UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE QUÍMICA

DEPARTAMENTO DE FÍSICO-QUÍMICA

UTILIZAÇÃO DE MÉTODOS QUIMIOMÉTRICOS EM DADOS DE NATUREZA MULTIVARIADA

THAIS FERNANDA PARREIRA

Dissertação apresentada ao Instituto de Química
como parte dos requisitos para obtenção
do título de Mestre em Química

Orientadora: Profa. Dra. Márcia Miguel Castro Ferreira

Campinas - SP

Julho – 2003

UNIDADE IQ
Nº CHAMADA Unicamp P248u
V ____ EX ____
TOMBO BC/ 58698
PROC. 16.127.04
C ☐ D ☒
PREÇO 11,00
DATA 29-06-04
Nº CPD ____

CM00198293-1
Bibid: 317484

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO INSTITUTO DE QUÍMICA UNICAMP

P248u

Parreira, Thais Fernanda.
Utilização de métodos quimiométricos em dados de natureza multivariada / Thais Fernanda Parreira. -- Campinas, SP: [s.n], 2003.

Orientadora: Márcia Miguel Castro Ferreira

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

1. Quimiometria. 2. Análise exploratória. 3. Calibração multivariada. 4. NIR. I. Ferreira, Márcia Miguel Castro. II. Universidade Estadual de Campinas. III. Título.

AGRADECIMENTOS

À Profa. Márcia, pelo incentivo em momentos de desânimo e pelas discussões sobre o trabalho.

À Cássia e aos outros autores do trabalho das tangerinas, ao IAC e ITAL, pelo empréstimo dos dados, bem como ao Henrique e à Henkel pelos dados do óleo de soja.

Aos colegas do grupo, pelo apoio e pela ajuda na elucidação de problemas que às vezes pareciam insolúveis.

À Luciana e à Fabiana, por toda a amizade, pelas longas discussões sobre o trabalho e pelos inúmeros favores prestados na parte burocrática da finalização da dissertacão.

Ao Marlon, Cristiano e Steve, pelas discussões e pela ajuda no desenvolvimento de rotinas no Matlab.

A todos os meus amigos que, muito pacientemente, conviveram comigo estes momentos, nem sempre tão fáceis: Val, Lu, Marcelo, Fabi, Márcia e Adriana.

À Bel da CPG por todo o auxílio prestado principalmente na parte final do trabalho.

À CAPES, pelo auxílio financeiro durante o início do projeto.

Aos meus pais, que sempre apoiaram minhas decisões, mesmo quando não concordavam com elas.

Ao Júlio, por todo apoio, carinho, paciência e dedicação.

E a todos aqueles que, de uma forma ou de outra, colaboraram para a execução e conclusão deste trabalho.

Thais Fernanda Parreira

Formação Acadêmica

Mestrado em Físico Química
Laboratório de Quimiometria Teórica e Aplicada
Instituto de Química – Universidade Estadual de Campinas
Título da Dissertação: Utilização de Métodos Quimiométricos em Dados Multivariados
Conclusão: Julho de 2003.

Bacharel em Química
Instituto de Química – Universidade Estadual de Campinas
Conclusão: Dezembro de 1997.

Publicações

T. F. Parreira, M. M. C. Ferreira, H. J. S. Sales e W. B. de Almeida, "*Quantitative determination of epoxidized soybean oil using near infrared spectroscopy and multivariate calibration*", Applied Spectroscopy, 56 (12), p. 1607-1614, ***2002***.

Trabalhos em Congressos

T. F. Parreira, M. M. C. Ferreira, H. J. S. Sales e W. B. de Almeida, "*Quantitative determination of epoxidized soybean oil using near infrared spectroscopy and multivariate calibration*", França, XXVI European Congress on Molecular Spectroscopy (2002) (apresentação oral).

T. F. Parreira, L. C. Sabino, A. T. Bruni e M. M. C. Ferreira, "*Um estudo QSAR da toxicidade aquática de anilinas e nitrobenzenos*", Caxambu, MG, Brasil, X Simpósio Brasileiro de Química Teórica (1999).

T. F. Parreira, C. R. L. Carvalho, J. V. Castro, M. M. C. Ferreira e P. R. N. Carvalho, "*Análise multidimensional aplicada ao estudo do efeito de desverdecimento da Tangor Murcote*", Campinas, SP, Brasil, III Simpósio Latino Americano de Ciência de Alimentos (1999).

T. F. Parreira, C. R. L. Carvalho, J. V. Castro, M. M. C. Ferreira e P. R. N. Carvalho,T. F. Parreira, "*Estudo Quimiométrico do efeito do desverdecimento na composição química interna da Tangor Murcote*", Santa Maria, RS, Brasil, 10º Encontro Nacional de Química Analítica (1999).

T. F. Parreira, M. M. C. Ferreira, H. J. S. Sales e W. B. de Almeida, "*Calibração Multivariada do óleo de soja epoxidado a partir de espectros na região do infravermelho próximo*", Santa Maria, RS, Brasil, 10º Encontro Nacional de Química Analítica (1999).

Experiência Profissional

Orema Ind. E Com. Ltda
Cargo: Formuladora Sênior
Desenvolvimento de novos produtos no segmento de tintas gráficas e vernizes. Análises físico-químicas instrumentais.
Desde Outubro 2002.

Deltech Control – Instrumentos de Precisão
Cargo: Especialista de Produto
Responsável pela parte técnica de instrumentos de espectroscopia NIR e UV-VIS, refratômetros e polarímetros. Instalação, suporte e treinamentos.
Agosto 2001 – Outubro 2001.

Buckman Laboratórios
Cargo: Estagiária – Pesquisa & Desenvolvimento
Desenvolvimento de novos produtos e novas metodologias analíticas em diversos segmentos da indústria.
Junho 2000 – Julho 2001.

RESUMO

Aplicações de alguns métodos quimiométricos de análise de dados serão mostrados neste trabalho a partir de dados experimentais cedidos pelo IAC (Instituto Agronômico de Campinas) e Henkel Indústrias Químicas. Inicialmente, é apresentada uma breve introdução dos métodos de análise multivariada, quali e quantitativos, os conceitos envolvidos, suas vantagens e desvantagens, com o objetivo de dar uma idéia geral de como os dados fornecidos serão tratados. O primeiro conjunto de dados refere-se ao estudo do processo artificial de desverdecimento aplicado a tangerinas do tipo Murcote. Alguns gases de tratamento foram aplicados para teste em diferentes temperaturas, e análises químicas foram realizadas durante um dado intervalo de tempo, com o objetivo de se avaliar a eficiência dos tipos de tratamentos utilizados. Nesse conjunto de dados, o estudo foi feito utilizando técnicas de análise exploratória multivariada (PCA e HCA), de ordem superior (PARAFAC), além da Análise de Variância (ANOVA), com o objetivo de comparar e avaliar o desempenho de cada uma delas. O segundo conjunto de dados é constituído por espectros de absorbância registrados na região do infravermelho próximo de amostras de óleo de soja epoxidado. Industrialmente, existe a necessidade de quantificar alguns analitos presentes no óleo de soja epoxidado com o objetivo de controlar a qualidade do produto final. Os analitos em questão são a porcentagem de água residual, o índice de iodo e de epóxi, sendo estes dois últimos indicativos da eficiência das propriedades estabilizantes do produto. Dessa forma, estes espectros foram utilizados na construção de modelos de calibração que propiciassem análises muito mais rápidas e econômicas para estes analitos. Utilizou-se para tanto o método PLS de regressão multivariada, explorando também técnicas de seleção de variáveis.

ABSTRACT

Chemometric methods of data analysis are applied in this work to experimental data provided by IAC (Instituto Agronômico de Campinas) and Henkel Chemical Industries. Firstly, a brief introduction of the quali and quantitatives multivariate methods, the concepts involved, their advantages and disadvantages, is presented with the goal to give a general idea on how the provided data will be treated. The first data set referees to the study of the artificial degreening process applied to Murcote tangerines. The fruits were treated with different gases, at different temperatures and chemical analyses were carried out during a time interval, with the goal to evaluate the efficiency of the treatments used. For this data set, the study was done using multivariate exploratory analysis (PCA and HCA), N-Way techniques (PARAFAC) and the analysis of variance (ANOVA). The methods were compared and their performance evaluated. The second data set consisted of absorbance spectra registered in the near infrared region from epoxide soybean oil samples. Industrially, it is necessary to quantify some analytes, present in the epoxide soybean oil, to control the final product quality. These analytes are the residual water percentage and the iodine and epoxide indices, these last two indicating the efficiency of the stabilizing properties of the product. In this way, these spectra were used in the construction of calibration models, saving time and money for chemical determination of the analytes. The multivariate regression method PLS was used, exploring some variables selection techniques as well.

ÍNDICE

LISTA DE ABREVIAÇÕES

ALS – Alternating Least Squares (Quadrados Mínimos Alternados)

ANOVA – Analysis of Variance (Análise de Variância)

AOAC - Association of Oficial Analytical Chemists (Associação Oficial dos Químicos Analíticos)

CLS – Classical Least Squares (Quadrados Mínimos Clássico)

PC – Principal Component (Componente Principal)

ESO – Epoxidized Soybean Oil (Óleo de Soja Epoxidado)

HCA – Hierarchical Cluster Analysis (Análise de Agrupamentos Hierárquicos)

ILS – Inverse Least Squares (Quadrados Mínimos Inverso)

MLR – Multiple Linear Regression (Regressão Linear Múltipla)

NIPALS – Non-Iterative Partial Least Squares (Quadrados Mínimos Parciais Não Iterativos)

NIRS – Near Infrared Spectroscopy (Espectroscopia no Infra-Vermelho Próximo)

PARAFAC – CANDECOMP – Parallel Factor Analysis-Canonical Decomposition (Análise Paralela de Fatores – Decomposição canônica)

PCA – Principal Component Analysis (Análise por Componentes Principais)

PCR – Principal Component Regression (Regressão por Componentes Principais)

PRESS – Prediction Error Sum of Squares (Soma dos Quadrados dos Erros de Previsão)

PLS – Partial Least Squares (Quadrados Mínimos Parciais)

PVC – Polyvinyl Chloride (Cloreto de Polivinila)

SEP – Standard Error of Prediction (Erro Padrão de Previsão)

SEV - Standard Error of Validation (Erro Padrão de Validação)

SECV – Standard Error of Cross-Validation (Erro Padrão de Validação Cruzada)

SVD – Singular Value Decomposition (Decomposição por Valor Singular)

LISTA DE TABELAS

LISTA DE FIGURAS

INTRODUÇÃO

Com o advento das inúmeras técnicas instrumentais de análises químicas, o crescente desenvolvimento computacional a estas acoplado e a consequente complexidade dos dados obtidos, têm sido de grande valia a utilização conjunta de métodos matemáticos com o objetivo de retirar-se o máximo proveito dos resultados disponíveis.

A aplicação de métodos matemáticos a um conjunto de dados por natureza multivariado, como por exemplo, a decomposição por valores singulares, permite uma simplificação do mesmo no sentido de comprimir o espaço dimensional a que este está confinado, possibilitando dessa forma uma melhor interpretação e visualização [1 - 4].

Os métodos quimiométricos podem ser aplicados em dados multivariados com os propósitos qualitativos (análise exploratória e reconhecimento de padrões) e quantitativos (calibração).

Neste trabalho, estas duas áreas de atuação foram estudadas separadamente a partir de dois conjuntos de dados. No primeiro, referente à avaliação de tratamentos artificiais de desverdecimento aplicados a frutos de tangerinas, foi avaliado o efeito de alguns gases utilizados com o intuito de acelerar este processo. Outras variáveis foram estudadas simultaneamente, como a temperatura, o tempo e algumas análises físico-químicas realizadas. A análise exploratória deste conjunto foi feita por meio da Análise de Componentes Principais (Principal Component Analysis - PCA) [5-7, 13] e Agrupamentos Hierárquicos (Hierarchical Cluster Analysis - HCA) [11, 12] para dados bilineares, e utilizando o modelo PARAFAC (PARalell FACtor Analysis) [15-20] para dados de ordem superior.

Num segundo conjunto de dados, espectros na região do Infravermelho Próximo de alguns analitos (água, iodo e epóxi) medidos no óleo de soja epoxidado foram utilizados. Estes analitos são parâmetros de controle de qualidade avaliados durante o processo de epoxidação do óleo. Foram construídos modelos de calibração para estes analitos utilizando o método dos Quadrados Mínimos Parciais (Partial Least Square - PLS) [10, 11, 13, 21, 22], explorando técnicas de seleção de variáveis, como o correlograma e a seleção visual a partir dos gráficos dos pesos e do vetor de regressão. A construção destes modelos de calibração teve como principais objetivos minimizar o tempo das análises realizadas em laboratório e o custo das mesmas, uma vez que a rapidez aliada à eficiência nas determinações torna-se imprescindível no controle de qualidade em processos químicos industriais.

Capítulo 1 - MÉTODOS

Notação:

Neste trabalho, escalares são indicados por letras minúsculas, em itálico, e vetores por letras minúsculas, em negrito. Para matrizes de duas dimensões, são empregadas letras maiúsculas, em negrito, e para tensores (com três ou mais dimensões), a mesma aparece sublinhada. Um vetor ou matriz seguido de apóstrofe indica a transposta do(a) mesmo(a).

Organização dos dados:

A análise multivariada requer a organização do conjunto de dados em questão numa matriz **X** (*n* x *m*), onde as linhas desta matriz representam o conjunto das amostras e as colunas, o das variáveis medidas, ou seja, resultados analíticos, os comprimentos de onda (no caso de dados espectrais) etc.

Uma matriz de dados **X,** contendo *m* medidas experimentais (variáveis) obtidas para *n* amostras, pode ser graficamente representada por *n* pontos num espaço *m*-dimensional [5] ou, em outras palavras, é possível representar-se esta matriz espacialmente, onde cada variável medida corresponde a uma dimensão do espaço e cada amostra um ponto neste mesmo espaço.

Para conjuntos com muitas variáveis, a alta dimensionalidade apresentada pode dificultar o tratamento dos dados e uma ferramenta matemática que possibilite uma melhor visualização espacial dos mesmos (como por exemplo a Análise de Componentes Principais e Agrupamentos Hierárquicos) torna-se de grande valia.

1. Métodos Qualitativos de Análise

Análise Exploratória Multivariada

1.1 Análise de Componentes Principais (Principal Component Analysis – PCA)

A Análise de Componentes Principais [5] é uma manipulação matemática da matriz de dados com o objetivo de reduzir a dimensionalidade original da mesma e está fundamentada na correlação entre as variáveis. Variáveis que apresentam grande redundância entre si são colineares e a alta colinearidade é uma forte indicação de que é possível encontrar-se novas bases que melhor representem as informações presentes nos dados que aquela definida pelas medidas. A alta colinearidade entre as variáveis também implica em que os dados residem em um subespaço do espaço total definido pelas medidas. Na construção de um novo conjunto de vetores de base, cria-se um conjunto de novas variáveis linearmente independentes para descrever estes dados. Cada novo vetor base é expresso em termos da combinação linear das antigas variáveis. Estes novos eixos, representados pelas chamadas componentes principais, são ortogonais entre si e ordenados em termos da quantidade de variância explicada pelos dados, sendo que o primeiro vetor encontra-se na direção de maior variância [6]. Assim, este novo conjunto de eixos de coordenadas no qual projetaram-se as amostras é muito mais informativo e, pelo fato de serem ordenados pela sua importância, é possível visualizar estas mesmas amostras num gráfico de baixa dimensionalidade.

Esta projeção em uma base ortogonal pode ser feita, entre outros métodos, por meio da decomposição por valores singulares (Singular Value Decomposition – SVD). Nessa, a

matriz original **X** (*n*, *m*) é decomposta e então representada pelo produto de três novas matrizes, duas delas ortonormais (**T** e **P**) e uma diagonal (**S**) (Equação 1) [7]:

$$\mathbf{X=TSP'} \qquad (1)$$

A matriz **S** é uma matriz diagonal com elementos diagonais não negativos arranjados em ordem decrescente. Os quadrados dos valores singulares correspondem aos autovalores da matriz **X'X** e medem a importância das componentes principais individuais (cada valor singular representa a porcentagem de variância explicada em cada uma de suas respectivas componentes) [8]. O pseudoposto ou posto químico é definido a partir da exclusão dos autovalores pouco ou nada significativos.

Após a seleção do pseudoposto *f*, temos a matriz **X** agora representada pelo produto das matrizes **T**, **S** e **P** (parte sistemática) acrescida de uma matriz de erros (parte não modelável), que representa o desvio (resíduos) com relação aos dados originais.

As colunas de **P** são autovetores da matriz **X'X,** abrangendo o espaço vetorial das colunas de **X**, enquanto que a matriz **T**, formada pelos autovetores da matriz (**XX'**), abrange o espaço vetorial das linhas de **X**. O produto **TS** define as coordenadas das amostras na nova base que são denominadas escores [9].

Os escores estão relacionados com a posição ocupada pelas amostras nos novos eixos e, a informação do quanto cada variável original contribui para a formação de cada novo eixo, está contida nos pesos. Os escores expressam as relações entre as amostras enquanto que os pesos mostram as relações entre as variáveis originais [10].

Pode-se entender o conceito de escores e pesos graficamente, utilizando para isso um exemplo de duas variáveis, num espaço bidimensional (Figura 1). Nesta figura, a

componente principal é o eixo que melhor se ajusta aos pontos do conjunto mostrado. O vetor linha $\mathbf{p}'_i$ possui dimensão 1 x 2 e seus elementos p_1 e p_2 são cossenos diretores, ou projeções do vetor unitário ao longo da componente principal nos eixos do gráfico. O vetor dos escores, nesta figura chamado de $\mathbf{t}_h$, é um vetor coluna *n* x 1 e seus elementos são as coordenadas dos respectivos pontos na linha da componente principal [11].

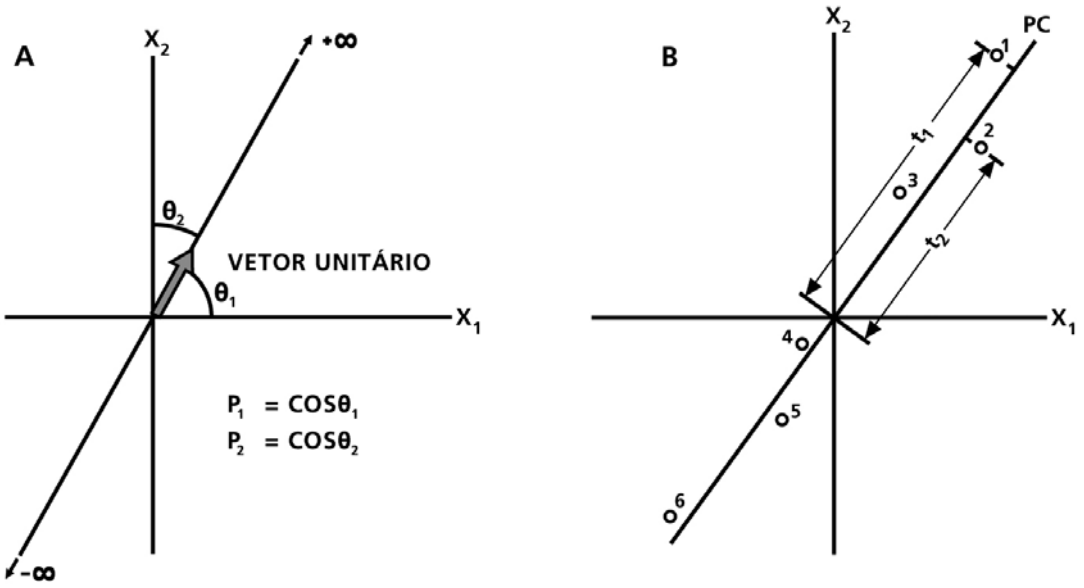


Figura 1: Componente principal (para duas variáveis): (A) Os pesos são os cossenos do vetor diretor; (B) Os escores são as projeções das amostras na direção da CP [11].

Sendo a PCA um método de quadrados mínimos, amostras anômalas influenciam fortemente no mesmo. Dessa forma, é essencial encontrá-las e eliminá-las ou corrigí-las antes de aplicar o método aos dados em questão [5].

Porém, antes mesmo disso, pode ser necessário um pré-processamento dos dados, com o objetivo de adequar as amostras do conjunto de maneira a maximizar ou minimizar o efeito de certas variáveis no todo. É o caso, por exemplo, de variáveis medidas em

diferentes unidades, com diferentes magnitudes. Outra circunstância onde se faz necessário tratar os dados seria no caso destes conterem informações correlacionadas, como ocorre, por exemplo, no caso de dados espectroscópicos.

Os pré-processamentos das variáveis podem ser feitos basicamente de três maneiras: centrando-os na média, escalando-os pela variância ou auto-escalando-os (centralização dos dados na média e posterior escalamento pela variância).

É possível também proceder-se a pré-tratamentos nas amostras do conjunto de dados, como por exemplo a 1ª e 2ª derivadas, o alisamento e a normalização, entre outros, sendo que o uso de cada método deve ser avaliado previamente, dependendo do conjunto em questão.

A centralização dos dados na média é convenientemente usada quando todas as variáveis forem medidas numa mesma unidade, possuindo uma mesma magnitude, como acontece normalmente no caso da espectroscopia. Esse tipo de pré-processamento permite que a presença de ruídos não afete negativamente na análise. Neste tipo de pré-processamento, o centróide da matriz de dados é levado à origem pela da subtração de cada elemento de cada coluna pela média da respectiva coluna (Equação 2)[12].

$$x_{ij(cm)} = x_{ij} - \overline{x}_j \qquad \textbf{(2)}$$

onde: $x_{ij(cm)}$ = valor centrado na média para a variável *j* na amostra *i*;

x_{ij} = valor da variável *j* na amostra *i*;

$\overline{x}_j$ = media dos valores das amostras na coluna *j*.

No escalamento pela variância, cada elemento de dada variável é dividido pelo desvio padrão dessa variável, levando dessa forma a variância à unidade. Esse tipo de escalamento conduz todos os eixos da coordenada ao mesmo comprimento, dando a cada variável a mesma influência no modelo (Equação 3)[12].

$$x_{ij\,(\mathrm{var})} = \frac{x_{ij}}{s_j} \qquad \textbf{(3)}$$

onde: $x_{ij\,(\mathrm{var})}$ = valor escalado pela variância para a variável *j* na amostra *i*;

x_{ij} = valor da variável *j* na amostra *i*;

s_j = desvio padrão dos valores da variável *j*.

E por último, o auto-escalamento, que é feito pela centralização dos dados na média e posterior escalamento pela variância. As variáveis terão dessa forma média zero e um desvio padrão igual a um (Equação 4)[12]. Estes dois últimos métodos são utilizados quando pretende-se dar o mesmo peso a todas as variáveis medidas, já que a PCA, por ser um método de quadrados mínimos, faz com que variáveis com alta variância possuam altos pesos [5].

$$x_{ij\,(as)} = \frac{\overline{x}_j}{s_j} \qquad \textbf{(4)}$$

onde: $x_{ij(as)}$ = valor auto-escalado da variável *j* para a amostra *i*;

$\overline{x}_j$ = média dos valores das amostras na coluna *j*;

S_j = desvio padrão dos valores da variável *j*.

Determinado o pré-processamento adequado, parte-se então para a análise exploratória propriamente dita.

1.2 Análise de Agrupamentos Hierárquicos (Hierarchical Clusters Analysis – HCA)

As técnicas de agrupamento são utilizadas com o objetivo de investigar as relações existentes dentro de um conjunto multivariado onde, *a priori*, nenhuma caracterização é conhecida. Ela pode ser de dois tipos: aglomerativa ou divisiva. No primeiro, cada amostra é considerada inicialmente um grupo e, de acordo com suas semelhanças, elas vão sendo agrupadas em subgrupos, até que todas elas formem um único grupo. Na técnica divisiva ocorre o contrário. Todas as amostras constituem um único grupo que será dividido em subgrupos, também de acordo com as similaridades entre as mesmas, até que cada amostra forme um único grupo [13].

A análise de agrupamentos hierárquicos (HCA) é uma técnica aglomerativa não supervisionada que examina as distâncias interpontuais entre todas as amostras do conjunto de dados e representa essa informação na forma de um gráfico bidimensional chamado dendrograma. Por meio do dendrograma pode-se visualizar os agrupamentos e similaridade entre as amostras e/ou variáveis.

A construção dos dendrogramas é feita com base na proximidade existente entre as amostras no espaço. Isso é feito calculando-se a distância entre todas as amostras

(agrupamentos) do conjunto, em pares, e então definindo uma matriz de similaridade cujos elementos são os chamados índices de similaridade que variam entre zero e um. Um índice alto indica uma distância pequena entre dois agrupamentos e, portanto, uma alta similaridade (Equação 5) [14]. A cada passo, os dois grupos mais similares vão se juntando e o processo vai se repetindo até que forme um único agrupamento [15].

$$s_{ij} = 1 - \frac{d_{ij}}{d_{\max}} \qquad \textbf{(5)}$$

onde : s_{ij} é a similaridade entre duas amostras (ou agrupamentos);

d_{ij} é a distância euclidiana entre as mesmas;

$d_{\max}$ é a maior distância encontrada entre todas as amostras do conjunto.

A escolha da proximidade entre dois agrupamentos pode ser feita basicamente por três métodos: do vizinho mais próximo, do vizinho mais distante ou pela média (que pode ser calculada de várias maneiras) e a mais simples medida de similaridade entre pontos num conjunto de dados é sua distância Euclideana [15].

A conexão pelo vizinho mais próximo é feita buscando inicialmente a maior similaridade (ou menor distância) entre dois grupos. A partir daí, a matriz de similaridade vai sendo continuamente atualizada, sempre procurando as menores distâncias e aproximando os agrupamentos mais similares, até que um único agrupamento seja formado.

Na conexão pela média, as amostras ligar-se-ão aos agrupamentos cujos centros estiverem mais próximos. Existem vários métodos de conexão pela média, sendo que a diferença entre eles está na maneira como este é calculado.

Pelo método do vizinho mais distante, busca-se sempre a maior distância entre as amostras e, dentre estas, o par de maior similaridade é agrupado [16].

Os agrupamentos encontrados podem então ser visualizados por um dendrograma.

No dendrograma exemplificado a seguir (Figura 2), as amostras são listadas do lado esquerdo do gráfico e os ramos indicam quais amostras estão em dado agrupamento. O eixo horizontal é uma medida da distância interagrupamentos e a posição das linhas verticais indica as distâncias entre dois destes pontos (ou similaridade) [14].

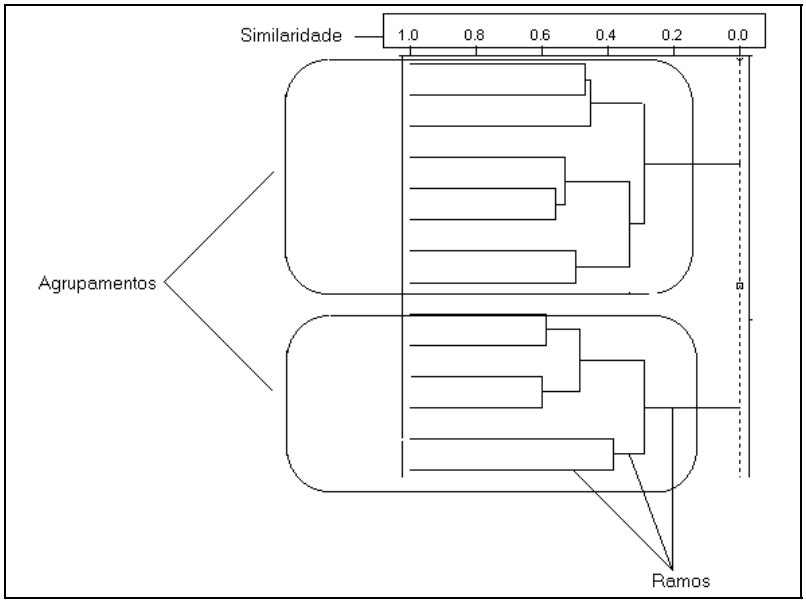


Figura 2: Exemplo de dendrograma

Este método pode ser aplicado na identificação de grupos dentro de um conjunto de dados, para testar hipóteses de agrupamentos, na identificação de membros de um dado

grupo ou na formação mais conveniente de agrupamentos com características um tanto quanto diversas [16].

Análise Exploratória Multi Way (Dados de Ordem Superior)

Durante as últimas décadas, o estudo dos chamados "arranjos n-dimensionais de dados" (mais conhecidos como arranjos de ordem superior) tem tido um enorme crescimento no ramo da análise de dados de natureza multivariada. Estes tipos de arranjos foram inicialmente estudados pela psicometria, sendo que, em 1963, L. Tucker [17] criou o primeiro modelo tridimensional de análise de componentes principais. Desde então, inúmeras extensões do método convencional têm sido desenvolvidas [18].

Este tipo de arranjo é caracterizado pela presença de três ou mais categorias de variáveis e/ou objetos num único conjunto de dados (Figura 3) [8].

Dentre os diversos modelos desenvolvidos para a análise deste tipo de dados, o mais simples é o denominado PARAFAC-CANDECOMP.

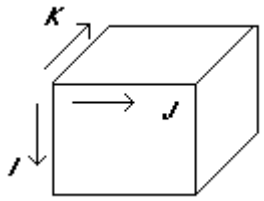


Figura 3: Arranjo tridimensional representado por um paralelepípedo de dimensões I x J x K, onde cada dimensão é caracterizada por uma dada categoria de variável e/ou objeto.

1.3 Método PARAFAC - CANDECOMP

Este modelo foi inicialmente proposto de maneira simultânea por Harshman (PARAllel FACtor analysis) [19] e Carrol & Chang (CANonical DECOMPosition) em 1970 [20].

O modelo é uma extensão da idéia da PCA bidimensional [21] e vem ganhando cada vez mais interesse por parte dos quimiometristas e pesquisadores de áreas afins [22].

É um método de decomposição n-dimensional onde o arranjo inicial é transformado em conjuntos de tríades ou componentes trilineares, que descrevem o mesmo de uma forma muito mais concisa que no arranjo original [23]. Ao contrário da PCA bilinear, cada componente consiste de um vetor de escores e dois de pesos. Porém, vale ressaltar que em se tratando de dados de ordem superior, usualmente não é feita a distinção entre escores e pesos, sendo estes tratados igualmente [24].

Um dado tensor **X** (*I, J, K*) é decomposto em duas contribuições, uma sistemática, descrita por *F* conjuntos de produtos externos de vetores (onde cada um é denominado uma tríade) e uma não modelável (**E**) representando os desvios do modelo com relação aos dados originais (Figura 4)[18].

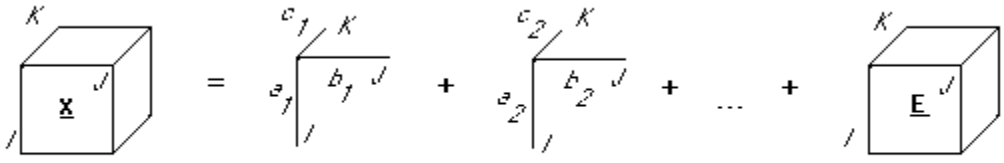


Figura 4: Decomposição do tensor X em suas contribuições - sistemática (tríades) e não modelável (E).

A decomposição originará três novas matrizes de pesos (**A**, **B** e **C**) com elementos a_{if}, b_{jf} e c_{kf}, relativas cada uma delas a um "modo" do arranjo (Equações 6 e 7). A modelagem ocorre de maneira a minimizar a soma dos quadrados dos resíduos e_{ijk} no modelo (Figura 5).

Este termo "modo" foi primeiramente utilizado por Tucker [17], indicando cada conjunto de índices pelos quais o arranjo de dados pode ser desdobrado.

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \qquad \textbf{(6)}$$

O termo a_{if} representa o peso da componente f no nível i do modo **A**.

Esta equação pode também ser escrita como:

$$\text{vec}(\underline{\mathbf{X}}) = \sum_{f=1}^{F} a_f \otimes b_f \otimes c_f + \text{vec}(\underline{\mathbf{E}}) \qquad \textbf{(7)}$$

onde a_f, b_f e c_f são os f-ésimas colunas das matrizes de pesos **A**, **B** e **C** respectivamente [21], vec é um operador de vetorização e $\otimes$ representa o produto de Kroneker[1].

[1] O símbolo $\otimes$ representa o produto de Kroneker, também chamado produto tensorial, e é definido como:
Dadas as matrizes $\mathbf{A}=a_{ij}(n,m)$ e $\mathbf{B}=b_{ij}(p,q)$, $\mathbf{A} \otimes \mathbf{B}=a_{ij}\mathbf{B}(np,mq)$ [25].

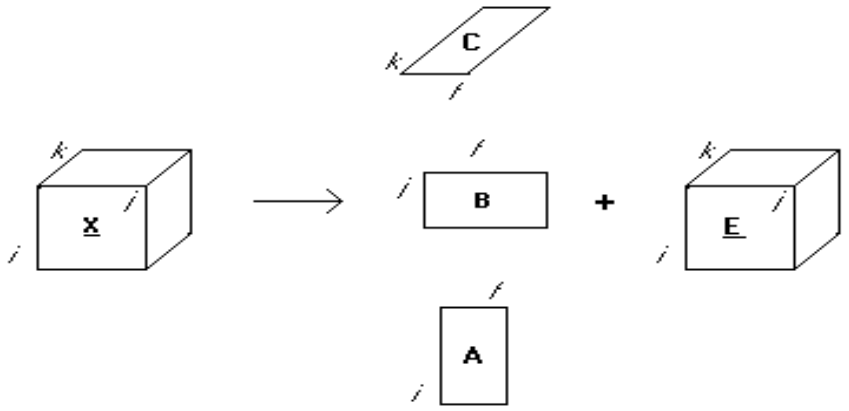


Figura 5: Representação gráfica da decomposição de um arranjo tridimensional $\underline{X}$ pelo modelo PARAFAC para *f* componentes.

Os modelos bilineares apresentam com frequência o já conhecido problema da liberdade rotacional. A grande vantagem do modelo PARAFAC com relação a estes, é sua singularidade de soluções [24]. Como toda rotação destrói a soma dos quadrados mínimos de uma solução ótima, o modelo propicia uma solução única que pode ser diretamente interpretável [26]. Porém, restrições baseadas em informações a respeito dos dados (por exemplo, não negatividade no caso de espectros), podem ser impostas para obter soluções estáveis na decomposição trilinear [25].

O algoritmo utilizado na solução do modelo PARAFAC é o ALS (Alternating Least Squares). Este assume sucessivamente que os pesos entre dois modos são conhecidos e então estima os parâmetros do último modo desconhecido por quadrados mínimos [18]. Este é um algoritmo iterativo cuja finalização ocorre apenas quando a diferença relativa nos ajustes entre duas iterações sucessivas estiver abaixo de um dado limite [18].

Uma dificuldade encontrada no modelo PARAFAC é a determinação do número ideal de componentes principais (F) a serem utilizadas (pseudoposto), isto é, as dimensões

das colunas das matrizes dos pesos. Essa determinação, contudo é ainda mais fácil que nas análises bidimensionais, devido às propriedades de singularidade do modelo. Em geral, as ferramentas utilizadas para este propósito são as mesmas que as usadas nas análises bidimensionais: a avaliação dos histogramas dos resíduos e a comparação com conhecimento externo (significado físico) dos dados a serem modelados [18], além do chamado diagnóstico de consistência do "core".

O diagnóstico de consistência de core é um indicativo de quão estável está o modelo criado. Este valor varia de acordo com o número de componentes principais escolhido. Para um dado número de componentes principais, a solução obtida pelo ajuste do modelo é utilizada no cálculo deste diagnóstico. Quanto maior o valor deste, mais estável é o modelo construído. Valores baixos podem indicar que o número de componentes escolhido não é o ideal ou ainda que pode ser necessário a imposição de restrições para melhorá-lo.

2. Métodos Quantitativos de Análise

A calibração tem por objetivo a construção de um modelo matemático que relacione a resposta de um dado instrumento analítico com alguma(s) propriedade(s) de interesse das respectivas amostras. A partir do modelo construído torna-se então possível a quantificação dessa(s) propriedade(s) em novas amostras onde elas são desconhecidas. A calibração univariada relaciona uma única resposta (variável) com essa propriedade de interesse, enquanto que na multivariada existe uma relação de um conjunto de respostas.

Calibração Multivariada

A calibração multivariada tem como idéia básica relacionar dois blocos de dados (**X** e **Y**), onde **X** (*n*, *m*) é o bloco dos dados experimentais contendo as variáveis independentes, representado em suas linhas pelas amostras do conjunto e em suas colunas pelas variáveis medidas e o outro bloco (**Y**), formado pela(s) variável(eis) dependentes(s) [27]. Dessa forma, na calibração multivariada, ao contrário do que ocorre na univariada, utilizam-se múltiplas respostas para relacionar com a(s) propriedade(s) medida(s) das amostras [14].

Este tipo de calibração tem como principais vantagens o fato de permitir determinações simultâneas de mais de um analito de interesse, permitir determinações mesmo na presença de interferentes e, apresentar uma diminuição do erro estimado no modelo devido ao fato de ser um método que utiliza múltiplas variáveis.

A calibração, de maneira geral, pode ser dividida em duas etapas consecutivas:

- Modelagem: estabelece-se uma relação entre o sinal medido e a propriedade que se deseja quantificar da amostra;
- Validação: assegura que o modelo reflete o comportamento do analito;

Basicamente, os métodos de regressão multivariada são classificados em dois tipos principais: Direto - Quadrados Mínimos Clássico (CLS – Classical Least Squares) e Inverso - (ILS – Inverse Least Squares). A construção do modelo de calibração pode ser feita por vários métodos e a escolha entre cada um deles está relacionada ao perfil do conjunto de dados [12]. Estes métodos serão apresentados a seguir.

Uma vez construído o modelo, este deve ser validado, ou seja, testado para garantir que os valores obtidos para a variável dependente sejam iguais ou bastante próximos dos experimentais.

A validação pode ser de dois tipos: utilizando um conjunto externo ou por meio da validação cruzada.

No primeiro caso, o conjunto de dados é dividido em dois subconjuntos, de calibração e de validação. O modelo é construído usando as amostras do conjunto de calibração e depois é validado utilizando o conjunto com as amostras restantes [12]. Os resultados obtidos para a variável dependente nessa etapa são comparados aos valores experimentais, sendo os resíduos calculados. A eficácia do modelo construído é maior quão menor forem os resíduos encontrados [12].

O método de Validação Cruzada funciona da seguinte forma: a matriz de dados é dividida em pequenos grupos. Um determinado grupo é então removido da matriz original e esta, agora reduzida, é decomposta normalmente em escores e pesos, e o modelo de calibração é estabelecido sobre esse novo conjunto. A partir deste, a propriedade das amostras removidas será prevista e os resíduos obtidos entre os valores reais e os estimados são computados. Isso é repetido para todos os pequenos grupos do conjunto de dados.

Existem diversas maneiras de dividir a matriz de dados em subgrupos. Se o número de amostras não é grande, a melhor maneira é utilizar a chamada “uma amostra fora por vez”. Nessa, como o próprio nome já diz, todas as amostras são consideradas um subgrupo individualmente. Esse procedimento torna-se bastante trabalhoso para conjuntos muito grandes, com muitas variáveis [28].

Em ambos os casos, os parâmetros comumente utilizados para avaliar a eficácia dos modelos construídos são o PRESS – Prediction Residual Error Sum of Squares (Equação 8)

e o SECV/SEP – Standard Error of Cross Validation / Prediction (Equações 9 e 10). A diferença básica destes dois últimos está no fato de o SECV levar em conta o número de componentes principais (fatores) utilizados.

$$PRESS = \sum_i (\hat{y}_i - y_i)^2 \qquad \textbf{(8)}$$

$$SECV = \sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{n - k - 1}} \qquad \textbf{(9)}$$

$$SEP = \sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{n}} \qquad \textbf{(10)}$$

onde: $\hat{y}_i$ é o valor previsto para a amostra *i* utilizando o modelo;

y_i é o valor medido para a amostra *i*;

k é o número de fatores utilizado;

n é o número de amostras do conjunto de calibração.

Após a validação, já é possível utilizar o modelo construído na previsão da propriedade modelada em amostras onde este valor é desconhecido [27].

2.1 Calibração por Quadrados Mínimos Clássico (CLS)

É utilizada em casos onde além do sistema apresentar-se simples, todos os seus analitos (componentes) são conhecidos e seus respectivos espectros podem ser obtidos separadamente. A grande restrição deste método é a exigência do conhecimento da concentração de cada espécie presente no sistema, o que muitas vezes pode ser inviável.

2.2 Calibração por Quadrados Mínimos Inversos (ILS)

Os métodos de calibração inversa podem ser divididos em três tipos: Regressão Linear Múltipla (MLR – Multiple Linear Regression), Regressão por Componentes Principais (PCR – Principal Component Regression) e Regressão por Quadrados Mínimos Parciais (PLS – Partial Least Squares) [27].

A característica básica dos métodos inversos está em como a relação entre as variáveis medidas e a concentração é modelada. As concentrações são tomadas como uma função das respostas (Equação 11):

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} \qquad \textbf{(11)}$$

onde: $\hat{\mathbf{y}}$ é o vetor das concentrações (*n* amostras x 1)

X é a matriz das variáveis medidas (*n* amostras x *m* variáveis)

$\hat{\mathbf{b}}$ é o vetor que contém os coeficientes do modelo (*n* variáveis x 1)

Esta equação pode ser usada para modelar a relação entre múltiplos analitos de interesse (diferentes vetores **y**) e a mesma matriz resposta (**X**) usando diferentes coeficientes do modelo ($\hat{\mathbf{b}}$) [14].

A determinação do vetor $\hat{\mathbf{b}}$ é feita pela resolução da equação: $\hat{\mathbf{b}} = (\mathbf{X'X})^{-1}\mathbf{X'}\,\hat{\mathbf{y}}$ e a chave para a resolução do mesmo está na inversão da matriz (**X'X**). Esta é uma matriz quadrada com número de linhas e colunas iguais ao número de variáveis medidas. Em teoria, um número de amostras no conjunto de calibração maior ou igual ao número de variáveis é necessário para essa inversão. Porém, geralmente o que ocorre em sistemas analíticos é um número de variáveis maior que o número de amostras ou ainda um grande número de variáveis altamente correlacionadas, tornando assim a matriz inversível (singular) [14].

2.2.1 Regressão Linear Múltipla - MLR

Quando se utiliza a MLR em casos como esses, uma solução para o problema poderia ser a redução do número de variáveis por meio de algum método de seleção, tornando a matriz não singular. Porém isso pode ser muito trabalhoso se o conjunto em questão for muito grande. Dessa forma, recomenda-se seu uso apenas quando o número de variáveis for pequeno, ou quando apenas um subconjunto das variáveis medidas é desejado.

Ao contrário deste, os métodos PCR e PLS, que são abordados no próximo tópico, não requerem a seleção de variáveis. Nestes, ocorre a transformação das variáveis medidas

em novas variáveis (componentes principais), que serão então utilizadas nos cálculos do modelo.

Porém, uma vantagem do método MLR, mesmo usando a seleção de variáveis, sobre os métodos que utilizam o espectro total (PCR e PLS), é que esse modelo é bastante simples, tornando-se ineficiente apenas no caso de uma seleção errônea de variáveis [14].

2.2.2 Regressão por Componentes Principais - PCR e Regressão por Quadrados Mínimos Parciais – PLS

Como mencionado anteriormente, a grande dificuldade encontrada no modelo inverso está na resolução de $\hat{\mathbf{b}}$ no caso de **X'X** não ser inversível (devido à redundância entre as variáveis) [14].

A idéia básica por trás destes dois métodos está em encontrar combinações lineares relevantes a partir dos valores dos espectros originais e então usá-las na equação de regressão. Dessa forma, toda informação irrelevante (informações aleatórias) dos espectros será descartada e apenas as partes mais relevantes serão utilizadas na modelagem. Assim, o problema da colinearidade é resolvido e uma equação de regressão mais objetiva é obtida [29].

Nestes modelos em específico, esta redundância é eliminada pela construção de uma nova matriz **P** com colunas formadas por combinações lineares das originais em **X**. Essa nova matriz agora torna **P'P** inversível, sendo assim possível encontrar o vetor que contém os coeficientes do modelo ($\hat{\mathbf{b}}$).

A diferença básica entre estes dois métodos (PCR e PLS) está em como essa nova base **P** é construída.

Na PCR, utiliza-se exclusivamente a matriz **X** na determinação da combinação linear das variáveis e as concentrações são usadas apenas quando os coeficientes de regressão são estimados. Essa mudança de bases é feita exclusivamente a partir de uma análise por componentes principais simples (PCA), e os resultados são então relacionados com as variáveis independentes (**y**), como é demonstrado a seguir.

Ao rearranjar a Equação (1), multiplicando ambos os lados por **P**, temos:

$$\mathbf{XP} = \mathbf{TP'P} \qquad \textbf{(12)}$$

Como **P** é ortonormal, **P'P = I** (**I** é matriz identidade) e portanto a Equação (1) pode ser escrita em função dos escores:

$$\mathbf{T} = \mathbf{XP} \qquad \textbf{(13)}$$

A etapa de correlação desta matriz, com a das variáveis dependentes, pode ser descrita pelas equações 14 e 15.

$$\mathbf{y} = \mathbf{Tb} + \mathbf{E} \qquad \textbf{(14)}$$

$$\mathbf{b} = (\mathbf{T'T})^{-1}\mathbf{T'y} \qquad \textbf{(15)}$$

Devido à ortogonalidade entre as linhas da matriz dos escores, a inversão de **T'T** não é mais problemática [12]. Deve-se lembrar que, como foi explicado anteriormente no item 1.1, a matriz **T** (*n*, *m*) foi truncada em **T** (*n*, *f*) após a seleção do pseudoposto *f* da matriz diagonal **S**.

Esta forma de relacionar o vetor **y** somente depois da decomposição da matriz **X**, pode constituir uma fragilidade do método no caso onde o analito de interesse tenha um sinal muito fraco. Neste caso, esse sinal não influencia fortemente na composição das primeiras componentes principais, fazendo com que haja a necessidade do uso de um número maior de componentes na construção do modelo [27].

O método PLS contorna essa dificuldade característica do PCR usando a informação das concentrações na obtenção dos fatores, o que só é justificável se tais concentrações tiverem valores confiáveis [27]. Neste método, a covariância nas medidas com as concentrações é utilizada em conjunto à variância em **X** para gerar **T**, tirando vantagem assim da correlação existente entre os dados da matriz original e a variável dependente [13]. No PLS, usa-se comumente o termo variável latente para designar as componentes principais. Isso se deve ao fato da construção das mesmas ser feita a partir de informações contidas no vetor das variáveis dependentes. O número de variáveis latentes que será utilizado é determinado durante o processo de validação cruzada.

Uma grande vantagem do método PLS é sua robustez. Isso significa que os parâmetros do modelo não se alteram de maneira significativa quando novas amostras são acrescentadas ou retiradas do conjunto de calibração [11]. Essa robustez permite que seja possível trabalhar com sistemas industriais cujas características nem sempre são mantidas rigorosamente da mesma maneira durante todo o processo, ou seja, é possível acrescentar estas "novas amostras" conforme elas apareçam sem alterar os parâmetros do modelo criado inicialmente.

O algoritmo comumente utilizado no método PLS é o NIPALS (Non-Iterative Partial Least Squares). Neste, como pode ser visto no exemplo a seguir (Exemplo 10), as variáveis latentes vão sendo calculadas sucessivamente de forma iterativa a partir da modelagem dos blocos **X** e **Y** [10].

Exemplo 1: Algoritmo NIPALS para cálculo PLS [10]

PREPROCESSE **X** e **y**

ESTIME o primeiro vetor peso **w**.

LOOP

BLOCO **X**

$\mathbf{X} = \mathbf{y}\,\mathbf{w}^T + \mathbf{E}_x \qquad (\mathbf{X}^T = \mathbf{w}\,\mathbf{y}^T)$

$\hat{\mathbf{w}} = \mathbf{X}^T (\mathbf{y}^T \mathbf{y})^{-1} \mathbf{y}^T$

$\hat{\mathbf{w}} = \hat{\mathbf{w}}^T / (\hat{\mathbf{w}}^T \hat{\mathbf{w}})^{1/2}$ (normalização)

$\mathbf{t} = \mathbf{X}\,\hat{\mathbf{w}}$ (escores)

$\mathbf{p} = \mathbf{X}^T \mathbf{t} (\mathbf{t}^T \mathbf{t})^{-1}$ (pesos)

Relacione o vetor **t** com a variável **y**

$q = (\mathbf{t}^T \mathbf{t})^{-1} \mathbf{t}^T \mathbf{y}$

Atualize **X** e **y**

$\mathbf{E} = \mathbf{X} - \mathbf{t}\,\mathbf{p}^T$

$\mathbf{e} = \mathbf{y} - q\mathbf{t}$

$\mathbf{X} = \mathbf{E}$

$\mathbf{Y} = \mathbf{e}$

Repita para o próximo vetor peso

CALCULE o vetor de regressão

$$\hat{\mathbf{b}} = \mathbf{W} (\mathbf{L}^T \mathbf{W})^{-1} \mathbf{q}^T$$

Mesmo nesses métodos, muitas vezes faz-se necessária a seleção de variáveis, porém não devido às limitações apresentadas no método MLR. O conjunto de dados muitas

vezes pode conter variáveis de alta colinearidade, como no caso de espectros, ou mesmo variáveis que não contribuem fortemente para a construção de um modelo de regressão, como muitas vezes ocorre em problemas de QSAR/QSPR (Quantitative Structure-Activity / Property Relationship) [12]. Nestes casos, é comum recorrer-se à algum método de seleção com o objetivo de minimizar estes problemas e otimizar a construção dos modelos propriamente ditos, como poderá ser visto na aplicação do Capítulo 4.

Para os dois trabalhos desenvolvidos, utilizou-se o software Matlab 5.2 da MathWorks [30], sendo que para o primeiro utilizou-se o Toolbox Nway 1.05 [31] e para o segundo o PLS – Toolbox [30], além do Pirouette v.2.02 da InfoMetrix [32].

APLICAÇÕES

Capítulo 2 - *Estudo quimiométrico do efeito do tratamento de desverdecimento na composição química interna do tangor Murcote (Citrus reticulata x Citrus sinensis)*

1. Objetivos

O objetivo deste trabalho foi fazer um estudo do processo artificial de desverdecimento de tangerinas Murcote. Alguns parâmetros que podem influenciar neste processo foram avaliados, como por exemplo, o tipo de gás aplicado durante o mesmo e a temperatura utilizada. As análises químicas realizadas foram feitas em diversos períodos. Para tanto, os dados fornecidos pelo Instituto Agronômico de Campinas (IAC) e do Instituto de Tecnologia de Alimentos (ITAL) foram analisados por métodos quimiométricos bilineares e de ordem superior, além da análise variância. Estes enfoques serão tratados a seguir.

2. Introdução

Um dos pontos de grande importância na aceitação e consequente comercialização das frutas cítricas é sua aparência externa. Busca-se sempre a uniformidade e uma perfeita coloração natural da casca destes frutos.

A mudança completa da coloração externa do tangor Murcote, de verde para o alaranjado intenso pode ser melhorada empregando-se técnicas que promovam o

desverdecimento destes frutos, ou seja, que acelerem o processo de degradação da clorofila e revelem os pigmentos carotenóides anteriormente encobertos [33].

A coloração dos frutos cítricos depende da relação entre a clorofila e os carotenóides presentes no flavedo [34]. A cor da clorofila predomina e, somente quando ela apresenta-se em baixa concentração é que predomina a cor dos carotenóides. A diminuição da concentração de clorofila pode ser obtida por um controle de temperatura e da concentração de etileno aplicada durante o processo artificial de desverdecimento.

O gás etileno tem sido usado com esse propósito, porém devido às dificuldades apresentadas com relação à necessidade de usarem-se câmaras especiais ou recipientes herméticos, tem-se testado o uso do Ethrel, que por apresentar-se na forma líquida, dispensa os cuidados anteriormente citados [35].

O Ethrel, cujo principal ingrediente ativo é o ethefon (ácido 2-cloroetil fosfônico), degrada-se abaixo da superfície da casca da fruta, liberando o etileno. A resposta ao tratamento com este composto é dependente da temperatura.

O objetivo deste trabalho foi analisar os dados referentes ao estudo do efeito da aplicação do gás etileno (Etil–5) e de Ethrel 240, em diferentes concentrações e temperaturas, na qualidade dos frutos de tangores Murcote (híbrido resultante do cruzamento das espécies *citrus reticulata* Blanco – tangerina - e *citrus sinensis* Osbeck - laranja). Essa qualidade pode ser avaliada a partir de algumas variáveis como o teor de ácido ascórbico, o ^{o}Brix (teor de sólidos solúveis), a coloração da casca dos frutos, entre outras. Para tanto, foram empregadas duas metodologias quimiométricas para dados multivariados: Análise de Componentes Principais (PCA) e de Agrupamentos Hierárquicos (HCA), além de uma metodologia para dados multidimensionais (de ordem superior), PARAFAC (Parallel Factor Analysis). Empregou-se também a Análise de Variância

(ANOVA – Analysis of variance) [36, 37] como um complemento no estudo das variáveis do conjunto de dados.

3. Materiais e Métodos

Frutos de tangor Murcote, com coloração de casca ainda verde, foram colhidos perfeitamente desenvolvidos em pomares comerciais do Estado de São Paulo. Com o objetivo de se acelerar o desverdecimento, os frutos foram submetidos aos seguintes tratamentos: a) acondicionamento dos frutos em um tambor hermético de 200L, no qual injetou-se o gás Etil-5 (etileno a 5%, em mistura com nitrogênio) na concentração de 10 μL/L de etileno; b) imersão dos frutos, por 3 minutos, em solução aquosa de Ethrel 240 (ácido 2-cloroetil fosfônico ou Ethefon, da Union Carbide do Brasil) nas concentrações de 250, 500, 750, 1000 e 2000 μL/L; c) tratamento controle – imersão dos frutos, por 3 minutos, em água.

No processo de desverdecimento, por 72 horas, os frutos correspondentes a cada um dos sete tratamentos citados, foram separados em dois lotes, os quais foram mantidos em câmaras de desverdecimento a 25°C e a 30°C, ambas com umidade relativa de 90%.

Após a fase de desverdecimento, as frutas foram transportadas para câmaras refrigeradas a 5-6°C, por um período de 5 semanas (fase de refrigeração), visando simular as condições de transporte marítimo aos principais países importadores do produto – Países Baixos e Arábia Saudita. Em seguida, foram deixadas à temperatura ambiente por mais duas semanas (fase de comercialização), com o intuito de simular as condições de comercialização, sem refrigeração.

Com o objetivo de avaliar a influência dos processos estudados na qualidade dos tangores, foram realizadas determinações, segundo AOAC (1970) [38], do teor de ácido ascórbico presente no suco dos frutos (titulação com 2,6-diclorofenolindofenol), pH (método potenciométrico), acidez total (pelo método acidimétrico, expressa em ácido cítrico) e ^{o}Brix (teor de sólidos solúveis, pelo método refratométrico). Determinou-se também a relação ^{o}Brix/acidez, a porcentagem de perda de peso fresco e a coloração externa dos tangores [38].

Devido à desuniformidade da coloração inicial das cascas dos frutos, estes foram separados em subgrupos A, B e C, sendo A composto de frutos de coloração verde mais escura, C, de coloração verde com traços de amarelo e B, de coloração intermediária aos subgrupos A e C. Para facilitar, nas análises foi utilizada uma média entre os valores A, B e C. Na avaliação subjetiva do desverdecimento foi empregada uma equipe de 4 julgadores. Para a avaliação da cor externa dos frutos utilizou-se uma escala de notas decrescente, cujos valores variavam de 14 a 1, sendo a nota 14 atribuída à cor totalmente verde e a nota 1, à cor laranja. Tanto as avaliações visuais quanto as avaliações de qualidade foram feitas ao 3^{o} dia (após a fase de desverdecimento) e semanalmente, durante as fases de refrigeração (4 semanas) e de comercialização (2 semanas), totalizando 7 períodos estudados [34].

4. Resultados e Discussão

4.1 Análise Multivariada (PCA e HCA) e Análise de Variância

Pelo fato dos dados coletados neste estudo serem de relativa complexidade, no sentido de abrangerem várias dimensões de estudo, como tempo, temperatura, tipo de tratamento aplicado aos frutos e análises físico-químicas e visuais realizadas, procurou-se estudar a influência de cada uma destas separadamente, a partir da construção de duas matrizes diferentes, alternando estas dimensões, ora como amostras, ora como variáveis. Os resultados mais significativos são discutidos a seguir. O conjunto de dados utilizado, na íntegra, pode ser visto nas tabelas mostradas a seguir (Tabelas 1 a 7).

Em todos os casos apresentados, as matrizes foram auto-escaladas. Esse tipo de escalamento, como já foi dito anteriormente, permite que seja dado um mesmo peso a todas as variáveis, ou seja, permite que uma dada variável não se sobreponha às outras de igual importância. Nesse caso em específico, isso se fez necessário devido à magnitude e às unidades das variáveis em questão, que diferiam entre si, em valores numéricos, entre 0,2 e 30 aproximadamente.

Tabela 1: Teores de ácido ascórbico (mg/100g) determinados em tangerinas Murcote, nos períodos de desverdecimento, armazenamento refrigerado e comercialização.

Fase	Desverdecimento		Refrigeração			Comercialização	
Temperatura	25°C		5-6°C			Ambiente	
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	20,3	19,3	19,3	18,9	16,5	14,9	20,0
Etil 10µL/L	20,7	19,6	19,4	18,8	15,5	12,0	16,6
Ethrel 250µL/L	20,1	21,5	20,1	19,8	18,6	16,9	18,0
Ethrel 500µL/L	22,1	22,4	21,5	20,9	18,8	16,6	18,9
Ethrel 750 µL/L	21,8	19,5	19,5	19,0	17,4	15,6	17,6
Ethrel 1000µL/L	19,3	19,4	18,8	18,3	16,7	14,4	18,5
Ethrel 2000µL/L	22,3	23,2	22,9	21,9	19,0	16,0	23,0

Fase	Desverdecimento		Refrigeração			Comercialização	
Temperatura	30°C		5-6°C			Ambiente	
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	20,5	20,4	19,6	18,8	17,8	16,8	17,2
Etil 10µL/L	19,8	20,5	19,0	18,4	14,9	10,9	13,1
Ethrel 250µL/L	19,7	20,5	19,4	18,2	17,9	17,4	16,9
Ethrel 500µL/L	18,7	18,6	18,3	17,8	16,4	15,0	16,6
Ethrel 750 µL/L	22,0	20,1	19,7	18,5	16,8	15,1	17,2
Ethrel 1000µL/L	22,1	20,4	19,8	19,1	18,0	17,0	17,0
Ethrel 2000µL/L	20,9	18,8	18,7	18,1	17,0	16,4	12,5

Tabela 2: Médias de porcentagem de perda de peso determinados em tangerinas Murcote, nos períodos de desverdecimento, armazenamento refrigerado e comercialização.

Fase	Desverdecimento	Refrigeração			Comercialização	
Temperatura	25°C	5-6°C			Ambiente	
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	0,37	3,93	4,90	6,78	7,99	10,14	11,83
Etil 10µL/L	0,07	1,36	2,57	4,68	5,72	7,36	8,94
Ethrel 250µL/L	0,57	1,79	3,02	5,05	5,95	7,58	9,37
Ethrel 500µL/L	0,28	0,72	1,20	2,57	3,55	5,17	6,87
Ethrel 750 µL/L	0,54	1,77	2,93	4,79	5,88	8,03	9,96
Ethrel 1000µL/L	0,41	1,70	2,67	4,78	5,68	8,03	9,89
Ethrel 2000µL/L	0,23	1,78	2,87	4,69	5,95	7,63	9,95

Fase	Desverdecimento	Refrigeração			Comercialização	
Temperatura	30°C	5-6°C			Ambiente	
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	1,33	3,69	4,67	6,34	7,26	8,63	10,24
Etil 10µL/L	0,31	1,90	2,96	4,80	5,97	8,09	10,03
Ethrel 250µL/L	0,52	1,19	2,26	4,15	5,07	6,91	8,76
Ethrel 500µL/L	1,22	2,08	3,01	4,82	5,79	7,22	9,08
Ethrel 750 µL/L	1,17	2,65	3,43	5,23	6,09	7,50	9,21
Ethrel 1000µL/L	1,38	1,99	3,10	5,13	6,15	8,38	10,38
Ethrel 2000µL/L	1,14	1,78	2,77	5,01	6,07	7,73	10,49

Tabela 3: °Brix determinados em tangerinas Murcote, nos períodos de desverdecimento, armazenamento refrigerado e comercialização.

Fase	Desverdecimento	Refrigeração			Comercialização		
Temperatura	25°C	5-6°C			Ambiente		
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	9,3	8,1	7,6	8,2	8,4	8,3	8,8
Etil 10µL/L	9,5	8,4	7,8	7,7	8,1	7,9	8,7
Ethrel 250µL/L	9,8	9,0	8,5	8,1	9,0	8,6	8,8
Ethrel 500µL/L	9,7	8,6	8,4	8,6	8,9	9,8	8,7
Ethrel 750 µL/L	9,8	8,6	8,9	8,3	8,5	8,9	9,0
Ethrel 1000µL/L	9,4	8,0	8,7	8,2	8,8	8,8	9,4
Ethrel 2000µL/L	9,4	8,6	8,2	8,4	9,0	8,8	9,8

Fase	Desverdecimento	Refrigeração			Comercialização		
Temperatura	30°C	5-6°C			Ambiente		
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	8,7	8,4	7,7	7,8	8,7	8,6	8,8
Etil 10µL/L	8,8	8,6	8,4	8,6	8,1	9,4	9,4
Ethrel 250µL/L	9,7	8,4	8,6	8,2	8,4	8,8	9,4
Ethrel 500µL/L	9,6	8,4	8,8	8,0	8,8	9,4	9,2
Ethrel 750 µL/L	8,9	8,8	7,7	8,6	8,7	9,1	9,0
Ethrel 1000µL/L	10,5	8,8	9,0	9,3	9,2	9,2	9,6
Ethrel 2000µL/L	8,9	8,4	7,8	7,4	7,7	8,4	8,0

Tabela 4: Relação °Brix/acidez total determinados em tangerinas Murcote, nos períodos de desverdecimento, armazenamento refrigerado e comercialização.

Fase	Desverdecimento	Refrigeração			Comercialização		
Temperatura	25°C	5-6°C			Ambiente		
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	12,2	9,8	12,5	12,2	13,5	11,1	13,3,
Etil 10µL/L	12,5	11,5	10,3	13,3	14,0	11,3	15,8
Ethrel 250µL/L	12,4	12,9	12,5	9,8	14,3	14,8	13,8
Ethrel 500µL/L	12,9	15,4	11,2	14,1	14,1	14,4	13,0
Ethrel 750 µL/L	12,4	11,6	13,7	11,7	14,4	12,7	18,4
Ethrel 1000µL/L	12,1	11,8	16,4	13,2	14,7	12,9	17,4
Ethrel 2000µL/L	11,9	12,1	10,9	12,9	19,1	20,0	16,6

Fase	Desverdecimento	Refrigeração			Comercialização		
Temperatura	30°C	5-6°C			Ambiente		
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	10,4	11,5	12,0	11,0	16,1	13,4	15,4
Etil 10µL/L	11,3	14,3	12,9	14,1	15,0	18,4	19,2
Ethrel 250µL/L	11,1	10,4	13,9	10,1	14,0	13,3	14,7
Ethrel 500µL/L	13,0	14,2	12,4	13,1	14,2	16,2	17,0
Ethrel 750 µL/L	11,4	13,3	12,4	13,7	13,4	15,2	14,1
Ethrel 1000µL/L	13,8	11,1	13,4	13,5	14,6	15,3	15,2
Ethrel 2000µL/L	11,6	13,1	11,5	12,5	14,5	15,3	21,6

Tabela 5: Acidez Total (expressa em g/100g de ácido cítrico) determinados em tangerinas Murcote, nos períodos de desverdecimento, armazenamento refrigerado e comercialização.

Fase	Desverdecimento	Refrigeração			Comercialização		
Temperatura	25°C	5-6°C			Ambiente		
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	0,76	0,83	0,61	0,67	0,63	0,75	0,66
Etil 10µL/L	0,76	0,73	0,76	0,58	0,58	0,70	0,55
Ethrel 250µL/L	0,79	0,70	0,68	0,83	0,63	0,58	0,64
Ethrel 500µL/L	0,75	0,56	0,75	0,61	0,63	0,68	0,67
Ethrel 750 µL/L	0,79	0,74	0,65	0,71	0,59	0,70	0,49
Ethrel 1000µL/L	0,78	0,68	0,53	0,62	0,60	0,68	0,54
Ethrel 2000µL/L	0,79	0,71	0,75	0,65	0,47	0,44	0,59

Fase	Desverdecimento	Refrigeração			Comercialização		
Temperatura	30°C	5-6°C			Ambiente		
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	0,84	0,73	0,64	0,71	0,54	0,64	0,57
Etil 10µL/L	0,78	0,60	0,65	0,61	0,54	0,51	0,49
Ethrel 250µL/L	0,87	0,81	0,62	0,81	0,60	0,66	0,64
Ethrel 500µL/L	0,74	0,59	0,71	0,61	0,62	0,58	0,54
Ethrel 750 µL/L	0,78	0,66	0,62	0,63	0,65	0,60	0,64
Ethrel 1000µL/L	0,76	0,79	0,67	0,69	0,63	0,60	0,63
Ethrel 2000µL/L	0,77	0,64	0,68	0,59	0,53	0,55	0,37

Tabela 6: Valores de pH determinados em tangerinas Murcote, nos períodos de desverdecimento, armazenamento refrigerado e comercialização.

Fase	Desverdecimento		Refrigeração			Comercialização	
Temperatura	25°C		5-6°C			Ambiente	
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	3,58	3,40	3,69	3,72	3,76	3,63	3,68
Etil 10µL/L	3,68	3,58	3,68	3,91	3,71	3,69	3,84
Ethrel 250µL/L	3,53	3,65	3,73	3,70	3,85	3,88	3,87
Ethrel 500µL/L	3,55	3,83	3,62	3,80	3,89	3,76	3,79
Ethrel 750 µL/L	3,51	3,50	3,56	3,72	3,85	3,81	3,82
Ethrel 1000µL/L	3,53	3,77	3,70	3,76	3,81	3,78	3,82
Ethrel 2000µL/L	3,66	3,63	3,66	3,75	3,93	4,14	3,80

Fase	Desverdecimento		Refrigeração			Comercialização	
Temperatura	30°C		5-6°C			Ambiente	
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	3,53	3,58	3,69	3,67	3,77	3,73	3,80
Etil 10µL/L	3,66	3,78	3,82	3,86	3,69	4,02	4,04
Ethrel 250µL/L	3,61	3,58	3,83	3,65	3,83	3,74	3,83
Ethrel 500µL/L	3,63	3,64	3,92	3,81	3,79	3,86	3,80
Ethrel 750 µL/L	3,53	3,65	3,76	3,75	4,04	3,89	3,77
Ethrel 1000µL/L	3,68	3,53	4,03	3,70	3,83	3,91	3,64
Ethrel 2000µL/L	3,67	3,76	3,66	3,95	4,04	3,97	4,16

Tabela 7: Médias dos valores de cor atribuídas à tangerinas Murcote, nos períodos de desverdecimento, armazenamento refrigerado e comercialização.

Fase	Desverdecimento	Refrigeração			Comercialização		
Temperatura	25°C	5-6°C			Ambiente		
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	8,4	7,0	6,8	5,8	5,6	4,6	3,2
Etil 10µL/L	8,5	7,5	7,1	6,0	5,5	4,1	4,0
Ethrel 250µL/L	7,2	6,5	5,7	5,1	5,1	3,0	2,8
Ethrel 500µL/L	8,2	7,7	6,7	5,8	5,8	4,4	3,7
Ethrel 750 µL/L	8,3	7,8	6,7	5,5	5,4	3,5	3,2
Ethrel 1000µL/L	8,4	7,3	6,6	5,8	5,3	3,4	3,0
Ethrel 2000µL/L	8,7	7,5	6,8	5,5	5,2	3,7	3,3

Fase	Desverdecimento	Refrigeração			Comercialização		
Temperatura	30°C	5-6°C			Ambiente		
	3 dias	1ª sem.	2ª sem.	4ª sem.	5ª sem.	1ª sem.	2ª sem.
Controle	9,1	8,2	8,2	7,5	7,3	6,6	4,5
Etil 10µL/L	8,3	8,0	7,7	6,6	6,5	4,2	3,8
Ethrel 250µL/L	7,2	7,2	6,5	5,9	5,9	4,8	4,1
Ethrel 500µL/L	8,2	7,7	7,3	6,9	6,5	4,9	4,0
Ethrel 750 µL/L	8,0	7,5	7,1	6,2	5,7	3,5	3,0
Ethrel 1000µL/L	8,3	8,0	7,6	7,1	6,9	4,5	3,3
Ethrel 2000µL/L	8,7	8,2	8,1	7,4	7,2	6,3	4,2

Em se tratando de dados de natureza química, muitas vezes o estudo acaba sendo feito sem um planejamento experimental adequado e muitas variáveis acabam sendo medidas em laboratório sem necessidade. Algumas delas podem não trazer informação relevante ao conjunto, ou podem ainda estar muito correlacionadas entre si [36]. Dessa forma, pressupondo que isso poderia ter ocorrido nesse caso, realizou-se uma análise prévia de variância, com o objetivo de se analisar a relevância das análises químicas realizadas na discriminação entre os tipos de tratamentos aplicados e na temperatura. Por meio dessa, é possível verificar de que maneira a variação entre os tipos de tratamento, tempo ou temperatura de estudo (efeitos) é significativa. A Análise de Variância é uma ferramenta estatística que estuda medidas dependentes de vários tipos de efeitos simultâneos, com o objetivo de decidir quais destes efeitos são realmente significativos [36, 37].

Por meio deste método, calcula-se a soma dos quadrados de cada um dos efeitos em questão e, dividindo este valor pelo número de graus de liberdade (número de parâmetros independentes para cada efeito), tem-se o chamado Quadrado Médio. Os valores de Quadrados Médios para cada efeito calculados serão utilizados na determinação do valor F. Este é um parâmetro estatístico utilizado para testar as hipóteses em análises de variância [36]. Utilizando uma confiança de 95%, compara-se os valores de F obtidos pelos dados experimentais aos valores teóricos. Um valor experimental de F superior ao teórico indica que o respectivo efeito é significativo, ou seja, a variância existente entre as amostras neste efeito é significativa. Nesse trabalho, foram considerados como efeitos o tempo e os tipos de tratamento, com o objetivo de facilitar a interpretação dos resultados obtidos. O efeito da temperatura, bem como os outros resultados obtidos podem ser vistos nas Tabelas 8 - 14. De acordo com estas, mostraram-se significativas de maneira simultânea para os dois

efeitos em questão as variáveis Teor de ácido ascórbico, °Brix, % de Perda de Peso e Coloração (Vide Tabelas 8, 9, 10 e 14, respectivamente).

Tabela 8: Análise de Variância, nas duas temperaturas utilizadas, tendo como efeitos o Tempo e os Tratamentos aplicados segundo a variável Teor de Ácido Ascórbico.

Fonte da variação	*Soma Quad (T=25°C)*	*Soma Quad (T=30°C)*	*Graus liberdade*	*Quadrados Médios (T=25°C)*	*Quadrados Médios (T=30°C)*	*F (T=25°C)*	*F (T=30°C)*	*Valor-P (T=25°C)*	*valor-P (T=30°C)*	*F crítico*
Tempo	176,8253	166,419	6	29,471	27,7365	**35,3713**	**21,5858**	1,14E-13	1,4E-10	2,3637
Tratamento	70,32531	32,8649	6	11,721	5,47748	**14,0675**	**4,262821**	3,56E-08	0,00242	2,3637
Resíduo	29,99469	46,258	36	0,8332	1,28494					
Total	277,1453	245,542	48							

Tabela 9: Análise de Variância, nas duas temperaturas utilizadas, tendo como efeitos o Tempo e os Tratamentos aplicados segundo a variável Porcentagem de Perda de Peso.

Fonte da variação	*Soma Quad (T=25°C)*	*Soma Quad (T=30°C)*	*Graus liberdade*	*Quadrados Médios (T=25°C)*	*Quadrados Médios (T=30°C)*	*F (T=25°C)*	*F (T=30°C)*	*Valor-P (T=25°C)*	*valor-P (T=30°C)*	*F crítico*
Tempo	448,032	409,723	6	74,67204	68,28715	**260,06**	**488,9724**	4,14E-28	5,93E-33	2,3637
Tratamento	47,6315	13,625	6	7,938584	2,270831	**27,6476**	**16,26036**	4,36E-12	5,94E-09	2,3637
Resíduo	10,3368	5,0276	36	0,287134	0,139654					
Total	506,001	428,375	48							

Tabela 10: Análise de Variância, nas duas temperaturas utilizadas, tendo como efeitos o Tempo e os Tratamentos aplicados segundo a variável °Brix.

Fonte da variação	*Soma Quad (T=25°C)*	*Soma Quad (T=30°C)*	*Graus liberdade*	*Quadrados Médios (T=25°C)*	*Quadrados Médios (T=30°C)*	*F (T=25°C)*	*F (T=30°C)*	*Valor-P (T=25°C)*	*valor-P (T=30°C)*	*F crítico*
Tempo	9,04204	6,88245	6	1,507007	1,147075	**14,4257**	**9,57705**	2,63E-08	2,67E-06	2,3637
Tratamento	2,82204	6,80245	6	0,47034	1,133741	**4,50228**	**9,465733**	0,001687	3,01E-06	2,3637
Resíduo	3,76082	4,31184	36	0,104467	0,119773					
Total	15,6249	17,9967	48							

Tabela 11: Análise de Variância, nas duas temperaturas utilizadas, tendo como efeitos o Tempo e os Tratamentos aplicados segundo a variável °Brix/Acidez.

Fonte da variação	*Soma Quad (T=25°C)*	*Soma Quad (T=30°C)*	*Graus liberdade*	*Quadrados Médios (T=25°C)*	*Quadrados Médios (T=30°C)*	*F (T=25°C)*	*F (T=30°C)*	*Valor-P (T=25°C)*	*valor-P (T=30°C)*	*F crítico*
Tempo	77,358	138,791	6	12,89306	23,13184	**3,69407**	**11,26066**	0,005808	4,7E-07	2,3637
Tratamento	34,361	33,691	6	5,726871	5,61517	**1,64084**	**2,733483**	0,164337	0,027183	2,3637
Resíduo	125,65	73,9518	36	3,490204	2,054218					
Total	237,37	246,434	48							

Tabela 12: Análise de Variância, nas duas temperaturas utilizadas, tendo como efeitos o Tempo e os Tratamentos aplicados segundo a variável Acidez Total.

Fonte da variação	*Soma Quad (T=25°C)*	*Soma Quad (T=30°C)*	*Graus liberdade*	*Quadrados Médios (T=25°C)*	*Quadrados Médios (T=30°C)*	*F (T=25°C)*	*F (T=30°C)*	*Valor-P (T=25°C)*	*valor-P (T=30°C)*	*F crítico*
Tempo	0,17636	0,26759	6	0,029394	0,044599	**5,06841**	**15,2235**	0,000737	1,36E-08	2,3637
Tratamento	0,03128	0,08745	6	0,005213	0,014575	**0,89887**	**4,97504**	0,89887	0,000843	2,3637
Resíduo	0,20878	0,10547	36	0,005799	0,00293					
Total	0,41642	0,46051	48							

Tabela 13: Análise de Variância, nas duas temperaturas utilizadas, tendo como o Tempo e os Tratamentos aplicados segundo a variável pH.

Fonte da variação	*Soma Quad (T=25°C)*	*Soma Quad (T=30°C)*	*Graus liberdade*	*Quadrados Médios (T=25°C)*	*Quadrados Médios (T=30°C)*	*F (T=25°C)*	*F (T=30°C)*	*Valor-P (T=25°C)*	*valor-P (T=30°C)*	*F crítico*
Tempo	0,4362	0,4748	6	0,07271	0,079133	**8,20951**	**6,81252**	1,25E-05	6,91E-05	2,3637
Tratamento	0,1101	0,1972	6	0,01836	0,032871	**2,07256**	**2,82987**	0,080952	0,023213	2,3637
Resíduo	0,3188	0,4182	36	0,00886	0,011616					
Total	0,8652	1,0902	48							

Tabela 14: Análise de Variância, nas duas temperaturas utilizadas, tendo como efeitos o Tempo e os Tratamentos aplicados segundo a variável Cor.

Fonte da variação	*Soma Quad (T=25°C)*	*Soma Quad (T=30°C)*	*Graus liberdade*	*Quadrados Médios (T=25°C)*	*Quadrados Médios (T=30°C)*	*F (T=25°C)*	*F (T=30°C)*	*Valor-P (T=25°C)*	*valor-P (T=30°C)*	*F crítico*
Tempo	134,956	107,34	6	22,49259	17,88986	**309,203**	**116,235**	1,97E-29	4,73E-22	2,3637
Tratamento	5,0498	13,016	6	0,841633	2,169388	**11,5698**	**14,095**	3,47E-07	3,48E-08	2,3637
Resíduo	2,61878	5,5408	36	0,072744	0,153912					
Total	142,624	125,9	48							

Para mostrar de que maneira as análises físico-químicas realizadas nos frutos variavam com o tempo, montou-se primeiramente uma matriz onde foram tomadas como amostras os sete de tratamentos aplicados nos sete períodos de estudo e como variáveis os sete tipos de análises realizadas nas duas temperaturas do tratamento (Figura 6).

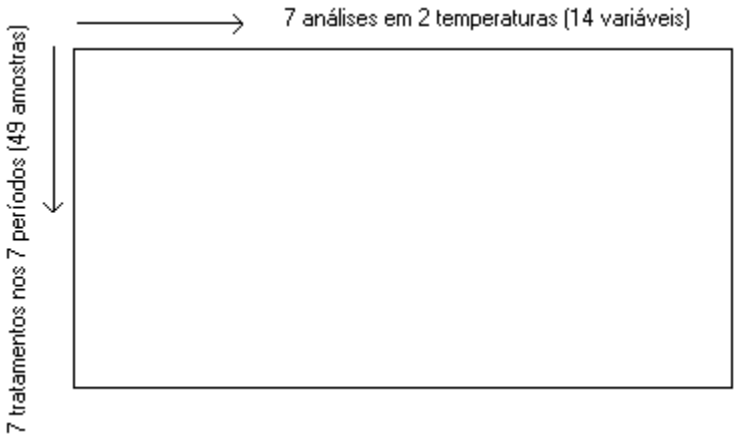


Figura 6: Esquema de construção da primeira matriz de dados (49 X 14)

Nesse conjunto, de acordo com as conclusões obtidas por meio da ANOVA, optou-se pela escolha das variáveis Porcentagem de perda de peso, as notas de coloração (usou-se uma cor média entre as cores A, B e C), o valor de °Brix e teor de ácido ascórbico, mais

significativas na discriminação entre as classes do conjunto. Essas variáveis foram selecionadas nas duas temperaturas em questão, totalizando um conjunto de 8 variáveis e 49 amostras.

Pela análise PCA, seis componentes principais descreveram um total de 99,30% da variância do conjunto original de dados.

Foi possível verificar por meio desta, uma significativa discriminação entre os períodos de estudo, como pode ser visto na Figura 7, relativa aos escores obtidos para as duas primeiras componentes principais (sendo a primeira delas a responsável pela efetiva separação). Nesse gráfico, pode-se perceber que o período de desverdecimento apresenta-se bastante diferenciado dos demais, enquanto que os dois primeiros períodos de refrigeração (1^a e 2^a semanas) apresentam características muito semelhantes. O mesmo observa-se para os dois períodos de comercialização.

A análise conjunta deste gráfico com o dos pesos (Figura 8), mostra como esses períodos estão relacionados às análises químicas. Pode-se notar, por exemplo, que a porcentagem de perda de peso aumenta com o tempo, enquanto que a avaliação de cor diminue. Isso é bastante razoável, uma vez que a perda de peso tende a aumentar em função da maior perda de umidade dos frutos com o tempo. As notas obtidas na avaliação da coloração da casca estão inversamente relacionadas com a eficiência do processo de desverdecimento, ou seja, quanto menores, mais amarelados os frutos, e consequentemente, mais eficiente o processo. O teor de ácido ascórbico diminue com o tempo e volta a aumentar durante o período de comercialização. Isso deve ocorrer devido ao fato de, neste período, o produto sofrer uma grande perda de peso e umidade e, consequentemente um aumento na concentração tanto do ácido quanto dos sólidos solúveis (°Brix).

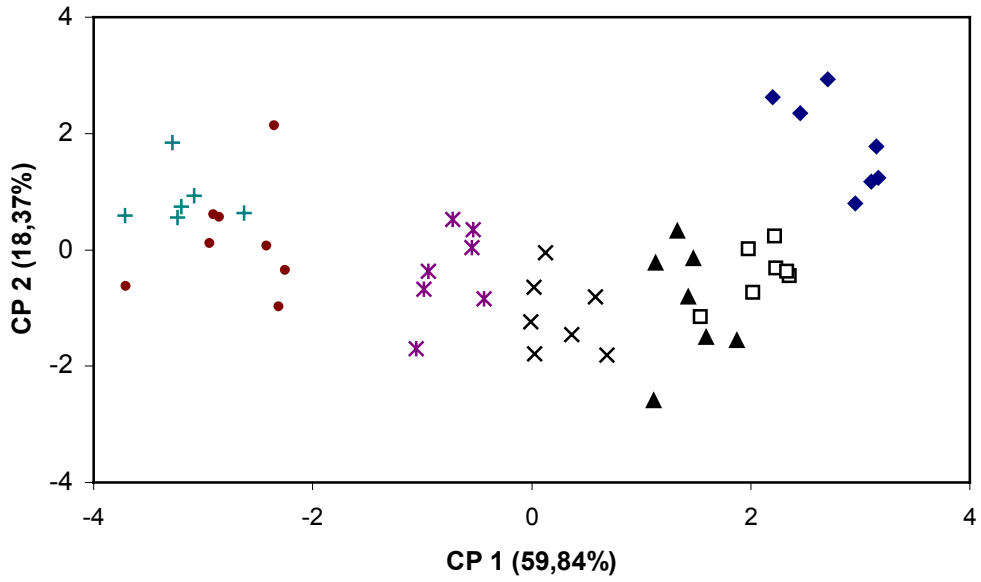


Figura 7: Escores – CP1 x CP2 – grupos discriminados segundo o tempo do processo

♦ - Desverdecimento, □ - 1ª. Sem. Refrigeração, ▲- 2ª. Sem. Refrigeração, × - 4ª. Sem. Refrigeração, ✳- 5ª. Sem. Refrigeração, ● - 1ª. Sem. Comercialização, + - 2ª. Sem. Comercialização.

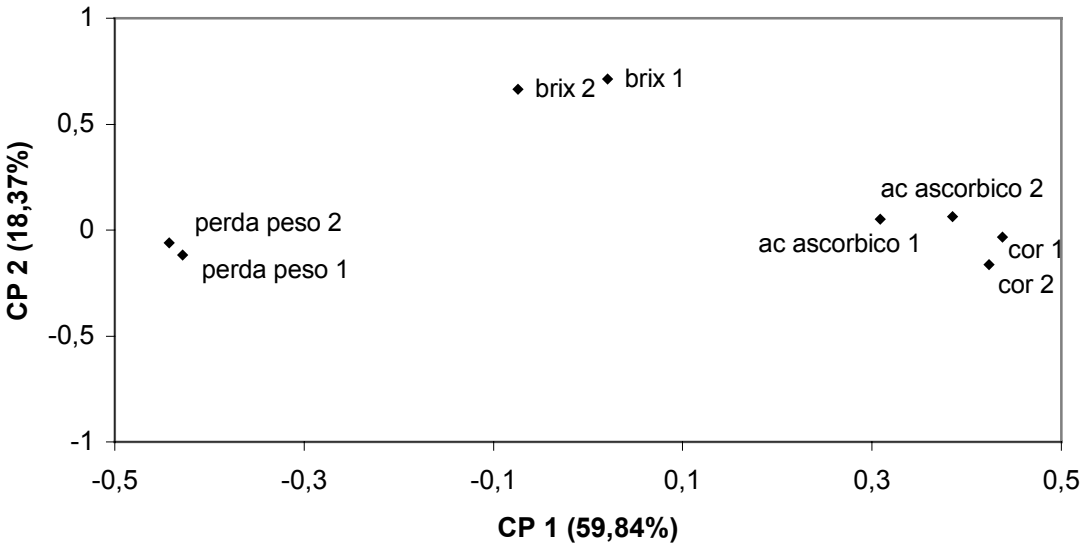


Figura 8: Pesos – CP1 x CP2 – classe tempo

Uma nova maneira de visualizar estes dados, porém agora considerando as fases do processo como um todo, ou seja, desverdecimento, refrigeração e comercialização pode ser vista na Figura 9 (Escores – CP1 x CP2). Pode-se notar mais claramente como estas fases separam-se significativamente, mostrando assim possuírem características bem distintas entre si.

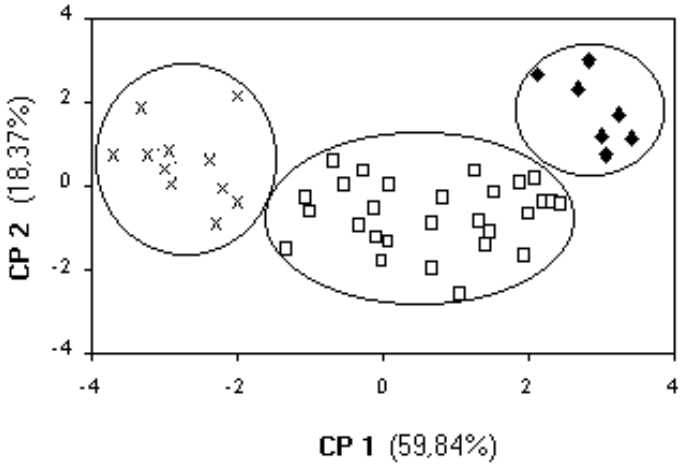


Figura 9: Escores – CP1 x CP2 – grupos discriminados segundo as fases do processo

♦ - Desverdecimento , □ - Refrigeração, ✕ - Comercialização

O dendrograma que nos mostra a separação entre as três fases do processo pode ser visto na Figura 10. O tipo de conexão interagrupamentos que melhor representou este conjunto de dados foi o Centróide. Esse é um tipo de conexão feita pelo centro do agrupamento, onde a distância entre um recém formado grupo A-B e outro grupo previamente formado C é calculada pela Equação 16, onde n_i é o número de amostras no grupo i.

$$d_{ab} \Rightarrow c = \sqrt{\frac{n_a d_{ac}^2}{n_a + n_b} + \frac{n_b d_{bc}^2}{n_a + n_b} + \frac{n_a n_b d_{ab}^2}{(n_a + n_b)^2}} \qquad \textbf{(16)}$$

n_i = número de amostras do grupo i

d_{ab} = distância entre as amostras a e b

Distâncias interagrupamentos pequenas, indicam alta similaridade entre os mesmos.

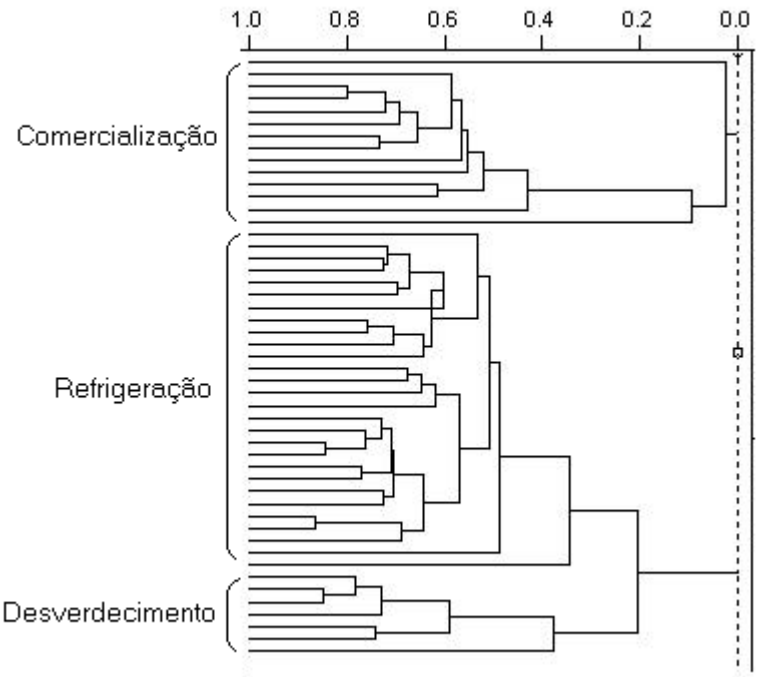


Figura 10: Dendrograma – Conexão Centróide (Os tempos do processo são representados pelas linhas verticais e a similaridade interagrupamentos pelas horizontais, ou no comprimento dos ramos)

O dendrograma obtido para este conjunto (Figura 10) apresenta dois grandes agrupamentos. O superior, representando as amostras na fase de comercialização e o

inferior, dividido em dois outros ramos, representando a fase de refrigeração ao centro e de desverdecimento abaixo. Todas as amostras do conjunto foram separadas de acordo com suas características, sendo que nenhuma foi classificada erroneamente. As amostras da fase de refrigeração apresentaram-se separadas em diversos subgrupos. Essa separação tem relação à mudança ocasionada nas propriedades dos frutos conforme aumentamos o período do processo, ou seja, as amostras das 1as semanas possuem características um tanto diferenciadas das últimas semanas do processo. As amostras das primeiras semanas do período de refrigeração assemelham-se mais às da fase de desverdecimento enquanto que as últimas assemelham-se às relativas ao período de comercialização. Essas amostras do início do processo de refrigeração aparecem no dendrograma na parte inferior, próximas ao grupo das amostras de desverdecimento, enquanto que as dos últimos períodos de refrigeração aparecem mais acima, próximas às amostras do período de comercialização.

Em uma segunda matriz, formada em suas linhas pelos tratamentos aplicados nas duas temperaturas em questão e em suas colunas pelas análises propriamente ditas nos sete períodos (Figura 11), pode-se avaliar o efeito dos gases de tratamento e da temperatura nas características dos frutos em função do tempo. Foram utilizadas as mesmas quatro variáveis da análise anterior, porém para os sete períodos de estudo, totalizando dessa forma 28 variáveis, em um conjunto de 14 amostras (7 tratamentos em 2 temperaturas).

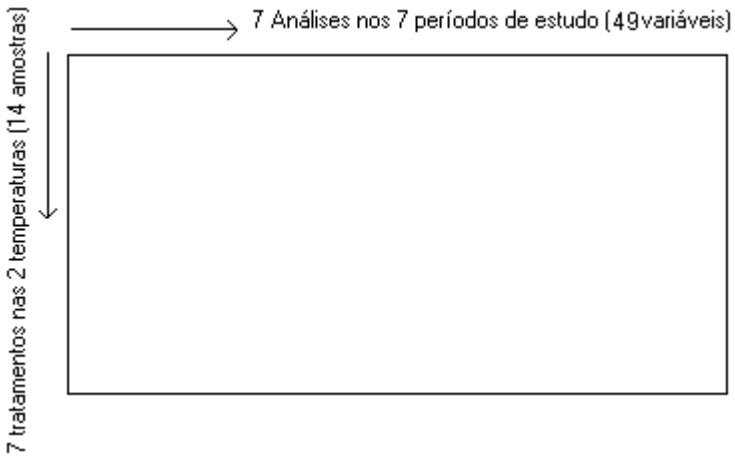


Figura 11: Esquema de construção da segunda matriz de dados (14 X 49).

A discriminação entre as duas temperaturas utilizadas no processo pode ser facilmente observada a partir do dendrograma mostrado na Figura 12. O tipo de conexão utilizada neste foi pela Média do Grupo. Este também é um método de conexão feita pelo centro do agrupamento, onde este centro é calculado pela Equação 17.

$$d_{ab} \Rightarrow c = \frac{n_a d_{ac}}{n_a + n_b} + \frac{n_b d_{bc}}{n_a + n_b} \qquad \textbf{(17)}$$

Onde:

n_i = número de amostras do grupo i

d_{ab} = distância entre as amostras a e b

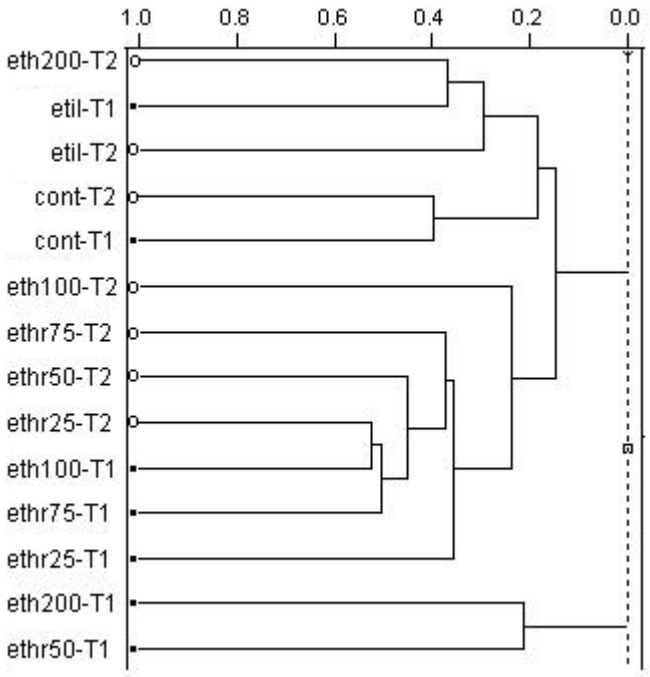


Figura 12: Dendrograma – Conexão pela Média do Grupo (os tempos do processo são representados pelas linhas verticais e a similaridade interagrupamentos pelas horizontais, ou no comprimento dos ramos)

▪ - T = 25°C**, o –** T = 30°C

Neste, apesar das amostras não se encontrarem formando agrupamentos distintos de acordo com as temperaturas do processo, é possível observar uma certa tendência, onde as amostras relativas aos tratamentos a 25°C encontram-se numa posição inferior no dendrograma às das amostras submetidas ao tratamento a 30°C. Algo que é possível notar de forma bastante clara é a similaridade entre as amostras do controle e às tratadas com Etil, para ambas temperaturas, indicando que a temperatura não influi fortemente nos resultados obtidos nesses tratamentos. O contrário pode-se dizer com as amostras submetidas ao tratamento com o Ethrel 500 e 2000μL/L, que se apresenta com características bastante diferenciadas para as duas temperaturas de estudo.

As mesmas conclusões podem ser tiradas ao observar os gráficos de Escores e Pesos (Figuras 13 e 14) obtidos pela da Analise por Componentes Principais. Existe uma tendência bastante clara de agrupamento entre as amostras submetidas ao mesmo tipo de tratamento, com exceção para aquelas citadas anteriormente, tratadas com Ethrel 500 e 2000μL/L. Ao levar em conta que as características ótimas para a comercialização dos frutos é determinada por uma coloração amarelada, um teor de ^{o}Brix alto, porcentagem de perda de peso e teor de acido ascórbico baixos, por meio destes gráficos podemos concluir que, com exceção dos frutos não submetidos a tratamento (controle) e dos tratados com Etil, todos os outros não apresentaram diferenças significativas entre si.

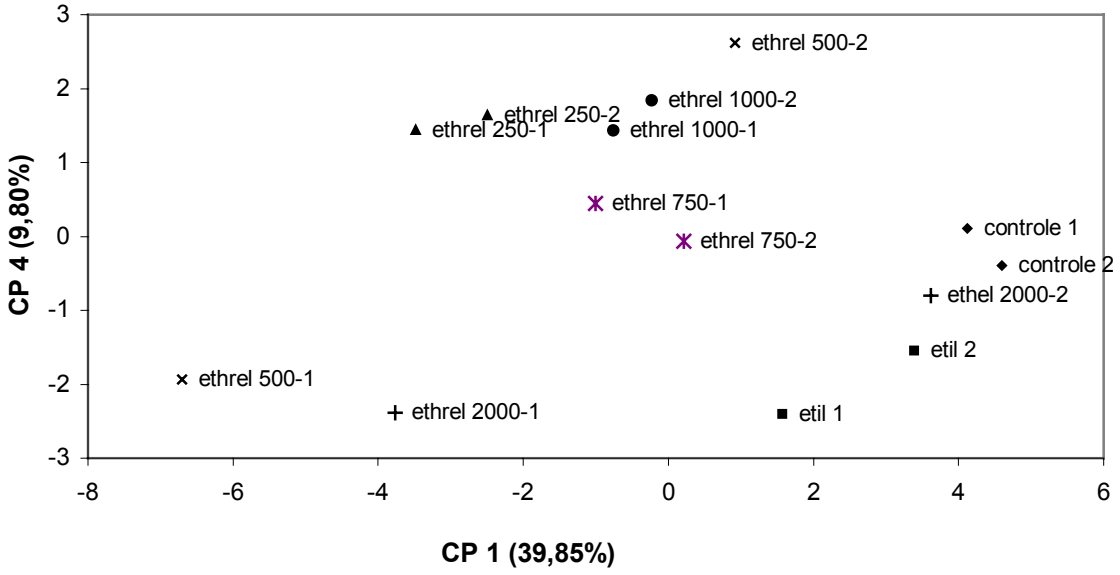


Figura 13: Escores – CP1 x CP4 – grupos discriminados segundo os tipos de tratamentos utilizados

◆ controle; ■ etil; ▲ - ethrel250ppm; × ethrel 500ppm; ∗ ethrel 750ppm; ● ethrel 1000ppm; + ethrel 2000ppm

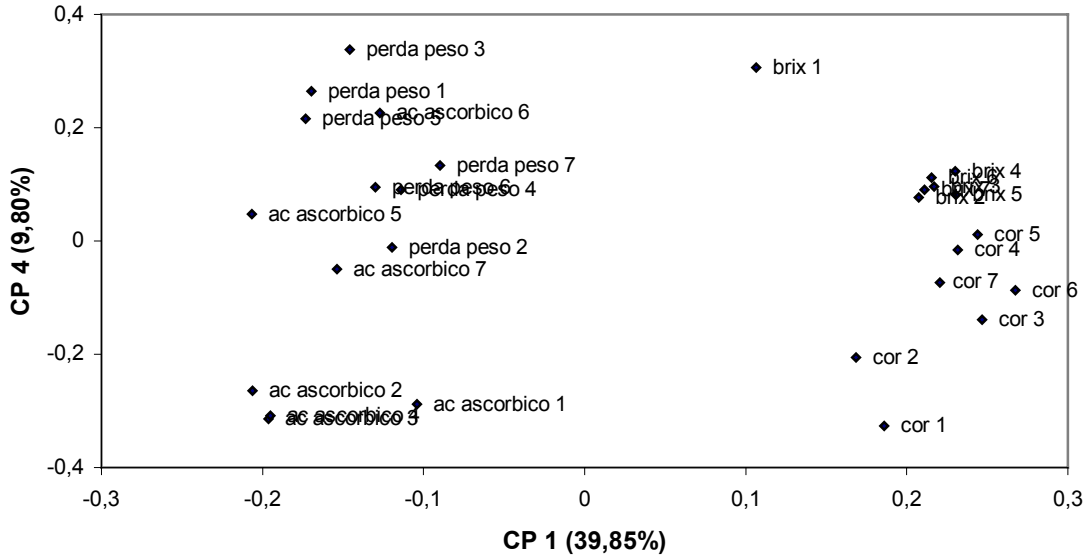


Figura 14: Pesos – CP1 x CP4 – classe: tratamento

4.2 Análise de Ordem Superior (PARAFAC)

Esse tipo de análise (assim como os outros métodos N-Way), tem como grande vantagem o fato de possibilitar o estudo simultâneo dos diversos modos (representados pelas dimensões do arranjo) de um conjunto de dados de n-dimensional. Isso significa, por exemplo, que é possível estudar um conjunto complexo como o apresentado anteriormente, de uma única vez, utilizando um único arranjo e, dessa forma, reduzindo o tempo de análise do mesmo.

Nesse caso específico, o conjunto de dados era constituído por quatro modos: análises químicas, determinadas em diversos períodos (tempo), tipos de tratamento aplicados aos frutos e temperaturas do processo de desverdecimento.

Dessa forma, o arranjo tetradimensional formado era constituído em cada uma de suas dimensões por cada um dos quatro modos, finalizando um arranjo (7 x 4 x 7 x 2), ou seja, sete variáveis no modo tempo, 4 no modo análises (foram utilizadas apenas as mesmas análises selecionadas no método anterior – PCA, para que uma comparação entre os dois métodos pudesse ser feita), 7 no modo tratamento e 2 no modo temperatura. As análises (variáveis) selecionadas foram teor de ácido ascórbico, °Brix, perda de peso (%) e nota de coloração.

O método utilizado, PARAFAC, foi então aplicado a esse arranjo, onde o modo tratamento foi centrado na média (uma vez que os valores deste modo possuiam a mesma magnitude e o modo análises foi auto-escalado.

Esse método, como dito anteriormente, baseia-se na decomposição dos modos do arranjo tetradimensional original **X** pelo método da decomposição por valor singular (SVD) em quatro novas matrizes bidimensionais **A**, **B, C** e **D**, cada uma relacionada a um modo do arranjo original.

A decomposição foi feita utilizando 2 componentes principais e restrição de não-negatividade em todos os modos, com exceção do modo tratamento, totalizando 22,79% de variância explicada e um diagnóstico de consistência de "core" de 66,72%. Esse diagnóstico, como foi dito anteriormente, é um indicativo do ajuste do modelo criado, sendo maior quanto melhor ou mais estável for este ajuste.

Estes valores obtidos são bastante baixos, indicando que o modelo criado não é muito bom. Porém, isso se deve ao fato do conjunto de dados em questão não ser adequado a este tipo de método. Em geral, para se obter um modelo de PARAFAC realmente bom, é necessária uma amostragem maior.

A análise do modelo pode ser feita visualmente pelos gráficos dos fatores de cada um dos modos. Estes gráficos podem ser vistos nas Figuras 15a - d, relativas aos modos tempo, análises, tratamento e temperatura respectivamente.

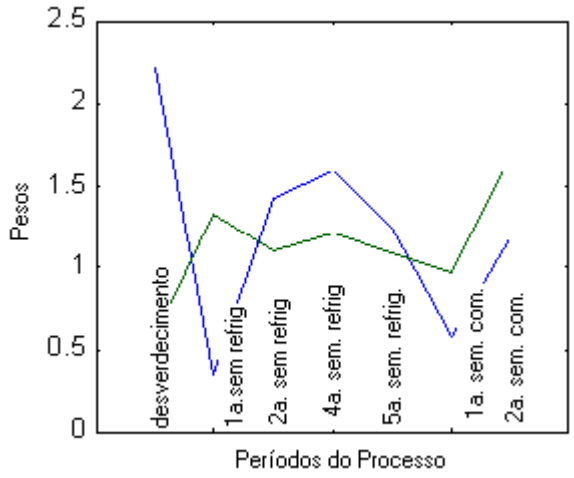
Ao visualizar inicialmente a Fig. 15d, é possível perceber de forma bastante clara o comportamento contrário entre as duas componentes principais, onde a primeira aumenta no sentido da temperatura 2 (30°C) e a segunda aumenta nos sentido da temperatura 1 (25°C). Pode-se dizer portanto que a primeira componente principal está diretamente relacionada à temperatura de 30°C, enquanto que a segunda relaciona-se à de 25°C. Dessa forma, ao analisar os outros gráficos, tendo como base esse pressuposto, é possível tirar conclusões a respeito dos outros modos com relação à temperatura.

Na Fig. 15a, por exemplo, pode-se distinguir claramente que os períodos onde há uma diferença considerável de comportamento com a temperatura são os dois primeiros, que representam o chamado período de desverdecimento e início do período de refrigeração. Nos períodos seguintes, a temperatura já não é mais tão importante no processo como um todo, uma vez que não provoca grandes alterações.

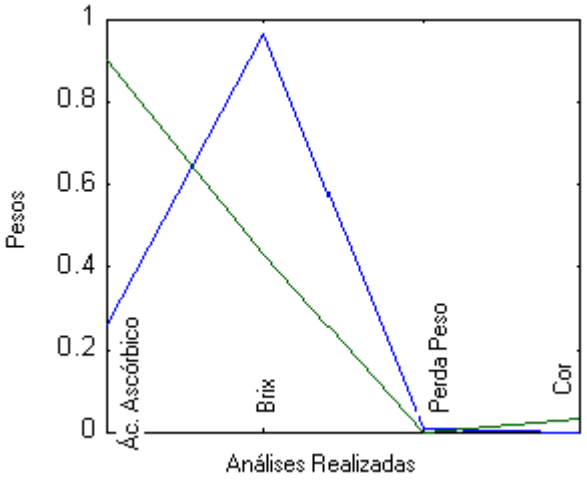
A Fig. 15b, relativa às análises físico-químicas realizadas mostra-nos uma significativa variação com a temperatura nas análises de °Brix e Ácido Ascórbico, enquanto que para a porcentagem de perda de peso e nota de coloração a temperatura já não é importante.

Por último, a Fig. 15c, relativa ao modo tratamento aplicado aos frutos traz-nos conclusões similares às obtidas anteriormente pelos métodos bidimensionais. Pode-se ver, por exemplo, que existe uma diferença considerável com relação às temperaturas para os tratamentos com Ethrel 500μL/L, Ethrel 1000 μL/L e Ethrel 2000 μL/L, enquanto que o

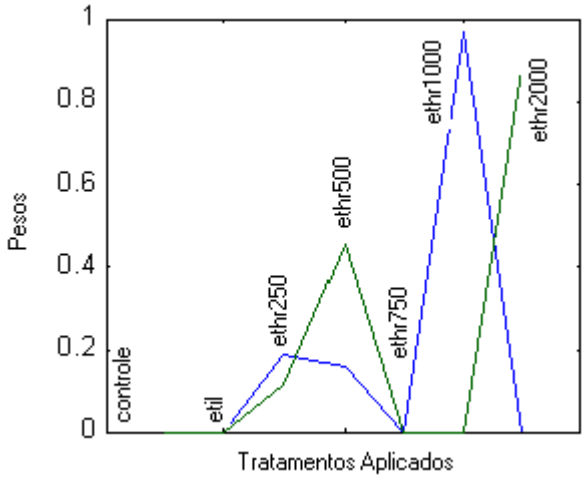
Controle e Etil são bastante similares. Esse resultado pôde ser visto no dendrograma apresentado na Figura 12.



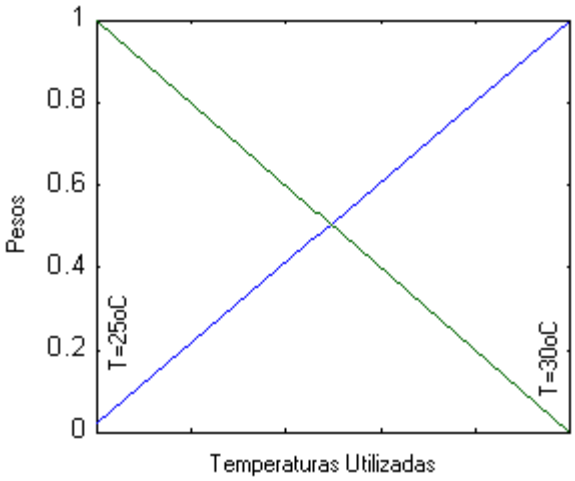
(Fig. 15a)



(Fig. 15b)



(Fig. 15c)



(Fig. 15d)

Figura 15: Resultados obtidos no modelo PARAFAC, usando 2 componentes principais (– PC1 e – PC2) e restrição de não-negatividade em todos os modos: (a) modo tempo, (b) modo análises, (c) modo tratamento, (d) modo temperatura.

5. Conclusões

Pelos métodos multivariados de análise foi-nos possível estudar o processo artificial de desverdecimento de maneira bastante simples e com ótimos resultados. Em se tratando de um conjunto de dados bastante grande e complexo, a utilização de métodos matemáticos com esse objetivo é de grande valia, uma vez que sem estes fica bastante difícil a análise conjunta de todas as variáveis envolvidas no processo. Foi possível tirarmos conclusões a respeito das variações ocorridas nos frutos com o aumento da temperatura do processo, tipo de tratamento artificial aplicado e períodos de estudo, além de quais variáveis eram realmente relevantes.

Entre os tratamentos aplicados, com exceção do controle e etil, que apresentaram um comportamento relativamente diferenciado dos demais, dentre aqueles onde foi utilizado o Ethrel, poucas alterações foram observadas nas características dos frutos. O que foi possível observarmos é que alguns tratamentos mostraram-se bastante diferenciados quando aplicados a temperaturas diferentes, como no caso do Ethrel a 500 e 2000μL/L.

Com exceção destes últimos, todos os tratamentos com Ethrel, a 250, 750 e 1000 μL/L, para ambas temperaturas de estudo, conduziram os frutos à características ótimas, como alto teor de Brix, baixo teor de Ácido Ascórbico e coloração bastante alaranjada, sendo o único inconveniente observado em alguns casos, a alta perda de peso dos frutos. Para o tratamento com Ethrel 500μL/L, apenas os frutos submetidos à temperatura de 25^{o}C apresentou bons resultados.

Sob todos estes aspectos, podemos recomendar, levando em conta o fato da maior praticidade no uso do Ethrel e os resultados obtidos, a utilização do mesmo, em qualquer

uma das concentrações mencionadas que conduziram aos melhores resultados, em qualquer uma das temperaturas.

Como pode ser visto, os métodos bi e multidimensionais em geral podem ser utilizados conjuntamente, uma vez que as conclusões obtidas por meio destes são algumas vezes complementares. As vantagens entre um ou outro, irá depender da complexidade dos dados a serem analisados e das informações que se queira tirar dos mesmos. Neste caso específico, não foi possível ajustar um bom modelo PARAFAC ao conjunto, porém uma grande vantagem do método é a possibilidade de analisar todos os modos simultaneamente, de uma maneira muito mais rápida.

Capítulo 3 - *Determinação Quantitativa do Óleo de Soja Epoxidado utilizando Espectroscopia na Região do Infravermelho Próximo.*

1 - Objetivos

A partir de um conjunto de dados proveniente da Henkel S/A Indústrias Químicas, constituído por espectros registrados na região do Infravermelho Próximo de amostras de óleo de soja epoxidado, foram construídas curvas analíticas para alguns analitos de interesse industrial, tendo como principais focos deste trabalho o uso do método PLS de calibração multivariada e o estudo comparativo entre alguns métodos de seleção de variáveis [39].

2. Introdução

2.1 O Óleo de Soja Epoxidado

O óleo de soja é um triglicerídeo formado pela combinação de seus ácidos graxos constituintes (aproximadamente 14% de ácido esteárico, 23% de ácido oleico, 55% de ácido linoleico e 8% de ácido linolênico). Dentre estes ácidos contendo 18 átomos de carbono, três deles (oleico, linoleico e linolênico) contêm 1, 2 e 3 insaturações por molécula respectivamente (Figura 16)[40].

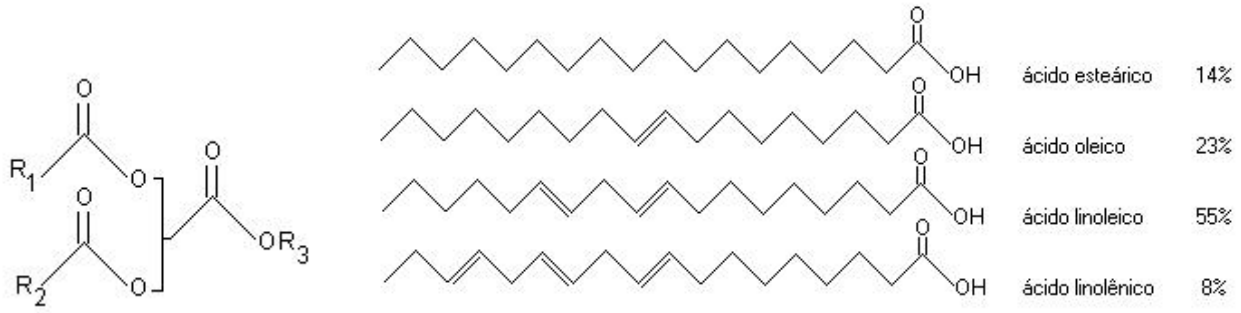


Figura 16: Estrutura dos ácidos graxos constituintes do óleo de soja

A modificação química no óleo de soja comercial, pela da epoxidação, por exemplo, pode alterar suas propriedades (reatividade) visando determinadas aplicações industriais.

O óleo de soja epoxidado (ESO – *epoxidized soybean oil*) é utilizado industrialmente como plastificante e estabilizante térmico para filmes de PVC (cloreto de polivinila). A epoxidação consiste em uma reação de adição na ligação dupla de um composto insaturado C=C com um composto com oxigênio ativo, normalmente um peróxido ou um perácido, adicionando assim um átomo de oxigênio ao composto original e convertendo a ligação dupla em um anel epóxido de três membros (anel oxirano) (Figura 17) [41]. O óleo de soja epoxidado tem sido amplamente utilizado em embalagens alimentícias como plastificante, aumentando sua flexibilidade, e como estabilizante térmico, minimizando a decomposição em produtos do PVC. Materiais como PVC e poliestireno normalmente contêm óleo epoxidado em níveis de 0,1 a 27% [42].

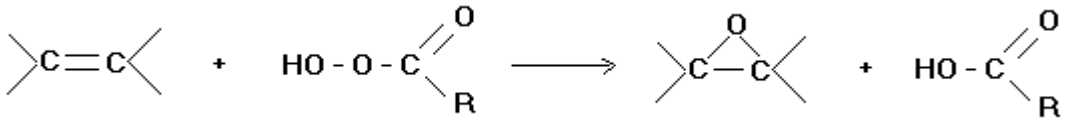


Figura 17: Reação de adição na ligação dupla de um composto insaturado com um composto com oxigênio ativo, formando um anel epóxido de três membros.

O PVC sofre uma reação de decomposição com a temperatura (desidrocloração), produzindo quase que exclusivamente HCl, além de ligações duplas conjugadas. O HCl liberado possui um efeito autocatalítico na reação de desidrocloração [43]. Além disso, essa decomposição é acompanhada pela alteração da coloração do polímero, indo do amarelo até o preto, passando pelo vermelho e marrom [44].

O papel dos diferentes tipos de estabilizantes (como o ESO), é prevenir a dehidrocloração e consequente perda de coloração do polímero pela ligação com o HCl liberado pelo filme (Figura 18)[44].

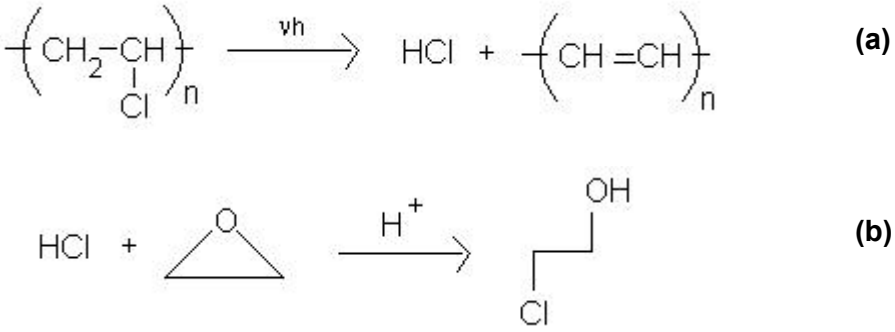


Figura 18: a-) Degradação térmica do PVC pela luz solar eliminando HCl; b-) Reação do anel oxirano com HCl gerado em a) inibindo o processo de degradação.

Com o objetivo de acompanhar o processo de epoxidação dos óleos vegetais, faz-se necessário dosar alguns índices relativos à eficiência deste processo. Os teores de epóxido, iodo e água são alguns destes índices. O primeiro está diretamente relacionado com a

propriedade estabilizante do produto, ou seja, quanto maior o teor de epóxido, mais eficiente é a estabilização térmica promovida pelo aditivo. O teor de iodo é um indicativo da quantidade de insaturações presentes no óleo de soja que a princípio é desconhecida. As duplas ligações do óleo são halogenadas com solução de bromo e o excesso deste reagente é dosado por iodometria. Outro analito, também dosado, é a água. Esta é proveniente dos resíduos do processo de lavagem a que o produto final é submetido. A presença de um nucleófilo como a água, por exemplo, em meio ácido, pode promover a abertura do anel epóxido gerando uma série de subprodutos, como dióis, dímeros, ésteres etc. Dessa forma sua concentração deve ser mínima para evitar essa degradação (Figura 19) [41].

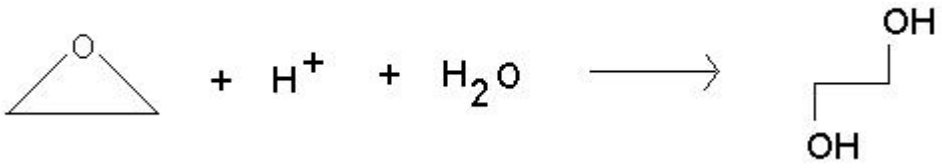


Figura 19: Degradação dos grupos epóxido pela água

2.2 A Espectroscopia na Região do Infravermelho Próximo

Nesse trabalho, foram construídos modelos de calibração a partir dos espectros registrados na região do infravermelho próximo (NIRS – Near Infrared Spectroscopy) dos três analitos acima mencionados.

A espectroscopia na região do Infravermelho Próximo tem sido amplamente utilizada nos últimos anos em diversos segmentos da indústria, como o alimentício, de rações, petroquímico, farmacêutico entre outros, com o propósito de quantificação [45-50].

Os espectros obtidos nessa região são provenientes de absorções moleculares resultantes de bandas de combinação (envolvendo 2 ou mais modos normais de vibração de um mesmo grupo funcional) e sobretons de bandas fundamentais de vibração da região do Infravermelho Médio. A atribuição de bandas nessa região não é simples, uma vez que uma única banda pode ser resultante da sobreposição de modos normais de vibração. Sendo que essa região é de certa forma limitada para a interpretação da estrutura qualitativa de compostos (ao contrário do que ocorre na região do Infravermelho Médio), ela tem sido muito utilizada para análises quantitativas de compostos contendo os grupos funcionais OH-, NH-, e CH- [51].

O NIRS apresenta diversas vantagens quando comparado aos métodos químicos tradicionais, como a rapidez, a não necessidade de preparação da amostra, o fato de ser um método não-destrutivo, além de permitir a multiplicidade de análises num espectro [52]. Porém, seu uso em análises quantitativas, popularizou-se apenas com a crescente disponibilidade de recursos computacionais e de "softwares" quimiométricos desenvolvidos para análise deste tipo de dados.

3. Materiais e Métodos

3.1. Determinação da porcentagem de água – Método de Karl Fischer

A titulação Karl Fischer é geralmente o método de referência para a determinação de baixos níveis de água em óleos. Foi utilizado neste caso um titulador KARL FISCHER AUTOMAT E547 e um MULTI DOSIMAT E415.

O reagente de Karl Fischer usado nessa titulação, é formado por iodo, dióxido de enxofre e piridina em excesso. Na presença de água, o dióxido de enxofre é oxidado pelo

iodo em ácido sulfúrico, como pode ser visto na Figura 20. A oxidação ocorre enquanto a água estiver presente no meio, sendo dessa forma possível determinar a sua quantidade inicial. Os resultados são expressos em termos de porcentagem de água e os níveis aceitáveis industrialmente neste processo devem ser inferiores a 0,3% [53].

$$H_2O + I_2 + SO_2 + 3C_5H_5N \rightarrow 2C_5H_5N.HI + C_5H_5N.SO_3$$

$$C_5H_5N.SO_3 + R - OH \rightarrow C_5H_5N.HSO_4R$$

Figura 20: Reações envolvidas na determinação de água – método Karl Fischer

3.2. Determinação do Índice de Iodo

A quantidade de iodo (índice de iodo – I.I.) foi determinado pela bromação das duplas ligações do óleo, pela adição de solução de bromo na presença de acetato de mercúrio II como catalisador. O excesso de reagente é determinado por iodometria. O resultado é expresso em gramas de iodo por 100g de amostra e é uma medida da quantidade de duplas ligações -C=C- não convertidas (Figura 21). O nível máximo aceitável para os propósitos industriais é de 4,0g de iodo/100g de amostra [54].

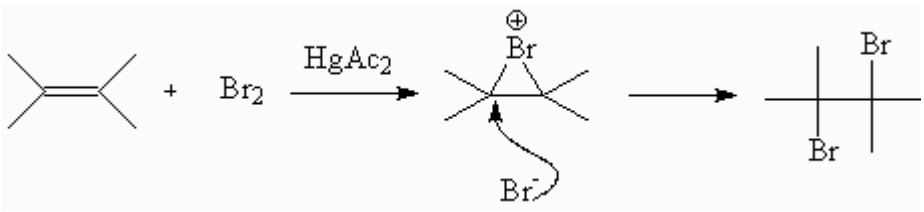


Figura 21: Reação de halogenação das duplas ligações do óleo de soja

3.3. Determinação do Índice de Epóxido

A determinação do índice de epóxido (oxigênio oxirano) é baseada na síntese do ácido bromídrico pela reação com brometo de tetraetilamônio e ácido perclórico. O ácido bromídrico rompe o anel epóxido pela bromação e consequente formação de hidroxila. O ponto final da titulação é marcado pela alteração de cor do indicador utilizado na presença de ácido bromídrico livre (produzido pelo excesso de ácido perclórico) como resultado da reação dos grupos epóxido (Figura 22). O resultado é expresso em porcentagem de oxigênio oxirano e deve ser superior a 6,3% [44].

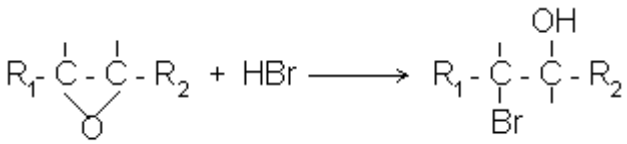


Figura 22: Reações envolvidas na determinação de epóxido (oxigênio oxirano)

3.4. Aquisição dos espectros NIR

Os espectros de absorção das amostras de óleo de soja epoxidado foram registrados "offline" na região do infravermelho próximo (de 10000 a 4500 cm^{-1}, com um incremento de 2 cm^{-1}), usando um espectrofotômetro FTIR BOMEM – MB160. O número de amostras diferiu de analito para analito (Portanto, são três conjuntos diferentes de amostras, um para cada analito). A Figura 23 mostra um espectro genérico registrado.

Tanto a aquisição dos espectros quanto a quantificação dos respectivos analitos foram feitas pela Henkel S/A, que produz o aditivo comercialmente, e fornecidos ao nosso grupo para o desenvolvimento e otimização das calibrações propriamente ditas.

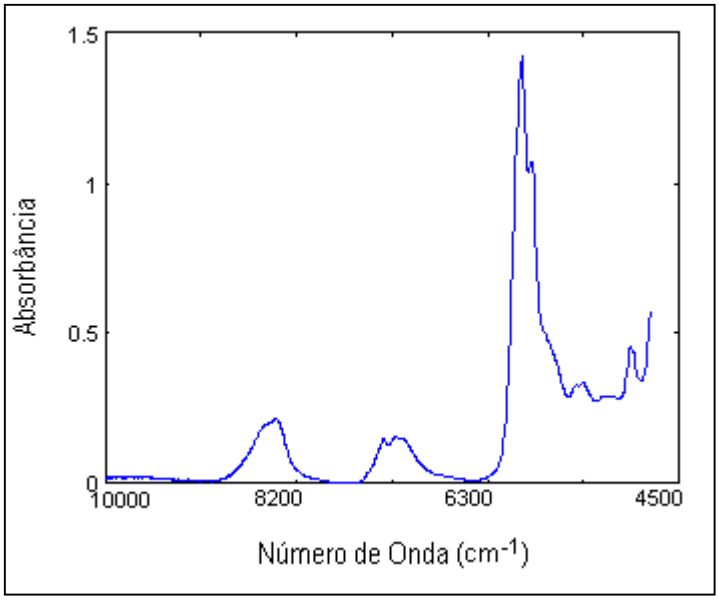


Figura 23: Espectro Genérico Registrado do Analito Água

3.5. A Calibração

As matrizes que constituíam os conjuntos de calibração eram formadas inicialmente por um total de 2530 variáveis, com números de amostras variando entre 43 (para o analito epóxi), 61 (para o analito água) e 73 (para o iodo). Desse total, foram rejeitadas as 130 primeiras variáveis, por não conterem sinais significativos para as análises e para diminuir as matrizes de trabalho.

O único pré-tratamento aplicado às matrizes foi o chamado alisamento pela média ("box car") [55]. Esse pré-tratamento nada mais é que um tipo de alisamento que conduz à uma redução da matriz original. Essa redução ocorre a partir da substituição de dados intervalos de números de onda dos espectros (variáveis), por suas médias. Foram testados alguns intervalos com esse objetivo, e aquele que conduziu aos melhores resultados, sem perda de informação espectral, foi o de 15 variáveis, correspondendo a 30 cm^{-1}, ou seja, calculou-se a média entre cada conjunto de quinze variáveis e utilizou-se este novo valor, reduzindo dessa forma o conjunto a um total de 160 variáveis.

Após essa redução na dimensão das matrizes, as mesmas foram centradas na média, com o objetivo de minimizar os efeitos externos obtidos durante a aquisição dos espectros, e só então foram construídos os modelos de calibração propriamente ditos [14].

O estudo realizado teve como objetivo a construção de modelos de regressão que propiciassem os menores resíduos de validação interna e externa possíveis. A base desse estudo foi feita mediante a exclusão de variáveis que fossem irrelevantes para a construção dos modelos. A escolha das variáveis passíveis de serem excluídas foi feita da seguinte forma: inicialmente foram testados dois métodos – análise simultânea dos gráficos dos pesos e do vetor de regressão (Seleção A) e comparação entre os espectros de maior e menor concentração do conjunto (Seleção B). Àquele que conduziu aos menores resíduos de validação, foi aplicado o chamado correlograma (ver Materiais e Métodos).

Os gráficos de pesos e vetor de regressão fornecem informação a respeito da importância das variáveis de um conjunto na construção de cada nova variável latente. Analisando os dois conjuntamente, é possível selecionarmos aquelas que contribuem significativamente na construção das variáveis latentes serão posteriormente utilizadas.

O correlograma é um gráfico que relaciona os números de onda de um determinado espectro com os respectivos coeficientes de correlação obtidos entre estes e o vetor da variável dependente. O correlograma é obtido da seguinte forma: determina-se o coeficiente de correlação de cada uma das variáveis independentes (números de onda) com o vetor da variável dependente. Relacionam-se então estes coeficientes de correlação obtidos com cada um destes números de onda (números de onda x coeficientes de correlação) [56]. Para cada analito, o correlograma foi testado com cortes em diferentes coeficientes de correlação, de maneira a avaliar qual propiciaria os melhores resultados, levando em consideração o número de variáveis latentes utilizadas, os valores de SECV e PRESS, o coeficiente de correlação da reta obtida e a porcentagem de variância explicada pelo modelo.

Nessa etapa, os modelos foram construídos utilizando a chamada validação cruzada. Esse tipo de validação é utilizado com o objetivo de determinar o número ideal de componentes principais a ser utilizado e detectar a existência de possíveis amostras com comportamento anômalo no conjunto de dados. É usado como um método de validação na ausência de amostras para um conjunto externo de validação. Nesse trabalho, utilizou-se a Validação Cruzada excluindo uma amostra por vez.

Após essa etapa, utilizando os modelos obtidos com as condições já otimizadas, foram testados conjuntos externos de amostras com o objetivo de validar os mesmos. O desempenho dos modelos foi avaliado a partir de dois parâmetros: a razão entre os desvios padrão das concentrações experimentais e dos resíduos, RPD (Ratio of Standard Deviation of Experimental Concentrations to the Standard Deviation of the Residuals), e a razão de intervalo de erro, RER (Ratio Error Range), dados pelas equações 12 e 13.

$$RPD = \frac{std(y_{exp})}{std(y_{exp} - y_{prev})} \qquad \textbf{(12)}$$

$$RER = \frac{range(y_{exp})}{std(y_{exp} - y_{prev})} \qquad \textbf{(13)}$$

onde: std é a estimativa do desvio padrão;

y_{exp} são as concentrações experimentais (medidas em laboratório) das amostras do conjunto externo de validação;

y_{prev} são os correspondentes valores estimados pelo modelo de calibração

$range(y_{exp})$ é o intervalo das concentrações experimentais.

4. Resultados e Discussão

4.1 Água

Um total de 61 amostras foram utilizadas para este analito, 51 como conjunto de calibração e 10 para validação externa. Inicialmente foram utilizadas (após a aplicação da média "boxcar") 160 variáveis. A faixa de concentração deste analito (porcentagem de água) variava entre 0,02 e 0,45%.

Esse conjunto é representado por amostras que podem ser distribuídas em três agrupamentos. O primeiro deles, contendo o maior número de amostras, representado por aquelas de menor concentração (0,02-0,1%); o segundo, formado por quatro amostras de

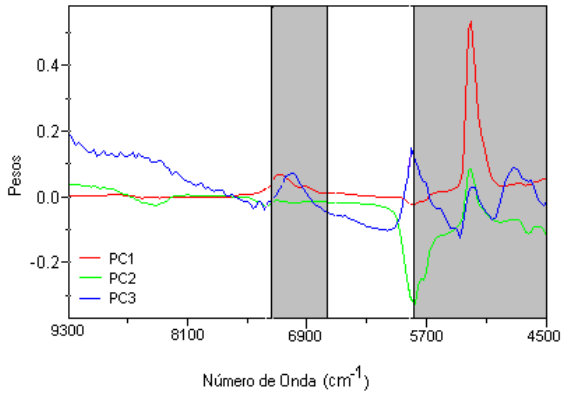
concentração no intervalo de 0,1 e 0,2% e o último, com duas amostras de concentração superior a 0,4%, acima do nível aceitável industrialmente. Essas últimas representam situações que, além de necessárias para a construção do modelo, podem eventualmente ser atingidas em nível industrial e portanto não podem ser consideradas como anômalas. São amostras de alto "leverage", ou seja, que têm um peso significativo na construção do modelo de regressão, uma vez que, diferem bastante do perfil médio das amostras de treinamento.

A seleção de variáveis foi feita, como foi dito anteriormente, por dois métodos distintos: Método A - análise simultânea dos gráficos de pesos e vetor de regressão (Figuras 24a e 24b) e Método B - seleção a partir da diferença entre os espectros de maior e menor concentração do conjunto (Figura 24c) Aplicou-se o correlograma ao método que conduziu aos melhores resultados (Figura 24d).

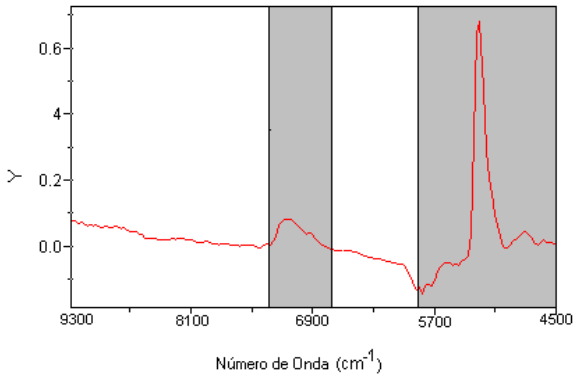
Para esse conjunto, o correlograma foi aplicado ao segundo método (subtração de espectros – Método B), que foi o que produziu os melhores resultados na primeira etapa.

Os resultados obtidos antes e depois do uso do correlograma podem ser vistos na Tabela 9.

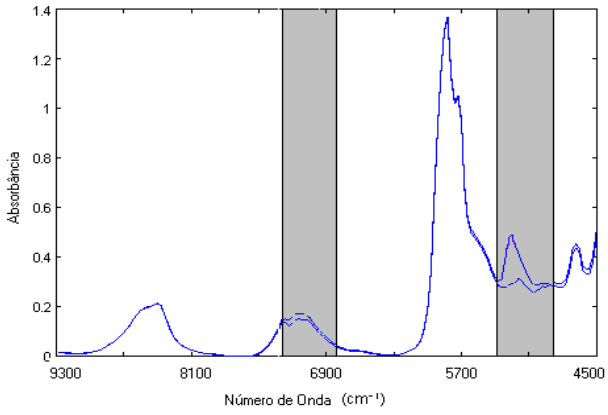
As regiões do espectro escolhidas a partir do método acima foram aquelas cujos números de onda apresentavam alta correlação com o vetor das concentrações, ou seja, aquelas que contribuiriam de forma positiva para a construção do modelo de calibração. Dessa forma, após alguns testes, foram escolhidas as regiões cujos números de onda correspondentes tivessem coeficientes de correlação superiores a 0,85.



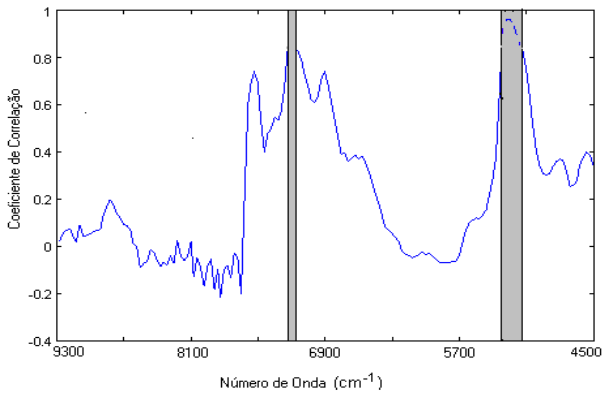
(Fig 24a)



(Fig 24b)



(Fig. 24c)



(Fig. 24d)

Figura 24: (a) Pesos, (b) Vetor de Regressão, (c) Espectros de Maior e Menor concentração (o de maior concentração possui a maior absorbância) e (d) Correlograma– Analito : Água
Regiões selecionadas em cada uma das etapas em destaque

Tabela 15: Resultados obtidos pelo método PLS para o analito água. A seleção pelo correlograma foi usada para coeficientes superiores a 0,85 sobre os resultados obtidos através do método B de seleção, baseada na subtração de espectros (maior e menor concentração) . A etapa de previsão foi executada a partir do modelo obtido após o correlograma.

Método	Variáveis	Var. Latentes	SECV	PRESS	Coeficiente Correlação	Região Selecionada (cm^{-1})
Média Boxcar	160	3	0,0111	0,0062	0,9929	9300-4500
Seleção A	65	2	0,0133	0,0088	0,9890	7250,25-6669,3; 5802,7-4500
Seleção B	22	2	0,0130	0,0085	0,9891	7221,3-6929,9; 5310,6-4961,3
Correlograma	**11**	**2**	**0,0131**	**0,0082**	**0,9883**	7163,4-7045,7; 5310,6-5106,02
Valid. Externa	11	2	0,0120[a]	0,0014	0,9144	

[a] SEP (Standard Error of Prediction)

Dessa forma, mantendo mais ou menos o mesmo nível de resíduos, foi possível construir um novo modelo, usando para isso apenas 11 variáveis, ao invés das 160 iniciais, e 2 variáveis latentes.

Foram utilizadas posteriormente 10 amostras com o objetivo de testar a chamada validação externa. Construiu-se um modelo de calibração com as condições anteriormente otimizadas e procedeu-se à previsão destas amostras. Os resultados obtidos podem ser vistos na Tabela 9. Os valores experimentais e previstos pelo modelo para este analito podem ser vistos na Tabela 10, assim como os parâmetros estatísticos que mostram a capacidade de previsão do modelo. O desvio padrão dos resíduos (0,012) é relativamente

menor que o desvio padrão dos valores experimentais (0,030), levando a um RPD (Equação 12) relativamente baixo de 2,50. Por outro lado, sendo o intervalo das concentrações experimentais grande, temos um RER (Equação 13) bastante satisfatório de 10,17.

Como pode ser visto pelo valor de SEP (Tabela 9), que é um indicativo do erro padrão das amostras com relação à curva de calibração (Standard Error of Prediction), o modelo obtido mostrou-se bastante eficiente na etapa de previsão das concentrações de água, já que este valor apresentou-se bastante baixo.

Tabela 16: Valores Experimental, Previsto e de Resíduos para % Água na Validação Externa.

Amostra	**Valores Experimentais**	**Valores Previstos**	**Resíduos**
1	0,06	0,057	0,003
4	0,05	0,072	-0,022
9	0,06	0,047	0,013
15	0,07	0,063	0,007
19	0,05	0,055	0,005
27	0,08	0,079	0,001
34	0,05	0,060	-0,010
50	0,13	0,111	0,019
56	0,06	0,045	0,015
60	0,008	0,004	0,004
Média	0,062	0,059	0,0035
Desvio Padrão	0,030	0,027	0,012
Intervalo	0,122		
RPD[a] 2,50		*RER*[b] 10,17	

[a]*RPD* = *std*(exp.)/*std*(resíduos)
[b]*RER* = *intervalo*(exp.)/*std*(resíduos)

O gráfico de valor medido *vs* valor previsto, para validação cruzada, utilizando 2 variáveis latentes, pode ser visto na Figura 25.

Dessa forma, foi possível a construção de um modelo de calibração eficiente, com baixos resíduos de validação e previsão, usando para tanto um número bastante reduzido de variáveis e uma variável latente a menos.

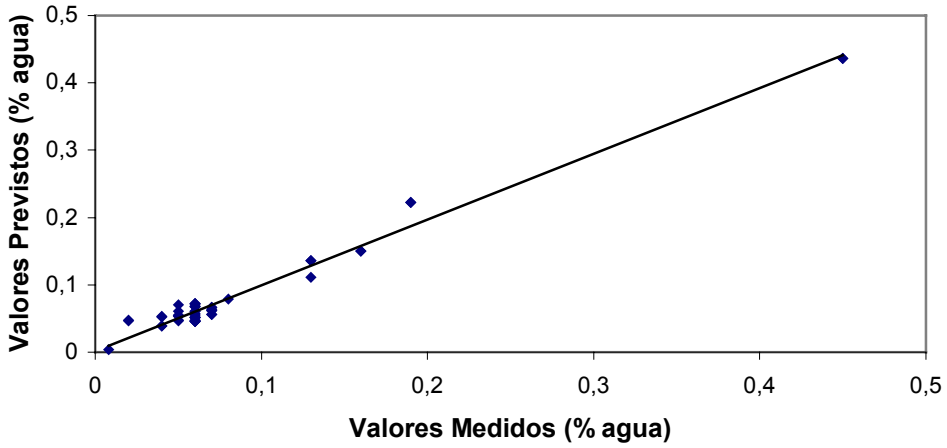


Figura 25: Valores Medidos Exerimentalmente *vs* Previstos pelo método PLS da concentração de água (%) usando Validação Cruzada– Analito : Água

As variáveis selecionadas encontram-se dentro da faixa original de 7163,4 – 7045,7 cm^{-1} e 5310,6 – 5106,02 cm^{-1}, ambas regiões de absorção características da água pura, sendo a de 5160 cm^{-1}, relativa às deformações e estiramento –OH e 7143 - 6667 cm^{-1}, ao seu primeiro sobreton [57].

4.2 Iodo

Para este segundo analito tem-se um total de 73 amostras, 60 para o conjunto de calibração e 13 para validação externa. Neste caso também, após a aplicação da média móvel, utilizou-se 160 variáveis. A faixa de medição deste analito (Índice de Iodo) varia entre 2,2 e 6,3g de iodo/100g de amostra, sendo que um grande número de amostras

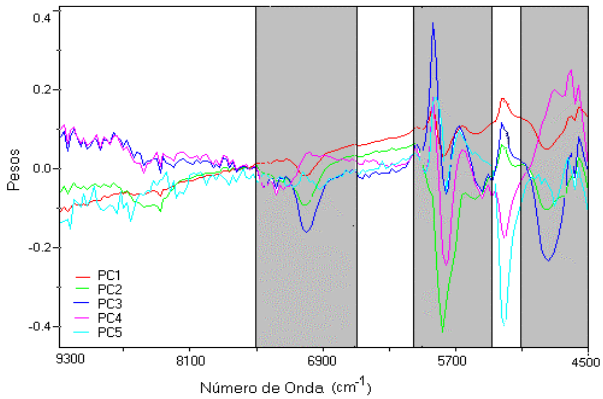
apresenta concentração entre 2,5 e 3,5g de iodo/100g amostra. Também nesse caso temos algumas amostras com I. I. acima dos níveis aceitáveis pelo controle de qualidade.

Tanto para este analito quanto para o próximo, os procedimentos foram os mesmos utilizados na calibração anterior. Após a aplicação da "Média Boxcar", testaram-se os dois métodos visuais de seleção de variáveis e aplicou-se o correlograma ao melhor deles.

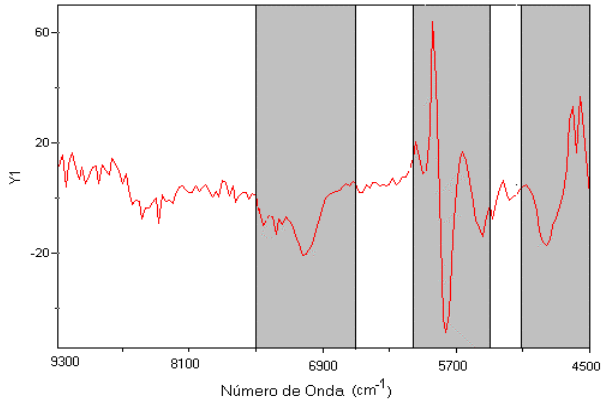
Os resultados obtidos podem ser vistos na Tabela 11. Para esse analito o primeiro método de seleção (a partir do gráfico dos pesos e do vetor de regressão - Método A) mostrou-se mais eficiente (Figura 26a e 26b). O gráfico de espectros de maior e menor concentração utilizado nessa etapa da seleção pode ser visto na Figura 26c.

O correlograma (Figura 26d), aplicado a um novo conjunto com 75 variáveis, referente a este não se mostrou tão bom quanto ao anterior. Foram testadas várias faixas de corte e a que levou aos melhores resultados foi para variáveis que apresentassem coeficiente de correlação com relação ao vetor das variáveis dependentes maior que 0,35.

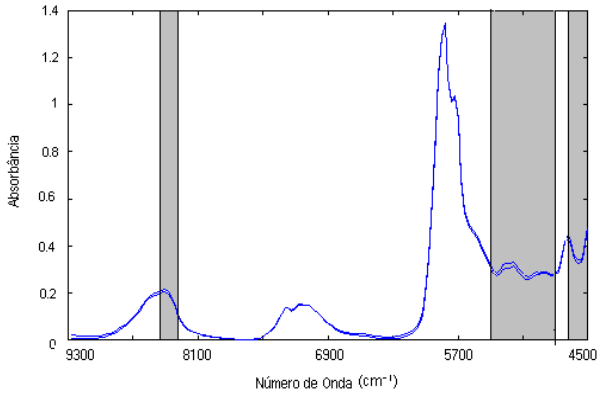
O método não foi tão eficiente quanto o do analito anterior uma vez que, apesar de reduzir o conjunto de 75 para 25 variáveis, o nível de resíduos obtidos com relação ao modelo da Seleção A é um pouco maior e o coeficiente de correlação da regressão é também inferior ao do anterior. Foram utilizadas 4 variáveis latentes para a construção deste modelo enquanto que para o primeiro, antes da seleção de variáveis, haviam sido utilizadas 5. Dessa forma, optou-se na escolha do modelo obtido na Seleção A como o melhor e este foi utilizado na etapa de previsão.



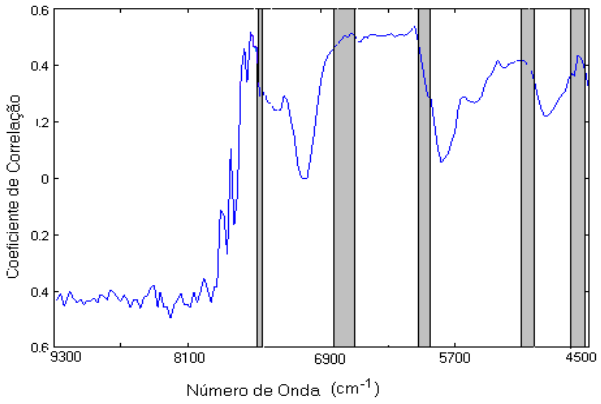
(Fig. 26a)



(Fig. 26b)



(Fig. 26c)



(Fig. 26d)

Figura 26: (a) Pesos, (b) Vetor de Regressão, (c) Espectros de Maior e Menor concentração (o de maior concentração possui a maior absorbância) e (d) Correlograma– Analito : Iodo Regiões selecionadas em cada uma das etapas em destaque

Tabela 17: Resultados obtidos pelo método PLS para o analito iodo. A seleção pelo correlograma foi usada para coeficientes superiores a 0,35 sobre os resultados obtidos através do método de seleção baseada na avaliação visual dos gráficos de pesos e vetor de regressão.. A etapa de previsão foi executada a partir do modelo obtido na Seleção A.

Método	Variáveis	Var. Latentes	SECV	PRESS	Coeficiente Correlação	Região Selecionada (cm^{-1})
Média Boxcar	160	5	0,2848	4,8651	0,9554	9300-4500
Seleção A	**75**	**4**	**0,2723**	**4,4485**	**0,9664**	**7425,9-6642,3; 6123,1-5368,5; 5138,8-4500**
Seleção B	35	5	0,3547	7,5503	0,9288	8435,1-8205,6; 5399,4-4789,5; 4675,6-4500
Correlograma	25	4	0,4835	14,0271	0,9219	7425,9-7366; 6817,9-6642,3; 6123,1-5947,5; 5138,8-4992,1; 4675,6-4528,9
Valid. Externa	75	4	0,2322[a]	0,7007	0,9714	

[a] SEP (Standard Error of Prediction)

Para este analito foram utilizadas 13 amostras para a validação externa. O procedimento foi o mesmo adotado no item anterior. Os resultados obtidos podem ser vistos na Tabela 11. A partir dos valores experimentais e previstos mostrados na Tabela 12, pode-se perceber que apenas 3 amostras foram previstas com um erro superior a 10%. Para

este analito, as concentrações eram bem uniformes e um RPD = 4,00 e RER = 14,19 são indicativos de que os resíduos obtidos são relativamente pequenos.

Tabela 18: Valores Experimental , Previsto e de Resíduos para I. I. na Validação Externa.

Amostra	Valores Experimentais	Valores Previstos	Resíduos
2	2,98	3,171	0,191
6	3,41	3,195	-0,214
9	3,02	3,050	0,030
14	3,25	3,578	0,328
20	2,61	2,578	-0,032
26	3,19	3,087	-0,103
30	3,38	2,959	-0,421
37	2,96	2,958	-0,002
50	3,82	3,741	-0,078
59	2,45	2,801	0,351
65	4,41	4,432	0,022
68	4,82	4,804	-0,016
72	5,85	5,414	-0,436
Média	3,550	3,521	0,029
Desvio Padrão	0,959	0,856	0,240
Intervalo	3,400		
RPD[a] 4,00		*RER*[b] 14,19	

[a]*RPD* = *std*(exp.)/*std*(residuos)
[b]*RER* = *intervalo*(exp.)/*std*(residuos)

O gráfico de valor medido *vs* valor previsto, utilizando 4 variáveis latentes, pode ser visto na Figura 27.

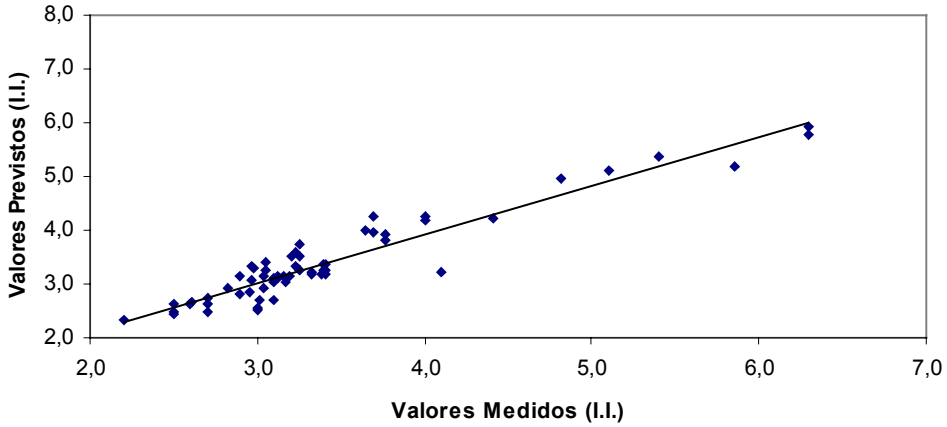


Figura 27: Valores Medidos Experimentalmente *vs* Previstos pelo método PLS do Ìndice de Iodo (I.I.) usando Validação Cruzada – Analito : Iodo

As variáveis selecionadas encontram-se na região entre 7430 – 6600; 6100 – 5400 e 5100 – 4500 cm^{-1}.

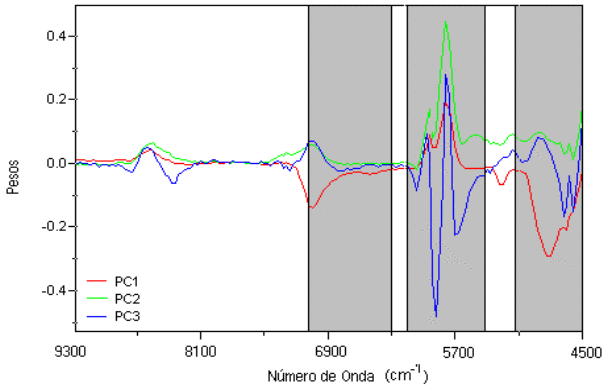
Esta primeira região (7430 - 6600 cm^{-1}) está relacionada à combinação dos estiramentos e deformações C-H, além do primeiro sobreton do estiramento – OH. A segunda (6100 - 5400 cm^{-1}) está relacionada ao estiramento – CH em olefinas, enquanto que a seguinte, entre 5100 e 4500 cm^{-1}, envolve a combinação de estiramentos de ligações duplas C=C de cadeias conjugadas [58].

4.3 Epóxido

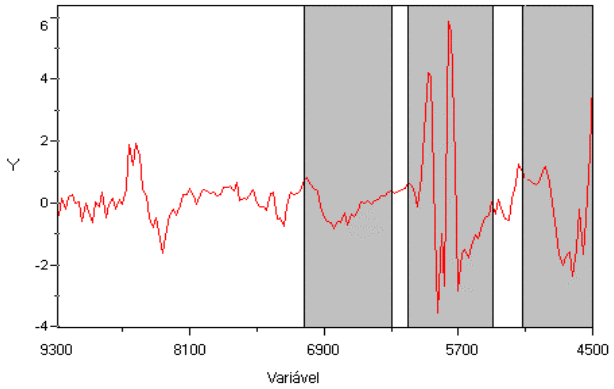
O procedimento para este último conjunto, de 35 amostras, foi idêntico ao adotado para os analitos anteriores. Neste caso, 42 amostras foram utilizadas (35 como conjunto de treinamento e 8 para validação externa). A faixa de índice de epóxido das amostras variou entre 6,21 e 6,61.

Com relação a este, o método de seleção que se mostrou mais eficiente foi a análise pela subtração - Seleção A (Figuras 28a, 28b e 28c), conduzindo a um conjunto de 65 variáveis.

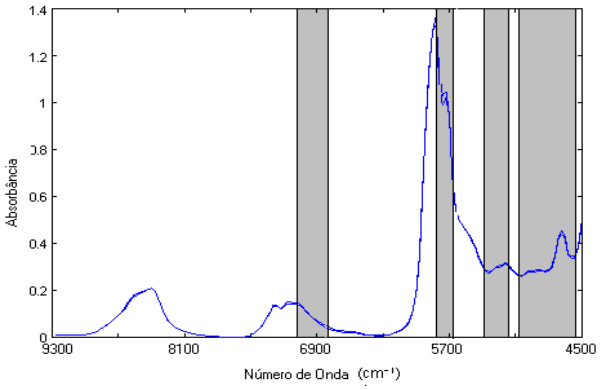
O uso do correlograma (Figura 28d) neste novo conjunto obtido (corte em variáveis com coeficientes maiores que 0,20) não foi de grande valia uma vez que, como no analito anterior, apesar de haver uma redução considerável no número de variáveis utilizadas, os resíduos obtidos são um pouco maiores e há um pequeno decréscimo no coeficiente de correlação. Os resultados obtidos podem ser vistos na Tabela 13.



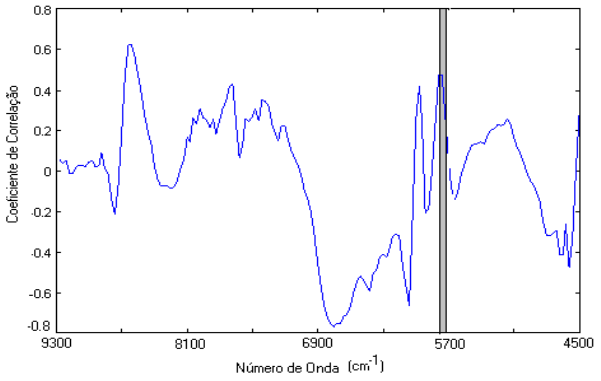
(Fig. 28a)



(Fig. 28b)



(Fig. 28c)



(Fig. 28d)

Figura 28: (a) Pesos, (b) Vetor de Regressão, (c) Espectros de Maior e Menor concentração (o de maior concentração possui a maior absorbância) e (d) Correlograma– Analito : Epóxido Regiões selecionadas em cada uma das etapas em destaque

Tabela 19: Resultados obtidos pelo método PLS para o analito epóxi. A seleção pelo correlograma foi usada para coeficientes superiores a 0,20 sobre os resultados obtidos através do método de seleção baseada na avaliação visual dos gráficos de pesos e vetor de regressão. A etapa de previsão foi executada a partir do modelo obtido na Seleção A.

Método	Variáveis	Var. Latentes	SECV	PRESS	Coeficiente Correlação	Região Selecionada (cm^{-1})
Média Boxcar	160	6	0,0349	0,0426	0,9841	9300-4500
Seleção A	**65**	**3**	**0,0410**	**0,0587**	**0,9600**	**7078,5-6256,5; 6152-5426,4; 5109,9-4500**
Seleção B	37	5	0,0379	0,0503	0,9757	7136,4-6816; 5833,6-5658; 5399,4-5223,7; 5052-4644,7
Correlograma	4	2	0,0860	0,2662	0,8787	5805-5687
Valid. Externa	37	5	0,0169[a]	0,0023	0,9939	

[a] SEP (Standard Error of Prediction)

Para este analito, foram utilizadas 8 amostras para a validação externa. Os resultados obtidos podem ser vistos também na Tabela 13. Os resultados mostrados indicam que é possível construir um modelo com uma capacidade de previsão muito boa e capaz de garantir o controle do produto.

O gráfico de valor medido *vs* valor previsto, utilizando 5 variáveis latentes, pode ser visto na Figura 29.

Os valores de E.I. medidos e previstos das amostras do conjunto externo de validação podem ser vistos na Tabela 14, assim como seus respectivos resíduos. Os erros de previsão mostraram-se bastante baixos (abaixo de 0,5%). Para este analito, tanto a razão RPD quanto RER conduziram a excelentes resultados (22,022 e 80,695, respectivamente, onde a razão ideal é de no mínimo 10 indicando precisão nas análises, principalmente pelo fato da concentração do analito variar num intervalo tão pequeno.

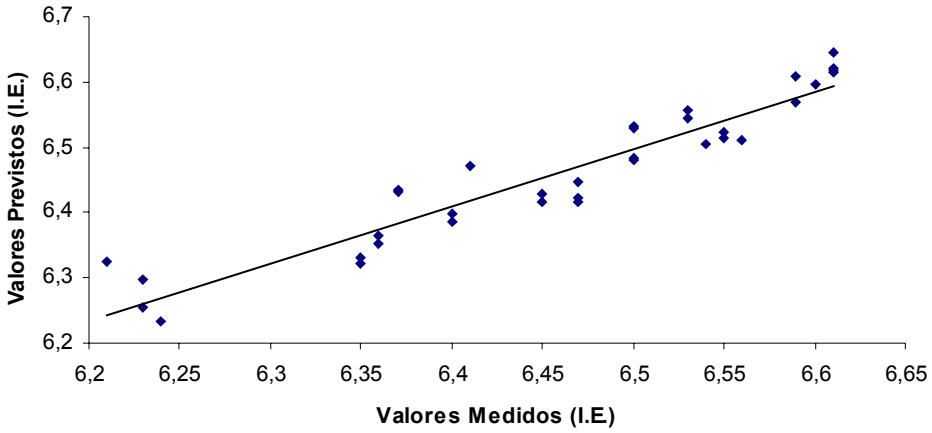


Figura 29: Valores Medidos Experimentalmente *vs* Previstos pelo método PLS do Ìndice de Epóxido (I.E.) usando Validação Cruzada – Analito : Epóxido

Tabela 20: Valores Experimental , Previsto e de Resíduos para I. E. na Validação Externa.

Amostra	Valores Experimentais	Valores Previstos	Resíduos
2	5,73	5,710	-0,020
3	6,21	6,199	-0,010
10	6,60	6,596	-0,004
15	6,56	6,540	-0,020
23	6,41	6,415	0,005
28	6,54	6,517	-0,023
41	6,47	6,443	-0,027
44	6,24	6,234	-0,006
Média	6,35	6,332	0,013
Desvio Padrão	0,29	0,288	0,011
Intervalo	0,89		
RPD[a]	26,022	*RER*[b]	80,695

[a]*RPD* = *std*(exp.)/*std*(residuos)
[b]*RER* = *intervalo*(exp.)/*std*(residuos)

Neste conjunto, as variáveis selecionadas encontram-se nos intervalos de 7100 – 6800; 5800 – 5600; 5400 - 5200 e 5000 – 4500 cm^{-1}. À primeira região (7100 - 6800 cm^{-1}) pode-se atribuir o primeiro sobreton do estiramento C-H. A segunda (5800 - 5600 cm^{-1}) está relacionada a estiramentos do grupo – CH presente em ácidos graxos, enquanto a terceira (5400 – 5200 cm^{-1}) pode ser atribuída ao segundo sobreton do estiramento C=O. A última região (5000 - 4500 cm^{-1}) deve-se à combinação de estiramentos de ligação dupla C=C [59].

5. Conclusões

Como pôde ser visto nesta aplicação, o uso da Espectroscopia NIR combinada à regressão multivariada é uma alternativa bastante viável às técnicas amplamente estabelecidas, especialmente em processos industriais. O uso do método PLS foi bem sucedido no sentido de construir modelos de regressão de alta qualidade. Foi mostrado também neste trabalho que usando métodos simples e intuitivos de seleção de variáveis, como a análise de pesos/vetor de regressão e o correlograma, o número de variáveis pôde ser significativamente reduzido sem prejudicar a qualidade do modelo.

Os parâmetros estatísticos utilizados, RPD e RER, indicaram grande precisão nas determinações para o I.E., e uma boa precisão para o I.I. Para o analito água os resultados mostraram-se satisfatórios.

A partir dos resultados obtidos, pode-se concluir que a metodologia proposta é apropriada para o monitoramento da reação de epoxidação do óleo de soja e avaliação da qualidade do aditivo em processos industriais, onde tempo, esforço e dinheiro são cruciais.

CONCLUSAO GERAL

A utilização de métodos quimiométricos em dados de natureza multivariada foi abordada de forma bastante ampla neste trabalho, explorando técnicas de analise quali e quantitativa.

Em ambos os casos, as abordagens empregadas conduziram a ótimos resultados, possibilitando a extração de informações dos conjuntos de dados que seriam de difícil análise sem o emprego das mesmas, como no primeiro caso, onde se tinha um conjunto de dados de relativa complexidade.

Neste caso específico, o uso da quimiometria possibilitou o estudo simultâneo das variáveis englobadas em diferentes categorias, de maneira a tirar conclusões práticas a respeito de quais seriam realmente relevantes no estudo e quais propiciaram os melhores resultados em termos de aceitabilidade do produto estudado frente ao consumidor final.

Da mesma forma, no segundo caso, modelos de calibração foram desenvolvidos com altas porcentagens de variância e baixos resíduos com relação aos valores experimentais, possibilitando dessa forma a otimização do processo de controle de qualidade do produto em questão.

De maneira geral, pode-se concluir que a utilização destes métodos em dados químicos é de grande valia, uma vez que atende aos principais interesses do mercado industrial, que são a otimização de custo e tempo das análises realizadas.

BIBLIOGRAFIA

[1] R. Tauler, D. Barcello, E. M. Thurman, "*Multivariate correlation between concentrations of selected herbicides and derivatives in outflows from selected US midwestern reservoirs*", Environ. Sci Technol., 34, p. 3307-3314, ***2000***.

[2] E. Csomos, K. Heberger, L. Simon-Sarkadi, "*Principal Component Analysis of biogenic amines and polyphenols in Hungarian wines*", J. Agr. Food Chem., 50, p.3768-3774, ***2002***.

[3] M. M. C. Ferreira, M. A. Morgano, S. C. D. de Queiroz, D. M. B. Mantovani, "*Relationships of the minerals and fatty acids contents in processed turkey meat products*", Food Chem., 69, p. 259-265, ***2000***.

[4] M. M. de Sena, R. J. Poppi, R. T. S. Frighetto, P. J. Valarini, "*Evaluation of the use of chemometric methods in soil analysis*", Quím. Nova, 23, p. 547-556, ***2000***.

[5] S. Wold, *"Principal Component Analysis"*, Chemom. Intell. Lab. Sys., 2, p.37-52, ***1987***.

[6] G. Strang, *"Linear Algebra and its Applications*", Academic Press, Nova Iorque, 2a ed., ***1976***.

[7] M. M. Reis e M. M. C. Ferreira, *"Separação de espectros simulados e de luminescência total através do método generalizado de anulação do posto (GRAM)"*, Quím. Nova, 22, p.11-17, ***1999***.

[8] A. K. Smilde, *"Three-Way analyses problems and prospects"*, Chemom. Intell. Lab. Syst., 15, p. 143-157, ***1992***.

[9] M. M. Reis, *"Aplicação de métodos quimiométricos em separação de espectros e reconhecimento de padrões"*, Dissertação de Mestrado, IQ-UNICAMP, ***1997***.

[10] H. Martens e T. Naes, "*Multivariate Calibration"*, John Wiley & Sons, Nova Iorque, ***1989***.

[11] P. Geladi e B.R. Kowalski, *"Partial Least-Squares regression: a tutorial"*, Anal. Chim. Acta, 185, p. 1-17, ***1986***.

[12] F. A. L. Ribeiro, "*Aplicação de métodos de análise multivariada no estudo de hidrocarbonetos policíclicos aromáticos*", Dissertação de Mestrado, IQ-UNICAMP, ***2001***.

[13] J. Smeyers-Verbeke, J.C. Den Hartog, W. H. Dekker, D. L. Massart, "*Clustering applied to an organic air pollutants data set*", Analusis, 12, p. 486-489, ***1984***.

[14] K. R. Beebe, R. J. Pell e M. B. Seasholtz, "*Chemometrics: a Pratical Guide"*, John Wiley & Sons, Nova Iorque, ***1998***.

[15] L. Nørskov-Lauritsena e H. B. Bürgi, "*Cluster Analysis of periodic distributions; application to conformational analysis*", J. Comput. Chem., 6, p. 216-228, ***1985***.

[16] N. Bratchell, *" Cluster Analysis"*, Chemom. Intell. Lab. Sys, 6, p.105-125, ***1989***.

[17] L. R. Tucker, *"Some mathematical notes on three-mode factor analysis"*, Psychometrika, 31, p. 279-311, ***1966***.

[18] R. Henrion, *"N-way principal component analysis – Theory, algorithms and applications"*, Chemom. Intell. Lab. Sys., 25, p.1-23, ***1994***.

[19] R. A. Harshman, "*Foundation of the PARAFAC procedure: model and conditions for an "explanatory" multi-mode factor analysis"*, UCLA Working Papers in Phonetics, 16, p. 1-84, ***1970***.

[20] J. D. Carrol e J. J. Chang, "*Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition*", Psychometrika, 35, p. 283-319, ***1970***.

[21] R. Bro, *" The N-Way Tutorial – Interactive Introduction to Multiway Analysis in MATLAB"*, http://www.models.kvl.dk/courses/parafac/chap0contents.htm, ***1998***.

[22] L. Munk, L. Nørgaard, S. B. Engelsen, R. Bro e C. A. Andersson, *"Chemometrics in food science – a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance"*, Chemom. Intell. Lab. Sys., 44, p. 31-60, ***1998***.

[23] R. Bro, *"Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis"*, Chemom. Intell. Lab. Sys., 46, p.133-147, ***1999***.

[24] R. Bro, *"PARAFAC. Tutorial and applications",* Chemom. Intell. Lab. Syst., 38, p.149 - 171, ***1997***.

[25] M. M. Reis, *"Desenvolvimento e Aplicações de Métodos Quimiométricos de Ordem Superior"*, Tese de Doutorado, IQ-UNICAMP, ***2002***.

[26] P. Geladi, *"Analysis of Multi-Way (Multi-Mode) Data"*, Chemom. Intell. Lab. Sys., 7, p. 11-30, ***1998.***

[27] M. M. C. Ferreira, A. M. Antunes, M. S. Melgo, P. L. O. Volpe, *"Quimiometria I: Calibração Multivariada, um tutorial"*, Quím. Nova, 22, p. 724-731, ***1999***.

[28] B. Grung e O. M. Kvalheim, *"Rank determination of spectroscopic profiles by means of cross validation. The effect of replicate measurements on the effective degrees of freedom"*, Chemom. Intell. Lab. Sys., 22, p. 115-125, ***1994***.

[29] T. Naes e T. Isaksson, *"Data compression by PLS/PCR"*, Nirnews 3, p.10, ***1992***.

[30] *"MATLAB® for Windows"*, The MathWorks, Inc., versão 5.1.0.421, ***1984-1997***.

[31] R. Bro, C. A. Andersson, " *The N-Way Toolbox for Matlab*", versão 1.04, ***1999***, http://www.models.kvl.dk/source/nwaytoolbox.

[32] *"Pirouette"*, InfoMetrix, Inc., versão 2.02, Woodinville, Washington, ***1990-1996***.

[33] I. Stewart, T. A. Wheaton, *"Carotenoids in citrus: their accumulation induced by ethylene"*, J. Agr. Food Chem., 20, p. 448-449, ***1972***.

[34] J. V. de Castro, V. L. P. Ferreira, R. M. Pio, *"Influência da temperatura no desverdecimento e qualidade do tangor murcote"*, Laranja, 12, p. 211-224, ***1991***.

[35] J. V. de Castro, V. L P. Ferreira, K. Yotsuyanagi, *"Aplicação pós-colheita de etileno e de ethrel no desverdecimento de tangor murcote"*, Rev. Bras. Frutic., 13, p. 237-242, ***1991***.

[36] M. Meloun, J. Militki, M. Forina, *"Chemometrics for Analytical Chemistry: PC - Aided Statistical Data Analysis"*, Ellis Horwood, Nova Iorque, Cap. 4***, 1992***.

[37] P. Lea, T. Naes, M. Rodbotten, "*Analysis of Variance for Sensory Data*", John Wiley and Sons, Nova Iorque, ***1997***.

[38] S. Williams, "*Official Methods of Analysis*", Association of Official Analytical Chemists, Washington, 11a ed, p. 155, 178, 371 e 777, ***1970***.

[39] T. F. Parreira, M. M. C. Ferreira, H. J. S. Sales, W. B. de Almeida, "*Quantitative determination of epoxide soybean oil using near infrared spectroscopy and multivariate calibration*", Appl. Spectr., 56, p. 1607-1614, ***2002***.

[40] B. A. Howell, S. R. Betso, J. A Meltzer, P. B. Smith and M. F. Debney, *"Thermal degradation of epoxidized soybean oil in the presence of chlorine-containing polymers"*, Thermochim. Acta, 166, p.207 –218, ***1990***.

[41] O. S. Kauder, *"Nontoxic Polyvinyl Chloride Processing Stabilizers"*, in: *Plastics Additives and Modifiers Handbook*, Van Nostrand Reinhold, Nova Iorque, Cap.18, p.297, ***1992***.

[42] P. G. Demertzis, K. A Riganakos e K. Akridademertzi, *"Gas chromatographic studies on polymer-plasticizer compability: interactions between food-grade PVC and epoxidized soybean oil",* Eur. Polym. J., 27, p. 231-235, ***1991***.

[43] H. O Wirth e H. Andreas, *"The stabilization of PVC against heat and light"*, Pure Appl. Chem., 49, p. 627-648**,** ***1977***.

[44] H. Baltacioglu e D. Balkose, *"Effect of zinc stearate and/or epoxidized soybean oil on gelation and thermal stability on PVC-DOP plastigels"*, J. Appl. Polym. Sci., 74, p. 2488-2498, ***1999***.

[45] N. M. Faber, *"Multivariate sensitivity for the interpretation of the effect of spectral pretreatment methods on near-infrared calibration model predictions"*, Anal. Chem., 71, p. 557-565, ***1999***.

[46] H. M. Heise e A. Buttner, *"Multivariate Calibration for near-infrared spectroscopy assays of blood substrates in human plasma based on variable selection using PLS-regression vector choices"*, Fresenius J. Anal. Chem., 362, p. 141-147, ***1998***.

[47] S. C. Rutan, O. E. Noord, R. R. Andréa, *"Characterization of the sources of variation affecting Near-Infrared spectroscopy using chemometric methods"*, Anal. Chem., 70, p.3198-3201, ***1998***.

[48] F. Cadet, D. Bertrand, P. Robert, J. Maillot, J. Dieudonné e C. Rouch, *"Quantitative determination of sugar cane sucrose by multidimensional statistical analysis of their Mid-infrared Attenuated Total Reflectance Spectra"*, Appl. Spec., 45, p.166-172, ***1991***.

[49] B. Vigerust, K. Kolset, S. Nordenson, A. Henriksen e K. Kleveland, *"Quantitative analysis of additives in low-density polyethylene using infrared spectroscopy and multivariate calibration"*, Appl. Spec., 45**,** p.173-177, ***1991***.

[50] U. Depczynski, K. Jetter, K. Molt e A. Niemöller, *"Quantitative analysis of near infrared spectra by wavelet coefficient regression using a genetic algorithm"*, Chem. Intell. Lab. Sys., 47, p.179-187, ***1999***.

[51] P.J. Brown, *"Wavelength selection in multicomponent near infrared calibration"*, J. Chemom., 6, p. 151-161, ***1992***.

[52] R. Hiukka, *"A multivariate approach to the analysis of pine-needle samples using NIR"*, Chemom. Intell. Lab. Sys., 44, p. 395-401, ***1998***.

[53] Application Bulletin 77/2e – *Karl Fischer water determinations* – Metrohm AG, Herisau, Suíça, ***2000***.

[54] H. J. S. Sales, *"Epoxidação de Óleo de Soja Catalisada por CH_3ReO_3"*, Dissertação de Mestrado, IQ-UNICAMP, ***2000***.

[55] H. H. Willard, L. L. Merrit Jr., J. A. Dean e F. A. Settle Jr., "*Instrumental Methods of Analysis",* Wadsworth Pub. Co., Belmont, Califórnia, 7a ed., p. 22, ***1998***.

[56] V. Bellon-Maurel, C. Vallat e D. Goffinet, *"Quantitative analysis of individual sugars during starch hydrolysis by FT-IR/ATR spectrometry – Part I: Multivariate Calibration study /repeatability and reproducibility"*, Appl. Spec., 49, p. 556-562, ***1995***.

[57] I. Murray e P. C. Williams, *"Chemical Principles of Near Infrared Technology"*, in: *Near-Infrared Technology in the Agricultural and Food Industries*, American Association of Cereal Chemists Inc., St. Paul , Minessota, 2a. ed., p. 17, ***1990***.

[58] J. J. Workman Jr., *"Interpretative spectroscopy of near infrared"*, Appl. Spec. Rev., 31, p. 251-320, ***1996***.

[59] U. Eschenauer, O. Henk, M. Hühne, P. Wu, I. Zebger e H. W. Siesler, *"Near-Infrared Spectroscopy in chemical research, quality assurance and process control"*, in: *Near Infrared Spectroscopy – Bridging the Gap between Data Analysis and NIR Applications,* Ellis Horwood Limited, Chichester, West Sussex, UK, Cap. 2, p. 11, ***1992***.