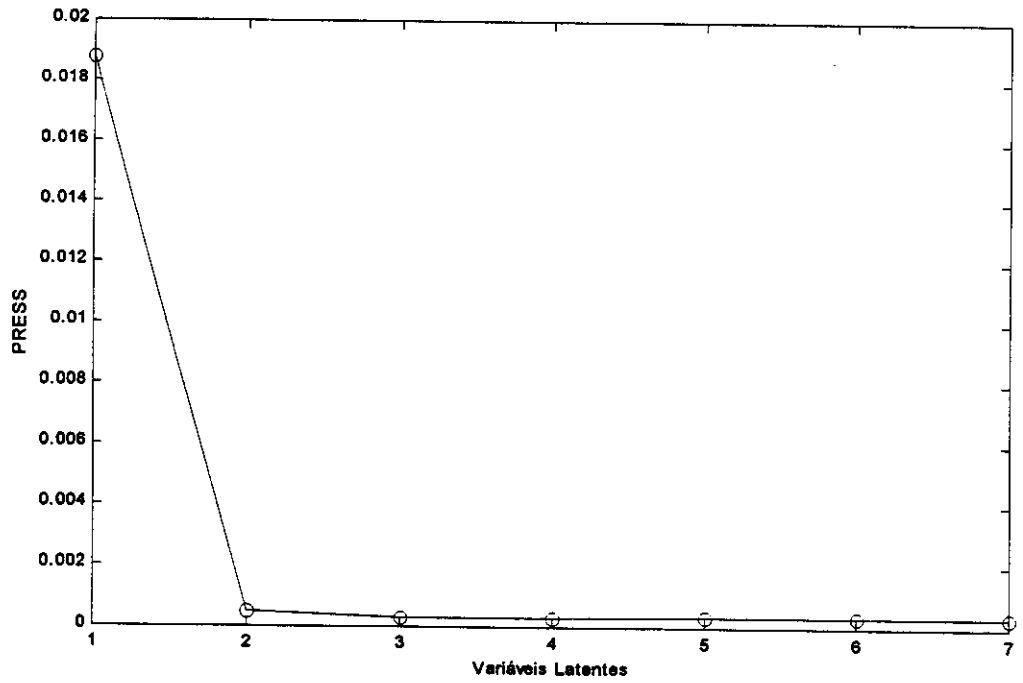


Figura 8: Erros de previsão *vs* Número de Variáveis Latentes

2.4. Aplicação do Método da Poda-PLS1

A seleção das variáveis importantes ao modelo de calibração foi feita submetendo-se o vetor de coeficientes ***b***, encontrados pelo método PLS1, à poda segundo os passos descritos no Capítulo 1, seção 1.8.

A Figura 9 mostra os erros de previsão determinados durante o processamento da poda. Como a escolha do número de variáveis importantes é feita automaticamente a partir da análise dos erros de previsão, tem-se para este caso, que o modelo que numericamente apresentou menor erro era composto por 31 variáveis.

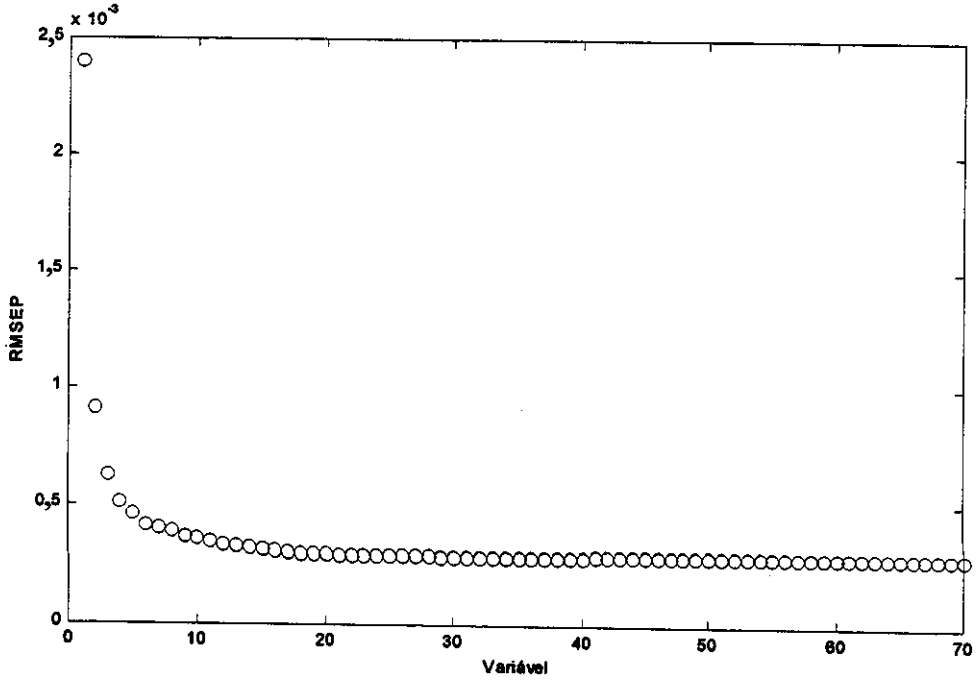


Figura 9: Erro de Previsão *vs* Número de Variáveis do Modelo ao executar a poda

A Figura 10 mostra as variáveis selecionadas para este modelo. Por se tratar de uma figura ilustrativa, está sendo representado apenas um único espectro, porém em todos, as variáveis mostradas serão as selecionadas.

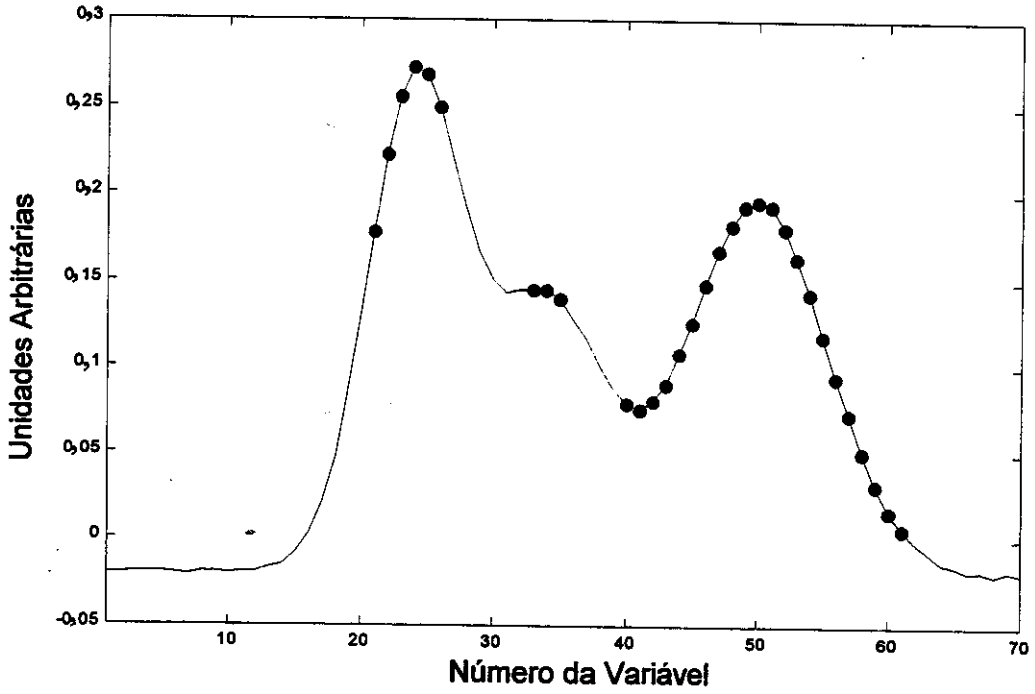


Figura 10: Variáveis selecionadas para o primeiro modelo construído

As Figuras 9 e 10 possibilitam algumas observações importantes. Em primeiro lugar, observa-se que visualmente não é fácil identificar qual é o número de variáveis que produz o modelo com menor erro de previsão, pois para uma faixa de 15 até 70 variáveis, o erro de previsão não se altera significativamente. Em conseqüência desta primeira observação, pode-se verificar na Figura 10 que algumas variáveis que compõem a primeira e a segunda bandas foram selecionadas neste caso. Intuitivamente esperava-se que apenas as variáveis da terceira banda fossem selecionadas. Isso pode ser explicado se levarmos em consideração que há uma rigidez bastante acentuada durante o processo de identificação do conjunto de variáveis que produz o modelo mais preciso. Nesta etapa, empregou-se como critério os valores numéricos dos erros associados aos números de variáveis. Foram selecionadas as variáveis que produziram o menor erro absoluto de previsão, ao invés de um conjunto com menor número de elementos, cujos erros de previsão fossem pouco maiores que este mínimo absoluto. Conclui-se, daí, que um modelo mais flexível selecionaria um conjunto de variáveis menor, que provavelmente corresponderia apenas às variáveis que compunham a terceira banda.

Surge, portanto, um problema a ser resolvido: como tornar o método mais flexível? Inicialmente poder-se-ia implementar um parâmetro que especificasse um percentual de tolerância para a variação do erro, isto é, se um conjunto de variáveis, com o menor número de elementos, apresentar um erro de previsão ligeiramente maior que o mínimo absoluto. Então, este poderia representar o melhor conjunto de variáveis a ser selecionado.

2.4.1. Flexibilização do Modelo

Neste sentido decidiu-se flexibilizar o modelo possibilitando uma tolerância de 5%. Os resultados obtidos podem ser observados nas Figuras 11 e 12. A Figura 12 corresponde à faixa de variáveis de 15 a 70 da Figura 9, com as escalas das ordenadas adequadamente ajustadas, a fim de facilitar a identificação do número de variáveis que produzem o erro absoluto mínimo.

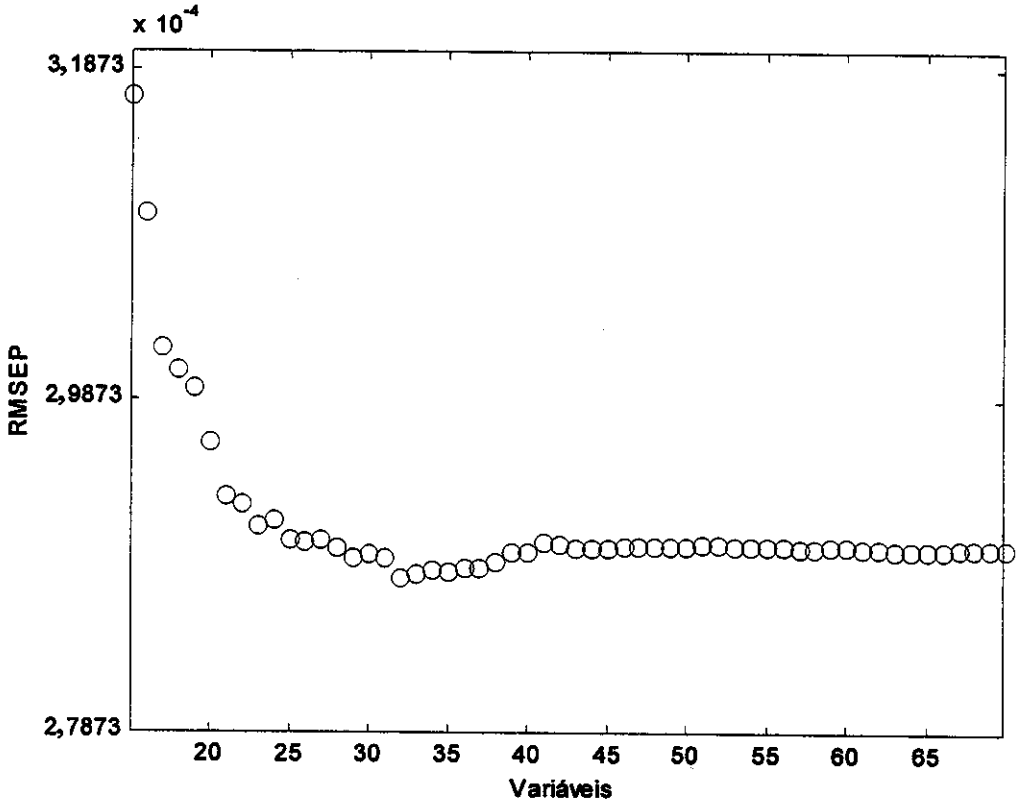


Figura 11: Ampliação da região da Figura 9 correspondente às variáveis 15 a 70

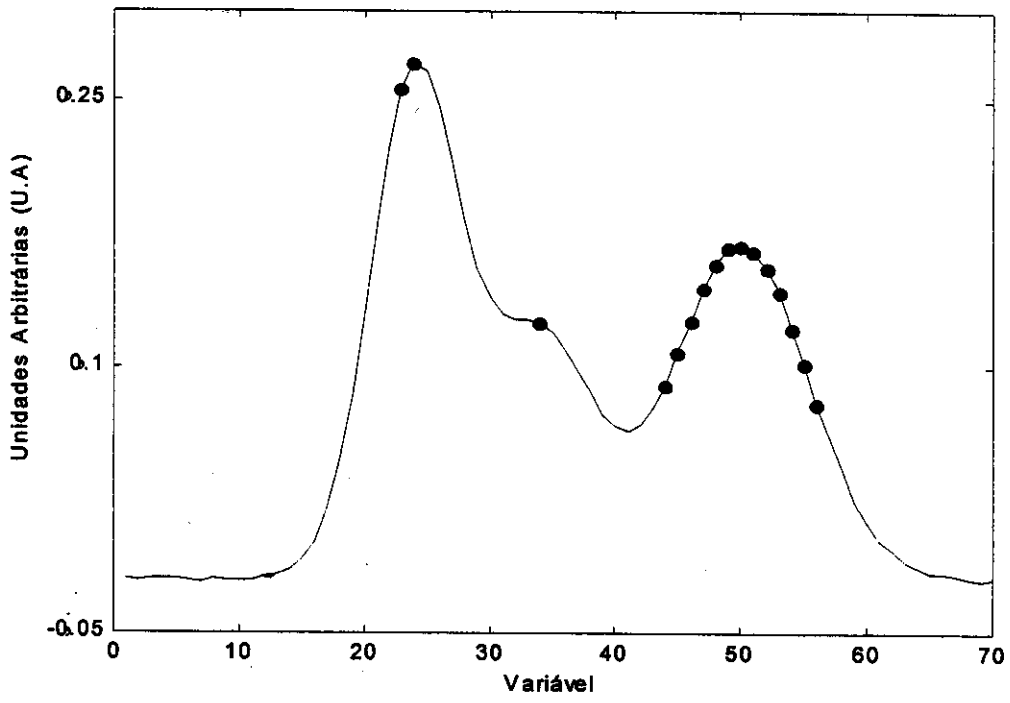


Figura 12: Variáveis selecionadas após flexibilização do primeiro modelo

A partir do estabelecimento do limite de tolerância de 5% para o erro de previsão, observou-se que o número de variáveis selecionadas usando este novo critério passou de 31 para 16, representando uma redução de quase 50%, isto é, apesar da redução de quase metade das variáveis o erro de previsão não sofreu aumento maior que 5% comparado com aquele obtido para as 31 variáveis. As variáveis selecionadas utilizando este artifício são mostradas na Figura 13.

2.4.2. Um Novo modelo

Uma vez que se observou que pequenas variações nos valores dos erros podiam provocar drásticas mudanças no número de variáveis selecionadas, imaginou-se que os resultados obtidos para este caso não deveriam ser genéricos, mas específicos para este caso. Sendo assim as mesmas 1200 amostras simuladas originalmente foram reorganizadas em ordem aleatória, a fim de gerar conjuntos de modelagem, previsão e teste, distintos dos primeiros. Este "novo" conjunto foi submetido às mesmas análises que o anterior, com o intuito de se verificar quais variáveis seriam selecionadas, possibilitando checar se a seleção de variáveis em conjuntos de dados distintos pode gerar resultados diferentes.

Os resultados obtidos para este caso podem ser verificados pelas Figuras 13 e 14. Aqui verificou-se duas situações: na primeira, o critério de seleção de variáveis foi o erro absoluto mínimo (Figura 13); na segunda, utilizou-se como critério de seleção de variáveis o erro absoluto mínimo possibilitando um limite de tolerância para variação deste valor de 5% (Figura 14).

Observa-se que a partir de um mesmo conjunto de amostras (1200 espectros simulados) houve, para o caso dos valores absolutos, uma intensa variação no número de variáveis selecionadas (cerca de 45%), isto é, houve uma variação de 31 – Figura 10 – variáveis para 57! (Figura 13). No caso em que o modelo era mais flexível, esse número variou de 16 (Figura 12) para 17 variáveis selecionadas (Figura 14).

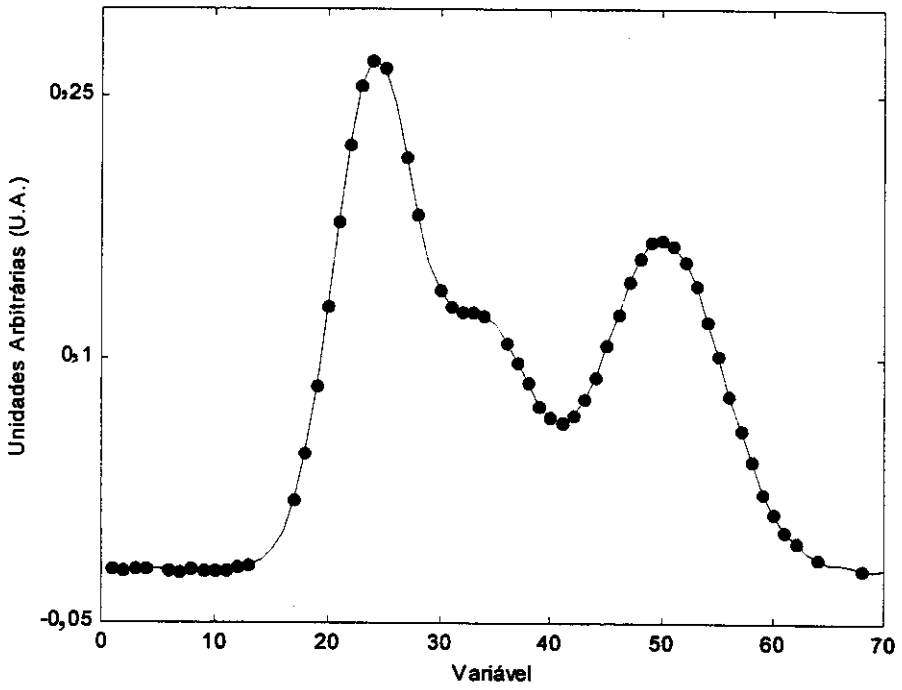


Figura 13: Variáveis selecionadas para o segundo modelo (valores absolutos)

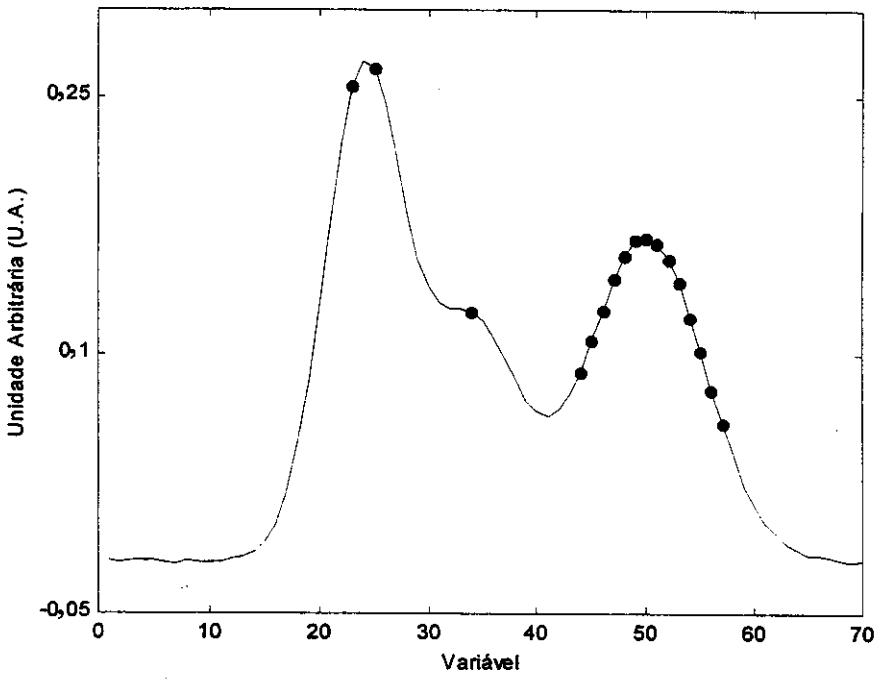


Figura 14: Variáveis selecionadas após flexibilização do segundo modelo

2.4.3. Construção de um Modelo Genérico

Como garantir a escolha de um conjunto de previsão que produza um modelo genérico?

Como o método *poda*-PLS1 seleciona as variáveis que minimizam os erros, espera-se que estas sejam as variáveis que concentrem a maior parte das informações das amostras. Conclui-se daí que se vários conjuntos de amostras de uma mesma espécie forem submetidas à "poda", certamente as variáveis realmente significativas para descrevê-las serão selecionadas na grande maioria dos casos. Tais variáveis terão, portanto, as maiores incidências quando se efetua a interseção dos conjuntos selecionados para cada modelo. Essa idéia é ilustrada na Figura 15.

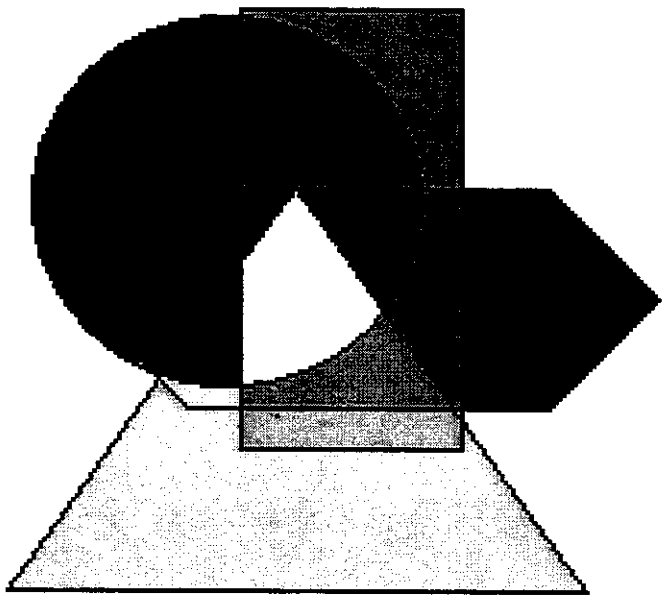


Figura 15: Esquema representativo da interseção de conjuntos – região em branco

A identificação deste conjunto interseção possibilita a construção de modelos mais abrangentes pois não estão associados a um único conjunto de amostras, mas a uma gama maior deles.

Este foi, portanto, o artifício empregado para garantir o poder de generalização, mas qual sua validade? Os modelos construídos são, de fato, genéricos ou SE restringem às amostras utilizadas?

Bem, efetivamente está-se construindo modelos equivalentes, pois o conjunto de amostras empregado para a calibração é sempre tirado da mesma população. Certamente essa equivalência restringe os modelos àquela população. Mas o que ocorreria se tal população representasse satisfatoriamente todo o universo de ocorrência? Seria garantida a generalidade dos modelos obtidos, não a uma dada população, mas a todo o universo de amostras possíveis.

A validação do artifício empregado pode ser verificada ao levar-se em conta amostras provenientes de fontes naturais (como por exemplo a cana-de-açúcar). Isso implica que as quantidades de analito possíveis de serem encontradas na natureza ficam restritas a uma faixa estreita (o universo de ocorrência é estreito – valores muito baixos ou muito altos nunca ocorrem). Assim, se um bom número de amostras for utilizado em cada conjunto, estas devem representar uma boa estimativa do universo possível.

Assim, dado seu poder de generalização, este procedimento passou a ser usado para estabelecer qualquer modelo que tenha sido simplificado através da seleção de variáveis pelo método da poda-PLS1.

2.5. Organização das Amostras

As amostras originais simuladas foram reorganizadas aleatoriamente em 40 diferentes conjuntos, gerando 40 conjuntos distintos de modelagem, 40 de previsão e 40 de teste. Dada a grande quantidade de amostras (1200) garante-se que a aleatoriedade da organização não produz um comprometimento significativos das respostas a serem adquiridas.

2.6. Características dos Modelos

Os modelos construídos em todos os casos correspondem aos vetores de regressão calculados segundo o PLS1. O número de variáveis latentes empregado na construção de cada um dos modelos era sempre o que produzia o menor erro de previsão para seus respectivos conjuntos de validação [24].

A avaliação do desempenho dos modelos, em prever novas amostras, deu-se com o emprego dos conjuntos teste, pois como já explicado, as amostras não participando da construção dos modelos, produzem erros de previsão que representam melhor a eficiência dos modelos construídos.

Tendo em mente que as variáveis selecionadas com maior incidência são as mais significativas para o modelo, determinou-se o percentual de incidência de cada uma das variáveis após proceder à seleção pelo método da poda para cada modelo.

Intuitivamente esperou-se que as variáveis importantes ao modelo tivessem incidência de 100%, isto é, a interseção das variáveis selecionadas para cada um dos 40 conjuntos de previsão, indicaria que as variáveis importantes sempre seriam selecionadas.

Neste ponto é importante ter em mente que o conjunto interseção irá restringir ainda mais o número de variáveis que o modelo de seleção já executou, isto é, o método da poda já colecionou um número reduzido de variáveis que se julgou importante ao modelo. Assim, quando se determina o conjunto interseção destas variáveis, tal número acaba sendo ainda menor. Ocorrendo isto, deve-se ter consciência que sua aplicação ficará restrita aos casos em que não haja um comprometimento significativo do poder de previsão obtido.

Tendo-se em mente que uma redução no número de variáveis já selecionadas, tidas como ótimas, pode comprometer o poder de previsão, deve-se tomar muito cuidado com a análise que permite um limite de tolerância e ainda é submetida à determinação do conjunto interseção, pois pode ser uma tratamento de dados muito drástico.

Comparando-se, portanto, os resultados obtidos quando o conjunto de interseção é determinado nos casos em que a análise foi mais rígida (erro absoluto mínimo – Figura 16) ou mais flexível (erros com tolerância de 5% – Figura 17), pode-se observar algumas tendências.

Inicialmente verifica-se que as 19 variáveis selecionadas, aplicando interseção de conjuntos, no caso dos modelos mais rígidos – Figura 16 – são aproximadamente iguais às variáveis selecionadas pelos modelos flexíveis sem aplicar interseção de conjuntos – Figuras 12 e 14, indicando que aplicar o método da interseção de conjuntos em modelos rígidos é aproximadamente equivalente a flexibilizá-lo pelo método do limite de tolerância de 5%.

Vale aqui, chamar a atenção de que esta observação está sendo feita para um caso particular de dados simulados bem comportados, assim, a equivalência aqui verificada pode não ser válida em análises de dados reais, visto que a complexidade do sistema pode não gerar gráficos de erros *vs* número de variáveis com uma tendência tão bem definida como a vista na Figura 9.

A Figura 17 mostra o caso em que as variáveis selecionadas são provenientes de modelos mais flexíveis e que foram submetidos ao método da interseção de conjuntos, onde se pode observar que 11 variáveis selecionadas foram selecionadas. Pode-se observar que tais variáveis encontram-se na região que compõe apenas a terceira banda, como era esperado. Porém, é necessário saber se este modelo ainda é equivalente aos anteriores. Isto deverá ser feito através de análise estatística.

Os resultados obtidos através destes modelos foram também comparados aos resultados obtidos pelos modelos em que todas as variáveis foram consideradas relevantes, isto é, antes da execução das podas. Antes, porém, da discussão desta etapa, será apresentado o parâmetro de comparação.

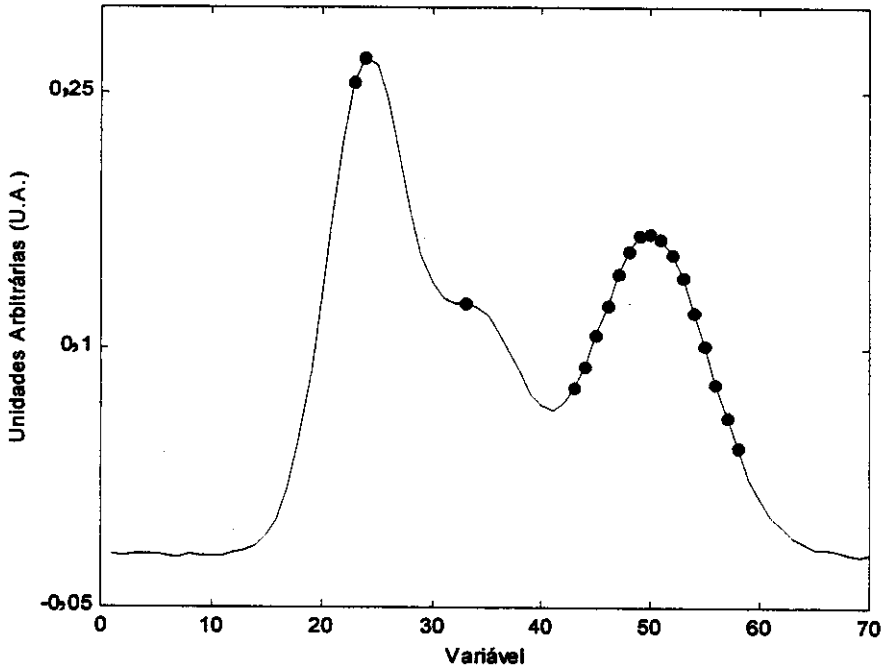


Figura 16: Variáveis selecionadas pelo método da interseção de conjuntos, aplicados a um modelo rígido – erro de previsão absoluto.

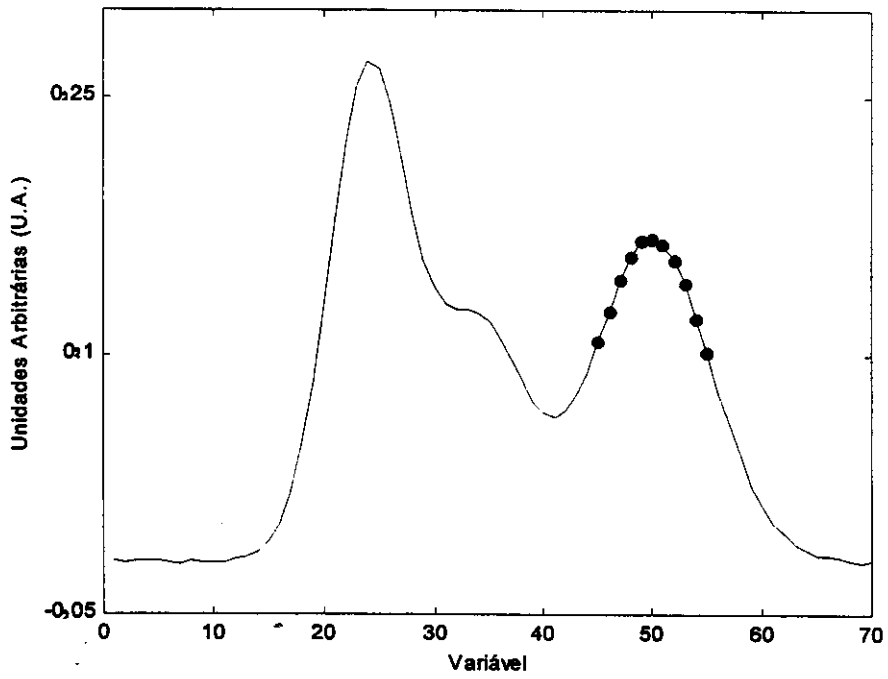


Figura 17: Variáveis selecionadas pelo método da interseção de conjuntos, aplicados a um modelo flexível – erro de previsão com tolerância de 5%

2.7. Avaliação das Respostas

A capacidade de previsão dos modelos pode ser estimada avaliando-se o SEP (Erro Padrão de Previsão) calculado segundo a seguinte fórmula:

$$SEP = \frac{\sum_{I}^{N}(y_{I\,real} - y_{I\,previsto})^2}{N} \qquad (29)$$

onde $y_{i\ real}$ corresponde ao i-ésimo valor simulado e $y_{i\ previsto}$ corresponde ao valor previsto pelo modelo, para a i-ésima amostra. N indica o número de amostras que compõem o conjunto avaliado.

O modelo será tanto melhor quanto menor o desvio entre os valores previstos e esperados (neste caso valores simulados), ou seja, quanto menor o valor SEP assumido para o conjunto teste.

2.8. Análise Estatística

A estimativa da variância amostral para um conjunto composto por um número superior a 30 objetos, pode ser dada por [26]:

$$s^2 = \frac{\sum_{I}^{N}(X_I - \overline{X})^2}{N} \qquad (30)$$

onde X_i corresponde ao i-ésimo termo do conjunto amostral, $\overline{X}$ corresponde à média aritmética dos *N* objetos do conjunto. Desta forma é possível estimar a

dispersão dos valores X_i em relação à média. Semelhantemente, pode-se dizer que a equação 29 também estima dispersões. Porém, essa é a dispersão dos valores $y_{iprevistos}$ em relação aos valores y_{ireal} correspondentes. Em outras palavras, têm-se tanto a equação 29 como a equação 30 descrevendo a mesma propriedade – a dispersão relativa de um conjunto de objetos que compõe uma amostra da população.

A comparação entre duas variâncias pode ser feita estatisticamente através do teste-*F*. Assim, para comparar-se os SEP's das duas classes de modelos, é necessário, por extensão, que se submeta ao mesmo teste estatístico.

Surge aqui um empecilho. Dispõem-se de quarenta modelos de calibração para cada classe de modelos (antes e depois da poda). Cada modelo produz seus correspondentes erros de previsão para os conjuntos teste, ou seja, deve-se comparar com o método tradicional os 40 erros de teste. Como compará-los todos?

A idéia empregada aqui simplifica a situação. Foram empregadas as mesmas variáveis para construção de cada um dos 40 modelos. Podemos dizer, portanto, que estes modelos se equivalem, conseqüentemente os erros de previsão gerados por esses modelos também deverão ser equivalentes entre si. Por esse raciocínio agruparam-se os erros de previsão de cada um dos vinte conjuntos de validação e teste, de modo a obter vetores com 12.000 elementos cada (40 conjuntos x 300 erros de validação) – nota-se que, neste caso, cada uma das 300 amostras contribui com um erro de previsão. Têm-se ao final deste processo, dois vetores com 12.000 elementos (um contendo os erros de teste antes a poda e outro após a poda).

A Figura 18 ilustra esta situação. Notar que para um primeiro conjunto de modelagem e previsão elaborou-se o primeiro modelo (b_1), que em seguida é utilizado para prever os valores para o primeiro conjunto teste. Encontra-se um erro de previsão para cada espectro do conjunto teste através da comparação dos valores previstos e conhecidos, assim, neste caso, ter-se-á 300 erros de previsão.

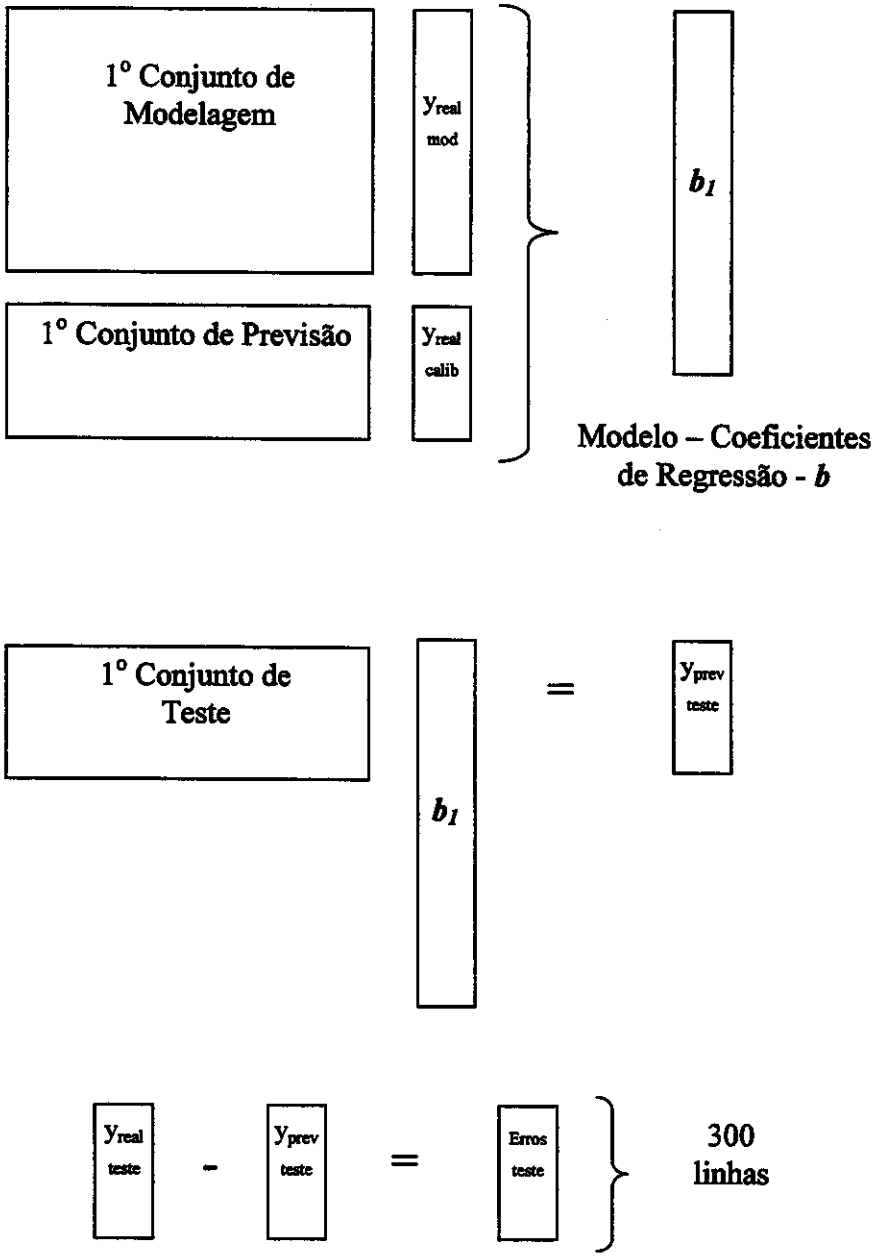


Figura 18: Representação esquemática da construção de um modelo de calibração, e subseqüente avaliação através do conjunto teste.

Emprega-se o mesmo raciocínio para os k-ésimos conjuntos de modelagem e previsão obtidos através da reorganização aleatória das amostras, (k = 2,3,4, ... , 40), portanto:

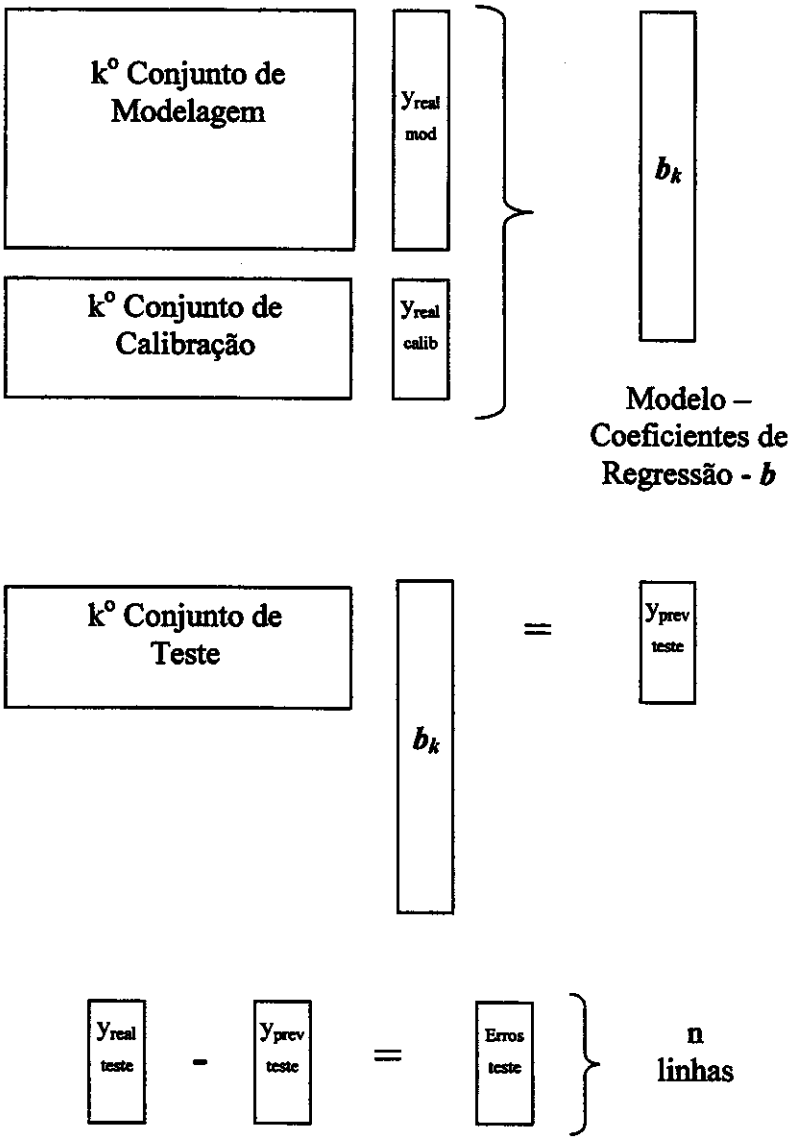


Figura 19: Representação esquemática da construção de um modelo genérico *k*, e subseqüente avaliação através do conjunto teste, composto por *n* amostras.

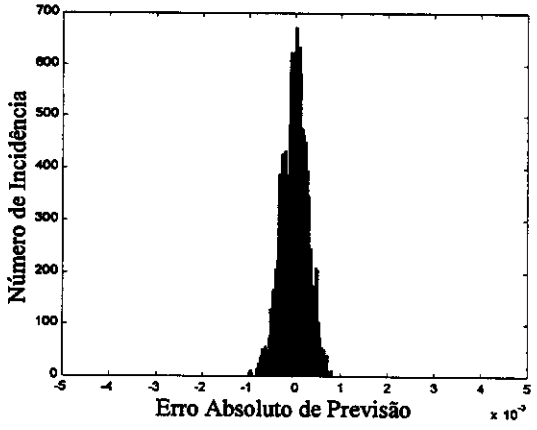
Desta forma, quando k = 40, ter-se-á uma coleção de 40 vetores contendo os 300 erros de previsão associados aos conjuntos testes que, se dispostos seqüencialmente, gera um vetor com as dimensões 12.000 x 1.

Importantíssimo salientar que os testes estatísticos que comparam dispersões podem ser usados, apenas e tão somente, se os dados assumirem valores que sigam uma distribuição normal.

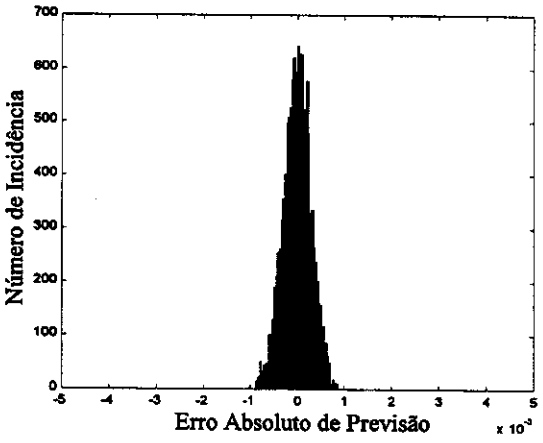
2.9. RESULTADOS

A análise dos resultados consistiu, como ponto fundamental, na comparação dos erros de previsão dos modelos construídos. A análise numérica não se mostraria muito evidente, assim decidiu-se apresentar os resultados graficamente.

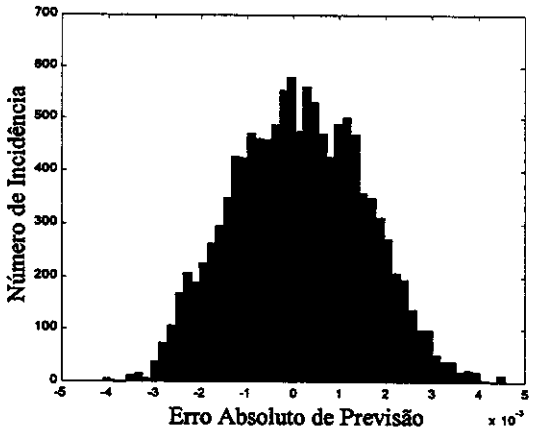
Duas comparações foram exploradas, inicialmente desejava-se verificar se os modelos mais flexíveis, quando submetidos a uma simplificação ainda maior devido ao procedimento da interseção de conjuntos, são equivalentes àqueles obtidos de forma menos drástica, ou seja, quando se aplica apenas o procedimento da interseção de conjuntos para modelos rígidos ou quando simplesmente emprega-se um pequeno limite de tolerância para os valores dos erros. Posterior, e principalmente, verificar se os modelos elaborados com as variáveis selecionadas pelo método da poda-PLS1 são equivalentes aos modelos construídos utilizando o método PLS1 antes da seleção das variáveis. As Figuras 20a a 20c mostram como se distribuem os erros de previsão para cada uma das amostras utilizadas na elaboração dos modelos e ajudarão na análise em questão, visto que dois modelos são tidos como equivalentes se estes se distribuírem segundo uma distribuição normal e se a análise da variância de seus erros estiverem dentro do intervalo definido pelo teste-*F* [27,28].



(a)



(b)



(c)

Figura 21: Distribuição do número de incidência do erro absoluto de previsão para: (a) o modelo construído antes da seleção de variáveis, (b) o modelo construído, após a seleção de variáveis em que se utilizou o método da interseção de conjuntos e (c) o modelo construído após a seleção de variáveis, em que se utilizou o método da interseção de conjuntos aliado ao limite de tolerância de 5%.

Uma análise mais simplificada das Figuras 20a e 20b mostra que há uma fortíssima semelhança entre a distribuição dos erros associados aos modelos construídos: 1 – com todas as variáveis, ou seja, antes da poda e 2 – sem um fator de tolerância. Já as Figuras 20a e 20c mostram uma total discrepância entre seus perfis, lembrando que os modelos que geraram os erros mostrados tanto na Figura 20b quanto na Figura 20c, foram submetidos ao procedimento da interseção de conjuntos. Os resultados numéricos, são mostrados na Tabela II.

Os cálculos realizados durante este estudo foram feitos levando-se em consideração as precisões adequadas a cada fator, isto é, a precisão máxima dos resultados (previsto) não pode ser maior que a precisão das variáveis dependentes (esperado).

Consultando o valor limite de $F_{\infty,\infty}$ com 95% de confiança, encontramos o valor 1,00 [29]. Assim, pode-se dizer que os modelos construídos pelo método da interseção de conjuntos, com 19 variáveis, apresentam as mesmas dispersões observadas para o método tradicional, visto que a validação desta hipótese é obtida se $SEP_A/SEP_B \leq 1{,}00$. Isso indica que reduzindo-se até 19 o número de variáveis, o modelo correspondente irá produzir resultados, com 95% de confiança, iguais ao modelo construído com todas as 70 variáveis, o que representa uma redução de 85% no número de variáveis. Para o método com tolerância de 5% e interseção de conjuntos o valor da razão entre os SEP's é igual a 24, não sendo, portanto, equivalente ao método tradicional (antes da poda). Resta apenas saber se o desvio médio entre os valores previstos e esperados, para cada modelo, também são estatisticamente equivalentes. Em outras palavras deseja-se saber se a média do desvio quadrático entre os valores estimados pelos modelos (tradicional e da poda) e os valores esperados são próximos.

Tabela II: Resultados comparativos do método tradicional e após seleção de variáveis.

	Método da Interseção de conjuntos			
	SEP_A^*	$RMSEP_A$	SEP_B^*	$RMSEP_B$
	$8{,}9.10^{-8}$	$3{,}0.10^{-4}$	$8{,}4.10^{-8}$	$2{,}9.10^{-4}$
$F_{v_A,v} = \frac{SEP_A}{SEP_B}$	**1,0**			
	Método com Tolerância de 5% + Interseção de conjuntos			
	SEP_A^*	$RMSEP_A$	SEP_B^*	$RMSEP_B$
	$2{,}0.10^{-6}$	$1{,}6.10^{-3}$	$8{,}4.10^{-8}$	$2{,}9.10^{-4}$
$F_{v_A,v} = \frac{SEP_A}{SEP_B}$	**24**			

$F_{\infty,\infty}$ = 1,00 [29]; *O índice A corresponde ao método tradicional de análise com todas as variáveis, enquanto o índice B corresponde ao método proposto (poda-PLS1)

Extraindo a raiz quadrada do SEP teremos o RMSEP (*Root Mean Square Error Prediction* – Raiz do Erro de Previsão Quadrático Médio) que indica qual o valor médio do erro de previsão. Comparando-se esse valor podemos avaliar se o erro praticado pelos modelos, em média, é próximo daquele obtido antes de executar a seleção de variáveis. Os RMSEP's obtidos são apresentados na Tabela II. Conclui-se, portanto, que reduzindo-se até 19 o número de variáveis no modelo rígido e empregando o método da interseção, obtém-se as mesmas dispersões e os mesmos erros quadráticos médios de previsão, que o método que utiliza todas as variáveis. Isso já não ocorre para o método com tolerância de 5% mais a interseção de conjuntos, onde os valores RMSEP são bastante diferentes.

A análise mostra que a utilização do método de interseção de conjuntos aliado a modelos mais flexíveis (com erros de previsão em níveis de tolerância de 5%) não é equivalente ao modelo tradicional, impedindo, portanto, o seu emprego em análises futuras. Assim, fica a partir daqui, estabelecido que se utilizará, para identificação das variáveis informativas, o método estatisticamente equivalente ao método tradicional, isto é, o método em que se emprega valores mínimos absolutos conjuntamente ao método da interseção de conjuntos.

A Figura 16 corresponde, então, aos 19 comprimentos de onda selecionados para construção do modelo simplificado.

Através destas análises foi possível avaliar o desempenho do *software* utilizado, bem como estabelecer os critérios e artifícios necessários à execução deste estudo. O caráter exploratório desta etapa foi bem sucedida no que diz respeito ao treinamento e direcionamento das análises.

O programa desenvolvido mostrou-se eficiente, pois foi capaz de reduzir um conjunto de variáveis em 85% sem, com isso, provocar uma diminuição na capacidade de previsão.

A partir deste resultado verificou-se a possibilidade de simplificar modelos potencializando sua utilização em análise de amostras reais.

As análises, apesar de demandaram um bom tempo de processamento (cerca de oito horas por conjunto analisado), mostraram-se muito robustas, pois elegeu o conjunto de variáveis que conhecidamente continham informações importantes sobre a amostra. Porém, depois que a seleção é realizada, os cálculos tornam-se extremamente mais rápidos. Tal conjunto mostrou-se genérico sendo capaz de prever satisfatoriamente amostras desconhecidas (como foi ocaso das amostras teste).

CAPÍTULO 3

3. Dados Reais

Efetuaram-se estudos com dados reais para aplicar o método proposto neste trabalho. Empregou-se este método a fim de realizar previsões de quantidades de açúcares por três diferentes metodologias: medições da porcentagem Brix (%Brix), polarimétricas (Pol) e de açúcares redutores (AR), em caldo de cana por espectroscopia no infravermelho próximo.

Todos os dados utilizados para realização destes estudos consistiam de amostras de caldo de cana-de-açúcar adquiridos na região do infravermelho próximo (1200 a 2500nm), resolução de 2nm, coletadas no Centro de Tecnologia Copersucar (CTC), na cidade de Piracicaba.

O equipamento utilizado foi um espectrofotômetro NIR900PLS da FEMTO, que possibilita a análise em cubetas (1mm) de fluxo contínuo. Os espectros foram adquiridos medindo-se a absorbância.

3.1. Análise pelo Método de Porcentagem Brix (%Brix)

O conjunto de dados utilizado para realização deste estudo era composto por espectros de 307 amostras de açúcares cujas quantidades foram determinadas, pelo método %Brix, a partir do caldo extraído da cana-de-açúcar, que indica o percentual de material sólido (soluto) em 100g de solução [30]. Este parâmetro é medido através de um sacarímetro.

3.1.1. Pré-Processamento

A análise PCA [7] possibilitou identificar, para esse conjunto, amostras cujos espectros apresentassem comportamento anômalo (*outlier*). Classificam-se como anômalas as amostras que apresentam altos valores de *leverage* e resíduo. A medida *leverage* representa o quanto uma amostra está distante da média das demais, isto é, é uma medida que informa se uma amostra é diferente das demais, porém ela só será definida como anômala se tiver um resíduo alto se comparado à média das demais amostras. Em outras palavras, pode-se dizer que se uma amostra for muito diferente das demais, e além disso o modelo não for capaz de

fazer boas previsões para ela (alto valor de resíduo) então trata-se de uma amostra anômala (*outlier*). A Figura 21 mostra o gráfico de influência, que consiste em representar graficamente os valores de *leverage* contra os valores de resíduo associados a cada amostra. Verifica-se que, num intervalo de 95% de confiança, há sete (7) amostras potencialmente comprometidas, sendo que duas delas (2 e 158) comportam-se de modo completamente diferente da média.

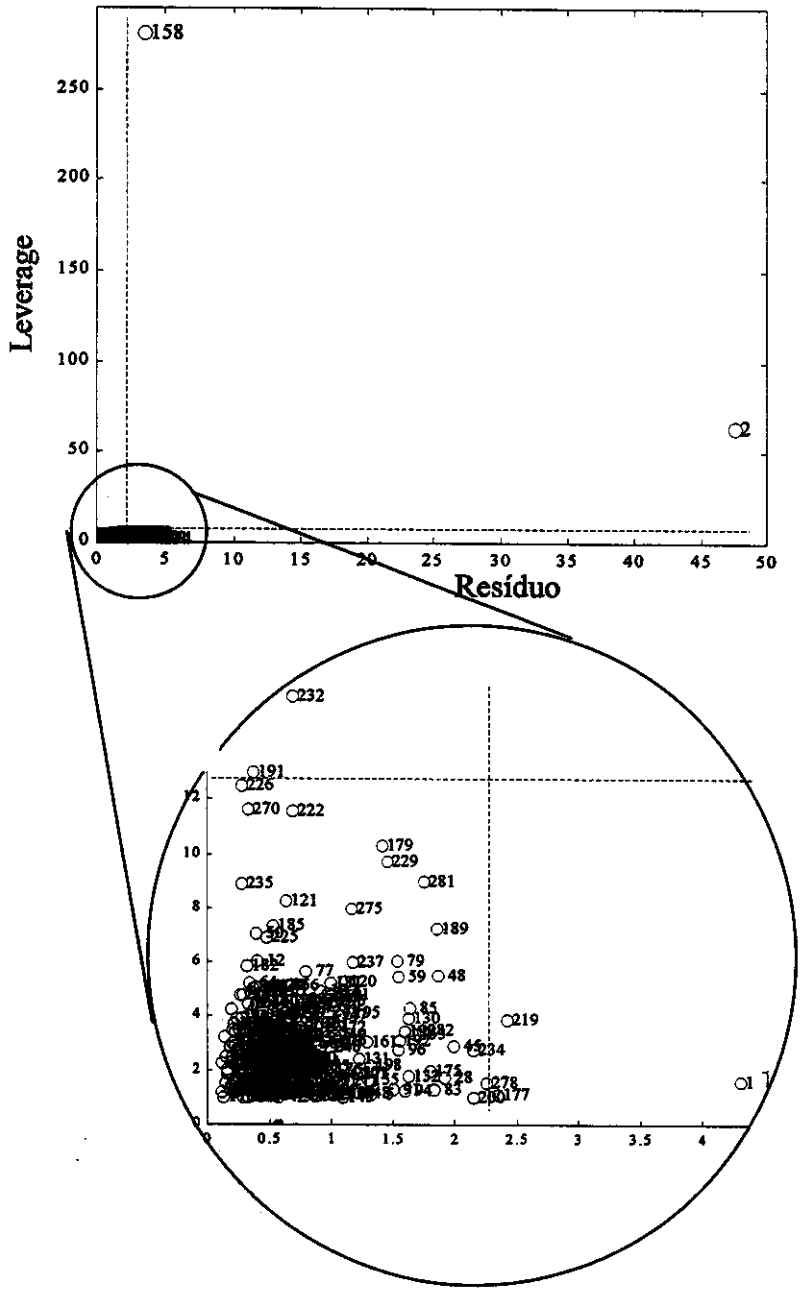


Figura 21: Gráfico de influência utilizado para identificação de *outliers*

A Figura 22 mostra os espectros mantidos.

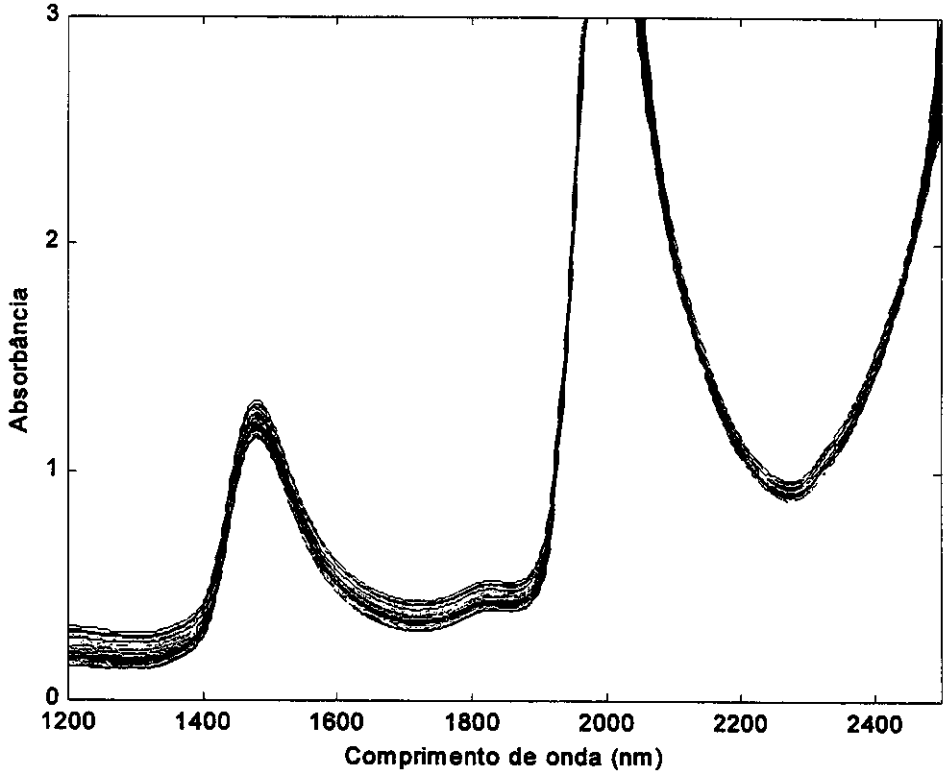


Figura 22: Espectros das 300 amostras de caldo de cana-de-açúcar adquiridos medindo-se a absorbância

Na Figura 22 verifica-se a existência de uma região do espectro em que o sinal está saturado devido à banda de água. As informações contidas nessa região não servem para construção do modelo, uma vez que não é possível estabelecer uma correlação linear entre os comprimentos de onda dessa região e a quantidade de analito correspondente. Decidiu-se, portanto, eliminar dentro dessa região as variáveis que não poderiam ser utilizadas. A Figura 23 mostra o resultado deste processamento:

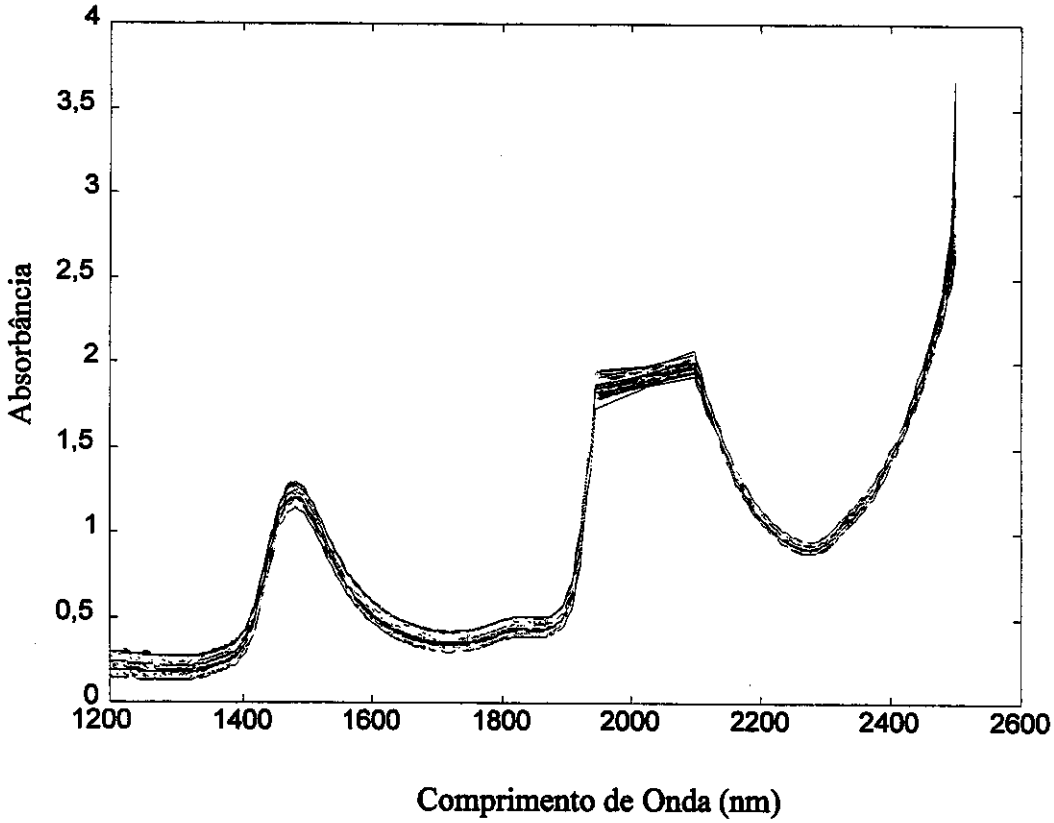


Figura 23: Espectros das 300 amostras de caldo de cana-de-açúcar após pré-processamento

Após tal corte, os espectros passaram a apresentar 555 variáveis, indicando, portanto que foram eliminadas 71 variáveis. A região que compreende o final do espectro foi mantida, embora seus valores de transmitância também fossem altos, e apresentavam pouco ruído, comparativamente à região eliminada, indicando que não estavam saturados.

Através dos resultados obtidos com os dados simulados decidiu-se, além da eliminação discutida, diminuir o número de variáveis pela metade, devido ao processamento matemático exigido pelo programa, que sendo iterativo, necessita de um tempo computacional geometricamente proporcional ao número de variáveis, para execução completa das podas. Desta forma decidiu-se restringir o número de variáveis (comprimentos de onda) pela metade, ou seja, 278 variáveis. Para isso foram considerados apenas os valores de absorbância dos

comprimentos de onda pares. Este procedimento buscou garantir que mesmo que comprimentos de onda importantes fossem eliminados, outros vizinhos poderiam representá-los, efeito garantido dada a resolução empregada (2nm).

Distribuiu-se as 300 amostras restantes (já pré-processadas) em três conjuntos distintos: Calibração; Validação e Teste. O critério empregado para proceder a esta distribuição se deu da seguinte forma: a partir da reorganização em ordem aleatória do conjunto original com 300 amostras, pôde-se observar, como ilustrado na Figura 24, que os índices %Brix estavam se distribuindo de maneira mais ou menos homogênea. Assim, decidiu-se atribuir as 150 primeiras amostras para compor o conjunto de Calibração; as próximas 75 amostras para compor o conjunto de Validação e as últimas 75 amostras para o conjunto Teste. A Figura 25 mostra o resultado deste procedimento.

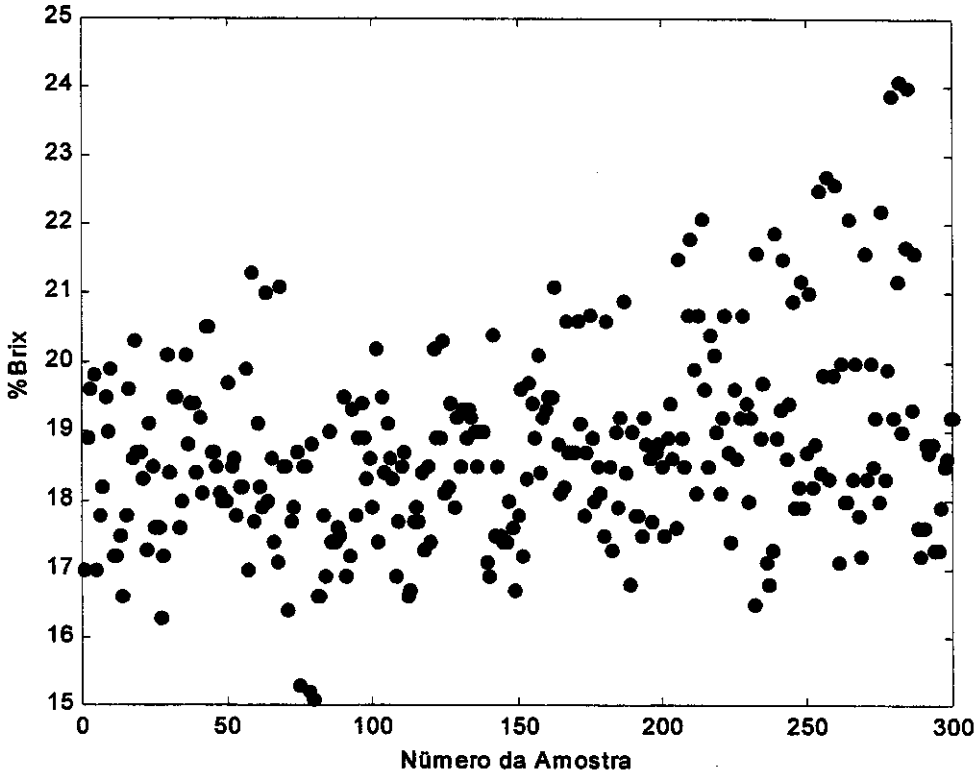


Figura 24: Distribuição das medidas do índice %Brix associadas aos espectros de caldo de cana-de-açúcar.

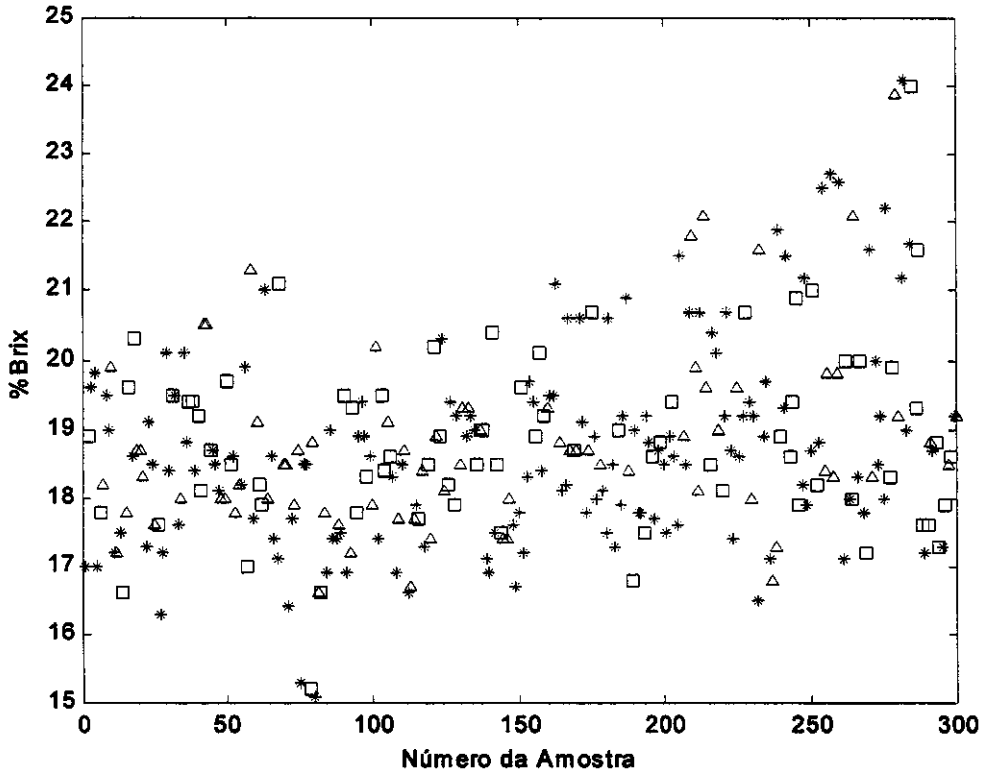


Figura 25: Distribuição das medidas do índice %Brix associadas aos espectros de caldo de cana-de-açúcar, sendo: ✳ – calibração; △ – previsão e □ teste.

Observando-se, na Figura 25, os valores dos índices %Brix referentes às amostras de previsão e teste, percebe-se que para cada uma delas há um suficiente número de amostras de modelagem que assumem valores próximos, favorecendo, assim, a elaboração de um modelo de calibração mais preciso.

O conjunto de calibração foi utilizado para construção do modelo via PLS1, validando-o com o conjunto de previsão. O conjunto teste serviu para verificar a capacidade de generalização do modelo pois, não estando presente diretamente na construção deste, possibilita avaliar como o modelo prevê as amostras "desconhecidas" - objetivo para qual foi construído.

3.1.2. RESULTADOS

Serão apresentados e discutidos aqui os resultados e os procedimentos que foram necessários para transpor as dificuldades que surgiram durante a execução do estudo deste novo método de seleção de variáveis (através da "poda" de coeficientes de regressão obtidos pelo PLS1). Tais artifícios já foram sucintamente descritos e empregados no Capítulo 2, para dados simulados, como por exemplo, a metodologia empregada para escolha do melhor número de variáveis latentes durante a construção do modelo (seção 2.6). As considerações e conclusões apresentadas no capítulo anterior serão prontamente assimiladas nesta etapa como, por exemplo, a necessidade de se gerar diversos conjuntos de dados equivalentes, através da reorganização aleatória das amostras, a fim tornar o procedimento de seleção de variáveis mais genérico (seções 2.7 e 2.8), bem como a utilização da metodologia da interseção de conjuntos que possibilita a identificação das variáveis informativas (seção 2.9). Discussões complementares para o caso específico de análise dos dados reais certamente serão feitas oportunamente, no decorrer desta exposição.

3.1.3. APLICAÇÃO DO MÉTODO DA PODA-PLS1

A seleção das variáveis importantes ao modelo de calibração foi feita submetendo-se o vetor de coeficientes $\boldsymbol{b}$, encontrados pelo método PLS1, à poda segundo os passos descritos no capítulo 1, seção 1.8.

A partir da construção do primeiro modelo de calibração para amostras reais de cana-de-açúcar, pôde-se chegar à Figura 26 que contém a representação gráfica do comportamento geral da variação dos erros à medida em que se diminui o número de variáveis que compõem os modelos de calibração (PLS1). O eixo das abscissas corresponde ao número de variáveis utilizadas na construção dos modelos e o eixo das coordenadas corresponde ao erro de previsão padrão que cada modelo produziu. Verifica-se nesta figura que o conjunto de calibração utilizado proporcionou não só a simplificação do modelo, que era composto por

278 variáveis para 11 variáveis, como também este modelo apresenta erros de previsão menores. Como o procedimento da poda é executado até que todas os coeficientes sejam eliminados, uma grande redução no número de variáveis que compõem o modelo (< 11) provoca um aumento brusco nos valores dos erros de previsão, indicando que um número muito reduzindo de variáveis não é capaz de explicar satisfatoriamente a correlação entre espectro e quantidade de amostra.

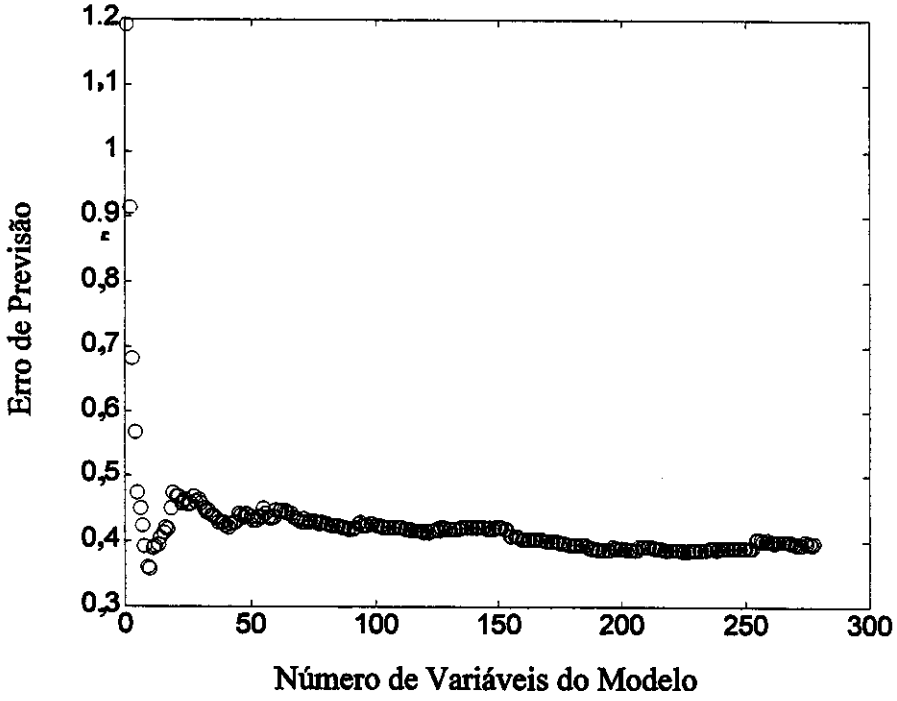


Figura 26: Erros de previsão do conjunto de validação *vs* número de variáveis que compõem o modelo obtido para o conjunto de dados obtidos pela medida do índice %Brix.

A Figura 27 mostra as variáveis selecionadas para este modelo.

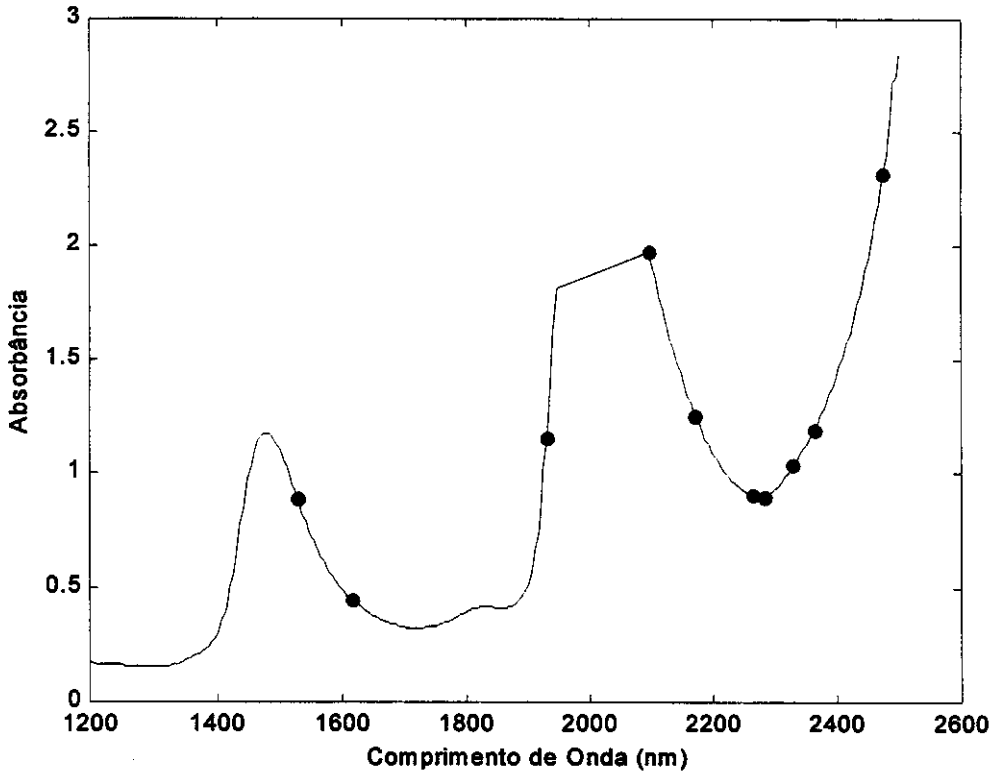


Figura 27: Variáveis selecionadas para o primeiro modelo construído

Como descrita anteriormente a poda dos coeficientes é feita de forma ordenada e seletiva visando, sempre, a minimização da função erro. O procedimento geral para execução da poda empregado nesta etapa é descrito na seção 1.8.

Lembrando que na eliminação de cada coeficiente de regressão do modelo pelo método da poda, associa-se um erro de previsão. A Figura 26, portanto, ilustra o monitoramento dos erros durante a execução das podas, para um conjunto de amostras de caldo de cana-de-açúcar cujas quantidades totais de açúcar foram determinados previamente pelas medidas dos índices Brix.

3.1.4. Análise de Repetibilidade

Uma das vantagens do método proposto sobre o método de poda tradicional (aplicado às redes neurais artificiais), está no fato dos pesos iniciais da

"rede" não serem aleatórios, mas provenientes de uma análise PLS1. Assim, um ***mesmo conjunto*** de dados apresenta sempre os mesmos resultados, independentemente do número de vezes em que se processe a análise, isso garante que as seleções das variáveis para tal conjunto serão absolutamente reprodutíveis, pois o método matemático para essa seleção dá-se, como já dito, de forma ordenada e seletiva.

Usando-se o mesmo conjunto de dados várias vezes, obtém-se sempre os mesmos resultados. O que ocorreria se fosse utilizado outro conjunto de dados para proceder a calibração de um mesmo analito, e com o uso das mesmas técnicas analíticas? Isso depende. Certamente o modelo construído não será idêntico ao primeiro, mas quanto mais parecidos forem os valores das variáveis nos dois conjuntos, mais semelhantes também serão os modelos (coeficientes de regressão) e, conseqüentemente, mais suas capacidades de previsões aproximar-se-ão. Se os valores assumidos pelas variáveis de dois conjuntos de calibração forem muito distintos, os modelos, por conseguinte, também o serão. Isto pode ocorrer, por exemplo, se um dos modelos apresentar baixíssima capacidade de generalização, isto é, o modelo é capaz de produzir erros baixos somente para as amostras utilizadas nos conjuntos de previsão ou teste, sendo, por outro lado, incapaz de realizar boas previsões para conjuntos externos. Isso em geral acontece quando o número de variáveis latentes for muito grande, o que corresponderia a dizer que está-se modelando, além das dependências lineares entre as varáveis dependentes e independentes, também o ruído aleatório inerente àquele conjunto de amostras [31,32]. Como a escolha do número de variáveis latentes é feita automaticamente é possível que isso ocorra, mesmo usando uma metodologia que busque evitar a ocorrência deste tipo de situação [24].

Surgem então questões importantes: Se os modelos forem muito diferentes entre si, ao serem submetidos à seleção de variáveis pelo método da poda, haverá diferença entre as variáveis selecionadas? Como garantir a escolha de um conjunto de modelagem que produza um modelo genérico?

Tanto para os dados simulados quanto para os dados reais, têm-se o método proposto selecionando as variáveis que minimizam os erros, que supostamente são as que concentrem a maior parte das informações das amostras. Assim, submetendo-se à poda-PLS1 vários conjuntos de amostras de uma mesma espécie (neste caso espectros de caldo de cana-de-açúcar), espera-se que aquelas variáveis realmente significativas ao modelo deverão ser selecionadas na maioria das vezes, isto é, estas variáveis terão um número maior de ocorrência, à medida em que se efetua a interseção dos conjuntos selecionados para cada modelo. Este artifício foi discutido e exemplificado na seções 2.6 e 2.9.

Por outro lado, em termos práticos, porém, nem sempre é possível obter uma vasta coleção de amostras, de modo a dispô-las em **diversos** conjuntos distintos.

Esta dificuldade foi encontrada durante a execução deste trabalho, onde espectros de infravermelho próximo (NIR) de 300 amostras de caldo de cana teriam que ser dispostos em três conjuntos: modelagem, validação e teste.

Se se quisesse montar pelo menos dez conjuntos de modelagem, dez de validação e dez de teste, distintos entre si, cada conjunto teria, respectivamente, cerca de 15, 8 e 7 amostras, números insuficientes para proceder-se adequadamente à análise PLS1.

Após a construção do primeiro modelo, apresentado na seção 3.1.3, reorganizou-se as 300 amostras originais em uma outra ordem (aleatória) e "gerou-se" um segundo conjunto de modelagem, de validação e teste "distintos" dos originais, pois houve permuta de amostras entre os três conjuntos: algumas amostras que antes pertenciam ao conjunto de modelagem passaram a pertencer ao conjunto teste, outras amostras do conjunto de validação passaram para o conjunto de modelagem e as amostras do conjunto teste passaram para o conjunto de validação. Este artifício possibilita a construção de um número relativamente grande de conjuntos de modelagem, validação e teste diferentes entre si, cujos números de amostras favorecem uma análise mais segura.

Apesar de se estar construindo modelos equivalentes, pois o conjunto de amostras empregado para a calibração é sempre tirado da mesma população, verifica-se que se trata de uma população que representa satisfatoriamente todo o universo de ocorrências possível, garantindo, assim, a generalidade dos modelos gerados.

Isto pode ser comprovado avaliando-se os valores extremos das concentrações observadas para a coleção das 300 amostras, que se tratando de amostras provenientes de fontes naturais (cana-de-açúcar), indicam estar restritas a uma faixa estreita de valores (μ = 19 ±1,4). Isso representa dizer que as quantidades de analito possíveis de ser encontradas na natureza ficam restritas a uma faixa estreita, devido ao universo de ocorrência ser estreito – valores muito baixos ou muito altos nunca ocorrem – assim as 300 amostras utilizadas aqui devem representar uma boa estimativa do universo possível.

Dada a representatividade deste conjunto de espectros de caldo de cana de açúcar (%Brix) inicial, a escolha das amostras que comporiam os conjuntos de calibração, validação ou teste, pode ser feita aleatoriamente, sem o comprometimento significativo das respostas adquiridas.

3.1.5. Organização das Amostras

Ordenando-se o conjunto das 300 amostras aleatoriamente, produziram-se vinte conjuntos distintos, contendo cada um: 150 amostras para modelagem, 75 amostras para validação e 75 amostras para teste.

3.1.6. Características dos Modelos

Os modelos construídos tinham todos as mesmas características básicas daqueles descritos na seção 2.4, quanto ao método de obtenção dos vetores de regressão; ao critério de escolha do número de variáveis latentes empregados; à metodologia de avaliação do desempenho dos modelos, bem como à metodologia empregada para identificação das variáveis informativas – método da interseção de conjuntos.

Aqui será discutido mais detalhadamente o método da interseção para o caso dos dados reais, visto que houve algumas diferenças.

No caso das amostras simuladas, foram selecionadas as variáveis que apresentavam incidência de 100%, isto é, as variáveis que foram selecionadas em todos os modelos gerados. Intuitivamente esperava-se que o mesmo ocorresse quando se tratasse de dados %Brix, isto é, após a seleção das variáveis dos vinte conjuntos de calibração, algumas das variáveis importantes deveriam estar presentes em todos os conjuntos. Verificou-se, no entanto, que nenhuma variável teve incidência de 100%. Isto pode ter ocorrido se alguns dos conjuntos de calibração fossem muito distintos dos demais e, comportando-se diferentemente média, eliminassem variáveis importantes.

Como então determinar o conjunto de variáveis que produza o melhor modelo? Para encontrar tal conjunto, estudou-se, 4 classes distintas de modelos, que correspondem aos modelos apodizados nos níveis de 80%, 85%, 90% e 95% de concordância entre as variáveis selecionadas.

Os níveis de apodização indicam que as variáveis selecionadas para construção dos modelos, apresentaram percentual de incidência maior ou igual a 80%, 85%, 90% e 95%. Isto corresponde a tornar o método um pouco mais flexível. Nestes níveis selecionou-se, respectivamente, 52, 24, 11 e 5 variáveis.

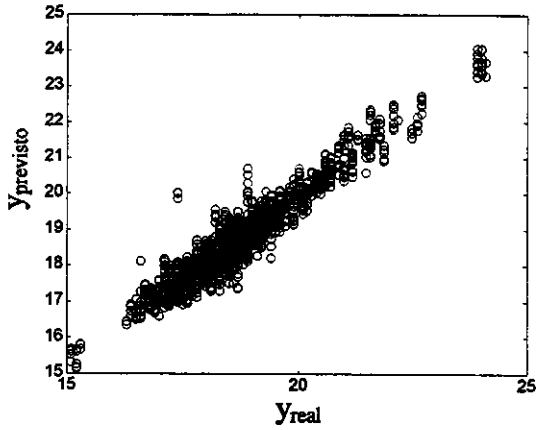
Os resultados obtidos através destes modelos foram comparados aos resultados obtidos pelos modelos onde todas as variáveis foram consideradas relevantes.

A comparação foi feita entre a capacidade de previsão dos conjuntos teste, antes e após a poda, nos quatro níveis de apodização. O parâmetro de comparação usado nesta etapa foi o mesmo usado na seção 2.8 (SEP), dado pela equação 30, através do teste estatístico teste-*F*.

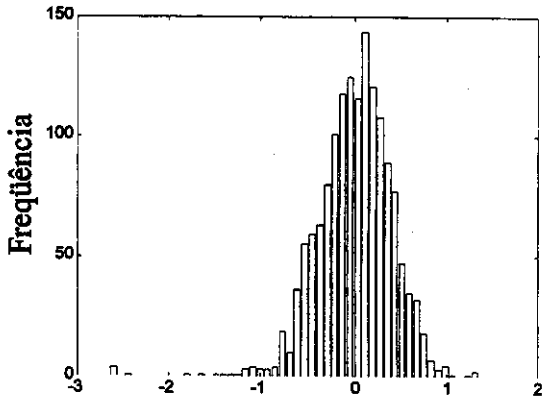
Surge aqui outro empecilho. Dispõem-se de vinte modelos de calibração para cada um dos quatro níveis de apodização (80% a 95%). Cada modelo produz seus correspondentes erros de previsão para os conjuntos teste. Como compará-los todos?

A idéia empregada aqui simplifica a situação. Para um mesmo nível de incidência (por exemplo 80%) foram empregadas as mesmas variáveis (52 comprimentos de onda) para construção de cada um dos vinte modelos e pode-se dizer, portanto, que estes modelos se equivalem. Conseqüentemente, os erros de previsão gerados por esses modelos também deverão ser equivalentes entre si. Por esse raciocínio agruparam-se os erros de previsão de cada um dos vinte conjuntos teste, de modo a obter vetores com 1500 elementos cada (20 conjuntos x 75 erros de previsão) – nota-se que, neste caso, cada uma das 75 amostras contribui com um erro de previsão. Têm-se ao final deste processo, oito vetores com 1500 elementos (cada um dos 4 níveis de incidência produzia 2 vetores contendo os erros teste antes da poda e teste após a poda).

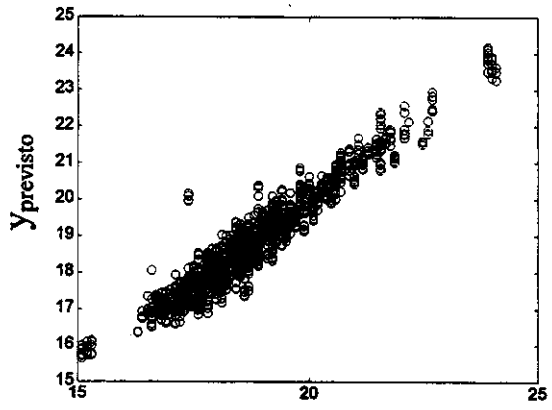
As Figuras 28a até 28j mostram os gráficos dos valores esperados *vs* previstos seguido do histograma dos erros obtido pelos modelos correspondentes ao método tradicional (empregando todas as variáveis) e aos níveis de incidência de 80%, 85%, 90%, 95%.



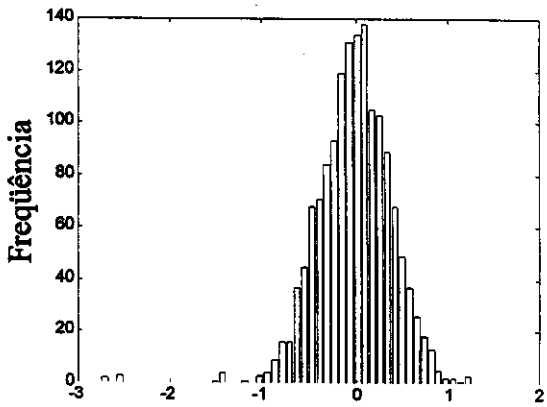
(a) PLS1 (todas as variáveis)



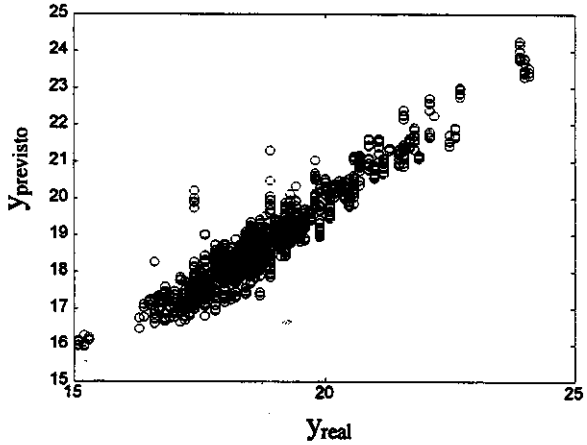
(b) PLS1 (todas as variáveis)



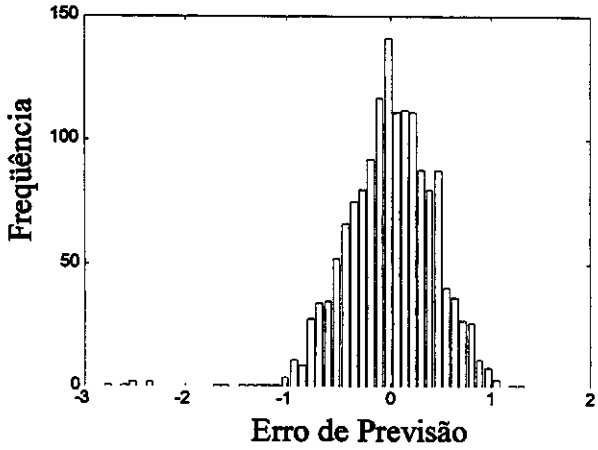
(c) 80%



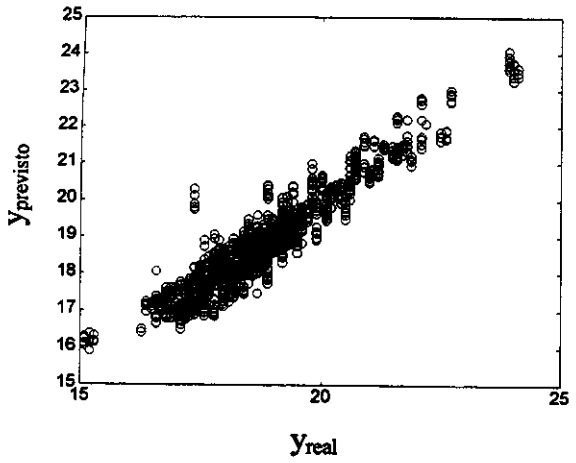
(d) 80%



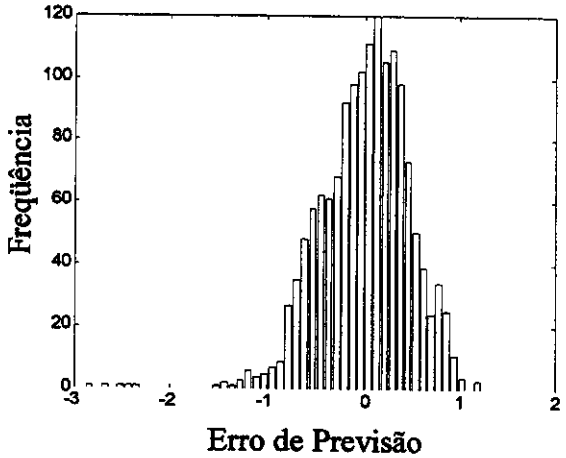
(e) 85%



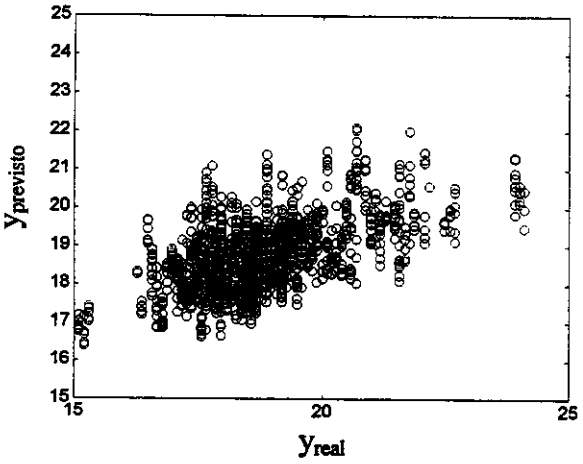
(f) 85%



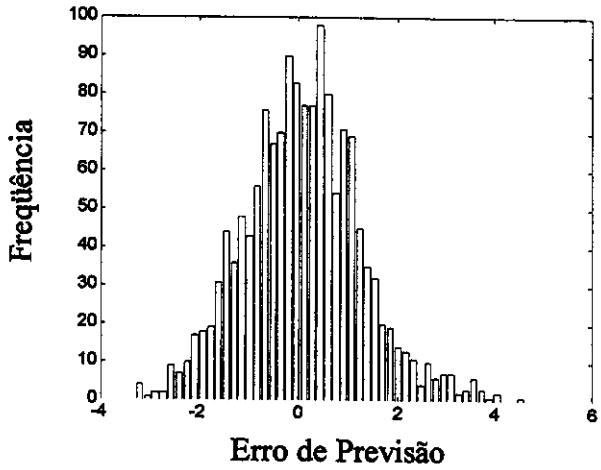
(g) 90%



(h) 90%



(i) 95%



(j) 95%

Figura 28: (a) a (j) Gráficos de valores reais *vs* previstos e respectivos erros de previsão para os métodos: PLS1 (todas as variáveis), poda-PLS1 apodizações de 80%, 85%, 90% e 95%

Verifica-se visualmente que há grande semelhança entre os perfis assumidos pela distribuição dos erros dos métodos tradicional e poda-PLS1 ($\mu \approx 0$ e $\sigma \approx 1$), onde os níveis de incidência são 80%, 85% e 90%, indicando que o método poda-PLS1 (mais simples) gera os mesmos resultados do método tradicional. No caso em que o nível de incidência é de 95%, que corresponde ao conjunto composto por apenas 5 variáveis, o perfil da distribuição muda

perceptivelmente ($\mu \approx 0$ e $\sigma \approx 4$), indicando que os erros obtidos para estes modelos já não são equivalentes aos do método tradicional.

A aleatoriedade dos erros também pode ser verificada visualmente, pois os erros mostram seguir uma distribuição normal, sendo que as maiores freqüências ocorrem em torno do valor zero.

Satisfeita esta condição (aleatoriedade dos dados) emprega-se o teste-*F*. Este teste esclarece se há diferença na precisão de dois métodos A e B. Para efetuar o teste, calcula-se a razão entre as variâncias dos métodos a serem comparados, a variância maior sempre aparece no numerador. [33]

$$F_{\nu_A,\nu_B} = \frac{s_A^2}{s_B^2} \qquad (31)$$

onde ν_A e ν_B correspondem aos graus de liberdade dos métodos A e B, s_A^2 e s_B^2 correspondem às estimativa dos desvios padrões para os métodos A e B, respectivamente.

No caso em questão deseja-se determinar se há diferença entre os métodos em estudo e o método tradicional, por isso chamaremos de método A o método em estudo e método B o método tradicional. Os graus de liberdade nos dois casos podem ser considerados infinitos devido ao grande número de medidas (1500).

Comparando-se o método tradicional (SEP_B = 0,2) com os modelos construídos nos quatro níveis de incidência, têm-se:

Tabela III: Resultados comparativos dos métodos tradicional e após seleção de variáveis.

	Método da Interseção para os Conjuntos Testes							
	80%		85%		90%		95%	
	SEP_A*	$RMSEP_A$	SEP_A*	$RMSEP_A$	SEP_A*	$RMSEP_A$	SEP_A*	$RMSEP_A$
	0,2	0,4	0,2	0,4	0,2	0,4	1,4	1,2
$F_{\nu_A,\nu} = \frac{SEP_A}{SEP_B}$	1		1		1		7	

$F_{\infty,\infty}$ = 1,00 [29]; *O índice A corresponde ao método proposto (poda-PLS1), enquanto o índice B corresponde ao método tradicional de análise.

Observa-se que os valores calculados são, para os graus de incidência de 80%, 85% e 90%, idênticos entre si. Isto ocorre porque os cálculos foram todos feitos levando-se em consideração as precisões adequadas a cada fator, isto é, a precisão máxima dos resultados (%Brix previsto) não pode ser maior que a precisão das variáveis dependentes (%Brix esperado).

Consultando o valor limite de $F_{\infty,\infty}$ com 95% de confiança, encontramos o valor 1,00 [29]. Conclui-se, portanto, que os modelos construídos com 52, 24, e 11 variáveis são equivalentes ao método tradicional. Assim, modelos com as 11 selecionadas irão produzir resultados, com 95% de confiança, iguais ao modelo construído as 278 variáveis totais.

Extraindo a raiz quadrada do SEP teremos o RMSEP (*Root Mean Square Error Prediction* – Raiz do Erro de Previsão Quadrático Médio) que indica qual o valor médio do erro de previsão. Comparando-se esse valor podemos avaliar se o erro praticado pelos modelos, em média, são próximos. Os modelos gerados pelo método da poda-PLS1, com incidência de 80% a 90%, apresentam SEP's idênticos aos tradicionais e, conseqüentemente, RMSEP's também idênticos.

A Figura 29 mostra os 11 comprimentos de onda selecionados para construção do modelo simplificado.

Na literatura [34], várias bandas de absorção, encontrados durante análise de caldo de cana realizadas por espectroscopia na região do infravermelho próximo, foram identificadas e suas atribuições foram compreendidas. A maioria dos comprimentos de onda selecionados podem ser atribuídos a sobretons e bandas de combinação de ligações –CH, $-CH_2$, $-CH_3$ e –OH. Estes comprimentos de onda estão correlacionados com os açúcares (glicose, frutose e sacarose) dissolvidos nas amostras de caldo de cana.

Há uma banda, nas proximidades da região de 2500nm atribuída ao estiramento combinado de C–H, C–C, C–O–C. Há, também, várias bandas na região e 2280–2330 nm provenientes da combinação dos estiramentos C–H e deformações $-CH_2$. Outra banda característica ocorre em 2100nm, a qual resulta da combinação do estiramento e deformação das ligações O–H. Finalmente

observa-se, na região de 1450nm, o aparecimento da banda devida ao primeiro sobreton do estiramento O–H.

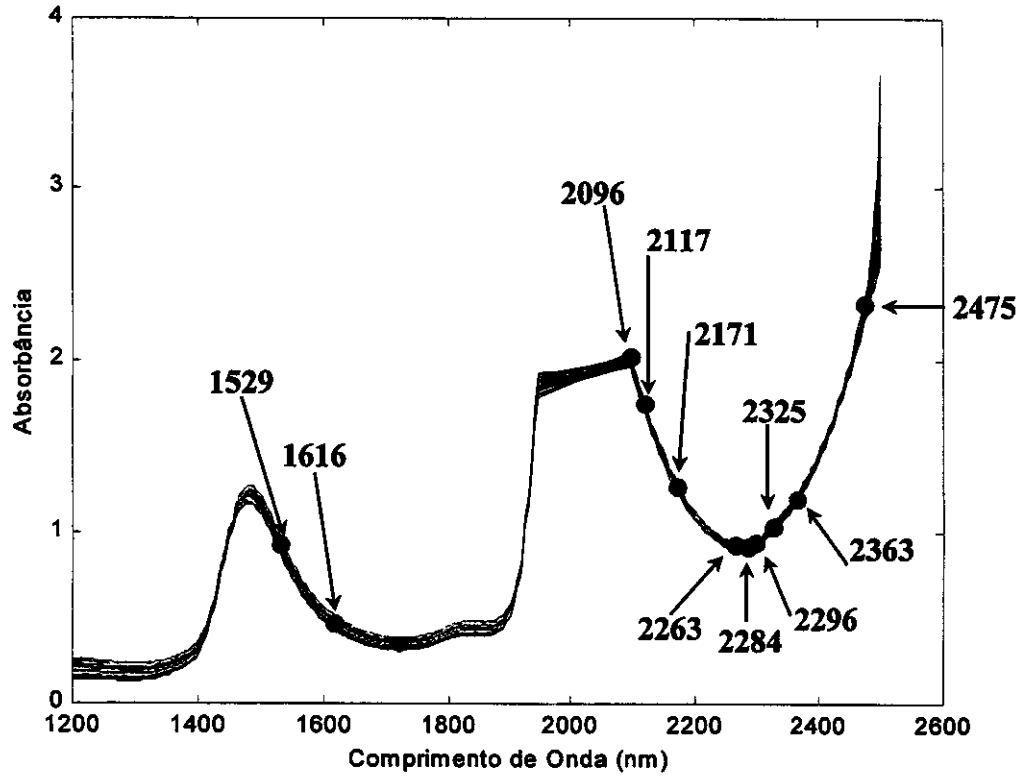


Figura 29: Valores e localização no espectro dos 11 comprimentos de onda selecionados pelo método da poda-PLS1

3.2. Análise pelo Método Polarimétrico (Pol)

As quantidades de açúcares presentes em um conjunto composto por 203 amostras foram determinadas através de um sistema analítico denominado Pol, que estabelece a concentração de substâncias opticamente ativas (açúcares) medindo-se a deflexão que um feixe de luz polarizada sofre ao atravessar a amostra, por determinado caminho óptico. Este parâmetro é medido através de um polarímetro [35].

3.2.1. PRÉ-PROCESSAMENTO

A primeira etapa suplantada consistiu de uma cuidadosa verificação dos dados a serem analisados. Observou-se o comportamento geral dos espectros, bem como as faixas de concentrações correspondentes, a fim de determinar se os dados eram bem comportados. Nesse processo pôde-se executar um adequado pré-processamento.

A análise PCA tornou possível a identificação e eliminação dos espectros com comportamentos anômalos (*outliers*). A Figura 30 indica os *Leverages* e os Resíduos para as 203 amostras analisadas pelo método Pol.

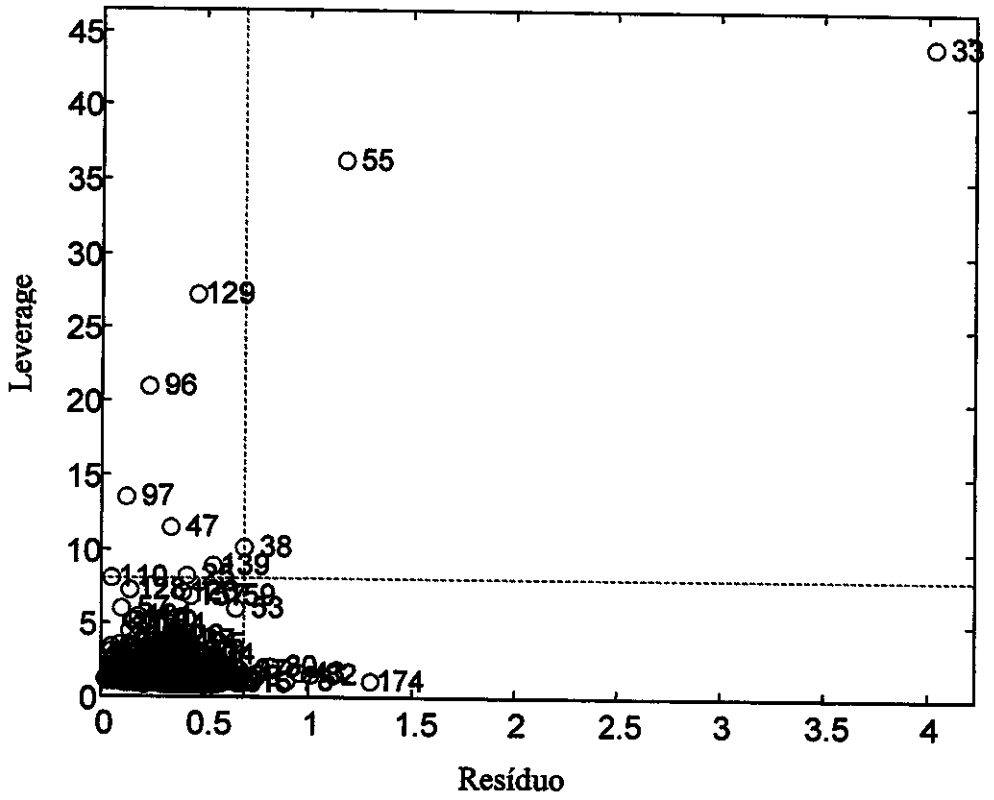


Figura 30: Gráfico dos resíduos *vs Leverage* para as amostras de cana-de-açúcar analisadas pelo método Pol.

As amostras eliminadas do conjunto foram: 33; 38; 43; 47; 55; 78; 82; 96; 97; 129; 174, pois estas estavam fora do intervalo de confiança de 95% - linha tracejada. Restringiu-se, assim, o número de amostras para 192.

A Figura 31 mostra os espectros mantidos

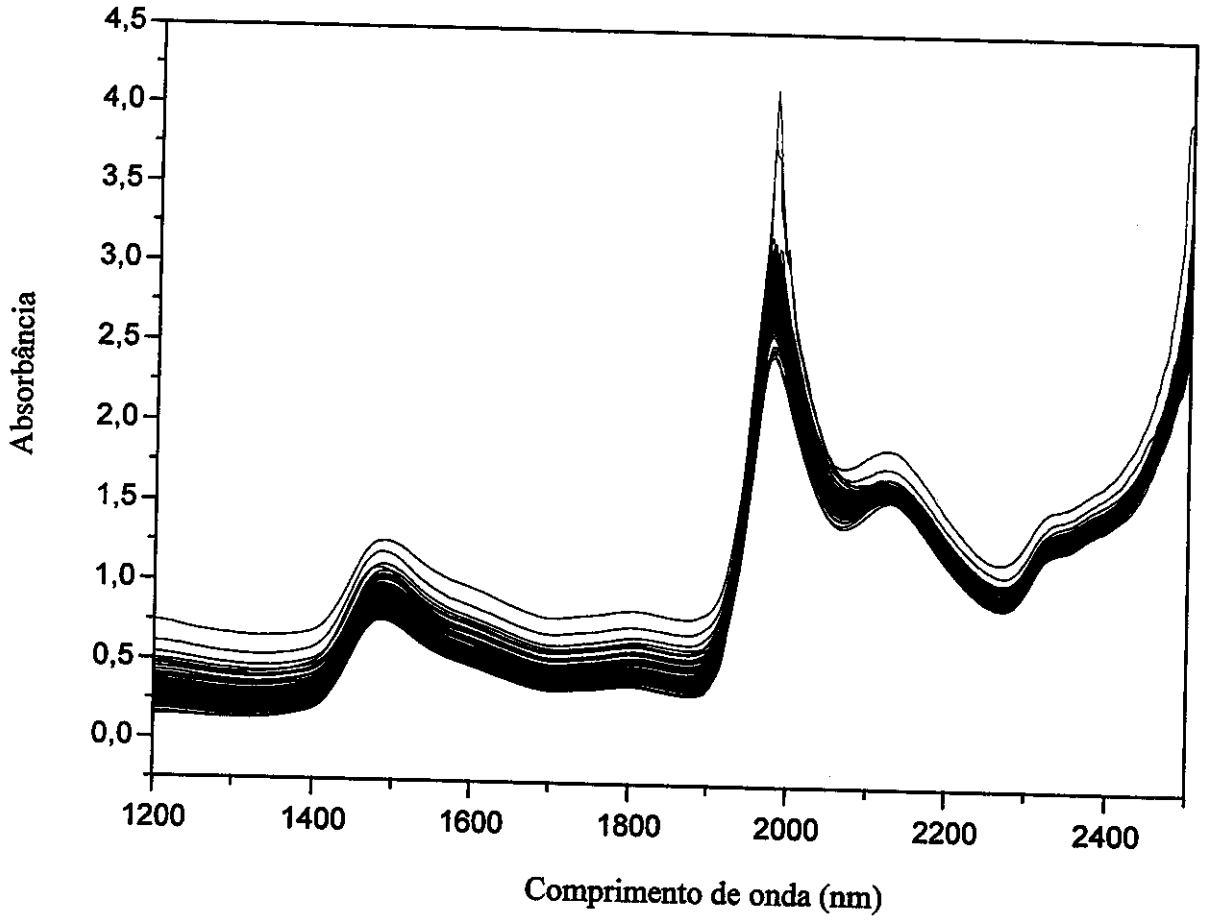


Figura 31: Espectros das amostras de cana-de-açúcar adquiridos medindo-se as absorbâncias, analisadas pelo método Pol.

Semelhantemente à análise realizada para o conjunto %Brix, anteriormente apresentado, foram considerados apenas os valores de absorbâncias dos comprimentos de onda pares, que restringiu o número de variáveis (comprimentos de onda) de 626 para 313. Para este conjunto de dados

não se observou nenhuma banda com sinal saturado, pois estas amostras eram provenientes de uma etapa do processo industrial, em que houve uma severa pré-concentração, de modo que o caldo de cana-de-açúcar passa por um estágio de concentração e passa a ser denominado "xarope". Uma vez que a quantidade de açúcar está presente em uma maior concentração, a quantidade de água nessa solução não é suficiente para saturar o sinal com sua banda correspondente.

Ao final do pré-processamento obteve-se 192 amostras com 313 variáveis cada. A distribuição das amostras nestes conjuntos deu-se de forma um pouco mais criteriosa que aquela empregada para os dados Brix. No caso anterior, onde conjuntos de dados foram analisados pelo método Brix, produziu-se 20 modelos de calibração distintos a partir da reorganização aleatória das amostras. Isso acarretou num custo computacional bastante significativo. Assim, visando minimizar o tempo de processamento para este novo conjunto (analisado segundo o método polarimétrico – Pol), decidiu-se diminuir de 20 para 10 o número de modelos construídos. Para que os resultados fossem tão representativos como os apresentados na seção 3.1.6, optou-se por não mais distribuir as amostras aleatoriamente, mas através de uma análise gráfica visual. A partir da distribuição dos valores dos índices Pol das amostras em análise foi possível escolher, visualmente, quais deveriam pertencer ao conjunto de modelagem (96 amostras), e quais deveriam compor os conjuntos de validação e conjuntos teste (48 amostras cada). O intuito deste artifício foi o de garantir a similaridade entre os modelos construídos, e com isso diminuir a probabilidade do método da poda-PLS1 selecionar conjuntos de variáveis muito discrepantes quanto às suas composições, por haver modelos de calibração muito distintos entre si. Um exemplo do resultado final de uma distribuição de índices Pol entre os conjuntos de modelagem, validação e teste, pode ser observado na Figura 32.

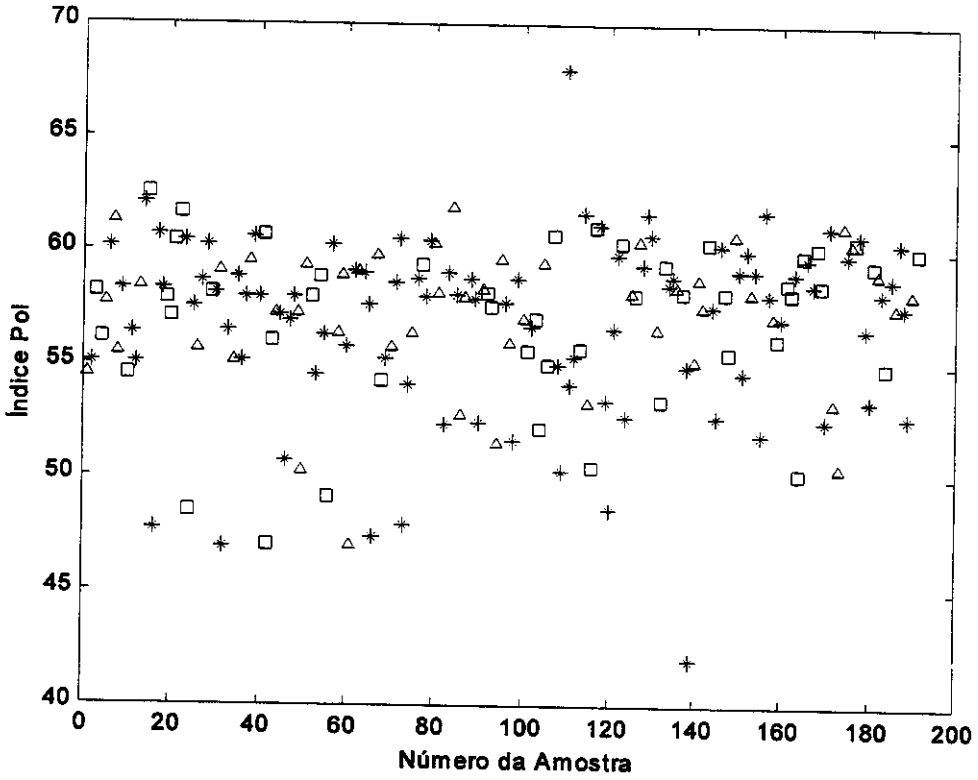


Figura 32: Distribuição das medidas do índice Pol associados aos espectros de caldo de cana-de-açúcar, sendo: ✳ – modelagem; △ – previsão e □ – teste.

Os modelos de calibração serão tanto melhores quanto mais homogênea for a distribuição das amostras do conjunto de modelagem, o que favorece uma maior generalização do modelo. A precisão medida durante a elaboração do modelo e durante sua validação pelo conjunto teste será tanto maior quanto mais próximos às amostras de modelagem estiverem das amostras de validação e teste. A Figura 32 mostra que esta situação pode ser verificada, isto é, as amostras do conjunto de modelagem apresentam-se, na maioria dos casos, próximas às amostras de validação e teste.

A construção do modelo deu-se de forma idêntica à empregada para o conjunto Brix, ou seja, o conjunto de modelagem foi utilizado para construção do modelo via PLS1, validando-o com o conjunto de validação. O conjunto teste serviu para verificar a capacidade de generalização do modelo.

3.2.2. Resultados

A análise deste conjunto de dados (Pol) deu-se de forma muito semelhante a do conjunto anterior (%Brix), os quais foram descritos e empregados no Capítulo 2, para dados simulados. Assim, considerações e conclusões obtidas para dados simulados serão empregadas nesta etapa, da mesma forma como ocorreu nas seções anteriores. As discussões complementares serão feitas à medida que forem sendo expostos os resultados.

3.2.3. Aplicação do Método da Poda-PLS1

Monitorando-se o erro de previsão padrão para o conjunto teste, pode-se determinar a "arquitetura" e, portanto, o conjunto de coeficientes, que produzem o erro mínimo. As variáveis associadas a esses coeficientes é que compõem o conjunto de variáveis selecionadas pelo método da poda-PLS1. A Figura 33 ilustra o monitoramento dos erros durante a execução das podas no caso da análise dos modelos construídos utilizando amostras de caldo de cana-de-açúcar, cujas quantidades de açúcares foram determinados pela medida do índice Pol.

A Figura 33 ilustra o comportamento geral da variação dos erros durante a simplificação dos modelos de calibração (PLS1). Observando os valores de erros mínimos absolutos, verifica-se, nesta figura, que houve uma redução do número de variáveis de 313 para apenas 3, e ainda que o valor numérico do erro de previsão neste caso é menor do que no caso em que todas as variáveis foram utilizadas na construção do modelo. Como já discutido anteriormente, o procedimento da poda é executado até que todos os coeficientes sejam eliminados, produzindo, portanto, um aumento substancial nos valores dos erros de previsão nas situações em que há uma grande redução no número de variáveis que compõem o modelo, o que corresponderia a dizer que um número muito reduzido de variáveis não é capaz de explicar satisfatoriamente a correlação entre espectro e quantidade de analito nas amostras em análise.

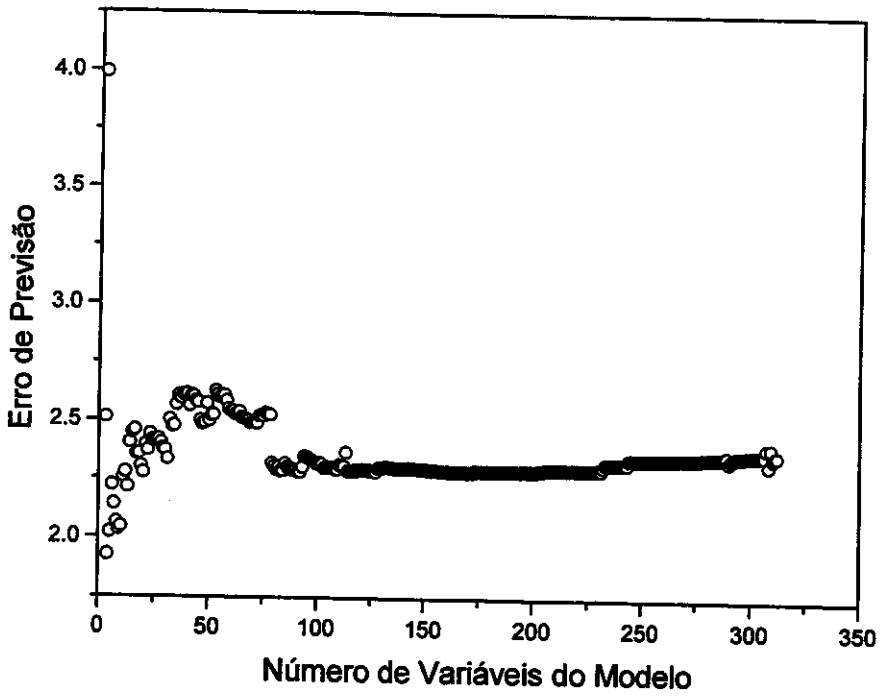


Figura 33: Erros de previsão do conjunto de validação *vs* número de variáveis que compõe os modelos do conjunto Pol.

3.2.4. Análise de Repetibilidade

Durante esta etapa foi verificado que, assim como nos casos anteriores, o conjunto de dados Pol, ao ser reorganizado diversas vezes, gera modelos equivalentes, porém distintos entre si. Com base na análise realizada durante a seção 2.4, será empregado o método da interseção de conjuntos para identificar o conjunto mais genérico de variáveis capazes de gerar modelos com erros de previsão semelhantes aos obtidos para os casos em que todas as variáveis são empregadas (método tradicional).

Após a construção do primeiro modelo, descrito na seção 3.2.3, reorganizou-se as 192 amostras originais para o conjunto Pol em uma outra ordem (segundo a metodologia descrita na seção 3.2.1) de modo a "gerar" um segundo conjunto de modelagem, de validação e de teste "distintos" dos originais. Este artifício foi repetido por mais oito vezes, possibilitando, portanto, a construção de dez conjuntos diferentes entre si.

3.2.5. Organização das Amostras

Ordenando-se os conjuntos das 192 amostras (Pol) pelo critério da seleção gráfica visual, produziu-se dez conjuntos distintos, contendo: 89 amostras para calibração, 45 amostras para validação e 44 amostras para teste.

Como os valores extremos das concentrações observadas para a coleção das amostras estavam restritos a uma faixa estreita de valores ($\mu_{Pol} = 57 \pm 4$) é possível concluir que 192 amostras utilizadas aqui devem representar uma boa estimativa do universo.

3.2.6. Características dos Modelos

Os modelos construídos em todos os casos correspondem aos vetores de regressão calculados segundo o PLS1. O número de variáveis latentes empregado na construção de cada um dos dez modelos era sempre o que produzia o menor erro de previsão para seus respectivos conjuntos de validação [24].

A avaliação do desempenho dos modelos em prever novas amostras deu-se com o emprego dos conjuntos teste, pois como já explicado, as amostras não participando da construção dos modelos, produzem erros de previsão que representam melhor a eficiência dos modelos construídos.

Tendo em mente que as variáveis selecionadas com maior incidência são as mais significativas para o modelo, determinou-se o percentual de incidência de cada uma das variáveis após proceder à seleção pelo método da poda-PLS1 para cada modelo.

Através da utilização do método da interseção de conjuntos buscou-se identificar quais as variáveis mais importantes ao modelo. Num caso ideal esperar-se-ia alcançar valores incidência iguais a 100%, isto é, as variáveis importantes para cada um dos conjuntos de calibração sempre seriam as mesmas. Porém, dada à complexidade de amostras reais, mas principalmente à menor precisão dos valores que o método polarimétrico gera, verificou-se que a construção dos

modelos de calibração devem ter ficado comprometidas, visto que nenhuma variável teve incidência de 100%, apesar de ter-se tido o cuidado de construir modelos similares, o que diminui a possibilidade deste comprometimento ter sido causado por grandes diferenças entre os modelos de calibração.

Semelhantemente ao que se realizou para o caso dos dados provenientes de uma análise Brix, estudou-se, diferentes classes de modelos, obtidos agora a partir dos dados submetidos ao método polarimétrico, que correspondem aos modelos apodizados em diferentes níveis de incidência. No caso em que o conjunto Pol foi empregado na análise do método da poda-PLS1 avaliou-se os seguintes níveis de concordância entre as variáveis selecionadas: 70%, 80%, 90% e 95%.

A idéia básica de se aplicar níveis de apodização está na flexibilização do modelo, assim, para os níveis de apodização apresentados acima, obteve-se, 76, 40, 22 e 1 variáveis selecionadas, respectivamente.

Assim, construiu-se quatro modelos diferentes, cada um com um número de variáveis selecionadas (76, 40, 22 e 1), a fim de compará-los aos resultados obtidos pelos modelos onde todas as variáveis foram consideradas relevantes.

A comparação foi feita entre a capacidade de previsão dos conjuntos teste, antes e após a poda, nos quatro níveis de apodização. Estes métodos foram comparados através do teste-*F*.

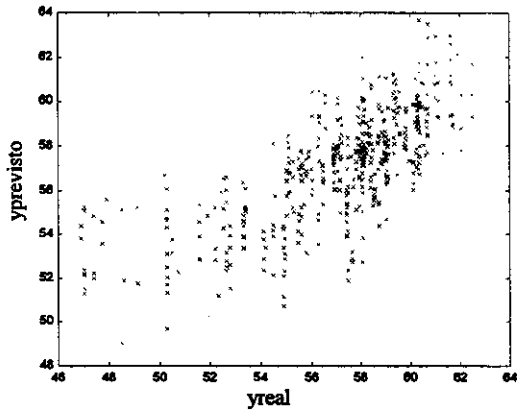
Como neste caso dispõe-se de 10 modelos de calibração para cada um dos níveis de apodização (70%, 80%, 90% e 95%), utilizou-se o mesmo procedimento descrito na seção 2.8 (esquematizado genericamente pela Figura 20. Têm-se, ao final deste processo, oito vetores com 490 elementos (cada um dos 4 níveis de apodização produzia 2 vetores contendo os erros de teste: antes e após o emprego do método da poda-PLS1).

Importantíssimo salientar que os testes estatísticos que comparam dispersões podem ser usados, apenas e tão somente, se os dados assumirem valores que sigam uma distribuição normal.

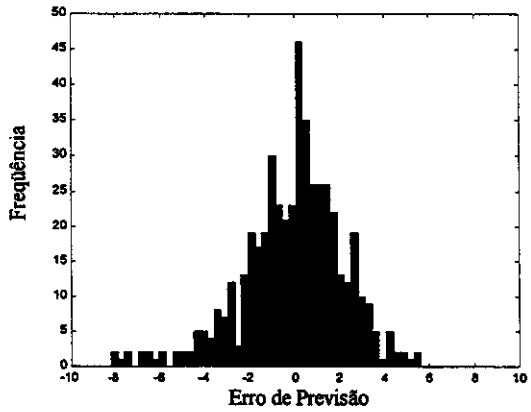
A análise dos resultados consistiu, como ponto fundamental, da comparação dos erros de previsão dos conjuntos teste pelos modelos construídos

quando há, e quando não há seleção de variáveis. A análise numérica não se mostraria muito evidente, assim, decidiu-se apresentar os resultados graficamente.

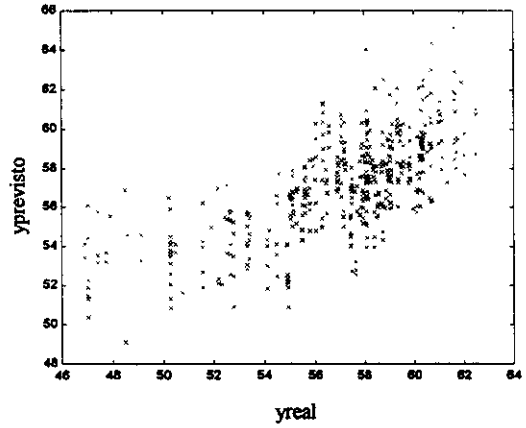
As Figuras 34a até 34j mostram os gráficos dos valores esperados *vs* previstos seguidos dos histogramas dos erros obtidos pelos modelos correspondentes ao método tradicional (empregando todas as variáveis) e aos níveis de incidência de 70%, 80%, 90% e 95% para o conjunto Pol.



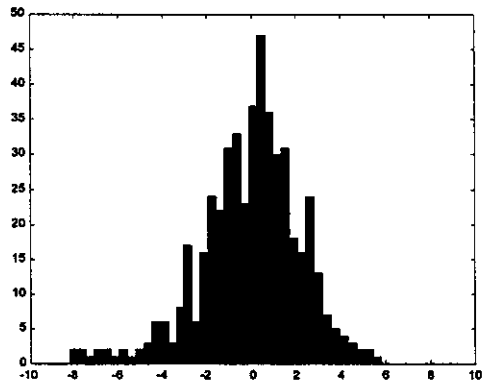
(a) PLS1



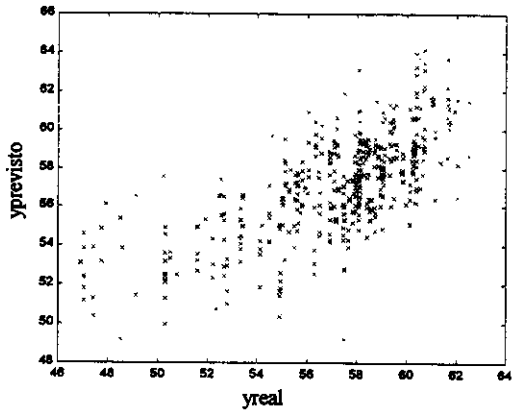
(b) PLS1



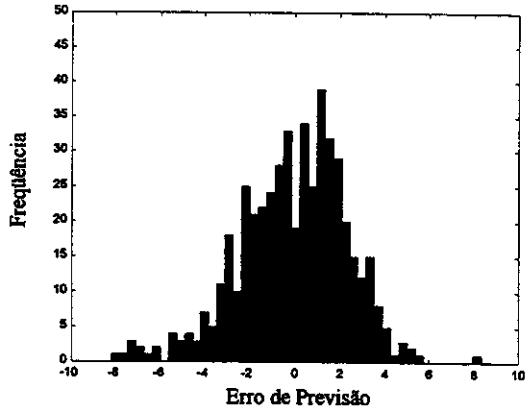
(c) 70%



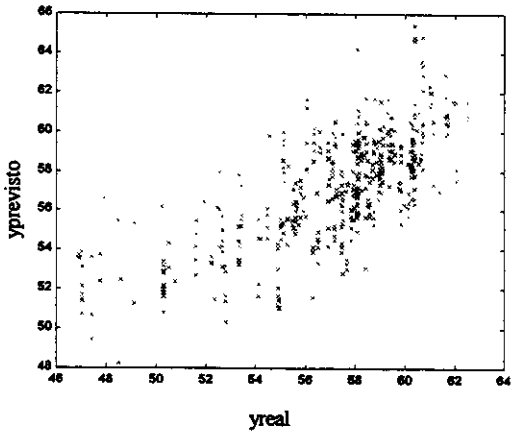
(d) 70%



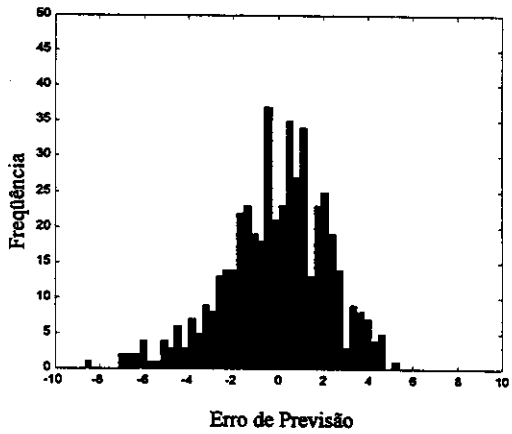
(e) 80%



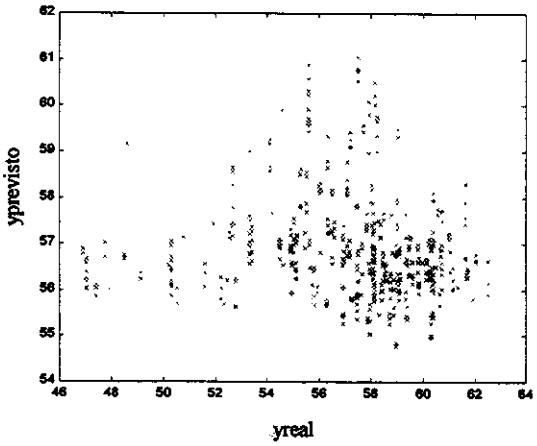
(f) 80%



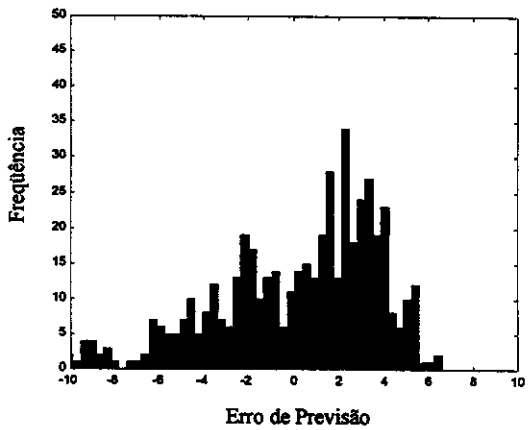
(g) 90%



(h) 90%



(i) 95%



(j) 95%

Figura 34: (a) a (j) Gráficos de valores reais *vs* previstos e respectivos erros de previsão para os métodos: PLS1 (todas as variáveis), poda-PLS1 70%, 80%, 90% e 95%

Nota-se visualmente uma semelhança entre os perfis assumidos pela distribuição dos erros dos métodos tradicional e o da poda-PLS1 ($\mu \approx 0$ e $\sigma \approx 2,4$, para os níveis de incidência de 70% a 90%), indicando que o método da poda-PLS1 (mais simples) gera praticamente os mesmos resultados do método tradicional. No caso em que o nível de incidência é de 95%, que corresponde ao conjunto composto por apenas 1 variável, o perfil da distribuição muda ($\mu \approx 0$ e $\sigma \approx 3,5$), indicando que os erros obtidos para estes modelos podem não ser equivalentes aos do método tradicional.

A aleatoriedade dos erros também pode ser verificada visualmente, pois os erros mostram seguir uma distribuição normal, sendo que as maiores freqüências ocorrem em torno do valor zero. Nota-se ainda que, comparativamente aos resultados obtidos pelo análise do método Brix, os valores dos índices Pol estão mais dispersos, mesmo para o caso em que todas as variáveis são empregadas na construção do modelo de calibração. Espera-se, portanto, que os modelos construídos a partir de dados polarimétricos, gerem resultados menos precisos que aqueles dados obtidos pelo método Brix.

Satisfeita a condição de aleatoriedade dos dados, emprega-se o teste-*F*.

No caso em questão deseja-se determinar se há diferença entre o método tradicional e os em estudo, por isso chamaremos de método A o método em estudo e método B o método tradicional quando são utilizadas todas as variáveis na construção do modelo de calibração. Os graus de liberdade nos dois casos podem ser considerados infinitos devido ao grande número de medidas (490).

Comparando-se, o erro de previsão para o conjunto teste obtido através do método tradicional ($SEP_{BPol} = 6,7$) para o conjunto Pol com os modelos construídos nos quatro níveis de apodização, têm-se:

Tabela IV: Resultados comparativos dos métodos tradicional e após seleção de variáveis.

	Método da Interseção para os Conjuntos Testes							
	70%		80%		90%		95%	
	SEP_A^*	$RMSEP_A$	SEP_A^*	$RMSEP_A$	SEP_A^*	$RMSEP_A$	SEP_A^*	$RMSEP_A$
	5,8	2,4	6,2	2,5	7,0	2,6	12	3,5
$F_{v_A,v} = \frac{SEP_A}{SEP_B}$	0,87		0,93		1,0		1,8	

$F_{\infty,\infty}$ = 1,00 [29]; *O índice A corresponde ao método proposto (poda-PLS1), enquanto o índice B corresponde ao método tradicional de análise.

Nesta etapa tomou-se o cuidado de efetuar-se os cálculos levando-se em consideração as precisões adequadas a cada fator, isto é, a precisão máxima dos resultados (previsto) não pode ser maior que a precisão das variáveis dependentes (esperado).

Consultando o valor limite de $F_{\infty,\infty}$ com 95% de confiança, encontramos o valor 1,00 [29].

Assim pode-se dizer que para o conjunto Pol, os modelos construídos com 76, 40 e 22 variáveis apresentam as mesmas dispersões observadas para o método tradicional. Isso indica que reduzindo-se até 22 o número de variáveis o modelo correspondente irá produzir resultados, com 95% de confiança, iguais ao modelo construído com todas as 313 variáveis. O valor médio do erro de previsão, dado pelo RMSEP, para o método tradicional foi obtido como sendo 2,6, verifica-se, portanto, que os valores encontrados para este parâmetro quando se trata dos modelos simplificados (após seleção de variáveis) estão bem próximos. O modelo construído com 22 variáveis, representando uma redução de 93% do número de variáveis, é estatisticamente equivalente ao modelo construído tradicionalmente em métodos multivariados, que levam em consideração todas as variáveis independentes.

A Figura 35 mostra as variáveis selecionadas.

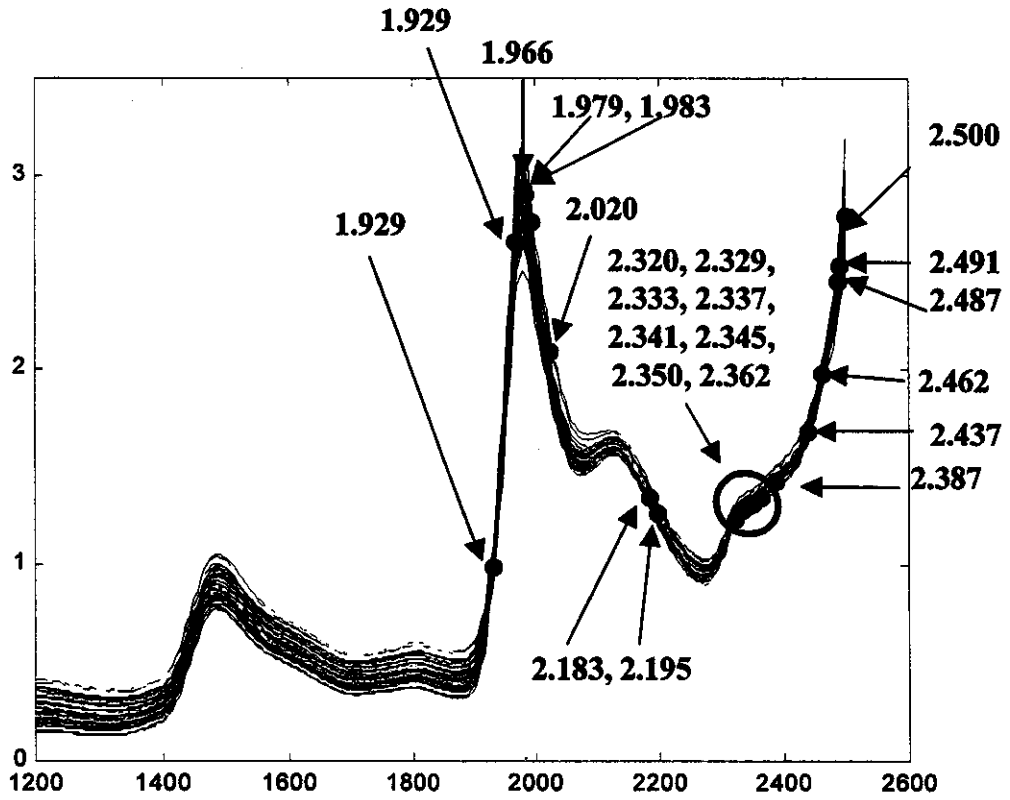


Figura 35: Valores e localização no espectro dos 24 comprimentos de onda selecionados pelo método da poda-PLS1

Verifica-se que as variáveis selecionadas, neste caso, apresentam boa concordância comparada às variáveis selecionadas para o caso das amostras analisadas pelo método Brix. Observando as Figuras 35 e 29, pode-se notar que as variáveis selecionadas nos dois casos dispõem-se praticamente na mesma região. Isto ocorre pois é justamente nestas regiões que ocorrem importantes e características absorções da radiação, na faixa de 1200nm a 2500nm (infravermelho próximo), pelos açúcares que compõem as amostras analisadas. Mais uma vez observa-se que estes comprimentos de onda também estão correlacionados a sobretons e combinação de bandas das correspondentes às ligações –CH, $-CH_2$, $-CH_3$ e –OH, uma vez que o Pol indica a concentração de açúcares em solução.

3.3. ANÁLISE PELO MÉTODO DOS AÇÚCARES REDUTORES (AR)

Neste método a determinação dos açúcares presentes em 193 amostras de xarope de cana-de-açúcar deu-se indiretamente, através da determinação das concentrações dos açúcares redutores, isto é, de todos os monossacarídeos presentes na amostra. O açúcar redutor pode ser oxidado através de reagentes brandos, como o Reagente de Tollens ou Fehling [36], resultando em ácidos aldônicos ou cetônicos [37], que podem ser quantificados por titulação de neutralização.

3.3.1. PRÉ-PROCESSAMENTO

Seguindo os mesmos passos apresentados para o caso anterior, determinou-se para este outro conjunto de amostras – Analisadas pelo Método AR – os *outliers*, para que se pudesse excluí-las. A Figura 36 indica os *Leverages* e os Resíduos para as 193 amostras analisadas pelo método AR.

As amostras eliminadas do conjunto foram: 04; 21; 31; 35; 36; 41; 45; 71; 75; 89; 121; 131; 149; 151; 165, pois estas indicavam, com um intervalo de confiança de 95%, ter comportamento anômalo. O número de amostras mantidas, foi, portanto, 178.

Neste caso o número de variáveis também foi reduzido pela metade, ou seja, cada amostra seria composta por uma coleção de 313 comprimentos de onda.

Ao final do processamento chegou-se, portanto, a um conjunto de 178 amostras de 313 variáveis cada.

A Figura 37 mostra os espectros mantidos.

A separação destas amostras nos conjuntos calibração, validação e teste deu-se usando-se o mesmo critério empregado para o conjunto Pol. A Figura 38 ilustra estas distribuições para um dos modelos construídos.

A construção do modelo de calibração foi feita da mesma forma que para os conjuntos Brix e Pol, isto é, utilizou-se o conjunto de modelagem para

construção do modelo (PLS1), validando-o com o conjunto de validação, enquanto o conjunto teste foi empregado para validação externa.

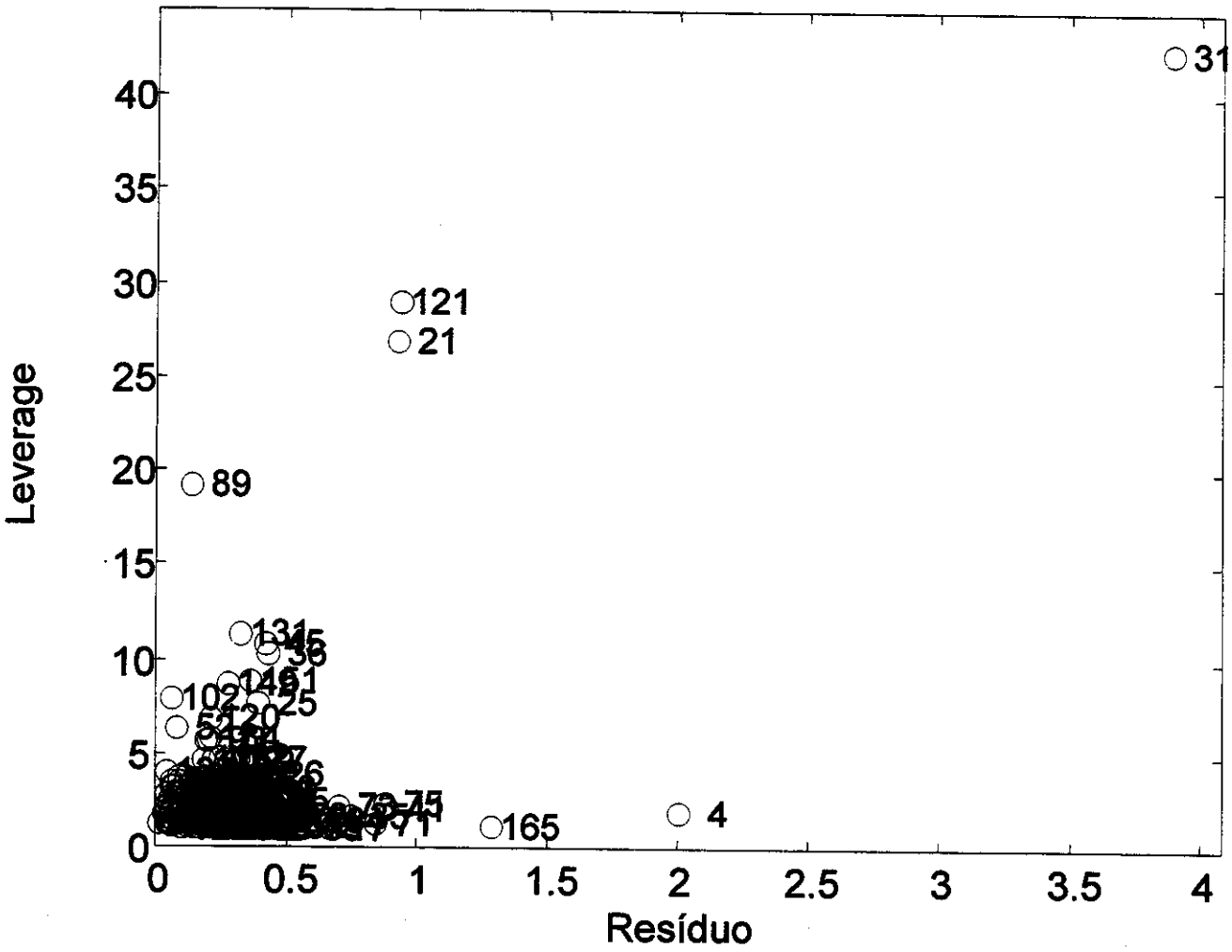


Figura 36: Gráfico dos resíduos *vs* Leverage para as amostras de xarope cana-de-açúcar analisadas pelo método AR.

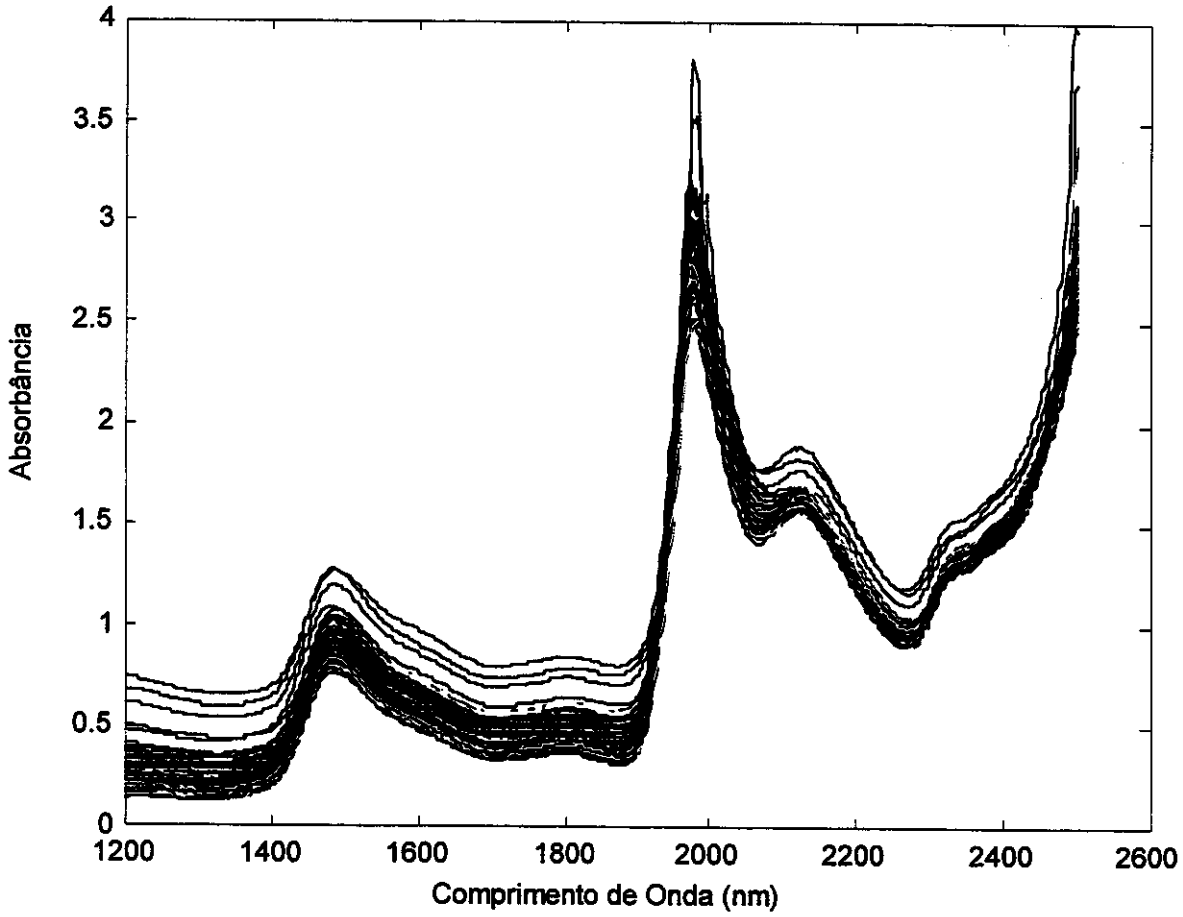


Figura 37: Espectros das amostras de cana-de-açúcar adquiridos medindo-se as absorbâncias, analisadas pelo método AR.

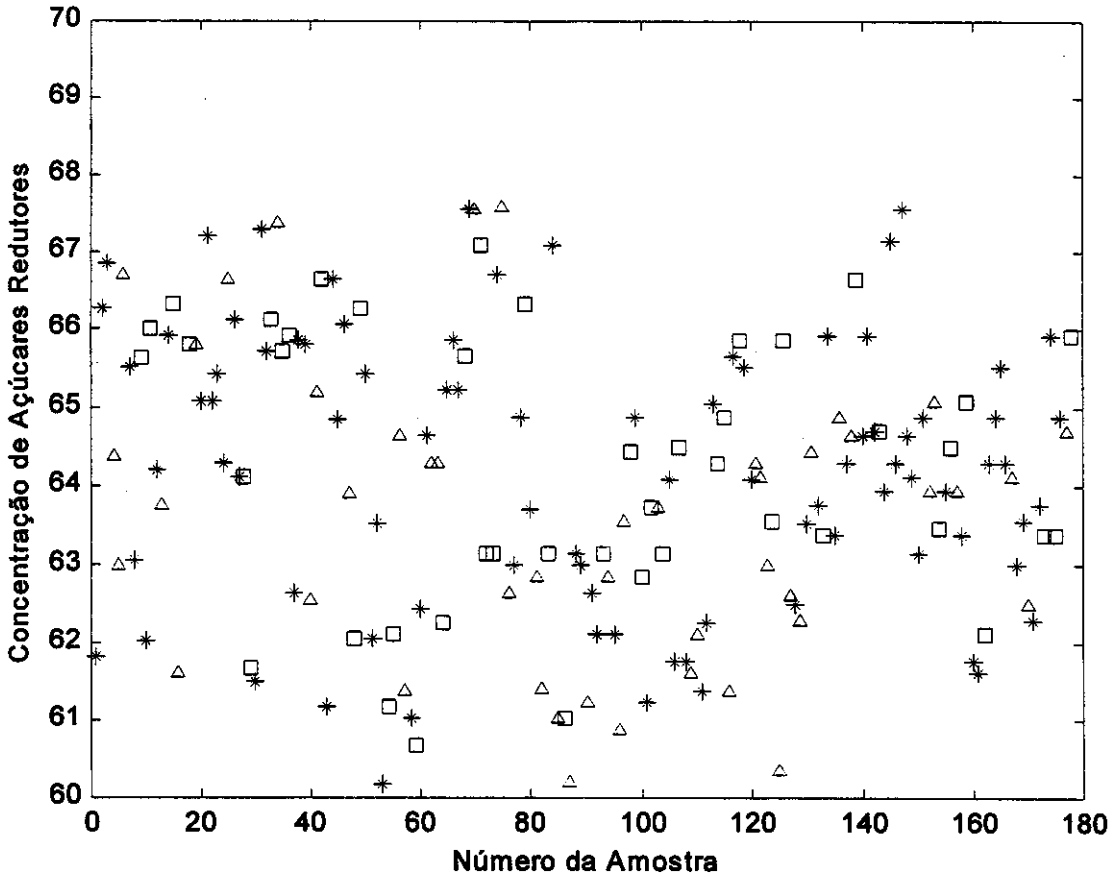


Figura 38: Distribuição das medidas das concentrações de açúcares redutores associadas aos espectros de caldo de cana-de-açúcar concentrado em xarope, sendo: * – modelagem; △ – previsão e □ – teste.

3.3.2. RESULTADOS

Empregou-se as mesmas análises descritas até aqui, sem que houvesse a necessidade de variar a metodologia. Portanto serão gerados dez conjuntos de dados equivalentes, por meio de uma reorganização das amostras, além de se empregar a metodologia da interseção de conjuntos para identificar, de maneira mais genérica, as variáveis informativas.

3.3.3. Aplicação do Método da Poda-PLS1

Monitorando-se os erros de previsão para o conjunto de validação pode-se construir o gráfico mostrado na Figura 39. Verifica-se que o erro de previsão absoluto mínimo ocorre quando o número de variáveis que compõe o modelo de calibração é igual a 3, representando, portanto, uma significativa diminuição no número de variáveis.

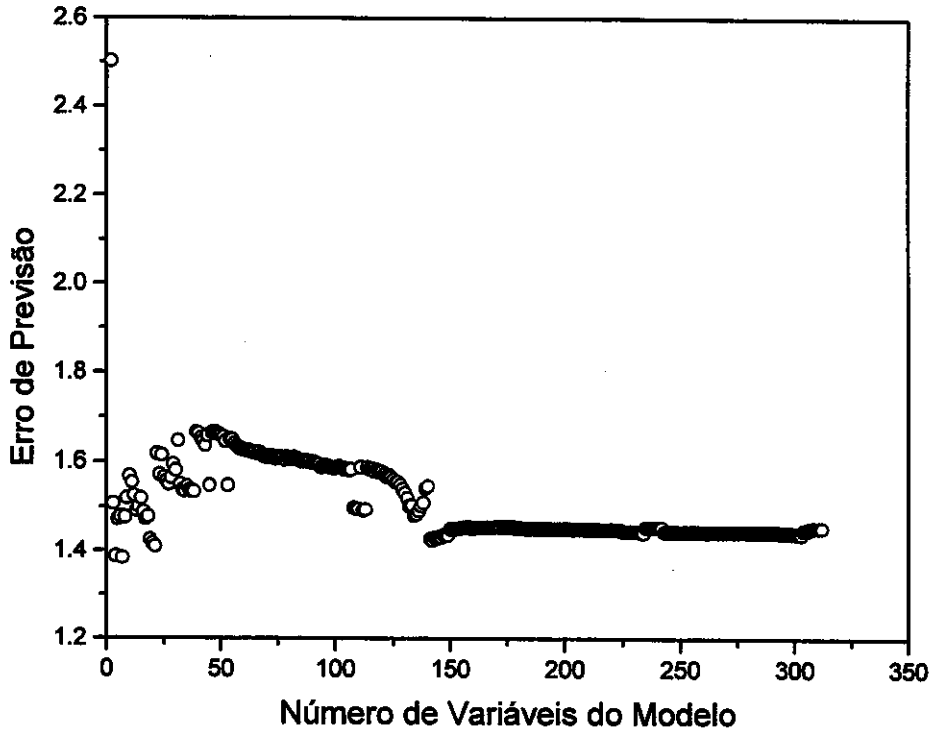


Figura 39: Erros de previsão do conjunto de validação *vs* número de variáveis que compõem os modelos do conjunto AR.

3.3.4. Análise de Repetibilidade

A partir do conjunto original de espectros adquiridos, cujas quantidades de açúcar foram determinadas pelo método dos açúcares redutores, procedeu-se a

reorganização das amostras do conjunto AR, de modo a gerar dez modelos distintos.

3.3.5. ORGANIZAÇÃO DAS AMOSTRAS

Uma vez que as amostras foram reorganizadas, elas produziram dez conjuntos de modelagem, previsão e teste, distintos, sendo que o número de amostras que compunha cada um destes conjuntos foi 89, 45 e 44, respectivamente.

Como os valores extremos das concentrações observadas para a coleção das amostras estavam restritos a uma faixa estreita ($\mu_{AR} = 64 \pm 2$), fornece bons indícios de que as 178 amostras, utilizadas para construção dos modelos, devem representar uma boa estimativa do universo possível.

3.3.6. CARACTERÍSTICAS DOS MODELOS

Todos os modelos foram construídos através do cálculo do vetor de regressão calculado pelo método PLS1, sendo que o número de variáveis latentes usados durante este processamento era sempre o que produzia o menor erro de previsão para seus respectivos conjuntos de validação. A avaliação do desempenho de deu através do monitoramento dos erros de previsão obtidos durante o processo de validação externa.

Da mesma forma que foi feita para os casos dos dados provenientes das análises Brix e Pol, estudou-se, diferentes classes de modelos, obtidos agora a partir dos dados submetidos ao método dos Açúcares Redutores (AR), que correspondem aos modelos apodizados em diferentes níveis de incidência, onde foram empregados os seguintes níveis de concordância entre as variáveis selecionadas: 40%, 50% e 60%. Estes níveis possibilitaram selecionar 46, 24 e 4 variáveis, respectivamente.

Utilizando estes três conjuntos de variáveis construiu-se novos modelos de calibração, cada um composto pelas variáveis selecionadas (46, 24 e 4), a fim de que se pudesse comparar suas capacidades de previsão com a capacidade de

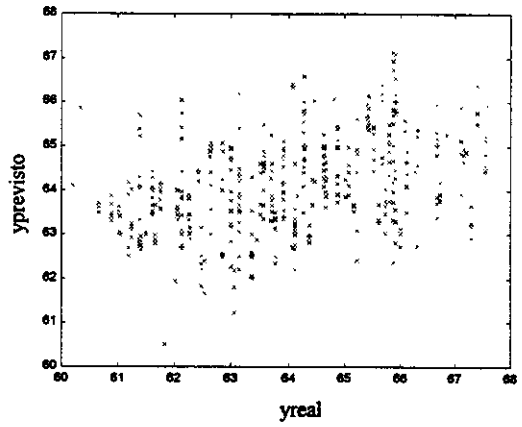
previsão daquele modelo em que todas as variáveis são empregadas para proceder a construção do modelo de calibração.

Comparou-se, então, a capacidade de previsão dos conjuntos teste, antes e após a poda, nestes três níveis de apodização, através do teste-*F*.

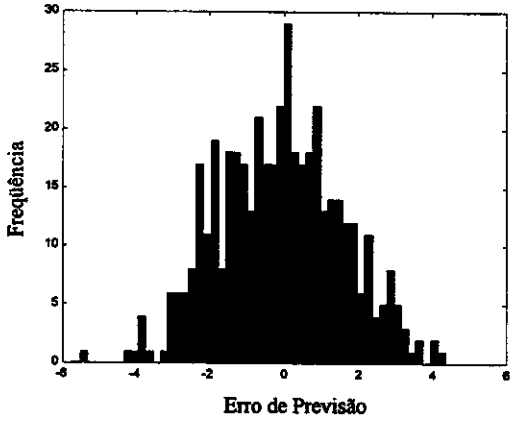
Têm-se dez modelos de calibração para cada nível de apodização (40%, 50% e 60%), que produziram seis vetores com 490 elementos (cada um dos 3 níveis de apodização produzia 2 vetores contendo os erros de teste: antes e após o emprego do método da poda-PLS1). Esta consideração já foi realizada nos processamentos dos resultados simulados, Brix e Pol (Figura 20).

Os resultados das comparações são qualitativamente verificados através da análise de representações gráficas.

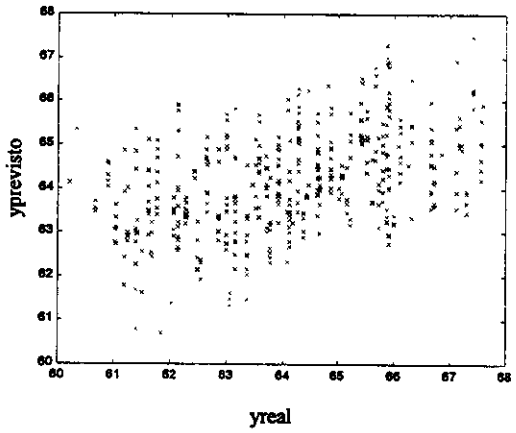
As Figuras 40a até 40h mostram, portanto, os gráficos dos valores esperados *vs* previstos seguidos dos histogramas dos erros obtido pelos modelos correspondentes ao método tradicional (empregando todas as variáveis), e aos níveis de incidência de 40%, 50% e 60% para o conjunto AR.



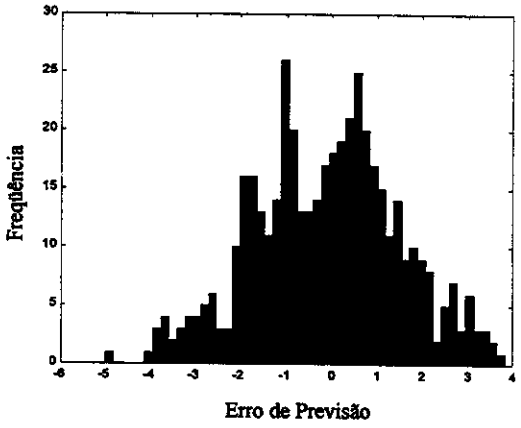
(a) PLS1



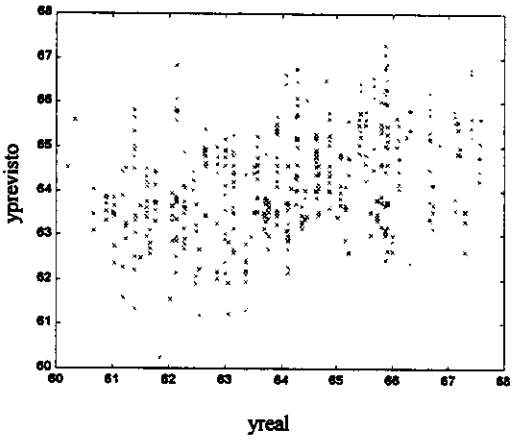
(b) PLS1



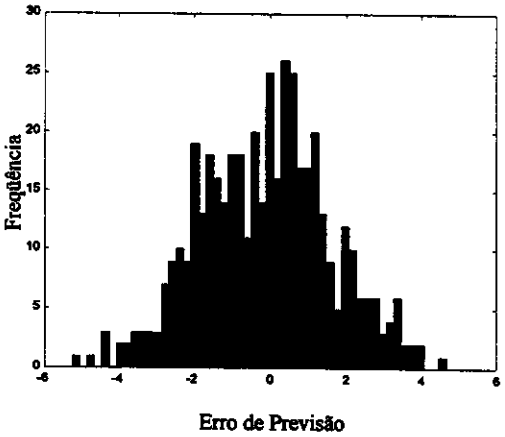
(c) 40%



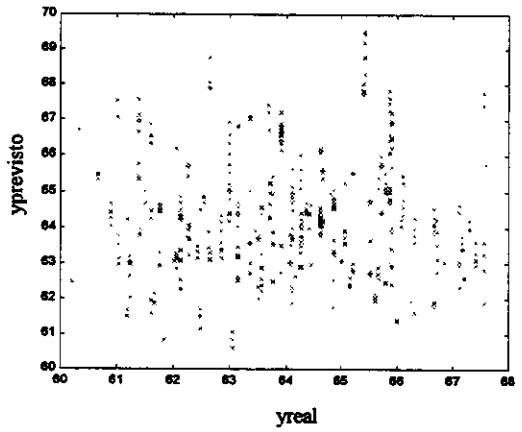
(d) 40%



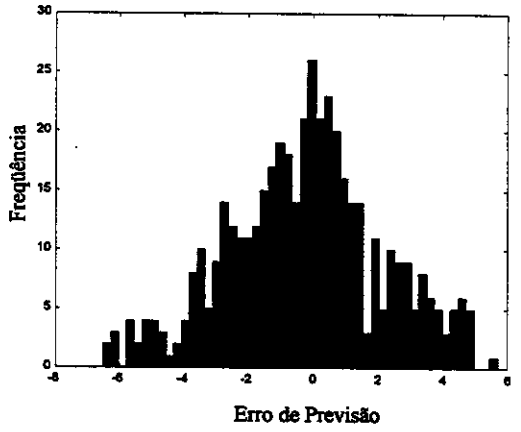
(e) 50%



(f) 50%



(g) 60%



(h) 60%

Figura 40: (a) a (h) Gráficos de valores reais *vs* previstos e respectivos erros de previsão para os métodos: PLS1 (todas as variáveis), poda-PLS1 40%, 50% e 60%.

A verificação visual permite observar que os perfis das distribuições dos erros de previsão dos métodos tradicional e da poda-PLS1 ($\mu \approx 0$ e $\sigma \approx 1,7$) tendem a exibir uma distribuição normal. Esta verificação, porém, não é facilmente identificada e parece que para 60% os erros não seguem uma distribuição normal. Verifica-se que os perfis apresentados, entre os métodos tradicional e da poda-PLS1, aproximam-se bastante, dando, assim, indícios de que os modelos com menos variáveis geram praticamente os mesmos resultados que o modelo com todas as variáveis. No caso em que o nível de apodização é 60%, têm-se uma situação em que além da distribuição não parecer ser exatamente normal, os resultados mostram uma dispersão (desvio padrão) ainda maior que nos casos anteriores. Conclui-se, assim, que houve um potencial comprometimento em considerar-se que os referidos desvios fossem realmente aleatórios e, neste nível de apodização não será realizado as análises estatísticas comparativas, uma vez que estas somente podem ser empregadas se for satisfeita a condição de aleatoriedade dos dados, caracterizada pela distribuição normal.

A fim de se proceder às comparações numéricas, será chamado de método A o aquele que corresponde método empregando seleção de variáveis, e método B o método tradicional (com todas as variáveis). Os graus de liberdade nos dois casos podem ser considerados infinitos devido ao grande número de medidas (450).

Comparando-se, o método tradicional (SEP_{BAR} = 2,6 para o conjunto teste) com os modelos construídos nos quatro níveis de incidência, têm-se:

Tabela V: Resultados comparativos dos métodos tradicional e após seleção de variáveis.

	Método da Interseção para os Conjuntos Testes			
	40%		50%	
	SEP_A^*	$RMSEP_A$	SEP_A^*	$RMSEP_A$
	2,8	1,7	3,0	1,7
$F_{\nu_A,\nu} = \frac{SEP_A}{SEP_B}$	1,1		1,2	

$F_{\infty,\infty}$ = 1,00 [29]; *O índice A corresponde ao método proposto (poda-PLS1), enquanto o índice B corresponde ao método tradicional de análise.

Conclui-se que para o conjunto AR, apenas os modelos construídos com 46 variáveis apresentam as mesmas dispersões observadas para o método

tradicional. Isso corresponde a dizer que é possível reduzir em cerca de 85% o número de variáveis e, ainda sim, produzir resultados, com 95% de confiança, iguais ao modelo construído com todas as 313 variáveis. Para o conjunto AR um número maior de variáveis foi necessário para se chegar a erros da mesma ordem dos encontrados nos modelos tradicionais. Isto deu-se devido à grande imprecisão do método AR, uma vez que se utilizam medidas indiretas para determinar a quantidade do analito de interesse. Calculando-se o RMSEP para o método tradicional ($RMSEP_{PLS1}$ = 1,6), pode-se comparar com aqueles produzidos pelos modelos simplificados pela seleção de variáveis ($RMSEP_{poda\text{-}PLS1}$ = 1,7). Nota-se que há boa concordância com estes valores.

As variáveis selecionadas neste caso podem ser vistas na Figura 41.

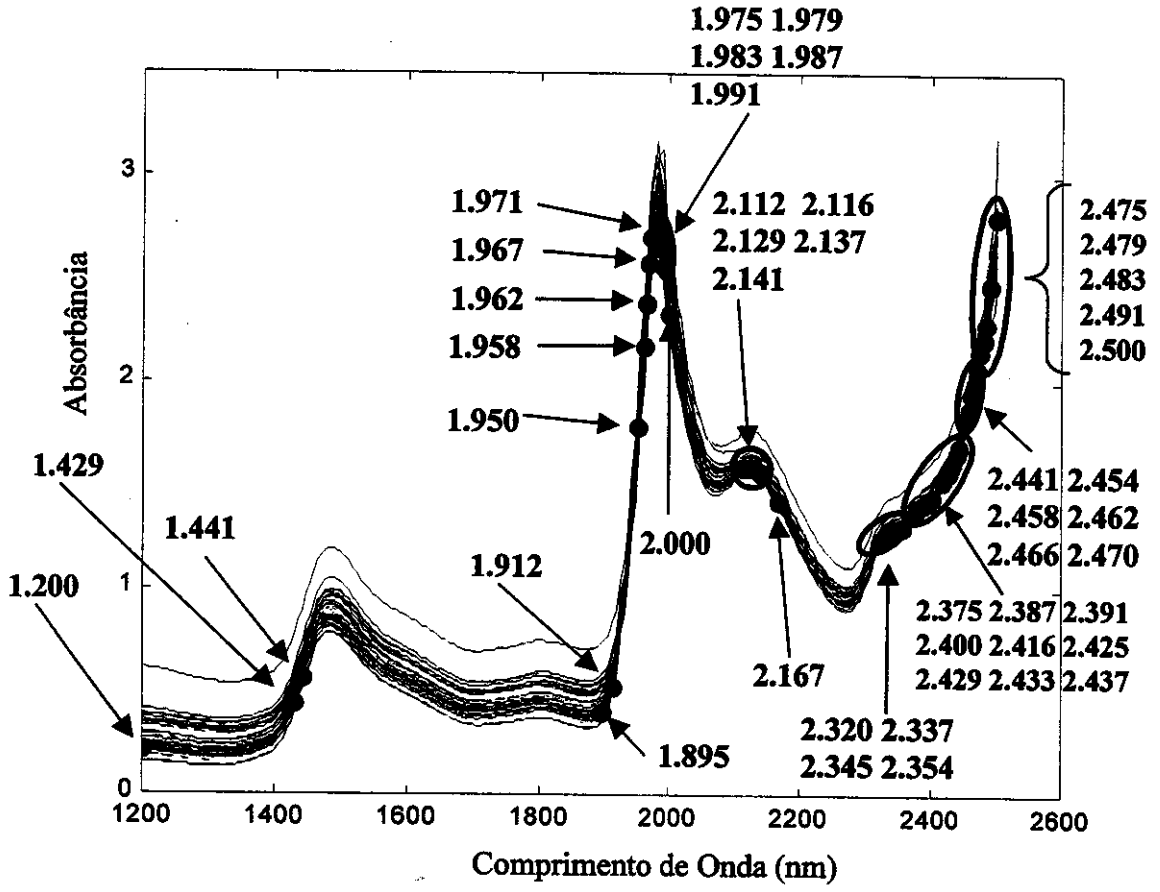


Figura 41: Valores e localização no espectro dos 46 comprimentos de onda selecionados pelo método da poda-PLS1

Aqui, mais uma vez, as variáveis selecionadas aparecem em regiões muito similares àquelas observadas para o método Brix e Pol. Como neste caso, o método padrão de análise (Açúcares Redutores) é mais impreciso, foi necessário um maior número de comprimentos de onda comparados aos outros dois métodos (Brix e Pol). Alguns comprimentos de onda ocorrendo na região do espectro compreendida entre os valores de 1890–2000nm, referete à banda de água provinda da solução, também foram selecionados neste caso. Verifica-se no entanto que os valores dos coeficientes de regressão, gerados durante a construção do modelo de calibração, possuem valores baixos para essas variáveis (cerca de 10% dos valores médios), indicando assim uma pequena influência no resultado produzido pelos modelos.

CONCLUSÕES

Através deste trabalho foi possível avaliar o desempenho de um novo método de seleção de variáveis, que baseia-se na análise da Hessiana dos erros de previsão obtidos durante a etapa de validação externa.

Os resultados obtidos para dados reais (amostras de caldo de cana-de-açúcar e amostras de xarope de açúcares) mostraram-se muito satisfatórios, comprovando a potencialidade deste método, uma vez que, apesar de reduzir-se em média 90% o número de variáveis, os modelos construídos após seleção de variáveis foram equivalentes aos modelos construídos pelo método PLS1 – solidamente estabelecido, sendo o mais amplamente empregado em calibração multivariada em espectroscopia.

A efetiva redução no número de variáveis tem impacto profundo em análises laboratoriais de controle de qualidade em processos industriais. Neste caso específico, verifica-se que o preço pago pela industria pela tonelada de cana fornecida pelos produtores, depende da quantidade de açúcar que aquela cana detém. Assim, quanto mais rápidas e mais precisas forem as análises para detecção desta quantidade melhores as relações entre produtores e indústrias.

Outro fator importante alcançado através da redução do número de variáveis nos modelos de calibração está na possibilidade de diminuir número de amostras necessárias ao emprego de métodos de calibração mais robustos, como é o caso da Regressão Linear Múltipla (MLR), uma vez que este método exige um número de amostras igual ao número de variáveis. Assim, restringindo-se o número de variáveis através do método de seleção descrito neste trabalho, obviamente, reduz-se o número de amostras exigidas para construção do modelo de calibração MLR.

Verificou-se que os erros de previsão obtidos para os três conjuntos de amostras analisados dependiam diretamente da precisão dos métodos empregados para determinação da quantidade de açúcares (Porcentagem Brix, Índice Polarimétrico e Açúcares Redutores), sendo, portanto, maiores para o caso em que foram empregados métodos mais imprecisos.

A imprecisão dos métodos refletiu na quantidade de variáveis selecionadas em cada caso. Como resultado têm-se que as variáveis

selecionadas para os conjuntos de amostras analisadas pelo método mais preciso (Brix) mostrou-se muito menor (11 variáveis) do que para o método menos preciso (46 variáveis no método dos Açúcares Redutores). O método polarimétrico apresentou desvios com valores intermediários, implicando na seleção de 24 variáveis. Apesar das imprecisões dos métodos padrões de análise e dos diferentes números de variáveis selecionadas, observa-se que as variáveis importantes ao modelo de calibração estão sempre nas mesmas regiões do espectro, independentemente do método analítico empregado. Estas variáveis estão correlacionadas com propriedades químicas intrínsecas à matéria (açúcares) presentes nas amostras, isto é, as absorbâncias na região do infravermelho próximo. Isto também pôde ser confirmado, através da verificação dos resultados obtidos por Costa Filho e Poppi [38], que empregou um outro método de seleção de variáveis (Algoritmo Genético [39]) em espectros adquiridos para misturas de açúcares na região do infravermelho próximo. Os comprimentos de onda selecionados estavam nas mesmas regiões aqui observadas.

Os modelos de calibração em que se empregaram amostras analisadas pelo método Brix mostraram-se mais precisos (desvio padrão dos erros de previsão = 0,4), mas não o mais exato. Para obter-se a exatidão dos métodos, tomou-se os valores correspondentes de RMSEP, que caracterizam o erro absoluto médio, e dividiu-os pelo valor médio das medidas realizadas por cada um dos método padrões, chegando-se a um parâmetro que define o Erro Médio Relativo (EMR).

As exatidões dos modelos de calibração foram: EMR_{Brix} = 0,073; EMR_{Pol} = 0,045 e EMR_{AR} = 0,031. Verifica-se, portanto que apesar dos modelos construídos a partir do método Brix serem os mais precisos, são os menos exatos Por outro lado, os modelos construídos a partir do método dos Açúcares Redutores, são menos precisos, porém mais exatos.

As análises demandaram um considerável tempo de processamento, mas mostraram-se muito robustas, pois elegeram o conjunto de variáveis que realmente contêm informações importantes sobre a amostra. Porém, depois que a seleção é realizada, os cálculos tornam-se extremamente mais rápidos. Tal

conjunto mostrou-se genérico sendo capaz de prever satisfatoriamente amostras desconhecidas (como foi ocaso das amostras teste).

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Tanaka, M.; Kojima,T.: "*Near-Infrared Monitoring of the Growth Period of Japanese Pear Fruit Based on Constituent Sugar Concentrations*", *J. Agric. Food Chem.* **44** (1996) 2272

[2] Li, W.; Goovaerts, P.; Meurens, M.: "*Quantitative Analysis of Individual Sugars and Acids in Orange Juices by Near-Infrared Spectroscopy of Dry Extracts*", *J. Agric. Food Chem.* **44** (1996) 2252

[3] Mirouze, F. de L.; Boulou, J. C.; Dupuy, N.: "*Quantitative Analysis of Glucose Syrups by ATR/FT-IR Spectroscopy*", *Appl. Spectrosc.* **47** (1993) 1187

[4] Cadet, F.; Offmann, B.: "*Extraction of Characteristic Bands of Sugars by Multidimensional Analysis of their Infrared Spectra*", *Spectroscopy Lett.* **29** (1996) 523

[5] Beebe, K. R.; Kowalski, B. R.: "*An Introduction to Multivariate Calibration and Analysis*", *Anal. Chem.* **59** (1987) 1007A

[6] Geladi, P.; Kowalski, B. R.: "*Partial Least-Squares Regression: A Tutorial*", *Anal. Chim. Acta* **185** (1986) 1

[7] Martens, H.; Næs, T.: Multivariate Calibration, Wiley, New York, 1996

[8] Hassib, B.; Stork, D. G.: "*Second Order Derivative for Network Pruning: Optimal Brain Surgeon*", Advances in Neural Information Processing System, vol. 5, S. J. Hanson, J. D. Cowan and C.L. Giles (eds), Morgan Kaufmann, San Mateo, 1993, 164

[9] Le Cun, Y.; Denker, J. S.; Solla, S. A.: "*Optimal Brain Damage*", Advances in Neural Information Processing System, vol. 2, D. S. Touretzsky (ed.), Morgan Kaufmann, San Mateo,1990, 598

[10] Poppi, R. J.; Massart, D. L.: "*The Optimal Brain Surgeon for Pruning Neural Network Architecture Applied to Multivariate Calibration*", *Anal. Chim. Acta* **375** (1998) 187

[11] Helland, I. S.: "*On the Structure of Partial Least Squares Regression*", *Commun. Statist. – Simula.* **17** (1988) 581

[12] Sekulic, S.; Seasholtz, M. B.; Wang, Z.; Kowalski, B. R.: "*Nonlinear Multivariate Calibration Methods in Analytical Chemistry*", *Anal. Chem.* **65** (1993) 835A

[13] Rossi, D. T.; Desilets, D. J.; Pardue, H. L.: "*Quantitation and Identification of Polynuclear Aromatic-Hydrocarbons by Liquid-Chromatography and Multiwavelength Absorption Spectrometry*", *Anal. Chim. Acta* **161** (1984) 191

[14] Rossi, D. T.; Pardue, H. L.: "*Effects of Wavelength Range on the Simultaneous Quantitation of Polynuclear Aromatic-Hydrocarbons with Absorption-Spectra*", *Anal. Chim. Acta* **175** (1985) 153

[15] Kalivas, J. H.; Roberts, N.; Sutter, J. M.: "*Global Optimization by Simulated Annealing with Wavelength Selection for Ultraviolet Visible Spectrophotometry*", *Anal. Chem.* **61** (1989) 2024

[16] Brown, P. J.: "*Wavelength Selection in Multicomponent Near-Infrared Calibration*", *J. Chemom.* **6** (1992) 151

[17] Leardi, R.: "*Application of a Genetic Algorithm to Feature-Selection under Full Validation Conditions and to Outlier Detection*", *J. Chemom.* **8** (1994) 65

[18] Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M.; Sterna, C.: "*Elimination of Uninformative Variables for Multivariate Calibration*", *Anal. Chem.* **68** (1996) 3851

[19] Osborne, S. D.; Jordan, R. B.; Künnemeyer, R.: "*Method of Wavelength Selection for Partial Least Squares*", *Analyst* **122** (1997) 1531

[20] Rumelhart, D. E.; McClelland, J. L.; (PDP Research Group): Parallel Distributed Processing: explorations in the microstructure of cognition, MIT editor, Cambridge, 1986

[21] Jang, J.; Sun, C.; Mizutani, E.: Neuro-Fuzzy and Soft Computing, Prentice-Hall, Upper Saddle River, 1997

[22] Hinton, G. E.: "*Connectionist Learning Procedures*", *Artiff. Intell.* **40** (1989) 185

[23] Weigend, A. S.; Rumelhart, D. E.; Huberman, B. A.: "*Generalization by Weight-Elimination with Application to Forecasting*", Advances in Neural Information Processing System, vol. 3, D. S. Touretzsky (ed.), Morgan Kaufmann, San Mateo, 1991, 875

[24] Messick, N. J.; Kalivas, J. H.; Lang, P. M.: "*Selecting Factors for Partial Least Squares*", *Microchem. J.* **55** (1997) 200

[25] Gemperline, P. J.; Long, J. R.; Gregoriou, V. G.: "*Nonlinear Multivariate Calibration Using Principal Components Regression and Artificial Neural Networks*", *Anal. Chem.* **63** (1991) 2313

[26] Spiegel, M. R.: Estatística, McGraw Hill (ed), São Paulo, 1984

[27] Neto, B. B.; Scarminio, I. S.; Bruns, R. E.: Planejamento e Otimização de Experimentos, Unicamp (ed.), Campinas, 1996

[28] Bussab, W. O.; Morettin, P. A.: Estatística Básica, Atual (ed), São Paulo, 1987

[29] Box, G. E.; Hunter, W. G.; Hunter, J. S.: Statistics for Experimenters. An Introduction to Design, Data Analysis and Model Building, Wiley, New York, 1978

[30] Hart, F. L.; Fisher, H. J.: Modern Food Analysis, Springer-Verlag, New York, 1971

[30] Adams, M. J.: Chemometrics in Analytical Spectroscopy, RSC Analytical Spectroscopy Monographs, N. Barnett (ed.), Deaking University, Victoria, 1995

[32] Haaland, D. M.; Thomas, E. V.: "*Partial Least-Squares Methods for Spectral Analysis: 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information*", *Anal. Chem.* **60** (1988) 1193

[33] Skoog, D. A.; West, D. M.; Holler, F. J.: Fundamentals of Analytical Chemistry, 6th Ed., Saunders, New York, 1992

[34] Workman Jr., J. J.: "*Interpretative Spectroscopy for Near Infrared*", *Appl. Spectrosc. Rev.* **31** (1996) 251

[35] Meade, G. P.; Chen, J. C. P.: Sugar Cane Handbook, 11th ed., Wiley, New York, 1985

[36] Morrison, R. T.; Boyd, R. N.: Organic Chemistry, 6th ed., Prentice-Hall, Englewood Cliffs, 1992

[37] Voet, D.; Voet, J. G.: Biochemistry, 2nd ed., Wiley, New York, 1995

[38] Costa Filho, P. A.; Poppi, R. J.: "*Use of Near Infrared Spectroscopy for Rapid Estimation of Sugar Cane Juice Quality Components*", Near Infrared Spectroscopy: Proceedings of the 9th International Conference, A. M. C. Davies and R. Giangiacomo editors, NIR Publications, Chichester, 2000, 897

[39] Costa Filho, P. A.; Poppi, R. J.: "*Algoritmo Genético em Química*", *Quim. Nova* **22** (1999) 405

APÊNDICE

A. Cálculo da inversa da Matriz Hessiana

A inversa da Matriz Hessiana ($\mathbf{H}^{-1}$) é fundamental para a formulação do procedimento do OBS. Quando o número de parâmetros livres, $\boldsymbol{B}$, na rede é muito grande, o problema para computar $\mathbf{H}^{-1}$ pode ser intratável. Será descrito a seguir um procedimento para realizar este cálculo, assumindo que a rede inicial já está totalmente treinada, o que representaria dizer que alcançou-se um mínimo na superfície de erros.

Para simplificar a representação, será suposto que a arquitetura da rede possui apenas um único neurônio de saída. Assim, para um dado conjunto de treinamento, pode-se expressar a função erro como:

$$E(\boldsymbol{b}) = \frac{1}{2N}\sum_{n=1}^{N}(d(n)-s(n))^2 \qquad \text{(a1)}$$

onde $s(n)$ é a saída atual da rede na apresentação do *n-ésimo* exemplo, $d(n)$ é a correspondente resposta desejada, e N é o número total de exemplos no conjunto de treinamento. A saída $s(n)$ pode ser expressa como:

$$s(n) = F(\boldsymbol{b},\boldsymbol{x}) \qquad \text{(a2)}$$

onde F é a função de mapeamento de entrada-saída realizada através das diversas camadas de neurônios até a saída, $\boldsymbol{x}$ é o vetor de entrada e $\boldsymbol{b}$ o vetor de pesos sinápticos da rede. A primeira derivada de $E(\boldsymbol{b})$ com relação a $\boldsymbol{b}$ é portanto:

$$\frac{\partial E(\boldsymbol{b})}{\partial \boldsymbol{b}} = -\frac{1}{N}\sum_{n=1}^{N}\frac{\partial F(\boldsymbol{b},\boldsymbol{x}(n))}{\partial \boldsymbol{b}}(d(n)-s(n)) \qquad \text{(a3)}$$

e a derivada segunda de $E(\boldsymbol{b})$ com relação a $\boldsymbol{b}$, ou seja, a Matriz Hessiana é:

$$\mathbf{H}(N) = \frac{\partial^2 E(\boldsymbol{b})}{\partial \boldsymbol{b}^2} = \frac{1}{N}\sum_{n=1}^{N}\left\{\left(\frac{\partial F(\boldsymbol{b},\boldsymbol{x}(n))}{\partial \boldsymbol{b}}\right)\left(\frac{\partial F(\boldsymbol{b},\boldsymbol{x}(n))}{\partial \boldsymbol{b}}\right)^{t} - \frac{\partial^2 F(\boldsymbol{b},\boldsymbol{x}(n))}{\partial \boldsymbol{b}^2}(d(n)-s(n))\right\} \qquad \text{(a4)}$$

enfatizando a dependência da Matriz Hessiana com o tamanho da amostra de treinamento, N.

Sob a suposição de que a rede está totalmente treinada, isto é, que a função custo $E(\boldsymbol{b})$ foi ajustada a um mínimo local na superfície de erros, é razoável dizer que $s(n)$ está perto de $d(n)$, assim pode-se ignorar o segundo termo e aproximar a Equação a4 para:

$$\mathbf{H}(N) \cong \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial F(\boldsymbol{b},\boldsymbol{x}(n))}{\partial \boldsymbol{b}}\right)\left(\frac{\partial F(\boldsymbol{b},\boldsymbol{x}(n))}{\partial \boldsymbol{b}}\right)^{t} \qquad \text{(a5)}$$

Para simplificar a notação, define-se um vetor Bx1

$$\xi(n) = \frac{1}{\sqrt{N}}\frac{\partial F(\boldsymbol{b},\boldsymbol{x}(n))}{\partial \boldsymbol{b}} \qquad \text{(a6)}$$

o qual pode ser computacionalmente determinado [1A]. Pode-se, então, reescrever a Equação a5 como:

$$\begin{aligned}\mathbf{H}(n) &= \sum_{k=1}^{n}\xi(k)\xi^{t}(k) \\ &= \mathbf{H}(n-1) + \xi(k)\xi^{t}(k), \qquad n = 1,2,3,\ldots,N\end{aligned} \qquad \text{(a7)}$$

Este recurso está na forma adequada para a aplicação da chamada inversão de matriz *lemma*.[2A].

Fazendo **A** e **B** denotar duas matrizes positivas definidas, relacionadas por:

$$\mathbf{A} = \mathbf{B}^{-1} - \mathbf{C}\mathbf{D}\mathbf{C}^{t} \qquad \text{(a8)}$$

onde **C** e **D** são outras duas matrizes. De acordo com a inversão de matriz *lemma*, a invesa da matriz **A** é definida por:

$$\mathbf{A}^{-1} = \mathbf{B} - \mathbf{B}\mathbf{C}(\mathbf{D} + \mathbf{C}^{t}\mathbf{B}\mathbf{C})^{-1}\mathbf{C}^{t}\mathbf{B} \qquad \text{(a9)}$$

A partir da expressão descrita na equação a7, têm-se:

$\mathbf{A} = \mathbf{H}(n)$

$\mathbf{B}^{-1} = \mathbf{H}(n-1)$

$\mathbf{C} = \xi(n)$

$\mathbf{D} = 1$

A aplicação da inversão de matriz *lemma* produz, portanto, a fórmula desejada para computar recursivamente a inversa da Hessiana:

$$\mathbf{H}^{-1}(n) = \mathbf{H}^{-1}(n-1) - \frac{\mathbf{H}^{-1}(n-1)\xi(n)\xi^{t}(n)\mathbf{H}^{-1}(n-1)}{1+\xi^{t}(n)\mathbf{H}^{-1}(n-1)\xi(n)} \qquad \text{(a10)}$$

Note-se que o denominador da equação a10 é um escalar, portanto, o cálculo de sua recíproca é obtida diretamente, assim, dado um valor anterior para a inversa da Matriz Hessiana, $\mathbf{H}^{-1}(n-1)$, pode-se computar seu valor atualizado $\mathbf{H}^{-1}(n)$ na apresentação do *n-ésimo* exemplo representado pelo vetor $\xi(n)$. Esta computação recursiva continua até que todo o conjunto dos N exemplos tenha sido considerado. A inicialização do algoritmo exige que se faça $\mathbf{H}^{-1}(0)$ grande [3A], uma vez que está sendo constantemente reduzido, de acordo com a equação a10. Este requisito é satisfeito se:

$$\mathbf{H}^{-1}(0) = \delta^{-1}\mathbf{I} \qquad \text{(a11)}$$

Onde δ é um pequeno número positivo e $\mathbf{I}$ é a matriz identidade. Esta forma de inicialização garante que $\mathbf{H}^{-1}(n)$ é sempre positiva definida [4A]. O efeito de δ torna-se progressivamente menor à medida que mais e mais exemplos são apresentados à rede.

REFERÊNCIAS BIBLIOGRÁFICAS DO APÊNDICE

[1A] Saarinen, S.; Bramley, R.; Cybenko, G.: "*Neural Networks, Backpropagation, and Automatic Differentiation*", Automatic Differentiation of Algorithms: Theory, Implementation, and Application, A. Griewank and G. F. Corliss (eds), Philadelphia, 1992, 31

[2A] Tylavsky, D. J.; Sohie, G. R. L.: "*Generalization of the Matrix-Inversion Lemma*", *Proceeding of the IEEE* **74** (1986) 1050

[3A] Haykin, S.: Neural Networks: a comprehensive foundation, MacMillan, New York, 1994

[4A] Beale, E. M. L.: Introduction to Optimization, Interscience in Discrete Mathematics and Optimization, Wiley , Chichester, 1988