UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE QUÍMICA

PÓS-GRADUAÇÃO EM FÍSICO-QUÍMICA

ANDRÉ MESSIAS KRELL PEDRO

DETERMINAÇÃO SIMULTÂNEA E NÃO-DESTRUTIVA DE SÓLIDOS TOTAIS E SOLÚVEIS, LICOPENO E BETA-CAROTENO EM PRODUTOS DE TOMATE POR ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO UTILIZANDO CALIBRAÇÃO MULTIVARIADA

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

Orientadora: Professora Doutora Márcia Miguel Castro Ferreira

CAMPINAS – SÃO PAULO

Setembro - 2004

Agradecimentos

Uma vez concluído um grande projeto, cabe destacar o apoio fundamental de pessoas que contribuíram, através de apoio direto ou simplesmente pela presença de espírito, compreensão e colaboração em momentos importantes de sua execução.

À Márcia Ferreira, obrigado por ter aceitado o desafio de conduzir-me neste trabalho; suas orientações foram certamente fundamentais para a conclusão desta dissertação. Agradeço também pelos estímulos e incentivos fornecidos ainda no período de graduação que definiram minha opção pela quimiometria.

Também agradeço àqueles, na Unilever, que acreditaram neste trabalho desde o início, incentivando-o financeira, política ou estrategicamente: Cristiane Gomes, Karina Teixeira e Daniela Paula. Pelo apoio, compreensão e orientação, agradeço ao colega e amigo Anísio Castilho, à Kézia Ferreira, à Roseli Pessoto e ao Ricardo Barreto. Ao Paulo Afonso, Sidnei Farias, Gorete Moraes, Nadia Tafarello, Gisele Bannwart, Ana Paula de Angelis e a todos os colegas que auxiliaram na coleta de informações, muito obrigado.

Agradeço ao amigo Colin Haine e sua esposa, Liz Woolner, que muito contribuíram na execução deste e de outros trabalhos, provavelmente mais até que eles próprios imaginam, dando-me exemplos marcantes de integridade, sensatez, honradez, disciplina e companheirismo.

Aos meus pais, Carlos Alexandre e Maria Izabel, agradeço pelo apoio não apenas neste último período, mas durante toda uma vida dedicada aos estudos, por terem-me fornecido tudo que uma boa alma necessita: saúde, educação e bons exemplos. À Carol, obrigado pela paciência e compreensão, especialmente nos momentos onde foram mais necessários.

ANDRÉ MESSIAS KRELL PEDRO

Formação Acadêmica

Mestre em Química

Universidade Estadual de Campinas – UNICAMP

Instituto de Química – Departamento de Físico-Química

Laboratório de Quimiometria Teórica e Aplicada

Dissertação defendida em 03/09/2004.

Bacharel em Química Tecnológica - 2002

Universidade Estadual de Campinas – UNICAMP

Instituto de Química

Agraciado com o Prêmio Lavoisier de Honra ao Mérito, oferecido pelo Conselho Regional de Química, pelo desempenho acadêmico no período de graduação.

Publicações

Pedro, A.M.K. and Ferreira, M.M.C.; "*Non-Destructive Determination of Solids and Carotenoids in Tomato Products by Near Infrared Spectroscopy and Multivariate Calibration*" – submetido para publicação.

Seminários

International Forum on Genetically Modified Organisms (OGM's) – 10/2003

Universidade Estadual de Campinas – UNICAMP

Palestra: "Food Industry and the OGM's – Implementing Effective Monitoring and Identity Preserved Programmes".

Semana de Alimentos – 07/2002

Faculdade de Engenharia de Alimentos

Universidade Estadual de Campinas – UNICAMP

Palestra: "Antioxidantes Naturais e Artificiais: Prevenindo a Oxidação de Alimentos à Base de Óleos e Gorduras"

International Workshop on Fats, Oils and Oilseeds Analysis – 11/2000

Rio de Janeiro

Palestra: "Oxidation Mechanisms on Fats and Oils: Measuring Oxidation Levels and Interpreting the Results".

Experiência Profissional

Unilever Bestfoods Brasil Ltda (2003 –)

LAFIC – Latin American Foods Innovation Centre

Coordenador de Pesquisa e Desenvolvimento

Unilever Bestfoods Brasil Ltda (1998 – 2002)

LAFIC – Latin American Foods Innovation Centre

Analista de Pesquisa e Desenvolvimento

Van den Berg Alimentos – Unilever (1995 – 1998)

Analista de Qualidade Assegurada de Fábrica

Resumo

DETERMINAÇÃO SIMULTÂNEA E NÃO-DESTRUTIVA DE SÓLIDOS TOTAIS E SOLÚVEIS, LICOPENO E BETA-CAROTENO EM PRODUTOS DE TOMATE POR ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO UTILIZANDO CALIBRAÇÃO MULTIVARIADA traz o desenvolvimento de modelos de calibração utilizando técnicas quimiométricas de regressão multivariada de dados. A determinação simultânea e não-destrutiva de importantes parâmetros de produtos derivados de tomates, como licopeno, beta-caroteno, sólidos totais e solúveis, foi realizada utilizando a técnica de espectroscopia no infravermelho próximo (NIR) e os métodos de calibração multivariada PCR e PLS. Os métodos de referência foram secagem em estufa para sólidos totais, refratometria para sólidos solúveis e HPLC para os carotenóides. Para que modelos com ótima performance preditiva fossem obtidos quatro métodos de seleção de variáveis foram empregados: o *Successive Projections Algorithm* (SPA), a *Dimension-Wise Selection* (DWS), a seleção por correlogramas e a divisão simétrica de espectros. Construiu-se um total de 64 modelos de calibração, nos quais observou-se que os melhores métodos de pré-processamento de espectros foram o alisamento pela média ou a correção multiplicativa de sinais (MSC). O método de regressão mais eficiente foi o PLS, enquanto que o procedimento de seleção de variáveis pela divisão simétrica de espectros forneceu os melhores modelos. Os modelos finais obtidos apresentam habilidades predidivas com desvios similares aos dos métodos de referência e podem ser largamente empregados na indústria alimentícia.

Abstract

SIMULTANEOUS AND NON-DESTRUCTIVE DETERMINATION OF TOTAL AND SOLUBLE SOLIDS, LYCOPENE AND BETA-CAROTENE IN TOMATO PRODUCTS BY NEAR INFRARED SPECTROSCOPY AND MULTIVARIATE CALIBRATION brings the development of calibration models by using chemometrical techniques of multivariate data regression. The simultaneous and non-destructive determination of important parameters in tomato products, like lycopene, beta-carotene and total and soluble solids, was accomplished by using the Near-Infrared (NIR) Spectroscopy technique and the PCR and PLS calibration algorithms. Reference methods were oven drying for total solids, refractometry for soluble solids and HPLC for the carotenoids. For the construction of optimal calibration models, four variable selection techniques were applied: the Successive Projections Algorithm (SPA), the Dimension-Wise Selection (DWS), the selection by the correlograms and the symmetrical spectra division. A total of 64 models were built, by which it could be concluded that the best pre-processing method was the average smoothing or the Multiplicative Signal Correction (MSC). The most effective regression method was the PLS, whilst the variable selection by symmetrical spectra division gave the best models. The final models obtained present predictive abilities that have similar deviations of those of the reference methods and, thus, can be broadly used in the food industry.

Índice de Figuras

Figura 1.1 – Ácido galacturônico e éster metílico do ácido galacturônico. 5

Figura 1.2 – Conversão de um éster metílico da pectina em ácido carboxílico pela ação da PME, liberan-do uma molécula de metanol 6

Figura 1.3 – Ação da PG na degradação da pectina do tomate pela clivagem das ligações entre os monômeros de ácido galacturônico. 6

Figura 1.4 – Carotenóides, como o β-caroteno e o licopeno, formados por repetições de unidades de isopreno 9

Figura 1.5 – Esquema simplificado do processamento de tomates 11

Figura 2.1 – Representação geométrica dos vetores **k**, **v** e **s** 24

Figura 2.2 – Exemplo do procedimento de alisamento pela média 44

Figura 2.3 – Aplicação da transformada de Fourier para o alisamento de uma curva gaussiana contendo ruído 46

Figura 2.4 – Exemplo de aplicação da eliminação de *offset* pelo método MSC 47

Figura 2.5 – Exemplo da aplicação de correção de linha de base pela aplicação de derivadas 49

Figura 2.6 – Exemplo de procedimento de seleção de variáveis pelo correlograma 52

Figura 2.7 – Gráfico de RMSEP em função do parâmetro de corte 52

Figura 2.8 – Exemplo de procedimento de divisão simétrica de espectros 53

Figura 2.9 – Exemplo de aplicação do algoritmo DWS 57

Figura 3.1 – Consistômetro Bostwick 66

Figura 3.2 – Cartão de *Blotter Test* para retenção de água em produtos de tomate 67

Figura 3.3 – Gráfico da variância explicada *vs.* número de componentes principais incluídas no modelo para a PCA 72

Figura 3.4 – Gráfico de *scores* para as quatro primeiras componentes principais da PCA 74

Figura 3.5 - Gráfico de *scores* para a PCA 75

Figura 3.6 – *Biplot* para a PCA 76

Figura 3.7 – 126 espectros originais para as 42 amostras de produtos de tomate. 78

Figura 3.8 – Espectros alisados pela média com janela de 15 pontos 80

Figura 3.9 – Gráfico dos resíduos de Student pela *leverage* para o modelo 1 (tabela 3.2) 81

Figura 3.10 – Gráfico de *loadings* para as 7 primeiras componentes principais do modelo 1, para sólidos totais (tabela 3.2)..... 82

Figura 3.11 – Espectros pré-processados por correção multplicativa de sinais (MSC) 83

Figura 3.12 - Gráfico de *loadings* para as sete primeiras componente principais do modelo de número 9, tabela 3.3..... 84

Figura 3.13 – Espectros pré-processados pela derivada-primeira e pela derivada-segunda 84

Figura 3.14 – Gráficos de *loadings* dos espectros pré-processados por derivadas-primeria e segunda..... 85

Figura 3.15 – a) um espectro bruto da amostra BR01; b) parte real da transformada de Fourier; c) parte imaginária da transformada de Fourier; d) espectro alisado; espectros após aplicação das derivadas e) primeira e f) segunda 87

Figura 3.16 – Gráfico de *loadings* para as sete primeiras componentes principais do modelo de calibração número 37, para licopeno..... 90

Figura 3.17 – Correlogramas entre os espectros e as propriedades: a) sólidos totais; b) sólidos solúveis; c) licopeno e d) β-caroteno..... 92

Figura 3.18 – vetores de regressão para a propriedade sólidos solúveis obtidos por PLS e por aplicação do método DWS..... 93

Figura 3.19 – Regiões dos espectros selecionadas pelo SPA..... 95

Figura 3.20 – Valores medidos *vs.* previstos para sólidos totais, sólidos solúveis, licopeno e β-caroteno em concentrados de tomate..... 97

Índice de Tabelas

Tabela 1.1 – Características físico-químicas do tomate *in natura* 4

Tabela 1.2 – Composição nutricional do tomate *in natura*. 7

Tabela 1.3 – Micronutrientes do tomate *in natura*. 8

Tabela 2.1 – Principais características, vantagens e desvantagens dos principais métodos de calibração multivariada..... 41

Tabela 3.1 – Autovalores e variâncias para cada componente principal da PCA 73

Tabela 3.2 – Modelos de calibração para sólidos e carotenóides em concentrados de tomate, utilizando os espectros inteiros, pré-processados por alisamento 80

Tabela 3.3 – Modelos de calibração para sólidos e carotenóides em concentrados de tomate, utilizando os espectros inteiros pré-processado por alisamento e MSC..... 82

Tabela 3.4 – Modelos de calibração para sólidos e carotenóides em concentrados de tomate, utilizando os espectros inteiros pré-processados pelas derivadas primeira e segunda 86

Tabela 3.5 – Modelos de calibração para sólidos e carotenóides em concentrados de tomate, utilizando a região entre 4500 a 9500 cm^{-1} dos espectros pré-processados por alisamento..... 88

Tabela 3.6 – Modelos de calibração para sólidos e carotenóides em concentrados de tomate, utilizando a região entre 4500 a 9500 cm^{-1} dos espectros pré-processados por MSC..... 89

Tabela 3.7 – Modelos de calibração PLS para sólidos e carotenóides em concentrados de tomate, utilizando a seleção de variáveis pelos correlogramas 91

Tabela 3.8 – Modelos de calibração PLS para sólidos e carotenóides em concentrados de tomate, utilizando a seleção de variáveis pelo método DWS..... 92

Tabela 3.9 – Modelos de calibração PLS para sólidos e carotenóides em concentrados de tomate, utilizando a seleção de variáveis pelo SPA 94

Tabela 3.10 – Modelos de calibração PLS para sólidos e carotenóides em concentrados de tomate, utilizando a seleção de variáveis por divisão simétrica de espectros 96

Tabela 3.11 – Validação dos modelos PLS para sólidos totais e solúveis, licopeno e β-caroteno em concentrados de tomate 96

Sumário

CAPÍTULO I - PRODUTOS DE TOMATE CARACTERÍSTICAS E PROCESSAMENTO

1.1 – Introdução 1
1.2 – Características do Fruto 4
1.3 – Cultivo e Processamento de Tomate 10
1.4 – Objetivos 14
1.5 – Referências 16

CAPÍTULO II-MÉTODOS QUIMIOMÉTRICOS DE CALIBRAÇÃO MULTIVARIADA

2.1 - Introdução 19
2.2 – A Quimiometria Aplicada à Química de Alimentos e aos Produtos de Tomate 20
2.3 – Noções de Álgebra Linear 23
    2.3.1 – Definições 23
    2.3.2 – Operações com Vetores e Matrizes 26
    2.3.3 – Inversão de Matrizes 29
    2.3.4 – Ortogonalidade e Ortonormalidade 31
    2.3.5 – A Decomposição em Valores Singulares (SVD) 32
2.4 – Métodos de Calibração Multivariada 33
    2.4.1 – Métodos de Regressão Linear Múltipla (MLR) 35
    2.4.2 – Métodos de Regressão por Compressão de Dados: PCR e PLS 36
2.5 – Métodos de Pré-Processamento de Dados 41
    2.5.1 – Os Processos de Centrar na Média e Autoescalamento 42
    2.5.2 – Técnicas de Alisamento 43
    2.5.3 – Correção Multiplicativa de Sinais (MSC) 47
    2.5.4 – Aplicação de Derivadas para Correções nas Linhas de Base 48
2.6 – Métodos de Seleção de Variáveis 50
    2.6.1 – Seleção de Variáveis pelos Correlogramas 50
    2.6.2 – Seleção de Variáveis pela Divisão Simétrica de Espectros 53
    2.6.3 – Algoritmo das Projeções Sucessivas (SPA) 54
    2.6.4 – *Dimension-Wise Selection* (DWS) 56
2.7 – Validação dos Modelos de Calibração 57
2.8 – Referências 60

CAPÍTULO III - ANÁLISE EXPLORATÓRIA DE DADOS FÍSICO-QUÍMICOS E DESENVOLVIMENTO DE MODELOS DE CALIBRAÇÃO PARA SÓLIDOS E CAROTENÓIDES EM CONCENTRADOS DE TOMATE

3.1 – Introdução 63
3.2 – Experimental 64
3.2.1 – Determinações Físico-Químicas e Instrumentais 65
3.2.2 – Aquisição dos Espectros de Infravermelho 69
3.2.3 – Técnicas Quimiométricas de Análise Exploratória e Calibração Multivariada . 70
3.3 – Análise por Componentes Principais na Identificação de Padrões nas Amostras 71
3.4 – Construção de Modelos de Calibração para Sólidos e Carotenóides 78
3.4.1 – Determinação dos Melhores Métodos de Calibração e Pré-Processamento 78
3.4.2 – Avaliação de Métodos de Seleção de Variáveis nos Modelos de Calibração 90
3.5 – Conclusões 98
3.6 – Referências 99

ANEXO I – RESULTADOS DOS ENSAIOS FÍSICO-QUÍMICOS NAS AMOSTRAS DE CONCENTRADO DE TOMATE 101

CAPÍTULO I
PRODUTOS DE TOMATE: CARACTERÍSTICAS E PROCESSAMENTO

1.1 - Introdução

O tomate (*Lycopersicon esculentum*) é um fruto originário das Américas, apesar de não haver concordância entre os autores se foi primeiramente cultivado pelos incas, no Peru, ou por tribos indígenas no México. Alguns autores acreditam que a maior parte das variedades comerciais da atualidade são originárias do tomate-cereja selvagem natural do México, enquanto outros chamam a atenção para as evidências da existência de variedades similares na América do Sul. O que se sabe ao certo é que o fruto foi levado do México para o Velho Continente no início do século XVI. O primeiro registro da planta por botânicos europeus data de 1554 e relata que o tomate era cultivado na Itália, onde foi chamado de *pomi d'oro* (maçã dourada). Posteriormente a planta tornou-se popular na França com o nome de *pomme d'amour* (maçã do amor) [1-3].

Todavia, o fruto do tomateiro não foi consumido como alimento até o início do século XIX porque acreditava-se que possuía componentes venenosos. De fato, foi inicialmente utilizado como planta ornamental e não foi antes de 1823 que sua utilização culinária tornou-se popular na Europa, apesar de haver registros do consumo do fruto por alguns povos em épocas tão remotas quanto 1596. Entretanto, assim que passou a ser empregado como alimento, seu sabor adocicado e levemente ácido fez grande sucesso e o tomate passou a integrar uma variada gama de receitas, indo da simples sopa de tomate aos mais requintados pratos da cozinha francesa [1-3].

O cultivo do tomate para fins de comercialização data do início do século XIX, mas detalhes das técnicas agrícolas só foram registrados em 1822. À época, quatro variedades vermelhas e duas amarelas eram conhecidas: a grande, a pequena, a grande-amarelo, a pêra, a cereja e a cereja-amarelo. Entretanto, conforme a popularidade do fruto como alimento foi crescendo, o número de variedades desenvolvidas também

aumentou proporcionalmente: em 1863, 23 variedades eram conhecidas; hoje, existe mais de uma centena delas. O rápido aumento no número de variedades deveu-se, principalmente, à introdução das variedades européias nos Estados Unidos, originando plantas adaptadas àquele país, e pela implementação de programas de desenvolvimento de novas variedades para atendimento da demanda de mercado [1, 4].

O tomate é um fruto sazonal, estando disponível por apenas três ou quatro meses no ano. Assim nasceu, em 1850, a indústria do tomate. Inicialmente era produzido apenas o tomate inteiro enlatado, mas a produção logo migrou para o tomate sem pele ou sementes, cortado em metades ou em óleo ou azeite. O processo de conservação em latas também popularizou o uso culinário do fruto em regiões de difícil acesso.

Apesar do emprego do tomate de mesa convencional na produção industrial, esta variedade apresentava problemas no processo fabril porque era difícil manter sua estrutura original, causando, inicialmente, dificuldades de aceitação pelos consumidores. Assim, a introdução do tomate "industrializado" desencadeou um novo período de desenvolvimento de novas variedades, mais adequadas às condições de processamento fabril [5].

No início do século XX a produção de tomate tornou-se um ramo industrial extremamente rentável, tanto para os fabricantes quanto para os engenheiros de processo, porque novas máquinas e equipamentos, específicos para esta indústria, deveriam ser construídos. Naquela época também tornou-se popular o uso de suco de tomate, envasado em garrafas de vidro e utilizado nas mais diversas aplicações. Entretanto, o ponto de inflexão da indústria do tomate deu-se com o desenvolvimento de equipamentos capazes de remover a grande quantidade de água da qual o fruto é constituído sem, no entanto, degradar sua cor, possibilitando a estocagem para utilização nos períodos de entressafra. Este produto ficou bastante popular no mercado com o nome de concentrado ou extrato de tomate [1, 5].

O tomate é constituído em sua maior parte por água, mas possui também açúcares, sais e fibras (vide seção 1.2). O método de concentração conhecido à época - e o mais utilizado até hoje - baseava-se na evaporação da água pela aplicação de calor.

Entretanto, o aquecimento direto do suco de tomate provoca a degradação térmica de açúcares, resultando num produto escuro e com sabor amargo. A aplicação de métodos de concentração sob pressão reduzida resolveu parcialmente o problema. Apesar de um produto aceitável ser obtido, sua característica final assumia notas de aroma e sabor de tomate cozido. Não obstante, o concentrado de tomate passou a ser largamente comercializado e teve ampla aceitação entre os consumidores, que não mais necessitavam picar ou bater o fruto. Entretanto, o que mais chamou a atenção dos consumidores para o novo produto foi seu elevado rendimento: porções diminutas do concentrado resultavam em preparos equivalentes a muitos tomates. O produto também passou a ser utilizado por grandes *chefs* de cozinha, atingindo *status* de ingrediente requintado [1-3].

Após o desenvolvimento do concentrado de tomate, a manufatura de produtos mais sofisticados, como molhos de tomate e o novo ketchup originário da China (*kat-si-up*, em chinês, significa "molho agridoce"), seguiu-se rapidamente. Os fabricantes perceberam depressa que o concentrado de tomate poderia ser estocado e utilizado para manufatura de produtos "secundários" durante a entressafra, diminuindo o custo fixo total das instalações. Deste modo, a partir de 1925, molhos de tomate, ketchup e concentrados de tomate de diversos tipos foram colocados no mercado. Estes novos produtos exigiram investimentos em equipamentos e na capacidade de produção, mas foram sucesso absoluto de mercado [1, 5].

O surgimento do mercado de produtos industrializados derivados de tomate gerou, então, uma nova busca por variedades mais adequadas para cada uso específico, onde eram utilizadas técnicas de cruzamento genético, plantio, preparo do solo e irrigação cada vez mais sofisticadas [1, 4, 5].

Hoje, os volumes envolvidos na produção e comercialização do tomate ainda impressionam. É o segundo vegetal mais comercializado globalmente, ficando atrás apenas da batata [1]. Em 2003, a produção mundial de tomate foi de aproximadamente 110 milhões de toneladas a um valor estimado médio de US$69,00 por tonelada. No

Brasil, produziram-se 3,3 milhões de toneladas. Os maiores produtores mundiais atualmente são Itália, Grécia, Espanha, China e Austrália [6, 7].

A maior parte do volume de tomate colhido é utilizada na manufatura de produtos como ketchup, molhos e concentrados de tomate, mas uma quantidade considerável ainda é atualmente consumida *in natura* [1, 4].

Atualmente existem mais de 150 variedades de tomate industriais disponíveis, sendo que apenas 15 respondem por mais de 90% da produção. Entretanto, este cenário tende a mudar com a utilização das modernas técnicas de manipulação genética disponíveis para aplicação na agricultura [8-10].

1.2 – Características do Fruto

O tomate pode ser descrito em termos de três grupos de propriedades: características físico-químicas, composição nutricional e teor de micro-nutrientes.

A caracterização físico-química do produto *in natura* consta na tabela 1.1. Nota-se que o fruto é composto basicamente por sólidos e água. Sólidos solúveis são constituídos principalmente por açúcares e sais dissolvidos no meio aquoso e normalmente são expressos em graus Brix (ºBrix). Sólidos insolúveis, por sua vez, são compostos por fibras vegetais, como celulose e material péctico [1, 11, 12].

Tabela 1.1 – Características físico-químicas do tomate *in natura*.

Parâmetro	%
Sólidos Totais	4,0 – 8,5
Sólidos Solúveis (º Brix)	4,0 – 6,0
Sólidos Insolúveis	0,9- 1,1
Açúcares totais	2,0 – 3,0
Frutose	1,1 – 1,5
Glicose	1,0 – 1,4
Acidez (ácido cítrico)[1]	0,3 – 0,5
Cloreto de Sódio	0,05 – 0,1
Minerais	0,3 – 0,6
Material péctico	0,17 – 0,23

Obs.: 1 – determinado como acidez total titulável.

Os açúcares do tomate são constituídos basicamente por glicose e frutose. Existem vários ácidos orgânicos no fruto, como acético, lático e málico. Todavia, o ácido cítrico está presente em concentrações cerca de trinta vezes mais elevadas que os demais e, assim, normalmente a acidez do tomate é expressa como porcentagem de ácido cítrico. O balanço entre acidez e açúcares é extremamente importante do ponto de vista sensorial porque estes compostos são os principais responsáveis pelo sabor característico do tomate. De fato, a relação acidez/Brix é comumente utilizada como parâmetro de comparação de performance entre frutos de diferentes origens ou variedades [1, 4, 13].

Os sólidos insolúveis do tomate são constituídos basicamente por fibras como celulose e pectina. Enquanto a celulose é uma fibra presente na membrana celular vegetal, a pectina constitui o material de ligação entre as células do fruto e, por isso, é também conhecida como cimento celular. Esta fibra também é a responsável pela estrutura rígida do fruto inteiro e pela viscosidade dos produtos acabados, característica intimamente relacionada ao rendimento no momento de sua aplicação culinária e, portanto, um atributo extremamente importante do ponto de vista do consumidor [1, 4].

A pectina é um polímero do ácido galacturônico (figura 1.1) formado durante o amadurecimento do fruto. Como a maioria dos materiais poliméricos, apresenta cadeias de diferentes comprimentos e pesos moleculares. Também pode ser esterificada em diferentes graus por grupos metila; as pectinas com baixo grau de esterificação (< 50%) são conhecidas como ácidos pécticos, enquanto que aquelas com elevado grau de esterificação são chamadas de ácidos pectínicos [1, 14, 15].

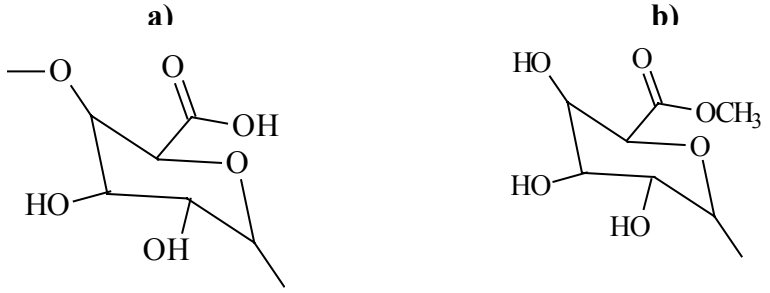

Figura 1.1 – a) ácido galacturônico e b) éster metílico do ácido galacturônico.

Em estágios avançados de amadurecimento o tomate começa a colapsar, perdendo sua estrutura rígida e tornando-se, consequentemente, desforme. Esta degradação estrutural é devida à ação de duas enzimas que atuam sobre a pectina no tecido celular conhecidas como pectinases. Uma delas, a metil-estearase péctica (PME) realiza uma hidroxilação (ou desmetilação) do ácido galacturônico (figura 1.2), o que aumenta consideravelmente a solubilidade da fibra e reduz significativamente a firmeza do fruto [1, 14, 15].

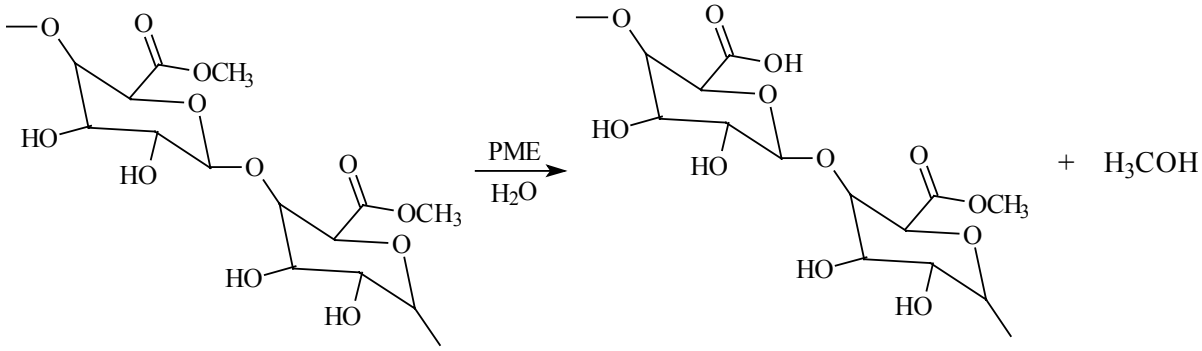

Figura 1.2 – Conversão de um éster metílico da pectina em ácido carboxílico pela ação da PME, liberando uma molécula de metanol.

A decomposição da pectina também é realizada pela poli-galacturonase (PG). Esta enzima atua clivando as ligações ozídicas entre os monômeros de ácido galacturônico, diminuindo o peso molecular médio das fibras (figura 1.3) [1, 11, 14, 15].

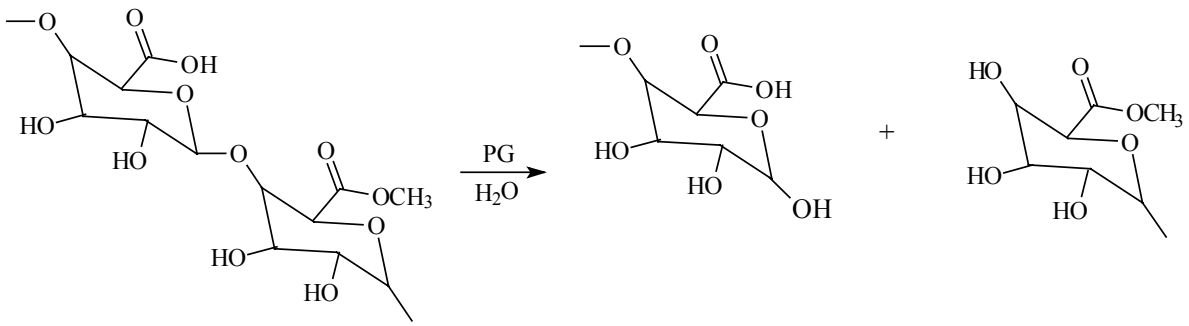

Figura 1.3 – Ação da PG na degradação da pectina do tomate pela clivagem das ligações entre os monômeros de ácido galacturônico.

A ação conjunta da PME e da PG, então, contribui para a desestruturação completa e irreversível do fruto, bem como para a perda da viscosidade da polpa de tomate. As enzimas pécticas continuam a atuar mesmo após a extração do suco de tomate no processamento fabril e, portanto, uma inativação enzimática apropriada deve ser executada (vide figura 1.5, seção 1.3) [1, 5, 14, 15].

A concentração de nutrientes em tomates varia consideravelmente de acordo com a variedade, condições de solo e adição de fertilizantes. A tabela 1.2 traz a composição nutricional média para uma mescla de variedades de tomates [1, 12]:

Tabela 1.2 – Composição nutricional do tomate *in natura*.

Nutriente	**Quantidade por 100 g**
Energia (kcal)	18,00
Água (%)	94,50
Proteínas (%)	0,88
Gordura total (%)	0,20
Gordura saturada (%)	0,04
Gordura poli-insaturada (%)	0,14
Carboidratos (%)	3,92
Açúcares totais (%)	2,63
Cinzas (%)	0,50

Obs.: valores referentes à média anual de uma mescla de variedades industriais.

O tomate guarda as principais características nutricionais da maioria dos vegetais de sua classe: possui baixo teor de calorias e gorduras, sendo descrito basicamente por água, açúcares e ácidos. Não é uma fonte significativa de minerais quando comparado com outros vegetais, sendo que potássio e fósforo estão presentes em maiores proporções. Entretanto, possui quantidades nutricionalmente significativas de ferro, provendo de 10 a 20% da Ingestão Diária Recomendada (IDR) para um adulto [1, 12, 16].

O fruto apresenta também outros micro-constituintes, como vitaminas C, pró-vitamina A (β-caroteno) e licopeno em concentrações nutricionalmente significativas (tabela 1.3). Por exemplo, um tomate médio supre aproximadamente 40% da IDR de um adulto em vitamina C e 20% em vitamina A [1, 16].

Tabela 1.3 – Micronutrientes do tomate *in natura*.

Nutriente	Quantidade por 100g
Minerais	
Cálcio (mg)	10,0
Ferro (mg)	0,27
Magnésio (mg)	11,0
Fósforo (mg)	24,0
Potássio (mg)	237
Sódio (mg)	5
Vitaminas	
A (UI)	833
C (mg)	12,7
E (mg)	0,54
Carotenóides	
Licopeno (μg)	2573
β-caroteno (μg)	449
α-caroteno (μg)	101

Obs.: valores referentes à média anual de uma mescla de variedades industriais.

Grande atenção vem sendo dada atualmente aos possíveis benefícios à saúde trazidos pelos carotenóides, especialmente o licopeno e o β-caroteno (figura 1.6). Estudos recentes trazem evidências de que estes compostos podem reduzir significativamente o risco para o desenvolvimento de alguns tipos de câncer como os de próstata, pulmão, garganta e intestino [17-20]. Sua estrutura altamente insaturada torna-os bons supressores de radicais livres (*quenchers*) e, principalmente, excelentes removedores (*scavangers*) de oxigênio singlete, principal responsável pelos processos oxidativos danosos no corpo [21, 22]; de fato, vários trabalhos vêm demonstrando seus efeitos antioxidativos no organismo, ajudando na redução do risco de ataques do coração e doenças como a arteriosclerose [18-20, 23-25]. Entretanto, estudos ainda estão em andamento para verificar se estes benefícios são devidos a um único tipo de carotenóide ou se existem efeitos sinergísticos entre eles ou com outros micro-nutrientes, como vitaminas C e E [20, 24, 25].

Alguns estudos reportam ainda que a biodisponibilidade, ou seja, a quantidade efetivamente disponível de carotenóides ao organismo, é maior em produtos processados que no vegetal *in natura*, contrariando o senso comum de que o processamento geralmente contribui para a redução de micronutrientes nos alimentos.

Este maior biodisponibilidade deve-se principalmente pelo fato de que os processos de concentração e aquecimento liberam as moléculas de carotenóides das lipoproteínas às quais estão conjugadas no tecido celular, tornando-as passíveis de serem absorvidas pelo organismo [26 - 28].

Os carotenos são constituídos por isoprenos organizados de tal maneira que as unidades isoprenóides são revertidas no centro da molécula. Podem existir no estado livre no tecido vegetal ou em combinação com lipoproteínas. Em plantas, também ocorrem na forma de ésteres de ácidos graxos [11, 29, 30].

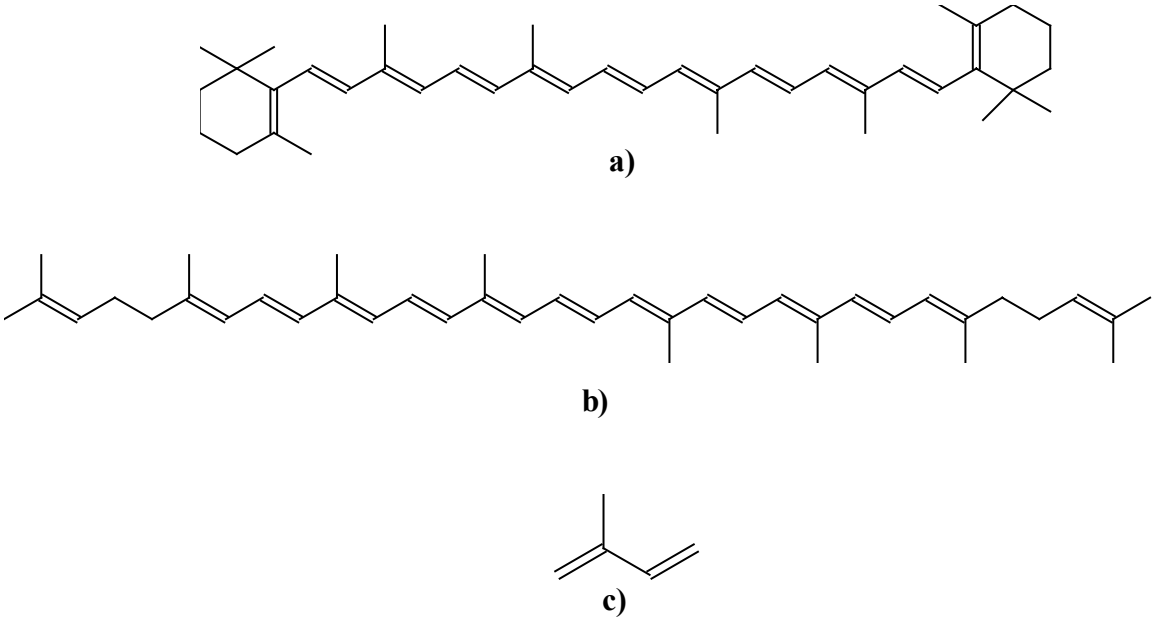

Figura 1.4 – Carotenóides, como o β-caroteno (a) e o licopeno (b), são formados por repetições de unidades de isopreno (c).

Apesar de não ser o vegetal que apresenta os maiores níveis destes carotenóides – a melancia tem mais licopeno e cenoura muito mais β-caroteno -, a maior parte da ingestão, principalmente do licopeno, vem do consumo de tomates, em grande parte porque este vegetal está presente em uma grande variedade de receitas e pratos em todos os continentes [1].

Devido aos possíveis benefícios nutricionais destes carotenóides, produtores de tomate, instituições de pesquisa e a indústria de alimentos trabalham fortemente no desenvolvimento de novas variedades com altos teores de licopeno e β-caroteno. Além de serem utilizados na indústria para manufatura de produtos funcionais, estes frutos poderiam servir como fonte barata de carotenóides para as indústrias farmacêutica e cosmética.

1.3 – Cultivo e Processamento de Tomate

O cultivo do tomate é uma atividade extremamente rentável, especialmente em um país como o Brasil, cuja produção agrícola tem papel importante na economia. Atualmente, as principais plantações de tomate estão na região centro-oeste, mais precisamente nas proximidades da cidade de Goiânia, estado de Goiás. Quantidades consideráveis ainda são produzidas na região de Uberlândia e Patos de Minas, em Minas Gerais, mas a tendência é que o volume principal migre para Goiânia. O fruto já foi extensivamente cultivado na região de São Paulo e, no passado, várias fábricas estavam instaladas naquele estado. Entretanto, conforme as novas variedades industriais foram melhor se adaptando às regiões mais quentes e secas, a produção começou a migrar.

A indústria alimentícia adquire os frutos diretamente dos produtores. As indústrias fabricantes de produtos de tomate fornecem as sementes das variedades apropriadas tanto aos processos quanto aos produtos finais que comercializarão, bem como os adubos, fertilizantes e defensivos agrícolas, enquanto que os produtores entram com a terra e a mão-de-obra. Os fabricantes também determinam quais as técnicas de plantio, tratamento do solo, irrigação e colheita deverão ser utilizadas nos produtos que irão comprar. Atualmente a colheita do tomate é realizada mecanicamente por colheitadeiras, o que também ajuda a reduzir seu preço por tonelada [1, 4].

No momento da contratação dos serviços, a indústria alimentícia e os produtores concordam parâmetros de qualidade para um tomate padrão, que inclui características físicas como cor, tamanho, defeitos visuais, doenças, folhas e pedúnculos. Entretanto, a quantidade média de sólidos, totais ou solúveis, é um dos parâmetros de maior peso na composição do preço final porque define o rendimento fabril. A quantidade de sólidos no fruto *in natura* pode variar dependendo das condições de solo e climáticas, principalmente da quantidade de chuvas durante o desenvolvimento e no período de colheita.

Como explanado anteriormente, o tomate é um fruto sazonal, estando o período de colheita compreendido entre os meses de julho ao início de setembro. Assim, a indústria alimentícia utiliza dois processos de fabricação distintos na manufatura de produtos de tomate, chamados de processo primário e secundário [1]. A figura 1.5 traz uma representação esquemática simplificada destes processos:

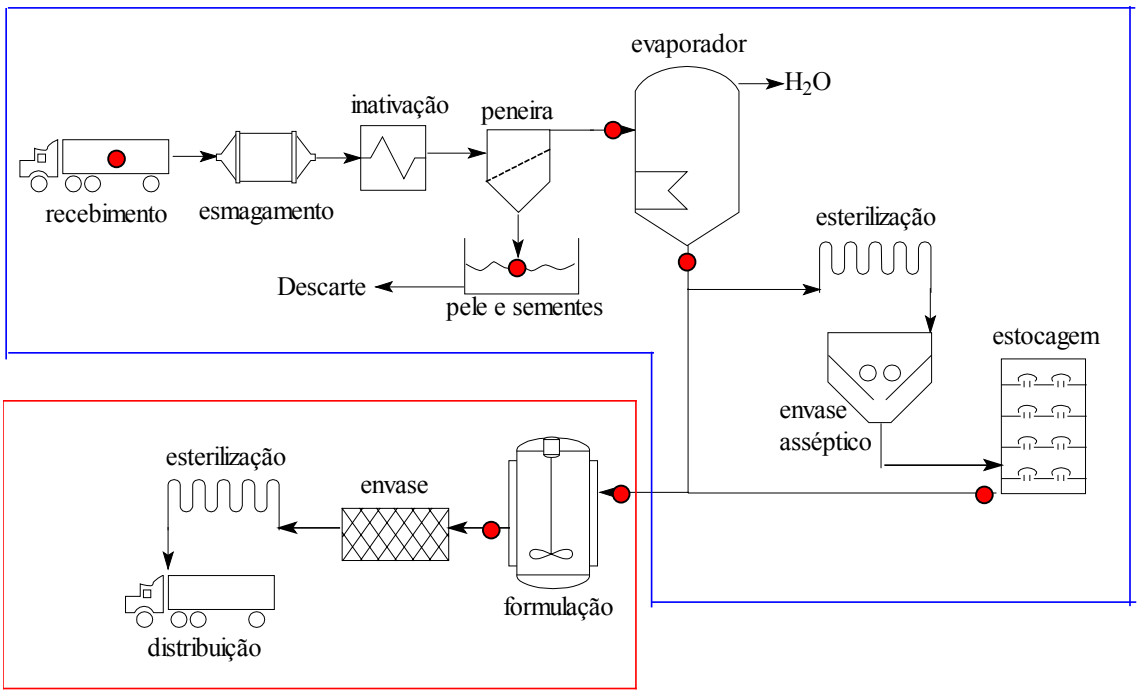

Figura 1.5 – Esquema simplificado do processamento de tomates. O processo primário está representado dentro do polígono em azul; o retângulo vermelho engloba o processo secundário. Os pontos vermelhos mostram os locais onde medidas de sólidos (totais ou solúveis) são obrigatórias.

O processo primário é empregado apenas durante o período de safra, pois consiste na transformação do fruto em polpa de tomate, uma matéria-prima extremamente viscosa e com elevado teor de sólidos totais (entre 25 e 35%), ou em cubos de tomate, que serão utilizados na produção dos molhos prontos. Esta etapa do processamento inicia-se com o recebimento do tomate. Os frutos são normalmente entregues em caminhões com carrocerias abertas, cobertas por lonas. Segue-se uma pesagem do veículo cheio para medição da quantidade entregue e neste momento realiza-se uma amostragem, utilizando sonda metálica automática. A amostra é homogeneizada e tem seus atributos visuais avaliados, além de realizarem-se medidas objetivas de cor, Brix e pH. Uma vez dentro dos parâmetros aceitáveis, o descarregamento do produto é autorizado. Outra porção da amostra segue para determinações mais sofisticadas, como sólidos insolúveis, material péctico, viscosidade, consistência, pesticidas, etc.

O tomate é lavado enquanto é descarregado em percursos hídricos e transportado para o interior da fábrica. Nesta etapa, um controle da vazão da água de transporte é fundamental para que não haja deterioração dos frutos. No percurso hídrico também são colocadas telas metálicas que vão da superfície da água até aproximadamente 1/5 da profundidade do leito. Estas telas servem como barreira para material estranho que porventura esteja flutuando entre os tomates, sendo manual ou mecanicamente removidos. Também possuem comportas que interrompem o fluxo de entrada do fruto na fábrica caso ocorra algum problema operacional que impeça a continuação do processo.

Os tomates são retirados do percurso hídrico por uma esteira de elevação, onde recebem um banho de hipoclorito de sódio em solução, que tem por finalidade reduzir a carga microbiana e remover qualquer material estranho mais fino, que não foi anteriormente retido nas telas. Da esteira elevatória passam por uma esteira de seleção manual, onde frutos defeituosos, doentes ou fora dos padrões de cor e textura são removidos.

Uma peneira de aço com pás rotatórias promove o esmagamento dos tomates, onde separam-se as sementes e o suco. Este último passa ainda por nova etapa de peneiramento para máxima remoção de peles e sementes e é então bombeado para um aquecedor, onde ocorre a inativação enzimática.

O suco inativado passa por uma bateria de pré-aquecimento, sendo em seguida concentrado em pressão reduzida. Usualmente são utilizados equipamentos de três estágios, podendo ser resfriados por efeito *flash cooler* [31]. O suco de tomate não pode ser concentrado em temperaturas muito elevadas para que não ocorra caramelização dos açúcares ou reações de escurecimento não enzimático. Deste modo, pressão e temperatura são parâmetros exaustivamente ajustados e monitorados durante o processamento para que o produto permaneça nos concentradores apenas pelo tempo suficiente para que uma dada quantidade de sólidos (totais ou solúveis) seja atingida [1].

O resultado do processo de concentração é a matéria-prima polpa de tomate, uma pasta viscosa com teor de sólidos de tomate variando entre 25 e 35%. Uma parte desta polpa é bombeada até esterilizadores, onde passa por etapas de aquecimento, *holding* e resfriamento. Após, é assepticamente envasada em *bags* apropriados, colocados em tambores de 250 kg ou bins de 1,2 tonelada, que são estocados em galpões ao abrigo das intempéries para serem utilizados durante a entressafra. Outra parte da polpa é encaminhada ao processo secundário. Note que em nenhum momento qualquer conservante é adicionado durante o processamento.

Pelo acima exposto, nota-se que o processo primário é de extrema importância na para a indústria de produtos de tomate porque a confecção de todos os demais produtos depende dele [1]. Além disso, caso o processo de concentração não fosse empregado durante a safra, não haveria espaço suficiente, a um custo aceitável, para estocagem da quantidade de tomate necessária para manufatura de produtos secundários durante a entressafra!

Durante a entressafra, a matéria-prima polpa de tomate é diluída para manufatura dos produtos secundários, como ketchup, molhos e concentrados de

tomate. Esta etapa do processo consiste basicamente de tanques de formulação, que são grandes recipientes de aço-inoxidável encamisados e dotados de gigantescos mexedores.

Os concentrados de tomate normalmente recebem em sua formulação apenas tomate, sal e açúcar. Dentro desta categoria incluem-se os purês de tomate (6 a 10% de sólidos solúveis), os simples concentrados (15 a 22% de sólidos solúveis) e os duplo-concentrados de tomate (25 a 30% de sólidos solúveis). Sua manufatura é realizada misturando-se, nos tanques de formulação, água, polpa de tomate, sal e açúcar, em quantidades pré-estabelecidas. O material é então misturado sob aquecimento e enviado às máquinas de envase. Após, passam por processo de esterilização, que consiste em uma seqüência de aquecimento e resfriamento, por um túnel de secagem e finalmente pela seção de embalagem. As caixas seguem então para os depósitos para posterior distribuição ao mercado.

1.4 - Objetivos

O objetivo deste trabalho é definir um método de análise rápido e não destrutivo para determinação simultânea de sólidos totais e solúveis, bem como dos carotenóides licopeno e β-caroteno em produtos de tomate, utilizando espectrometria no infravermelho próximo (NIR) e calibração multivariada.

Torna-se evidente, pelas informações apresentadas nas seções anteriores, que a concentração de sólidos de tomate é um parâmetro de extrema relevância, não apenas para controle da matéria-prima *in natura* mas também durante todas as fases de processamento, indo até o controle de qualidade dos produtos finais que chegam aos consumidores. Entretanto, os métodos de análise hoje disponíveis são destrutivos e, no caso de sólidos totais, demorados.

Sólidos solúveis são comumente determinados por refratometria: uma alíquota líquida da amostra é obtida por filtração ou centrifugação e colocada entre os dois prismas do refratômetro, formando um filme delgado. Luz de uma lâmpada de halogênio é forçada a atravessar o filme e o índice de refração da substância é

determinado. Como a concentração de sólidos solúveis é diretamente proporcional ao índice de refração, o instrumento é calibrado com soluções-padrão de sacarose e os resultados são expressos em porcentagem de sólidos solúveis, ou º Brix. Como o índice de refração é dependente da temperatura, os instrumentos modernos possuem sistemas de correção para pequenas diferenças entre as temperaturas de calibração e da amostra no momento da leitura [1, 11].

Todavia, a determinação de sólidos solúveis torna-se instável quando amostras com Brix elevado são tratadas por simples filtração. Nestes casos uma quantidade suficiente de material particulado passa através do meio filtrante, fazendo parte da composição do filme entre os prismas do refratômetro. Este material causa o espalhamento de luz e a medida do índice de refração torna-se imprecisa.

Apesar de a clarificação da amostra poder ser conseguida também por centrifugação, este é um procedimento lento demais para aplicação no controle de processo. Por outro lado, a calibração de um método espectroscópico pode ser realizada utilizando o tratamento por centrifugação como método de referência, obtendo-se, assim, resultados precisos e de maneira rápida.

Sólidos totais são determinados por secagem em estufa a vácuo, a 70 ºC. O processo pode levar de três a cinco horas, dependendo da quantidade de água e da eficiência do vácuo do instrumento. Atualmente, existem balanças analíticas acopladas a fornos de microondas capazes de realizar a determinação em poucos minutos para amostras com baixa concentração de sólidos solúveis. Todavia, para amostras com quantidade de sólidos solúveis acima de aproximadamente 15 Brix é necessária uma etapa de diluição, o que torna o método muito lento e dependente do analista, características indesejáveis para um parâmetro de controle de processos. Caso esta diluição não seja realizada, ou a eficiência de secagem será muito baixa, principalmente devido à viscosidade do produto, ou ocorrerá carbonização local de açúcares, ambos os efeitos fornecendo resultados inexatos. Além disso, esses equipamentos não podem ser utilizados para automação de processo devido à impossibilidade de sua implantação *on line*.

Planos de monitoramento e controle de processo em tempo real são consideravelmente onerosos e de eficiência questionável quando os métodos de análise são relativamente lentos em relação à demanda exigida pelos volumes de uma fábrica de produtos de tomate. Por outro lado, análises espectrométricas, quando aliadas a métodos de calibração multivariada, podem realizar estas determinações de maneira rápida, confiável e não destrutiva, fornecendo informação em tempo real sobre as condições de processo.

Além dos planos de monitoramento, técnicas espectroscópicas aliadas à quimiometria, com especial atenção à espectroscopia no infravermelho próximo, podem ser empregadas em programas de desenvolvimento de novas variedades de tomates, reduzindo consideravelmente o tempo e o volume de recursos aplicados até o desenvolvimento de um novo fruto com as características desejáveis.

1.5 – Referências

1. Gould, W.A.; "Tomato Production, Processing & Technology"; 3rd Ed.; CTI Pub. Inc.; Baltimore 1992.
2. Morrison, G.; "Tomato Varieties"; Michigan State Coll.; Michigan 1938.
3. May, E.C.; "The Canning Clan."; The MacMillan Co.; Westminster; 1937.
4. Alencar, R.C.S., "O Tomateiro"; LTC; São Paulo; 1979.
5. Judje, A.J.; "The History of Canning Industry"; Edward E. Judge & Sons; Westminster 1914.
6. USDA online, http://www.fas.usda.gov, 11/06/2004.
7. IBGE online, http://www.ibge.gov.br, 11/06/2004.
8. Robinson, C.; "Genetic Modification Technology and Food"; ILSI Europe; Brussels 2001.
9. Lajolo, F.M., Nutti, M.R.; "Transgênicos: Bases Científicas da sua Segurança"; SBAN; São Paulo 2003.
10. Voet, D., Voet, J., Pratt, C.W.; "Fundamentos de Bioquímica"; Artmed Editora; Porto Alegre 2000.
11. Fennema, O.R.; "Food Chemistry"; 3rd. Ed.; Marcel Dekker, Inc.; New York; 1996.
12. USDA Nutrient Data Laboratory; http://www.nal.usda.gov/fnic/foodcomp/; 20/07/2004.

13. Nielsen, S.S.; "Food Analysis"; 2nd Ed.; Gaithersburg; Aspen 1998.

14. Porreta, S., Poli, G., Palmieri, L.; "The Effect of PME on Tomato Structure During Ripening and the Use of Calcium Salts During tomato Processing"; *Sciences Aliments*; **1994**; *46*; 100-108.

15. Barret, D.M., Garcia, E. Wayne, J.E.; "Optimization of the Calcium Addition Process of Canned Tomatoes"; *Crit. Rev. Food Sci. Nutr.*; **1998**; *38*; 173-258.

16. ANVISA; RDC 33/98; 13/01/1998.

17. Giovannucci, E.; "Tomatoes, Tomato-Based Products, Lycopene, and Cancer: Review of the Epidemiological Literature"; *J. Natl. Cancer Inst.*; **1999**; 91; 317-331.

18. Clinton, SK.; "Lycopene: Chemistry, Biology and Implications for Human Health and Disease"; *Nutr. Rev.*; **1998**, *56*; 35-51.

19. Bramley, P.; "Is Lycopene Beneficial to Human Health?"; *Phytochem.*; **2000**; *54*; 233-236.

20. Tapiero, H., Townsend, D.M., Tew, K.D.; "The Role of Carotenoids in the Prevention of Human Pathologies"; *Biomed. Pharmacotherapy*; **2004**; *58*; 100 – 110.

21. Di Mascio, P., Kaiser, S., Sies, H.; "Lycopene as the Most Efficient Biological Carotenoid Singlet Oxygen Quencher"; *Arch. Biochem. Biophys.*; **1989**, 274, 532-538.

22. Sies, H., Stahl, W.; "Lycopene: Antioxidant and Biological Effects and its Bioavailability in the Human"; *Proc Soc. Exp. Biol. Med.*; **1998**; 218; 121-124.

23. Willcox, J., Catignani, G.L., Lazarus, S; "Tomato and Cardiovascular Health"; *Crit. Rev. Food Sci. Nutr.*; **2003**; *43(1)*; 1-18.

24. Gerster, H.Y.; "The Potential Role of Lycopene for Human Health"; *J. Am. Coll. Nutr.*; **1997**; 16; 109-126.

25. Stahl W, Sies H.; "Lycopene: A Biologically Important Carotenoid for Humans?"; *Arch. Biochem. Biophys.*; **1996**, *336*; 1-9.

26. Gärtner, C., Stahl, W., Sies, H.; "Lycopene is More Bioavailable from Tomato Paste than from Fresh Tomatoes"; *Am. J. Clin. Nutr.*; **1997**; *66*; 116-122.

27. Stahl, W., Sies, H.; "Uptake of Lycopene and its Geometrical Isomers is Greater from Heat-Processed than from Unprocessed Tomato Juice"; *J. Nutr.*; **1992**; 122; 2161-2166.

28. van het Hof, K.H., *et alli*; "Carotenoid Bioavailability in Humans from Tomatoes Processed in Different Ways Determined from the Carotenoid Response in the Triglyceride-Rich Lipoprotein Fraction of Plasma After a Single Consumption and in Plasma After Four Days of Consumption"; *J. Nutr.*; **2000**; *130*; 1189-1196.

29. Isler, O.; "Carotenoids"; cap. I; Birkhäuser Verlag Basel, Stuttgart 1971.

30. Jain, C.K., Argawal, S., Rao, V.; "The Effect of Dietary Lycopene on Bioavailability, Tissue Distribution, *in vivo* Antioxidant Properties and Colonic Preneoplasia in Rats"; *Nutr. Res.*; **1999**; *19*; 1383-1391.

31. Amaya, D.B.R.; "A Guide to Carotenoid Analysis in Foods"; ILSI; Washington 1999.

32. Brennan, J.G. *et alli*; "Food Engineering Operations"; Elsevier; London 1990.

CAPÍTULO II

MÉTODOS QUIMIOMÉTRICOS DE CALIBRAÇÃO MULTIVARIADA

2.1 - Introdução

A moderna instrumentação de análises químicas é capaz de gerar uma quantidade considerável de dados, sobre uma única amostra, em um curto espaço de tempo: um espectrômetro pode registrar sinais provenientes de mais de mil comprimentos de onda ou um único cromatograma pode apresentar mais de cem picos [1, 2]. Assim, para que informação útil seja obtida deste grande volume de dados é necessário que se utilizem técnicas matemáticas adequadas, sendo a quimiometria um dos campos de estudo da química que fornece tais ferramentas [1-3].

A quimiometria pode ser definida como a aplicação de métodos matemáticos e estatísticos a dados de origens distintas para obtenção de informação química. Consiste de um conjunto de técnicas de cálculo com o objetivo de promover a obtenção de informação útil de um conjunto complexo de dados, englobando conceitos de planejamento experimental, pré-processamento de dados, estatística e análise multivariada [1-4].

As técnicas quimiométricas podem ser genericamente divididas em três classes distintas: análise exploratória de dados, construção de modelos quantitativos de calibração e construção de modelos qualitativos de classificação. A análise exploratória utiliza basicamente os métodos de Análise das Componentes Principais (*Principal Component Analysis* – PCA) e a Análise de Agrupamentos Hierárquicos (*Hierarquical Clustering Analisys* – HCA), enquanto que os modelos de classificação mais utilizados são o *Soft Independent Modelling by Class Analogy* (SIMCA) e a análise pelo K-ésimo Visinho mais Próximo (K^{th}-*Nearest Neighbour* - KNN). Modelos de classificação estão fora do escopo deste trabalho, mas boas referências são encontradas no final deste capítulo [3, 5, 6].

Dentre as técnicas quimiométricas, as de calibração multivariada destacam-se porque possibilitam a realização de determinações quantitativas de compostos

presentes na ordem de porcentagem, ou até de micro-constituintes, em matrizes complexas. Outras vantagens de sua aplicação são a redução considerável da necessidade de preparo de amostras, bem como da utilização de reagentes químicos por vezes nocivos ao homem ou ao meio-ambiente, e a possibilidade da utilização de métodos não-destrutivos que podem ser aplicados *on line* em processos fabris. Além disso, a aplicação de métodos quimiométricos reduz a necessidade de seletividade instrumental através da utilização de um número relativamente grande de sinais, englobando uma gama maior de informação no modelo matemático [1, 2].

Devido a sua natureza inerentemente multivariada, as técnicas quimiométricas utilizam-se extensivamente da manipulação de vetores e matrizes, objeto de estudo do ramo da matemática denominado álgebra linear (excelentes obras de referência são fornecidas ao final deste capítulo) [7-10]. Entretanto, com o objetivo de tornar clara a explanação sobre as técnicas de calibração multivariada, um breve resumo das principais operações matriciais da álgebra linear utilizadas em quimiometria são expostas na seção 2.3.

2.2 – A Quimiometria Aplicada à Química de Alimentos e aos Produtos de Tomate

Métodos quimiométricos são largamente empregados em química de alimentos, não apenas na forma de modelos de calibração mas também para realização de análises exploratórias e construção de modelos de classificação. Descrições destas técnicas são fornecidas no capítulo 2.

Análises exploratórias são comumente utilizadas para determinar possíveis relações entre propriedades físicas e constituintes químicos em matrizes alimentícias, correlações entre as diversas variáveis instrumentais ou subjetivas ou na definição do comportamento dos alimentos em função de variações em sua composição ou na forma de seu processamento. Entretanto, as técnicas de análise sensorial são as que se utilizam destes métodos com maior freqüência, provavelmente devido a sua natureza inerentemente multivariada. Nestas análises muitos atributos são avaliados em

relativamente poucas amostras, gerando matrizes de dados com características muito similares àquelas com as quais os quimiometristas estão acostumados a trabalhar [1, 2].

Métodos de calibração multivariada encontraram terreno fértil para o desenvolvimento de aplicações em alimentos devido principalmente à complexidade e variedade de matrizes que esses produtos apresentam, tornando demorados e caros os métodos clássicos de análise química. Mesmo os instrumentos mais sofisticados dependem de complicados procedimentos de tratamento de amostras. Assim, uma das principais vantagens da aplicação da calibração multivariada a matrizes alimentícias é a redução ou até eliminação do tratamento de amostras. Estas etapas são normalmente substituídas por técnicas matemáticas de tratamento do sinal instrumental [1, 2].

A quimiometria também começou a se popularizar entre os químicos de alimentos graças aos vários instrumentos que já são fornecidos com softwares de tratamento de dados. Também existem programas independentes que apresentam interfaces gráficas extremamente amigáveis e de fácil utilização, como o Unscrumbler (Camo Software, Inc.), o Pirouette (Infometrix, Inc.) e o PLS Toolbox (Eigenvector Resarch, Inc.), sendo que este último funciona na plataforma Matlab (The MathWorks, Inc.).

Devido à complexidade das matrizes alimentícias, os métodos de análise são, em sua grande maioria, destrutivos, impedindo sua aplicação para controle *on line* de processo. Deste modo, a utilização de métodos espectroscópicos, aliados à calibração multivariada, são atualmente extensivamente estudados em química de alimentos, visando principalmente a redução do tempo e da extensiva manipulação das amostras [1, 2].

A espectroscopia estuda a interação da matéria com a radiação eletromagnética. Dependendo da energia da radiação incidente, um ou mais dos seguintes processos pode ocorrer: reflexão, espalhamento, fluorescência/fosforescência, reação fotoquímica ou absorção. A energia associada à radiação eletromagnética é diretamente proporcional à sua freqüência (ν) e, portanto, inversamente proporcional ao comprimento de onda (λ), de modo que $E = h\nu = h(c/\lambda)$, em que "c" é a velocidade da

luz e "*h*" é a constante de Planck. Também é comum em espectroscopia a utilização do *número de onda*, $\tilde{\nu}$, definido como 1/λ, no lugar da freqüência. Deste modo, $E = hc\tilde{\nu}$ [11, 12].

A absorção de energia da radiação eletromagnética pode ocorrer devido a transições eletrônicas entre orbitais atômicos ou moleculares ou a mudanças de estados rotacionais ou vibracionais das moléculas. Na região do infravermelho ocorrem absorções principalmente devido a estas duas últimas transições, sendo que para o estado condensado as rotacionais podem ser negligenciadas por apresentarem intensidades muito baixas [11, 12].

A região do infravermelho no espectro eletromagnético pode ser ainda dividida em duas sub-regiões distintas. O infravermelho próximo (NIR, *Near Infrared*), compreende a radiação com números de onda entre 4000 e 10000 cm^{-1} e o infravermelho médio (MIR, *Mid Infrared*), entre 650 e 4000 cm^{-1}. No infravermelho médio ocorrem as *transições fundamentais*, assim chamadas porque a molécula passa do estado fundamental (ou de menor energia) para o estado excitado imediatamente superior. Esta característica faz da espectroscopia no infravermelho médio uma técnica excelente na caracterização de compostos orgânicos, pois cada ligação característica de um grupo funcional apresenta uma banda de vibração em uma freqüência específica. Entretanto, quantificações com esta região do espectro só foram possíveis com a utilização de técnicas quimiométricas [11, 12].

A região do infravermelho próximo foi por muitos anos quase que abandonada pela maioria dos químicos porque nela ocorrem apenas transições de sobretom (*overtones*), ou seja, do estado fundamental para dois ou até três níveis excitados acima, cujas intensidades de absorção são bastante reduzidas, dificultando a utilização na quantificação de compostos. Além disso, não há definição de bandas características de grupos funcionais devido à ocorrência de sobreposições e também por fenômenos de recombinação e ressonâncias de Fermi [11, 13, 14].

Entretanto, com o desenvolvimento dos métodos quimiométricos e com a popularização dos computadores pessoais, a técnica de infravermelho próximo passou

a ser amplamente utilizada na determinação quantitativa de uma série de compostos em matrizes alimentícias complexas. Além de não ser destrutiva, podendo ser implementada *on line*, apresenta vantagens como rapidez na aquisição dos espectros e a redução ou eliminação do tratamento de amostras, com proporcional diminuição no uso de reagentes e solventes químicos [1, 12].

Em alimentos, utilizando a técnica de infravermelho próximo [11, 13, 14] foram determinados: gordura em leite [15]; acidez, sólidos solúveis e firmeza em maçãs [16]; o teor de açúcares em suco de laranja [17] e as concentrações de açúcares em Satsuma chinesa [18], kiwi [19] e pêssegos inteiros [20].

O primeiro trabalho envolvendo calibração multivariada de parâmetros de qualidade para tomates utilizando NIR foi publicado em 1998 por Hong e Tsou [21]. Calibraram-se sólidos solúveis, acidez total e a razão a/b da cor Hunter [22], utilizando regressão linear múltipla (MLR) em 4 ou 5 comprimentos de onda, dependendo do parâmetro. Em 2003, Goula e Adamopoulos [23] estimaram umidade, açúcares, acidez, proteína e sal em produtos de tomate utilizando também MLR sobre 3 dos 9 comprimentos de onda fornecidos pelo espectrômetro de infravermelho. Mais recentemente, Shyam e Matsuoka [24] determinaram a razão ácido/brix por NIR, utilizando PLS.

2.3 – Noções de Álgebra Linear [7-10]

2.3.1 - Definições

A álgebra linear lida basicamente com vetores e matrizes. Os primeiros são usualmente representados por letras minúsculas em negrito (**x**), enquanto que matrizes por letras maiúsculas em negrito (**X**).

Um escalar (representado por letras minúsculas em itálico, *u*), é definido como uma quantidade descrita por uma magnitude, ou seja, um único número, enquanto que um vetor é constituído por um arranjo ordenado de escalares, sendo cada um deles um *elemento* do vetor:

$$\mathbf{s} = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} \tag{2.1}$$

sendo que, neste exemplo, **s** é um *vetor coluna*. Entretanto, um vetor também pode ser representado na forma de uma linha e, neste caso, é denominado *vetor linha*:

$$\mathbf{t}^{\mathrm{T}} = \begin{bmatrix} t_1 & t_2 & t_3 & t_4 \end{bmatrix} \tag{2.2}$$

Vetores são usualmente representados como vetores coluna. Um vetor linha é representado pela *transposição* (vide seção 2.2.2) de um vetor coluna.

A posição de um elemento em um vetor é determinada por um sub-índice em itálico, *i.* Assim, t_3 é o terceiro elemento do vetor $\mathbf{t}^{\mathrm{T}}$.

A principal e marcante diferença entre vetores e escalares reside no fato de que os primeiros são definidos por uma magnitude e uma direção, trazendo, deste modo, uma quantidade maior de informação sobre o sistema que representam. Tomem-se como exemplos os vetores $\mathbf{k}^{\mathrm{T}} = [1 \quad 0 \quad 0]$, $\mathbf{v}^{\mathrm{T}} = [0 \quad 1 \quad 0]$ e o vetor $\mathbf{s}^{\mathrm{T}} = [1 \quad 1 \quad 1]$; suas representações geométricas são mostradas na figura 2.1:

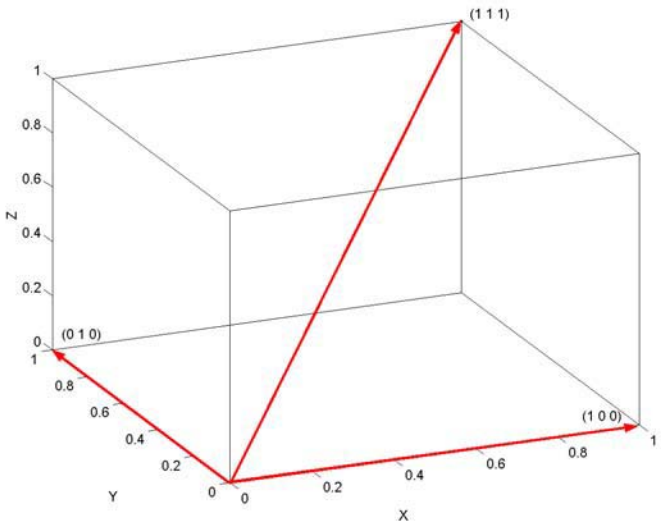

Figura 2.1 – Representação geométrica dos vetores **k**, **v** e **s**.

Um conceito extremamente importante em álgebra linear é o de *dimensionalidade*, que é definida como sendo o número de escalares ou elementos de um vetor. Por exemplo, o vetor **s** tem dimensão três e o vetor **t** tem dimensão igual a quatro. A dimensionalidade de um vetor é um conceito importante porque define o tamanho ou dimensão do *espaço* onde este vetor está contido. A representação da dimensão do espaço definido por um vetor segue a notação $\Re^n$, sendo n a dimensionalidade do vetor. Assim, $\mathbf{s} \in \Re^3$ e $\mathbf{t} \in \Re^4$.

Apesar de ser impossível representar geometricamente vetores com dimensão maior que três, esta limitação não implica em absoluto que tais vetores não existam. De fato, o objetivo da álgebra linear é justamente realizar operações com entidades de dimensionalidade superior a três.

Outro conceito extremamente importante em álgebra linear e que é bastante explorado em quimiometria é o de *sub-espaço*. Um sub-espaço é um espaço de dimensão menor contido em outro, de dimensão maior. Por exemplo, o vetor **t** tem espaço de dimensão quatro ($\in \Re^4$) e, por isso mesmo, contém os espaços de ordem inferior: $\Re^3$, $\Re^2$ e $\Re^1$.

Matrizes são entidades definidas pelo arranjo apropriado de um número de vetores. Tomem-se como exemplos as matrizes **A** e **B** abaixo:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} \tag{2.3}$$

em que a posição de cada elemento na matriz é definido por seus sub-índices. Assim, a_{ij} é o elemento da matriz **A** que se encontra na linha i e coluna j.

Os elementos de uma matriz podem ser números reais, imaginários, funções, outras matrizes ou quaisquer outras entidades. Como em química as matrizes são resultantes de ensaios analíticos, normalmente os elementos das matrizes são definidos por números reais.

No caso de matrizes podem-se distinguir dois espaços: o *espaço linha* e o *espaço coluna*. O primeiro é definido pelos vetores que compõe as linhas e o segundo por aqueles com compõe as colunas de uma matriz. Cada um destes espaços possui sua própria dimensionalidade. Por exemplo, o espaço linha da matriz **B** tem dimensão dois, enquanto que seu espaço coluna tem dimensão igual a três. Ambos os espaços são representados numa única notação, $\Re^{mxn}$, em que *m* representa o número de linhas e *n* o número de colunas. Assim, $\mathbf{A} \in \Re^{2x2}$ e $\mathbf{B} \in \Re^{3x2}$. Note que o número total de elementos em uma matriz é dado pelo produto das dimensões dos espaços linha e coluna. Assim, a matriz **A** tem quatro elementos (2 x 2) enquanto que a matriz **B** possui seis elementos (3 x 2).

O valor do espaço de menor dimensão de uma matriz é denominado *posto*. Assim, ambas as matrizes em 2.3 possuem posto igual a dois. Na seção 2.2.3 será fornecida uma extensão do conceito de posto de uma matriz.

2.3.2 – Operações com Vetores e Matrizes

A transposição de matrizes e vetores constitui uma operação extensivamente utilizada em álgebra linear. A transposta converte vetores linhas em vetores colunas, e *vice-versa*. Analogamente, matrizes têm suas linhas convertidas em colunas, e colunas em linhas. A notação para transposição de uma matriz genérica **G** é $\mathbf{G}^T$ e a operação pode ser definida como:

$$g_{ij}^{\ T} = g_{ji} \qquad \textbf{(2.4)}$$

Note que a transposta de uma matriz corresponde à rotação de seus elementos em torno da diagonal principal, sendo que os elementos da diagonal principal de uma matriz são definidos por g_{ij} para $i = j$. Para uma matriz genérica $\mathbf{G} \in \Re^{3x4}$:

$$\mathbf{G} = \begin{bmatrix} g_{11} & g_{12} & g_{13} & g_{14} \\ g_{21} & g_{22} & g_{23} & g_{24} \\ g_{31} & g_{32} & g_{33} & g_{34} \end{bmatrix}, \qquad \mathbf{G}^{\mathrm{T}} = \begin{bmatrix} g_{11} & g_{21} & g_{31} \\ g_{12} & g_{22} & g_{32} \\ g_{13} & g_{23} & g_{33} \\ g_{14} & g_{24} & g_{34} \end{bmatrix} \tag{2.5}$$

Uma das operações mais simples efetuadas em álgebra linear é a multiplicação de um vetor ou uma matriz por um escalar. Esta operação é realizada multiplicando-se cada elemento do vetor ou matriz pelo escalar. Assim:

$$c\mathbf{A} = c\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} ca_{11} & ca_{12} \\ ca_{21} & ca_{22} \end{bmatrix} \tag{2.6}$$

Matrizes **A** e **B** também podem ser somadas, desde que suas dimensões sejam iguais (**A** e **B** $\in \Re^{mxn}$), pois a operação é realizada elemento a elemento:

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & a_{13}+b_{13} \\ a_{21}+b_{21} & a_{22}+b_{22} & a_{23}+b_{23} \end{bmatrix} \tag{2.7}$$

Pela equação 2.7 também é possível demonstrar que tanto a propriedade comutativa quanto a propriedade associativa aplicam-se a soma de matrizes, ou seja, **A + B = B + A** e **C + (A + B) = (C + A) + B**.

A subtração de matrizes segue as mesmas premissas de dimensionalidade de sua adição, de modo que a operação **A – B** pode ser interpretada como **A** + (-1)**B**, em que a operação de multiplicação pelo escalar –1 é aplicada à matriz **B**.

O produto de matrizes é uma das operações mais importantes em álgebra linear e pode ser melhor compreendida através da extensão do produto escalar de dois vetores. Sejam **u** e **v** dois vetores genéricos tais que **u** $\in \Re^{1x3}$ e **v** $\in \Re^{3x1}$. O produto escalar **u.v** é dado por:

$$\mathbf{u}.\mathbf{v} = \begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \left[\left(u_1 \times v_1 \right) + \left(u_2 \times v_2 \right) + \left(u_3 \times v_3 \right) \right] = uv \qquad \textbf{(2.8)}$$

em que *uv* é um escalar e o número de colunas de **u** é igual ao número de linhas de **v**.

O produto escalar entre dois vetores também pode ser utilizado para calcular a norma-2 ou norma Euclidiana de um vetor. A norma-2 de um vetor **u** é expressa pela notação $\|\mathbf{u}\|$ e é definida como a raiz quadrada da soma quadrática dos elementos deste vetor:

$$\|\mathbf{u}\| = \sqrt{\sum_{j=1}^{J} u_j^{\,2}} = \sqrt{\mathbf{u}^{\mathrm{T}}\mathbf{u}} \qquad \textbf{(2.9)}$$

em que *J* é o número de elementos do vetor.

O conceito de produto escalar entre dois vetores pode ser estendido para o produto entre duas matrizes. Sejam duas matrizes genéricas $\mathbf{A} \in \Re^{3x2}$ e $\mathbf{B} \in \Re^{2x2}$. O produto $\mathbf{AB} = \mathbf{C}$, sendo $\mathbf{C} \in \Re^{3x2}$ é dado por:

$$\mathbf{AB}=\mathbf{C}=\begin{bmatrix} a_{11} & a_{21} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} (a_{11}\times b_{11})+(a_{21}\times b_{21}) & (a_{11}\times b_{12})+(a_{21}\times b_{22}) \\ (a_{21}\times b_{11})+(a_{22}\times b_{21}) & (a_{21}\times b_{12})+(a_{22}\times b_{22}) \\ (a_{31}\times b_{11})+(a_{32}\times b_{21}) & (a_{31}\times b_{12})+(a_{32}\times b_{22}) \end{bmatrix} \qquad \textbf{(2.10)}$$

em que o número de colunas de **A** é igual ao número de linhas de **B**.

Dadas duas matrizes genéricas $\mathbf{A} \in \Re^{mxn}$ e $\mathbf{B} \in \Re^{nxq}$, um elemento qualquer c_{ij} da matriz $\mathbf{C} \in \Re^{mxq}$, resultado do produto escalar entre o vetor da *i*-ésima linha de **A** com a *j*-ésima coluna de **B**, pode ser definido como:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} \times b_{kj} \qquad \textbf{(2.11)}$$

Note que a propriedade comutativa não se aplica à multiplicação de matrizes. Assim, geralmente $\mathbf{AB} \neq \mathbf{BA}$.

2.3.3 – Inversão de Matrizes

A inversa de uma matriz genérica **G** é representada por $\mathbf{G}^{-1}$. Quando a matriz **G** é multiplicada por sua inversa a resultante é a matriz identidade:

$$\mathbf{G}^{-1}\mathbf{G} = \mathbf{G}\mathbf{G}^{-1} = \mathbf{I}, \text{ para } \mathbf{G} \in \Re^{m\text{x}m} \tag{2.12}$$

em que a condição $\mathbf{G} \in \Re^{m\text{x}m}$ revela que **G** deve ser quadrada.

A *matriz identidade*, **I**, caracteriza-se por possuir todos os elementos da diagonal principal iguais a um e os demais iguais a zero:

$$\begin{aligned} a_{ij} = \delta, \quad & \text{sendo } \delta = 1 \text{ se } i = j \\ & \text{e } \delta = 0 \text{ se } i \neq j. \end{aligned} \tag{2.13}$$

A matriz identidade recebe esta denominação porque, quando multiplicada por qualquer outra matriz, resulta na própria matriz:

$$\begin{aligned} \mathbf{AI} &= \mathbf{A} \\ \mathbf{IB} &= \mathbf{B} \end{aligned} \tag{2.14}$$

Assim, nota-se imediatamente, na equação 2.12, a semelhança entre a representação da multiplicação de uma matriz por sua inversa com a multiplicação de um número escalar por seu recíproco, em que $a \times a^{-1} = a^{-1} \times a = 1$.

Para que uma matriz possua inversa deve ser *lineramente independente*, ou seja, nenhuma linha ou coluna pode ser expressa como uma combinação linear de quaisquer outras. A matriz **E** abaixo é um exemplo de matriz onde a terceira coluna é uma combinação linear das duas primeiras:

$$\mathbf{E} = \begin{bmatrix} 1 & 3 & 4 \\ 2 & 4 & 6 \\ 3 & 2 & 5 \end{bmatrix}$$

Neste caso, o posto da matriz é igual a dois, e não três como o seria se a definição fornecida em 2.2.1 fosse aplicada. Deste modo, **E** é tida como uma matriz de *posto incompleto* ou *singular*. Matrizes singulares não são passíveis de inversão. Deste modo, a definição precisa do posto de uma matriz é dada pelo menor número de linhas ou colunas linearmente independentes desta matriz.

Quando a matriz não é quadrada pode-se calcular uma matriz denominada *pseudo-inversa*. Uma forma clássica de cálculo da pseudo-inversa de uma matriz $\mathbf{X} \in \Re^{mxn}$, $m \neq n$, é denominada pseudo-inversa de *Moore-Penrose*, representada pelo símbolo $\mathbf{X}^{+}$:

$$\mathbf{X}^{+} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \qquad \textbf{(2.15)}$$

em que $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ é quadrada ($\in \Re^{nxn}$) e terá inversa caso não seja singular. Esta condição certamente não será atendida se **X** possuir mais colunas que linhas; ou seja, para o cálculo de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ o número de amostras deve ser maior que o número de variáveis na matriz **X**.

Também deve-se atentar para o fato de que a inversa da matriz $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ torna-se instável quando **X** é *mal-condicionada*. Um resultado instável significa que qualquer mínima alteração nos valores de **X** promove resultados bastante diferentes para sua inversa.

Uma matriz mal-condicionada tem posto completo, mas está bastante próxima de uma matriz singular. De fato o *número de condição*, *K*, é um parâmetro usualmente empregado para avaliação da *condição* do uma matriz e pode ser interpretado como o recíproco da distância da matriz **X** original de uma matriz ligeiramente modificada $\mathbf{X}_{\boldsymbol{m}}$, singular. Deste modo, quanto maior o número de condição, mais instável é a inversa da matriz **X**. Segue abaixo exemplo de uma matriz 2 x 2 mal condicionada:

$$\mathbf{X} = \begin{bmatrix} 1 & 2{,}0001 \\ 2 & 3{,}9999 \end{bmatrix},\ \mathbf{X}^{-1} = 10^{4}\begin{bmatrix} -1{,}3333 & 0{,}6667 \\ 0{,}6667 & -0{,}3333 \end{bmatrix},\ K = 8{,}3\times 10^{4}.$$

Nota-se que, apesar de ter posto completo, $\mathbf{X} \cong \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$, que tem posto incompleto e, portanto, é singular.

Se **X** for ligeiramente modificada, por exemplo, no elemento x_{22}, a inversa assume valores completamente diferentes, dez vezes menores que os anteriores:

$$\mathbf{X}_m = \begin{bmatrix} 1 & 2{,}0001 \\ 2 & 3{,}9996 \end{bmatrix},\ \mathbf{X}_m^{-1} = 10^3 \begin{bmatrix} -6{,}6660 & 3{,}3335 \\ 3{,}3333 & -1{,}6667 \end{bmatrix},\ K = 4{,}2 \times 10^4.$$

em que $\mathbf{X}_m$ continua mal-condicionada.

2.3.4 – Ortogonalidade e Ortonormalidade

Diz-se que dois vetores são *ortogonais* quando seu produto escalar é igual a zero. O produto escalar entre dois vetores quaisquer, $\mathbf{u}^T$ e **v,** também pode ser escrito como (vide equação 2.8):

$$\mathbf{u.v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos\theta \qquad \textbf{(2.16)}$$

em que θ é o ângulo entre os **u** e **v**.

Assim, para vetores não nulos, **u.v** = 0 se, e somente se, cosθ = 0, implicando em θ = 90°.

Dois vetores **a** e **b** quaisquer serão *ortonormais* se o produto escalar entre eles for igual a um.

Uma matriz **G** qualquer será ortogonal quando $\mathbf{G}^T\mathbf{G} = \mathbf{I}$; ou seja, os vetores que definem suas colunas devem ser ortonormais. A inversa de uma matriz ortogonal é igual a sua transposta ($\mathbf{G}^{-1} = \mathbf{G}^T$). Esta é uma importante propriedade em álgebra linear, pois computacionalmente o cálculo de inversas de matrizes requer muito mais operações que sua transposição [10].

Uma implicação relevante da ortogonalidade de uma matriz é que suas colunas são linearmente independentes, ou seja, nenhuma coluna pode ser escrita como uma combinação linear de qualquer outra (seção 2.2.3). De fato, esta é uma propriedade extensivamente explorada, não apenas em quimiometrica como na álgebra linear como um todo.

2.3.5 – A Decomposição em Valores Singulares (*Singular Value Decomposition* - SVD)

Matrizes podem ser decompostas como o produto de duas ou mais matrizes diferentes, mas que possuem propriedades interessantes. Esta estratégia é utilizada em uma série de aplicações em álgebra linear, incluindo a resolução do problema dos quadrados mínimos utilizado nos métodos de calibração [25].

A decomposição em valores singulares (*Singular Value Decomposition* – SVD) é aplicada em vários algoritmos quimiométricos (seção 2.3.2). Esta decomposição transforma uma matriz $\mathbf{X} \in \Re^{ixj}$ no produto de três matrizes, **U, S** e **V**:

$$\mathbf{X} = \mathbf{USV}^{T} \qquad \textbf{(2.17)}$$

em que a matriz $\mathbf{U} \in \Re^{ixi}$, $\mathbf{S} \in \Re^{ixj}$ e $\mathbf{V} \in \Re^{jxj}$.

As matrizes **U** e **V** são ortogonais. Suas colunas trazem, respectivamente, os autovetores dos produtos cruzados $\mathbf{XX}^{T}$ e $\mathbf{X}^{T}\mathbf{X}$**.** A matriz **S** é diagonal e traz os valores singulares destas matrizes, ordenados em ordem decrescente. Os valores singulares relacionam-se com os autovalores (λ) da matriz **X**, em que:

$$s_j = \sqrt{\lambda_j} \qquad \textbf{(2.18)}$$

A matriz **U** relaciona-se com a matriz $\mathbf{XX}^{T}$ de correlação para as amostras e **V** relaciona-se com a matriz $\mathbf{X}^{T}\mathbf{X}$ de correlação para as variáveis e, sendo assim, seus autovalores serão idênticos.

Um algoritmo para realização da decomposição SVD pode ser encontrado em Golub e Van Loan [10]. Entretanto, a maioria dos pacotes em álgebra linear já possui rotinas otimizadas para execução desta decomposição [26, 27].

2.4 – Métodos de Calibração Multivariada

Há muitos anos os químicos se utilizam de vários métodos de calibração para determinação indireta de analitos em uma variada gama de matrizes. Estes métodos cresceram consideravelmente com a instrumentação analítica, pois as respostas destes dispositivos são constituídas por sinais eletrônicos, que são posteriormente correlacionados matematicamente com propriedades ou concentrações de substâncias de interesse [1, 4].

Todavia, até aproximadamente meados da década de 1970 os métodos de calibração eram constituídos basicamente pela regressão de uma propriedade química qualquer, mais usualmente a concentração de um analito, sobre uma única variável, geralmente um único canal de um instrumento analítico [4].

Os métodos de calibração univariada exigiam, e ainda exigem, considerável seletividade instrumental porque o único canal selecionado deve apresentar sinal exclusivamente dependente da propriedade de interesse, sendo livre de interferência devido a outras substâncias. Deste modo, a aplicação destas técnicas requer o desprendimento de quantidades consideráveis de recursos na determinação de procedimentos de tratamento de amostras para isolamento do analito de interesse e/ou na construção de instrumentos cada vez mais seletivos [1-4].

Com o surgimento dos métodos de calibração multivariada, entretanto, o cenário em análise química começou a mudar. Ao invés de desenvolver métodos analíticos seletivos, passou-se a investir recursos no desenvolvimento de técnicas matemáticas capazes de transformar sinais oriundos de misturas de componentes em informação útil sobre uma ou várias propriedades de interesse. Estes métodos ganharam força com o desenvolvimento de computadores pessoais equipados com maiores recursos e mais acessíveis aos pesquisadores e aos profissionais da química [1-3].

Outra característica dos métodos multivariados na determinação de uma propriedade de interesse é a chamada *vantagem multicanal*. Quando apenas um canal ou variável é utilizado, a calibração torna-se extremamente sensível a ruídos nesta variável. A utilização de diversos canais minimiza ou mesmo elimina esta interferência, tornando a calibração mais robusta [1-3, 12, 28].

Entre os métodos de calibração multivaridada encontram-se o MLR (*Multiple Linear Regression*), o PCR (*Principal Components Regression*) e o PLS (*Partial Least Squares*). Todos assumem relação linear dos dados instrumentais com a propriedade de interesse [1-4, 12, 14, 28-30]:

$$y_i = \sum_{j=1}^{J} b_j x_j + e_i \qquad \textbf{(2.19)}$$

sendo b_k a constante de proporcionalidade da x_k-ésima variável instrumental, y_i o valor da propriedade para a i-ésima amostra e e_i o resíduo da estimativa. Usualmente várias amostras são empregadas na construção de um modelo de calibração, originando um sistema de equações. Assim, a equação 2.19 pode ser escrita, em notação matricial, como [1, 7-10]:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \qquad \textbf{(2.20)}$$

em que **y** é o vetor com os valores medidos da propriedade de interesse para I amostras, **X** é a matriz de espectros com I linhas e J colunas de comprimentos de onda, **e** é o vetor de resíduos e **b** é o vetor de regressão com as constantes de proporcionalidade que relacionam as variáveis com a propriedade de interesse. Efetuar uma calibração, então, consiste em encontrar os valores numéricos do vetor **b** [1, 3, 13, 28, 29].

As equações 2.19 e 2.20 mostram que os métodos de calibração multivariada comumente utilizam o que se convencionou denominar *métodos inversos de calibração*. Nestes métodos o vetor de regressão é obtido pela projeção da propriedade de interesse sobre o espaço gerado pelas colunas da matriz de variáveis, enquanto que nos *métodos diretos* projetam-se as variáveis na propriedade de interesse [1, 4]:

$$\mathbf{X} = \mathbf{yb} + \mathbf{e} \qquad (2.21)$$

Apesar da aparente semelhança, as equações 2.20 e 2.21 trazem implícita a informação de onde o erro experimental se encontra em maior monta. O método de quadrados mínimos preconiza que a variável independente é livre ou possui erro desprezível frente àquele apresentado pela variável dependente. Esta afirmativa procede quando padrões de calibração são construídos para determinação de analitos em matrizes razoavelmente simples num processo de calibração clássico. Com os avanços da instrumentação científica, entretanto, os desvios analíticos estão normalmente contidos nos métodos de referência utilizados na calibração multivariada e, deste modo, o método inverso é o mais amplamente utilizado [1, 4].

2.4.1 – Método da Regressão Linear Múltipla (MLR)

O método MLR opera resolvendo o sistema expresso pelas equações 2.19 e 2.20 pela estratégia dos quadrados mínimos, utilizando as equações normais [1, 4, 7, 31]:

$$\begin{aligned} \mathbf{X}\hat{\mathbf{b}} &= \mathbf{y} \\ \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{b}} &= \mathbf{X}^{\mathrm{T}}\mathbf{y} \\ \hat{\mathbf{b}} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \end{aligned} \qquad (2.22)$$

em que o símbolo "^" indica que o vetor de regressão **b** assim obtido é um estimativa.

O principal senão desta estratégia é que, como visto anteriormente a pseudoinversa de *Moore-Penrose* só existe se **X** tiver mais linhas que colunas (seção 2.2.3); ou seja, um número maior de amostras deve estar disponível para que a calibração seja realizada. Além disso, a matriz **X** deve estar livre de colinearidade – suas linhas e colunas devem ser linearmente independentes -, o que geralmente não ocorre em espectroscopia. Nota-se, assim, que para utilização do método MLR a dados espectroscópicos um número reduzido de comprimentos de onda deve ser empregado na calibração e, quando um volume considerável de informação é fornecido pelo instrumento, uma criteriosa seleção de variáveis deve ser executada [1, 32, 33].

Outra característica do método MLR é que toda a informação disponível nas variáveis empregadas nos cálculos é utilizada pelo modelo. Assim, apesar de menos tendenciosos, modelos obtidos por MLR incluem também o ruído instrumental [1-3].

2.4.2 – Métodos de Regressão por Compressão de Dados: PCR e PLS

Para resolver as limitações do método MLR criaram-se ferramentas de compressão de dados que não apenas eliminam a colinearidade, mas também reduzem a quantidade de ruído inserida no modelo matemático final. Os métodos PCR e PLS utilizam a compressão de dados por Componentes Principais (*Principal Components*, PCs) para este fim. O método PCA (*Principal Components Analysis*) decompõe a matriz **X** em duas matrizes distintas, **T** e **L** [1-3, 28, 30, 34]:

$$\mathbf{X} = \mathbf{T}\mathbf{L}^{\mathrm{T}} \tag{2.23}$$

em que **T** é a matriz de *scores* e **L** é a matriz de *loadings*. Cada coluna de **T** e **L** corresponde a uma componente principal (PC).

Existem basicamente dois algoritmos pelos quais as matrizes **T** e **L** podem ser obtidas. O principal deles utiliza a decomposição SVD da matriz **X** (seção 2.2.5), sendo que:

$$\begin{aligned} \mathbf{T} &= \mathbf{U}\mathbf{S} \\ \mathbf{L} &= \mathbf{V} \end{aligned} \tag{2.24}$$

Assim, as colunas da matriz de *scores*, **T**, trazem os autovetores da matriz de correlação $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ multiplicados pelos autovalores desta mesma matriz [1, 2, 34]:

$$\lambda_j = \sum_{i=1}^{I} t_{ij}{}^2 = \mathbf{t}_j{}^{\mathrm{T}}\mathbf{t}_j \tag{2.25}$$

sendo que λ_j é o autovalor para a *j*-ésima coluna de **T**.

As colunas de **T** e **L** são ordenadas pela grandeza dos autovalores, ou seja, quanto maior a somatória quadrática expressa na equação 2.25, mais à esquerda esta coluna estará na matriz correspondente. Assim, seleciona-se o número de componentes principais (*A*) a manter no modelo pela comparação entre estes autovalores. Componentes principais com autovalores pequenos trazem pouca informação (ou apenas ruído) para o modelo e podem ser desprezadas.

A determinação do número de componentes principais a manter é melhor visualizada através de um gráfico de autovalores pelo número de componentes principais inclusas no modelo. Quando λ assume valores comparáveis com o resíduo experimental, ou quando seu valor não mais diminui com o aumento do número de componentes principais, significa que apenas ruído está sendo incluso no modelo, e componentes subsequentes podem ser desprezadas. O número de componentes principais mantidas no modelo final também é denominado de *pseudo-posto* da matriz **X** [1-3, 34].

A matriz de *loadings*, **L**, traz os pesos de cada variável original para a construção das componentes principais: variáveis com um peso grande na construção de cada PC contribuem de maneia substancial para aquela PC, enquanto que valores pequenos pouco contribuem. Cabe salientar que uma variável pode possuir um peso considerável para uma dada PC, mas desprezível para outra [1-3, 28-30].

Outra característica importante das matrizes **T** e **L** e que é explorada em quimiometria, é a ortogonalidade dos vetores que definem as colunas destas matrizes (seções 2.2.3 e 2.2.4). Por possuírem colunas ortogonais não existe colinearidade entre os novos eixos definidos pelas componentes principais [1-3, 28-30, 34].

Assim, com uma única decomposição matricial consegue-se reduzir o número de variáveis (normalmente, $A << K$), eliminar a colinearidade em **X** e eliminar ou reduzir o ruído incluso no modelo [1-3, 28-30, 34].

Um outro método utilizado na obtenção de *loadings* e *scores* é o algoritmo NIPALS desenvolvido por Hermann Wold em 1966 [35]. Trata-se de um método iterativo que extrai uma componente principal por vez. O algoritmo segue as seguintes etapas:

1. Toma-se como estimativa inicial dos primeiros *scores*, $\mathbf{t}_1$, a coluna de $\mathbf{X}$ com o maior somatório quadrático, ou seja, a coluna com a maior variância;
2. Estima-se, por quadrados mínimos, os *loadings* $\mathbf{l}_1$ para estes *scores*;
3. Normalizam-se os *loadings* dividindo $\mathbf{l}_1$ por $\|\mathbf{l}_1\|$;
4. Novos *scores* $\mathbf{t}_1^n$são então calculados a partir de $\mathbf{l}_1$: $\mathbf{t}_1^n = \mathbf{X}\mathbf{l}_1^T$;
5. Repetem-se as etapas 2 a 4 até que haja convergência dos *scores*;
6. Calcula-se a matriz $\mathbf{X}_r$, recomposta pela primeira componente principal: $\mathbf{X}_r = \mathbf{t}_1\mathbf{l}_1^T$;
7. Calcula-se uma matriz de resíduos: $\mathbf{E}_1 = \mathbf{X}_r - \mathbf{X}$;
8. Repete-se o processo para extração de todas as A componentes principais, substituindo $\mathbf{X}$ por $\mathbf{E}_n$, para $1 \leq n \leq A$.

A principal vantagem deste algoritmo é que torna possível determinar *a priori* o número de componentes principais (A) a extrair de $\mathbf{X}$, economizando recursos computacionais e tempo. Esta característica é particularmente vantajosa quando o número de variáveis ou comprimentos de onda da matriz $\mathbf{X}$ é muito grande [1, 30, 35].

A análise de componentes principais também é largamente utilizada como método de análise exploratória de dados. Gráficos de *loadings* e *scores* revelam padrões característicos do comportamento das amostras em função de um conjunto de variáveis que dificilmente seriam reconhecidos pela observação de valores tabelados [1-3, 34]. De fato, no capítulo 3 deste trabalho uma análise exploratória utilizando gráficos de PCA é apresentada.

Pela compressão por PCA, o modelo matemático da regressão linear (PCR) pode ser expresso como [1-3, 28, 30]:

$$\mathbf{y} = \mathbf{T}^{A}\mathbf{b} + \mathbf{e} \qquad \textbf{(2.26)}$$

em que $\mathbf{T}^{A}$ indica que a matriz de *scores* original foi truncada nas A primeiras componentes principais. Alguns autores [2] sugerem que os *loadings* também sejam considerados no modelo, obtendo-se assim um vetor de regressão **b'** diretamente relacionado com as variáveis originais:

$$\begin{aligned} \mathbf{y} &= \mathbf{T}^{A}\mathbf{L}^{A\mathrm{T}}\mathbf{b} + \mathbf{e} \\ \mathbf{y} &= \mathbf{T}^{A}\mathbf{b}' + \mathbf{e} \end{aligned} \qquad \textbf{(2.27)}$$

em que $\mathbf{b'} = \mathbf{L}^{\mathrm{AT}}\mathbf{b}$.

A principal desvantagem do método PCR é que apenas as variáveis da matriz **X** são consideradas na extração das componentes principais, sendo a propriedade de interesse negligenciada neste processo. O método PLS, por outro lado, considera a propriedade de interesse no processo de extração das componentes principais, que neste caso são denominadas *variáveis latentes* ou *fatores* [1-3, 29, 36].

Assim, a principal diferença entre os métodos PCR e PLS é que, enquanto o primeiro busca maximizar apenas a variância das colunas da matriz **X** na extração das componentes principais, o segundo maximiza a covariância da propriedade de interesse em **y** com as variáveis de **X** para determinação dos fatores.

O algoritmo para realização da calibração pelo método PLS segue basicamente as mesmas premissas do método NIPALS [1, 29, 30, 36]:

1. O vetor **y** é utilizado como uma primeira estimativa para os *scores*: $\mathbf{t}_1 = \mathbf{y}$;
2. Um vetor de pesos **w** é extraído pela projeção por quadrados mínimos de $\mathbf{t}_1$ em **X**;
3. O vetor **w** é normalizado: $\mathbf{w}^n = \mathbf{w}/||\mathbf{w}||$;
4. Um novo vetor de *scores* $\mathbf{t}_1^{\,n}$ é estimado por quadrados mínimos, utilizando **X** e **w**;
5. Novo **w** é calculado e checa-se pela convergência de $\mathbf{t}_1$.

6. Calculam-se os *loadings* dos espectros **X,** por quadrados mínimos, utilizando **t**: $\mathbf{l}_1 = \mathbf{X}^T\mathbf{t}_1 / (\mathbf{t}_1{}^T\mathbf{t}_1)$;
7. Calcula-se o *loading* (q_1) da propriedade **y**, por quadrados mínimos, utilizando $\mathbf{t}_1$: $q_1 = \mathbf{y}^T\mathbf{t}_1 / (\mathbf{t}_1{}^T\mathbf{t}_1)$;
8. Calculam-se as matrizes de resíduos, $\mathbf{E} = \mathbf{X} - \mathbf{t}_1\mathbf{l}_1{}^T$ e $\mathbf{f} = \mathbf{y} - \mathbf{t}_1 q_1$;
9. Os cálculos são repetidos substituindo-se **X** por **E** e **y** por **f** até que os A fatores tenham sido extraídos.

Nota-se que a maximização da covariância entre a matriz **X** e o vetor **y** é conseguida através da matriz **W** de pesos para os *loadings* de **X**. Nota-se também que enquanto a matriz **W** é ortogonal, **L** não o é e, assim, a interpretação destes *loadings* deve ser realizada com cautela. Usualmente **L** e **W** são visualizadas em um mesmo gráfico. Normalmente assemelham-se para os primeiros fatores, mas **L** tende a se diferenciar de **W** conforme o número de fatores aumenta [1, 36].

Uma vez extraídos os *scores* (**T**)*, loadings* (**L** e **q**), e a matriz de pesos (**W**) para todos os A fatores, pode-se calcular o vetor de regressão do método PLS como [1, 29, 36]:

$$\mathbf{b} = \mathbf{W}(\mathbf{L}^T\mathbf{W})^{-1}\mathbf{q} \qquad \textbf{(2.28)}$$

A tabela 2.1 traz as principais características, vantagens e desvantagens dos métodos de calibração multivariada aqui apresentados [1-3, 13, 28-36]:

Tabela 2.1 – Principais características, vantagens e desvantagens dos principais métodos de calibração multivariada.

Método	Vantagens	Desvantagens
MLR	- Rápido; - Simples – quadrados mínimos; - De fácil interpretação; - Utiliza toda informação de **X** no modelo; - Não é tendencioso.	- Exige mais amostras que variáveis; - **X** deve estar livre de colinearidade; - Inclui resíduos no modelo.
PCR	- Elimina colinearidade em **X**; - O número de variáveis pode ser maior que o de amostras; - Relativamente rápido; - De fácil interpretação;	- Exige um número maior de PC's que PLS; - Tendencioso.
PLS	- Elimina colinearidade em **X**; - O número de variáveis pode ser maior que o de amostras; - Tende a fornecer modelos com menos PCs que o método PCR.	- Muito tendencioso: desvantajoso quando a precisão em **y** é relativamente pequena; - Complexo.

É importante notar que os métodos de regressão por compressão de dados, PCR e PLS, fornecem resultados idênticos ao método MLR quando todas as componentes principais ou fatores são empregados na calibração [1-3]. De fato, esta propriedade é utilizada no cálculo de regressão linear de matrizes mal-condicionadas ou de posto incompleto.

2.5 – Métodos de Pré-Processamento de Dados

Os métodos de pré-processamento são utilizados para padronizar a entrada de dados aos métodos quimiométricos de análise, por exemplo, tornando variáveis obtidas por técnicas analíticas distintas, que possuem ordens de grandeza diferentes, comparáveis entre si, ou para que informação não relevante seja eliminada dos modelos. Dentre as técnicas empregadas em espectroscopia destacam-se os processos de centrar na média ou auto-escalar os dados, os métodos de alisamento, a correção multiplicativa de sinais (*Multiplicative Signal Correction* - MSC) e as derivadas primeira e segunda [1-3, 32, 33, 37].

2.5.1 – Os Processos de Centrar na Média e Auto-Escalamento [1, 3, 37]

O procedimento de centrar os dados na média é bastante simples: basta subtrair o valor de cada elemento da *j*-ésima coluna pela média calculada para esta mesma coluna:

$$x_{ij}{}^{c} = x_{ij} - \bar{x}_j \qquad \textbf{(2.29)}$$

em que $x_{ij}{}^{c}$ é o elemento da *i*-ésima linha da coluna *j* já centrado na média e $\bar{x}_j$ é o valor médio da coluna *j*.

Deste modo, a média de cada coluna da matriz $\mathbf{X}^{c}$ será zero. O resultado deste procedimento é o deslocamento da origem do hiper-espaço para o centro de informação descrito pelos dados originais, não havendo qualquer distorção da informação disponível. Abaixo segue um exemplo de aplicação do procedimento de centrar na média para uma matriz **A**:

$$\mathbf{A} = \begin{bmatrix} 1 & 5 & -7 \\ 3 & 4 & 12 \\ 12 & -4 & 1 \\ -1 & -2 & 15 \end{bmatrix},\ \mathbf{m} = \begin{bmatrix} 3{,}75 & 0{,}75 & 5{,}25 \end{bmatrix},\ \mathbf{A}^{c} = \begin{bmatrix} -2{,}75 & 4{,}25 & -12{,}25 \\ -0{,}75 & 3{,}25 & 6{,}75 \\ 8{,}25 & -4{,}75 & -4{,}25 \\ -4{,}75 & -2{,}75 & 9{,}75 \end{bmatrix}$$

em que **m** é o vetor linha contendo as médias das colunas de **A** e $\mathbf{A}^{c}$ é a matriz resultante, centrada na média.

O processo de auto-escalamento, como o próprio nome sugere, consiste em padronizar a escala de todas as variáveis para desvio-padrão igual a um. É obtido tomando-se os dados centrados na média e dividindo cada elemento pelo desvio-padrão dos elementos de sua coluna correspondente na matriz **X**:

$$x_{ij}{}^{a} = \frac{x_{ij} - \bar{x}_j}{s(x_j)} \qquad \textbf{(2.30)}$$

em que x_{ij}^{a} representa o elemento da linha i e coluna j já auto-escalado e s(x_j) representa o desvio-padrão dos elementos da j-ésima coluna de **X**. Para a matriz **A:**

$$\mathbf{s} = \begin{bmatrix} 5{,}7373 & 4{,}4253 & 10{,}1448 \end{bmatrix},\ \mathbf{A}^{a} = \begin{bmatrix} -0{,}4793 & 0{,}9604 & -1{,}2075 \\ -0{,}1307 & 0{,}7344 & 0{,}6645 \\ 1{,}4380 & -1{,}0734 & -0{,}4189 \\ -0{,}8279 & -0{,}6214 & 0{,}9611 \end{bmatrix}$$

sendo que **s** é o vetor contendo as variâncias das colunas de **A** e $\mathbf{A}^{a}$ é a matriz **A** auto-escalada.

Em espectroscopia os canais são comprimentos de onda e, deste modo, qualquer variação espectral contém informação relevante sobre a amostra. Assim, o procedimento de auto-escalamento raramente é empregado para dados desta origem. Por outro lado, este procedimento é bastante útil quando os dados são provenientes de determinações de técnicas analíticas distintas, tornando-os comparáveis. De fato, este processo é largamente empregado em análise exploratória de dados.

2.5.2 – Técnicas de Alisamento [1, 3, 38]

Técnicas de alisamento são empregadas para que ruído instrumental aleatório seja removido dos espectros, aumentando a razão sinal/ruído. Um dos procedimentos mais simples é o alisamento pela média, também conhecido como *box car*. Baseia-se na definição de um intervalo ou janela de N pontos (comprimentos de onda) dos quais a média do sinal instrumental será calculada. O ponto central desta janela é substituído pelo valor médio das absorbâncias. A janela é deslocada para o ponto $N + 1$ e uma média das absorbâncias deste novo intervalo é calculada. O procedimento é repetido até que a janela percorra todo o espectro.

A figura 2.2 exemplifica este processo para uma curva gaussiana onde ruído foi artificialmente adicionado. Em (a) é mostrada a curva com baixa razão sinal/ruído exatamente na região de maior informação espectral; uma janela de 15 pontos (b) é selecionada para percorrer todo o espectro ruidoso. Neste exemplo, os 15 primeiros comprimentos de onda são substituídos por um único ponto alocado no centro da janela e possuindo o valor médio das absorbâncias (c). O espectro resultante (d) tem exatamente as mesmas características do original, porém possui uma relação sinal/ruído bastante superior.

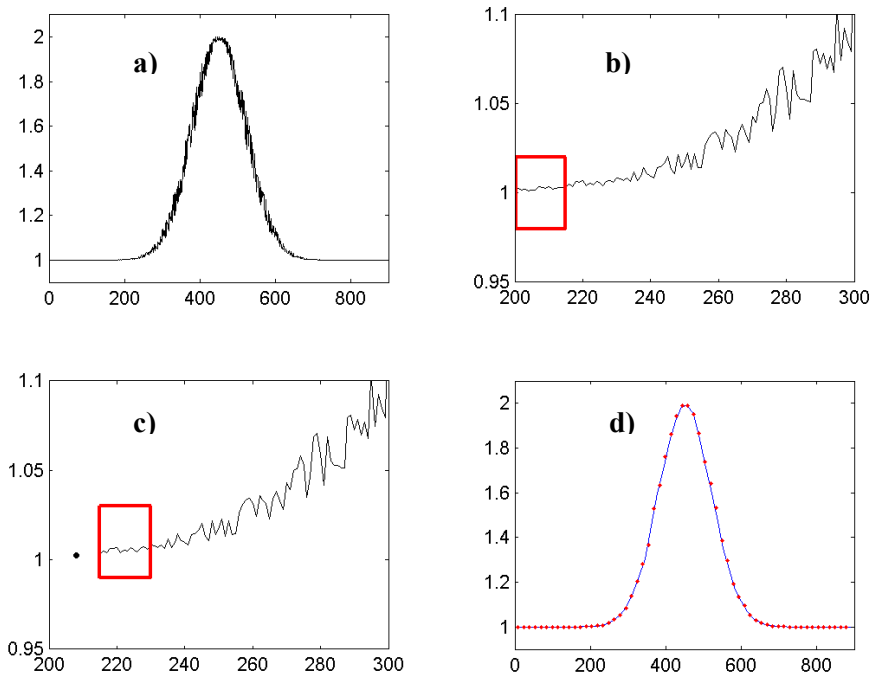

Figura 2.2 – Exemplo do procedimento de alisamento pela média: a) curva gaussiana onde ruído aleatório foi artificialmente adicionado; b) janela entre os pontos 200 e 215; c) deslocamento da janela para os pontos entre 216 e 230, e o ponto resultante do alisamento da região imediatamente anterior, e; d) espectro resultante alisado – os pontos vermelhos correspondem aos centróides das janelas utilizadas no processo de alisamento.

Note que o alisamento pela média também consiste num método de redução do número de variáveis. A curva em 2.2 (a) possuía originalmente 900 pontos, enquanto que a curva resultante em 2.2 (c) possui $900/15 = 60$ pontos.

Outro método para realizar o alisamento de curvas utiliza a transformada de Fourier para levar os espectros, que estão representados no domínio do tempo, para a dimensão das freqüências (equação 2.31). O resultado desta transformação matemática é o interferograma de cada espectro, que tem então as oscilações de alta freqüência eliminadas levando a zero valores dentro de janelas aplicadas em seus extremos. Após, os interferogramas alisados são novamente convertidos para o domínio do tempo através da aplicação da inversa da transformada de Fourier (equação 2.32), reconstruindo os espectros.

$$X(k) = \sum_{1}^{N} x(n) \mathrm{e}^{(-i \times 2\pi \times (k-1)(n-1))/N} \qquad 1 \le \mathrm{k} \le N \qquad \textbf{(2.31)}$$

$$x(n) = \sum_{1}^{N} X(k) \mathrm{e}^{(i \times 2\pi \times (k-1)(n-1))/N} \qquad 1 \le n \le N \qquad \textbf{(2.32)}$$

A figura 2.3 mostra a aplicação deste procedimento para uma curva gaussiana ruidosa:

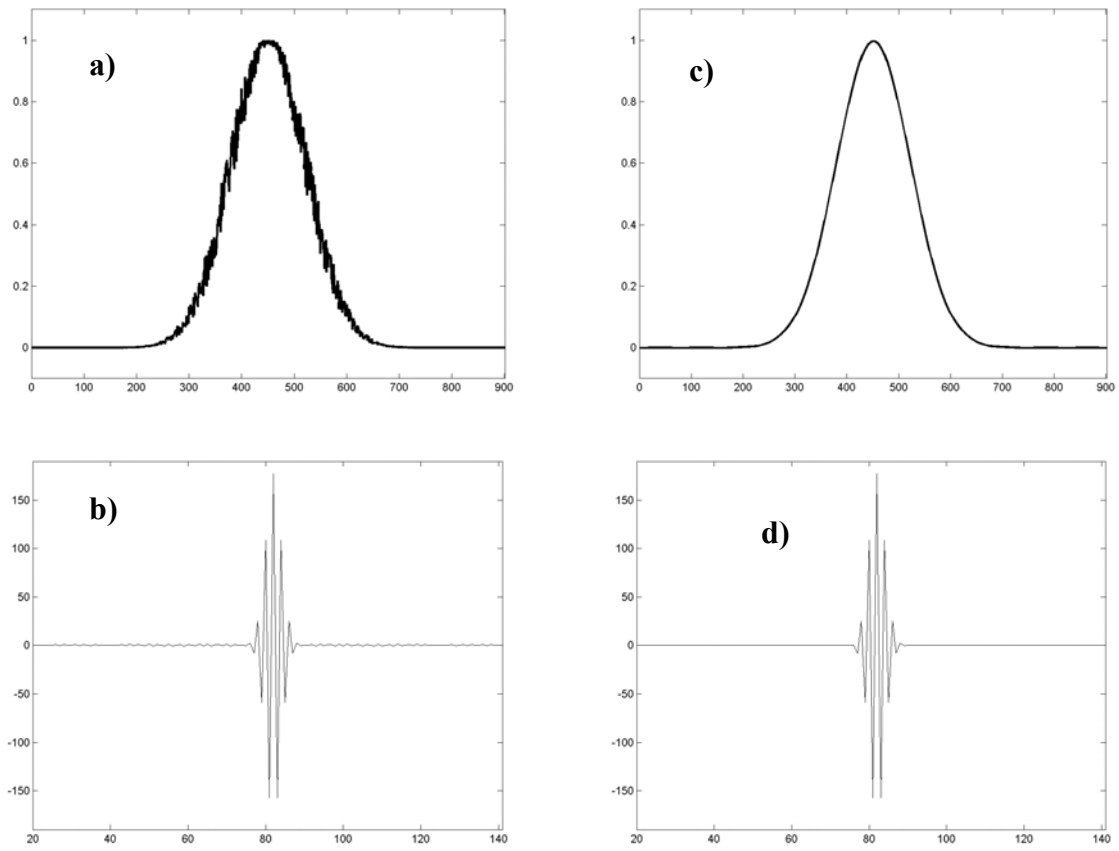

Figura 2.3 – Aplicação da transformada de Fourier para o alisamento de uma curva gaussiana contendo ruído: a) curva original; b) interferograma (parte real); c) espectro após alisamento e d) interferograma após remoção de elementos de alta freqüência, pela utilização de janela de 75 pontos, partindo dos extremos.

É usual ignorar a parte imaginária das transformadas de Fourier e realizar o procedimento apenas na parte real do interferograma, pois esta traria a maior parte da informação relevante do sistema. Como ver-se-á adiante, em alguns casos a parte imaginária pode trazer informação importante e, portanto, deve ser incluída no processo de alisamento.

2.5.3 – Correção Multiplicativa de Sinais (*Multiplicative Signal Correction* – MSC) [1, 3]

Espectros coletados utilizando-se técnicas de reflectância difusa são comumente sujeitos a desvios na linha de base (*offset*) devido a diferenças entre os padrões de espalhamento da luz da referência e da amostra analisada [25-28]. Este desvio é usualmente indesejado em calibração multivariada, podendo ser reduzido ou eliminado pela aplicação da correção multiplicativa de sinais. Neste método, todos os espectros são ortogonalmente projetados, utilizando as equações normais (2.22), em um espectro mediano, obtido a partir dos espectros originais.

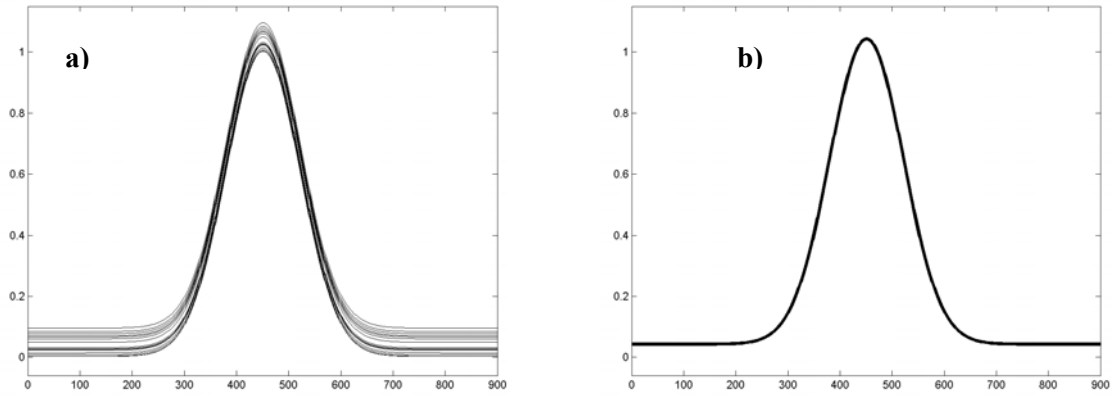

Figura 2.4 – Exemplo de aplicação da eliminação de *offset* pelo método MSC: a) curvas gaussianas apresentando desvios das linhas de base; b) eliminação do *offset* pela projeção das curvas originais na curva média.

É importante salientar que o método MSC não remove inclinações nas linhas de base (*bias*), pois o espectro médio resultante também estará sujeito a esta inclinação. A implementação deste algoritmo também é bastante simples:

1. Calcula-se um espectro médio, calculando-se as médias de cada coluna (comprimentos de onda) da matriz de espectros;
2. Insere-se um vetor cujos elementos são todos iguais a um como primeira coluna (este vetor acomodará o termo constante da regressão, *b*);
3. Por quadrados mínimos, calculam-se coeficientes de regressão *a* e *b*;

4. Os espectros corrigidos por MSC são obtidos subtraindo-se o termo constante *b* e dividindo o resultado pela inclinação *a* para cada comprimento de onda. O processo é repetido para os espectros de todas as amostras.

2.5.4 – Aplicação de Derivadas para Correções nas Linhas de Base [1, 3, 39]

Inclinação (*bias*) e desvios nas linhas de base (*offset*) também podem ser removidos pela aplicação de derivadas. As linhas de base das curvas que apresentam tais características podem ser aproximadas por um modelo linear do tipo:

$$A = a\lambda + b \tag{2.33}$$

em que a absorbância *A* é influenciada por uma inclinação *a* e por um termo constante *b*, que representa o desvio da linha de base. A derivada primeira da expressão acima fornece, então:

$$\frac{dA}{d\lambda} = a \tag{2.34}$$

e, deste modo, o termo constante *b* (*offset*) é removido.

Aplicando a derivada-segunda sobre a expressão 2.33 obtém-se também a remoção do termo *a*, que representa a inclinação da linha de base:

$$\frac{d^2A}{d\lambda^2} = 0 \tag{2.35}$$

A expressão 2.35 revela que, pela aplicação da derivada-segunda, uma completa regularização da linha de base é obtida.

Todavia, os gráficos resultantes das derivações não possuem mais a mesma forma dos espectros originais, guardando as características inerentes das derivadas. A figura 2.5 abaixo mostra a aplicação das derivadas primeira e segunda sobre curvas

gaussianas apresentando modelos de linha de base similares à equação 2.33. Nota-se, em (b), o termo constante na linha de base expresso pela equação 2.34; em (c), a linha de base coincide com o zero.

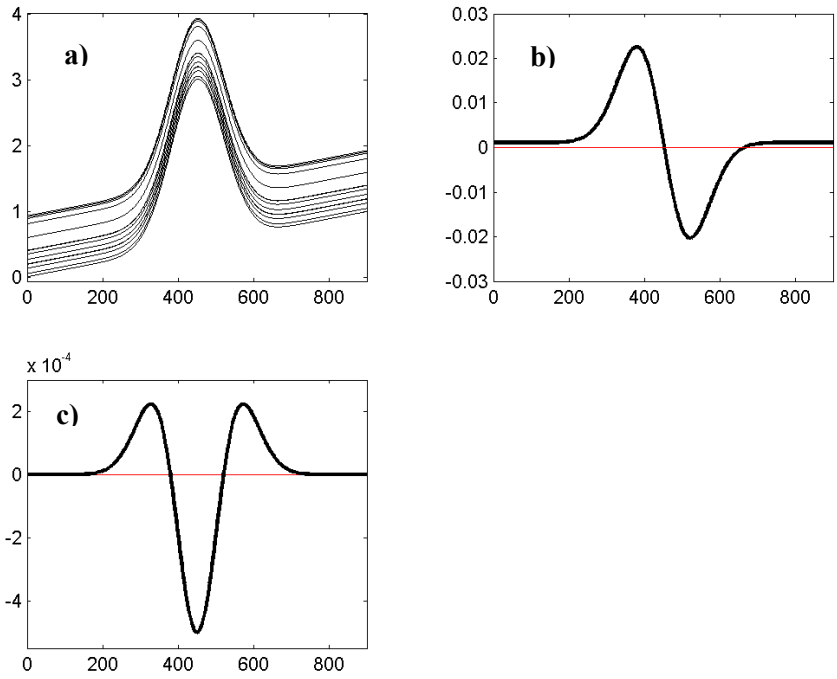

Figura 2.5 – Exemplo da aplicação de correção de linha de base pela aplicação de derivadas: a) curvas gaussianas apresentando inclinação (*bias*) e desvios de linha de base (*offset*), segundo modelo da equação 2.33; b) aplicação da derivada primeira, corrigindo o *offset* e c) aplicação da derivada-segunda, corrigindo *offset* e *bias*.

A implementação das técnicas de derivação de espectros é bastante simples. Os algoritmos mais empregados foram sugeridos por Savitzky e Golay [39], que utilizam intervalos de comprimento de onda ($\Delta\lambda$) constantes, de modo que:

$$\frac{dA}{d\lambda} \cong \frac{\Delta A}{\Delta \lambda} = A_{\lambda(1+\delta)} - A_{\lambda(1-\delta)} \qquad \textbf{(2.36)}$$

$$\frac{d^2 A}{d\lambda^2} = A_{\lambda(1-\delta)} + A_{\lambda(1+\delta)} - 2A_{\lambda(1)} \qquad \textbf{(2.37)}$$

Deve-se observar que os métodos de regularização de linha de base pela aplicação de derivadas apresentam a desvantagem de diminuir a razão sinal/ruído e assim, quando espectros apresentam regiões ruidosas, deve-se realizar um alisamento prévio eficiente. Quando este procedimento for impossível ou ineficaz, cabe ao pesquisador decidir pela aplicação ou não de derivadas.

2.6 – Métodos de Seleção de Variáveis

As técnicas de compressão de dados descritas anteriormente podem ter sua eficiência consideravelmente aumentada pela eliminação de variáveis irrelevantes do conjunto de calibração, incrementando também a performance dos procedimentos de cálculo. Esta melhora vem basicamente do fato de que a variância destes comprimentos de onda não-relevantes é eliminada dos *loadings* nos métodos de compressão de dados (vide seção 2.3.2). Existem diversos métodos de seleção de variáveis disponíveis atualmente, cada um possuindo características intrínsecas e sendo desenvolvido para um conjunto de aplicações específico. Neste trabalho, foram explorados os métodos de seleção pelo correlograma e pela divisão simétrica dos espectros, bem como o algoritmo das projeções sucessivas (*Successive Projections Algorithm* – SPA) e o *dimension-wise selection* (DWS) [1, 3, 32, 33, 39, 40].

2.6.1 – Seleção de Variáveis pelos Correlogramas [1, 4]

Correlogramas são gráficos que trazem os valores dos coeficientes de correlação entre cada comprimento de onda de um espectro e a propriedade de interesse. Deste modo, comprimentos de onda apresentando correlação significativa com a propriedade de interesse devem ser importantes para a calibração.

A primeira etapa do processo consiste no cálculo dos coeficientes de correlação entre as variáveis ou comprimentos de onda e a propriedade de interesse:

$$r = \frac{(x_j - \bar{x}_j)(y - \bar{y})}{\sqrt{s^2(x_j) \times \mathrm{s}^2(\mathrm{y})}} \qquad \textbf{(2.38)}$$

em que $s^2(x_j)$ e $s^2(y)$ são as variâncias do *j*-ésimo comprimento de onda e da propriedade de interesse, respectivamente.

O cálculo pode ser realizado de maneira mais eficiente, utilizando álgebra linear, seguindo os seguintes procedimentos:

1. A matriz **X** e o vetor **y** são primeiramente autoescalados (seção 2.4.1);
2. Insere-se o vetor **y** após a última coluna de **X,** gerando uma nova matriz **X'**;
3. Calcula-se $\mathbf{K} = \mathbf{X'}^{\mathrm{T}}\mathbf{X'}$, em que **K** é a matriz de correlação normalizada pelo número de graus de liberdade;
4. Calcula-se $\mathbf{C} = \mathbf{K}/(I - 1)$, em que $(I - 1)$ é o número de graus de liberdade;
5. Toma-se a última linha de **C** para construção do gráfico de correlação (correlograma), sendo fácil demonstrar que este procedimento resulta na equação 2.38.

Calculado o correlograma, escolhe-se um parâmetro de corte, *k*. A figura 2.6(a) mostra um exemplo de correlograma onde o parâmetro de corte escolhido foi 0,8. Nota-se que, como coeficientes de correlação podem assumir valores negativos, este parâmetro é tomado em módulo (neste exemplo, $|k| = 0{,}8$).

A figura 2.6(b) mostra as variáveis que seriam inclusas no modelo de calibração pela utilização deste *k* (áreas em cinza):

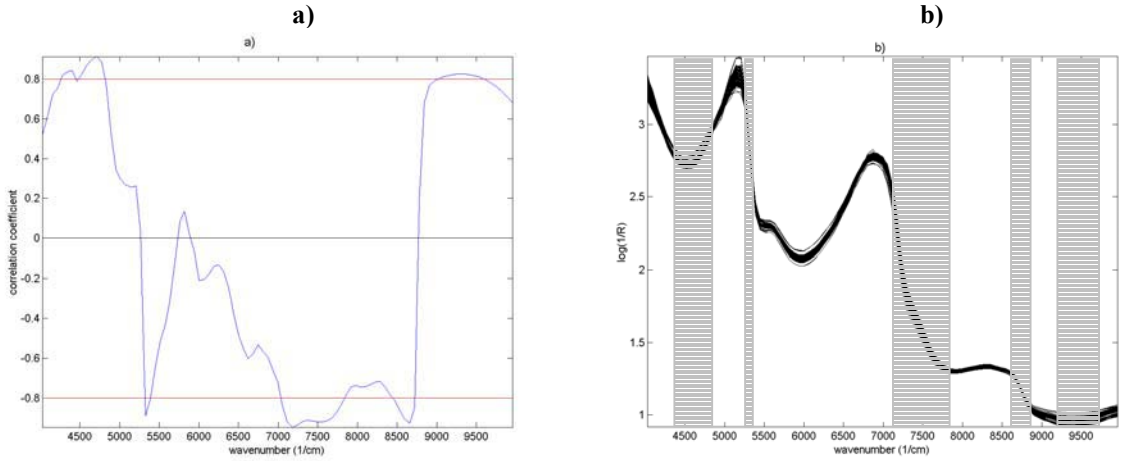

Figura 2.6 – Exemplo de procedimento de seleção de variáveis pelo correlograma: a) variáveis apresentando coeficientes de correlação menores que |0,8| (parâmetro de corte) são descartadas do modelo; assim, b) apenas os comprimentos de onda destacados em cinza são selecionados.

Selecionados os comprimentos de onda relativos a um dado parâmetro de corte, um modelo é construído e sua performance avaliada através do valor de RMSEP de validação externa (vide seção 2.6).

Os cálculos são repetidos para valores cada vez menores de parâmetros de corte e um gráfico de RMSEP em função de *k* é construído, conforme figura 2.7:

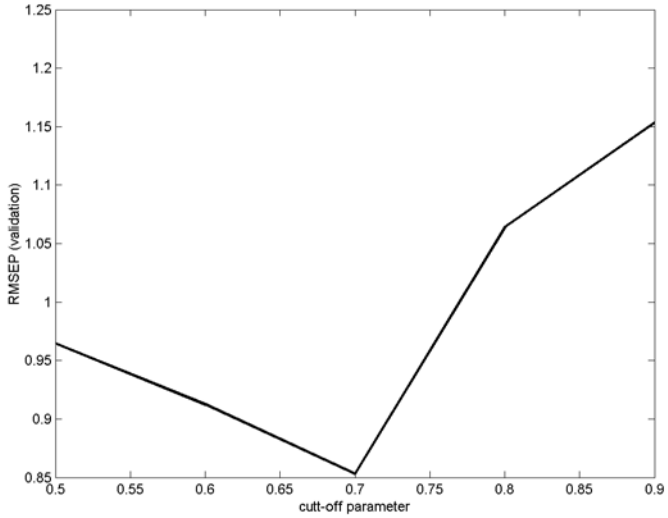

Figura 2.7 – Gráfico de RMSEP em função do parâmetro de corte. Neste exemplo, o melhor modelo é claramente aquele cujos comprimentos de onda incluídos no modelo correspondem a um parâmetro de corte do coeficiente de correlação $|k|$ = 0,7.

Nota-se que, quanto menor o valor de *k*, mais os modelos tenderão para aquele obtido utilizando-se o espectro inteiro e, deste modo, é usual utilizar o processo para valores não inferiores a 0,3.

Como os coeficientes de correlação são dependentes do pré-processamento previamente realizado nos espectros, esta técnica de seleção de variáveis deve ser repetida para cada um deles, individualmente, quando vários são testados.

2.6.2 – Seleção de Variáveis pela Divisão Simétrica dos Espectros [1, 18, 19]

Este procedimento de seleção de variáveis é bastante simples. Consiste em dividir os espectros originais em um número arbitrário de faixas de comprimentos de onda distintas, cada uma contendo características espectrais próprias. Um exemplo de divisão de espectros em três regiões simétricas é mostrado na figura 2.8 para um espectro de infravermelho próximo.

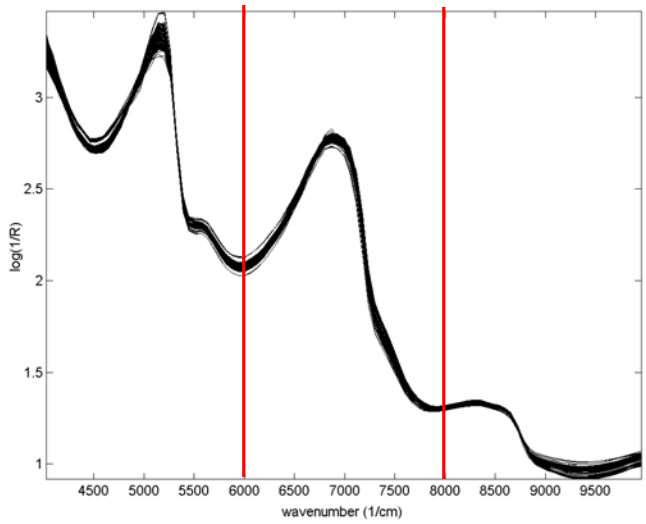

Figura 2.8 – Exemplo de procedimento de divisão simétrica de espectros em três regiões distintas: 4000 a 6000 cm^{-1}, 6000 a 8000 cm^{-1} e 8000 a 10000 cm^{-1}.

As calibrações são realizadas em cada uma das regiões individualmente, bem como em combinações destas, até que um modelo com o menor RMSEP (equação 2.40) e maior coeficiente de correlação de validação (equação 2.44) seja obtido.

2.6.3 – Algoritmo das Projeções Sucessivas (*Successive Projections Algorithm* – SPA) [40]

O SPA foi criado por M.C.U. Araújo e R.K.H. Galvão, do Instituto de Tecnologia Aeronáutica (ITA) do Brasil. Consiste num procedimento de busca sistemática de variáveis informativas na medida em que inicia-se com um comprimento de onda adequadamente escolhido e incorpora novas variáveis a cada iteração. A seleção é efetuada realizando-se projeções sucessivas de todos os comprimentos de onda no espaço ortogonal da variável inicial selecionada.

O procedimento é relativamente simples. Partindo-se de um comprimento de onda inicial ($k_{(0)}$) e conhecendo-se o número total de comprimentos de onda a serem extraídos (N):

1. Cria-se um conjunto **S** de comprimentos de onda que ainda não foram selecionados: $\mathbf{S} = \{j, 1 \leq j \leq J \text{ e } j \notin \{k_{(0)}, ..., k_{(n-1)}\}\}$.
2. Calcula-se a projeção de todos os vetores $\mathbf{x}_j$ em **S** no espaço ortogonal a $k_{(n-1)}$ como:

$$\mathbf{P}\mathbf{x}_j = \mathbf{x}_j - (\mathbf{x}_j{}^{\mathrm{T}}\mathbf{x}_{k(n-1)})\mathbf{x}_{k(n-1)}(\mathbf{x}_{k(n-1)}{}^{\mathrm{T}}\mathbf{x}_{k(n-1)})^{-1} \qquad \textbf{(2.39)}$$

em que **P** é o operador linear de projeção.

3. Um comprimento de onda $k_{(n)}$, cujo vetor possui maior projeção ortogonal a $k_{(n-1)}$, é selecionado: $k_{(n)} = \arg(\max(\|\mathbf{P}\mathbf{x}_j\|), j \in \mathbf{S}$.
4. O processo é repetido para $\mathbf{x}_j = \mathbf{P}\mathbf{x}_j$, até que $n = N$.

Como os próprios autores do algoritmo destacam "apesar do SPA ser muito parecido com o procedimento de ortogonalização de Gran-Schmidt, estes algoritmos possuem propósitos diferentes. O algoritmo de Gram-Schmidt manipula os dados para gerar um novo conjunto de vetores ortogonais que, geralmente, não possuem

significado físico. O SPA, ao contrário, não modifica o conjunto de vetores originais uma vez que as projeções são utilizadas apenas para propósitos de seleção. Assim, a relação entre as variáveis espectrais e os dados é preservada" [40].

A performance do algoritmo é dependente do número de vetores (*N*) e do comprimento de onda inicial ($k_{(0)}$) no qual as projeções são realizadas. Para determinar estes parâmetros os autores sugerem o seguinte procedimento:

1. Especifica-se um conjunto de misturas calibração e outro de validação;
2. Especifica-se um número mínimo (*Nmin*) e um número máximo (*Nmax*) de comprimentos de onda a selecionar, onde acredita-se que o número ótimo (*N**) encontrar-se-á;
   - Para *N* indo de *Nmin* a *Nmax:*
   - Para *inicial* indo de *j* até *J*:
3. Utilizando os passos de projeção descritos acima, selecionam-se *N* comprimentos de onda, começando por $k_{(0)} = inicial;$
4. Constrói-se um modelo por regressão linear múltipla (MLR, vide seção 2.3.2) com os comprimentos de onda selecionados;
5. O modelo é utilizado para prever as amostras do conjunto de validação e calcula-se o RMSEP;
6. O valor calculado de RMSEP é armazenado na variável ρ(*inicial*);
7. Seja r(*N*) = min[ρ(*inicial*)], *inicial* = 1,...,*J* e s(*N*) = arg[min(ρ(*inicial*)], *inicial* = 1, ..., *J*.
8. Os cálculos são repetidos para todos os *N* entre *Nmin* e *Nmax* e os valores ótimos *N** e $k^*_{(0)}$ são dados por *N** = arg(min(r(*N*)), *N* = *Nmin*, ..., *Nmax*) e $k^*_{(0)}$ = s(*N**).

O autor desta dissertação realizou uma pequena mas fundamental alteração neste procedimento. Porque os comprimentos de onda selecionados são *os mais ortogonais possíveis entre si*, mas não necessariamente formalmente ortogonais, o método de seleção descrito acima pode gerar matrizes mal-condicionadas e a inversa de *Moore-*

Penrose torna-se bastante instável (vide seção 2.2.3). Deste modo optou-se por construir modelos por MLR através da PCR, utilizando todas as componentes principais no cálculo (vide seção 2.3.2).

2.6.4 – *Dimension-Wise Selection* (DWS) [1, 41]

O procedimento de seleção de variáveis *dimension-wise selection* é aplicável principalmente aos modelos de calibração obtidos pelo método PLS, mas existem adaptações para outros métodos. Não se trata de um procedimento formal de seleção de variáveis, mas sim da remoção de comprimentos de onda irrelevantes do modelo. No método PLS isto é conseguido fazendo com que os elementos dos vetores de pesos (**w**) cujos valores sejam relativamente pequenos sejam igualados a zero no momento da extração dos *scores*.

A implementação computacional deste procedimento é bastante simples, bastando inserir o seguinte elemento de operação iterativa condicional no algoritmo PLS:

se $w_i <$ parâmetro de corte,
$w_i = 0$;
fim

Com este novo vetor **w,** que possui elementos truncados, calculam-se os *scores* conforme algoritmo PLS.

A etapa mais complexa deste método é a definição dos parâmetros de corte. Para defini-los, modelos de calibração são construídos a partir de *loadings* e *scores* obtidos com vetores **w** que foram submetidos a vários parâmetros de corte. Gráficos de RMSEP em função do parâmetro de corte são construídos para cada modelo e o melhor parâmetro é aquele que fornecer o menor valor dos erros de predição. O procedimento é repetido para cada componente principal individualmente.

A principal diferença entre este método e as outras técnicas de seleção aqui expostas é que a informação trazida pelas variáveis "eliminadas" é transferida para outros comprimentos de onda que contribuem mais efetivamente para a predição da

propriedade de interesse. Matematicamente esta operação pode ser definida como a rotação das componentes principais na direção de uma solução otimizada para a regressão por quadrados mínimos. A figura 2.9 mostra um exemplo da aplicação do DWS. Para efeito de ilustração, valores extremos de parâmetros de corte, resultando numa rotação significativa das componentes principais, foram propositadamente aplicados.

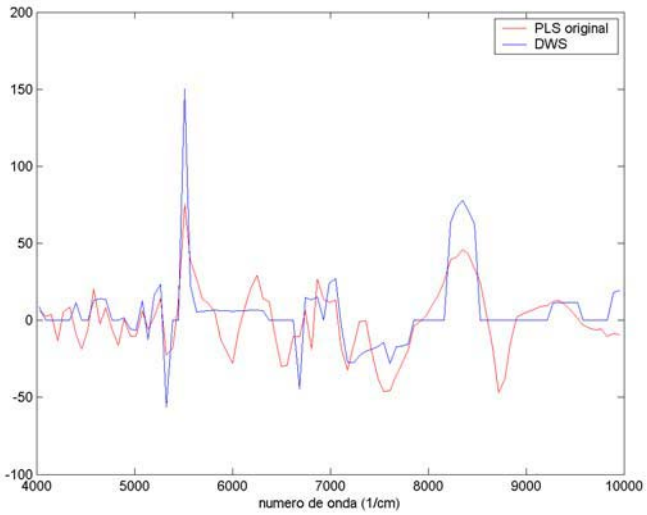

Figura 2.9 – Exemplo de aplicação do DWS. Note como os valores dos coeficientes de regressão aumentam para variáveis mais relevantes ao modelo (que possuem maiores pesos em **w**) enquanto que aquelas que menos contribuem têm seu peso diminuído.

2.7 – Validação dos Modelos de Calibração [1-3, 28-30]

Uma vez determinado o modelo de regressão é necessário que se valide sua capacidade de predição. Este procedimento é especialmente relevante para a verificação do nível de ajuste do modelo no momento da previsão de novas amostras. Um modelo de calibração pode estar tão bem ajustado aos dados de calibração que suas previsões são falhas. Este comportamento é chamado de *overfiting*. O reverso também pode ocorrer: quando o modelo de calibração não incluiu variância suficiente das amostras de calibração, suas previsões poderão ser falhas. Neste caso, tem-se o chamado *underfiting*.

Os principais métodos são o de validação externa, onde um número de amostras é removido do conjunto de calibração e utilizado para posterior predição, e o de validação cruzada, onde usualmente uma amostra (*leave one out*) é removida do conjunto de calibração de cada vez e *I* modelos de calibração construídos. A principal vantagem deste método é que torna-se possível determinar o número de variáveis latentes a manter no modelo, bem como definir se há alguma amostra anômala (*outlier*).

Pelo método de validação cruzada a *i*-ésima amostra é removida do conjunto de calibração e um modelo construído. O valor desta amostra é previsto pelo modelo e comparado com o valor encontrado pelo método de referência. Após as *I* amostras terem sido removidas, calcula-se o *Root Mean Square Error of Prediction* (RMSEP – Raiz Quadrada dos Erros Médios de Predição) [3, 4]:

$$RMSEP = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{I}} \qquad \textbf{(2.40)}$$

sendo que $\hat{y}_i$ é o valor estimado da concentração da *i*-ésima amostra e $PRESS = \sum (y_i - \hat{y}_i)^2$ é o *Prediction Error Sum of Squares* (PRESS – Somatório Quadrático dos Erros de Predição). Note que PRESS tem unidades de variância e RMSEP pode ser tomado como um desvio-padrão médio.

Outras figuras de mérito importantes para detecção de amostras anômalas nos modelos de calibração são a *leverage* e os resíduos de Student. A *leverage* pode ser entendida como a distância entre o centro de informação dos dados ($\bar{\mathbf{x}}, \bar{\mathbf{y}}$) e o ponto definido pela *i*-ésima amostra, e pode ser calculada como:

$$h_{ii} = \frac{1}{N} + (\mathbf{x}_i - \bar{\mathbf{x}})^{\mathbf{T}} (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \qquad \textbf{(2.41)}$$

Os resíduos de Student, por sua vez, são determinados pelas expressões:

$$Lresc_i = \sqrt{\frac{(yi - \hat{y}_i)^2}{(N-1)(1-h_{ii})}}$$

$$RS_i = \frac{(yi - \hat{y}_i)}{Lresc\sqrt{(1-h_{ii})}} \qquad \textbf{(2.42)}$$

sendo $Lresc_i$ o resíduo da concentração corrigido pela *leverage* e RS_i o resíduo de Student para a *i*-ésima amostra [4].

Amostras anômalas são trivialmente identificadas através de um gráfico de resíduos de Student pela *leverage*, pois apresentam valores elevados para ambos os parâmetros. Para facilitar a identificação destas amostras, um parâmetro de *leverage* crítico é calculado:

$$h_{crit} = \frac{\alpha \times A}{N} \qquad \textbf{(2.43)}$$

em que a definição de α varia de autor para autor, sendo mais usual $\alpha = 2$ ou $\alpha = 3$ [1, 2].

Amostras extremas podem apresentar *leverage* acima do valor crítico, desde que possuam resíduos de Student moderados. Este caso é comum quando um conjunto de amostras apresenta valores elevados ou diminutos para a propriedade de interesse. Neste caso, podem ser essenciais à calibração, não constituindo, todavia, amostras anômalas.

Detectadas e excluídas as amostras anômalas, constroem-se os modelos de calibração utilizando diferentes métodos matemáticos e com processos distintos de seleção de variáveis. Seu poder de predição é comparado utilizando o PRESS e o coeficiente de correlação de validação [7].

$$r = \frac{(y - \hat{y})^2}{\sqrt{s^2(y) \times s^2(\hat{y})}} \qquad \textbf{(2.44)}$$

Observa-se que 2.44 é apenas um caso específico de 2.38.

2.8 - Referências

1. Martens, H., Naes, T.; "Multivariate Calibration"; John Wiley & Sons; Chichester 1993.
2. Ferreira, M.M.C., *et alli*; "Quimiometria I – Calibração Multivariada, um Tutorial"; *Química Nova*; **1999**; *22*; 724-731.
3. Beebe, K.R., Pell, R.J., Seasholtz, M.B.; "Chemometrics: a Practical Guide"; John Wiley & Sons; New York 1998.
4. Box, G.E.P., Hunter, W.G., Hunter, J.S.; "Statistics for Experimenters"; John Wiley & Sons; New York 1978.
5. Derde, M.P., Massart, D.L.; "Multivariate Calibration by Data Compression"; *Anal. Chim. Acta*; **1986**; *191*; 1-16.
6. De Maesschalkd, *et alli*; "The Development of Calibration Models for Spectroscopic Data Using Principal Component Regression"; *Anal. Chem.*; **1972**; *44*; 72 - 84.
7. Strang, G.; "Introduction to Linear Algebra"; Wellesley-Cambridge Press; Wellesley 1993.
8. Strang G.; "Linear Algebra and its Applications"; 3rd Ed.; Harcourt Brace Jovanovich, Pub.; San Diego 1988.
9. Noble, B., Daniel, J.W.; "Applied Linear Algebra"; Prentice Hall; Englewood Cliffs 1988.
10. Golub, G., van Loan, C.F.; "Matrix Computations"; 2nd Ed.; Johns Hopkins Press; London 1989.
11. Pasquini, C.*;* "Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications"; *J. Braz. Chem. Soc.*; **2003**; *14*; 198-219.
12. Williams, P., Norris, K. (ed.); "NIR Technology in the Agricultural and Food Industries"; Am. Ass. Cereal Chemists; St. Paul; 1990.
13. Olinger, J.M., Griffiths, P.R.; "Effects of Sample Dilution and Particle Size/Morphology on Diffuse Reflectance Spectra of Carbohydrate Systems in the Near- and Mid-Infrared. Part I: Single Analytes; *Appl. Spectr.*; **1993**; *47*; 687-694.
14. Olinger, J.M., Griffiths, P.R.; "Effects of Sample Dilution and Particle Size/Morphology on Diffuse Reflectance Spectra of Carbohydrate Systems in the Near- and Mid-Infrared. Part II: Durum Wheat"; *Appl. Spectr.*; **1993**; *47*; 695-701.
15. Chen, J.Y., Iyo, C., Kawano, S.; "Development of Calibration with Sample Cell Compensation for Determination of Fat Content of Unhomogenised Raw Milk by a Simple Near Infrared Transmitance Method"; *J. Near Infrared Spectrosc.*; 1999; *7*; 265-273.

16. Lammertyn, J., *et alli*; “Non-Destructive Measurement of Acidity, Soluble Solids and Firmness of Jonagold Apples Using NIR Spectroscopy”; *Transac. ASAE*; **1998**; *41*; 1089-1094.

17. Jha, S.N., Matsuoka, T., Kawano, S.; “Nondestructive Techniques for Quality Evaluation of Intact Fruits and Vegetables”; *J. Near Infrared Spectrosc.*; **1995**; *3*; 211-218.

18. Kawano, S., Fujiwara, T., Iwamoto, M.; “Non-Destructive Determination of Sugar Content in Satsuma Mandarin Using Near Infrared Transmittance”; *J. Japanese Soc. Horticult. Sci.*; **1993**; *62*; 465-470.

19. Osborne, S.D., Kunnemyer, R. Jordan, R.B.; “A Low-Cost System for the Grading of Kiwifruit”; *J. Near Infrared Spectrosc.*; **1999**; *7*; 9-15.

20. Kawano, S., Watanabe, H., Iwamoto, M.; “Determination of Sugar Content in Intact Peaches by Near Infrared Spectroscopy with Fibre Optics in Iteractance Mode”; *J. Japanese Soc. Horticultural Sci.*; **1992**; *61*; 445-451.

21. Hong, T.L., Tsou, S.C.S.; “Determination of Tomato Quality by Near Infrared Spectroscopy”; *J. Near Infrared Spectrosc.*; **1998**; *6*; A321-A324.

22. HunterLab; “HunterLab PC2D Intructions Manual”; HunterLab, Inc.; Reston; 1980.

23. Goula, A.M., Adamopoulos, K.G.; “Estimating the Composition of Tomato Juice Products by Near Infrared Spectroscopy”; *J. Near Infrared Spectrosc.*; **2003**; *11*; 123-136.

24. Jha, S.N., Matsuoka, T.; “Non-Destructive Determination of Acid-Brix Ratio of Tomato Juice Using Near Infrared Spectroscopy”; *J. Food Sci. Tech.*; **2004**; *39*; 425-430.

25. Bjork, A; "Numerical Methods for Least Squares Problems"; SIAM; Philadelphia 1996

26. Hanselman, D., Littlefield, B.; “Matlab 6 – Curso Completo”; Prentice Hall; São Paulo 2003.

27. INRIA Meta 2 Project/ENPC Cergrene; “Introduction to Scilab”; INRIA; 2003.

28. Kowalki, B.R.; “Chemometrics: Mathematics and Statistics in Chemistry”; D. Reidel; Dordrecht 1984.

29. Geladi, P., MacDougall, D., Mantens, H.; “Linearization and Scatter Correction for NIR Reflectance Spectra of Meat”; *Appl. Spectrosc.*; **1985**; *39*; 491 – 500.

30. Brereton, R.G.; “Introduction to Multivariate Calibration in Analytical Chemistry”; *Analyst*; **2000**; *125*; 2125-2154.

31. Wold, S.; “Principal Component Analysis”; *Chemom. Intel. Lab. Syst.*; **1987**; *2*; 37-52.

32. Oliveira, S.C.C., *et alli*; “A Escolha da Faixa Espectral no Uso Combinado de Métodos Espectroscópicos e Quimiométricos”; *Quim. Nova*; **2004**; *27*; 219-225.

33. Höskuldsson, A.; "Variable and Subset Selection in PLS Regression"; *Chemom. Intel. Lab. Syst.*; **2001**; *55*; 23-38.

34. Malinowski, E.R.; "Factor Analysis in Chemistry"; John Wiley & Sons; New York; 1991.

35. Wold, H.; "Mulrivariate Analysis"; Academic Press; New York 1966.

36. Robusté, J.R., *et alli*; "Looking for the Best Model with PLS"; *Analusis*; 1993; *21*; 299-304.

37. Bro, R., Smild, A.K.; "Centering and Scaling in Component Analysis"; *J. Chemom.*; **2003**; *17*; 16-33.

38. Guidorizzi, H.L.; "Um Curso de Cálculo"; vol. 3; 2ª Ed.; LTC; São Paulo; 2001.

39. Savistsky, A., Golay, M.; "Smoothing and Differentiation of Data by Simplified Least Squares Procedure"; *Anal. Chem*.; **1964**; *36*; 1627.

40. Araújo, M.C.U, *et alli*; "The Successive Projections Algorithm for Variable Selection in Spectroscopic Multicomponent Analysis"; *Chemom. Intel. Lab. Syst.*; **2001**; *57*; 65-73.

41. Lindgren, F., *et alli*; "Interective Variable Selection (IVS) for PLS. Part 1: Theory and Algorithms"; *J. Chemom.*; **1994**; *8*; 349-363.

CAPÍTULO III

ANÁLISE EXPLORATÓRIA DE DADOS FÍSICO-QUÍMICOS

DESENVOLVIMENTO DE MODELOS DE CALIBRAÇÃO PARA SÓLIDOS E CAROTENÓIDES EM CONCENTRADOS DE TOMATE

3.1 - Introdução

Enquanto sólidos de tomate são parâmetros vitais no recebimento, controle de processo e formulação de produtos secundários, os carotenóides vêm chamando a atenção, tanto da indústria alimentícia quanto do mercado consumidor, devido aos possíveis benefícios à saúde que podem proporcionar (vide capítulo I). Apesar de haver ainda grande discussão sobre a efetividade destes benefícios, as indústrias buscam desenvolver alimentos funcionais ricos em licopeno e β-caroteno e a indústria farmacêutica já adiciona estes carotenóides em complementos alimentares.

A maior parte dos programas de variedades desenvolvidos pela indústria de alimentos busca produzir tomates com características que melhorem sua performance durante o processamento fabril ou forneçam produtos com aplicações culinárias diferenciadas. Dentre tais características destacam-se os frutos com maior teor de sólidos, principalmente solúveis, cores variadas, balanço ácido/brix otimizado ou maior resistência mecânica - pela produção de elevados teores de pectinas ou através da supressão da expressão de enzimas -, etc. Entretanto, uma segunda classe de estudos de novas variedades visa a utilizar o fruto como fonte de nutrientes na manufatura de alimentos funcionais. Neste campo existem ainda duas frentes de estudos: uma que visa aumentar a quantidade de nutrientes já naturalmente produzidos pelo tomate e uma outra, mais desafiadora, que tem por finalidade promover a biossíntese de nutrientes que atualmente não são entregues pelo fruto [1-3].

Atualmente, as técnicas de engenharia genética podem gerar, com um nível espantoso de precisão, híbridos apresentando características desejáveis diversas em um fruto. De fato, a variedade de tomate *Flavr-Savr* tem a produção de enzimas pécticas suprimidas. A precisão das técnicas de manipulação do DNA, apesar de diminuir o número e quantidade de plantas nos cultivares para estudos de novas variedades, não reduz a necessidade das determinações analíticas clássicas para comprovação da expressão da propriedade de interesse. Essas determinações são, em sua grande maioria, realizadas pelos métodos clássicos de análise. Deste modo, métodos de análise rápidos e não-destrutivos são bem-vindos mesmo dentro de estudos envolvendo técnicas de manipulação genética [1].

Neste trabalho, coletaram-se 15 propriedades físico-químicas de produtos concentrados de tomate, com as quais realizou-se uma análise de componentes principais (PCA) para que padrões de performance destes produtos fossem identificados. Quatro propriedades - sólidos totais e solúveis, licopeno e β-caroteno - foram eleitas para construção de modelos de calibração, principalmente devido à relevância de sua aplicação no monitoramento de processo fabril e/ou em programas de desenvolvimento de novas variedades de tomates.

3.2 – Experimental

Quarenta e duas amostras de produtos concentrados de tomate com teor de sólidos variando entre 6,9 a 35,9% (6,8 a 31,1 °Brix, respectivamente) foram adquiridas em vários mercados no Brasil, Argentina, Estados Unidos e Europa (Holanda, Itália e Grécia). Estas amostras foram classificadas da seguinte maneira:

- Simples-concentrado (17,0 a 22,0 °Brix de tomate);
- Duplo-concentrado (25,0 a 31,5 °Brix de tomate);
- Base de tomate para molhos (12,5 a 17,0 °Brix de tomate);
- Matéria-prima polpa de tomate (25,0 a 31,1 °Brix de tomate);
- Purê de tomate (6,9 a 12,0 ºBrix de tomate).

3.2.1 – Determinações Físico-Químicas e Instrumentais

Sólidos totais, em porcentagem, foram determinados em triplicata por secagem em estufa Fanem EV8, a 70°C, sob pressão reduzida (150mmHg) pela utilização de uma bomba de vácuo Edwards modelo E2M8. Sólidos solúveis (°Brix) foram determinados em duplicata utilizando refratômetro de bancada digital Abbe (American Optical, Inc.) com correção de temperatura pela utilização de um banho termostático com circulação de água Lauda RM6. Todas as leituras foram realizadas em temperatura ambiente (23-25°C) após filtração simples através de algodão hidrofílico [4, 5].

O pH foi medido em pHmetro digital Micronal modelo B474 com correção de temperatura, equipado com eletrodo combinado de vidro Metrohm (ref 6.0232.100).

Acidez total foi determinada em porcentagem de ácido cítrico (vide capítulo I) utilizando titulador automático Metrohm modelo 702 equipado com unidade intercambiável 806 e agitador magnético modelo 728. A solução titulante utilizada foi NaOH 0,1 mol L^{-1} (Merck 1.09141.1000) e a detecção do ponto de equivalência deu-se potenciometricamente no modo MET (*Monotonic Equivalence Point*), utilizando eletrodo combinado de vidro para pH Metrohm (ref 6.0232.100) [6].

Sal foi determinado na forma de cloreto de sódio (NaCl) pelo método de Mohr. Utilizou-se o mesmo titulador automático Metrohm modelo 702, substituindo apenas a unidade intercambiável, que continha solução titulante de $AgNO_3$ 0,1 mol L^{-1} (Merck 1.07286.1000). O ponto de equivalência foi determinado potenciometricamente no modo MET, utilizando a diferença de potencial detectada por eletrodo combinado íon-seletivo de prata Metrohm (ref 6.0450.100) [5 - 7].

A consistência Bostwick é uma medida rápida da viscosidade do produto, determinada pelo escoamento de uma quantidade estabelecida (50 mL) de produto através de uma calha em aço inoxidável de dimensões 30 x 5 x 3 cm, possuindo um fundo graduado em centímetros. Esta calha possui um sistema de comporta que separa o compartimento onde a amostra é colocada inicialmente daquele onde o escoamento se processará (figura 3.1) [4, 5].

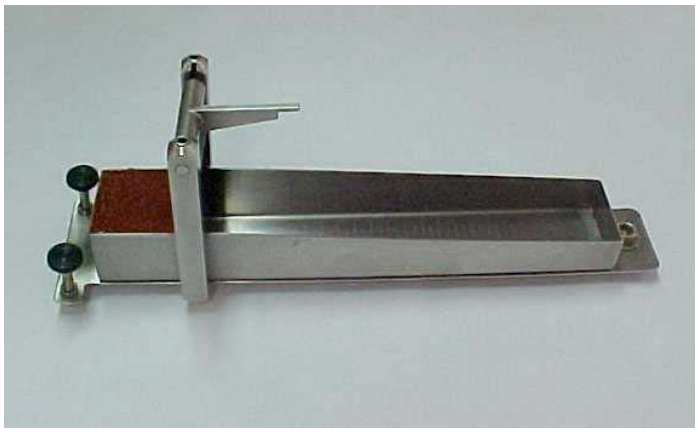

Figura 3.1 – Consistômetro Bostwick.

O produto é inicialmente diluído a 12 ºBrix de tomate, uma medida também conhecida como NTSS (*Natural Tomato Soluble Solids*, Sólidos Solúveis Naturais de Tomate) [4]. Após nivelamento do instrumento na bancada, uma quantidade suficiente do material diluído é adicionada ao compartimento da amostra, mantendo a comporta fechada. A comporta é finalmente aberta e o material é deixado escoar sobre a calha por 30 segundos, tempo após o qual a leitura é diretamente efetuada na escala graduada. Como forma-se um perfil de escoamento laminar [8] a leitura é efetuada no máximo do perfil de escoamento, que deve estar centrado na calha graduada caso o equipamento seja bem nivelado [4, 5].

O autor desta dissertação realizou uma pequena alteração no método original. Nota-se que, como os produtos são padronizados para uma quantidade de tomate, a medida descrita acima é capaz de revelar apenas diferenças em termos da qualidade da matéria-prima porque todas as amostras possuem a mesma quantidade de tomate. Por outro lado, não revela diferenças relativas à formulação dos produtos, parâmetro que afeta principalmente sua performance durante a aplicação culinária. Assim, a medida de consistência Bostwick também foi determinada com os produtos diluídos a 12 ºBrix propriamente ditos. Apesar destas medidas apresentarem-se algo correlacionadas, diferenças significativas são observadas em produtos onde quantidades relativamente elevadas de açúcar invertido foram adicionadas. Neste estudo, ambas as determinações, a convencional e a modificada, foram efetuadas.

As medidas de cor foram efetuadas em colorímetro *tri-stimulus* Hunter Lab.® modelo PC2Δ com iluminante C e geometria iluminante/observador 45º/2º. Utilizou-se escala Hunter L, a, b e o TPS (*Tomato Paste Score*) foi calculado pelo instrumento como [9]:

$$TPS = -46{,}383 + 1{,}0211a + 10{,}607b - 0{,}42198b^2 \qquad \textbf{(3.1)}$$

O *Blotter Test*, também conhecido como "teste do mata-borrão", é uma medida da capacidade de retenção de água (ou, contrariamente, do dessoramento) de produtos de tomate. Esta determinação está relacionada com a quantidade e qualidade do material péctico presente e é um parâmetro extremamente importante do ponto de vista de performance culinária. Para as determinações, utilizaram-se cartões de papel absorvente (Bridge & Co., Londres) de gramatura e espessura rigorosamente controladas, contendo impressos tanto um círculo de raio 1,9 cm, quanto quatro escalas milimetradas em posições ortogonal ou oposta umas às outras (figura 3.2).

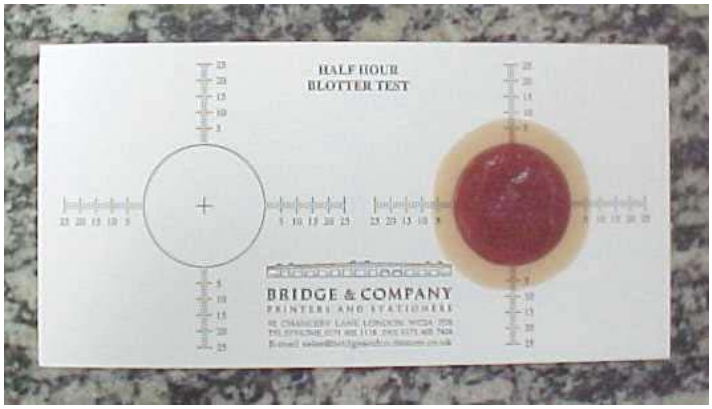

Figura 3.2 – Cartão de *Blotter Test* para determinação da capacidade de retenção de água em produtos de tomate.

As amostras foram diluídas a 12 NTSS e 7 mL adicionados uniformemente à área circular demarcada. O material foi mantido em contato com o papel absorvente por 30 minutos e, após, a leitura do deslocamento da água através do papel foi realizada nas escalas graduadas. O resultado foi reportado como a média, em

milímetros, do escoamento nas quatro escalas.

Açúcares (frutose, glicose e sacarose) foram determinados por cromatografia líquida utilizando HPLC Shimatzu equipado com um forno para coluna CTO-10A, um injetor automático Sil-10A, bombas de pistão SPD-10AV e um detector de índice de refração modelo RID-6A. Utilizou-se coluna cromatográfica Shodex NH_2P-50 4E (5μm, 25 x 0,46 cm) com guarda-coluna Shodex NH_2P-50G (4,6 x 10 mm), fase móvel acetonitrila:água 75:25, isocrático a 1 mL min^{-1}. Os tempos de retenção obtidos foram 6,9 minutos para frutose, 8,7 minutos para glicose e 11,5 minutos para sacarose. A quantificação foi realizada pela construção de curvas de calibração utilizando padrões Sigma-Aldrich (S8501 para sacarose, D9434 para glicose e F9048 para frutose) previamente secos em estufa a 70 ºC, sob pressão reduzida de 150 mmHg durante 6 horas.

Pesaram-se 5 gramas da amostra, em balança analítica, em béquer de 100 mL. Transferiu-se o material quantitativamente para um balão volumétrico de 100 mL, utilizando aproximadamente 50 mL de água destilada. Após homogeneização, adicionaram-se 10 mL de solução a 0,25 mol L^{-1} de ferrocianeto de potássio [$K_4Fe(CN)_6.3H_2O$] e 10 mL de solução a 1,0 mol L^{-1} de acetato de zinco [$(H_3CCOO)_2Zn.2H_2O$] para clarificação. Agitou-se o balão vigorosamente, completando o volume com água destilada em seguida. Após homogeneização, o conteúdo do balão foi filtrado através de papel de filtro Wattmann 41 para um erlenmeyer de 250 mL. Uma alíquota foi transferida para *vial* de 2 mL com septo de silicone e 5 μL foram injetados.

Licopeno e β-caroteno foram determinados utilizando-se o mesmo sistema cromatográfico empregado nas determinações de açúcares mas, neste caso, um detetor UV-VIS Shimatzu SPD-10AV com comprimento de onda ajustado para 473 nm foi empregado. A separação foi realizada em coluna Zorbax RP18 ODS (5 μm, 15 x 0,46 cm). A fase móvel empregada foi MetOH:THF:H_2O (67:27:6), isocrático a 1,0 mL min^{-1}. Tempos de retenção típicos foram 17,5 minutos para licopeno e 20,8 minutos para β-caroteno. Padrões Sigma-Aldrich foram empregados na construção de curvas de

calibração de licopeno (C0251) e β-caroteno (L9879).

Para a extração dos carotenóides seguiu-se o procedimento de extração sugerido por Sadler *et alli* [10], onde a etapa de saponificação não é executada [11]. Assim, pesaram-se 5 g da amostra em erlenmeyer de 250 mL, adicionando-se em seguida 100 mL de uma mistura de solventes contendo hexano:acetona:etanol 50:25:25. O erlenmeyer foi mantido ao abrigo da luz pela utilização de uma folha de papel alumínio. A mistura de solventes foi deixada em contato com a amostra, sob agitação vigorosa de um agitador magnético, durante 10 minutos. Após, 15 mL de água destilada foram adicionados e a agitação mantida por mais 5 minutos. Após, a agitação é desligada e espera-se pela separação das fases orgânica (50 mL) e aquosa (65 mL). Uma alíquota da fase orgânica foi imediatamente coletada em um *vial* de 2 mL com septo de silicone e 5 μL injetados para análise. Uma pequena fração da forma *cis* foi detectada para licopeno e β-caroteno. As áreas foram somadas com as das formas *all trans*, e os resultados são aqui expressos como licopeno ou β-caroteno total.

3.2.2 - Aquisição dos Espectros de Infravermelho

Os espectros de infravermelho foram coletados imediatamente após a abertura das amostras. Uma alíquota adequada da amostra foi adicionada à parte do fundo de uma placa de Petri de vidro marca Schott (23 755 48 05). A placa foi inserida no acessório de reflectância difusa modelo MSC-100 do espectrômetro FT-NIR marca Büchi NIRLab N-200. O acessório MSC-100 possui sistema de rotação de amostra, assim evitando variações na leitura devido a diferenças na espessura da placa de Petri, bem como eliminando problemas de aquecimento local da amostra.

Três espectros foram coletados para cada amostra, na região entre 4000 e 10000 cm^{-1} com resolução espectral de 4 cm^{-1}. A temperatura das amostras não foi termostaticamente controlada durante a aquisição dos espectros pois a temperatura ambiente do laboratório variou entre 22 e 24 ºC.

3.2.3 – Técnicas Quimiométricas de Análise Exploratória de Dados e Calibração Multivariada

Para as análises exploratórias por PCA a matriz de propriedades físico-químicas foi pré-processada por auto-escalamento. A análise de PCA foi realizada através de rotina baseada no algoritmo da decomposição SVD de matrizes.

Para as calibrações, os espectros de infravermelho apresentaram ruído considerável na região entre 4000 e 7300 cm^{-1} e, assim, técncias de alisamento foram empregadas. Utilizou-se alisamento pela média com janela de 15 pontos e alisamento pela transformada de Fourier. Este último método não apresentou-se mais eficiente que o primeiro, mais simples.

Calibrações pelos métodos PCR e PLS foram realizadas nestes espectros alisados. O procedimento de centrar na média foi empregado, tanto aos espectros quanto às propriedades de interesse, antes da construção dos modelos de calibração. As amostras foram divididas em dois grupos distintos, um de calibração, contendo 111 espectros (37 amostras) e outro de validação, contendo 15 espectros (5 amostras). Invariavelmente, o método PLS apresentou modelos com melhor capacidade preditiva e com menor número de fatores que o PCR.

O número ótimo de fatores a manter nos modelos de calibração foi determinado utilizando o método de validação cruzada, enquanto que sua performance preditiva foi avaliada por validação externa. Quaisquer amostras anômalas identificadas pelos gráficos de *leverage vs.* Resíduos de Student foram avaliadas e eliminadas dos modelos finais de calibração, quando pertinente.

Para remoção do deslocamento (*offset*) e inclinação (*bias*) das linhas de base dos espectros realizaram-se, após alisamento, pré-processamentos pela correção multiplicativa de sinais (MSC) e pelas derivadas primeira e segunda de acordo com algoritmo de Savitzky e Golay. Modelos obtidos com espectros pré-processados por derivadas apresentaram capacidade preditiva inferior àqueles obtidos com espectros tratados por alisamento ou MSC, principalmente pela redução da razão sinal/ruído em regiões importantes do espectro (vide figura 3.13).

Novas calibrações foram realizadas onde os extremos dos espectros de

infravermelho foram desconsiderados, sendo utilizada, então, apenas a região compreendida entre 4500 a 9500 cm^{-1}. Este procedimento teve por objetivo avaliar o efeito dessas regiões, sujeitas a oscilações instrumentais, nos modelos de calibração.

Seleções de variáveis foram executadas pelos métodos de fracionamento simétrico de espectros, pelo correlograma e pelos algoritmos DWS e SPA, e modelos PLS construídos. As habilidades preditivas destes modelos foram comparadas através do RMSEP e do coeficiente de correlação de validação externa (r_{val}) com aqueles obtidos sem seleção de variáveis.

Todos os cálculos foram realizados utilizando-se o software Matlab versão 6.1 (The Mathworks, Inc.) com rotinas de pré-processamento, seleção de variáveis, regressão multivariada, validação e gráficas desenvolvidas pelo autor. Os resultados destas rotinas foram comparados com aqueles obtidos pelo software quimiométrico NIRCal versão 4.2 [12] fornecido juntamente com o equipamento e resultados semelhantes foram obtidos. Aqui, apenas os resultados conseguidos com o software Matlab serão apresentados.

3.3 – Análise por Componentes Principais (PCA) na Identificação de Padrões de Comportamento das Amostras

A tabela com todos os resultados das análises físico-químicas consta no anexo I. Nota-se que é extremamente complexo encontrar padrões de comportamento das amostras pela observação direta dos valores nela constantes. A Análise por Componentes Principais (PCA), por outro lado facilita o reconhecimento humano de padrões pela compressão de dados, conforme exposto no capítulo 2.

Uma forma visual rápida de avaliar como cada componente principal descreve a informação contida no conjunto original de dados é obtida construindo-se um gráfico da variância trazida por cada uma delas juntamente com a variância acumulada, conforme figura 3.3:

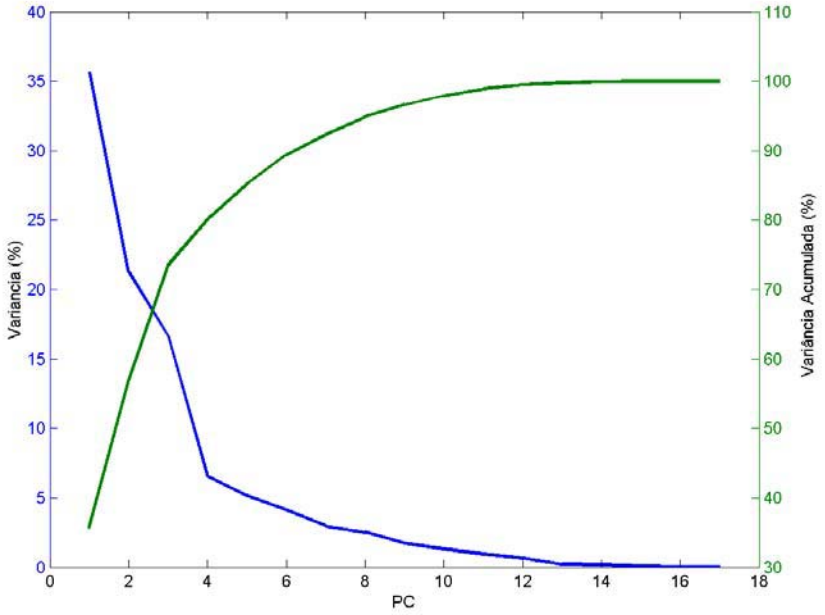

Figura 3.3 – Gráfico da variância explicada *vs.* número de componentes principais incluídas no modelo para a PCA.

A seleção gráfica do número de componentes principais a manter no modelo pode ser realizada também observando onde a inclinação do gráfico de variância sofre uma inflexão brusca. No gráfico da figura 3.3 observa-se que a variância cai acentuadamente até a quarta componente principal, tendendo a estabilizar. Neste ponto, aproximadamente 80% da informação dos dados originais é explicada, mas esta PC traz apenas 6% da variância total dos dados originais.

Também pode-se observar, através da tabela 3.1 abaixo, como os autovalores, a variância e variância acumulada comportam-se em função de cada componente principal:

Tabela 3.1 – Autovalores e variâncias para cada componente principal da PCA.

PC	Autovalores	Variância (%)	Variância Acumulada (%)
1	248,4926	35,6517	35,6517
2	148,0524	21,2414	56,8931
3	115,9617	16,6373	73,5304
4	45,6476	6,5492	80,0795
5	35,9382	5,1561	85,2356
6	28,8470	4,1387	89,3744
7	20,6876	2,9681	92,3425
8	17,5876	2,5233	94,8658
9	12,0489	1,7287	96,5945
10	9,2548	1,3278	97,9223
11	6,6210	0,9499	98,8722
12	4,4738	0,6419	99,5141
13	1,6101	0,2310	99,7451
14	1,1849	0,1700	99,9151
15	0,3899	0,0559	99,9710
16	0,1368	0,0196	99,9907
17	0,0651	0,0093	100,0000

Três componentes principais trazem aproximadamente 73% da informação constante nos dados originais, uma quantidade satisfatória quando se leva em consideração o fato de que os dados são oriundos de determinações com naturezas analíticas diferentes, com desvios relativos de até 15%.

Uma vez determinado o número de componentes principais a manter no modelo construíram-se gráficos de *scores* (figura 3.4) para avaliar a relação entre as amostras de produtos de tomate no novo espaço gerado por estas componentes principais. Nota-se que, de fato, é possível observar um padrão de distribuição das amostras em função das classes utilizando apenas as componentes 1 a 3. A quarta componente principal foi inserida apenas a título de verificação de que não traz informação relevante para a classificação das amostras.

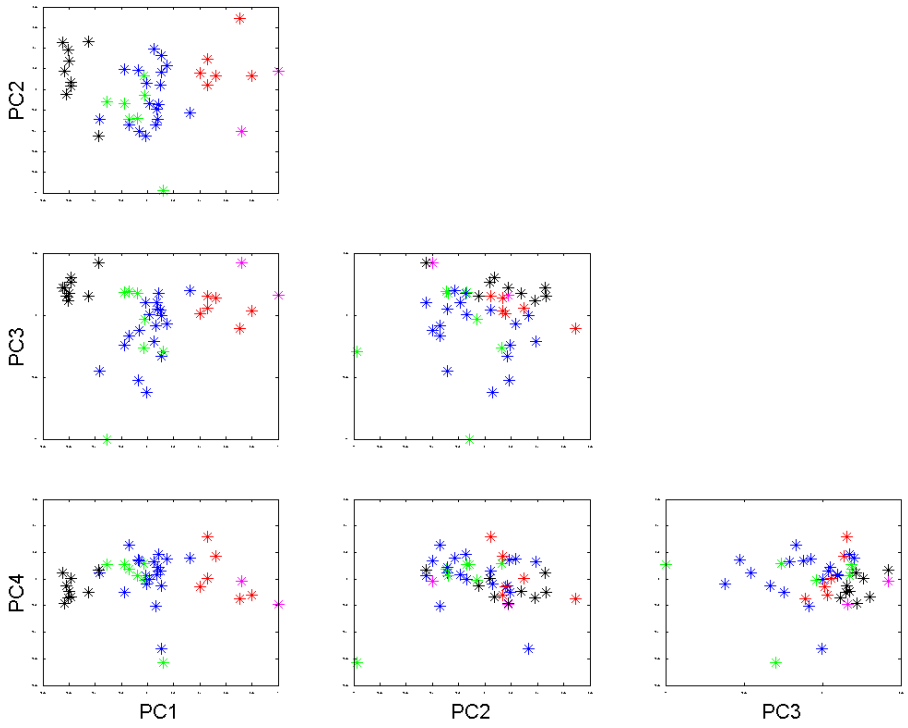

Figura 3.4 – Gráfico de *scores* para as quatro primeiras componentes principais da PCA. As cores dos pontos seguem a seguinte classificação: preto = purê de tomate, verde = base de tomate, azul = simples concentrado, vermelho = duplo-concentrado e magenta = polpa de tomate.

Como os produtos estão classificados basicamente em função da quantidade de sólidos de tomate que possuem, é possível observar que a primeira componente principal descreve basicamente esta grandeza e outras variáveis a ela relacionadas.

Um gráfico em três dimensões das três primeiras componentes principais é mostrado na figura 3.5.

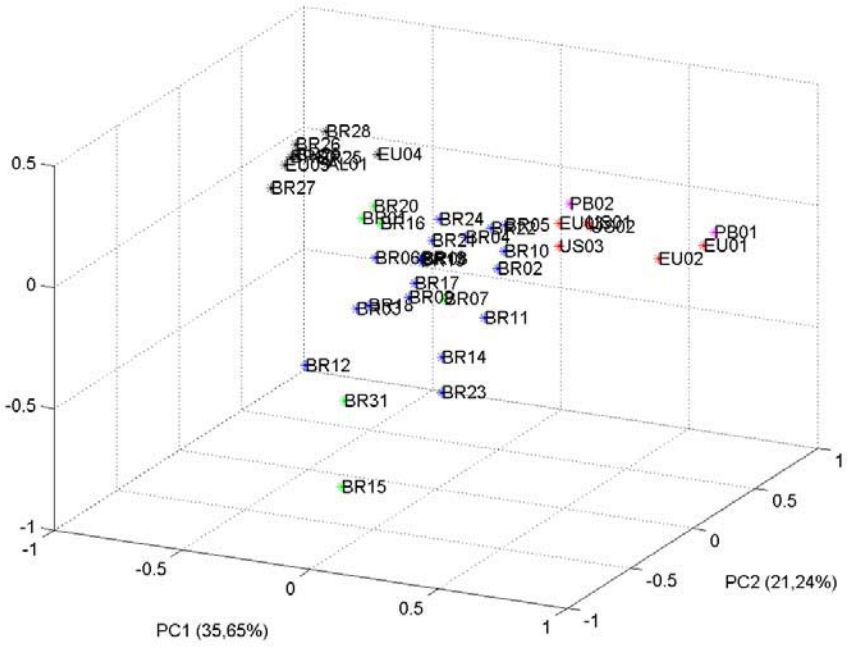

Figura 3.5 - Gráfico de *scores* para a PCA construída com as determinações físico-químicas como variáveis.

Nota-se que as amostras de purê de tomate (preto) agrupam-se no canto superior esquerdo do gráfico. Por outro lado, as amostras de base de tomate (verde) e simples-concentrados (azul) estão bastante próximas entre si. Duplo-concentrados (vermelho) e polpas de tomate (magenta) agrupam-se num terceiro bloco, à direita da primeira componente principal. De fato, duplo-concentrados têm basicamente a mesma quantidade de tomate que as matérias-primas que os originaram, sendo apenas acrescidos de sal e/ou açúcar.

Também é possível observar que as amostras de diferentes países agrupam-se dentro de suas respectivas classes, não havendo evidência de agrupamento em função de sua origem. Este comportamento era um tanto esperado, uma vez que a maioria dos fabricantes de produtos de tomate no mundo utiliza-se basicamente das mesmas variedades. Talvez análises de minerais fossem eficazes em classificar a origem das amostras, mas estas determinações estavam fora do escopo deste trabalho.

Para entender a distribuição das amostras no espaço dos *scores* geraram-se gráficos de *loadings*. Em química de alimentos é comum levar ambos, *scores* e *loadings*, para uma escala comum e construir um gráfico único que mostra, simultaneamente, o espaço das amostras e das variáveis que gerou este espaço. Gráficos como este são denominados *biplots* (figura 3.6).

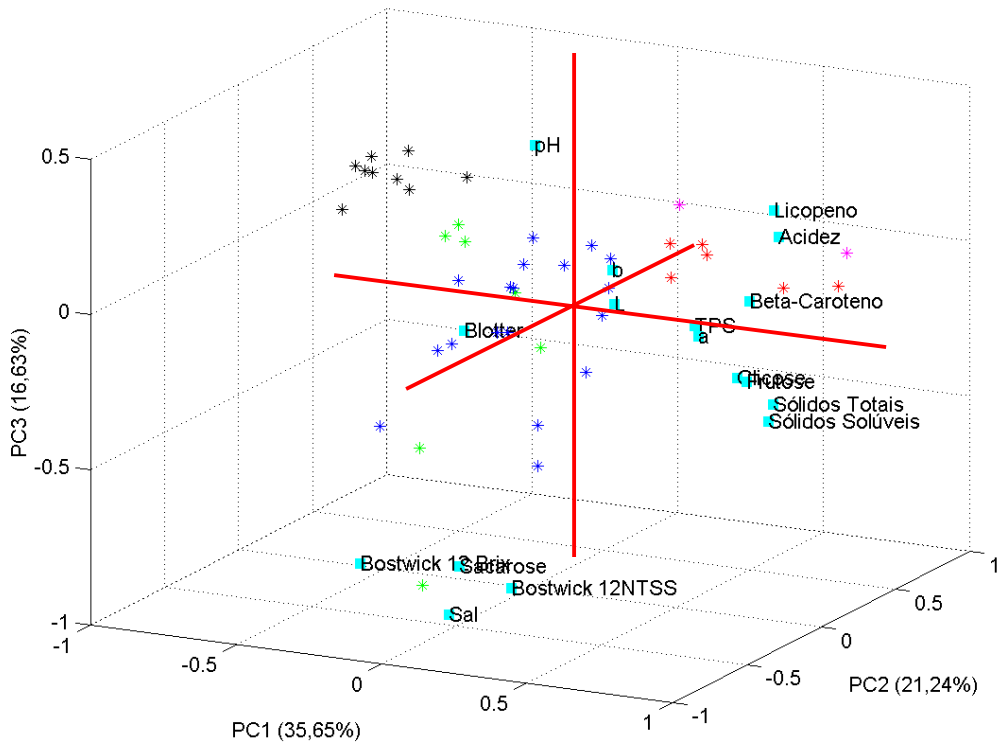

Figura 3.6 – *Biplot* para a PCA das propriedades físico-químicas dos concentrados de tomate. O ponto vermelho representa a origem dos eixos definidos pelas PCs.

A primeira componente principal tem, como observado anteriormente, peso relevante de atributos ligados à quantidade de tomate. Quanto mais à direita neste eixo estiverem as amostras, maiores suas quantidades de sólidos totais e solúveis, açúcares naturais do fruto (glicose e frutose) e carotenóides. Pode-se interpretar que a segunda componente principal define o balanço de sal, açúcar e a acidez dos produtos, importantes atributos sensoriais, enquanto que a terceira componente principal

classifica, grosso modo, principalmente amostras viscosas das fluídas. *Blotter Test* e os parâmetros de cor Hunter "L" e "b" apresentaram basicamente as mesmas variações para todas as amostras e não foram parâmetros relevantes em sua classificação, fato revelado por seu posicionamento próximo à origem do espaço gerado pelas componentes principais.

É importante também notar a coerência do gráfico da figura 3.6; como a PCA agrupa variáveis correlacionadas entre si no novo espaço gerado. Assim, variáveis como sólidos totais e solúveis, glicose e frutose, Bostwick 12 Brix e Bostwick 12 NTSS e os parâmetros de cor Hunter "a" e "TPS" encontram-se muito próximos, como esperado (na figura, glicose e frutose estão sobrepostos, e sacarose está ao lado do Boswick 12 Brix). Parâmetros de formulação, como sal e sacarose, também se agrupam. Este comportamento é extremamente útil para avaliação do cenário de produtos de concorrentes, permitindo que decisões de mercado robustas sejam tomadas pelos estrategistas da indústria de alimentos.

Nos gráficos das figuras 3.5 e 3.6 a amostra BR15 chama a atenção por estar isolada das demais. De fato, trata-se de um produto extremamente fluído, até mesmo comparável com as amostras de purê de tomate. Entretanto, não se agrupou com os purês de tomate porque apresentou teor de sólidos maior, além de elevada concentração de açúcares, principalmente mas não apenas sacarose.

3.4 – Construção de Modelos de Calibração para Sólidos e Carotenóides em Concentrados de Tomate

3.4.1 – Determinação dos Melhores Métodos de Calibração e Pré-Processamento de Espectros

A figura 3.7 mostra os 126 espectros brutos coletados para as 42 amostras de produtos de tomate. São compostos basicamente por três picos em aproximadamente 5100, 6800 e 8500 cm^{-1} e por dois ombros em aproximadamente 5500 e 7200 cm^{-1}. Nota-se também que a razão sinal/ruído é menor nas regiões entre 4000 e 5500 cm^{-1} e 6300 a 7300 cm^{-1}. Também são evidentes o considerável deslocamento (*offset*) e inclinação (*bias*) da linha de base.

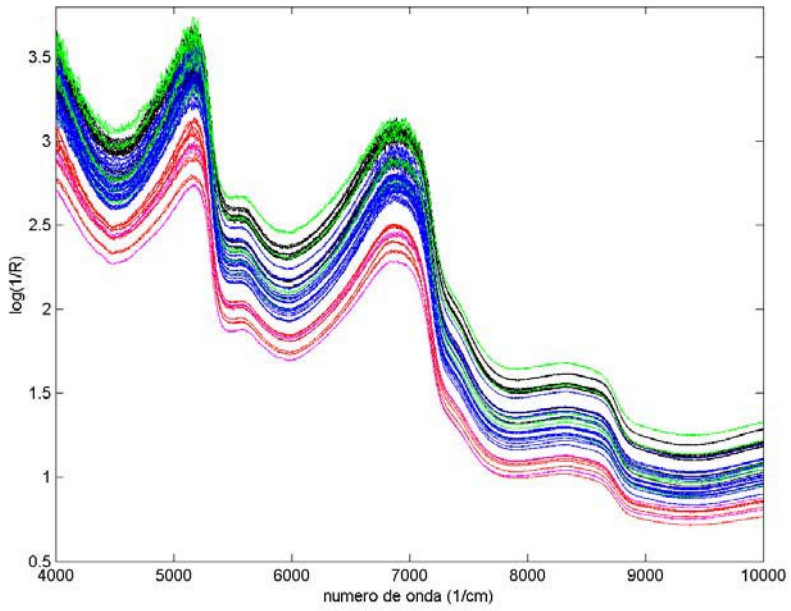

Figura 3.7 – 126 espectros originais para as 42 amostras de produtos de tomate. Classificação de cores: preto = purê de tomate, verde = base de tomate, azul = simples-concentrado, vermelho = duplo-concentrado e magenta = polpa de tomate.

Deslocamento e inclinação na linha de base são características comuns em técnicas espectroscópicas que utilizam aquisição de espectros por reflectância difusa. A reflectância de uma superfície é, de acordo com a teoria de Kubelka-Munk, função de uma constante de absorção (K) e de um coeficiente de espalhamento (S) [13 - 15]:

$$R = \frac{K}{S} \qquad \textbf{(3.2)}$$

Instrumentalmente, a reflectância (R) é obtida pela razão entre a intensidade da luz que atinge o detector após ser refletida pela amostra (I_a) e a de um material de referência (I_0), cujo coeficiente de absorção na região do infravermelho seja desprezível [13 - 15]:

$$R = \frac{I_a}{I_0} \qquad \textbf{(3.3)}$$

Nota-se, deste modo, que quaisquer diferenças entre os padrões de espalhamento da amostra e da referência causará um deslocamento constante da linha de base do espectro. No caso das amostras de produtos concentrados de tomate as diferenças entre as dimensões e, principalmente, entre as quantidades de material insolúvel particulado, são as responsáveis pelos deslocamentos observados. De fato, a figura 3.7 evidencia a existência de considerável correlação entre a concentração de tomate nos produtos e os deslocamentos nas linhas de base dos espectros. Esta correlação foi útil na construção de pelo menos um modelo de calibração, como ver-se-á adiante.

A inclinação nas linhas de base também é comum em espectros coletados por reflectância difusa porque o caminho óptico percorrido pela luz é inversamente proporcional à freqüência (e, consequentemente, à energia). Quanto maior a energia, menor o comprimento de onda da luz e mais o feixe de luz incidente torna-se comparável com as dimensões do material particulado presente na amostra, aumentando o espalhamento e diminuindo a componente de absorção da reflexão (equação 3.2) [13, 15].

O ruído presente em algumas regiões dos espectros, todavia, teve de ser minimizado ou eliminado por técnicas de alisamento. A figura 3.8 mostra os espectros alisados pela média com janela de 15 pontos. O número total de variáveis (comprimentos de onda) caiu de 1557 para 97 e, assim, pode-se dizer que uma primeira seleção de variáveis foi já executada.

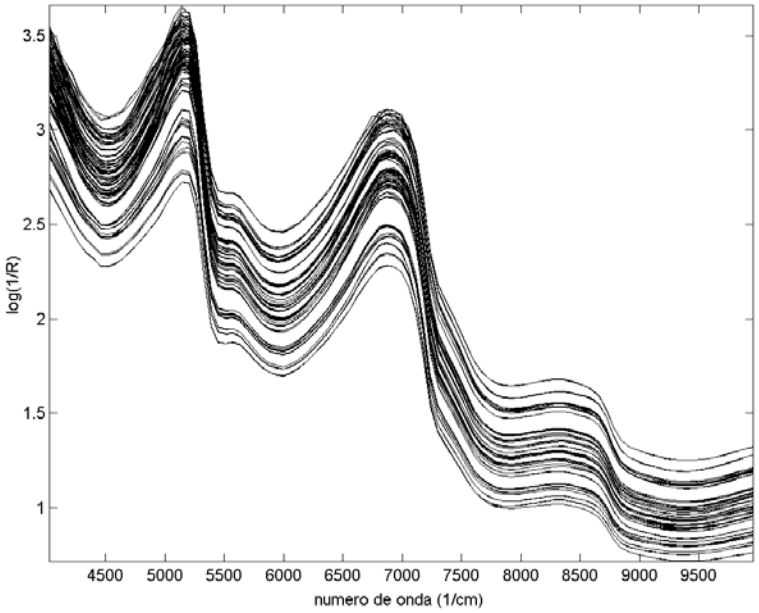

Figura 3.8 – Espectros alisados pela média com janela de 15 pontos.

Com os espectros alisados, modelos PLS e PCR foram construídos. A tabela 3.2 mostra as capacidades de predição de cada um deles:

Tabela 3.2 – Modelos de calibração para sólidos e carotenóides em concentrados de tomate, utilizando os espectros inteiros, pré-processados por alisamento.

Propriedade	**Modelo**	**Método de Calibração**	**Fatores**	**RMSEP**	**r_{val}**	***Outliers***
Sólidos Totais (%)	1	PLS	9	1,1749	0,9982	3
	2	PCR	13	2,0976	0,9943	3
Sólidos Solúveis (° Brix)	3	PLS	9	0,9576	0,9987	0
	4	PCR	13	1,9381	0,9947	0
Licopeno (mg kg^{-1})	5	PLS	7	24,5538	0,9996	0
	6	PCR	7	24,7641	0,9995	0
β-Caroteno (mg kg^{-1})	7	PLS	7	0,8349	0,9975	0
	8	PCR	9	0,8967	0,9972	0

Os três espectros anômalos (*outliers*) são referentes à amostra EU05 e foram removidas porque, além de apresentarem *leverage* acima do valor crítico e resíduos de Student maiores que 2 (figura 3.9), um modelo com melhor capacidade preditiva foi obtido sem elas.

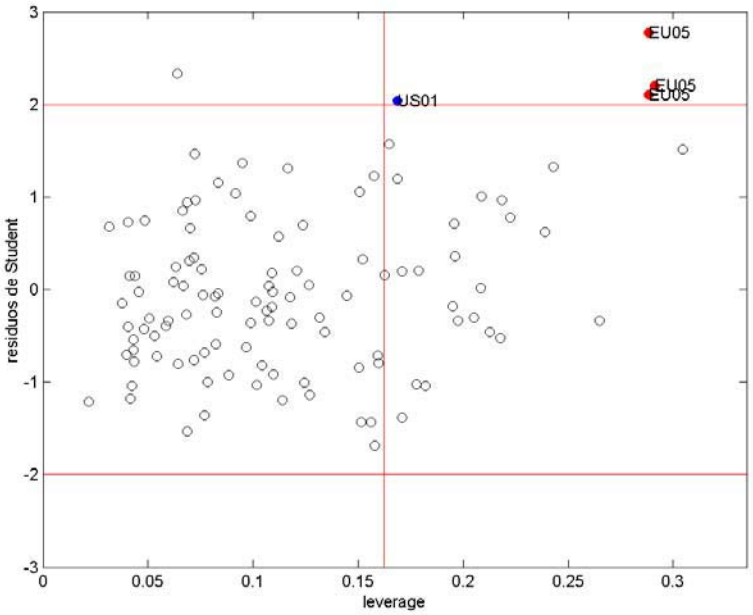

Figura 3.9 – Gráfico dos resíduos de Student pela *leverage* para o modelo 1 (tabela 3.2).

O espectro da amostra US01 (ponto azul) também foi inicialmente removido, mas o modelo resultante apresentou performance inferior à do modelo 1 e, assim, esta amostra foi mantida. Cabe ressaltar que amostras alocadas na região de elevados valores absolutos de resíduos de Student e de *leverage* são *possíveis outliers*, devendo seu efeito no modelo ser cuidadosamente examinado, como realizado com a amostra US01.

Como visto no capítulo 2, desvios e inclinação da linha de base (*offset* e *bias*) são características indesejáveis na calibração porque uma ou mais componentes principais serão utilizadas em sua descrição. A figura 3.10 mostra o gráfico de *loadings* para as 7 primeiras componentes principais do modelo 1. Nota-se que os *loadings* da primeira componente principal apenas adicionam um termo constante a todos os *scores* do modelo construído para a propriedade sólidos totais. Nota-se também que a segunda, a terceira e a sétima componentes estão sujeitas a inclinações, provavelmente devido ao *bias* dos espectros.

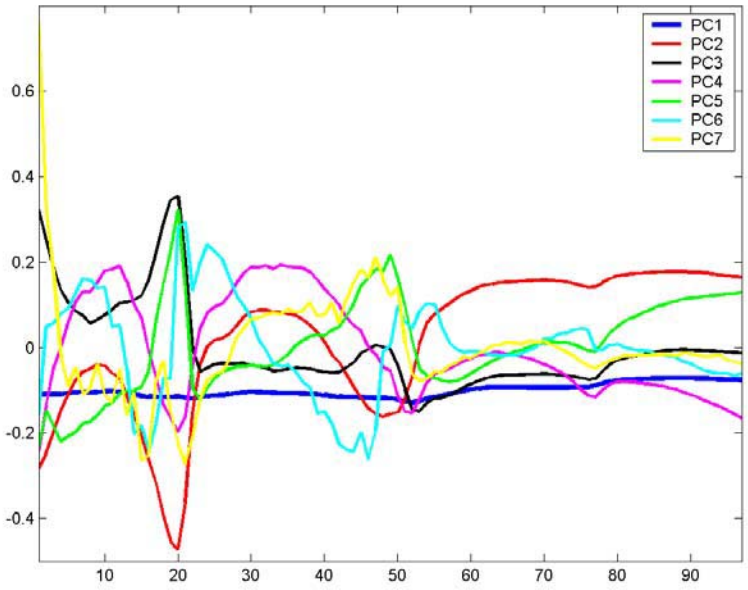

Figura 3.10 – Gráfico de *loadings* para as 7 primeiras componentes principais do modelo 1, para sólidos totais (tabela 3.2).

Para corrigir o efeito do deslocamento da linha de base dos espectros utilizou-se pré-processamento por MSC (figura 3.11). A tabela 3.3 traz os modelos PLS e PCR obtidos com estes espectros:

Tabela 3.3 – Modelos de calibração para sólidos e carotenóides em concentrados de tomate, utilizando os espectros inteiros pré-processado por alisamento e MSC.

Propriedade	**Modelo**	**Método de Calibração**	**Fatores**	**RMSEP**	**r_{val}**	***Outliers***
Sólidos Totais (%)	9	PLS	10	0,5523	0,9996	0
	10	PCR	10	0,9783	0,9988	0
Sólidos Solúveis (° Brix)	11	PLS	10	0,6750	0,9994	0
	12	PCR	11	1,4757	0,9969	0
Licopeno (mg kg^{-1})	13	PLS	9	61,8745	0,9974	0
	14	PCR	11	65,6345	0,9957	0
β-Caroteno (mg kg^{-1})	15	PLS	5	0,7326	0,9981	0[1]
	16	PCR	8	0,8052	0,9977	0[1]

Obs.: 1 - três *outliers* foram detectados, mas sua remoção não produziu modelos com melhor capacidade prediditiva e, então, esta amostra foi mantida no modelo.

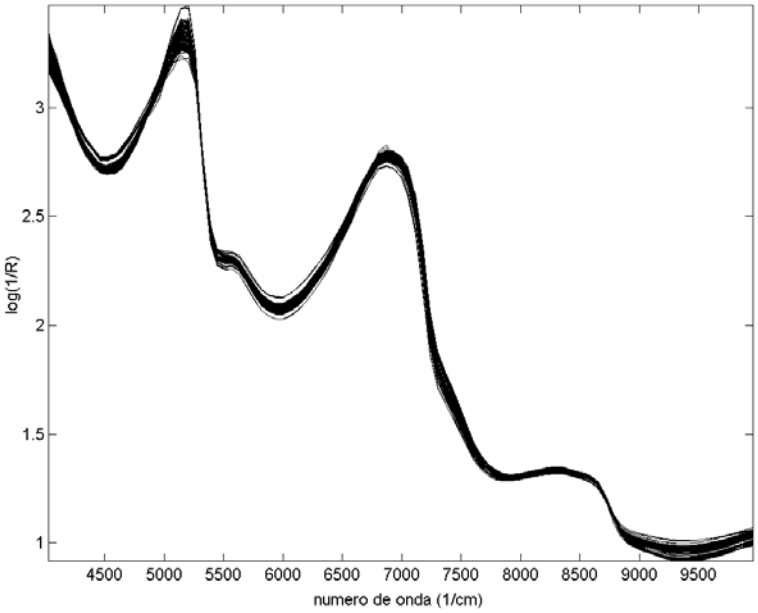

Figura 3.11 – Espectros pré-processados por correção multplicativa de sinais (MSC).

O deslocamento da linha de base, apesar de não ter sido completamente eliminado, foi consideravelmente reduzido. Já as inclinações nas linhas de base não foram removidas porque o pré-processamento por MSC consiste na projeção de todos os espectros no espectro médio, que possui essa mesma inclinação dos espectros que o originaram, conforme visto no capítulo 2.

A figura 3.12 mostra o gráfico de *loadings* do modelo de número 9, para a propriedade sólidos totais. Nota-se que os *loadings* de todas as componentes principais mostradas apresentam características que lembram o comportamento dos espectros; nenhuma descreve deslocamento da linha de base. Entretanto, as componentes principais 5, 4 e 7, principalmente, ainda apresentam valores elevados para os primeiros números de onda dos espectros, descrevendo ainda algum *bias*.

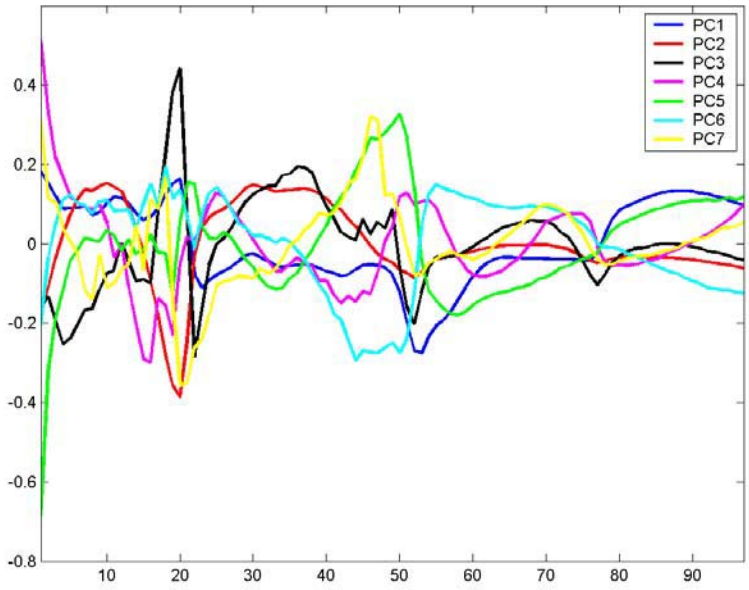

Figura 3.12 - Gráfico de *loadings* para as sete primeiras componente principais do modelo de número 9, tabela 3.3.

Um dos procedimentos possíveis para eliminação do *bias* dos espectros de infravermelho é a aplicação de derivadas. A figura 3.13 mostra o resultado deste procedimento para os espectros pré-processados por alisamento pela média:

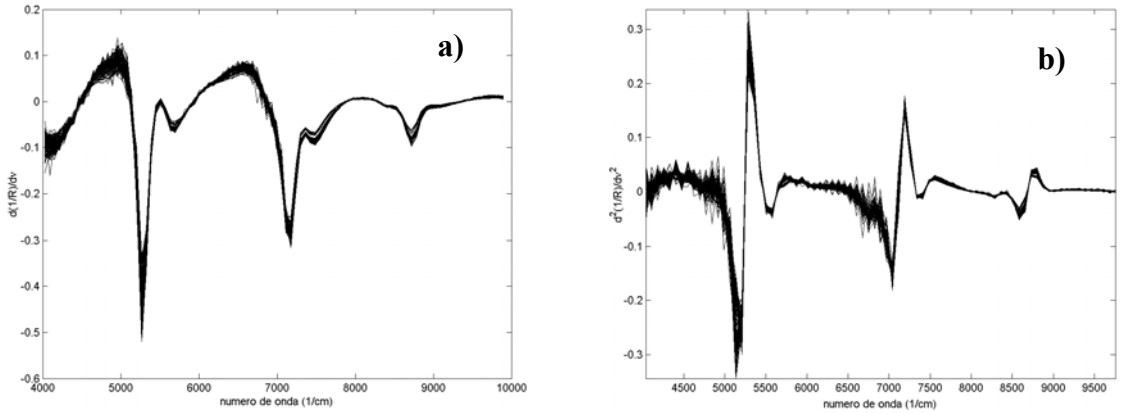

Figura 3.13 – Espectros pré-processados a) pela derivada-primeira e b) pela derivada-segunda.

Nota-se que o ruído resultante do procedimento de alisamento pela média foi amplificado quando as derivadas primeira e segunda foram aplicadas. A diminuição da razão sinal/ruído foi mais pronunciada para a derivação de maior ordem. Observa-se também que a inclinação nas linhas de base dos espectros foi convertida em um termo constante na figura 3.13(a) enquanto que a aplicação da derivada-segunda eliminou completamente tanto a inclinação quanto o deslocamento das linhas de base.

Muitos fatores foram geralmente requeridos para que os modelos de calibração PLS e PCR construídos com as derivadas dos espectros originais fornecesse capacidade preditiva ótima (tabela 3.4). De fato, os gráficos de *loadings* (figura 3.14) para a propriedade sólidos totais mostram que estes modelos foram consideravelmente afetados pela baixa razão sinal/ruído em regiões espectrais importantes.

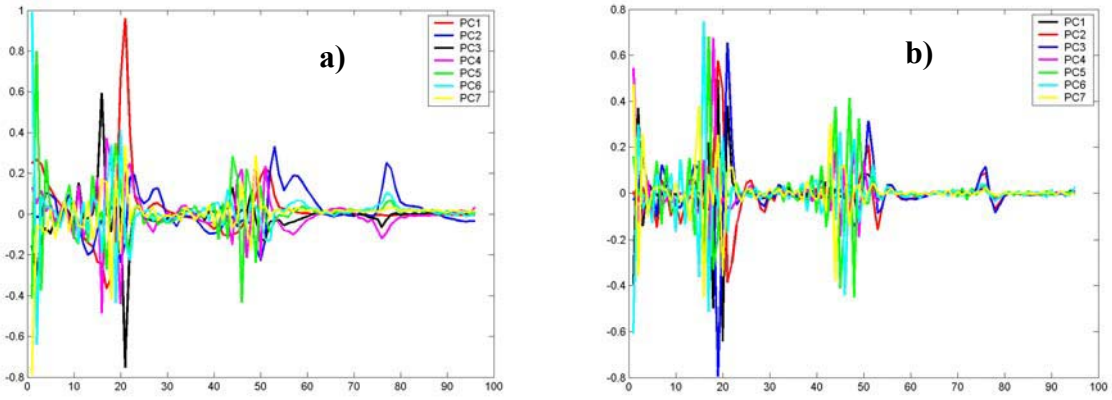

Figura 3.14 – Gráficos de *loadings* para os espectros pré-processados por: a) derivada-primeria e b) derivada segunda.

Tabela 3.4 – Modelos de calibração para sólidos e carotenóides em concentrados de tomate, utilizando os espectros inteiros pré-processados pelas derivadas primeira e segunda.

Propriedade	Modelo	Pré-Processamento	Método de Calibração	Fatores	RMSEP	r_{val}
Sólidos Totais (%)	17	Derivada-Primeira	PLS	20	1,2903	0,9980
	18		PCR	13	2,0550	0,9947
Sólidos Solúveis (° Brix)	19		PLS	20	0,9380	0,9989
	20		PCR	14	2,3357	0,9924
Licopeno ($mg\ kg^{-1}$)	21		PLS	14	55,7059	0,9976
	22		PCR	19	76,4512	0,9956
β-Caroteno ($mg\ kg^{-1}$)	23		PLS	10	1,2528	0,9943
	24		PCR	13	1,4061	0,9933
Sólidos Totais (%)	25	Derivada-Segunda	PLS	13	1,2201	0,9981
	26		PCR	9	1,0731	0,9989
Sólidos Solúveis (° Brix)	27		PLS	13	1,3761	0,9974
	28		PCR	9	1,3728	0,9978
Licopeno ($mg\ kg^{-1}$)	29		PLS	9	86,0096	0,9936
	30		PCR	9	107,0276	0,9901
β-Caroteno ($mg\ kg^{-1}$)	31		PLS	3	2,0562	0,9846
	32		PCR	6	2,2398	0,9815

Nota-se que nenhum dos modelos acima apresentou performance preditiva superior àqueles obtidos com os espectros processados por MSC. Além disso, ao contrário do que seria esperado com a remoção das características de linha de base, um número muito superior de fatores foi necessário para obtenção de vetores de regressão com capacidade preditiva ótima, gerando modelos mais complexos.

Viu-se anteriormente que os deslocamentos das linhas de base guardam alguma relação com a quantidade de material particulado presente nas amostras. Uma maneira de comprovar se esta informação estaria sendo relevante na predição das propriedades seria avaliar como os modelos de calibração se comportariam quando este deslocamento fosse removido: modelos com aplicação de derivadas seriam menos eficientes na predição de amostras externas que aqueles obtidos sem estes pré-processamentos. No presente caso, devido à presença marcante de ruído em regiões importantes dos espectros, a comparação dos modelos constantes na tabela 3.4 com aqueles das tabelas 3.2 e 3.3 é inconclusiva para este fim, uma vez que é impossível determinar de maneira inequívoca se a capacidade de predição dos modelos foi inferior devido ao ruído ou à remoção do *offset*.

O principal motivo pelo qual a razão sinal/ruído foi tão afetada nos espectros das derivadas foi um alisamento de eficiência inadequada para o método. Outra forma de realizar o alisamento de espectros é pela transformada de Fourier. Este método foi empregado a um dos espectros da figura 3.7 e derivadas calculadas, conforme mostrado na figura 3.15:

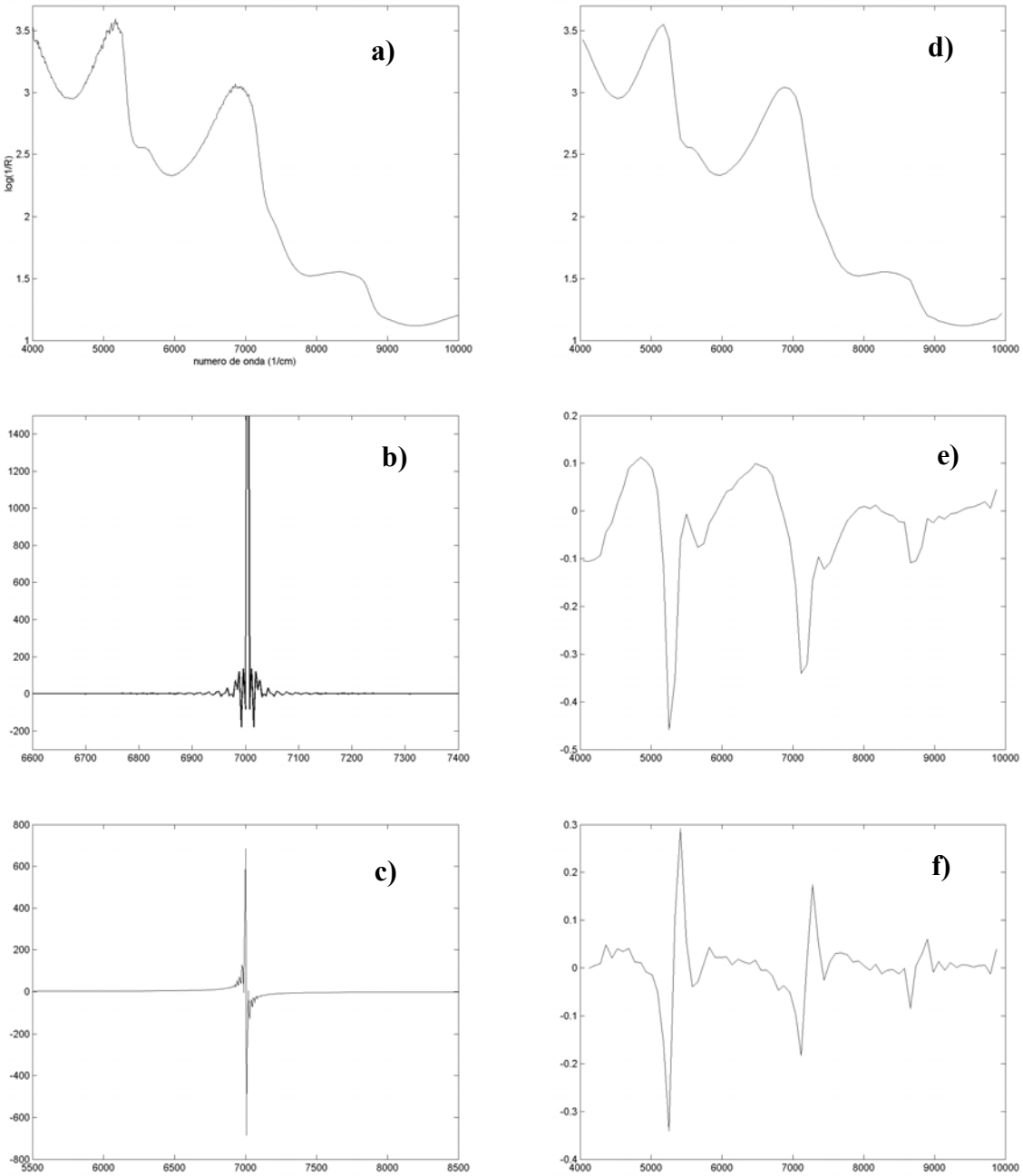

Figura 3.15 – a) um espectro bruto da amostra BR01; b) parte real da transformada de Fourier; c) parte imaginária da transformada de Fourier; d) espectro alisado; espectros após aplicação das derivadas e) primeira e f) segunda.

Uma característica interessante observada foi que a transformada de Fourier dos espectros de concentrados de tomate possuía informação relevante tanto na parte real quanto na parte imaginária, como pode-se visualizar na figura 3.15(c). Usualmente trabalha-se apenas com a parte real do interferograma; neste caso, a negligência da parte imaginária gerou espectros reconstruídos distorcidos ou completamente descaracterizados quando comparados com os originais.

O espectro resultante do processo de alisamento por Fourier (figura 3.13d) teve o elevado ruído das regiões espectrais anteriormente apontadas removido mas, por outro lado, pequenas ondulações foram inseridas por quase toda a curva. Deste modo, quando as derivadas foram aplicadas aos espectros reconstruídos, curvas insatisfatórias foram obtidas e, assim, modelos de calibração com capacidade preditiva inferior foram obtidos (não mostrado).

Regiões espectrais extremas podem apresentar instabilidade devido a características instrumentais e, assim, não raras vezes modelos são construídos sem elas. Por exemplo, estas regiões normalmente correspondem ao máximo (ou mínimo) de rotação de espelhos, movimento de hastes do interferômetro, etc. Assim, modelos de calibração PCR e PLS foram construídos para as regiões compreendidas entre 4500 e 9500 cm^{-1} para os espectros pré-processados por alisamento e por MSC:

Tabela 3.5 – Modelos de calibração para sólidos e carotenóides em concentrados de tomate, utilizando a região entre 4500 a 9500 cm^{-1} dos espectros pré-processados por alisamento.

Propriedade	Modelo	Método de Calibração	Fatores	RMSEP	r_{val}	*Outliers*
Sólidos Totais (%)	33	PLS	10	0,9701	0,9989	3
	34	PCR	10	1,6775	0,9965	3
Sólidos Solúveis (° Brix)	35	PLS	8	1,0506	0,9984	0
	36	PCR	10	2,0331	0,9941	0
Licopeno (mg kg^{-1})	37	PLS	6	21,7289	0,9996	0
	38	PCR	8	22,2211	0,9996	0
β-Caroteno (mg kg^{-1})	39	PLS	5	0,8394	0,9976	0[1]
	40	PCR	12	0,9079	0,9970	0[1]

Obs.: 1 - um *outlier* foi detectado, mas sua remoção não produziu modelos com melhor capacidade prediditiva e, então, esta amostra foi mantida no modelo.

O modelo de número 37, para licopeno, apresentou capacidade preditiva significativamente superior quando as regiões extremas dos espectros foram removidas. Além disso, o pré-processamento por alisamento pela média foi superior àquele obtido por MSC (tabela 3.6) revelando, assim, que o deslocamento da linha de base contém informação relevante para a calibração desta propriedade, contrariando o senso comum de que esta característica dos espectros pode ou até deve ser removida em busca do melhor modelo de calibração.

Tabela 3.6 – Modelos de calibração para sólidos e carotenóides em concentrados de tomate, utilizando a região entre 4500 a 9500 cm^{-1} dos espectros pré-processados por alisamento e MSC.

Propriedade	Modelo	Método de Calibração	Fatores	RMSEP	r_{val}	*Outliers*
Sólidos Totais (%)	41	PLS	8	0,8704	0,9990	0
	42	PCR	8	0,6846	0,9994	0
Sólidos Solúveis (° Brix)	43	PLS	9	0,7949	0,9991	0
	44	PCR	10	1,0440	0,9985	0
Licopeno (mg kg^{-1})	45	PLS	6	51,7676	0,9974	0
	46	PCR	8	79,9831	0,9930	0
β-Caroteno (mg kg^{-1})	47	PLS	5	0,7588	0,9979	0[1]
	48	PCR	7	0,7745	0,9979	0

Obs.: 1 - um *outlier* foi detectado, mas sua remoção não produziu um modelo com melhor capacidade prediditiva e, então, esta amostra foi mantida no modelo.

A figura 3.16 mostra o gráfico de *loadings* para as sete primeiras componentes principais obtidas do modelo de calibração PLS (número 37) para a propriedade sólidos totais. Nota-se que a primeira componente principal continua explicando basicamente os deslocamentos, enquanto que a segunda e a quarta componentes estão sujeitas às inclinações das linhas de base.

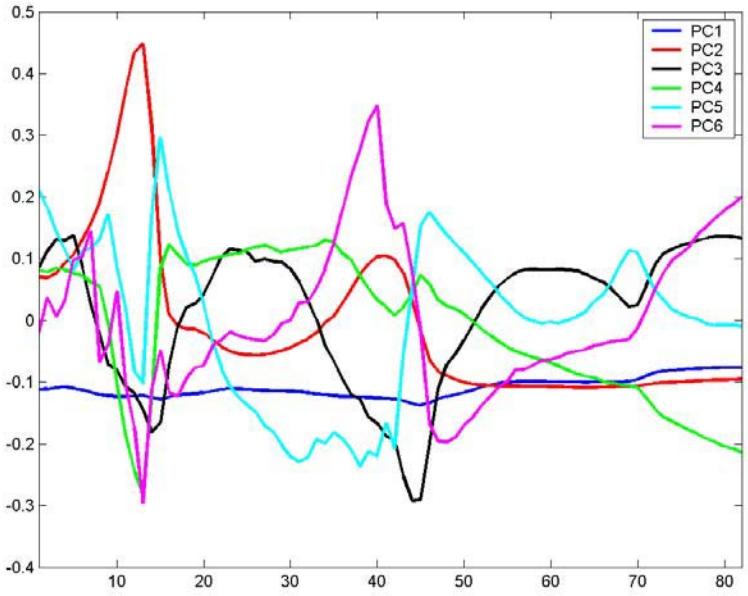

Figura 3.16 – Gráfico de *loadings* para as sete primeiras componentes principais do modelo de calibração número 37, para licopeno.

Conclui-se, pela análise das tabelas 3.2 a 3.6, que os melhores modelos de calibração foram, em sua maioria, obtidos pela aplicação de pré-processamento por alisamento seguido de correção multiplicativa de sinais, e que o melhor método de calibração é inequivocamente o PLS. Para licopeno, a remoção da linha de base por MSC produziu modelos com menor capacidade preditiva; deste modo, para esta propriedade, o deslocamento da linha de base contém informação relevante para o modelo de calibração.

3.4.2 – Avaliação de Métodos de Seleção de Variáveis na Complexidade e Capacidade Preditiva de Modelos de Calibração

Na seção precedente, determinou-se que o melhor método de calibração para as propriedades dos produtos de tomate foi o PLS, e que os métodos de pré-processamento dos espectros mais eficientes foram o alisamento pela média (para licopeno), seguido de correção multiplicativa de sinais (MSC) para as demais propriedades.

Os métodos de seleção de variáveis pelo correlograma, por D*imension-Wise Selection* (DWS), pelo *Successive Projections Algorithm* (SPA) e pela divisão simétrica de espectros foram aplicados aos modelos que apresentaram as melhores performances em termos de capacidade preditiva construídos na seção 3.4.1.

O objetivo destes cálculos de seleção de variáveis foi incrementar ainda mais a performance dos modelos, pela escolha de comprimentos de onda relevantes na calibração das propriedades de interesse, facilitando também a interpretação física do comportamento dos espectros em função destas propriedades.

As tabelas 3.7 a 3.10 mostram os modelos de calibração obtidos por cada método de seleção de variáveis.

Tabela 3.7 – Modelos de calibração PLS para sólidos e carotenóides em concentrados de tomate, utilizando a seleção de variáveis pelos correlogramas.

Propriedade	Modelo	Parâmetro de Corte	Número de Variáveis	Fatores	RMSEP	r_{val}
Sólidos Totais (%)	49	0,6	66	10	0,7085	0,9994
Sólidos Solúveis (°Brix)	50	0,5	74	10	0,7313	0,9996
Licopeno (mg kg^{-1})	51	0,9	76	9	34,1495	0,9990
β-Caroteno (mg kg^{-1})	52	0,5	85	6	0,8740	0,9972

A figura 3.17 traz os correlogramas para cada propriedade de interesse. As estruturas dos correlogramas para sólidos totais, sólidos solúveis e β-caroteno são bastante similares. De fato, as quantidades de nutrientes guardam alguma relação com a quantidade de tomate presente na amostra: quanto maior o nível de diluição, menor a concentração de nutrientes no produto.

O correlograma para licopeno apresenta relativamente pouca estrutura e mostra valores negativos para todos os comprimentos de onda. Este comportamento é ditado principalmente pela relação desta propriedade com o deslocamento de linha de base: grosso modo, quanto mais diluída é a amostra, menor é a quantidade de licopeno presente e maior o *offset* (vide figura 3.7).

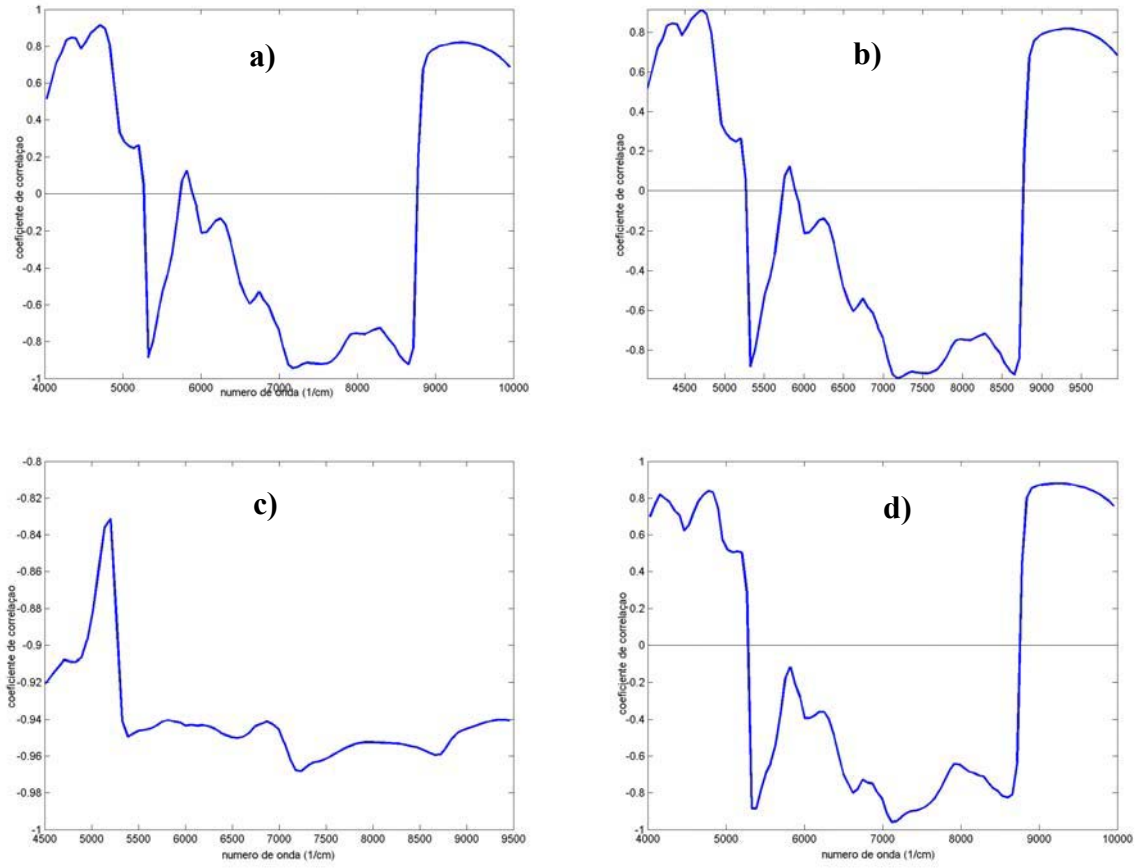

Figura 3.17 – Correlogramas entre os espectros e as propriedades: a) sólidos totais; b) sólidos solúveis; c) licopeno e d) β-caroteno.

A aplicação do método DWS forneceu modelos de calibração possuindo exatamente as mesmas características dos modelos-base, conforme tabela abaixo:

Tabela 3.8 – Modelos de calibração PLS para sólidos e carotenóides em concentrados de tomate, utilizando a seleção de variáveis pelo método DWS.

Propriedade	Modelo	Modelo-Base	Fatores	RMSEP	r_{val}
Sólidos Totais (%)	53	9	10	0,5523	0,9996
Sólidos Solúveis (°Brix)	54	11	10	0,6750	0,9994
Licopeno (mg kg^{-1})	55	37	6	21,7289	0,9996
β-Caroteno (mg kg^{-1})	56	15	5	0,7326	0,9981

Estes resultados devem-se ao fato de que as rotações proporcionadas às componentes principais pelo cancelamento de valores irrelevantes de **w** foram extremamente pequenas, quando efetivamente ocorreram, porque as soluções obtidas pelos modelos-base já estarem originalmente orientadas para a direção de melhor

performance e, assim, nenhuma rotação adicional resultou em modelos com melhores capacidades preditivas.

A figura 3.18 mostra o gráfico dos vetores de regressão, para a propriedade sólidos solúveis, obtidos pelos modelos PLS com e sem a aplicação da seleção de variáveis por DWS. Os vetores de regressão são bastante semelhantes, em ambos os casos. Para as demais propriedades, os vetores de regressão apresentaram nível ainda maior de concordância.

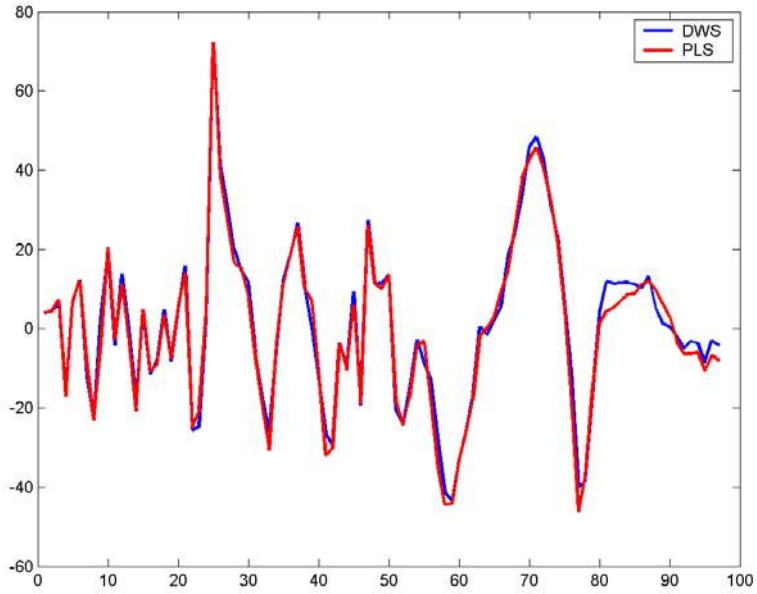

Figura 3.18 – vetores de regressão para a propriedade sólidos solúveis obtidos por PLS e por aplicação do método DWS.

O *Successive Projections Algorithm* foi também aplicado sobre os espectros de infravermelho das amostras de concentrados de tomate. Apesar do fato de o algoritmo principal apresentar-se razoavelmente eficiente, o custo computacional do processo de busca sistemática para determinação do vetor que iniciaria o processo de seleção (k_0) e do número ótimo de variáveis a extrair (N) foi consideravelmente alto: aproximadamente trinta minutos para os espectros pré-processados por MSC (com 97

comprimentos de onda) e cerca de dezoito minutos para os espectros pré-processados apenas por alisamento.

Pelo método SPA obtiveram-se modelos que não apresentaram maior eficiência na predição das propriedades de interesse que os anteriormente construídos, conforme demonstra a tabela 3.9. Cabe salientar também que o algoritmo consiste na busca das variáveis do conjunto de calibração que sejam o *mais* ortogonais *entre si* e, portanto a propriedade de interesse não é levada em consideração. Como nem sempre as variáveis mais ortogonais entre si são as que melhor explicam a propriedade de interesse, não é de todo surpreendente que este algoritmo não tenha fornecido modelos com melhores capacidades preditivas.

Tabela 3.9 – Modelos de calibração PLS para sólidos e carotenóides em concentrados de tomate, utilizando a seleção de variáveis pelo SPA.

Propriedade	Modelo	k_0 (cm^{-1})	N	Fatores	RMSEP	r_{val}
Sólidos Totais (%)	57	5266	36	13	1,2200	0,9983
Sólidos Solúveis (°Brix)	58	7116	36	11	1,6915	0,9960
Licopeno (mg kg^{-1})	59	8782	71	10	26,9269	0,9995
β-Caroteno (mg kg^{-1})	60	4030	45	5	1,0679	0,9960

Obs.: k_0 refere-se ao comprimento de onda por onde as projeções sucessivas se iniciaram, enquanto que N é o número ótimo de comprimentos de onda a extrair. Vide capítulo 2 para detalhes do método.

A figura 3.19 mostra, em azul, o comprimento de onda inicial (k_0) e, em vermelho, as demais variáveis selecionadas pelo SPA. Com exceção de β-caroteno, a região entre 8000 e 10000 cm^{-1} não foi considerada nos modelos de calibração.

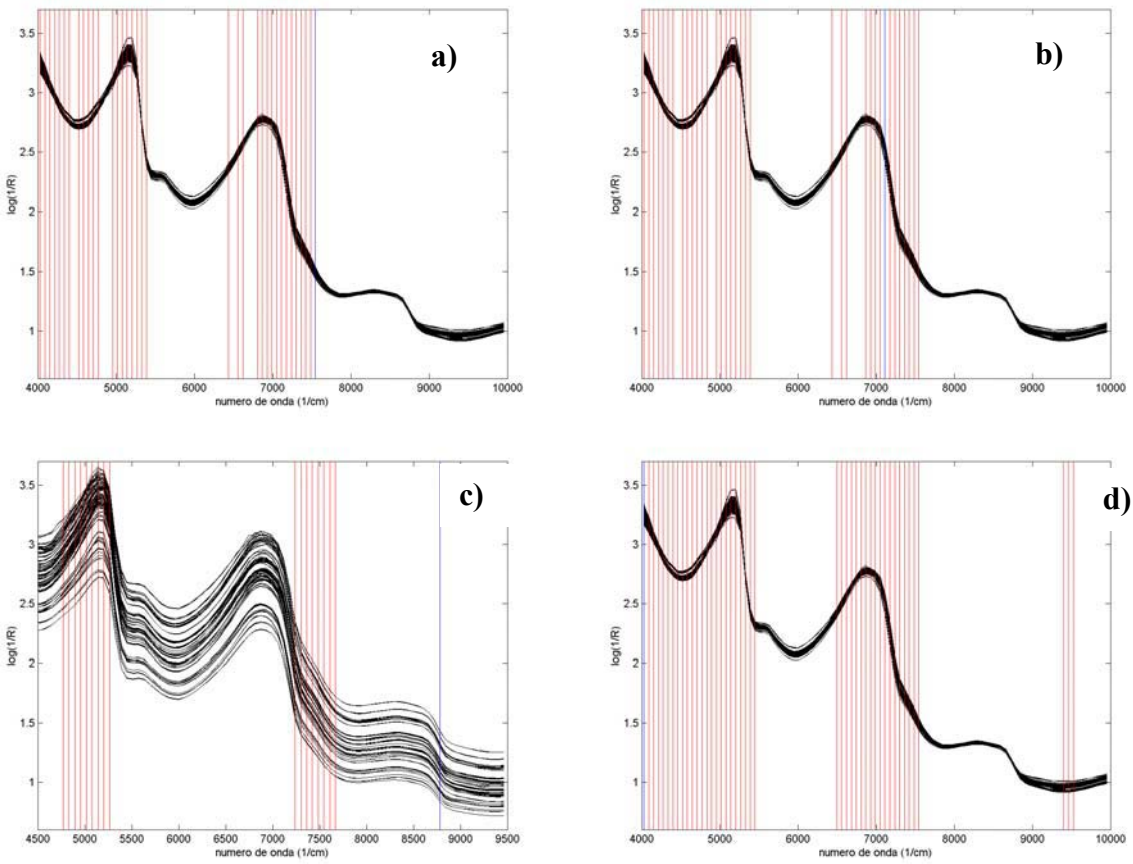

Figura 3.19 – Regiões dos espectros selecionadas para os modelos de calibração para as propriedades a) sólidos totais; b) sólidos solúveis; c) licopeno e d) β-caroteno. As linhas azuis correspondem a k_0 e as linhas vermelhas aos demais N-1 comprimentos de onda selecionados pelo método SPA.

A divisão simétrica de espectros pode ser entendida como um caso particular do IPLS (*Interval Partial Least Squares*) desenvolvido por Osborne *et alli* [16]. As frações dos espectros, bem como suas combinações, foram submetidas ao método PLS para construção de modelos de calibração. A tabela 3.10 mostra os melhores modelos obtidos por este método. A região 1 é compreendida entre os números de onda 4000 a 6000 cm^{-1}, a região 2 entre 6000 a 8000 cm^{-1} e a região 3 entre 8000 a 10000 cm^{-1}.

Tabela 3.10 – Modelos de calibração PLS para sólidos e carotenóides em concentrados de tomate, utilizando a seleção de variáveis por divisão simétrica de espectros.

Propriedade	Modelo	Região de Calibração	Fatores	RMSEP	r_{val}
Sólidos Totais (%)	61	1 e 3	10	0,4157	0,9998
Sólidos Solúveis (°Brix)	62	2 e 3	11	0,6333	0,9996
Licopeno (mg kg^{-1})	63	1 e 2	5	21,5779	0,9996
β-Caroteno (mg kg^{-1})	64	1 e 2	5	0,7455	0,9981

Modelos com melhores capacidades preditivas foram obtidos para sólidos totais, sólidos solúveis e licopeno. Para β-caroteno o modelo-base continuou apresentando melhor performance. Além disso, esse método mostrou-se o mais rápido e menos custoso computacionalmente.

Uma vez construídos os melhores modelos de calibração, sua capacidade preditiva foi avaliada para as cinco amostras do conjunto de validação, conforme tabela 3.11:

Tabela 3.11 – Validação dos modelos PLS para sólidos totais e solúveis, licopeno e β-caroteno em concentrados de tomate.

Propriedade	Número do Modelo	Amostra	Valor Medido	Valor Previsto	Erro Absoluto	Erro Relativo (%)
Sólidos Totais (%)	61	BR27	7,74	7,78	-0,04	0,57
		BR16	16,80	17,18	-0,50	3,00
		BR23	22,40	21,81	0,58	2,63
		BR25	19,06	18,63	0,42	2,24
		PB02	27,10	27,13	-0,03	0,11
Sólidos Solúveis (º Brix)	62	BR27	7,3	7,9	-0,6	7,76
		BR16	15,4	16,6	-1,2	7,61
		BR23	22,3	22,2	0,1	0,27
		BR25	18,7	18,7	0,0	0,17
		PB02	25,0	25,3	-0,3	1,11
Licopeno (mg kg^{-1})	63	BR27	144	135	9	6,45
		BR16	602	608	-6	1,03
		BR23	601	596	5	0,75
		BR25	779	734	44	5,69
		PB02	921	916	5	0,49
β-Caroteno (mg kg^{-1})	15	BR27	5,1	5,0	0,1	1,50
		BR16	11,1	11,4	-0,3	3,11
		BR23	11,0	11,1	-0,1	0,93
		BR25	10,3	10,8	-0,5	4,79
		PB02	17,5	16,3	1,19	6,83

A figura 3.20 mostra os gráficos dos valores previstos pelos medidos para cada uma das propriedades submetidas à calibração. A reta de 45° corresponde a valores previstos e experimentais idênticos; deste modo, reflete uma condição ideal de predição e, assim, quanto mais agrupados estiverem os resultados desta reta, melhor o modelo.

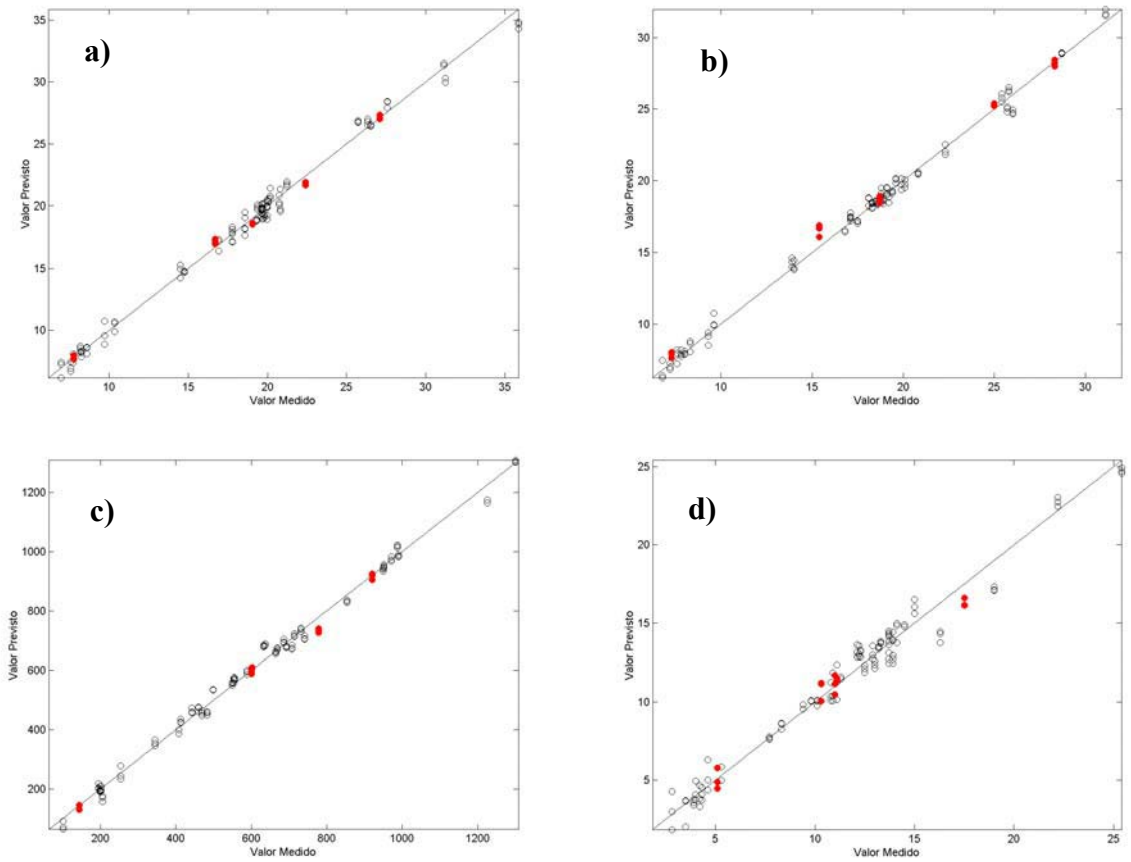

Figura 3.20 – Valores medidos *vs.* previstos para a) sólidos totais; b) sólidos solúveis; c) licopeno e d) β-caroteno em concentrados de tomate. Os pontos pretos correspondem aos espectros de calibração e os vermelhos às amostras de validação externa.

Os valores previstos para sólidos totais e solúveis, licopeno e β-caroteno forneceram erros de previsão consistentes com os métodos de referência (entre 1 a 6%, aproximadamente) e, assim sendo, podem ser empregados a diversas aplicações nas indústrias de alimentos.

3.5 – Conclusões

A análise por componentes principais (PCA) sobre as propriedades físico-químicas permitiu, através de inspeção visual rápida, a identificação de padrões de comportamento em um conjunto de amostras de diferentes origens onde determinações distintas foram realizadas, economizando tempo e recursos na avaliação estatística numérica destes resultados. Esta análise pode ainda ser incrementada com dados mercadológicos e, deste modo, avaliações robustas das estratégias de concorrentes podem ser realizadas.

Modelos de calibração relacionando características espectrais de produtos concentrados de tomate foram construídos utilizando procedimentos de pré-processamento de espectros e seleção de variáveis relativamente simples. Os métodos de pré-processamento mais eficientes foram o alisamento pela média, para licopeno, e a correção multiplicativa de sinais (MSC) para sólidos e β-caroteno.

O método PLS mostrou-se mais eficaz na calibração das propriedades de interesse que o PCR, fornecendo modelos mais simples e com melhor capacidade preditiva.

O fato do deslocamento da linha de base apresentar correlação com a concentração de tomate mostrou-se útil em pelo menos um modelo de calibração. O licopeno, uma das propriedades mais interessantes deste estudo – por apresentar maior potencial econômico para a indústria, apresentou melhor modelo de calibração quando este *offset* não foi removido. É usual, em calibração multivariada, eliminar este termo constante das linhas de base pela aplicação de derivadas ou outras técnicas de subtração. Este estudo demonstrou, contudo, que tal procedimento deve ser adotado com cautela.

Dos métodos de seleção de variáveis estudados, o de divisão simétrica de espectros, o mais simples tanto computacional quanto matematicamente, forneceu os melhores modelos PLS de calibração, demonstrando que não é a complexidade matemática do método que define sua eficiência.

Deve-se observar, também, que o fato de um método de seleção ter fornecido resultados mais interessantes para uma aplicação não significa sobremaneira que este seja mais robusto que os demais. É o número de aplicações em diferentes situações que rotulará a robustez de um método de seleção de variáveis. De fato, a avaliação criteriosa dos espectros e do problema em mãos, juntamente com o conhecimento prévio do sistema químico em estudo, continuam constituindo a base da análise química, independente da aplicação de ferramentas matemáticas avançadas. Todavia, quando aplicada conscientemente, a quimiometria pode fornecer resultados de alto impacto não apenas na rotina de um laboratório analítico, como em questões econômicas de relevância.

Estima-se que o tempo de avaliação de novas variedades de tomate ricas em sólidos ou carotenóides possa ser reduzido em um terço com a aplicação dos modelos de calibração aqui desenvolvidos.

3.6 – Referências

1. Robinson, C.; "Genetic Modification Technology and Food"; ILSI Europe; Brussels 2001.
2. Voet, D., Voet, J., Pratt, C.W.; "Fundamentos de Bioquímica"; Artmed Editora; Porto Alegre 2000.
3. Lajolo, F.M., Nutti, M.R.; "Transgênicos: Bases Científicas da sua Segurança"; SBAN; São Paulo 2003.
4. Gould, W.A.; "Tomato Production, Processing & Technology"; 3rd Ed.; CTI Pub. Inc.; Baltimore 1992.
5. Nielsen, S.S.; "Food Analysis"; 2nd Ed.; Gaithersburg; Aspen 1998.
6. Metrhom Ltd.; "792 Basic Titrino – Instructions for Use"; 8.792.1003.
7. Vogel, A.I.; "Vogel – Análise Química Quantitativa"; 2ª Ed.; Guanabara Koogan; Rio de Janeiro; 1992.
8. Deman, J.M. *et. alli*; "Rheology and Texture in Food Quality"; Avi; Westport; 1976.
9. HunterLab; "HunterLab PC2D Intructions Manual"; HunterLab, Inc.; Reston; 1980.

10. Sadler, G., Davis, J., Dezman, D.; "Rapid Extraction of Lycopene and Beta-Carotene from Reconstituted Tomato Paste and Pink Grapefruit Homogenates"; *J. Food Sci.*; **1990**; *55*; 1460 – 1461.

11. Amaya, D.B.R.; "A Guide to Carotenoid Analysis in Foods"; ILSI; Washington 1999.

12. Büchi Labortechnik; "Büchi Nirlab Chemometrics Software Manual", fornecido com o software NIRCal.

13. Williams, P., Norris, K.; "NIR Technology in the Agricultural and Food Industries"; Am. Ass. Cereal Chemists; St. Paul 1990.

42. Pasquini, C.*;* "Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications"; *J. Braz. Chem. Soc.*; **2003**; *14*; 198-219.

43. Olinger, J.M., Griffiths, P.R.; "Effects of Sample Dilution and Particle Size/Morphology on Diffuse Reflectance Spectra of Carbohydrate Systems in the Near- and Mid-Infrared. Part I: Single Analytes; *Appl. Spectr.*; **1993**; *47*; 687-694.

14. Osborne, S.D. *et alli*; "The Intermediate Partial Least Squares Algorithm"; *Analyst*; **1997;** *122*; 1531-1537.

Anexo I – Resultados das Determinações Físico-Químicas em Produtos Atomatados

Amostras	Sólidos Totais (%)	Sólidos Solúveis (º Brix)	Sólidos Insolúveis (%)	pH	Sal (% NaCl)	Acidez (% Ác. Cítrico)	Bostwick 12NTSS (cm)	Bostwick 12 Brix (cm)	Blotter Test (cm)
BR01	14,76	14,0	0,76	4,10	1,04	0,88	4,8	8,5	14,0
BR02	19,71	18,8	0,91	4,23	1,47	0,89	6,6	7,7	11,0
BR03	18,56	17,5	1,06	4,04	2,94	0,68	6,8	11,9	2,5
BR04	19,60	18,3	1,30	4,21	1,46	0,80	5,1	6,2	13,0
BR05	20,73	19,9	0,83	5,07	1,12	1,36	8,4	9,9	15,0
BR06	19,32	18,3	1,02	4,25	1,21	0,78	6,4	7,0	14,5
BR07	19,66	18,6	1,06	4,22	1,03	1,03	8,9	10,0	8,0
BR08	19,67	18,8	0,87	4,18	1,27	1,09	8,1	9,4	13,5
BR09	20,04	19,4	0,64	4,11	1,19	0,85	8,8	10,1	>25
BR10	18,58	18,1	0,48	4,19	2,01	0,86	6,1	7,9	9,0
BR11	19,97	19,1	0,87	4,24	2,21	0,65	7,4	10,9	15,0
BR12	16,92	16,8	0,12	4,07	3,71	0,51	7,8	13,9	25,0
BR13	19,38	18,5	0,88	4,23	1,26	0,87	6,1	6,9	8,0
BR14	20,18	19,6	0,58	4,04	2,19	0,77	8,8	12,0	14,0
BR15	21,23	20,8	0,43	4,16	3,67	0,49	11,8	19,0	23,0
BR16	16,68	15,4	1,28	4,37	1,42	0,69	4,3	6,6	11,0
BR17	17,79	17,1	0,69	4,21	2,37	0,82	5,9	10,8	>25
BR18	17,98	17,1	0,88	4,26	2,95	0,59	6,8	11,9	2,0
BR19	19,66	18,9	0,76	4,29	1,18	0,87	6,7	8,5	12,0
BR20	14,48	13,9	0,58	4,28	1,27	0,67	4,6	6,8	6,0
BR21	19,82	19,2	0,62	4,28	1,26	0,92	5,9	7,7	8,0
BR22	19,97	19,1	0,87	4,07	1,23	1,20	4,6	4,8	9,0
BR23	22,40	22,3	0,10	4,04	4,72	0,77	11,1	11,8	24,0
EU01	31,24	28,3	2,94	4,22	0,33	1,99	5,5	5,5	>25
EU02	31,16	28,7	2,46	4,09	0,46	1,95	5,6	5,6	>25
BR24	20,80	20,1	0,70	4,33	2,11	0,34	>25	12,9	>25
US01	27,57	25,8	1,77	4,08	0,42	2,20	4,50	4,9	9,0
EU03	26,52	26,0	0,52	4,33	0,33	1,40	4,10	4,6	7,0
US02	26,32	25,7	0,62	3,97	0,41	2,46	4,30	5,0	10,0
US03	25,74	25,4	0,34	4,29	0,91	1,26	5,60	5,9	8,0
BR25	19,06	18,7	0,36	4,21	1,25	0,78	4,5	6,4	7,0
PB01	35,86	31,1	4,76	4,46	0,55	1,42	4,1	4,1	6,5
PB02	27,10	25,0	2,10	4,22	0,47	1,50	4,2	4,2	7,0
EU03	9,69	9,3	0,39	4,36	0,15	0,66	9,5	9,5	>25
EU04	10,35	9,6	0,75	4,22	1,05	0,58	8,0	8,0	13,0
BR26	8,18	8,0	0,18	4,40	0,60	0,37	8,4	8,4	24,0
BR27	7,74	7,3	0,44	4,44	0,51	0,36	7,8	7,8	24,0
AL01	8,58	8,3	0,28	4,35	0,25	0,28	10,8	10,8	>25
BR28	8,23	7,8	0,43	4,18	0,62	0,36	12,6	12,6	>25
BR29	6,94	6,8	0,14	4,34	0,12	0,34	6,9	6,9	17,2
BR30	7,56	7,2	0,36	4,49	0,50	0,31	9,0	9,0	>25
BR31	7,73	7,6	0,13	4,31	0,46	0,42	8,8	8,8	18,0

Amostras	Cor Hunter				Açúcares			Licopeno ($\mu g\ kg^{-1}$)	Beta-Caroteno ($\mu g\ kg^{-1}$)
	L	a	b	TPS	Sacarose (%)	Frutose (%)	Glicose (%)		
BR01	22,7	18,6	11,7	38,9	0,31	3,33	2,63	198	4,0
BR02	25,9	26,8	13,2	47,5	1,57	4,11	2,91	201	4,2
BR03	23,2	19,9	11,9	40,4	0,83	3,62	3,06	195	3,9
BR04	24,4	21,4	12,7	42,2	0,78	4,56	3,74	101	3,5
BR05	25,6	23,3	13,3	43,9	0,51	4,83	4,13	199	4,6
BR06	21,9	17,6	11,2	37,5	1,82	3,96	3,39	205	2,8
BR07	24,4	25,5	12,7	46,0	2,14	4,66	3,57	463	4,3
BR08	25,0	19,1	13,3	39,5	0,46	4,40	3,80	588	9,8
BR09	22,1	19,7	11,4	39,9	0,75	6,05	4,73	636	11,3
BR10	23,8	27,7	11,9	48,4	0,26	4,24	3,06	555	11,1
BR11	23,8	26,8	12,4	47,7	1,26	5,88	4,09	633	13,7
BR12	24,2	20,6	11,8	41,1	1,97	3,15	2,67	741	13,0
BR13	22,6	20,1	11,4	40,2	1,59	5,18	3,87	686	13,9
BR14	27,0	24,1	13,5	44,6	2,56	5,06	4,34	345	8,3
BR15	24,6	24,9	12,2	45,6	5,31	3,65	3,16	669	12,9
BR16	21,7	20,1	11,4	40,2	1,22	3,95	2,57	602	11,1
BR17	25,2	24,5	12,9	45,2	0,51	3,65	3,27	408	12,2
BR18	22,3	19,4	11,6	39,7	1,29	4,57	2,63	554	12,5
BR19	23,2	21,0	12,0	41,6	1,16	4,27	3,76	550	13,2
BR20	23,3	20,6	12,0	41,2	0,55	3,05	2,52	498	10,1
BR21	22,6	21,3	11,5	41,6	0,58	4,37	3,46	665	13,3
BR22	21,6	20,4	11,4	40,6	0,65	5,22	3,07	460	10,8
BR23	25,7	23,7	13,4	44,2	0,68	5,01	4,57	601	11,3
EU01	26,0	28,7	13,5	49,2	0,00	7,11	5,04	693	13,9
EU02	23,2	23,5	12,0	44,2	0,00	8,39	6,54	854	16,3
BR24	19,9	17,1	9,8	34,6	0,25	6,88	6,99	601	11,0
US01	24,6	22,7	12,3	43,5	0,00	5,72	5,63	1301	25,4
EU03	25,2	24,9	12,8	45,6	0,50	6,26	6,88	1226	22,2
US02	23,1	23,3	11,7	43,7	0,45	4,18	4,12	254	10,9
US03	25,2	23,1	12,8	43,8	0,00	6,73	7,26	972	12,3
BR25	23,0	21,1	11,5	41,4	0,79	2,64	3,60	779	10,3
PB01	23,0	23,9	12,0	44,5	0,00	9,55	8,18	987	15,0
PB02	21,9	17,1	10,5	36,0	0,00	8,19	6,18	921	17,5
EU03	24,5	27,4	13,0	48,2	0,00	1,78	1,61	952	13,7
EU04	22,1	16,8	11,0	36,4	0,00	2,26	2,11	990	19,0
BR26	25,3	23,9	12,9	44,6	0,39	1,75	1,40	750	12,1
BR27	23,8	21,8	12,3	42,5	0,25	1,71	1,34	144	5,1
AL01	25,4	25,1	13,5	45,5	0,00	2,09	1,86	413	5,3
BR28	23,7	20,7	12,3	41,3	0,25	1,71	1,34	714	13,7
BR29	27,7	23,7	14,0	43,6	0,47	0,92	1,30	442	9,1
BR30	24,4	22,7	12,7	43,5	0,49	1,72	1,35	708	14,1
BR31	24,4	20,4	12,8	41,1	0,52	1,80	1,30	469	7,70