

**PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E  
PROCESSOS INDUSTRIAIS**

Gilson Augusto Helfer

**CHEMOSTAT: DESENVOLVIMENTO DE *SOFTWARE* PARA ANÁLISE  
EXPLORATÓRIA DE DADOS MULTIVARIADOS**

Santa Cruz do Sul  
2014

Gilson Augusto Helfer

**CHEMOSTAT: DESENVOLVIMENTO DE *SOFTWARE* PARA ANÁLISE  
EXPLORATÓRIA DE DADOS MULTIVARIADOS**

Dissertação apresentada ao Programa de Pós-Graduação em Sistemas e Processos Industriais – Mestrado, como requisito parcial para obtenção do Título de Mestre em Sistemas e Processos Industriais.

Orientador: Prof. Dr. Luciano Marder

Coorientador Prof. Dr. João Carlos Furtado

Santa Cruz do Sul

2014

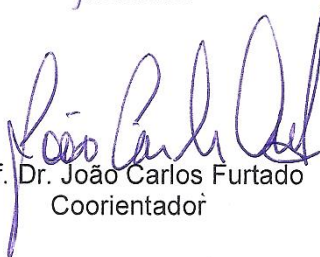
Gilson Augusto Helfer

**CHEMOSTAT: DESENVOLVIMENTO DE SOFTWARE PARA ANÁLISE  
EXPLORATÓRIA DE DADOS MULTIVARIADOS**

Esta dissertação foi submetida ao Programa de Pós-Graduação em Sistemas e Processos Industriais – Mestrado, Universidade de Santa Cruz do Sul – UNISC, como requisito parcial para obtenção do título de Mestre em Sistemas e Processos Industriais.



Prof. Dr. Luciano Marder  
Orientador



Prof. Dr. João Carlos Furtado  
Coorientador



Prof. Dr. Adilson Ben da Costa  
Examinador - UNISC



Prof. Dr. Marco Flôres Ferrão  
Examinador - UFRGS

## AGRADECIMENTOS

À Deus, pela sabedoria e discernimento necessários para chegar ao final deste trabalho.

Ao meu orientador, e ex-colega de graduação, Prof. Dr. Luciano Marder, pela atenção, pelas velhas fotos e pelo incentivo durante a trajetória desta investigação.

Ao meu coorientador e Prof. Dr. João Furtado pelas sugestões no decorrer da pesquisa; ao Prof. Dr. Adilson Ben da Costa, pela contribuição e disponibilização de algoritmos do Matlab; e aos professores do PPGSPI que de alguma forma contribuíram para o crescimento do meu conhecimento, em especial ao Prof. Dr. Ruben Panta pela sabedoria transmitida em suas simples lições.

Ao Prof. Dr. Marco Flôres Ferrão pelo incentivo constante e por acreditar em minha capacidade acadêmica desde há muito tempo – mais até do que eu próprio –, características dignas de um verdadeiro mestre.

À funcionária do PPGSPI Janaína Haas pela disponibilidade em auxiliar sempre que necessário.

Aos colegas do PPGSPI, que compartilharam conhecimento e horas de estudo, em especial a turma do “diurético”, Manoel Mazzuchi, Charles Neu, Giuliano Forgiarini, Roberta Kaufmann, Cátia Machado, Jaqueline Krüger, Edu Mazzini e Fábio Paz.

À minha noiva Karina Meneghetti Brendler, pela compreensão nos momentos em que estive ausente, mesmo presente. Pelo amor e carinho de forma contínua.

À minha família, meus pais Elemar João e Elaine Helfer, pelos braços sempre abertos; à minha filha, Manuela Helfer, a quem busco ser exemplo; à Pedro Franz e Sueli Meneghetti Brendler pelo apoio incondicional; ao Tenente Jorge Brendler (*in memoriam*), pelas boas lembranças (e frases) que restaram vivas.

Aos amigos Samuel Weis, Laone Kuentzer, Anderson e Alex Muller cuja amizade ultrapassa o tempo.

À Fernanda Bock e Lilian Ferreira pela dedicação e auxílio no laboratório.

À Triângulo Alimentos, pela doação das amostras.

À Capes, pelo apoio financeiro.

À todos aqueles que colaboraram direta ou indiretamente para a elaboração deste trabalho.

## RESUMO

Este trabalho, motivado pelo vasto uso da quimiometria associado em grande parte a dependência de aplicativos que requerem licença de operação e/ou uso de rotinas, teve com objetivo desenvolver um *software* gratuito, de uso acadêmico, de fácil instalação e manuseio, sem necessidade de programação em nível de usuário, para análise exploratória de dados. O *software* desenvolvido e denominado ChemoStat, contempla as técnicas de análise de agrupamento hierárquico (HCA), análise de componentes principais (PCA), análise de componentes principais por intervalos (iPCA), assim como, técnicas de correção, transformação dos dados e detecção de amostras anômalas. Os dados podem ser importados através da área de transferência, arquivos de texto, ASCII ou do FT-IR Perkin-Elmer (.sp). É possível gerar uma grande variedade de gráficos e tabelas que permitem a análise dos resultados os quais podem ser exportados em inúmeros formatos. As principais funcionalidades do *software* foram exploradas utilizando espectros no infravermelho médio e próximo de óleos vegetais e imagens digitais de diferentes tipos de óleo diesel comercial. Como forma de validar os resultados do *software*, os mesmos conjuntos de dados foram analisados utilizando o Matlab<sup>®</sup> e os resultados em ambos os aplicativos coincidiram nas mais diversas combinações. Além da versão *desktop*, o reuso dos algoritmos permitiu disponibilizar uma versão *online* que oferece uma experiência única via web.

**Palavras-chave:** *software*, quimiometria, análise exploratória de dados.

## ABSTRACT

The objective of this work is to develop an exploratory data analysis software for free and academic use that is easy to install and can be handled without user-level programming due the extensive use of chemometrics and its association with the applications that require purchased license or routines. The developed software, named Chemostat, employs Hierarchical Cluster Analysis (HCA), Principal Component Analysis (PCA), intervals Principal Component Analysis (iPCA), as well as correction methods, data transformation and outlier detection. The data can be imported from the clipboard, text files, ASCII or FT-IR Perkin-Elmer “.sp” files. It generates a variety of charts and tables that allows the analysis of the results which can be exported in several formats. The main features of the software were tested using mid-infrared and near-infrared spectra in vegetable oils and digital images obtained from different types of commercial diesel. In order to validate the software results, the same sets of data were analysed using Matlab<sup>®</sup> and the results in both applications coincided in various combinations. In addition to the desktop version, the reuse of algorithms allowed to provide an online version that offers a unique experience on the web.

**Keywords:** software, chemometrics, exploratory data analysis.

## LISTA DE ILUSTRAÇÕES

Figura 1. Reflexão interna em um elemento de ATR .....	24
Figura 2. Reflexões internas em um acessório NIRA.....	25
Figura 3. Espectro eletromagnético, com destaque para as subdivisões da região de luz visível.....	26
Figura 4. Modelo de cor RGB.....	28
Figura 5. Imagem em escala de cinza e seu histograma. ....	29
Figura 6. (a) Modelo HSV. (b) Corte horizontal do modelo HSV. ....	29
Figura 7. Representação de um vetor e de uma matriz de dados.....	32
Figura 8. Espectro UV/Vis com ruído (a) e o mesmo após filtro Savitzky-Golay (b)..	36
Figura 9. Representação de um modelo de ligação simples. ....	38
Figura 10. Representação de um modelo de ligação completa.....	38
Figura 11. Representação de um modelo de ligação pela média.....	39
Figura 12. Representação da matriz de dados decomposta em produto de matrizes de posto 1.....	40
Figura 13. Uma componente principal no caso de duas variáveis onde ângulos representam os <i>loadings</i> (a) e as projeções das amostras representam os <i>scores</i> (b). .....	40
Figura 14. Placa de vidro e acessório de alumínio para a aquisição da leitura do branco. ....	49
Figura 15. a) Impressora multifuncional e papel-máscara, b) recipiente de vidro com amostra, c) tampa de plástico branca e d) tampa com fundo preto fosco para evitar entrada de luz.....	50
Figura 16. Tela de desenvolvimento da IDE Microsoft Visual Studio® 2010.....	51
Figura 17. Tela principal do ChemoStat - padrão espectroscopia.....	54
Figura 18. Tela principal do ChemoStat - padrão imagens. ....	55
Figura 19. Menu de operação - cabeçalho da tela principal.....	55
Figura 20. Janela de seleção de arquivos padrão Windows®.....	56
Figura 21. Tela principal padrão espectroscopia – identificação da seção 1.....	57
Figura 22. Detalhe da seção 2 na tela principal padrão espectroscopia. ....	58
Figura 23. Tela principal padrão espectroscopia com grade de dados – seção 3.....	58

Figura 24. Tela principal padrão espectroscopia com grade de dados – detalhe do menu de operações acionado. ....	59
Figura 25. Gráfico de espectros de óleos vegetais obtidos via espectroscopia no infravermelho próximo, sem tratamento de dados, a partir da função Plot 2D. ....	60
Figura 26. Menu de operações do gráfico via função Plot 2D. ....	60
Figura 27. Gráfico de espectros de óleos vegetais com opção “selecionar região” executada. ....	61
Figura 28. Grade de dados com rótulo das funções atribuídas. ....	62
Figura 29. Gráfico do conjunto de espectros dos óleos vegetais sem tratamento de dados na faixa entre 5500 e 6000 $\text{cm}^{-1}$ . ....	63
Figura 30. Menu principal de operações - funções de conversão. ....	64
Figura 31. Menu principal de operações - funções de espectro médio. ....	65
Figura 32. Menu principal de operações - funções de normalização. ....	66
Figura 33. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR) normalizados entre os limites zero e um de absorvância, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – ChemoStat. ....	67
Figura 34. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR) normalizados entre os limites zero e um de absorvância, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – Matlab®. ....	67
Figura 35. Menu principal de operações - funções de transformação. ....	68
Figura 36. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados e com aplicação de SNV, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – ChemoStat. ....	69
Figura 37. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados e com aplicação de SNV, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – Matlab®. ....	69
Figura 38. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com aplicação de SNV e primeira derivada (5 pontos), na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – ChemoStat. ....	70
Figura 39. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com aplicação de SNV e primeira derivada (5 pontos), na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – Matlab®. ....	70
Figura 40. Menu principal de operações - funções de pré-processamento. ....	71



Figura 41. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – ChemoStat.....	72
Figura 42. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – Matlab®.....	72
Figura 43. Tela para identificação das amostras por classe.....	73
Figura 44. Menu principal de operações - funções de pré-processamento para PCA. ....	74
Figura 45. Gráfico de <i>scores</i> PC1 x PC2 do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – ChemoStat. ....	75
Figura 46. Gráfico de <i>scores</i> PC1 x PC2 do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – Matlab®. ....	75
Figura 47. Gráfico de <i>loadings</i> (PC1) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – ChemoStat.....	77
Figura 48. Gráfico de <i>loadings</i> (PC1) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – Matlab®.....	77
Figura 49. Gráfico de <i>loadings</i> (PC2) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – ChemoStat.....	78
Figura 50. Gráfico de <i>loadings</i> (PC2) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – Matlab®.....	78
Figura 51. Gráfico <i>biplot</i> de <i>scores</i> e <i>loadings</i> (PC1 x PC2) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – ChemoStat. ....	79
Figura 52. Gráfico <i>biplot</i> de <i>scores</i> e <i>loadings</i> (PC1 x PC2) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – Matlab®.....	79

Figura 53. Janela de diálogo para entrada de valor referente ao <i>alpha</i> para distribuição de Fisher-Snedecor.....	80
Figura 54. Gráfico para $T^2$ de Hotelling (PC1 x PC2) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – ChemoStat. ....	80
Figura 55. Gráfico para $T^2$ de Hotelling (PC1 x PC2) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000 $\text{cm}^{-1}$ – Matlab®.....	81
Figura 56. Menu principal de operações - funções de pré-processamento para “PCA by region” . ....	82
Figura 57. Menu principal de operações - funções de pré-processamento para iPCA. ....	83
Figura 58. Gráfico de scores (PC1 x PC2) do conjunto de espectros dos óleos vegetais (FT-MIR), autoescalados e previamente normalizados, com SNV e primeira derivada, na faixa entre 650 e 4000 $\text{cm}^{-1}$ – ChemoStat. ....	84
Figura 59. Gráfico de scores (PC1 x PC2) do conjunto de espectros dos óleos vegetais (FT-MIR), autoescalados e previamente normalizados, com SNV e primeira derivada, na faixa entre 650 e 4000 $\text{cm}^{-1}$ – Matlab®.....	84
Figura 60. Caixa de diálogo para entrada de valores referente ao intervalo de iPCA. .	85
Figura 61. Janela com 32 gráficos de scores referente aos intervalos aplicados nos espectros de óleos vegetais (FT-MIR) - ChemoStat. ....	86
Figura 62. Janela com 4 gráficos de scores referente aos intervalos 25, 26, 27 e 28 aplicados nos espectros de óleos vegetais (FT-MIR) – ChemoStat.....	87
Figura 63. Janela com 4 gráficos de scores referente aos intervalos 25, 26, 27 e 28 aplicados nos espectros de óleos vegetais (FT-MIR) – Matlab®.....	87
Figura 64. Janela com o gráfico de scores referente ao intervalo 28 aplicados nos espectros de óleos vegetais (FT-MIR) - ChemoStat. ....	88
Figura 65. Janela com o gráfico de scores referente ao intervalo 28 aplicados nos espectros de óleos vegetais (FT-MIR) - Matlab®.....	88
Figura 66. Menu de opções sobre o gráfico de scores do método iPCA.....	89
Figura 67. Caixa de diálogo para entrada de valores referente ao número de componentes principais.....	89

Figura 68. Variação percentual das componentes principais divididas em 32 intervalos aplicados nos espectros de óleos vegetais (FT-MIR) - ChemoStat. ....	90
Figura 69. Variação percentual das componentes principais divididas em 32 intervalos aplicados nos espectros de óleos vegetais (FT-MIR) - Matlab® .....	90
Figura 70. Menu principal de operações – função HCA. ....	92
Figura 71. Dendrograma HCA – ligação completa – das amostras de óleos vegetais (FT- MIR), na região entre 1066 e 1169 cm <sup>-1</sup> - ChemoStat. ....	93
Figura 72. Dendrograma HCA – ligação completa – das amostras de óleos vegetais (FT- MIR), na região entre 1066-1169 cm <sup>-1</sup> - ChemoStat.....	93
Figura 73. Dendrograma HCA – ligação pela média – das amostras de óleos vegetais (FT-MIR), na região entre 1066-1169 cm <sup>-1</sup> - ChemoStat. ....	94
Figura 74. Dendrograma HCA – ligação pela média – das amostras de óleos vegetais (FT-MIR), na região entre 1066-1169 cm <sup>-1</sup> - Matlab® .....	95
Figura 75. Dendrograma HCA – ligação simples – das amostras de óleos vegetais (FT-MIR), na região entre 1066-1169 cm <sup>-1</sup> - Matlab® .....	96
Figura 76. Dendrograma HCA – ligação simples – das amostras de óleos vegetais (FT-MIR), na região entre 1066-1169 cm <sup>-1</sup> - Matlab® .....	96
Figura 77. Detalhe da seção 2 na tela principal padrão espectroscopia. ....	97
Figura 78. Tela principal padrão imagem com grade de dados – seção 3. ....	98
Figura 79. Tela principal padrão imagem com grade de dados – detalhe do menu de operações acionado. ....	99
Figura 80. Janela com gráficos de histograma e imagem de uma amostra de óleo diesel tipo S1800 escaneada. ....	100
Figura 81. Menu principal de operações - funções de exportação de dados. ....	100
Figura 82. Menu principal de operações - funções para imagem média. ....	101
Figura 83. Gráfico de <i>scores</i> PC1 x PC2 de histogramas R, G e B das médias das replicatas de óleos diesel escaneados – ChemoStat. ....	103
Figura 84. Gráfico de <i>scores</i> PC1 x PC2 de histogramas R, G e B das médias das replicatas de óleos diesel escaneados – Matlab® .....	103
Figura 85. Gráfico de <i>loadings</i> da PC1 de histogramas R, G e B das médias das replicatas de óleos diesel escaneados – ChemoStat. ....	104
Figura 86. Gráfico de <i>loadings</i> da PC1 de histogramas R, G e B das médias das replicatas de óleos diesel escaneados – Matlab® .....	104

Figura 87. Gráfico de <i>scores</i> PC1 x PC2 dos modelos de cores R, G, B, r%, g%, b%, S, V, L, I de óleos diesel escaneados – ChemoStat.....	106
Figura 88. Gráfico de <i>scores</i> PC1 x PC2 dos modelos de cores R, G, B, r%, g%, b%, S, V, L, I de óleos diesel escaneados – Matlab®.....	106
Figura 89. Gráfico de <i>loadings</i> da PC1 dos modelos de cores R, G, B, r%, g%, b%, S, V, L, I de óleos diesel escaneados – ChemoStat.....	107
Figura 90. Gráfico de <i>loadings</i> da PC1 dos modelos de cores R, G, B, r%, g%, b%, S, V, L, I de óleos diesel escaneados – Matlab®.....	107
Figura 91. Gráfico para $T^2$ de Hotelling PC1 x PC2 dos modelos de cores R, G, B, r%, g%, b%, S, V, L, I de óleos diesel escaneados – ChemoStat.....	108
Figura 92. Gráfico para $T^2$ de Hotelling PC1 x PC2 dos modelos de cores R, G, B, r%, g%, b%, S, V, L, I de óleos diesel escaneados – Matlab®.....	108
Figura 93. Dendrograma HCA – ligação completa – das amostras de óleos diesel escaneadas – Chemostat.....	109
Figura 94. Dendrograma HCA - ligação completadas - amostras de óleos diesel escaneadas – Matlab®.....	110
Figura 95. Tela de entrada - ChemoStat versão web.....	111
Figura 96. Tela de registro de usuários - ChemoStat versão web.....	112
Figura 97. Detalhe da mensagem de recuperação de senha - ChemoStat versão web.....	112
Figura 98. Tela principal para análise exploratória de dados - ChemoStat versão web.....	113
Figura 99. Caixa de diálogo para permissão de acesso à área de transferência. ...	114
Figura 100. Gráfico de <i>scores</i> (PC1 x PC2) para espectros de óleos vegetais (FT-MIR), na região entre 1066 e 1169 $\text{cm}^{-1}$ - ChemoStat versão web. ....	115
Figura 101. Gráfico de <i>loadings</i> (PC1) para espectros de óleos vegetais (FT-NIR), na região entre 5500 e 6000 $\text{cm}^{-1}$ - ChemoStat versão web. ....	116
Figura 102. Dendrograma HCA – ligação completa – para espectros de óleos vegetais (FT-MIR), na região entre 1066 e 1169 $\text{cm}^{-1}$ - ChemoStat versão web. ...	117

## LISTA DE TABELAS

Tabela 1. Composição de alguns óleos vegetais comercializados no Brasil.....	21
Tabela 2. Identificação e origem das amostras de óleos vegetais utilizadas no NIR e MIR.....	47
Tabela 3. Identificação e origem das amostras de óleo diesel utilizadas no NIR, MIR e escâner.....	48
Tabela 4. Valores para variância e variância acumulada das seis primeiras componentes principais dos dados de óleos vegetais - ChemoStat e Matlab®.....	91
Tabela 5. Valores para variância e variância acumulada das seis primeiras componentes principais dos dados das imagens de óleo diesel (histograma) - ChemoStat e Matlab®.....	102
Tabela 6. Valores para variância e variância acumulada das seis primeiras componentes principais dos dados das imagens de óleo diesel (modelos de cores) - ChemoStat e Matlab®.....	105

## LISTA DE ABREVIATURAS

ANP	Agência Nacional do Petróleo, Gás Natural e Combustível
ATR	Espectroscopia por reflexão total atenuada
B	Azul ( <i>blue</i> )
b%	Azul relativo
C#	C-Sharp
FT-IR	Espectroscopia no infravermelho por transformada de Fourier
G	Verde ( <i>green</i> )
g%	Verde relativo
GC-FID	Cromatografia gasosa com detector de ionização por chama
GHz	Giga-hertz
H	Matiz ( <i>hue</i> )
HCA	Análise por agrupamento hierárquico
I	Intensidade
IDE	Ambiente de desenvolvimento integrado
iPCA	Análise das componentes principais por intervalos
Mb	<i>Megabyte</i>
MCR	<i>Matlab<sup>®</sup> Compiler Runtime</i>
MIR	Infravermelho médio
MSC	Correção de espalhamento multiplicativo
NIR	Infravermelho próximo
NIRA	Acessório de infravermelho próximo
PCA	Análise das componentes principais
PCs	Componentes principais
R	Vermelho ( <i>red</i> )
r%	Vermelho relativo
RAM	Memória de acesso aleatório
S	Saturação
SNV	Varição Normal Padrão
V	Valor
L	Luminância
UATR	Refletância total atenuada universal

## SUMÁRIO

1	INTRODUÇÃO .....	18
2	OBJETIVOS .....	20
2.1	Objetivo geral .....	20
2.2	Objetivos específicos .....	20
3	REFERENCIAL TEÓRICO .....	21
3.1	Óleos vegetais .....	21
3.2	Óleo diesel .....	22
3.3	Espectroscopia no infravermelho .....	23
3.3.1	Medidas de refletância .....	23
3.3.2	Medidas de transreflectância .....	25
3.4	Processamento de imagens .....	26
3.5	Modelos de representação de cores .....	27
3.5.1	Modelo RGB .....	27
3.5.2	Histograma RGB .....	28
3.5.3	Modelo HSV .....	29
3.5.4	Brilho: intensidade, iluminação e luminância .....	30
3.6	Quimiometria .....	31
3.7	Tratamento dos dados .....	32
3.7.1	Primeira e segunda derivadas .....	33
3.7.2	Varição normal padrão (SNV) .....	33
3.7.3	Correção do espalhamento de luz (MSC) .....	33
3.7.4	Normalizações .....	34
3.7.5	Suavizações .....	35
3.8	Pré-processamentos dos dados .....	36
3.9	Análise por Agrupamento Hierárquico – HCA .....	37
3.10	Análise de Componentes Principais - PCA .....	39
3.11	Técnicas de otimização (seleção de variáveis) .....	40
3.12	Análise de Componentes Principais por intervalos – iPCA .....	41
3.13	Detecção de outlier - método $T^2$ de Hotelling .....	41
3.14	Aplicações da análise exploratória de dados .....	42
4	METODOLOGIA .....	46

4.1 Amostragem.....	47
4.1.1 Origem e identificação das amostras de óleos vegetais .....	47
4.1.2 Origem e identificação das amostras de óleos diesel .....	47
4.1.3 Obtenção dos espectros dos óleos vegetais.....	48
4.1.4 Obtenção das imagens das amostras de óleo diesel.....	49
4.2 Desenvolvimento e validação do <i>software</i> .....	50
4.3 Requisitos mínimos do <i>software</i> .....	52
5 RESULTADOS E DISCUSSÕES .....	54
5.1 Tela principal.....	54
5.2 Importação de dados espectrais .....	56
5.2.1 Menu de ferramentas .....	59
5.2.2 Função “Plot 2D” .....	59
5.2.3 Função “Export Excel/ Text” .....	61
5.2.4 Função “Export ASCII” .....	61
5.2.5 Função “Extract region” .....	62
5.2.6 Funções de conversões de unidades .....	63
5.2.7 Função espectro médio.....	64
5.2.8 Funções de normalização de espectros .....	65
5.2.9 Funções de transformações.....	68
5.2.10 Funções de pré-processamentos .....	71
5.2.11 Identificação de amostras .....	73
5.2.12 Algoritmo PCA .....	74
5.2.13 Função PCA por regiões (“by regions”) .....	81
5.2.14 Análise por Componentes Principais em intervalos – iPCA.....	82
5.2.15 Algoritmo HCA .....	92
5.3 Importação de dados de imagens.....	97
5.3.1 Menu de ferramentas .....	99
5.3.2 Função “Histogram” .....	99
5.3.3 Função “Export Excel/ Text” .....	100
5.3.4 Função “imagem média” .....	101
5.3.5 Identificação de amostras .....	101
5.3.6 Algoritmo PCA .....	101
5.3.7 Algoritmo HCA .....	109



5.4 Solução <i>online</i> .....	110
5.4.1 Tela de acesso e registro .....	111
5.4.2 Registro de usuários novos.....	111
5.4.3 Perda da senha de acesso .....	112
5.4.4 Tela principal para análise exploratória .....	113
5.5 Dados de outras origens.....	117
5.6 Perspectivas futuras.....	117
6 CONCLUSÕES .....	119
REFERÊNCIAS .....	120
ANEXO A: Programa “Plot 3D” .....	125
ANEXO B: Fórum de discussão sobre algoritmos da PCA.....	127
ANEXO C: Editor de imagens .....	130
ANEXO D: Imagens escaneadas das amostras de óleo diesel comercial.....	132
ANEXO E: Apêndices de publicações.....	137

## 1 INTRODUÇÃO

A palavra quimiometria surgiu na década de 1970 e seu desenvolvimento baseava-se na computação científica, envolvendo principalmente métodos estatísticos multivariados para dados da química analítica. Os primeiros quimiometristas eram necessariamente programadores Fortran ou Basic e utilizavam *mainframes* e bibliotecas estatísticas nas sub-rotinas. As aplicações quimiométricas iniciaram para conjuntos de dados analíticos químicos, por vezes simples, como numa cromatografia líquida de alta eficiência (HPLC), em conjuntos de dois ou três picos (BRERETON, 2009).

Na década de 1980 a quimiometria se organizou como uma disciplina, surgindo as primeiras publicações, associações e cursos dedicados ao tema. As aplicações industriais foram particularmente importantes nessa fase de seu desenvolvimento enquanto que a fronteira entre quimiometria e outras disciplinas tornava-se gradualmente estabelecida. Já na década de 1990, a aplicação de quimiometria começou a se expandir, especialmente na indústria farmacêutica. Desde então, devido à capacidade de instrumentos analíticos em adquirir grandes quantidades de dados rapidamente e o aumento da capacidade de processamento dos computadores, a quimiometria tornou-se uma ferramenta indispensável para mineração e análise de dados. Arelado ao avanço tecnológico e à demanda na área da pesquisa, muitos *softwares* comerciais surgiram (BRERETON, 2007).

Atualmente quimiometristas realizam parte do desenvolvimento de suas pesquisas utilizando Matlab® (The Mathworks, Natick, E.U.A.), ou uma versão equivalente livre chamada GNU Octave (<http://www.octave.org/>), que são ambientes flexíveis para computação matemática. No entanto, estes aplicativos exigem algum investimento em tempo para a familiarização e interpretação de suas sintaxes. Assim como o Fortran na década de 1980, estes aplicativos requerem conhecimentos de programação e não são necessariamente simples para pessoas com pouca experiência algorítmica. Outros programas específicos como Pirouette® (Infometrix, Bothell, E.U.A.), Unscrambler® (CAMO, Woodbridge, E.U.A.), Evince® (UmBio, Umea, Suécia) e ferramentas tipo *add-ins*, como PLS\_Toolbox® (Eigenvector Research, Wenatchee, E.U.A), devem necessariamente ser registradas mediante compra de licenças, inviabilizando, muitas vezes, seu uso acadêmico generalizado.

Recentemente surgiu o Chemoface<sup>®</sup>, um aplicativo gratuito e baseado no Matlab<sup>®</sup>, porém sem necessidade de licença, tendo como requisito principal a instalação do MCR (*Matlab Compiler Runtime*) (NUNES et al., 2012). A vantagem do uso deste compilador é a utilização em várias plataformas como Windows<sup>®</sup>, Linux<sup>®</sup> e Mac<sup>®</sup>. Como desvantagem, apresenta a própria dependência do MCR e seu suporte, do uso de uma grande capacidade de memória física (versão 8.2 possui 447 Mb), além da necessidade de privilégios de administrador do sistema operacional para instalação. Há ainda outros *softwares* da área estatística aplicada à biologia ou geografia, alguns gratuitos, outros baseados em linha de comando, entretanto desprovidos de alguns recursos específicos utilizados na quimiometria (JARVIS, 2006).

Neste sentido, buscou-se desenvolver um *software* de fácil adoção, instalação e manuseio, destinado a alunos, professores e pesquisadores, e que abrangesse, primeiramente, uma das áreas mais utilizadas na quimiometria: a análise exploratória de dados, além de uma solução *online* básica que abrangesse dispositivos móveis, como *tablets*, e outros sistemas operacionais.

## 2 OBJETIVOS

### 2.1 Objetivo geral

O objetivo principal deste trabalho foi desenvolver um *software* gratuito, de uso acadêmico, de fácil instalação e manuseio, sem necessidade de programação em nível de usuário, para análise exploratória de dados, além de uma solução *online* dotada de alguns recursos básicos da quimiometria.

### 2.2 Objetivos específicos

Os objetivos específicos foram:

- Desenvolver um *software*, também chamado de versão *desktop*, que contemple as técnicas de análise de agrupamento hierárquico (HCA), análise de componentes principais (PCA), análise de componentes principais por intervalos (iPCA), assim como, técnicas de correção, transformação dos dados e detecção de amostras anômalas (*outliers*).
- Adquirir espectros no infravermelho médio e próximo de diferentes óleos vegetais como conjunto de dados para avaliação e validação das ferramentas contempladas pelo *software*.
- Adquirir imagens digitais de diferentes óleos diesel comerciais e implementação de uma função para geração de histogramas e decomposição de pixels nos modelos de cores RGB, HSV, valores de iluminação e intensidade de brilho.
- Validar as ferramentas contempladas pelo *software* utilizando Matlab<sup>®</sup> versão 7.11 (The Mathworks Inc.).
- Desenvolver uma solução *online* com alguns recursos básicos de tratamento de dados além das técnicas de análise de agrupamento hierárquico (HCA) e análise de componentes principais (PCA).

### 3 REFERENCIAL TEÓRICO

O referencial teórico abordado neste trabalho está dividido em cinco tópicos principais. O primeiro e o segundo compreendem uma breve revisão sobre óleos vegetais e óleo diesel, matérias-primas utilizadas na construção dos conjuntos de dados para avaliação e validação do *software*. O terceiro e o quarto tópicos abordam a espectroscopia no infravermelho e o processamento de imagens digitais, técnicas utilizadas para aquisição do conjunto de dados. O quinto, e último tópico, estão relacionados aos métodos multivariados de análise contemplados pelo *software*.

#### 3.1 Óleos vegetais

Os óleos e gorduras, também chamados de lipídios, são substâncias insolúveis em água (hidrofóbicas), de origem animal, vegetal ou até mesmo microbiana. São formadas a partir da condensação entre glicerol e ácidos graxos, chamados triglicerídeos, e ácidos graxos livres, que chegam a representar até 96% do peso total dessas moléculas. São eles os principais combustíveis da maioria dos organismos e constituem, na verdade, uma das mais importantes formas de armazenamento de energia química. As unidades fundamentais da maioria dos lipídios são os ácidos graxos, que podem ser saturados (ligações simples) e insaturados (duplas ligações) (MORETTO & FETT, 1998; VISENTAINER & FRANCO, 2006). Zambiazzi et al. (2007), via cromatografia gasosa (GC-FID), identificaram a composição dos ácidos graxos em diversos óleos vegetais comercializados no Brasil, sendo os resultados apresentados na Tabela 1.

Tabela 1. Composição de alguns óleos vegetais comercializados no Brasil

Tipo de óleo	Saturado (%)	Monoinsaturado (%)	Poli-insaturado (%)
Canola	6,98	64,42	28,60
Girassol	12,36	15,93	71,71
Milho	13,87	24,76	61,37
Soja	15,10	21,73	63,17
Amendoim	18,38	50,33	31,29
Arroz	20,68	41,41	37,91
Algodão	25,73	17,49	56,78

Fonte: ZAMBIAZI et al., 2007.

### 3.2 Óleo diesel

O óleo diesel, ou gasóleo, é obtido através do refino do petróleo, sendo sua constituição formada principalmente por hidrocarbonetos e, em baixas concentrações, por enxofre, nitrogênio e oxigênio. É utilizado em motores de combustão interna e ignição por compressão, sendo empregados nas mais diversas aplicações, tais como: automóveis, furgões, ônibus, caminhões, embarcações marítimas, etc. O Brasil comercializa três tipos de diesel:

- Diesel S10: óleo com máximo teor de enxofre de 10 mg/kg.
- Diesel S500: óleo com máximo teor de enxofre de 500 mg/kg.
- Diesel S1800: óleo com máximo teor de enxofre de 1800 mg/kg (PETROBRÁS DISTRIBUIDORA, 2014).

A Resolução nº 6, de 16 de setembro de 2009, do Conselho Nacional de Política Energética (CNPE), Ministério de Minas e Energia, estabeleceu em cinco por cento, em volume, o percentual mínimo obrigatório de adição de biodiesel ao óleo diesel comercializado ao consumidor final, de acordo com o disposto no artigo 2º da Lei no 11.097, de 13 de janeiro de 2005. O uso de biodiesel favorece a agregação de valor às matérias-primas oleaginosas de origem nacional, além de benefícios em toda sua cadeia produtiva, como a geração de emprego, renda e o desenvolvimento da indústria nacional de bens e serviços. O biodiesel é uma matriz energética renovável e sua mistura no óleo diesel favorece a redução das emissões de gases responsáveis pelo efeito estufa. Além disso, possibilita a redução da importação de diesel derivado de petróleo, com efetivos ganhos na Balança Comercial (BRASIL, 2009).

A partir de 1º de julho de 2013, o óleo diesel S500 recebeu corante vermelho, ocorrendo a proibição da adição de corante ao óleo diesel S1800. Essa disposição consta na Agência Nacional do Petróleo, Gás Natural e Combustível (ANP), através da Resolução nº 65 de 2011 (RANP 65/11), que estabelece as especificações e as obrigações quanto ao controle da qualidade a serem cumpridas para todo o óleo diesel comercializado no território nacional. Ainda de acordo a resolução, o diesel S10 possui uma coloração incolor à amarelada enquanto que o S1800 uma tonalidade amarelo à alaranjada, podendo ainda variar para marrom (BRASIL, 2011).

### 3.3 Espectroscopia no infravermelho

É uma técnica que permite identificar uma amostra através da radiação infravermelha. Essa radiação ao ser absorvida causa alteração nos modos rotacionais e vibracionais das moléculas. A diferença entre a radiação emitida pela fonte e a radiação absorvida pela amostra é registrada por um detector, gerando um espectro de absorção no infravermelho (BARBOSA, 2007).

Assim como ocorre em outros tipos de absorção de energia que caracterizam um processo quantizado, as moléculas são excitadas para atingir um estado maior de energia quando absorvem radiação no infravermelho. Uma molécula absorve apenas determinadas frequências (energias) selecionadas de radiação, ocorrendo sua vibração por deformações axiais e/ou angulares (PAVIA, LAMPMAN & KRIZ, 2010).

Na região do infravermelho ocorre absorção pela maioria dos compostos orgânicos e inorgânicos que possuem ligações covalentes. A região do infravermelho próximo ( $4.000$  a  $14.290\text{ cm}^{-1}$ ) tem como características picos largos e de baixa intensidade, enquanto no infravermelho médio ( $200$  a  $4.000\text{ cm}^{-1}$ ) aparecem picos muito intensos e geralmente estreitos (PASQUINI, 2003).

A espectroscopia se destaca pela versatilidade e adaptabilidade para analisar amostras de natureza diferentes. Para as amostras em forma sólida, a medição pode ser realizada por refletância difusa. Já as amostras líquidas são medidas por meio de transmissão de radiação. Um caso intermediário são as amostras que são analisadas por transreflectância, em que parte da luz incidente é refletida sobre a amostra e a outra parte a atravessa, sendo refletida por um dispositivo transreflectante, concebido de tal modo que também delimita o caminho ótico (SANCHEZ, 2010).

#### 3.3.1 Medidas de refletância

A técnica de reflexão total atenuada (ATR) se caracteriza pelas múltiplas reflexões da radiação infravermelha que ocorrem no interior de cristais, de materiais com alto índice de refração como, por exemplo, o seleneto de zinco (ZnSe),

interagindo apenas com a amostra que estiver superficialmente no cristal (FERRÃO, 2001).

A Figura 1 ilustra o funcionamento de um elemento de ATR. Após sofrer a difração quando passa do meio  $n_1$  para o meio  $n_2$ , o feixe de infravermelho é direcionado para um cristal opticamente denso com alto índice de refração ( $n_2$ ). O cristal deve assegurar a reflexão interna desse feixe ao entrar em contato com a amostra ( $n_3$ ), de forma a permitir que este atravesse o cristal e seja medido pelo detector.

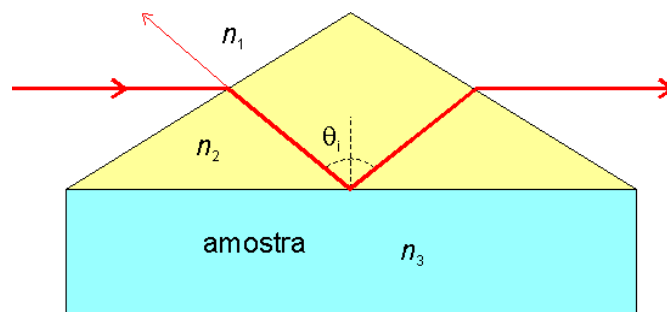


Figura 1. Reflexão interna em um elemento de ATR  
Fonte: Ferrão, 2001.

Essa refletância interna cria uma onda evanescente que se estende para além da superfície do cristal no interior da amostra mantida em contato com o cristal. Tal onda evanescente sobressai apenas poucos microns (0,5 a 5  $\mu\text{m}$ ) além da superfície cristalina e no interior da amostra. Em regiões do espectro de infravermelho onde a amostra absorve energia, a onda evanescente será atenuada ou alterada. A energia atenuada de cada onda evanescente retorna para o feixe de infravermelho, que então sai pela extremidade oposta do cristal e atinge o detector do espectrômetro, gerando o espectro de infravermelho (ALISKE, 2010).

O acessório de refletância atenuada universal - UATR (*Universal Attenuated Total Reflectance*), utilizado no infravermelho médio, oferece uma análise de fácil execução e limpeza para a maioria das amostras, como os óleos em geral, com um mínimo de preparação e alta reprodutibilidade (PERKIN-ELMER, 2010).



### 3.3.2 Medidas de transreflectância

Na transreflectância a radiação passa pela amostra duas vezes (ela passa uma vez ao penetrar na amostra e outra ao ser refletida), o que resulta em um espectro duas vezes mais intenso que o obtido pelo método de transmitância normal, através de um filme. Tanto o ângulo de incidência quanto o de reflexão sobre a superfície são importantes, uma vez que eles impactam a intensidade do espectro final obtido, e assim, quanto maior o ângulo de incidência, maior a intensidade da radiação refletida (BARBOSA, 2007).

A amostra, normalmente líquida ou semilíquida, é colocada num recipiente de vidro juntamente com uma superfície refletora. O feixe de radiação incidente entra no recipiente, passa através da amostra e é refletido no refletor. Assim, voltando novamente através da amostra, a radiação é lida pelo detector, conforme ilustra a Figura 2.

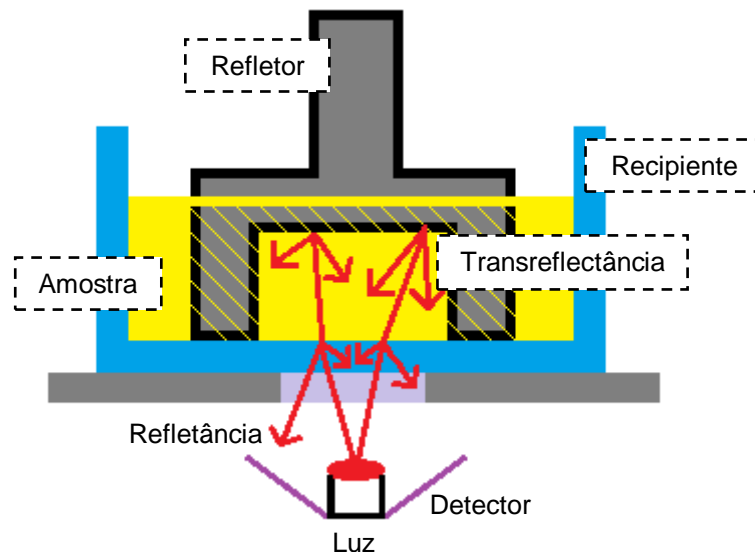


Figura 2. Reflexões internas em um acessório NIRA.  
Fonte: Autor, adaptado de Sanchez, 2010.

A espectroscopia no infravermelho próximo (NIR) com esfera de integração apresenta medida de refletância absoluta e eficiente combinando especular, promovendo uniformidade na detecção de amostras heterogêneas e redução dos efeitos de polarização oriundos do feixe de iluminação e da amostra. Possui um detector de Índio Gálio-Arsênio (InGaAs) que proporciona uma alta relação sinal-ruído (PERKIN-ELMER, 2010).

### 3.4 Processamento de imagens

O espectro de luz visível ocupa uma faixa muito estreita do espectro total de radiações eletromagnéticas (Figura 3). Para a cor ser vista, é necessário que o olho seja atingido por uma energia eletromagnética através da luz refletida por ele. A teoria de percepção cromática pelo olho humano baseia-se numa hipótese formulada por Young em 1801, que estabelece que os cones (células fotossensíveis que compõem a retina juntamente com os bastonetes) se subdividem em três classes, com diferentes máximos de sensibilidade situados em torno do vermelho (R, do inglês *red*), do verde (G, do inglês *green*) e do azul (B, do inglês *blue*). Assim, todas as sensações de cor percebidas pelo olho humano são combinações das intensidades dos estímulos recebidos por cada um destes tipos de cones (GONZALEZ & WOODS, 2008).

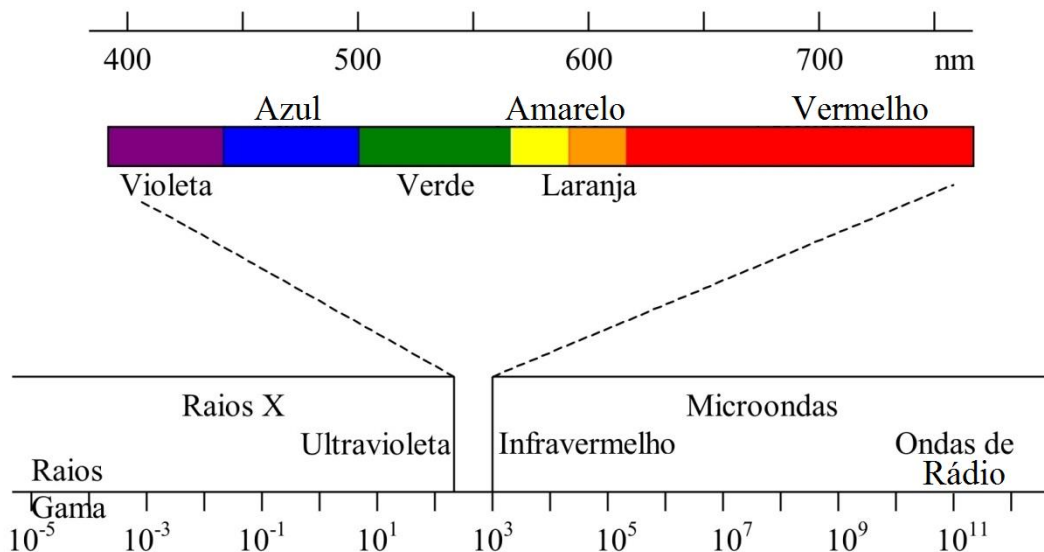


Figura 3. Espectro eletromagnético, com destaque para as subdivisões da região de luz visível. Fonte: Adaptado de Universidade Federal de Santa Catarina, 2014.

As cores RGB são denominadas cores primárias aditivas pois é possível obter qualquer outra cor a partir de uma combinação aditiva de uma ou mais delas, em diferentes proporções. A mistura das cores primárias, duas a duas, produz as chamadas cores secundárias, que são: magenta (R+B), amarelo (R+G) e ciano (G+B). A mistura das três cores primárias ou de uma secundária com sua cor

primária “oposta” produz a luz branca, e ao contrário, na subtrativa a união das três cores primárias ou de uma secundária com sua primária oposta produz o preto (MARQUES FILHO & VIEIRA NETO, 1999).

Os percentuais de vermelho, verde e azul, presentes em uma cor recebem o nome de coeficientes tricromáticos e são dados pelas equações 1, 2 e 3 (GONZALEZ & WOODS, 2008).

$$r = \frac{R}{R+G+B} \quad (1)$$

$$g = \frac{G}{R+G+B} \quad (2)$$

$$b = \frac{B}{R+G+B} \quad (3)$$

Onde R, G e B representam a quantidade de luz vermelha, verde e azul, respectivamente, normalizada entre 0 e 1. Logo, a soma dos três coeficientes tricromáticos é 1 (GONZALEZ & WOODS, 2008).

### 3.5 Modelos de representação de cores

#### 3.5.1 Modelo RGB

O mais comum sistema para imagens coloridas é o RGB. Neste espaço de cores, cada pixel é definido pelos valores de intensidade do vermelho (*Red*), verde (*Green*) e azul (*Blue*) na região do visível do espectro eletromagnético (ANTONELLI et al., 2004).

Para efeito de padronização, o CIE (*Commission Internationale de l'Eclairage*, do francês, Comissão Internacional de Iluminação) atribuiu, em 1931, os seguintes comprimentos de onda a estas cores primárias: azul = 435,8 nm, verde = 546,1 nm, vermelho = 700 nm (MARQUES FILHO & VIEIRA NETO, 1999).

O modelo RGB é baseado em um sistema de coordenadas cartesianas, que pode ser visto como um cubo onde três de seus vértices são as cores primárias, outros três as cores secundárias, o vértice junto à origem é o preto, e o mais

afastado da origem corresponde à cor branca, conforme ilustra a Figura 4. Neste modelo, a escala de cinza se estende através de uma linha (a diagonal do cubo) que sai da origem (preto) até o vértice mais distante dela (branco). Por conveniência, geralmente assume-se que os valores máximos de R, G e B estão normalizados na faixa de 0 a 1 (GONZALEZ & WOODS, 2008).

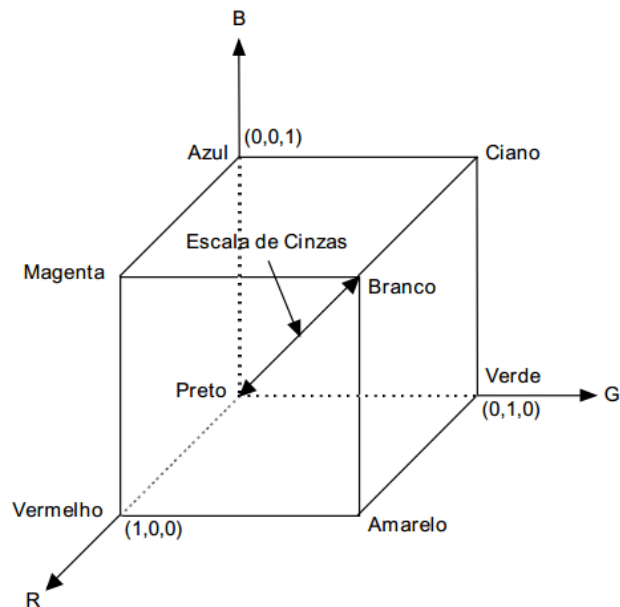


Figura 4. Modelo de cor RGB.  
Fonte: Gonzalez & Woods, 2008.

### 3.5.2 Histograma RGB

O histograma de uma imagem corresponde à distribuição dos níveis de cinza da mesma, os quais podem ser representados por um gráfico indicando o número de pixels na imagem para cada nível de cinza. Estes valores são normalmente representados por um gráfico de barras ou de distribuição de frequência. Através da visualização do histograma de uma imagem obtém-se uma indicação de sua qualidade quanto ao nível de contraste e quanto ao seu brilho médio (se a imagem é predominantemente clara ou escura) pelo número de vezes que o nível de cinza ocorre na imagem (PEDRINI & SCHARTZ, 2008).

A Figura 5 demonstra uma imagem em escala e seu histograma. Neste caso há uma baixa exposição de luz branca e o gráfico do histograma tende à esquerda (intensidades mais escuras) (BURGER & BURGE, 2009).

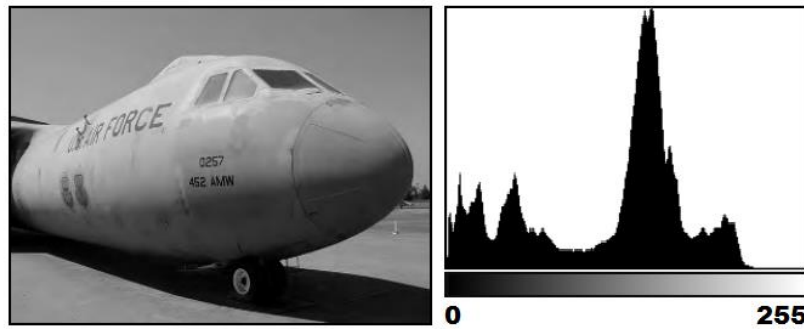


Figura 5. Imagem em escala de cinza e seu histograma.  
Fonte: Burger & Burge, 2009.

### 3.5.3 Modelo HSV

Quando se referencia imagens RGB, o modelo HSV permite separar as componentes de matiz, saturação e valor (luminância) da informação de cor em uma imagem, da forma como o ser humano as percebe. Sua utilização é mais intensa em sistemas de visão artificial fortemente baseado no modelo de percepção de cor pelo ser humano, como por exemplo, um sistema automatizado de colheita de frutas, em que é preciso determinar se a fruta está suficientemente madura para ser colhida a partir de sua coloração externa (ANTONELLI et al., 2004).

Geometricamente, o modelo HSV pode ser visto como um sólido, indicado na Figura 6 (a), cujos cortes horizontais produzem triângulos, Figura 6 (b), nos quais os vértices contêm as cores primárias e o centro corresponde à combinação destas cores em iguais proporções. Esta combinação estará mais próxima do preto ou do branco, conforme a altura em que o corte tenha sido efetuado (BURGER & BURGE, 2009).

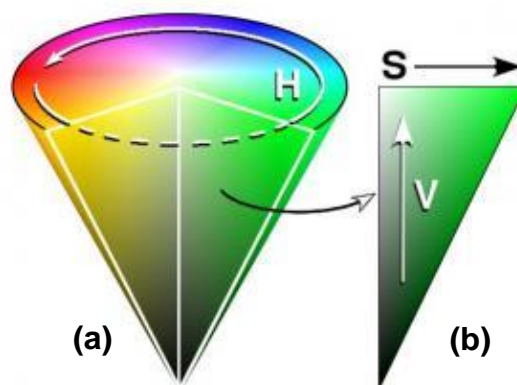


Figura 6. (a) Modelo HSV. (b) Corte horizontal do modelo HSV.  
Fonte: Autor, adaptado de Burger & Burge, 2009.

Matiz (H, do inglês, *Hue*) é um atributo que descreve a cor pura como o amarelo, laranja ou vermelho, compreendido entre um valor normalizado de 0 e 1, relativo à 360°. A saturação (S, do inglês, *Saturation*) é uma medida de quanto uma cor pura é diluída por uma luz branca, normalizados entre 0 e 1, e componente valor (V, do inglês, *Value*) refere-se ao brilho da cor, também normalizados entre 0 e 1 (GONZALEZ & WOODS, 2008).

### 3.5.4 Brilho: intensidade, iluminação e luminância

Dependendo da finalidade e objetivos da representação, existem várias possibilidades de obtenção do valor que representa o brilho. A definição mais simples é encontrada no modelo HSI, onde I representa intensidade (*intensity*), aplicando-se apenas a média dos três componentes RGB. Na teoria, consiste na projeção de um ponto sobre o eixo neutro ou a altura vertical de um ponto no cubo RGB inclinado (HANBURY, 2007).

No modelo de HSV o V indica valor (*value*) e é definido como o maior componente de uma cor RGB. Isto coloca todas as três cores primárias, e também todas as "cores secundárias" - ciano, amarelo e magenta - em um mesmo plano com o branco, formando uma pirâmide hexagonal fora do cubo RGB (SMITH, 1978).

Já no modelo de HSL, a iluminação é definida como a média dos maiores e menores de componentes de cor RGB. Esta definição também coloca as cores primárias e secundárias em um mesmo plano, mas num plano que passa no meio do caminho entre o branco e o preto (AGOSTON, 2005).

A conversão entre os modelos RGB e HSV, e os componentes I e L pode ser obtida através das equações 4, 5, 6, 7 e 8.

$$H = \begin{cases} 60 \times \frac{(G-B)}{(M-m)}, & \text{se } M = R \\ 60 \times \frac{(B-R)}{(M-m)} + 120, & \text{se } M = G \\ 60 \times \frac{(R-G)}{(M-m)} + 240, & \text{se } M = B \end{cases} \quad (4)$$

$$S = \begin{cases} \frac{(M-m)}{M}, & \text{se } M \neq 0 \\ 0, & \text{se } M = 0 \end{cases} \quad (5)$$

$$V = M \quad (6)$$

$$I = \frac{(R+G+B)}{3} \quad (7)$$

$$L = \frac{M+m}{2} \quad (8)$$

Onde  $M = \max(R, G, B)$  e  $m = \min(R, G, B)$ . Os valores de R, G e B devem estar normalizados entre zero e um.

### 3.6 Quimiometria

A quimiometria não é uma disciplina da matemática, da estatística ou da computação, mas sim da química. Os problemas que ela se propõe a solucionar são de interesse e originados na química, ainda que as ferramentas de trabalho provenham principalmente da matemática, estatística e computação. Informações químicas tais como, voltamogramas, espectros, cromatogramas, curvas de titulação e outras fontes podem ser digitalizadas e agrupadas em vetores e matrizes (TEÓFILO, 2013).

Segundo Ferreira et al. (1999), a quimiometria é frequentemente utilizada para maximizar as informações de um conjunto de dados, discretos ou instrumentais, extraídos de matrizes multivariadas, como por exemplo, informações provenientes da espectroscopia.

Outra abordagem da quimiometria é baseada na análise multivariada de imagens a partir uma imagem digital de uma dada cena, pela extração dos elementos de figura, chamado pixels, onde cada pixel é caracterizado por uma série de variáveis espectrais, ou também chamados canais (GELADI et al., 1992).

Os dados multivariados, tanto de informações químicas ou a partir de imagens, geralmente correspondem a uma matriz  $X$  de valores, correspondendo a  $m$  variáveis para  $n$  amostras, conforme ilustra Figura 7.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & \cdots & x_{nm} \end{bmatrix}$$

Figura 7. Representação de um vetor e de uma matriz de dados.  
Fonte: Autor, adaptado de Ferreira et al., 1999.

Quando da aquisição dos espectros por técnicas de reflexão, alguns fenômenos não desejados podem ocorrer, como por exemplo ruídos das mais diversas origens, sejam provocados pela não homogeneidade da amostra, sejam os que acompanham os sinais analíticos obtidos por técnicas instrumentais. Na tentativa de eliminação ou minimização desses ruídos, que podem dificultar a interpretação dos dados, podem ser empregadas técnicas de tratamento do espectro como transformações (técnicas da primeira e segunda derivada, algoritmo de Savitzky-Golay, correção do espalhamento de luz, variação normal padrão), normalizações e pré-processamentos (FERRÃO, 2000).

Já na aquisição de imagens, a utilização da média ou mediana de um conjunto de pixels, busca, de certa forma, atenuar ruídos do tipo impulsivo que possam ocorrer (GONZALEZ & WOODS, 2008).

### 3.7 Tratamento dos dados

Refere-se à transformação dos mesmos com objetivo de distribuí-los adequadamente, possibilitando a extração de informações úteis e facilitando a interpretação. Diversos tipos de tratamentos de dados podem ser aplicados aos dados originais antes de realizar alguma análise exploratória, pois a distribuição dos mesmos pode não ser adequada para a extração das informações (FERREIRA et al., 1999).

Diversos tipos de tratamentos de dados podem ser aplicados na espectroscopia como: primeira e segunda derivadas, variação normal padrão (SNV), correção do espalhamento de luz (MSC), normalizações e suavizações. Já para as



imagens foram desenvolvidos métodos de agrupamento utilizando a média e a mediana de pixels.

### 3.7.1 Primeira e segunda derivadas

A primeira e a segunda derivadas são transformações de alisamento baseadas em um filtro polinomial de Savitzky-Golay. Este método aplica uma convolução para as variáveis independentes em uma janela contendo um ponto central de dados e  $n$  pontos de cada lado. Um polinômio de segunda ordem ponderado é ajustado a esses  $2n + 1$  pontos onde o ponto central é substituído pelo valor calculado (INFOMETRIX, 2011).

### 3.7.2 Variação normal padrão (SNV)

A variação normal padrão (SNV, do inglês, *Standard Normal Variate*) é outra abordagem para compensar o espalhamento da luz pela amostra, muito utilizada em espectrometria NIR. Numa matriz, pode ser descrito como um escalonamento de linha. A média e o desvio padrão de uma amostra são primeiramente calculados com base nas variáveis espectrais. Após, o valor para cada variável é corrigida subtraindo-se a média e, em seguida, dividindo-se pelo desvio padrão (eq.9). O resultado é muitas vezes semelhante ao MSC (CAMO, 2006).

$$f(x) = \frac{x_i - \text{média}(x)}{\text{desvio\_padrão}(x)} \quad (9)$$

### 3.7.3 Correção do espalhamento de luz (MSC)

A correção do espalhamento de luz (MSC, do inglês, *Multiplicative Scatter Correction*) é uma abordagem padrão para compensar espalhamento da luz presente nos espectros obtidos por técnicas de reflexão. Cada espectro da amostra original é regredido linearmente para proporcionar uma equação de reta (eq.10). O espectro da amostra é em seguida corrigido em cada comprimento de onda,

primeiramente subtraindo-o pelo coeficiente linear (intercessão) e, em seguida, realizando a divisão pelo coeficiente angular (inclinação) (eq. 11) (FURTADO, 2002).

$$y = a * x + b \quad (10)$$

$$f(x) = \frac{x_i - a}{b} \quad (11)$$

### 3.7.4 Normalizações

As normalizações consistem da divisão das variáveis espectrais por uma constante, pelo valor máximo espectral, ou mesmo defini-lo numa margem variando de zero a um. Elas são geralmente empregadas quando os dados derivam de instrumentos diferentes, ou quando replicatas possuem variação em sua magnitude por alteração na linha de base. Uma maneira fácil de colocar essas medidas em uma escala comparável é subtrair cada variável pelo valor mínimo amostral, dividindo pela faixa de valores que compreendem a amostra, chamado de correção 1-0 (eq. 12) (INFOMETRIX, 2011; CAMO, 2006).

$$f(x) = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (12)$$

Dividindo-se pelo valor máximo da amostra, escalam-se os dados de uma forma típica da espectrometria de massa, onde é dado um valor de 100% ao fragmento de massa mais abundante (eq. 13) (INFOMETRIX, 2011).

$$f(x) = \frac{x_i}{\max(x)} \quad (13)$$

Às vezes é desejável se obter uma compensação na escala dos dados, principalmente quando os valores das variáveis são extremamente grandes ou pequenas, por exemplo, quando a aplicação da segunda derivada atenua a magnitude dos dados de forma significativa, ou para os dados de RMN de alguns instrumentos cujos valores encontram-se na ordem dos milhões. Neste caso,

dividem-se as variáveis por uma constante menor que zero (aumentando a escala) ou maior que zero (diminuindo a escala) (eq. 14) (INFOMETRIX, 2011).

$$f(x) = \frac{x_i}{K} \quad (14)$$

### 3.7.5 Suavizações

Muitos tipos de dados químicos consistem em série sequenciais de ruídos. Misturados no interior dos ruídos estão os sinais, tais como os picos cromatográficos ou espectroscópicos, e a informação sobre a qualidade de um produto fabricado ou concentrações de um composto. Uma dos principais requisitos da quimiometria é a obtenção de um sinal tão informativo quanto possível após a remoção desse ruído. Uma técnica importante envolve a suavização ou alisamento dos dados, entretanto seu uso demasiado pode reduzir o sinal em intensidade e resolução (BRERETON, 2007).

Os métodos mais simples envolvem filtros lineares em que os dados suavizados resultantes são uma função linear dos dados brutos. Uma dos métodos mais clássicos de alisamento é a média móvel (M.A., do inglês *Moving Average*), que substitui cada observação, ou ponto, com uma média das observações adjacentes (incluindo o próprio). Quanto mais pontos, mais suave o sinal se torna, maior a redução de ruído, porém, maior a chance de “borrar” o sinal. O número de pontos no filtro é muitas vezes chamado de "janela" (GLASBEY & HORGAN, 1995).

Filtros de média móvel têm a desvantagem de que eles exigem que os dados sejam aproximados por uma linha reta, porém picos de espectro, por exemplo, são muitas vezes definidos por curvas e obtém uma melhor aproximação através de polinômios (JONSSON, 2011).

Como a janela é movida ao longo dos dados, uma nova curva de melhor ajuste é calculada para cada intervalo. Este processo é realizado pelo filtro Savitzky-Golay, desenvolvido em 1964, o qual baseia-se na realização de uma regressão de mínimos quadrados linear através de um ajuste polinomial em torno de cada ponto do espectro. Trata-se de um método muito útil para remover eficazmente picos de

ruído espectral, enquanto a informação química pode ser mantida, como mostram as Figuras 8 – (a) e (b) (CAMO, 2006; JONSSON, 2011).

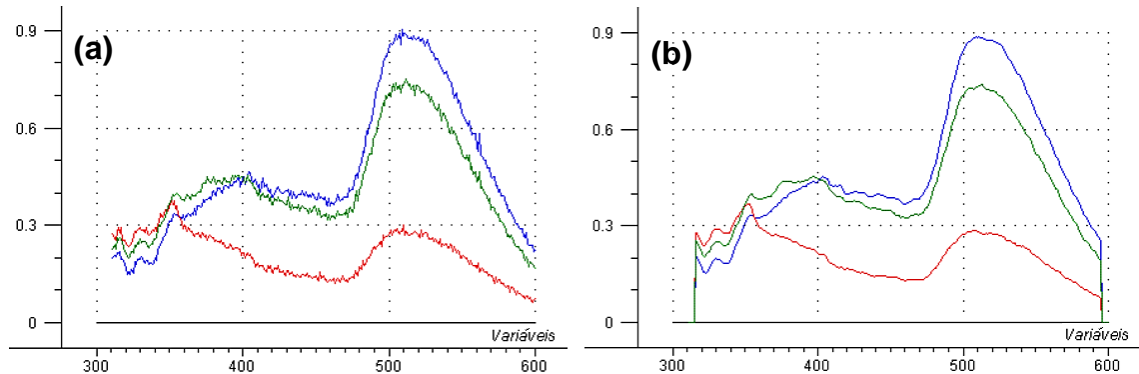


Figura 8. Espectro UV/Vis com ruído (a) e o mesmo após filtro Savitzky-Golay (b).  
Fonte: Camo, 2006.

### 3.8 Pré-processamentos dos dados

Pré-processamento, diferente das transformações, é uma operação orientada entre as variáveis para um conjunto de amostras de duas ou mais amostras, permitindo compará-las em diferentes dimensões. Numa matriz, pode ser descrito como um escalonamento de coluna. A adição de uma amostra a um conjunto de dados pode influenciar grandemente no efeito dessa técnica de pré-processamento. O pré-processamento é necessário porque vários algoritmos multivariados calculam resultados impulsionados por padrões de variância nas variáveis independentes (INFOMETRIX, 2011).

Muitas vezes variáveis com unidades distintas e diferentes variâncias podem produzir resultados enganosos quando um pré-processamento inadequado é realizado (CAMO, 2006).

Dentre os pré-processamentos mais utilizados em análise multivariada destacam-se o processo de autoescalar os dados e processo de centrá-los na média (BRO & SMILDE, 2001). Centrar os dados na média consiste em subtrair cada intensidade pelo respectivo valor médio para cada comprimento de onda (eq. 15). Já autoescalar os dados, representa centrar os dados na média e dividir pelo respectivo desvio padrão (eq. 16) (MATOS et al., 2003).

$$f(x) = x_i - \bar{x} \quad (15)$$

$$f(x) = \frac{x_i - \bar{x}}{\sigma(x)} \quad (16)$$

### 3.9 Análise por Agrupamento Hierárquico – HCA

As técnicas hierárquicas aglomerativas partem do princípio de que no início do processo de agrupamento tem-se  $n$  grupos, ou seja, cada elemento do conjunto de dados observado é considerado como sendo um objeto isolado (MINGOTI, 2005).

Os grupos não são conhecidos antes da análise matemática e não são realizadas hipóteses sobre a distribuição das variáveis. A análise de *cluster* procura objetos que estão próximos uns dos outros no espaço variável, ou seja, numa distância entre dois pontos no espaço  $n$ -dimensional (MILLER & MILLER, 2010).

Entre as medidas mais empregadas para estabelecer a relação de distância entre duas amostras, encontram-se a correlação de Pearson, a distância Manhattan e, em destaque, a distância de Euclidiana (eq.17).

$$D_{ab} = \sqrt{\sum_{i=1}^m (x_{ai} - x_{bi})^2} \quad (17)$$

Para que os agrupamentos sejam efetuados define-se matematicamente o conceito de similaridade (eq.18), onde  $D_{ab}$  é igual a distância entre as amostras  $a$  e  $b$  e  $D_{max}$  a maior distância no conjunto de dados (TRINDADE et al, 2005).

$$Similaridade_{ab} = 1 - \frac{D_{ab}}{D_{max}} \quad (18)$$

Posteriormente, existem várias possibilidades para determinar a formação dos agrupamentos. Uma opção é tomar a menor distância entre dois objetos a cada interação, ilustrado pelo exemplo da Figura 9, chamado de método de ligação simples - *Single Linkage* (WEHRENS, 2011).

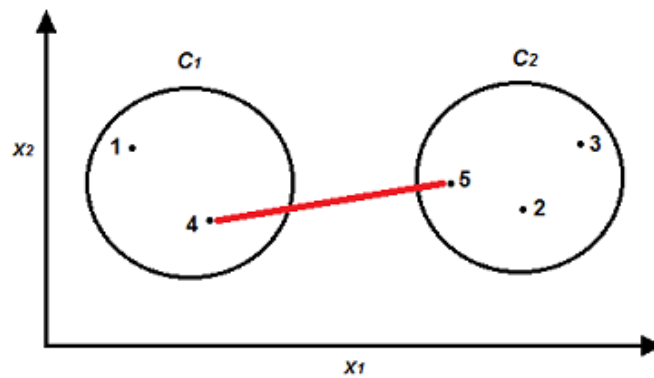


Figura 9. Representação de um modelo de ligação simples.  
 Fonte: Autor, adaptado de Wehrens, 2011.

A estratégia oposta, determinada pelos objetos no respectivo *cluster* que estão mais afastados, demonstrada pelo exemplo da Figura 10, também pode ser realizada e é definida pelo método de Ligação Completa - *Complete Linkage* (WEHRENS, 2011).

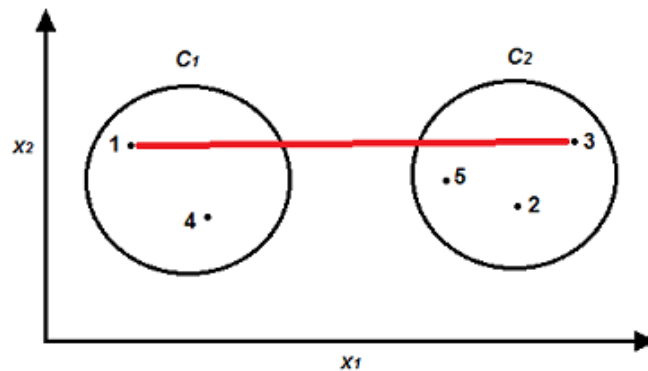


Figura 10. Representação de um modelo de ligação completa.  
 Fonte: Autor, adaptado de Wehrens, 2011.

Outro método também utilizado para agrupamentos é o da Ligação pela Média (*Average Linkage*). Neste caso, a distância entre dois objetos é tratada como a média das distâncias entre todos os pares de elementos que podem ser formados por cada um destes objetos, como demonstra a Figura 11 (WEHRENS, 2011).

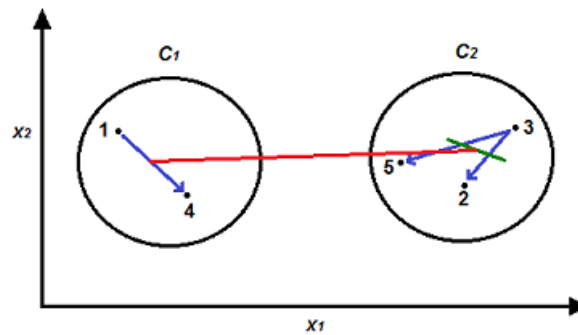


Figura 11. Representação de um modelo de ligação pela média.  
 Fonte: Autor, adaptado de Wehrens, 2011.

### 3.10 Análise de Componentes Principais - PCA

A Análise de Componentes Principais (PCA, do inglês, *Principal Component Analysis*), ou expansão Karhunen-Loeve, é um método clássico para redução de dimensionalidade ou análise exploratória de dados (SMITH, 2002).

A PCA é base fundamental da maioria dos métodos modernos para tratamento de dados multivariados e consiste numa manipulação da matriz de dados com objetivo de representar as variações presentes em muitas variáveis, através de um número menor de "fatores". Constrói-se um novo sistema de eixos (denominados rotineiramente de fatores, componentes principais, variáveis latentes ou ainda autovetores) para representar as amostras, no qual a natureza multivariada dos dados pode ser visualizada em poucas dimensões (FERREIRA et al., 1999).

A análise de fatores é realizada sobre uma matriz de dados que relaciona um conjunto de variáveis a diversos experimentos (amostras). Esta matriz de dados  $A$  pode ser centrada na média ou escalonada, sendo decomposta no produto de três matrizes através do algoritmo de decomposição de valor singular (SVD) (SMITH, 2012) (eq.19).

$$A=U \times S \times V' \quad (19)$$

A matriz  $V$  (transposta) é denominada de matriz dos *loadings*, as colunas da matriz  $V$  correspondem aos autovetores e  $S$  é uma matriz diagonal. As matrizes  $U$  e  $V$  são ortogonais entre si e o produto entre matrizes  $U$  e  $S$  é denominado de *scores*. Como resultado da análise de componentes principais o conjunto de dados originais

é agrupado em função da correlação existente entre as variáveis gerando um novo conjunto de eixos (componentes principais) ortogonais entre si e de mais simples manipulação matemática (FERRÃO, 2000).

Na Figura 12 tem-se a representação da matriz de dados decomposta em matrizes de *loadings* e *scores*.

$$\begin{matrix} & m \\ \boxed{\mathbf{A}} & \\ n & \end{matrix} = \begin{matrix} & 1 \\ \boxed{\mathbf{US}_1} & \\ n & \end{matrix} \begin{matrix} & m \\ \boxed{\mathbf{V}'_1} & \\ 1 & \end{matrix} + \begin{matrix} & 1 \\ \boxed{\mathbf{US}_2} & \\ n & \end{matrix} \begin{matrix} & m \\ \boxed{\mathbf{V}'_2} & \\ 1 & \end{matrix} + \dots + \begin{matrix} & 1 \\ \boxed{\mathbf{US}_r} & \\ n & \end{matrix} \begin{matrix} & m \\ \boxed{\mathbf{V}'_r} & \\ 1 & \end{matrix}$$

Figura 12. Representação da matriz de dados decomposta em produto de matrizes de posto 1. Fonte: Autor, adaptado de Laqqa, 2006.

Para exemplificar  $US_h$  e  $V'_h$ , a Figura 13 ilustra no plano bidimensional duas variáveis  $x_1$  e  $x_2$ . A Figura 13 (a) mostra uma componente principal que é a reta que aponta para a direção de maior variabilidade das amostras da Figura 13 (b). Os *scores*  $US_h$  são as projeções das amostras na direção da componente principal e os  $V'_h$  *loadings* são os cossenos dos ângulos formados entre a componente principal e cada variável (LAQQA, 2006).

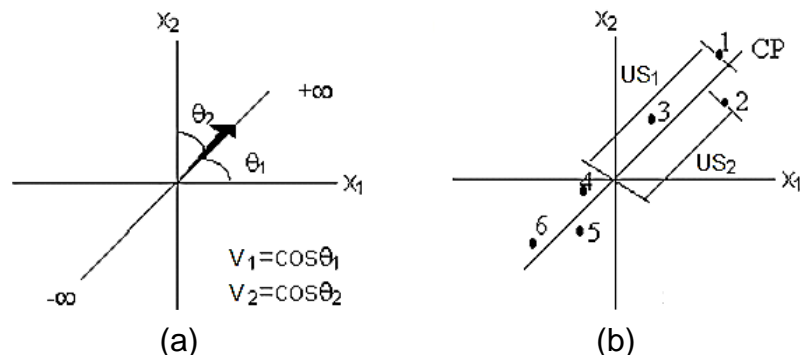


Figura 13. Uma componente principal no caso de duas variáveis onde ângulos representam os *loadings* (a) e as projeções das amostras representam os *scores* (b). Fonte: Autor, adaptado de Laqqa, 2006.

### 3.11 Técnicas de otimização (seleção de variáveis)

Técnicas de otimização são procedimentos de seleção que permitem eliminar os termos que não são relevantes na modelagem. Isso gera um subconjunto com o melhor número de variáveis, e que apresente maior sensibilidade e linearidade para



o(s) analito(s) de interesse, minimizando e até eliminando, desta forma, características potenciais dos interferentes, bem como não linearidades (COSTA FILHO & POPPI, 2002).

Dentre as técnicas utilizadas para seleção de variáveis, pode-se destacar o iPCA (PEREIRA et al., 2008).

### 3.12 Análise de Componentes Principais por intervalos – iPCA

O método iPCA é uma derivação do PCA, onde o espectro representado pelo conjunto de dados é dividido em um número de intervalos equidistantes. Em cada intervalo é realizado um modelo PCA apresentando resultados em gráficos de scores. Este método é utilizado para dar uma visão geral dos dados sendo útil na interpretação de quais regiões selecionadas são mais representativas na construção de um bom modelo de calibração multivariada (LEARDI e NORGAARD, 2004).

Norgaard et al. (2000) advertem sobre os tamanhos dos intervalos. Caso o tamanho for muito grande, existe uma probabilidade maior de englobar regiões que não são representativas para o problema. O fato inverso ocorre quando se utilizam intervalos pequenos, ou seja, suprimir as informações necessárias para prever adequadamente a propriedade de interesse.

A iPCA é considerada uma técnica de otimização pois permite identificar regiões de maior sensibilidade e linearidade para os analitos de interesse (COSTA FILHO & POPPI, 2002).

### 3.13 Detecção de outlier - método $T^2$ de Hotelling

O método  $T^2$  de Hotelling foi baseado na generalização da estatística  $t$  de Student para o caso multivariado, de acordo com as estimativas amostrais das matrizes de covariâncias. É um dos métodos que mede a variação dentro do modelo PCA, podendo identificar possíveis *outliers*.  $T^2$  é a soma das pontuações quadradas (eq. 20) (TAVARES, 2003).

$$T^2 = (X_k - \bar{X}) \cdot S^{-1} \cdot (X_k - \bar{X})^t \quad (20)$$

Onde  $X_k$  é o valor da componente principal analisada, geralmente duas (PC1 e PC2) de uma amostra,  $\bar{X}$  a média dessas componentes principais,  $S^{-1}$  o inverso da matriz de covariância das duas componentes principais envolvidas no cálculo, e  $t$  a matriz transposta.

Se algum ponto cair fora dos limites de um intervalo de confiança específico ( $\alpha=95\%$ , por exemplo), este ponto pode ser considerado um *outlier*, ou seja, não ser representativo no conjunto de dados do modelo PCA, caso o número de componentes exceda a dois. (VINZI et. al, 2010). Limites estatísticos podem ser desenvolvidos para os *scores* de uma amostra (eq. 21 e eq.22) (GORAYEB, 2010)

$$LIC = 0 \quad (21)$$

$$LSC = \frac{(m-1)^2}{m} \times \frac{(p/(m-p-1))F(\alpha/2;p;m-p-1)}{1+(p/(m-p-1))F(\alpha/2;p;m-p-1)} \quad (22)$$

Onde  $p$  é a quantidade de componentes principais envolvidas no momento, geralmente duas (PC1 e PC2),  $m$  a quantidade de amostras e  $F(\alpha/2;p;m-p-1)$  significa o percentil da distribuição de Fisher-Snedecor (F) com  $p$  e  $m-p-1$  graus de liberdade.

### 3.14 Aplicações da análise exploratória de dados

Panero, F. S. et al (2009) utilizaram a análise exploratória na discriminação geográfica do quiabo do Rio Grande do Norte, nos municípios de Ceará-Mirim, Macaíba e Extremoz, e de Pernambuco, nos municípios de Caruaru e Vitória de Santo Antão. Para tanto, foram determinados via Espectrofotômetro de Absorção Atômica, os seguintes metais: Cu, Zn, Na, Fe, K, Ca, Mn e Mg. Os dados obtidos foram submetidos a análise de componentes principais - PCA e análise de agrupamentos hierárquicos – HCA. Na PCA verificou-se que as duas primeiras componentes descreveram 83,27% da variação total dos dados, agrupando as amostras conforme sua região. Já a HCA, confirmou os resultados da PCA, discriminando também geograficamente as amostras. Ambos os métodos, comprovaram neste estudo, que análise exploratória de dados permite a obtenção

de informações rápidas e eficientes sobre a similaridade entre as amostras pela visualização gráfica.

Bicudo, Pinto & Cyrino (2010) propuseram realizar o agrupamento de alimentos de acordo com o perfil de aminoácidos essenciais, determinando quais mostram perfis mais próximos do requerimento da tilápia do Nilo (*Oreochromis niloticus*), e estudar a relação entre os aminoácidos dentro dos agrupamentos obtidos. Neste trabalho, a busca de uma ração balanceada, que proporcione maior crescimento aos peixes, passa pela escolha adequada das fontes proteicas disponíveis. Foram utilizadas, então, composições de aminoácidos em relação ao conteúdo de lisina, de 40 alimentos muito utilizados como ingredientes na formulação de dietas para peixes. Os ingredientes foram agrupados de acordo com o perfil de aminoácidos utilizando a análise de agrupamento por meio da distância Euclidiana, enquanto a análise de componentes principais foi utilizada para determinar a relação entre os aminoácidos em cada grupo obtido. Três grupos de ingredientes foram formados e apenas dois ingredientes, sorgo baixo tanino e farelo de glúten de milho 60%, não entraram em nenhum dos três grupos. A análise de componentes principais conseguiu resumir e explicar 75% da variância total com apenas três componentes principais.

Müeller et al. (2011) indicaram que é possível empregar espectroscopia no infravermelho médio com acessório de refletância total atenuada (ATR) associada à análise de componentes principais (PCA) na classificação e diferenciação de óleos vegetais. Foram utilizados os óleos vegetais de soja, milho, canola, girassol, arroz e azeite de oliva. A partir destes óleos foram obtidas misturas binárias (blendas) em proporção de 10, 20, 40, 60, 80 e 100% m/m totalizando 1 g de amostra. Um total de 64 blendas foi preparado, sendo que todas as amostras de óleos vegetais foram misturadas entre si. Neste caso, a PCA mostrou que com três componentes principais é possível descrever 75,75% da variância dos dados. Outra evidência observada é a influência da quantidade de amostra presente na blenda. Blendas que apresentam 60% ou mais de um determinado óleo vegetal tendem a se aproximar do óleo vegetal de origem.

Soares et al. (2011) demonstraram que a análise de componentes principais (PCA) é uma ferramenta quimiométrica adequada para classificar misturas de biodiesel com soja crua. Foram adulteradas com óleo de soja cru três origens

diferentes de biodiesel (algodão, mamona e palma), em concentrações variando de 1 a 40% (m/m). As amostras foram analisadas por espectrometria de infravermelho médio (MIR) e os seus espectros foram estudados em três diferentes faixas espectrais: espectro inteiro ( $4000-665\text{ cm}^{-1}$ ), e nas faixas de  $1800-1700\text{ cm}^{-1}$  e  $1800-1000\text{ cm}^{-1}$ . Utilizando a PCA para determinar a origem do biodiesel utilizado no sistema adulterado, a melhor segregação das origens foi obtida para o espectro inteiro com uma variância explicada de 99% para os três primeiros componentes.

Já Viera et al. (2010) avaliaram a adulteração de misturas biodiesel diesel empregando espectroscopia no infravermelho e análise por componentes principais. Neste estudo, foram elaboradas 81 amostras binárias (blendas) a partir de adições de percentuais crescentes de biodiesel, óleo de girassol bruto, óleo de soja degomado ou óleo residual de fritura em diesel, sendo empregadas uma amostra de interior e outra metropolitana, cedida pela REFAP, Esteio - RS, Brasil. As concentrações compreenderam a faixa de concentração de 0,5 a 30% (para o biodiesel, óleo de girassol bruto, óleo de soja degomado e óleo residual de fritura). Os espectros destas amostras foram adquiridos em dois espectrofotômetros distintos, Nicolet Magna 550 e Shimadzu IR Prestige - 21. Para selecionar e avaliar as faixas espectrais foi utilizado o algoritmo iPCA (análise por componentes principais por intervalos). De acordo com os resultados foi possível visualizar a separação dos grupos formados pelos óleos vegetais (adulterantes) das blendas contendo biodiesel, usando uma faixa específica do espectro selecionada por meio da análise de iPCA. Observou-se também que utilizando outras faixas do espectro podem-se separar as amostras contendo diesel interior e metropolitano, bem como a diferenciação dos dois equipamentos utilizados nas análises. Este estudo amplia as potencialidades da espectroscopia no infravermelho, a qual pode ser amplamente empregada no monitoramento de outros adulterantes frequentemente encontrados em misturas comerciais de biodiesel/diesel.

Godinho et al. (2008) classificaram refrigerantes através de análise de imagens e análise de componentes principais (PCA). Através de um escâner de mesa, foram geradas imagens de 29 marcas de refrigerantes dos tipos Cola, Guaraná e Laranja, sendo possível estabelecer padrões de similaridade dentro dos gráficos dos *scores* das componentes principais, com base nos valores médios dos histogramas dos canais de cor R, G e B. A mudança na cor das imagens foi acompanhada no gráfico

das componentes principais conforme o peso do índice de cor na componente principal, como resultado da tonalidade média de cada marca. Resultados de análises físico-químicas como o teor de sacarose, de ácido sórbico e o pH podem ser correlacionados com as imagens classificadas pela PCA. Diferentes marcas de refrigerantes puderam ser classificadas pelas suas imagens.

## 4 METODOLOGIA

De acordo com Santos (2000), explorar é investigar, é informar ao pesquisador a real importância do problema, o estágio em que se encontram as informações já disponíveis a respeito do assunto, podendo revelar ao pesquisador novas fontes de informação. Tendo como base os objetivos específicos este trabalho será realizado a partir de uma pesquisa exploratória, uma pesquisa descritiva, pela descrição dos fatos e fenômenos observados.

Em relação aos procedimentos de coleta, Santos (2000) afirma que quando um fato ou fenômeno da realidade é conduzido de forma controlada, com objetivo de descobrir fatores que o produzem que por esses são produzidos, caracterizarão este trabalho como uma pesquisa experimental, tendo como fontes de informação as pesquisas bibliográficas e as interpretações dos dados resultantes das análises laboratoriais, durante todo o período da dissertação.

Este trabalho foi desenvolvido nos laboratórios do Curso de Química da Universidade de Santa Cruz do Sul – UNISC e foi dividido em cinco etapas: a primeira consistiu na aquisição de espectros no infravermelho médio e próximo de diferentes óleos vegetais como conjunto de dados para avaliação e validação das ferramentas contempladas pelo *software*. A segunda etapa focou no desenvolvimento do *software* e suas ferramentas, a terceira etapa contemplou a validação das ferramentas implementadas pelo *software* utilizando Matlab<sup>®</sup> versão 7.1 (The Mathworks Inc.). A quarta etapa, consistiu na aquisição de imagens digitais de diferentes óleos diesel comerciais e implementação de uma função para geração de histogramas e decomposição de pixels nos modelos de cores RGB, HSV, valores de iluminação e intensidade de brilho, e a última etapa, contemplou o desenvolvimento de uma solução *online* com alguns recursos básicos de tratamento de dados além das técnicas de análise de agrupamento hierárquico (HCA) e análise de componentes principais (PCA).

## 4.1 Amostragem

### 4.1.1 Origem e identificação das amostras de óleos vegetais

Para o desenvolvimento da primeira etapa do trabalho, foram utilizadas amostras de diferentes óleos vegetais disponíveis comercialmente (algodão, amendoim, arroz, canola, girassol, mamona, milho e soja). A descrição dos óleos vegetais utilizados está apresentada na Tabela 2.

Tabela 2. Identificação e origem das amostras de óleos vegetais utilizadas no NIR e MIR.

<b>Tipo de óleo</b>	<b>Código</b>	<b>Marca</b>	<b>Lote</b>	<b>Fabricação</b>	<b>Validade</b>
Soja	SOYB	Bunge	L0912-066368	15/09/2012	06/2013
Soja	SOYL	Leve	C5112	16/06/2012	06/2013
Canola	CANB	Bunge	L0312-021459	28/03/2012	12/2012
Canola	CANL	Liza	L05 C	11/05/2012	05/2013
Girassol	SFWB	Bunge	L0412-029698	29/04/2012	01/2013
Girassol	SFWL	Liza	L06 C	08/06/2012	06/2013
Arroz	RISE	Carreteiro	L180	06/2012	06/2013
Milho	CRNB	Bunge	L0312-020615	26/03/2012	12/2012
Milho	CRNL	Liza	L08 C	27/08/2012	08/2013
Algodão	COTT	Triangulo	PD-071	12/06/2012	-
Amendoim	PEAT	Triangulo	PD-069	12/06/2012	-

Fonte: Autor, 2013.

Os espectros obtidos no infravermelho médio tiveram a letra “M” acrescida na frente do código de identificação das amostras, enquanto no infravermelho próximo tiveram a letra “N”.

### 4.1.2 Origem e identificação das amostras de óleos diesel

Para o desenvolvimento da quarta etapa foram utilizadas 14 amostras de óleo diesel comerciais. Destas 14 amostras, 11 amostras foram coletas em postos de combustíveis na região de Santa Cruz do Sul e Porto Alegre e 3 amostras provenientes de uma distribuidora no Estado do Rio Grande do Sul, chamadas de amostras-padrão. Na Tabela 3 encontra as informações referentes às amostras de óleo diesel comerciais.

Tabela 3. Identificação e origem das amostras de óleo diesel utilizadas no NIR, MIR e escâner.

Tipo de diesel	Código	Distribuidora	Cidade	Data coleta
S10	S1BSS	BR	Santa Cruz do Sul	30/07/2013
S10	S1BGP	BR	Porto Alegre	03/08/2013
S10	S1STS	Shell	Santa Cruz do Sul	30/07/2013
S10	S1SAP	Shell	Porto Alegre	03/08/2013
S500	S5BSS	BR	Santa Cruz do Sul	30/07/2013
S500	S5BGP	BR	Porto Alegre	03/08/2013
S500	S5ICP	Ipiranga	Santa Cruz do Sul	30/07/2013
S500	S5SAP	Shell	Santa Cruz do Sul	31/07/2013
S1800	S8IBS	Ipiranga	Santa Cruz do Sul	30/07/2013
S1800	S8IRS	Ipiranga	Santa Cruz do Sul	31/07/2013
S1800	S8VUS	24 Horas	Santa Cruz do Sul	09/08/2013

Fonte: Autor, 2013.

#### 4.1.3 Obtenção dos espectros dos óleos vegetais

Os espectros foram adquiridos no espectrofotômetro PERKIN ELMER modelo Spectrum 400 nas regiões do infravermelho médio (MIR), intervalo de 4000 - 650  $\text{cm}^{-1}$ , e no infravermelho próximo (NIR), intervalo de 10000 - 4000  $\text{cm}^{-1}$ .

No MIR foi utilizado um acessório de refletância total atenuada universal (UATR-FTIR), com resolução de 4  $\text{cm}^{-1}$  e 32 varreduras. O cristal utilizado nesta técnica contém na sua base superior diamante e elemento focalizador de seleneto de zinco. Para a aquisição dos espectros no MIR foi realizado primeiramente a leitura do branco (*background*) para cada 5 replicatas, todas com um auxílio de uma pipeta.

Já no NIR foi utilizado um acessório contendo esfera de integração (NIRA), com resolução de 4  $\text{cm}^{-1}$  e 32 varreduras. Para a aquisição dos espectros no NIR, primeiramente, foi realizado a leitura do branco (*background*) contendo apenas a placa de vidro, a superfície refletora de alumínio e uma tampa para encobri-los, não permitindo a entrada de luz externa.

Após a leitura do branco, a amostra é colocada na placa de vidro com o auxílio de uma pipeta até formar uma camada em toda parte inferior da placa, conforme apresentado na Figura 14.



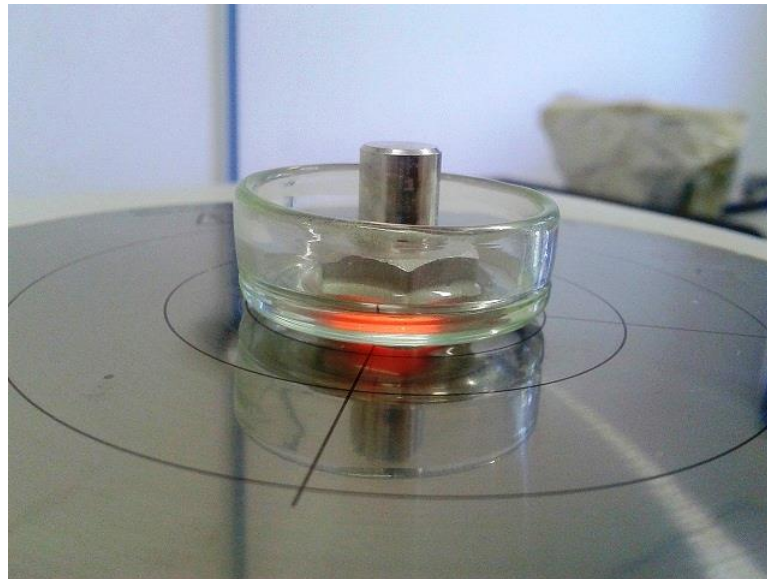


Figura 14. Placa de vidro e acessório de alumínio para a aquisição da leitura do branco.  
Fonte: Autor, 2013.

#### 4.1.4 Obtenção das imagens das amostras de óleo diesel

As imagens das amostras de óleo diesel foram adquiridas numa impressora multifuncional HP modelo F4400 dotada de um escâner de mesa. Para o escaneamento das imagens utilizou-se uma máscara de cartolina branca com furo central a 1/3 de altura do papel, gerando assim imagens sempre na mesma posição, Figura 15 (a). Foi adicionado então aproximadamente 2 mL de óleo diesel num vidro tipo recipiente, utilizado para aquisição dos espectros no infravermelho próximo, Figura 15 (b), além de uma tampa branca, Figura 15 (c), e outra com fundo preto fosco como cobertura de proteção a luz externa, Figura 15 (d).

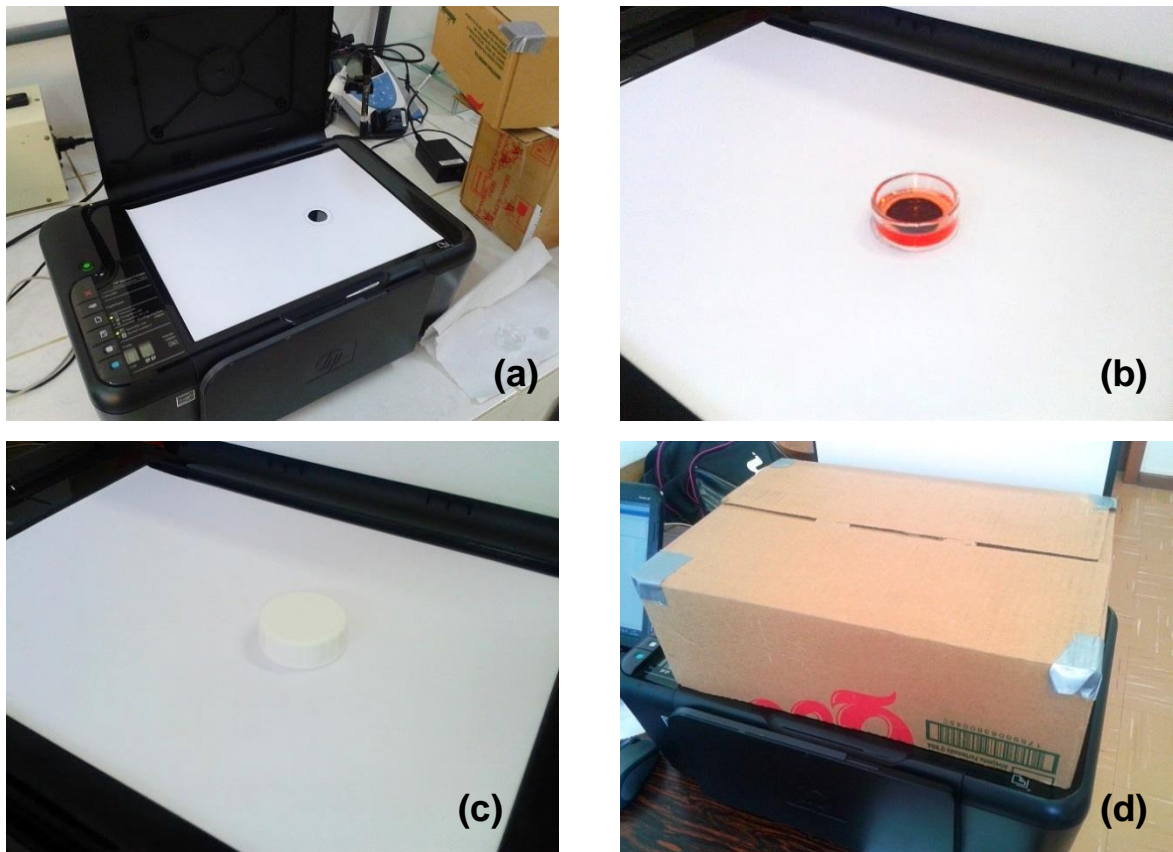


Figura 15. a) Impressora multifuncional e papel-máscara, b) recipiente de vidro com amostra, c) tampa de plástico branca e d) tampa com fundo preto fosco para evitar entrada de luz.  
Fonte: Autor, 2013.

As imagens foram escaneadas em quintuplicata, logo após serem analisadas pelo infravermelho próximo, numa resolução de 600 dpi's com contraste de 75%.

## 4.2 Desenvolvimento e validação do *software*

O *software* foi gerado num ambiente de desenvolvimento integrado (IDE, do inglês, “*Integrated Development Environment*”). A função da IDE é reunir características e ferramentas de apoio à construção de *software* com o objetivo de agilizar este processo. Para tanto foi utilizado a IDE Microsoft Visual Studio 2010<sup>®</sup> versão *Professional*, que possui um alto nível de abstração de controles e classes, decorrente do uso do pacote *Microsoft .NET Framework 4.0*, ilustrado na Figura 16.

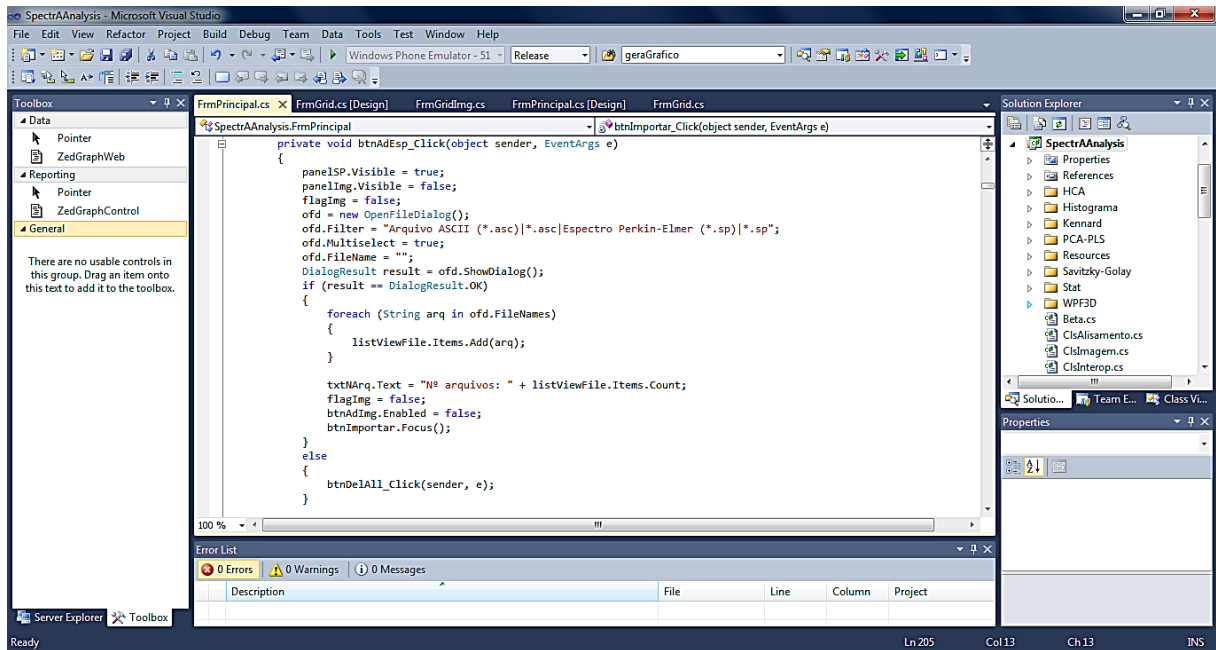


Figura 16. Tela de desenvolvimento da IDE Microsoft Visual Studio® 2010.  
Fonte: Autor, 2013.

As linguagens de programação adotadas foram C# (C-Sharp) e VB (Visual Basic). Foram utilizados algoritmos e bibliotecas de terceiros, como *ZedGraph* (CHAMPION, 2012) para plotagens de gráficos e o *Accord.NET Framework* (SOUZA, 2012) que possui inúmeros algoritmos da área da estatística, todos de código aberto. A solução *online* também foi desenvolvida no Visual Studio 2010 e contou com a vantagem de reutilização dos algoritmos da versão *desktop*.

Para importação de dados ao *software* são aceitos arquivos de espectros gerados pelo espectrofotômetro de infravermelho Perkin-Elmer nos formatos “sp” e “asc” (ASCII), e arquivos de imagens nos formatos “bmp”, “jpg” e “png”, além da opção da área de transferência, comumente conhecido como “copiar-colar” através dos atalhos de teclas “Control+C” (copiar) e “Control+V” (colar). Após importado, se forem dados espectrais, poderão ser segregadas apenas algumas regiões espectrais ou todo o conteúdo do arquivo. A solução *online* conta apenas com a importação através da área de transferência.

Além disso, foram desenvolvidas algumas ferramentas matemáticas de transformação e pré-processamento de sinais, como correções, suavizações e normalizações de acordo com os métodos da primeira e segunda derivadas, correção de espalhamento de luz, método de variação normal padrão e método de Savitzky-Golay, a partir de dados centrados na média ou escalonados. Conversões

de medidas também foram implementadas entre as unidades absorvância e transmitância.

Foi introduzida a opção de geração de médias para replicatas e identificação de classes de amostras pela análise sintática de sua denominação (nome do arquivo).

Os dados, no formato de planilhas provenientes das análises, podem ser exportados para extensão “.xls” (Microsoft Excel<sup>®</sup>), “.txt” (texto tabulado) e “.asc” (ASCII). Já os gráficos, possuem opção de salvar nos formatos de figuras “.bmp”, “.png” e “.jpg”.

No caso de importação de imagens, podem ser selecionados dados do histograma R, G e B, separadamente, ou, a média ou mediana de todo o *frame* nos modelos de cores R, G, B, R relativo, B relativo, G relativo, H, S, V, I e L, também separadamente.

As técnicas desenvolvidas na versão *desktop* foram:

- Análise por agrupamento hierárquico (HCA);
- Análise por componentes principais (PCA), e detecção de amostras anômalas (*outliers*) pelo método  $T^2$  de Hotelling;
- Análise por componentes principais por intervalos (iPCA), muito utilizado como técnica de seleção de variáveis espectrais.

Já a versão *online* possui habilitada apenas as técnicas de HCA e PCA (*scores*). A linguagem utilizada para ambas às versões é inglesa.

A validação das ferramentas contempladas pelo *software* foi realizada utilizando o aplicativo Matlab<sup>®</sup> versão 7.11 (The Mathworks Inc.).

### 4.3 Requisitos mínimos do *software*

Para o funcionamento do *software* ChemoStat são necessários alguns requisitos mínimos para instalação, tais como:

- Windows<sup>®</sup> XP Service Pack 3, Windows<sup>®</sup> Vista, Windows<sup>®</sup> 7, Windows<sup>®</sup> 8 ou 8.1, nas versões de 32 bits ou 64 bits.
- Pacote Microsoft .NET Framework 4.0, disponível em: <http://www.microsoft.com/pt-BR/download/details.aspx?id=24872>, ou superior.

- Processador de 1GHz com 512 Mb de memória RAM e espaço livre em disco de no mínimo 10 Mb.
- Microsoft Internet Explorer 9.0, ou superior, para versão *online*.

O *software* ChemoStat consiste em três arquivos e, para seu funcionamento, todos devem estar no mesmo diretório. São eles:

- ChemoStat.exe, arquivo executável.
- ChemoLib.dll, biblioteca de algoritmos.
- ZedGraph.dll, biblioteca de gráficos.

Para iniciar o programa basta executar o arquivo "ChemoStat.exe". A falta de uma instalação "formal" do Windows<sup>®</sup> é intencional e permite o uso do ChemoStat sem privilégios de administrador.

## 5 RESULTADOS E DISCUSSÕES

Neste capítulo são apresentadas as características e as funcionalidades do *software* desenvolvido utilizando o conjunto de dados obtidos por espectroscopia no infravermelho médio e próximo de óleos vegetais e o conjunto de dados obtidos através de escaneamento de imagens de diferentes tipos de óleo diesel. Simultaneamente, os resultados obtidos através do *software* desenvolvido são comparados àqueles obtidos via Matlab<sup>®</sup> versão 7.11.

### 5.1 Tela principal

A tela principal do *software* ChemoStat possui 3 seções, ilustradas na Figura 17.

- Seção 1: Destinado ao gerenciamento de arquivos, marcada em vermelho.
- Seção 2: Destinado ao gerenciamento das variáveis (comprimento de ondas ou modelos de cores), marcada em verde.
- Seção 3: Área destinada à grade ou matriz de dados, marcada em azul.

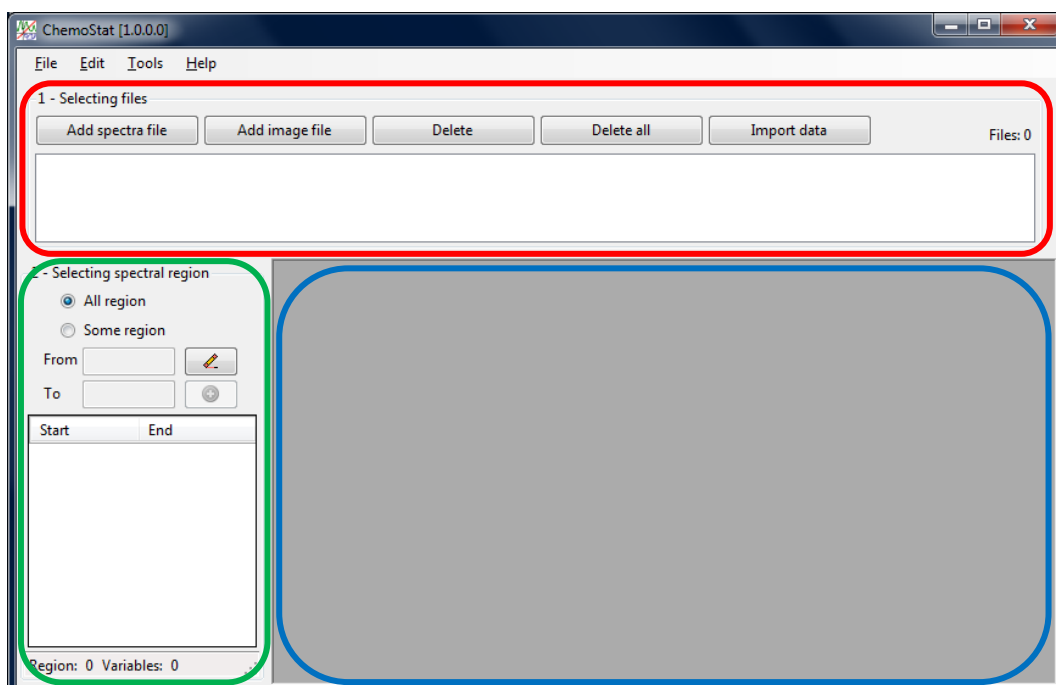


Figura 17. Tela principal do ChemoStat - padrão espectroscopia.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

A tela principal padrão é de uso para espectroscopia. O botão “*Add image file*”, permite alterar a seção 2 para utilização de imagens, conforme marcado em verde na Figura 18.

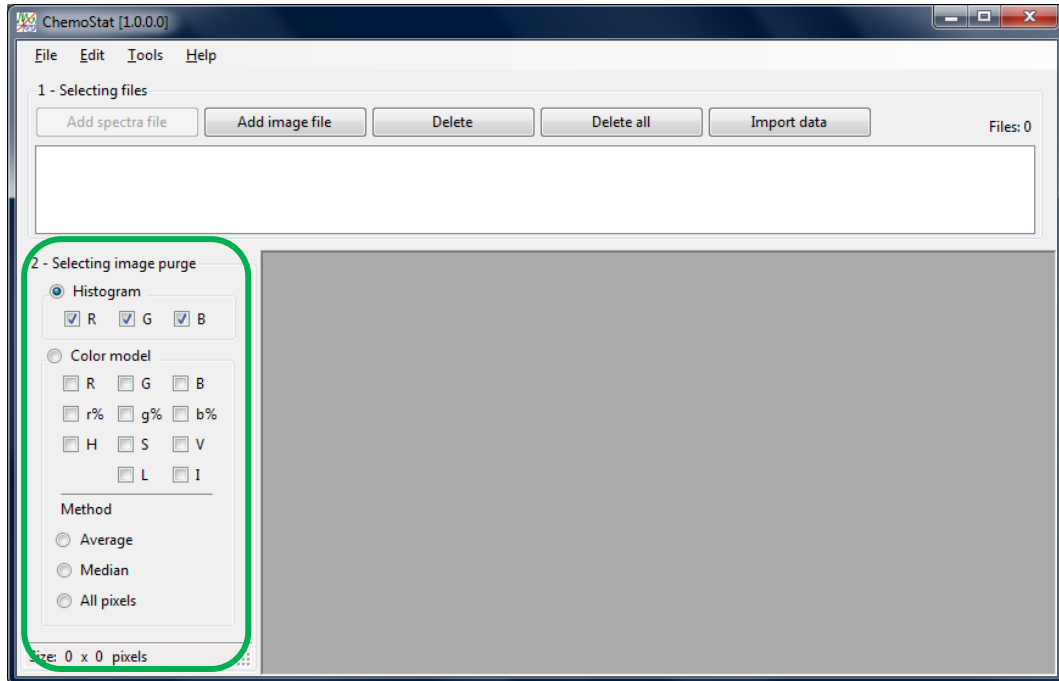


Figura 18. Tela principal do ChemoStat - padrão imagens.  
Fonte: Autor, extraído do *software* ChemoStat, 2014

Existe ainda outra forma de selecionar os modos de operação para espectroscopia ou imagens. Basta acessar via menu “*File*” e logo após “*Operation mode*”, no campo superior direito, conforme demonstra a Figura 19.

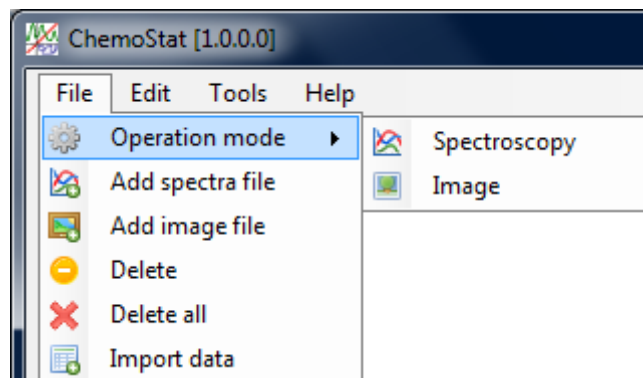


Figura 19. Menu de operação - cabeçalho da tela principal.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

A opção “*Spectroscopy*” habilita a tela principal para dados espectroscópicos, enquanto que a opção “*Image*” habilita para dados de imagens. Os demais campos “*Add spectra file*”, “*Add image file*”, “*Delete*”, “*Delete all*” e “*Import data*”, possuem as mesmas funções dos botões da seção 1, a seguir discutidos.

## 5.2 Importação de dados espectrais

Os dados podem ser importados via área de transferência (comumente conhecido pelo atalho das teclas “*Control+V*”) através do Excel, na disposição dos dados onde as colunas representam as amostras e as linhas representam as variáveis, ou via botão “*Add spectra file*” para seleção de arquivos.

Ao pressionar o botão “*Add spectra file*” será apresentada uma janela para seleção dos arquivos (padrão Windows<sup>®</sup>, conforme Figura 20) com as opções de filtro de extensão ASCII ou SP, formatos gerados pelo equipamento Perkin-Elmer ou pelo *software* ChemoStat. Após selecioná-los basta clicar em “*Open*” ou “*Abrir*” para que os mesmos sejam transferidos para a seção 1.

Vale ressaltar que os arquivos no formato “sp”, até então, poderiam ser lidos apenas no *software* da Perkin-Elmer, fabricante do espectrômetro. Foi desenvolvido um algoritmo para interpretação dessa extensão, trazendo assim rapidez na execução das tarefas do ChemoStat por não necessitar de conversões para formato texto.

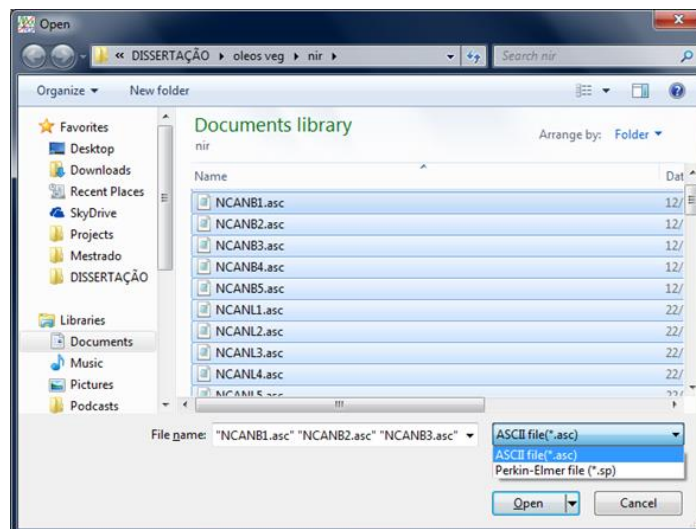


Figura 20. Janela de seleção de arquivos padrão Windows<sup>®</sup>.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.



Uma vez adicionados na seção 1, marcado em vermelho na Figura 21, os arquivos podem ser deletados individualmente ou em sua totalidade, função exercida pelos botões “Delete” e “Delete all”, respectivamente. O teclado do computador também poderá ser utilizado para exclusão desses arquivos, desde que sejam selecionados pelo “mouse”.

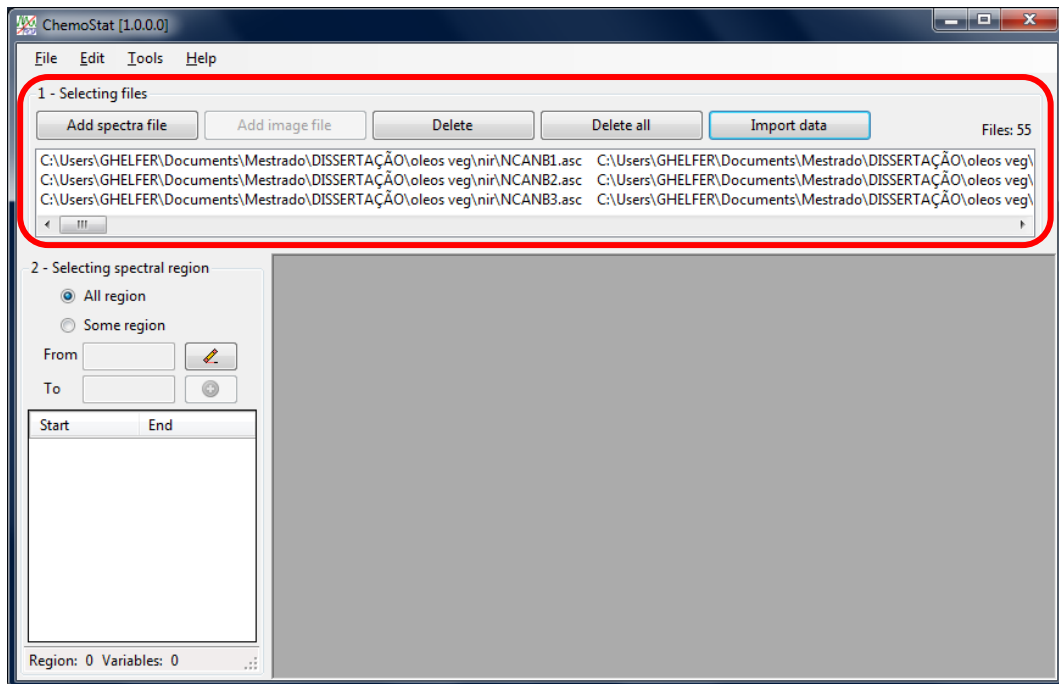


Figura 21. Tela principal padrão espectroscopia – identificação da seção 1.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

Antes da leitura dos arquivos, o *software* permite a segregação de determinadas regiões espectrais. Para tanto, é necessário que seja habilitada a seção 2, selecionando a opção “Some region” e preenchendo os campos “From” e “To” com o comprimento de onda inicial e final, respectivamente.

Nesta etapa, podem ser segregadas uma ou mais regiões, clicando no botão com símbolo de adição. A opção “All region”, *default*, carrega toda a informação espectral contida nos arquivos, desabilitando este comando, mesmo que a grade “Start-End” esteja preenchida. A Figura 22 demonstra o preenchimento das regiões.

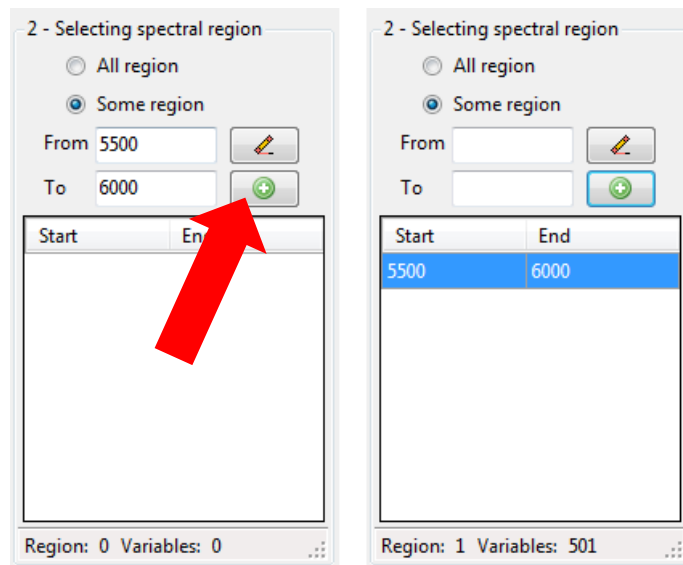


Figura 22. Detalhe da seção 2 na tela principal padrão espectroscopia.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

Após selecionada a opção de importação, o botão “*Import data*” realiza a leitura dos arquivos e extração das informações referentes aos espectros na seção 3, em azul. Vale ressaltar que o nome do arquivo é utilizado como nome da amostra, excluindo-se a extensão. A Figura 23 ilustra a grade dos dados importados e o nome do arquivo no cabeçalho da mesma.

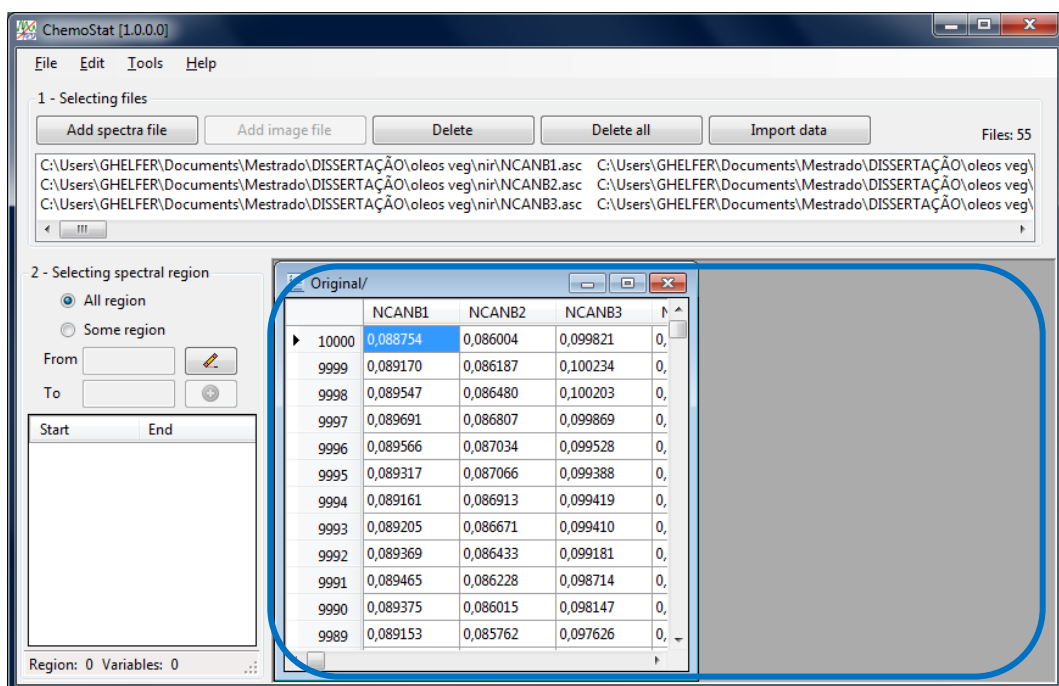


Figura 23. Tela principal padrão espectroscopia com grade de dados – seção 3.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

### 5.2.1 Menu de ferramentas

Na grade ou matriz de dados, ao clicar com o botão direito do “mouse”, será apresentado um menu com as opções de plotagem de gráfico, exportação para excel, texto e ASCII, conversões, tratamentos e pré-processamentos dos dados, identificação de amostras por classe, além das técnicas HCA, PCA e iPCA, conforme demonstra a Figura 24, e a seguir discutidas.

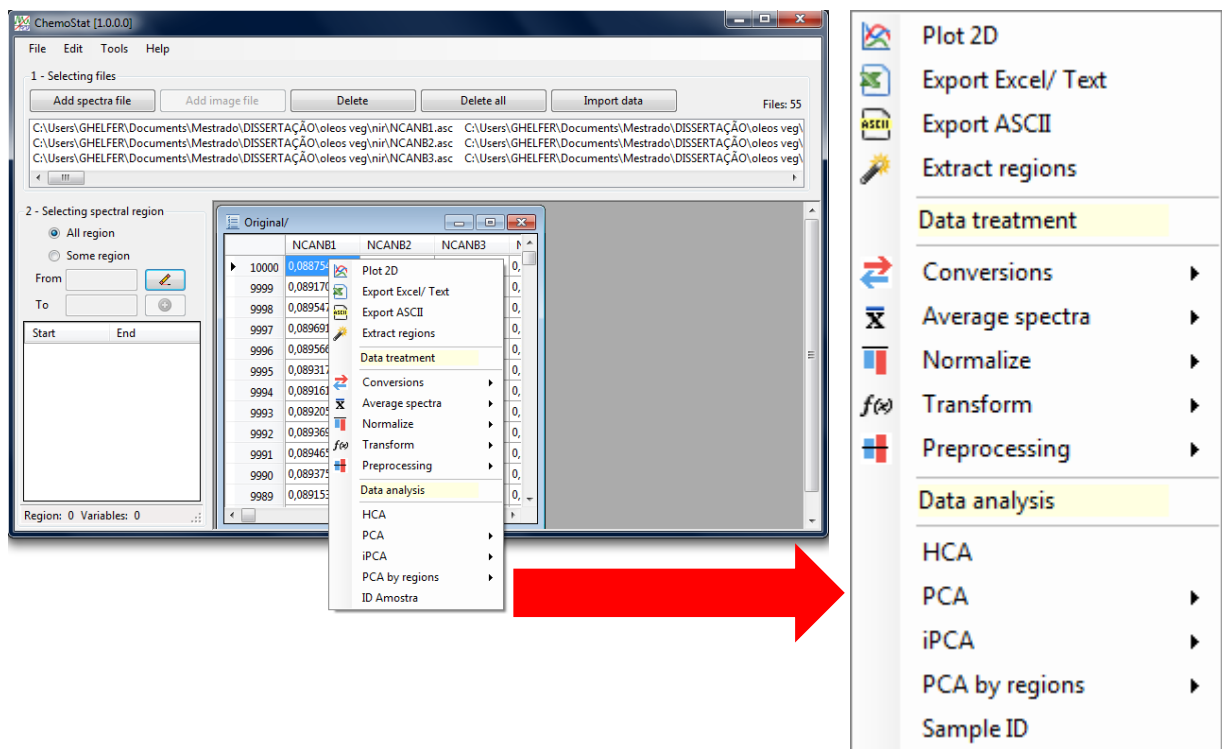


Figura 24. Tela principal padrão espectroscopia com grade de dados – detalhe do menu de operações acionado.

Fonte: Autor, extraído do software ChemoStat, 2014.

### 5.2.2 Função “Plot 2D”

A opção “Plot 2D”, quando aplicada no grade de dados, exibe uma nova janela apresentando o gráfico dos espectros, conforme mostra a Figura 25.

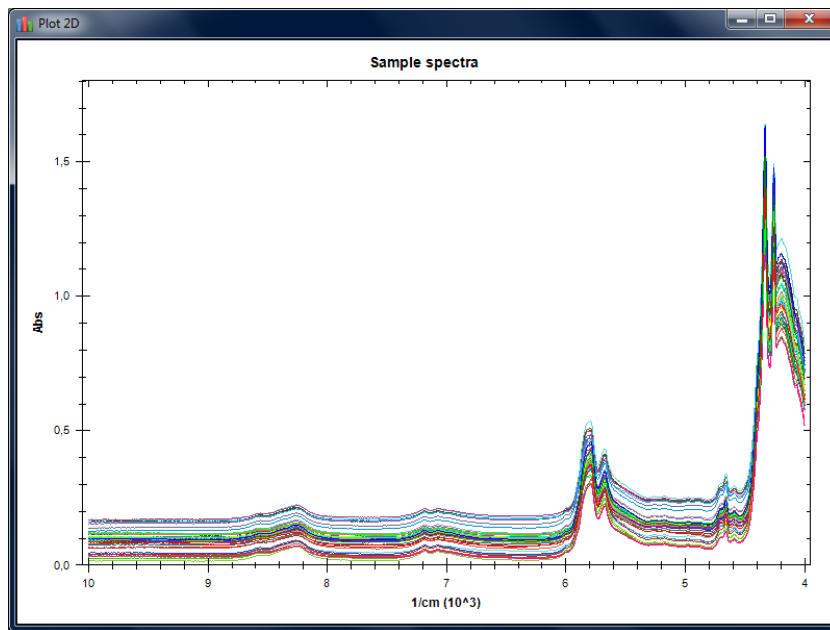


Figura 25. Gráfico de espectros de óleos vegetais obtidos via espectroscopia no infravermelho próximo, sem tratamento de dados, a partir da função Plot 2D.  
 Fonte: Autor, extraído do *software* ChemoStat, 2014.

Na área do gráfico, existe mais um menu de comandos, acionado ao clicar com o botão direito do “mouse”. Além do menu padrão, encontrado em todos os gráficos do *software*, exceto no HCA, existem ainda as opções específicas do Plot2D, como “*Select region*” (Selecionar região) e “*Show/ hide legend*” (Mostrar/ocultar legenda). O menu padrão consiste nas seguintes rotinas: “*Copy*” (copiar), “*Save image as*” (salvar imagem como), “*Page setup*” (configuração da página), “*Print*” (imprimir), “*Show point values*” (mostrar pontos) e “*Set scale to default*” (setar escala para padrão), apresentado na Figura 26.

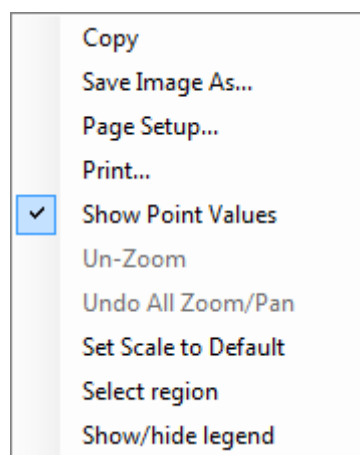


Figura 26. Menu de operações do gráfico via função Plot 2D.  
 Fonte: Autor, extraído do *software* ChemoStat, 2014.

O comando específico “Selecionar região” demarca, no gráfico de espectros, as regiões informadas na seção 2, conforme ilustra a Figura 27, enquanto o item “Mostrar/ocultar legenda” adiciona um quadro informativo com os nomes das amostras plotadas no gráfico.

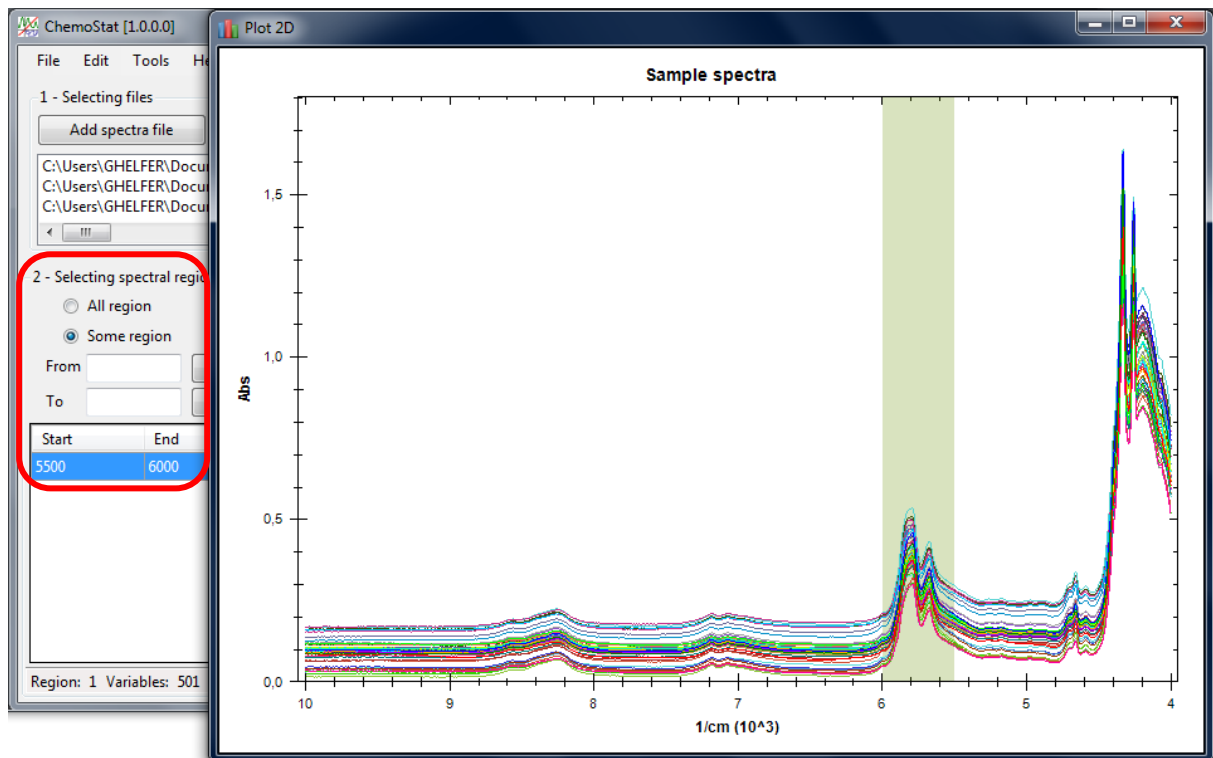


Figura 27. Gráfico de espectros de óleos vegetais com opção “selecionar região” executada. Fonte: Autor, extraído do *software* ChemoStat, 2014.

### 5.2.3 Função “Export Excel/ Text”

O item “*Export Excel/ Text*”, quando acionado, solicita ao usuário um nome de arquivo e um diretório onde será gerado o mesmo no formato Excel ou Texto, extensões “xls” ou “txt”, de acordo com o filtro de extensão selecionado, que agregará todas as informações contidas na grade de dados.

### 5.2.4 Função “Export ASCII”

A opção “*Export ASCII*” permite apenas a escolha de um diretório. Neste local, essa função gera um arquivo formato ASCII (“.asc”) para cada amostra, de acordo

com a disposição dos dados na grade. A extensão “.asc”, assim como a extensão “.sp”, são formatos que o ChemoStat utiliza como entrada de dados.

### 5.2.5 Função “Extract region”

O item “*Extract region*”, permite realizar uma função semelhante à importação de dados quando selecionado a opção “*Some region*” da seção 2. Nesse caso, ao ser acionada, extrai, para uma nova grade de dados, somente aquelas regiões indicadas na seção 2, caso existam. Cada função acrescentada aos dados originais vai sendo rotulada no cabeçalho da grade de dados, conforme marcado em vermelho na Figura 28.

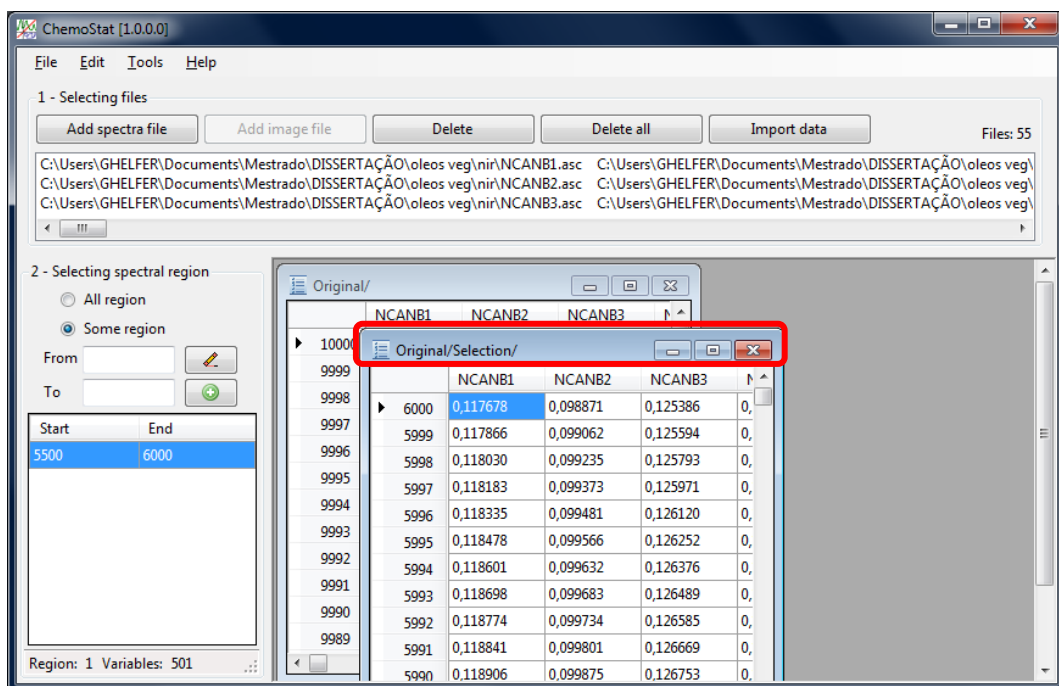


Figura 28. Grade de dados com rótulo das funções atribuídas.  
Fonte: Autor, extraído do software ChemoStat, 2014.

A Figura 29 exhibe o conjunto de espectros dos óleos vegetais sem tratamento de dados na faixa entre 5500 e 6000  $\text{cm}^{-1}$  após execução das opções “*Extract region*”, nessa faixa de onda, e “*Plot 2D*”.

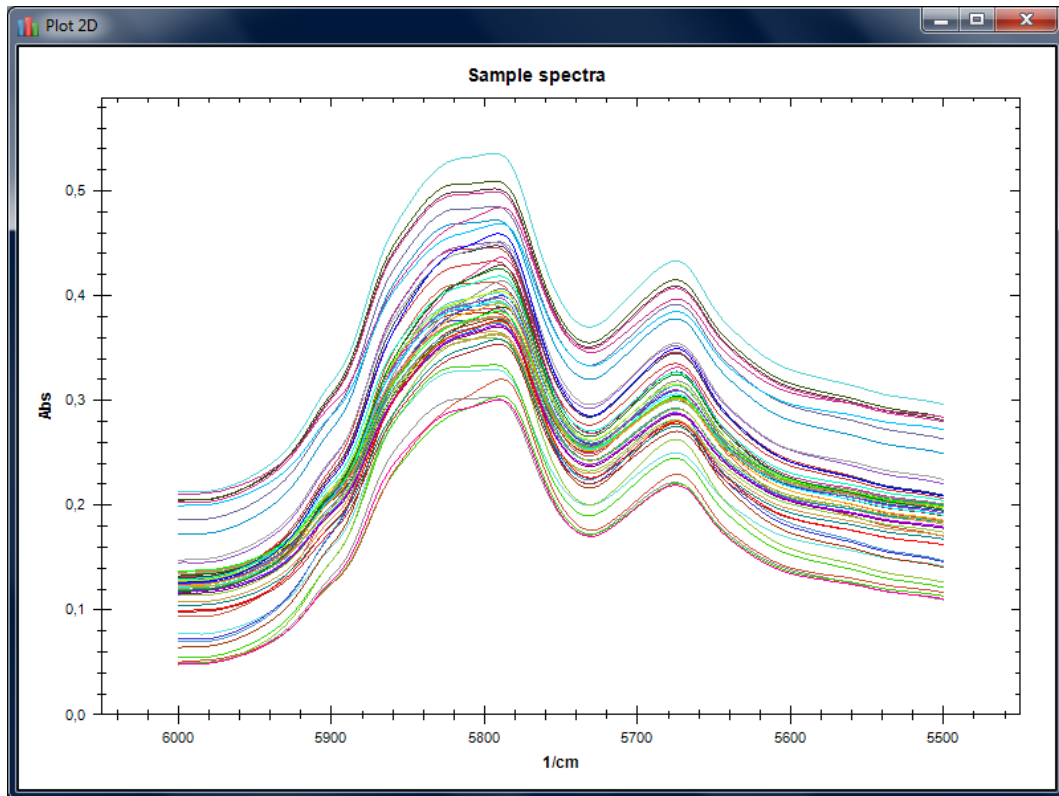


Figura 29. Gráfico do conjunto de espectros dos óleos vegetais sem tratamento de dados na faixa entre 5500 e 6000  $\text{cm}^{-1}$ .

Fonte: Autor, extraído do *software* ChemoStat, 2014.

Segundo Yang, Irudayaraj e Paradkar (2005) e Xiaobo et al. (2010), a faixa entre 5.500 e 6.000  $\text{cm}^{-1}$  representa a vibração axial (estiramento) da ligação C–H dos grupos funcionais  $-\text{CH}_2$ ,  $-\text{CH}_3$  e  $-\text{CH}=\text{CH}$  nos óleos e gorduras comestíveis. Como os óleos vegetais do conjunto de dados diferem com relação as saturações e insaturações, essa região deve permitir a discriminação dos óleos.

### 5.2.6 Funções de conversões de unidades

O *software* permite a conversão entre as unidades absorvância e transmitância. Para executá-los basta acionar no menu o item “*Conversions*” e clicar na opção “*Abs >> T%*” para conversão de absorvância para transmitância ou “*T% >> Abs*” de transmitância para absorvância, de acordo com a Figura 30.

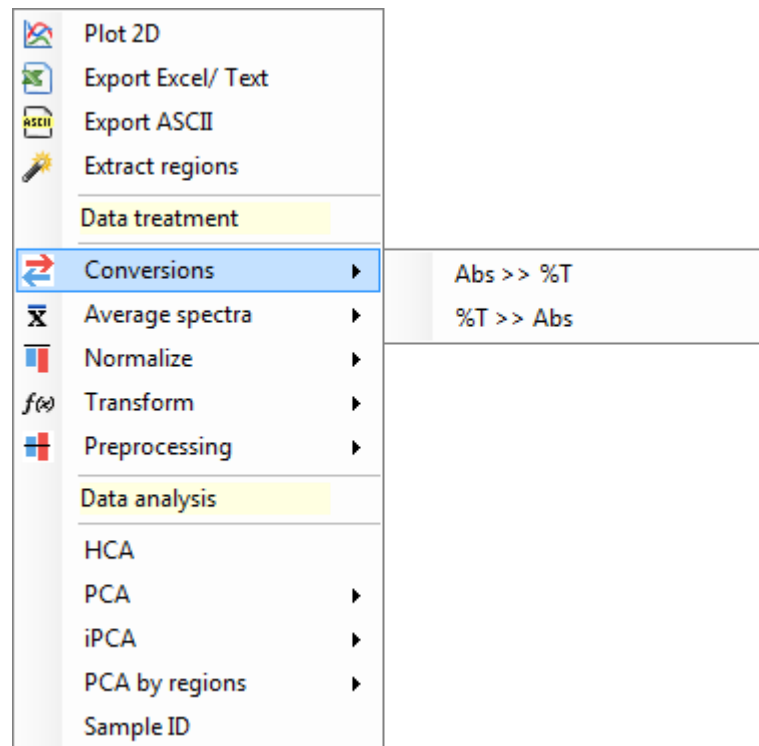


Figura 30. Menu principal de operações - funções de conversão.  
 Fonte: Autor, extraído do *software* ChemoStat, 2014.

### 5.2.7 Função espectro médio

Outro procedimento muito utilizado em matrizes de espectros na quimiometria chama-se espectro médio, ou “*Average spectra*”, de acordo com a Figura 31. Essa rotina realiza a média das replicatas das amostras. Os itens “*2nd-Duplicate*”, “*3rd-Triplicate*”, “*4th-Quadruplicate*” e “*5th-Quintuplicata*” realizam as médias de duas, três, quatro e cinco amostras, quando dispostas lado a lado na grade de dados.

Caso o número de replicatas seja maior que cinco, o item “*By value*” permite informar a quantidade de replicatas para operação da média, através de uma janela de diálogo imposta ao usuário. Já a opção “*By class*” possibilita realizar a média das replicatas de acordo com a identificação das amostras por classes, função previamente definida pelo usuário e posteriormente discutida. O item “*All collection*” realiza a média de todas as amostras contidas na grade de dados.



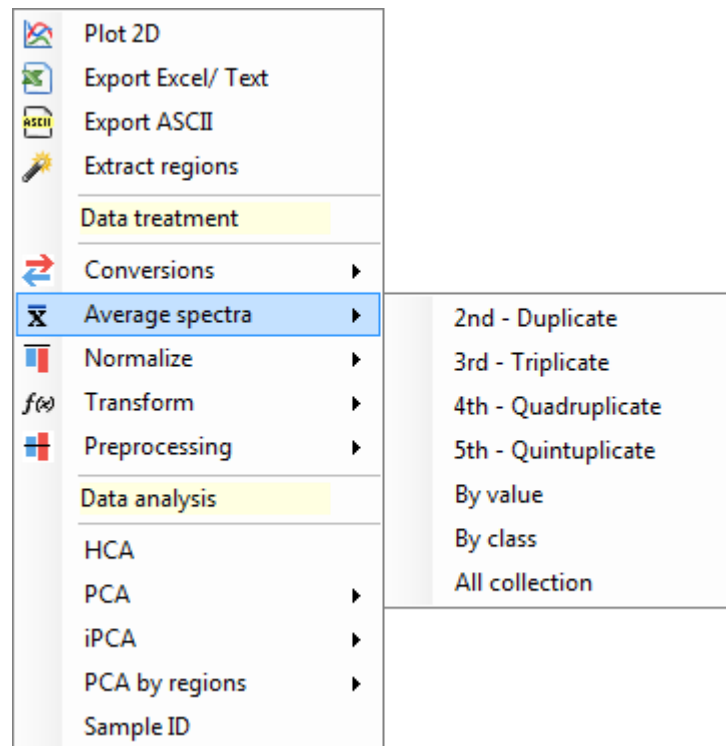


Figura 31. Menu principal de operações - funções de espectro médio.  
 Fonte: Autor, extraído do *software* ChemoStat, 2014.

### 5.2.8 Funções de normalização de espectros

A função de normalização contida no *software* ChemoStat corresponde a três opções, conforme ilustrado na Figura 32. A primeira, descrita como “*By range(1-0)*” realiza a correção entre os espectros de modo que todos fiquem numa escala comparável, compreendidos entre 1 e 0. Já a opção “*By max*” divide os espectros pelo valor máximo da amostra, enquanto que “*By value*” solicita ao usuário, por meio de uma caixa de diálogo, um valor pelo qual serão divididos os espectros.

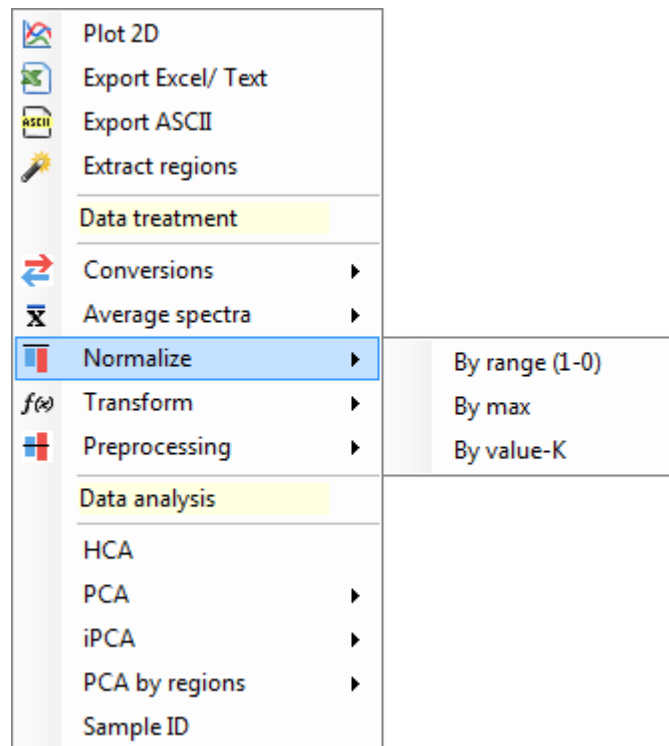


Figura 32. Menu principal de operações - funções de normalização.  
 Fonte: Autor, extraído do *software* ChemoStat, 2014.

As Figuras 33 e 34 exibem o conjunto de espectros dos óleos vegetais normalizados entre os limites de zero e um de absorvância, pela opção “*By range (1-0)*”, empregando ChemoStat e Matlab<sup>®</sup>, respectivamente.

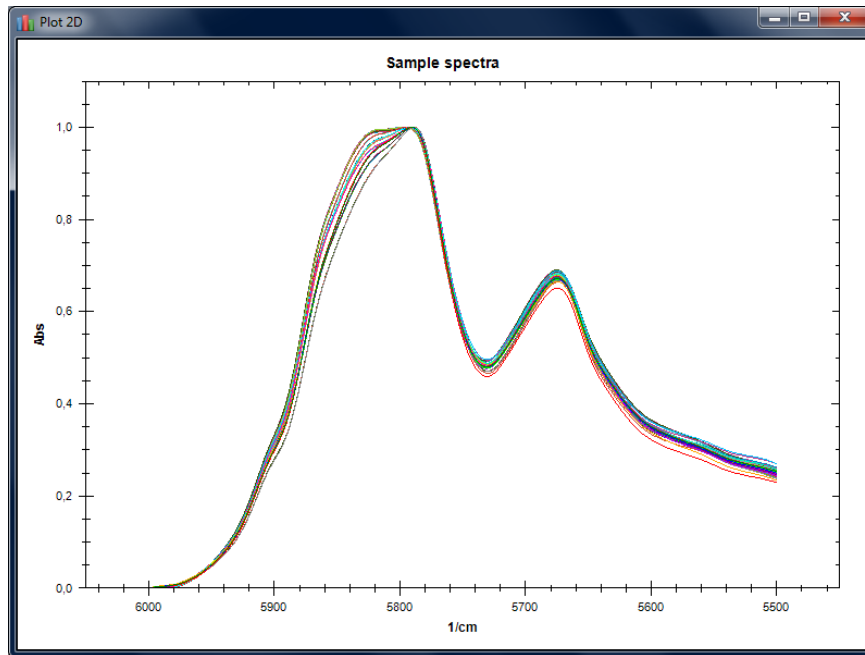


Figura 33. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR) normalizados entre os limites zero e um de absorvância, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – ChemoStat.  
 Fonte: Autor, extraído do software ChemoStat, 2014.

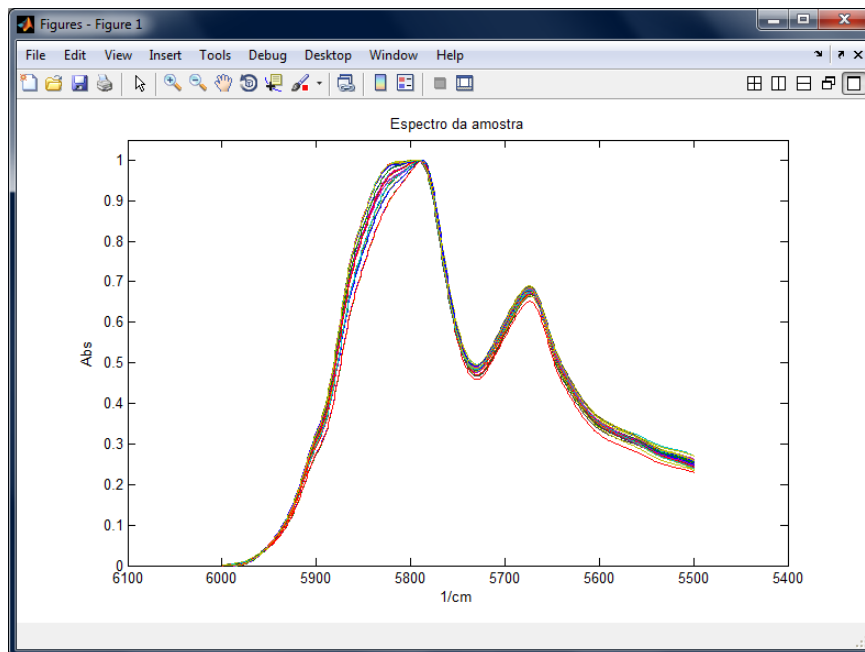


Figura 34. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR) normalizados entre os limites zero e um de absorvância, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – Matlab®.  
 Fonte: Autor, extraído do software Matlab®, 2014.

### 5.2.9 Funções de transformações

A Figura 35 apresenta as funções de transformação disponíveis no *software*. A opção “SNV” aplica sobre a grade de dados o método de Variação Normal Padrão (do inglês, *Standart Normal Variate*), enquanto que a opção “MSC” executa o método chamado Correção do Espalhamento de Luz (do inglês, *Multiplicative Scatter Correction*).

O item “1st derivative” possibilita realizar a primeira derivada sobre os dados. Neste caso, assim como os itens “2nd derivative”, da segunda derivada, e “Moving Average”, média móvel, é requisitada ao usuário, via caixa de diálogo, a quantidade de pontos (número de ondas) nos quais será aplicado o algoritmo.

Já a opção “Savitsky-Golay”, como o nome menciona, realiza o método de Savitsky-Golay para suavização dos dados. Antes de realizar a aplicação do algoritmo, o *software* solicita ao usuário, por meio de uma caixa de dialogo, três valores. O primeiro diz respeito a ordem derivativa, o segundo valor significa a ordem polinomial e a terceira a janela de pontos nos quais será aplicado o algoritmo.

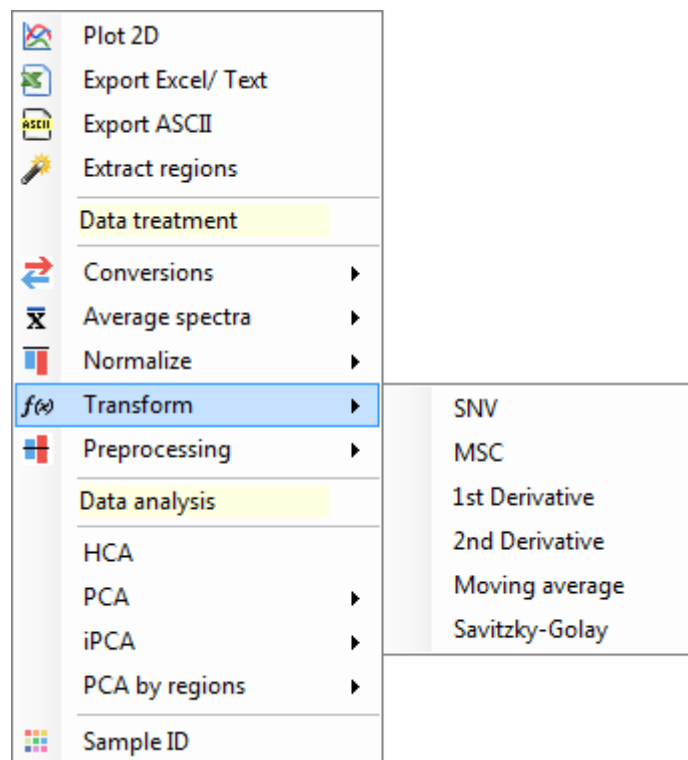


Figura 35. Menu principal de operações - funções de transformação.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

As Figuras 36 e 37 exibem o conjunto de espectros dos óleos vegetais na faixa entre 5500 e 6000  $\text{cm}^{-1}$ , normalizados entre os limites de zero e um e corrigidos pelo método do valor normal padrão, pela opção “SNV”, empregando ChemoStat e Matlab<sup>®</sup>, respectivamente.

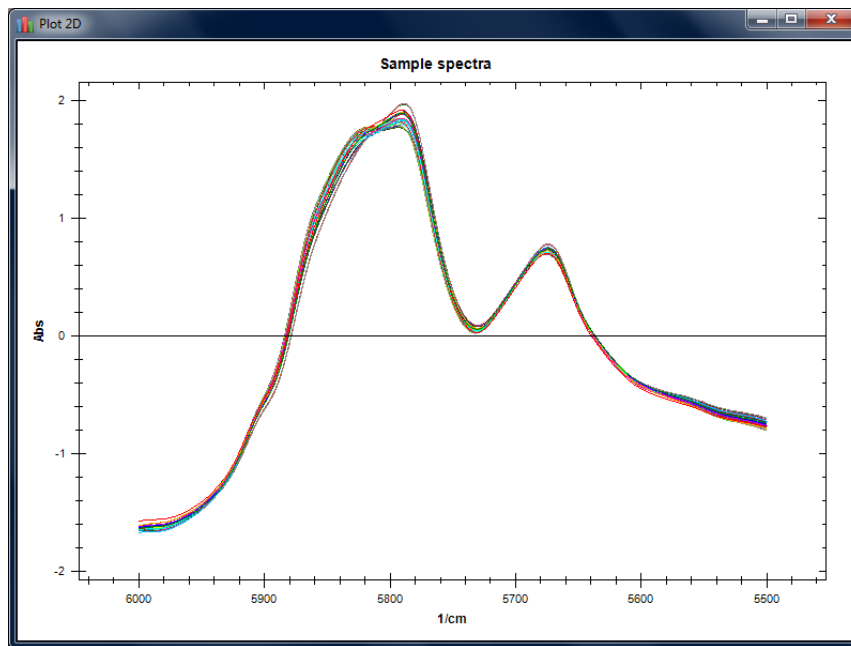


Figura 36. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados e com aplicação de SNV, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – ChemoStat.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

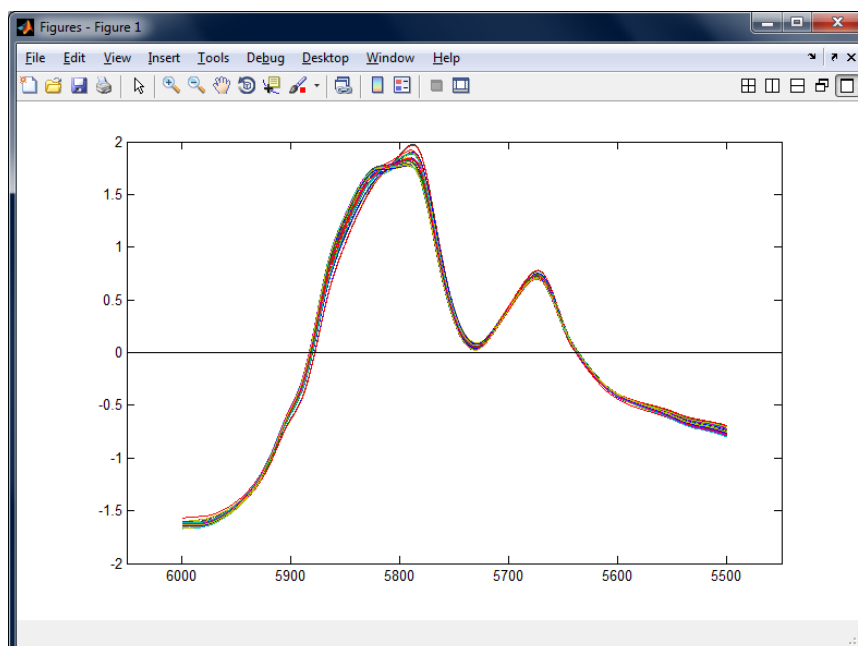


Figura 37 Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados e com aplicação de SNV, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – Matlab<sup>®</sup>  
Fonte: Autor, extraído do *software* Matlab<sup>®</sup>, 2014.

Já as Figura 38 e 39 aplicam a primeira derivada numa janela de 5 pontos nos dados da Figura 36 e 37, empregando ChemoStat e Matlab®, respectivamente.

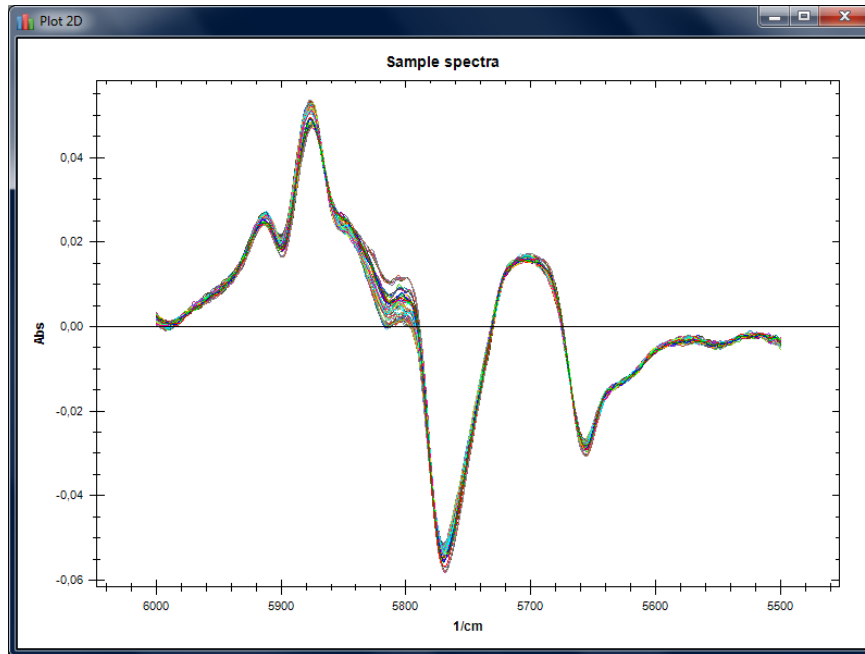


Figura 38. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com aplicação de SNV e primeira derivada (5 pontos), na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – ChemoStat. Fonte: Autor, extraído do *software* ChemoStat, 2014.

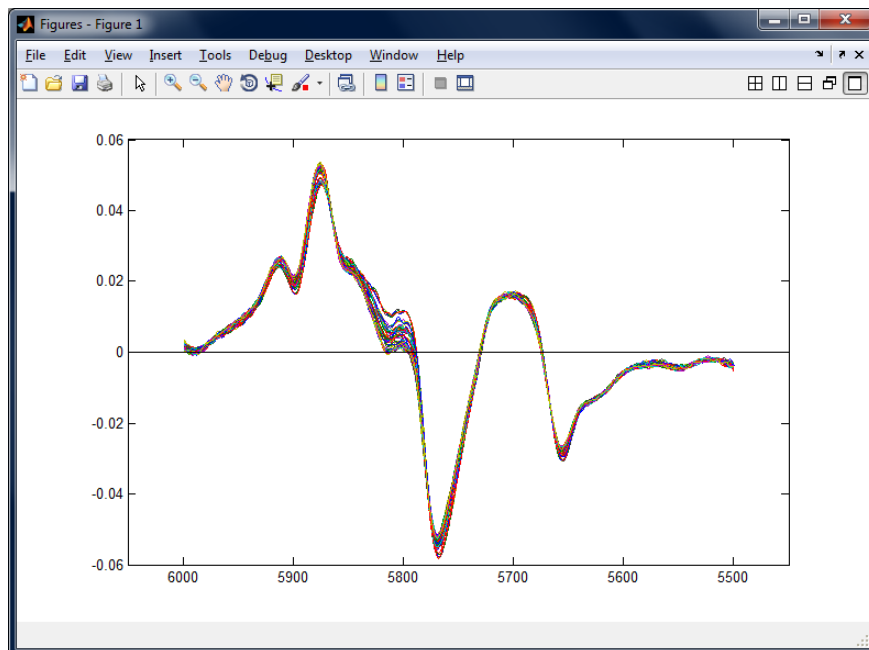


Figura 39. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com aplicação de SNV e primeira derivada (5 pontos), na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – Matlab®. Fonte: Autor, extraído do *software* Matlab®, 2014.

### 5.2.10 Funções de pré-processamentos

O *software* permite também autoescalar os dados “*Autoscale*” ou centrá-los na média “*Meancenter*”. Para que isto ocorra basta acionar as opções disponíveis no item “*Pre-processing*”, apresentada na Figura 40. Vale ressaltar que essas rotinas já estão pré-estabelecidas nos métodos PCA e iPCA, e caso necessitam serem utilizadas na grade de dados, ao utilizar a PCA ou iPCA, deve-se escolher a opção “*None*”, discutida posteriormente.

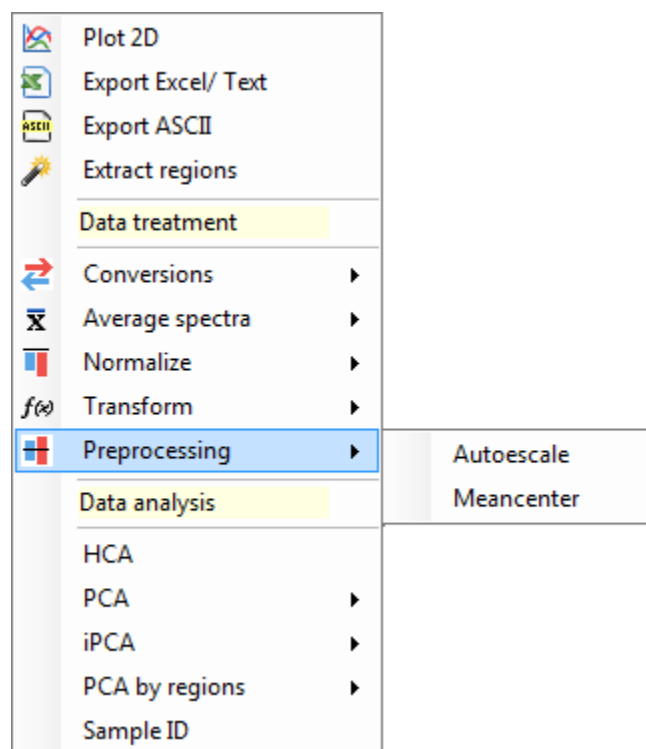


Figura 40. Menu principal de operações - funções de pré-processamento.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

As Figuras 41 e 42 exibem o conjunto de espectros dos óleos vegetais na faixa entre  $5500$  e  $6000\text{ cm}^{-1}$ , normalizados entre os limites zero e um, corrigidos por SNV e primeira derivada, além do método de centrar os dados na média, acionado pela opção “*Meancenter*”, empregando ChemoStat e Matlab®, respectivamente.

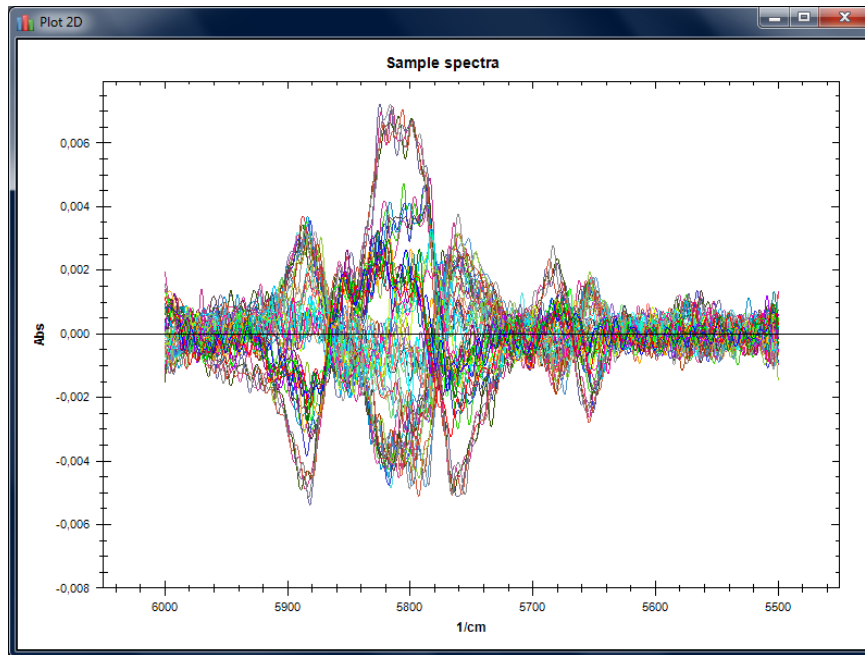


Figura 41. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – ChemoStat  
 Fonte: Autor, extraído do *software* ChemoStat, 2014.

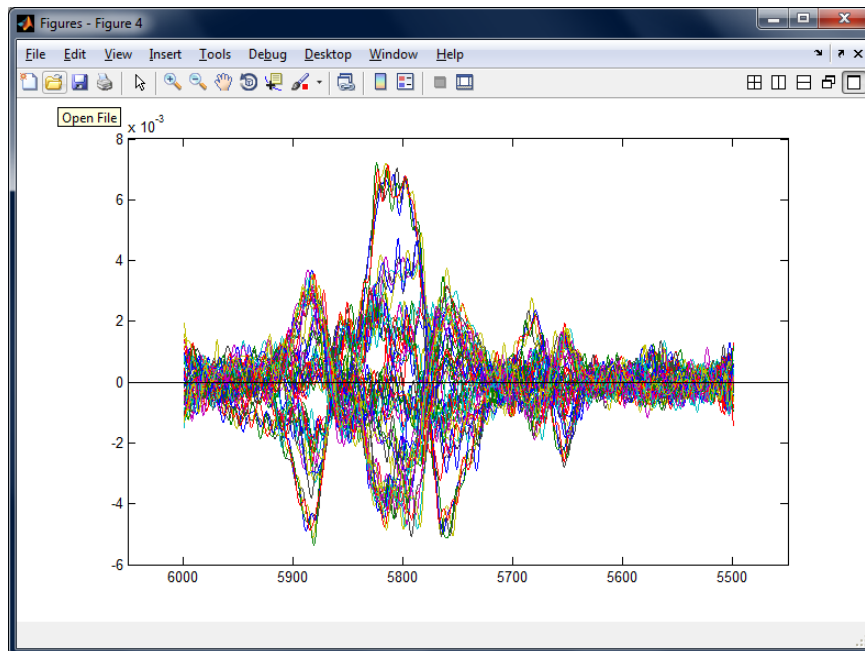


Figura 42. Gráfico do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – Matlab®.  
 Fonte: Autor, extraído do *software* Matlab®, 2014.



### 5.2.11 Identificação de amostras

O *software* apresenta uma função de classificação das amostras. Chamada de “*Sample ID*”, a mesma permite a identificação a partir de cores pré-estabelecidas através da análise sintática do nome da amostra. Para que isto ocorra, deve ser preenchido o campo “*Characters*” com um número de caracteres semelhantes entre as replicatas. O botão “*Check name*” executa a função, predefinindo uma cor de acordo com o código de identificação da amostra (tabela de cores). O usuário pode alterar a identificação, e consequentemente as cores, digitando um número no campo “*Code*”, entre os campos “*Name*” e “*Color*”. O botão “*Save & Close*” salva as configurações de cor e fecha a janela.

Na Figura 43 são ilustrados o momento em que a janela é apresentada e após o botão “*Check name*” ser acionado. Neste caso foi realizada a identificação das amostras por cores pela análise sintática de 5 caracteres.

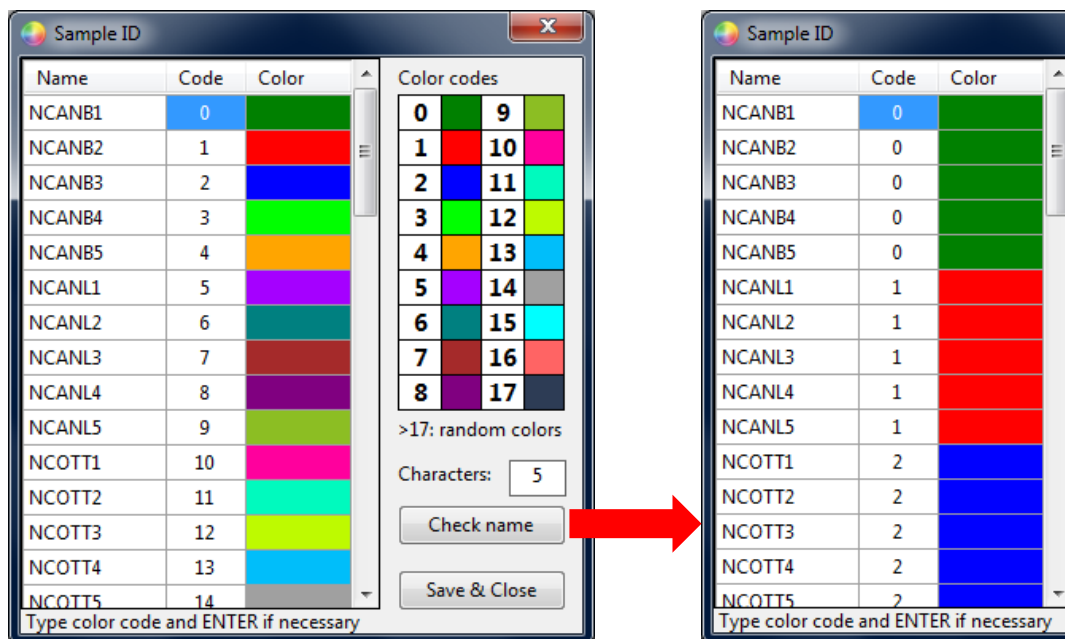


Figura 43. Tela para identificação das amostras por classe.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

As amostras classificadas nessa função apresentarão as cores correspondentes quando plotadas na PCA, iPCA (ambos resultados de scores) e HCA, além de servir como orientação na realização do espectro médio por classe (“*Average spectra >> By class*”).

### 5.2.12 Algoritmo PCA

O *software* ChemoStat possibilita a análise de componentes principais (“PCA”). Para que isto ocorra, basta clicar no item PCA no menu e escolher uma das opções: “*Meancenter*”, para centrar os dados na média; “*Autoscale*” para autoescalar os dados; ou “*None*”, para nenhum pré-processamento, ou seja, para quando algum dos pré-processamentos já tenha sido realizado em etapas anteriores à grade de dados, de acordo com a Figura 44.

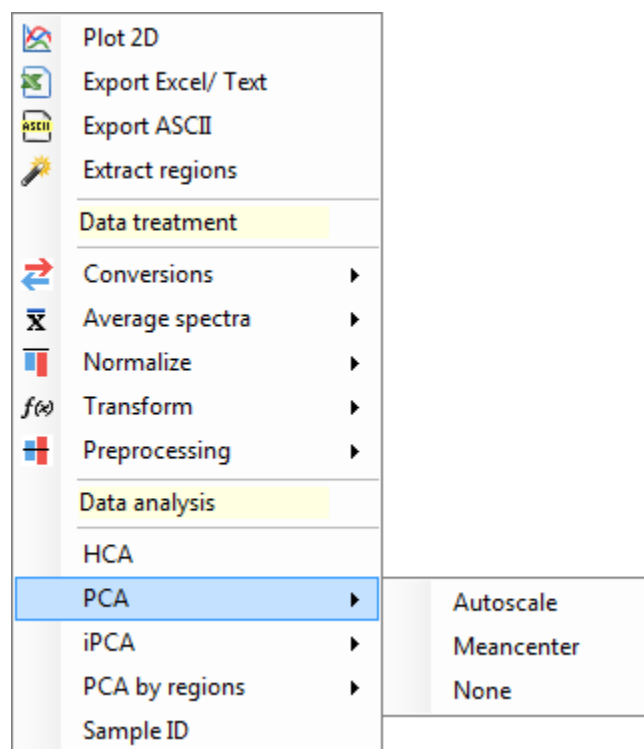


Figura 44. Menu principal de operações - funções de pré-processamento para PCA.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

As Figuras 45 e 46 exibem o gráfico de scores a partir da análise dos componentes principais (PCA) aplicada ao conjunto de espectros dos óleos vegetais na faixa entre 5500 e 6000  $\text{cm}^{-1}$ , normalizados entre os limites zero e um, corrigidos pelo método do valor normal padrão (“*SNV*”) e, posteriormente, centrados na média, empregando ChemoStat e Matlab<sup>®</sup>, respectivamente.

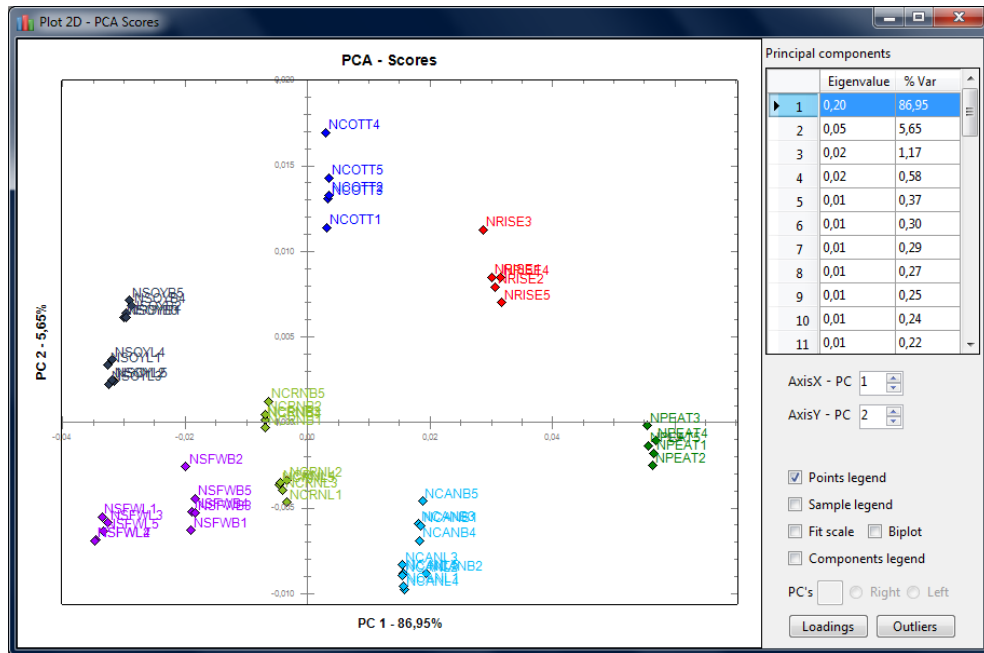


Figura 45. Gráfico de *scores* PC1 x PC2 do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – ChemoStat.

Fonte: Autor, extraído do *software* ChemoStat, 2014.

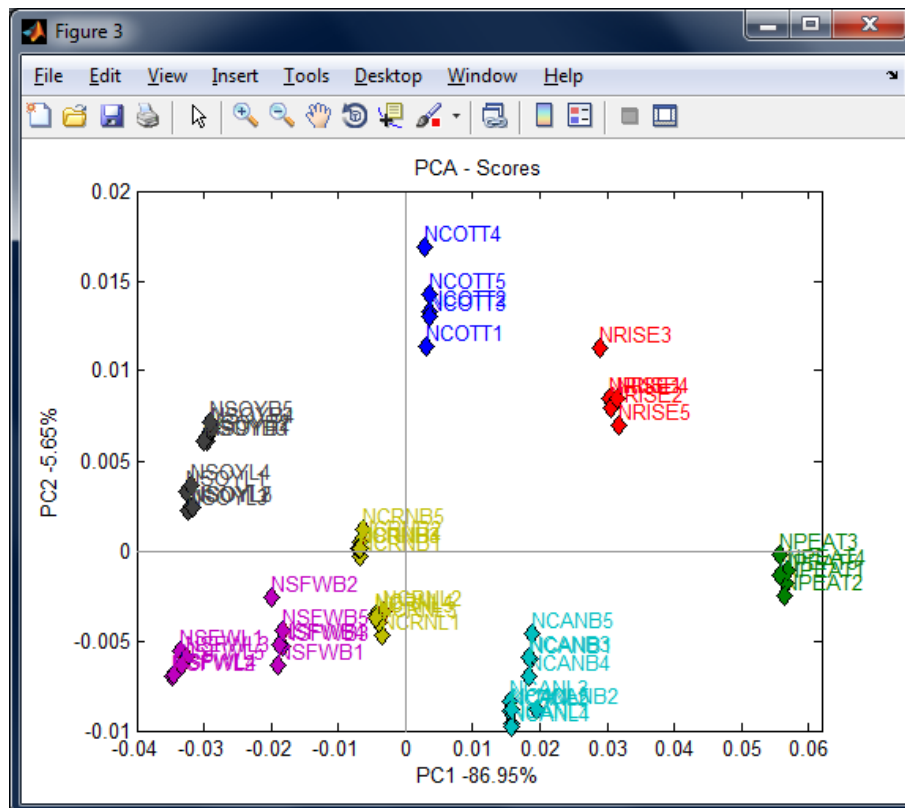


Figura 46. Gráfico de *scores* PC1 x PC2 do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – Matlab®.

Fonte: Autor, extraído do *software* Matlab®, 2014.

A tela com gráfico de *scores* apresenta comandos adicionais, conforme ilustra a Figura 45. Os campos “*AxisX – PC*” e “*AxisY – PC*” permitem navegar entre as componentes principais calculadas. A caixa de checagem “*Points legends*” insere o nome das amostras sobre os pontos no gráfico, enquanto que “*Sample legend*” inclui um campo com o nome das amostras abaixo do título do gráfico. Para configurar os eixos do gráfico na mesma escala basta acionar a opção “*Fit scale*”, o ajuste ocorre automaticamente, assim como “*Biplot*” permite a visualização de dois gráficos simultaneamente, além do gráfico dos *scores*, é adicionado o gráfico dos *loadings*, ambos nas mesmas componentes apresentadas. A caixa de checagem “*Components legends*” inclui o valor das variáveis latentes e sua soma acumulada, ambos em valores percentuais, no lado direito, se acionado o campo “*Right*”, ou esquerdo, caso acionado o campo “*Left*”. O campo “*PC’s*” define quantas variáveis serão mostradas. O botão “*Loadings*”, quando pressionado, abre uma nova janela, apresentando os resultados dos *loadings* enquanto que o botão “*Outliers*” exhibe dados relativos às amostras anômalas calculadas através do método multivariado  $T^2$  de Hotelling, posteriormente ilustrados neste capítulo.

Clicando com o botão direito do “*mouse*” em cima da área do gráfico aparecerão opções como ajuste de escala, copiar figura para área de transferência, exportar para formatos de imagens, além da opção “*Export scores*” para exportação dos valores de *scores*. Os valores de *scores* exportados podem ser plotados em três dimensões. Para isso foi desenvolvido em recurso extra, chamado “*Plot 3D*”, cujas especificações encontram-se no Anexo A.

Os resultados obtidos pelo gráfico dos *scores* da PC1 x PC2 (Figura 45), é possível observar que a PC1 separa os óleos de milho (verde-claro), soja (preto) e girassol (rosa), em valores positivos, das amostras de algodão (azul), amendoim (verde), arroz (vermelho) e canola (azul-claro), em valores negativos. As amostras de milho, soja e girassol se agrupam por possuírem maiores quantidades de ácidos graxos poli-insaturados em sua composição (Tabela 1). Em relação às demais, a PC2 separa os óleos com maior concentração de ácidos graxos monoinsaturados (canola e amendoim), em valores positivos, dos óleos com maior concentração de ácidos graxos saturados (algodão e arroz), em valores negativos.

As Figuras 47 e 48 exibem os gráficos de *loadings* para a PC1, e as Figuras 49 e 50 para a PC2, empregando ChemoStat e Matlab<sup>®</sup>, respectivamente.

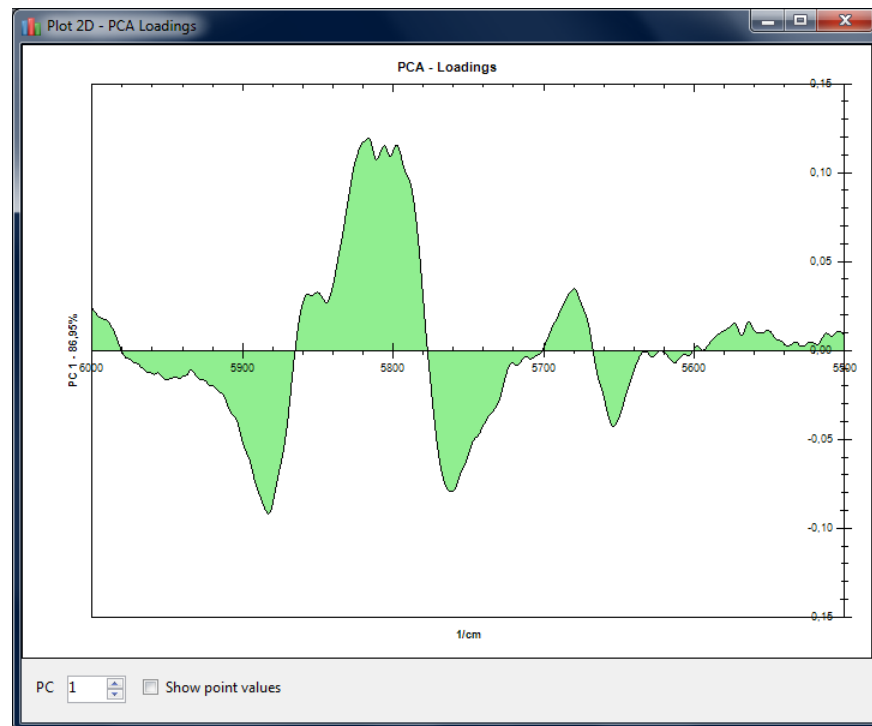


Figura 47. Gráfico de *loadings* (PC1) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – ChemoStat.

Fonte: Autor, extraído do software ChemoStat, 2014.

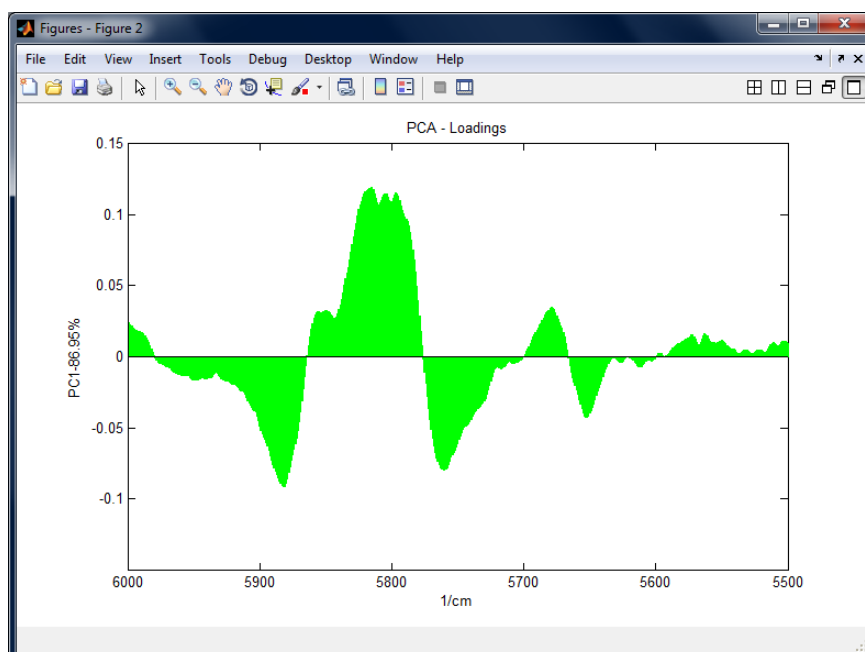


Figura 48. Gráfico de *loadings* (PC1) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – Matlab®.

Fonte: Autor, extraído do software Matlab®, 2014.

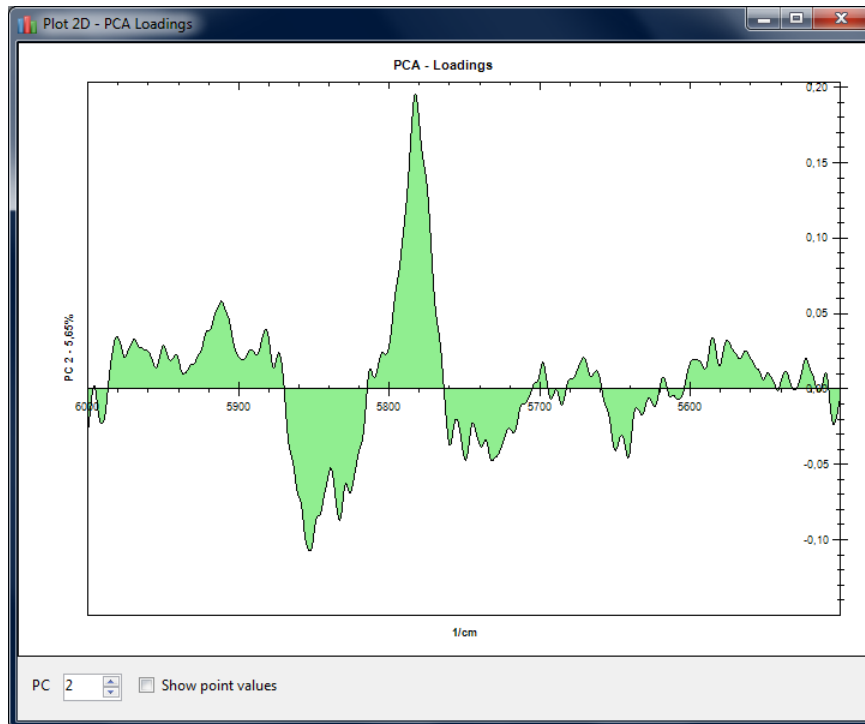


Figura 49. Gráfico de *loadings* (PC2) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – ChemoStat.

Fonte: Autor, extraído do *software* ChemoStat, 2014.

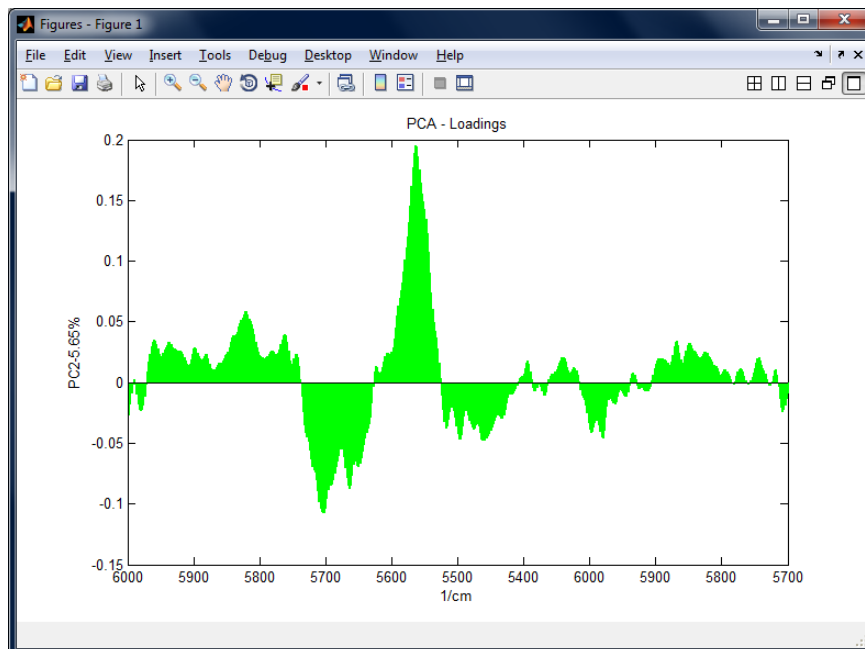


Figura 50. Gráfico de *loadings* (PC2) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – Matlab®.

Fonte: Autor, extraído do *software* Matlab®, 2014.

As Figuras 51 e 52 apresentam as projeções dos *scores* e *loadings* utilizando a opção “*Biplot*”, para os óleos vegetais estudados na faixa espectral entre 5500 e 6000  $\text{cm}^{-1}$ , empregando ChemoStat e Matlab®, respectivamente.

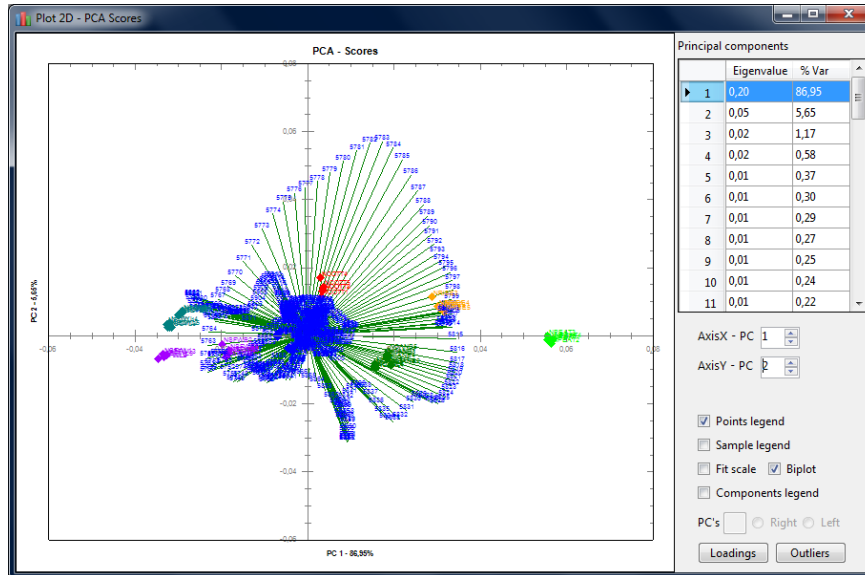


Figura 51. Gráfico *biplot* de *scores* e *loadings* (PC1 x PC2) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – ChemoStat.

Fonte: Autor, extraído do software ChemoStat, 2014.

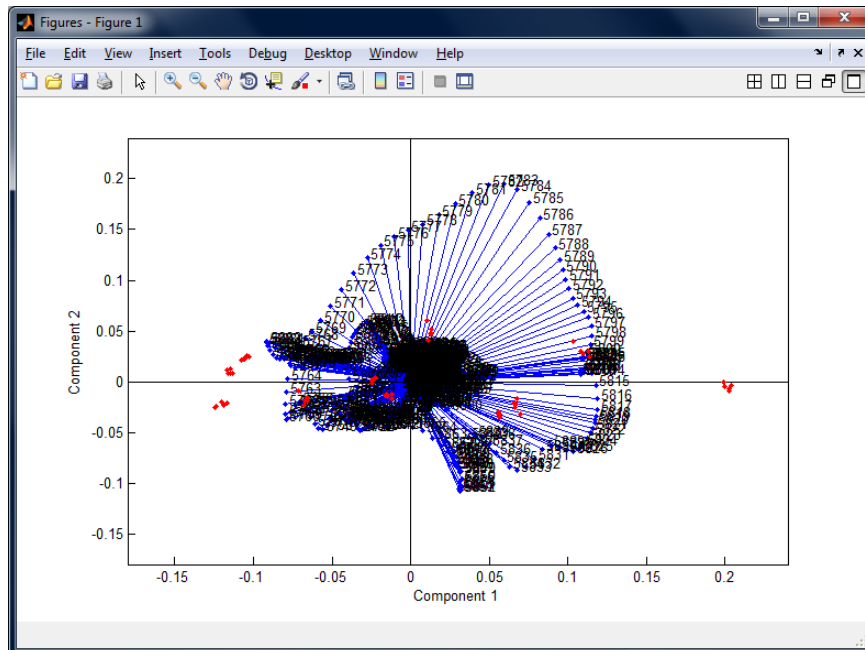


Figura 52. Gráfico *biplot* de *scores* e *loadings* (PC1 x PC2) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – Matlab®.

Fonte: Autor, extraído do software Matlab®, 2014.

Por último, para detecção de amostras anômalas, a função executada pelo botão “Outliers” primeiramente solicita ao usuário o nível de confiança, ou o *alpha* da distribuição de Fisher-Snedecor, para o cálculo da decomposição do  $T^2$  de Hotelling, de acordo com a janela representada pela Figura 53.

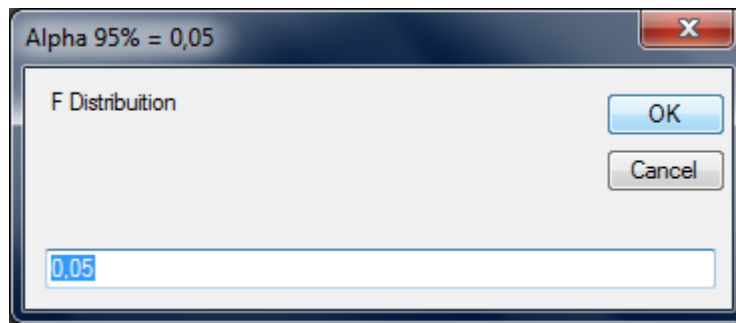


Figura 53. Janela de diálogo para entrada de valor referente ao *alpha* para distribuição de Fisher-Snedecor.

Fonte: Autor, extraído do software ChemoStat, 2014.

Após informado o *alpha*, aparecerá o gráfico de outliers calculado a partir da decomposição do  $T^2$  de Hotelling para dados multivariados, conforme ilustra a Figura 54.

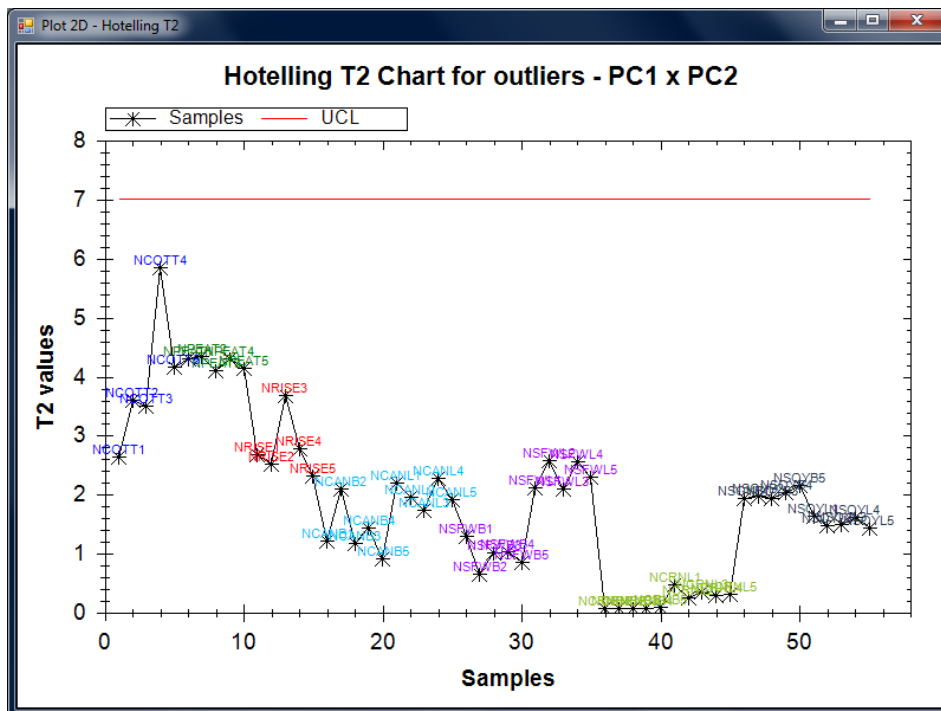


Figura 54. Gráfico para  $T^2$  de Hotelling (PC1 x PC2) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados, com SNV, primeira derivada e centrados na média, na faixa entre 5500 e 6000  $\text{cm}^{-1}$  – ChemoStat.

Fonte: Autor, extraído do software ChemoStat, 2014.



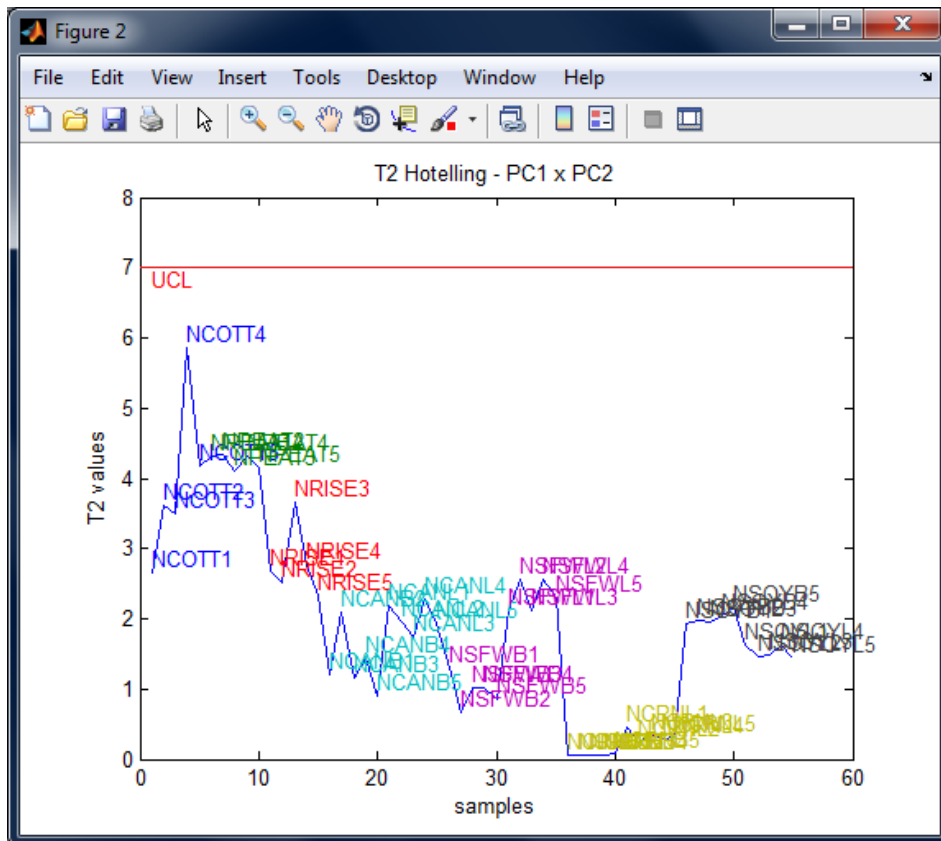


Figura 55. Gráfico para  $T^2$  de Hotelling (PC1 x PC2) do conjunto de espectros dos óleos vegetais (FT-NIR), normalizados com SNV, primeira derivada e centrados na média, na faixa entre  $5500$  e  $6000\text{ cm}^{-1}$  – Matlab®.

Fonte: Autor, extraído do software Matlab®, 2014.

Vale ressaltar, caso houvesse alguma amostra anômala (*outlier*), que a(s) mesma(s) ficaria(m) acima do limite superior de controle (*UCL*, do inglês, “*Upper Control Limit*”). O mesmo procedimento foi adotado no Matlab®, tendo os mesmos resultados, de acordo com a Figura 55.

Como pode ser observado nas comparações realizadas anteriormente, os resultados obtidos na análise de componentes principais (PCA) a partir do software ChemoStat estão em completa concordância com os resultados obtidos no Matlab®.

### 5.2.13 Função PCA por regiões (“by regions”)

A função “*PCA by regions*”, acionada conforme demonstra a Figura 56, realiza uma operação de PCA para cada região informada na seção 2, separadamente. As opções disponíveis são “*Meancenter*”, “*Autoscale*” e “*None*”.

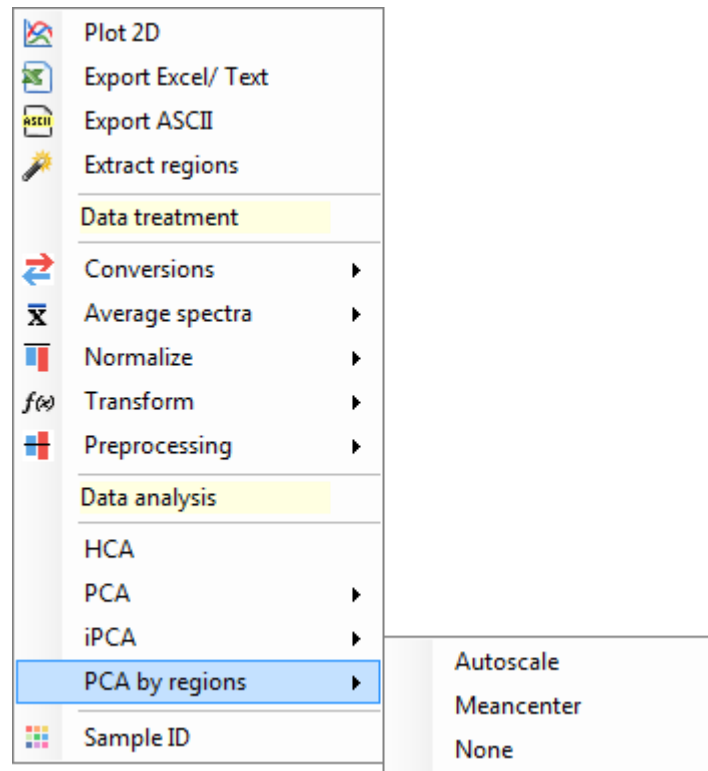


Figura 56. Menu principal de operações - funções de pré-processamento para “PCA by region”.  
 Fonte: Autor, extraído do *software* ChemoStat, 2014.

#### 5.2.14 Análise por Componentes Principais em intervalos – iPCA

O *software* ChemoStat permite a análise dos componentes principais por intervalo (iPCA), muito utilizada para seleção de variáveis espectrais. Ao selecionar o item “iPCA” no menu, aparecem três opções, referentes ao pré-processamento de dados, autoescalados (“*Autoscale*”), centrados na média (“*Meancenter*”) ou nenhum (“*None*”), de acordo com a Figura 57.

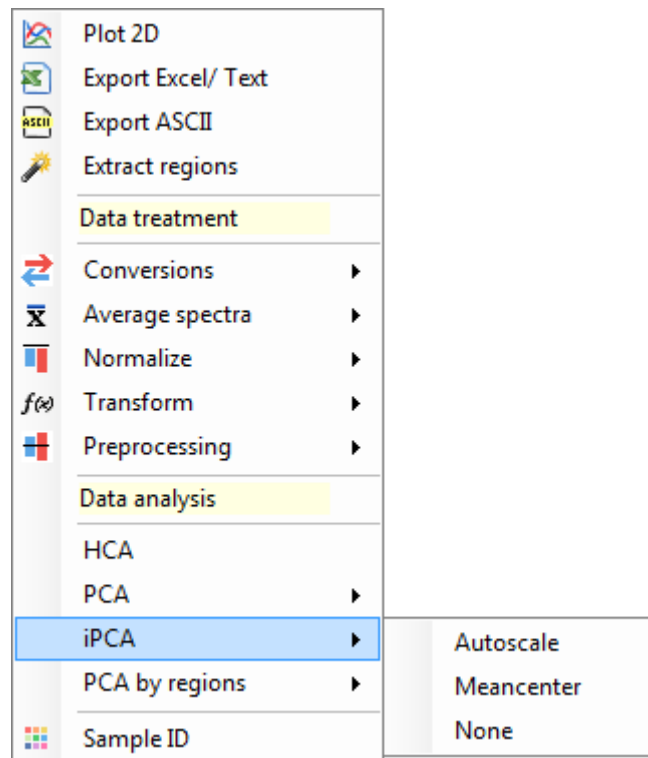


Figura 57. Menu principal de operações - funções de pré-processamento para iPCA.  
 Fonte: Autor, extraído do *software* ChemoStat, 2014.

Antes de proceder com a iPCA, foi realizada uma PCA para os dados autoescalados no infravermelho médio dos óleos vegetais com toda a faixa espectral, ou seja, de  $650$  a  $4.000\text{ cm}^{-1}$ , previamente normalizados na faixa entre 1 e 0, aplicada SNV e primeira derivada com janela de 5 pontos. Como pode ser verificado nas Figuras 58 e 59, utilizando toda faixa espectral, não se obteve uma satisfatória separação dos óleos vegetais, empregando ChemoStat e Matlab®, respectivamente.

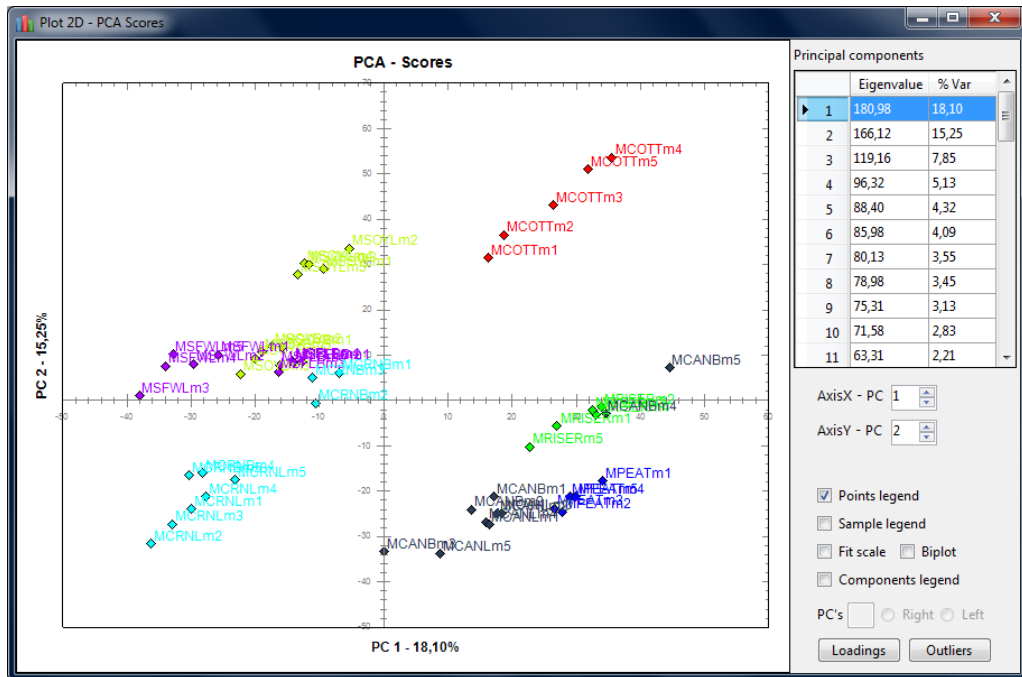


Figura 58. Gráfico de scores (PC1 x PC2) do conjunto de espectros dos óleos vegetais (FT-MIR), autoescalados e previamente normalizados, com SNV e primeira derivada, na faixa entre 650 e 4000  $\text{cm}^{-1}$  – ChemoStat.

Fonte: Autor, extraído do *software* ChemoStat, 2014.

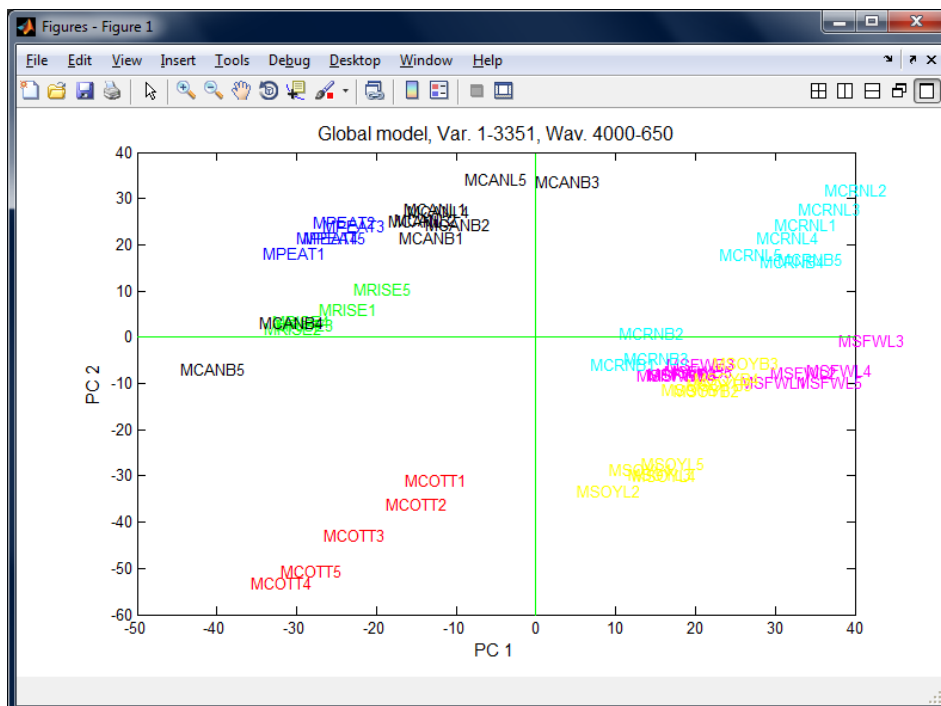


Figura 59. Gráfico de scores (PC1 x PC2) do conjunto de espectros dos óleos vegetais (FT-MIR), autoescalados e previamente normalizados, com SNV e primeira derivada, na faixa entre 650 e 4000  $\text{cm}^{-1}$  – Matlab®.

Fonte: Autor, extraído do *software* Matlab®, 2014.

Analisando as Figuras 58 e 59, observam-se que os resultados da PC1 e da PC2 do *software* ChemoStat estão espelhadas, ou seja, possuem sinal oposto em

relação ao *software* Matlab<sup>®</sup>. Este efeito pode ocorrer somente em uma, nas duas primeiras ou em mais componentes principais. Scott Ramos, cientista-chefe da empresa Infometrix<sup>®</sup>, comentou, no fórum de algoritmos do Pirouette – *software* de quimiometria - em 14/11/2005, Anexo B, que isto é um resultado esperado. O algoritmo principal por trás da etapa de decomposição dos dados em PCA é baseado no método de potenciação. Este algoritmo faz a decomposição de um autovetor de cada vez. Para derivar o primeiro autovetor, o algoritmo constitui um vetor de trabalho que é então otimizado através de uma série de iterações. Esse vetor inicial de trabalho pode ser um vetor de uns, ou uns e zeros, ou números aleatórios, ou ainda um vetor extraído do conjunto de dados. Uma vez que as iterações convergem e o primeiro autovetor é produzido, a informação representada neste primeiro fator é "removida" do conjunto de dados, e esta nova matriz de dados é, então, processada para extrair o segundo autovetor, etc. Este processo é repetido em número de vezes iguais ao *ranking* da matriz até que todos os autovetores sejam computados, sendo possível reconstruir a matriz de dados original exatamente pela multiplicação dos pontos (*ranking* completo) por estes autovetores. Caso seja trocado o sinal em todos os valores em qualquer par de autovetores, a matriz reconstruída será a mesma. Assim, o sinal não é crítico, mas as magnitudes são.

Selecionado o item iPCA no menu principal de operações (Figura 57), após executada uma das opções, “Autoscale” ou “Meancenter”, o *software* solicitará ao usuário, via caixa de diálogo (Figura 60), o número de intervalos pelo qual será dividido a região espectral.

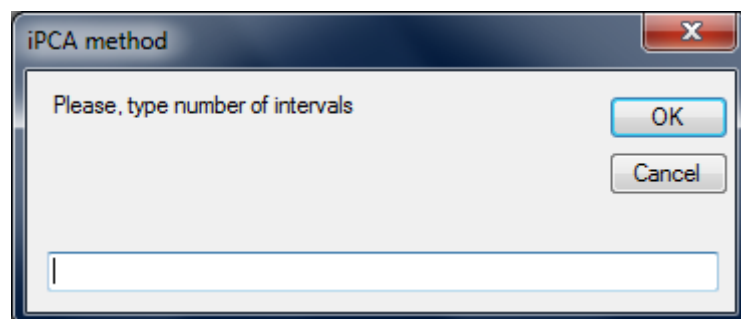


Figura 60. Caixa de diálogo para entrada de valores referente ao intervalo de iPCA.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

Logo após informado o número de intervalos e pressionado o botão “OK”, o *software* realiza a PCA para cada intervalo separadamente, permitindo uma

visualização global, num mesmo gráfico e em várias janelas mostrando gráficos de scores em múltiplos de quatro.

Com os dados dos óleos vegetais obtidos por FT-MIR, foram realizadas iPCA com as subdivisões em 8, 16 e 32 intervalos. O melhor desempenho foi alcançado dividindo o espectro em 32 intervalos. Na Figura 61 é exibida a janela de visão global com os 32 gráficos de scores (PC1 x PC2).

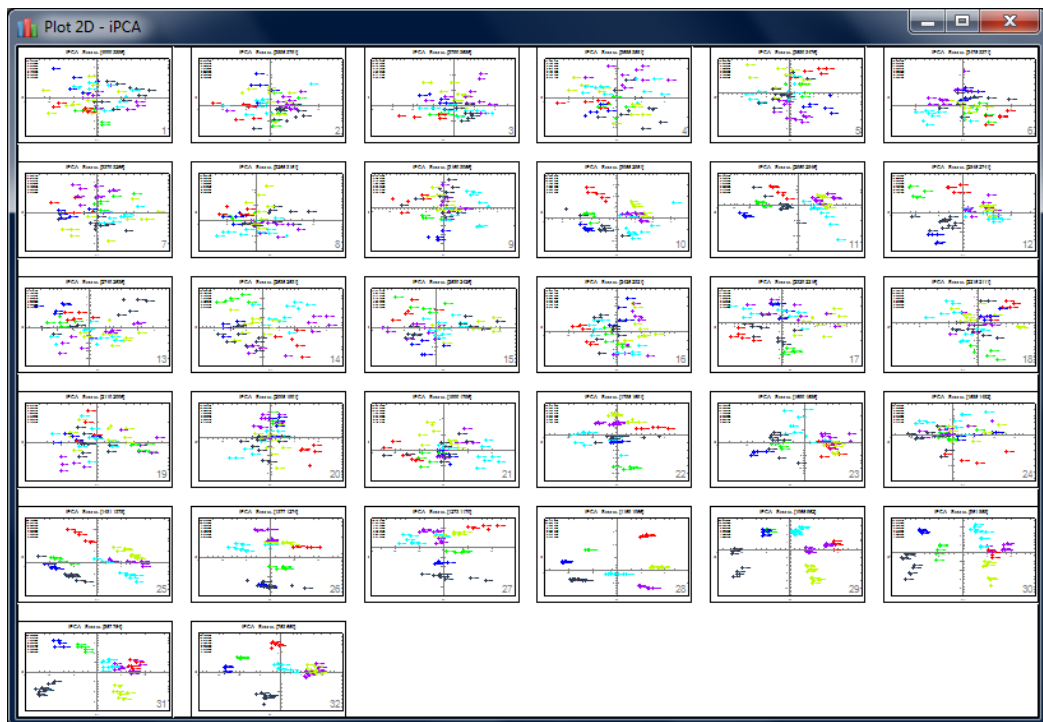


Figura 61. Janela com 32 gráficos de scores referente aos intervalos aplicados nos espectros de óleos vegetais (FT-MIR) - ChemoStat.

Fonte: Autor, extraído do *software* ChemoStat, 2014.

O *software* ChemoStat permite também a visualização de 4 gráficos de scores, ilustrado pela Figura 62. Já a Figura 63 apresenta o mesmo procedimento para o mesmo conjunto de dados a partir do aplicativo Matlab<sup>®</sup> com o pacote *iToolbox* (NORGAARD, 2006),

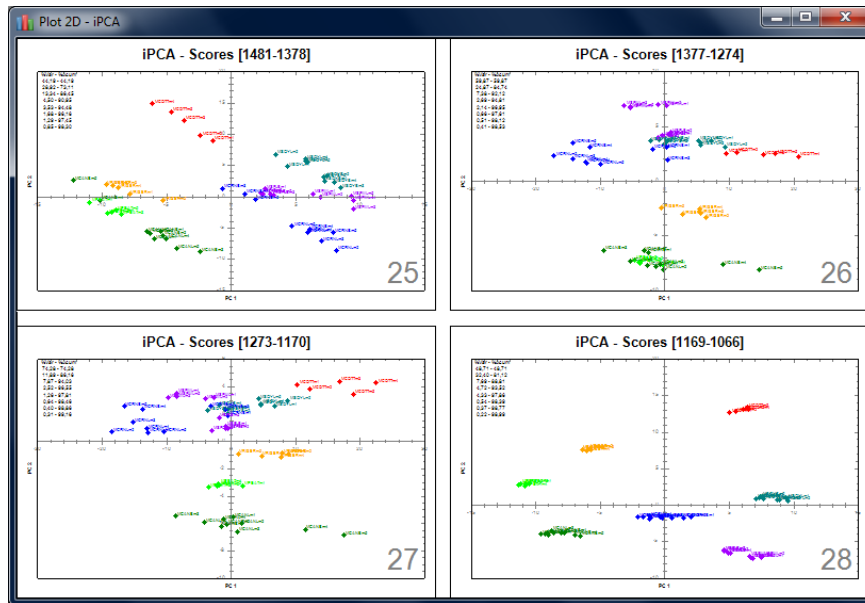


Figura 62. Janela com 4 gráficos de scores referente aos intervalos 25, 26, 27 27 e 28 aplicados nos espectros de óleos vegetais (FT-MIR) – ChemoStat.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

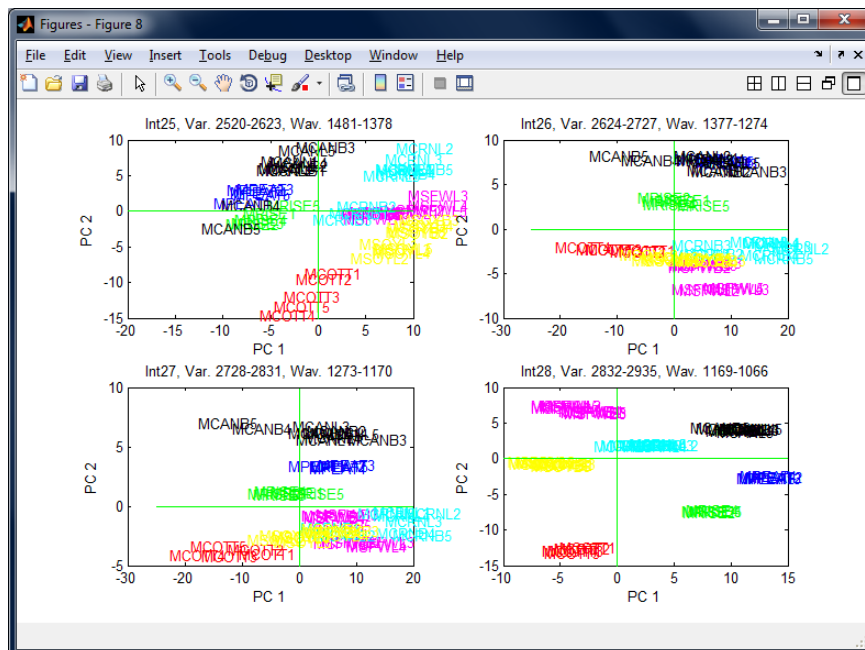


Figura 63. Janela com 4 gráficos de scores referente aos intervalos 25, 26, 27 e 28 aplicados nos espectros de óleos vegetais (FT-MIR) – Matlab®.  
Fonte: Autor, extraído do *software* Matlab®, 2014.

Tanto na visão global, quanto nas múltiplas de quatro, caso algum gráfico seja clicado, o mesmo será aberto de forma individualizada, como ilustra a Figura 64 sobre o intervalo de número 28, referente à sub-região que compreende a faixa entre 1066 e 1169  $\text{cm}^{-1}$ , empregando ChemoStat. A Figura 65 corresponde ao mesmo intervalo, porém utilizando o aplicativo Matlab®.

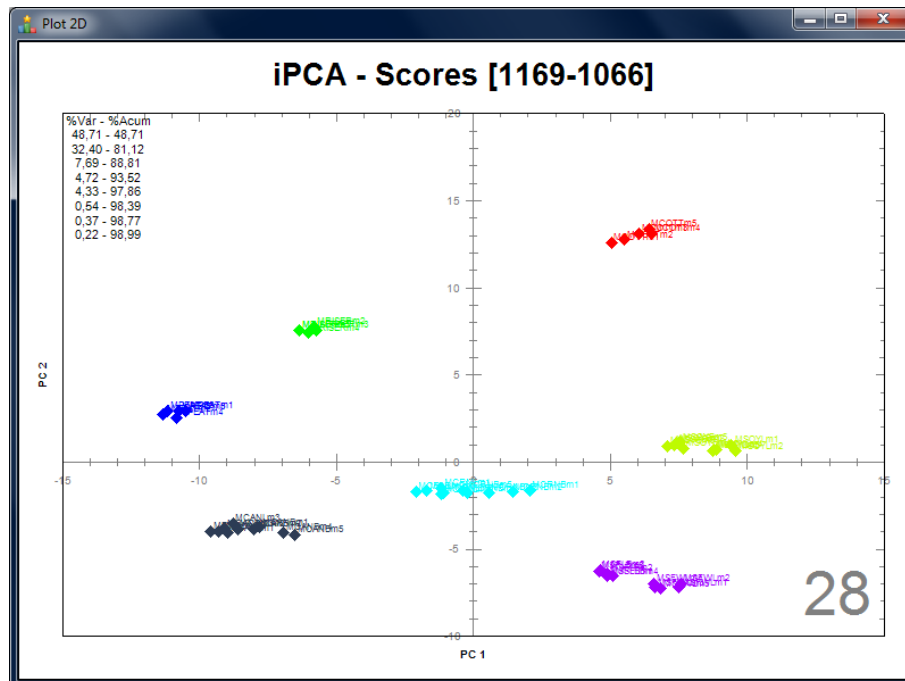


Figura 64. Janela com o gráfico de scores referente ao intervalo 28 aplicados nos espectros de óleos vegetais (FT-MIR) - ChemoStat.

Fonte: Autor, extraído do *software* ChemoStat, 2014.

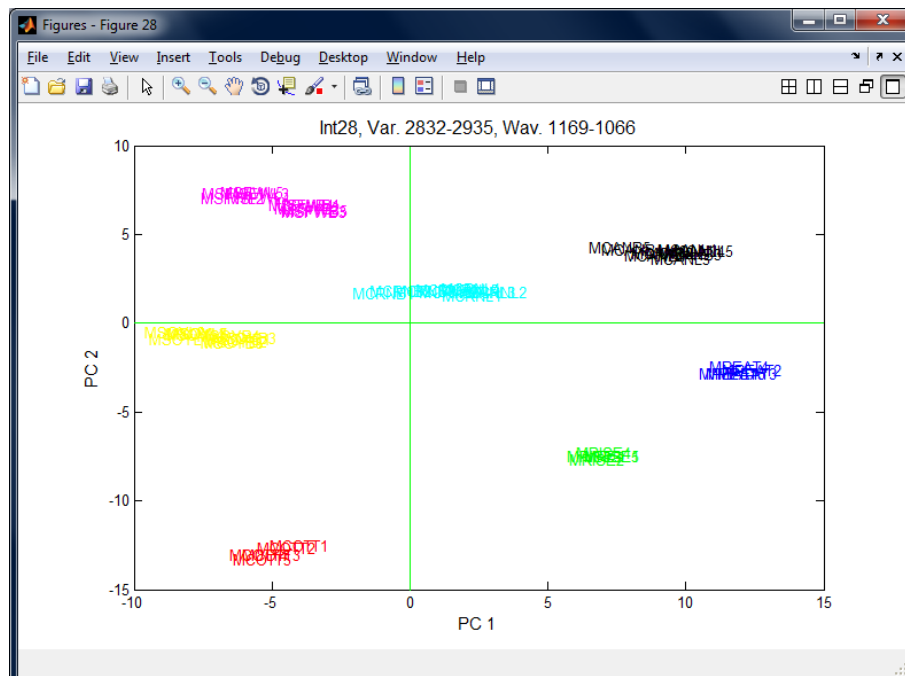


Figura 65. Janela com o gráfico de scores referente ao intervalo 28 aplicados nos espectros de óleos vegetais (FT-MIR) - Matlab®.

Fonte: Autor, extraído do *software* Matlab®, 2014.



Ainda na janela da visão global, quando utilizado o botão direito do “*mouse*” para visualização do menu de opções padrão, aparecerá ainda o item “*Spectra plotting*” (Figura 66).

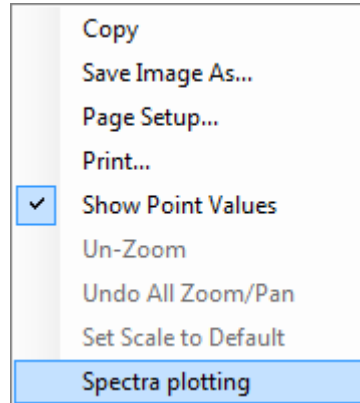


Figura 66. Menu de opções sobre o gráfico de scores do método iPCA.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

Esse algoritmo permite a junção do gráfico dos espectros divididos pelos intervalos informados assim como suas componentes principais. Para tanto, basta informar o número de componentes principais desejadas através de uma caixa de diálogo. Nessa mesma caixa, Figura 67, o *software* calculará o número máximo de componentes principais calculadas.

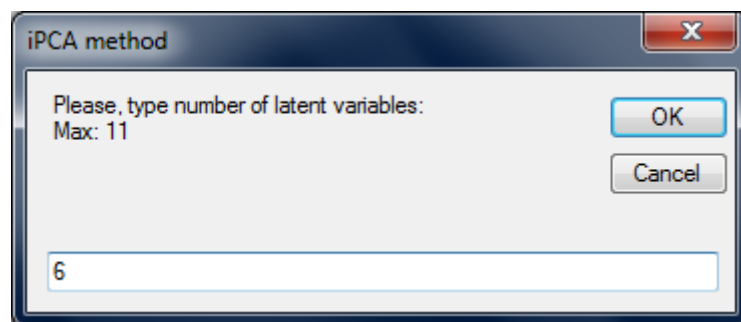


Figura 67. Caixa de diálogo para entrada de valores referente ao número de componentes principais.

Fonte: Autor, extraído do *software* ChemoStat, 2014.

O resultado dessa operação é ilustrado na Figura 68, onde fora escolhido 6 componentes principais. A mesma operação foi realizada empregando Matlab® conforme mostra a Figura 69.

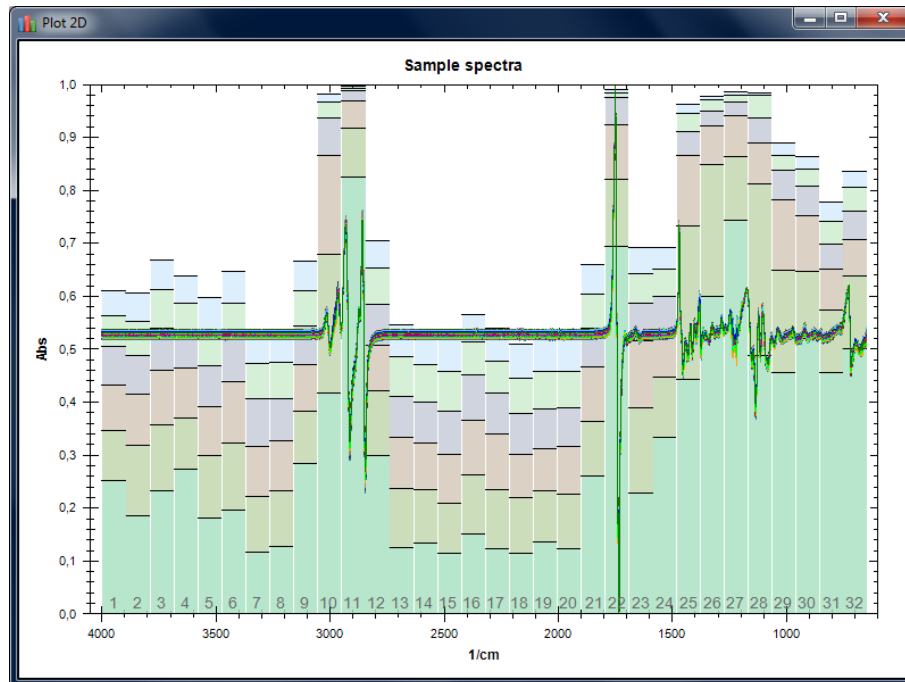


Figura 68. Variação percentual das componentes principais divididas em 32 intervalos aplicados nos espectros de óleos vegetais (FT-MIR) - ChemoStat.  
 Fonte: Autor, extraído do *software* ChemoStat, 2014.

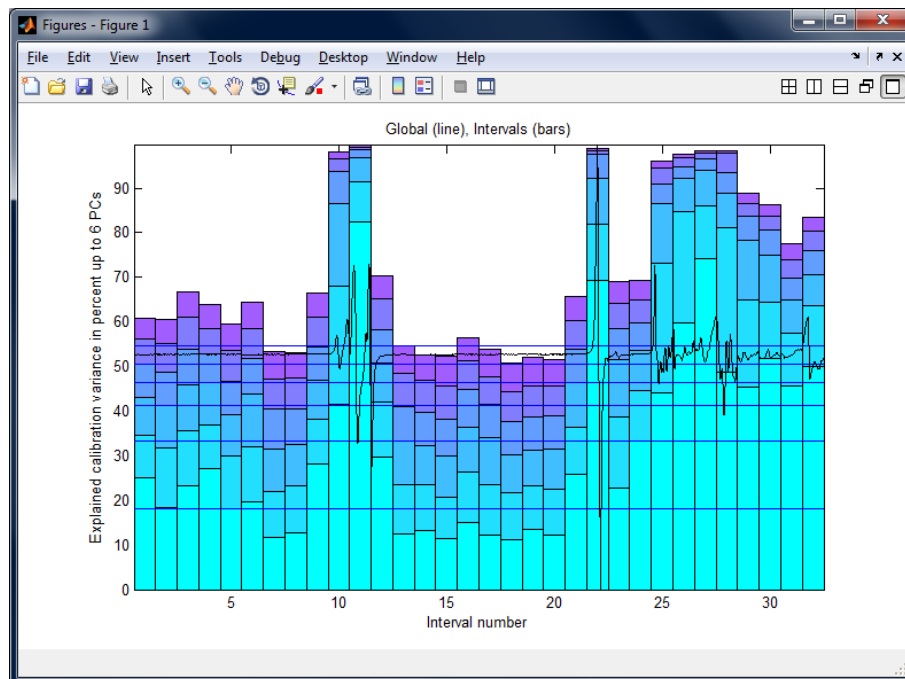


Figura 69. Variação percentual das componentes principais divididas em 32 intervalos aplicados nos espectros de óleos vegetais (FT-MIR) - Matlab®.  
 Fonte: Autor, extraído do *software* Matlab®, 2014.

As alturas das barras representam, em forma percentual, a variância contida em cada componente principal para cada intervalo. A linha traçada horizontalmente

representa a variância de cada uma das componentes principais para a análise de PCA para toda a informação do espectro. Os valores das variâncias referentes ao intervalo 28 são apresentados na Tabela 4, empregando ChemoStat e Matlab<sup>®</sup>.

Tabela 4. Valores para variância e variância acumulada das seis primeiras componentes principais dos dados de óleos vegetais - ChemoStat e Matlab<sup>®</sup>.

PC	ChemoStat		Matlab <sup>®</sup>	
	var (%)	var acum (%)	var (%)	var acum (%)
1	48,71	48,71	48,71	48,71
2	32,40	81,12	32,40	81,12
3	7,69	88,81	7,69	88,81
4	4,72	93,52	4,72	93,52
5	4,33	97,86	4,33	97,86
6	0,54	98,39	0,54	98,39

Fonte: Autor, 2014.

A região espectral onde se obteve uma melhor discriminação dos óleos vegetais correspondeu à faixa entre 1.169 e 1.066  $\text{cm}^{-1}$  (intervalo 28). O intervalo 28 acumula 81,12% das informações nas duas primeiras componentes principais e observa-se a discriminação dos óleos vegetais. Segundo Yang, Irudayaraj & Paradkar (2005) e Guillén & Cabo (1997), a região entre 1.100 e 1.200  $\text{cm}^{-1}$  representa a vibração axial (estiramento) da ligação C-O dos ésteres (C-C(=O)-O) e a vibração angular (flexão) da ligação C-H dos grupos funcionais -CH<sub>2</sub> nos óleos e gorduras comestíveis. Semelhante ao observado no infravermelho próximo (FT-NIR) é possível observar no gráfico dos scores da PC1 x PC2 (Figura 63) que a PC1 separa os óleos de milho, soja e girassol, em valores positivos, das amostras de algodão, amendoim, arroz e canola, em valores negativos. Da mesma forma, a PC2 separa os óleos de canola e amendoim, em valores positivos, dos óleos de algodão e arroz, em valores negativos. A separação dos óleos vegetais está associada ao fato de que cada óleo/gordura difere na composição, comprimento e grau de insaturação de ácidos graxos nas cadeias dos triglicerídeos (MORETTO & FETT, 1998).

Os resultados obtidos na análise de componentes principais por intervalos (iPCA) no *software* ChemoStat estão em completa concordância com os resultados obtidos no Matlab<sup>®</sup>.

### 5.2.15 Algoritmo HCA

A opção “HCA”, Figura 70, quando aplicada na grade de dados, exhibe uma nova janela apresentando o resultado da Análise por Agrupamento Hierárquico. Caso seja necessário o emprego de alguma correção, transformação ou pré-processamento dos dados, os mesmos devem ser realizados previamente sobre a grade de dados.

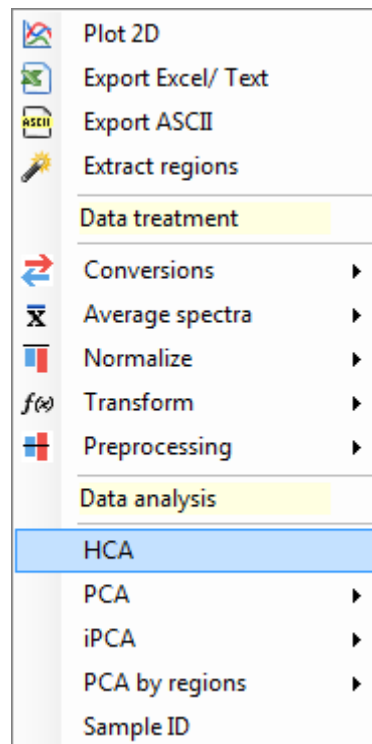


Figura 70. Menu principal de operações – função HCA.  
 Fonte: Autor, extraído do *software* ChemoStat, 2014.

Para realizar a validação da HCA com ligação completa (*complete linkage*), utilizaram-se as mesmas amostras e as mesmas regiões analisadas pela PCA (Figura 63). As Figuras 71 e 72 ilustram o resultado da HCA pelo método de ligação completa (“*Complete-Linkage*”), empregando ChemoStat e Matlab<sup>®</sup>, respectivamente, e as opções disponíveis no seletor como a ligação simples (“*Single-Linkage*”) e ligação pela média (“*Average-Linkage*”) no *software* ChemoStat.

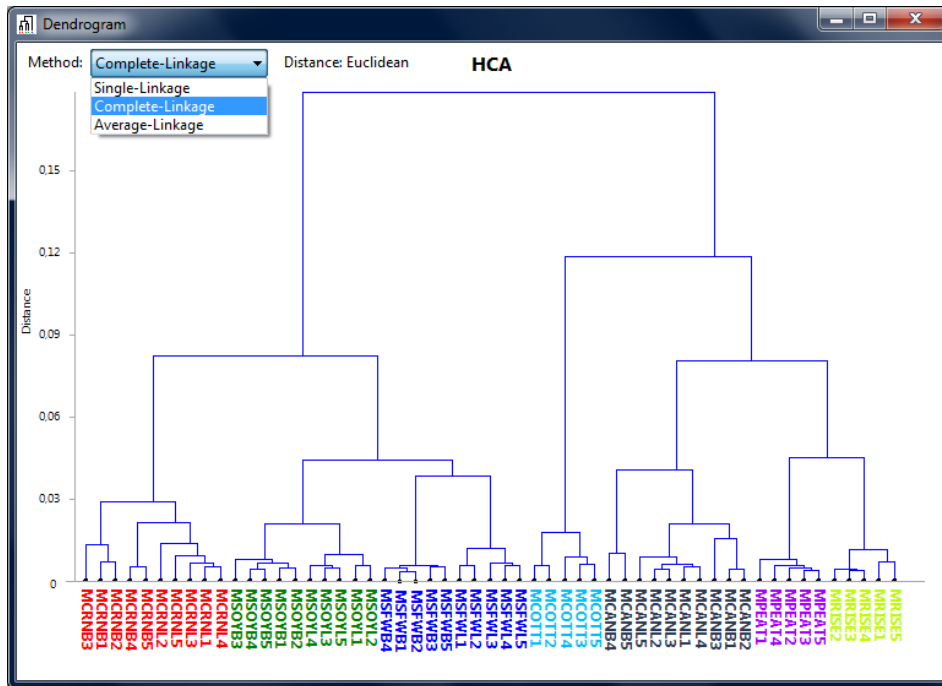


Figura 71. Dendrograma HCA – ligação completa – das amostras de óleos vegetais (FT- MIR), na região entre 1066 e 1169  $\text{cm}^{-1}$  - ChemoStat.

Fonte: Autor, extraído do software ChemoStat, 2014.

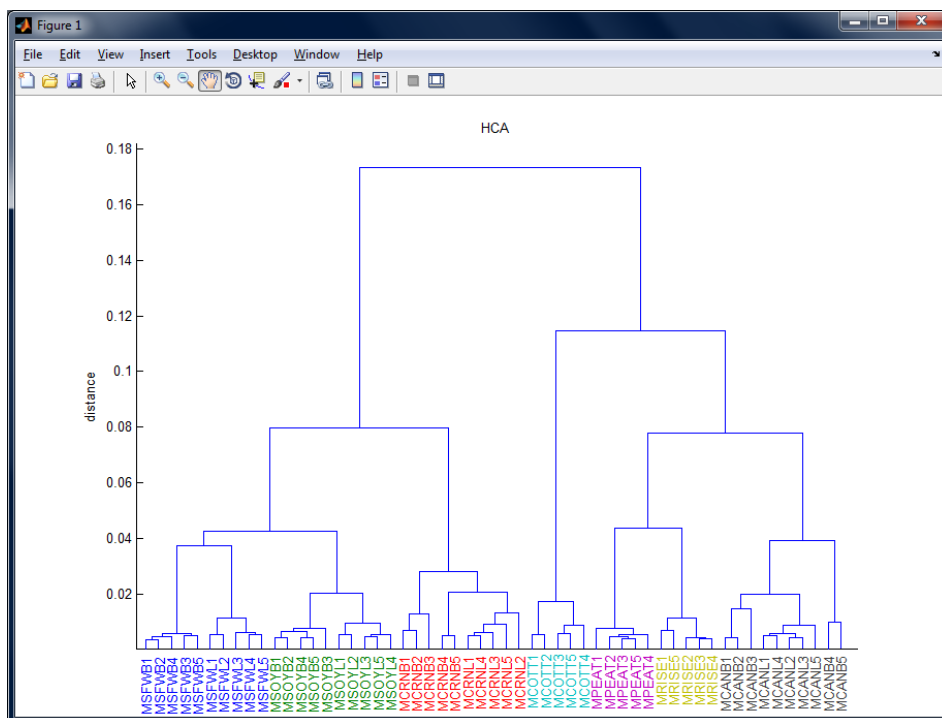


Figura 72. Dendrograma HCA – ligação completa – das amostras de óleos vegetais (FT- MIR), na região entre 1066-1169  $\text{cm}^{-1}$  - ChemoStat.

Fonte: Autor, extraído do software Matlab®, 2014.

No dendrograma da Figura 71 (ChemoStat) verifica-se a presença de dois grandes agrupamentos, um maior, que corresponde às 30 amostras mais poli-

insaturadas, ou seja, as amostras de soja (verde), milho (vermelho) e girassol (azul) e outro menor formado pelas amostras de algodão (azul-claro), amendoim (lilás), canola (preto) e arroz (verde-claro), todas separadas pela primeira componente principal.

Tendências semelhantes ocorreram na validação do método HCA com ligação pela média, como mostram as Figuras 73 e 74, empregando ChemoStat e Matlab®, respectivamente.

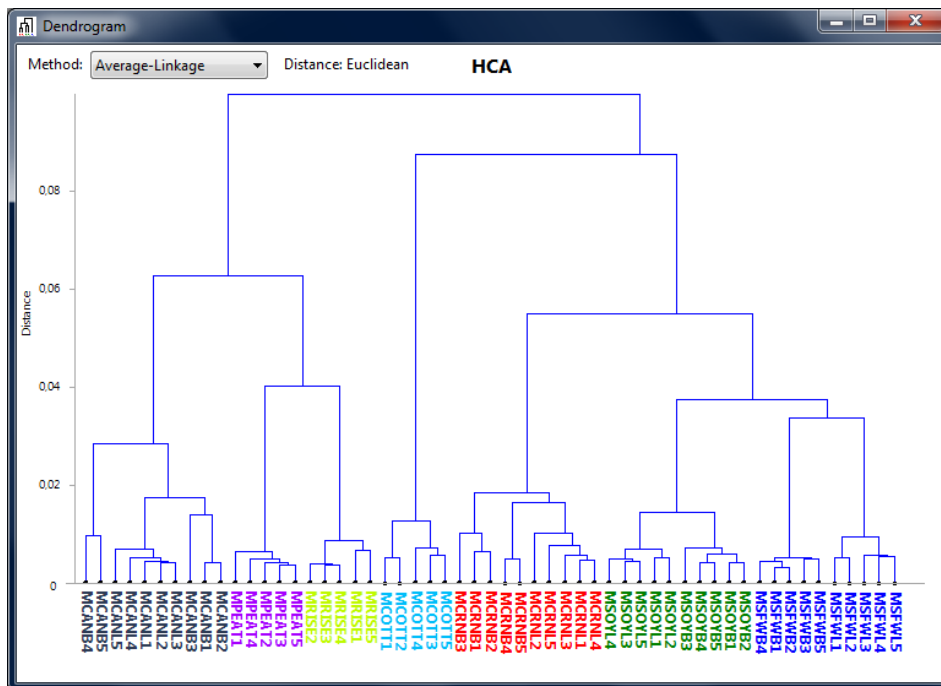


Figura 73. Dendrograma HCA – ligação pela média – das amostras de óleos vegetais (FT-MIR), na região entre 1066-1169  $\text{cm}^{-1}$  - ChemoStat.

Fonte: Autor, extraído do software ChemoStat, 2014.

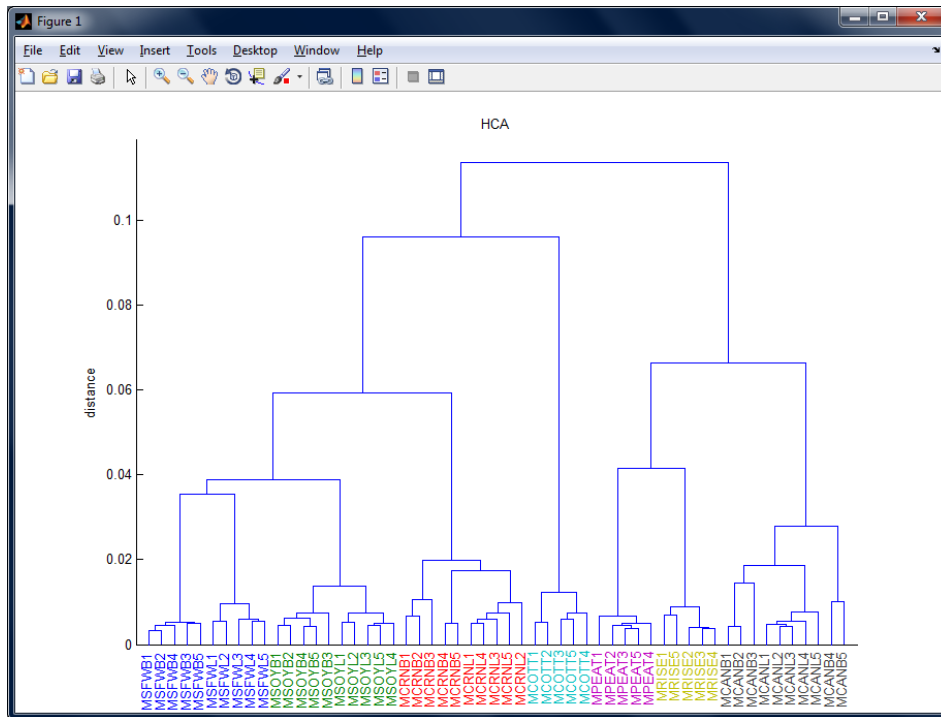


Figura 74. Dendrograma HCA – ligação pela média – das amostras de óleos vegetais (FT-MIR), na região entre 1066-1169  $\text{cm}^{-1}$  - Matlab®.  
 Fonte: Autor, extraído do software Matlab®, 2014.

Por fim, foi realizada a HCA pelo método de ligação simples (*single linkage*), onde observou-se que a amostra de algodão, que possui maior índice de ácidos graxos saturados, formou um grupo separado (azul-claro), antes da junção com dois grandes grupos, um deles formados pelos óleos vegetais que possuem maior índice de poliinsaturados, soja (verde), milho (vermelho) e girassol (azul), como mostra a Figura 75, no software ChemoStat, e o comparativo, Figura 76, no Matlab®.

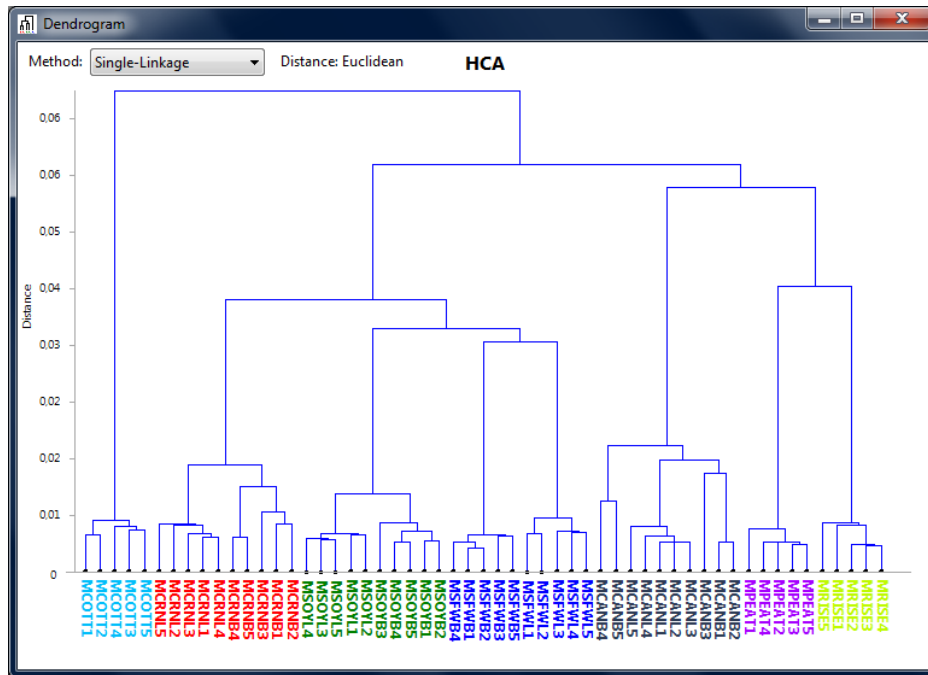


Figura 75. Dendrograma HCA – ligação simples – das amostras de óleos vegetais (FT-MIR), na região entre 1066-1169  $\text{cm}^{-1}$  - Matlab®.

Fonte: Autor, extraído do *software* ChemoStat, 2014.

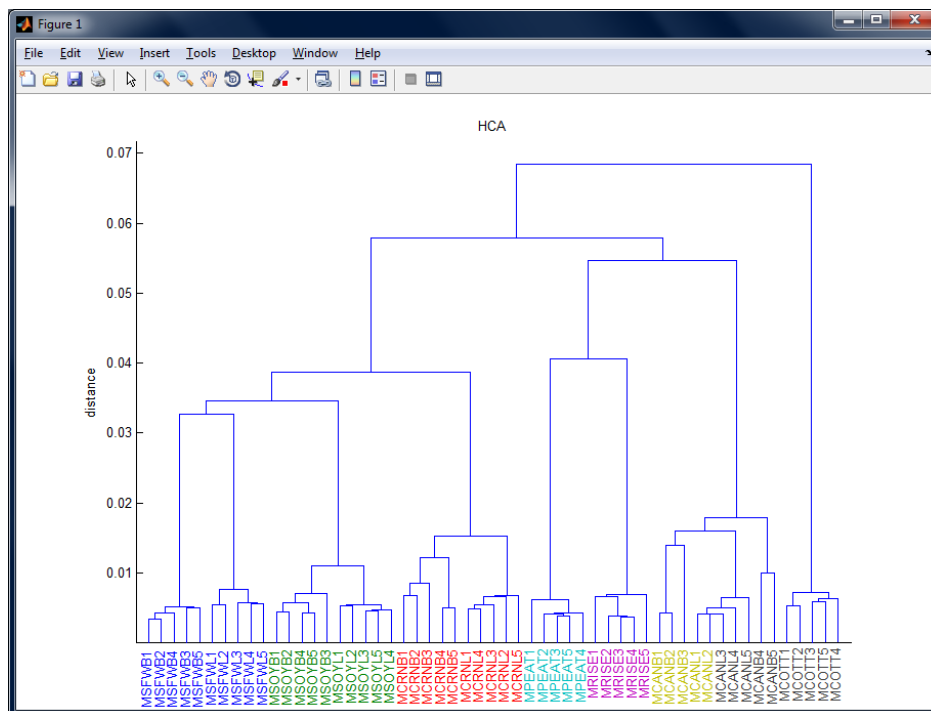


Figura 76. Dendrograma HCA – ligação simples – das amostras de óleos vegetais (FT-MIR), na região entre 1066-1169  $\text{cm}^{-1}$  - Matlab®.

Fonte: Autor, extraído do *software* Matlab®, 2014.

Os resultados obtidos na análise por agrupamento hierárquico (HCA) no *software* ChemoStat estão em completa concordância com os resultados obtidos no Matlab®.



### 5.3 Importação de dados de imagens

Os dados podem ser importados via área de transferência (comumente conhecido pelo atalho das teclas “*Control+V*”) através do Excel, na disposição dos dados onde as colunas representam as amostras e as linhas representam as variáveis, ou via botão “*Add image file*” para seleção de arquivos.

Ao pressionar o botão “*Add image file*” será apresentada uma janela para seleção dos arquivos (padrão *Windows*<sup>®</sup>) com as opções de filtro de extensão “*bmp*”, “*gif*”, “*jpg*” e “*png*”. Após selecioná-los basta clicar em “*Open*” ou “*Abrir*” para que os mesmos sejam adicionados para a seção 1.

Semelhante ao modo de espectroscopia, os arquivos podem ser deletados individualmente ou em sua totalidade, função exercida pelos botões “*Delete*” e “*Delete all*”, respectivamente. O teclado do computador também poderá ser utilizado para exclusão desses arquivos, desde que sejam selecionados pelo “*mouse*”.

Antes da leitura dos arquivos, o *software* permite a escolha de dois tipos de segregação de pixels, conforme exhibe a Figura 77.

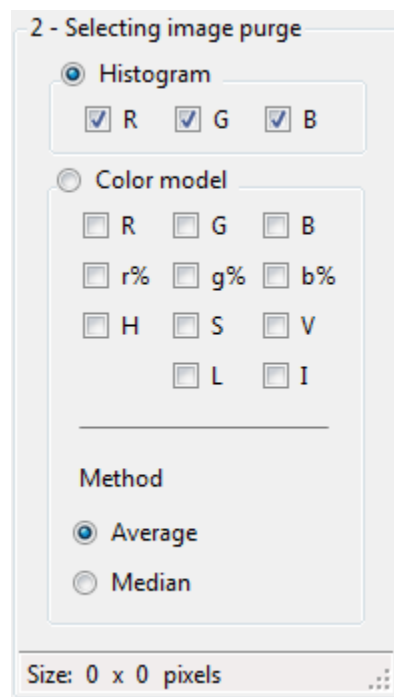


Figura 77. Detalhe da seção 2 na tela principal padrão espectroscopia.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

A primeira opção, chamada “*Histogram*”, realiza uma distribuição de frequência baseada em 256 tons, chamada histograma, para cada componente de cor RGB. Basta selecionar individualmente R, G ou B para que sejam extraídos seus respectivos histogramas em sequência num mesmo vetor.

A segunda opção, chamada “*Color model*” permite extrair as informações dos modelos de cor RGB, RGB relativo, definidos como “r%”, “g%” e “b%”, HSV, incluindo ainda as informações de intensidade (“I”) e luminância (“L”), quando selecionados na caixa de marcação. É possível optar por um ou mais componentes, separadamente. Os dados extraídos pixel a pixel podem ser agrupados numa média, caso a opção “*Average*” tenha selecionada, ou numa mediana, referente à opção “*Median*”.

O botão “*Import data*” realiza a leitura dos arquivos e extração das informações referentes às imagens. Vale ressaltar que o nome do arquivo é utilizado como nome da amostra, excluindo-se a extensão. A Figura 78 ilustra a grade dos dados importados e o nome do arquivo no cabeçalho da mesma na seção 3.

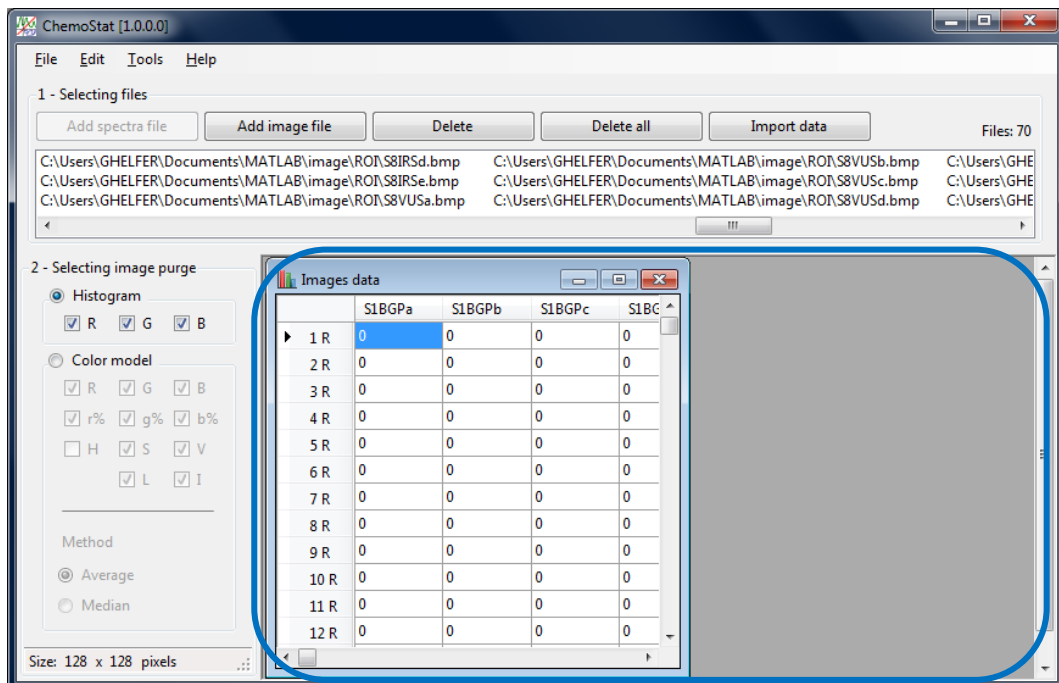


Figura 78. Tela principal padrão imagem com grade de dados – seção 3.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

Caso haja a interesse de utilizar uma área específica dentro da imagem, ao invés de sua totalidade, foi desenvolvido um editor de imagens. Para acessá-lo

basta acessar a barra de menu “*Tools >> Image cutter*”. O Anexo C descreve suas funcionalidades.

### 5.3.1 Menu de ferramentas

Na grade ou matriz de dados, ao clicar com o botão direito do “*mouse*”, será apresentado um menu com as opções de histograma, exportação para Excel e texto, identificação de amostras por classe, além das técnicas HCA e PCA, conforme detalha a Figura 79, a seguir discutidas.

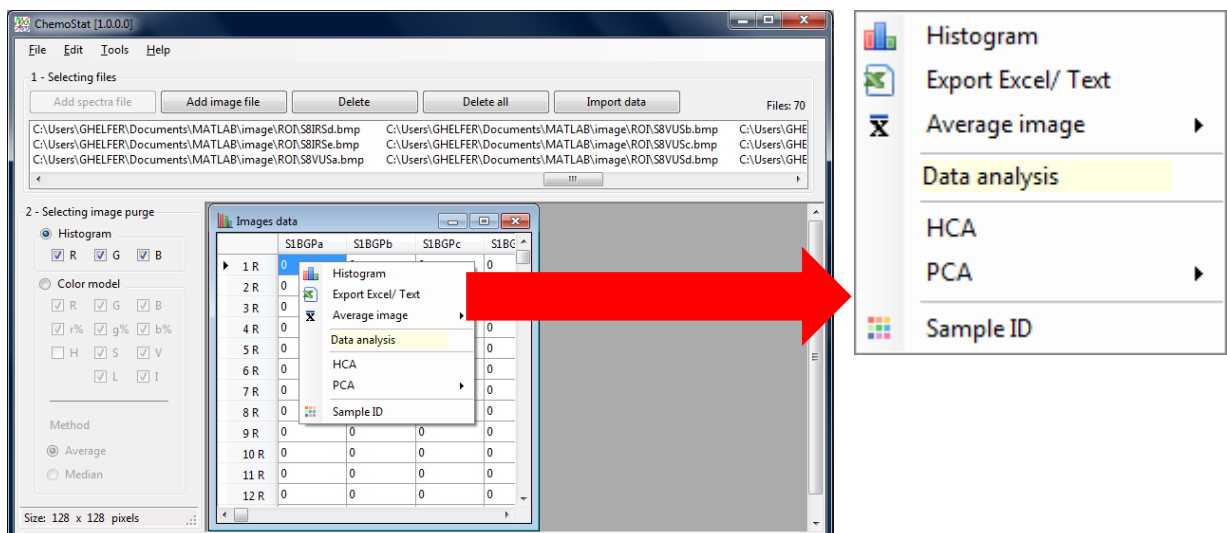


Figura 79. Tela principal padrão imagem com grade de dados – detalhe do menu de operações acionado.

Fonte: Autor, extraído do *software* ChemoStat, 2014.

### 5.3.2 Função “Histogram”

A opção de Histograma, quando aplicada na grade de dados, exibe uma nova tela apresentando a imagem, cujo nome encontra-se na barra da janela, e gráficos de histogramas dessa imagem, referentes as componentes R (vermelho), G (verde), B (azul), S (saturação) e L (luminância), conforme mostra a Figura 80.

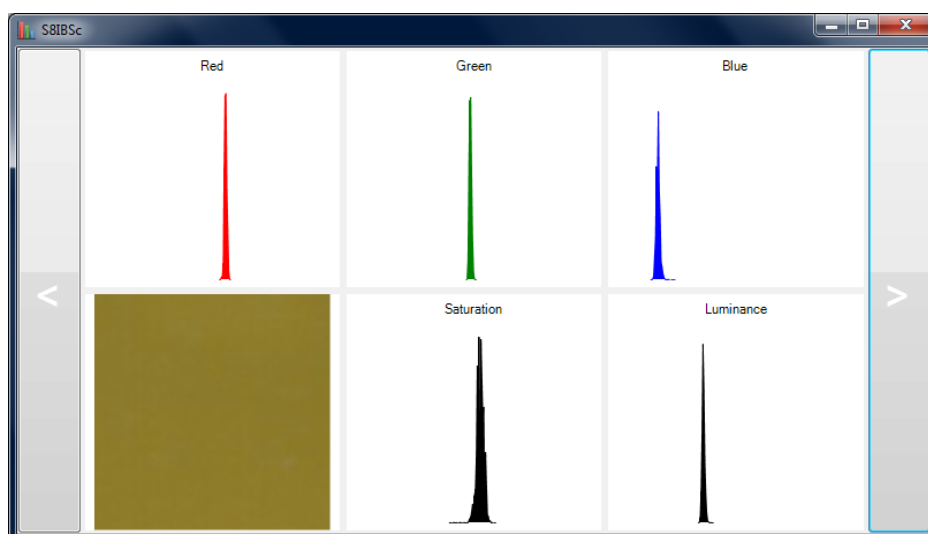


Figura 80. Janela com gráficos de histograma e imagem de uma amostra de óleo diesel tipo S1800 escaneada.

Fonte: Autor, extraído do *software* ChemoStat, 2014.

Nas laterais da janela existem dois botões de navegação entre as imagens, para avançar para a próxima imagem basta clicar no botão da direita “>”, enquanto que, para retroceder, o botão da esquerda, “<”.

No Anexo D encontram-se as todas as imagens escaneadas da primeira replicata de óleo diesel comercial e seus respectivos histogramas.

### 5.3.3 Função “Export Excel/ Text”

O item “*Export Excel/ Text*”, Figura 81, quando acionado, solicita ao usuário nome de arquivo e um diretório onde será gerado o mesmo no formato Excel ou Texto, extensões “.xls” ou “.txt”, de acordo com o filtro de extensão selecionado, que agregará todas as informações contidas na grade de dados.

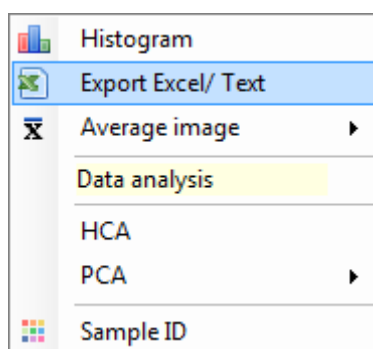


Figura 81. Menu principal de operações - funções de exportação de dados.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

### 5.3.4 Função “imagem média”

O item “*Average image*” realiza a média das replicatas das amostras. Os itens “*2nd-Duplicate*”, “*3rd-Triplicate*”, “*4th-Quadruplicate*” e “*5th-Quintuplicate*” realizam as médias de duas, três, quatro e cinco amostras, quando dispostas lado a lado na grade de dados. Caso o número de replicatas seja maior que cinco, o item “*By value*” permite informar a quantidade de replicatas a ser executada a média, através de uma janela de diálogo imposta ao usuário. Já a opção “*By class*” possibilita realizar a média das replicatas de acordo com a identificação das amostras por classes, função previamente definida pelo usuário e anteriormente discutida. O item “*All collection*” realiza a média de todas as amostras contidas na grade de dados. A Figura 82 exibe todas as opções disponíveis.

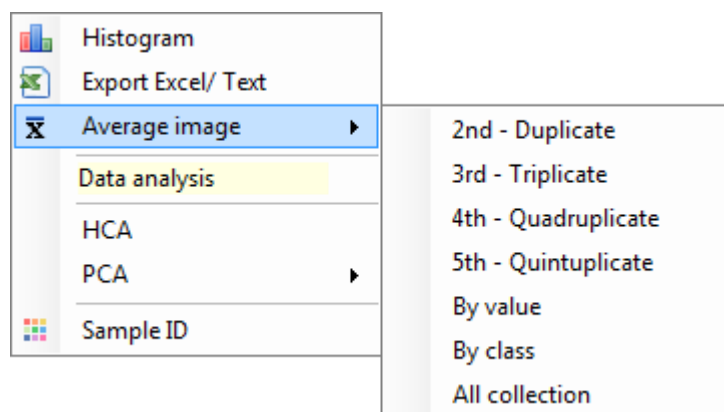


Figura 82. Menu principal de operações - funções para imagem média.  
Fonte: Autor, extraído do *software* ChemoStat, 2014.

### 5.3.5 Identificação de amostras

A função de identificação de amostras é semelhante à utilizada no modo espectroscopia, descrita anteriormente no item 5.2.11.

### 5.3.6 Algoritmo PCA

O *software* ChemoStat permite a análise de componentes principais (PCA) também para as imagens, apresentando as mesmas janelas de gráficos de *scores*, *loadings*, *outliers* e demais opções disponíveis para os espectros

Para validá-la, foram analisados um total de 10 amostras comerciais de diesel coletados em Santa Cruz do Sul e em Porto Alegre dos tipos S10, S500 e S1800, além de 3 amostras, uma de cada tipo, adquiridas junto à uma refinaria no Estado do Rio Grande do Sul, denominadas como amostras-padrão.

Foram empregadas duas técnicas diferentes de extração de dados das imagens, via histograma R, G e B (média das replicatas) e a partir das médias dos pixels dos modelos de cores R, G, B, R relativo (r%), G relativo (g%), B relativo (b%), S, V, L e I, de cada amostra separadamente.

### 5.3.6.1 Histogramas

A PCA executada a partir do histograma R, G e B (média das replicatas) nos *softwares* ChemoStat e Matlab<sup>®</sup> demonstraram que com duas componentes principais é possível descrever 50,52% da variância dos dados, conforme apresentado na Tabela 5.

Tabela 5. Valores para variância e variância acumulada das seis primeiras componentes principais dos dados das imagens de óleo diesel (histograma) - ChemoStat e Matlab<sup>®</sup>.

PC	ChemoStat	ChemoStat	Matlab <sup>®</sup>	Matlab <sup>®</sup>
	var (%)	var acum (%)	var (%)	var acum (%)
1	36,55	36,55	36,55	36,55
2	13,97	50,52	13,97	50,52
3	10,77	61,29	10,77	61,29
4	9,65	70,94	9,65	70,94
5	6,53	77,47	6,53	77,47
6	5,29	82,76	5,29	82,76

Fonte: Autor, 2014.

Analisando os gráficos dos *scores* da PC1 x PC2 dos histogramas centrados na média, como mostram as Figuras 83 e 84, empregando ChemoStat e Matlab<sup>®</sup>, respectivamente, é possível observar a separação das amostras S10, S500 e S1800 e o agrupamento das amostras comerciais em seus respectivos tipos, com base nas amostras-padrão.

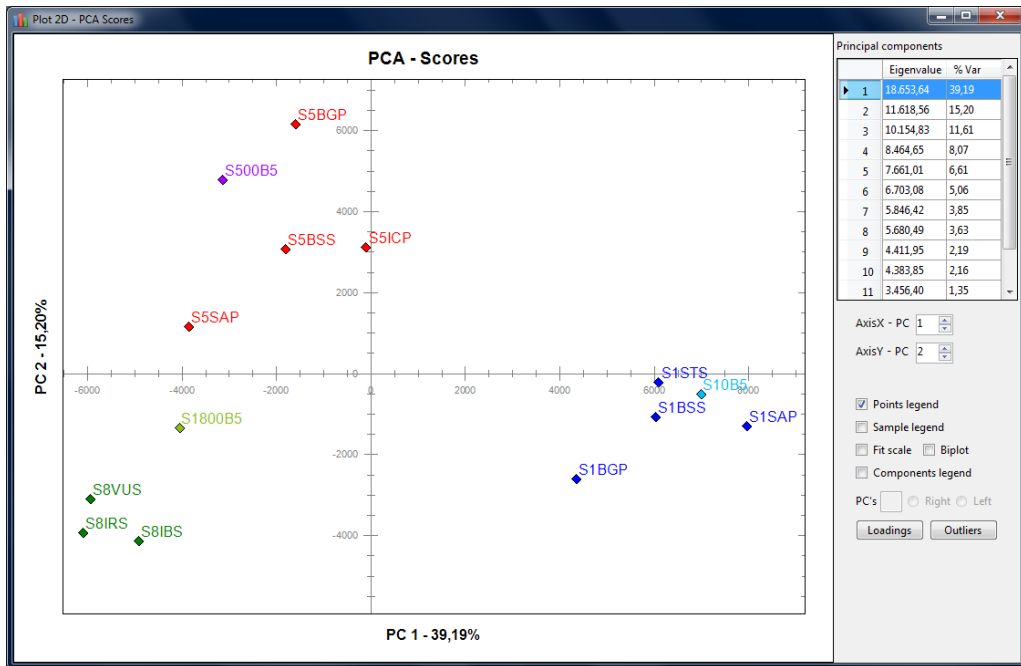


Figura 83. Gráfico de scores PC1 x PC2 de histogramas R, G e B das médias das replicatas de óleos diesel escaneados – ChemoStat.  
 Fonte: Autor, extraído do software ChemoStat, 2014.

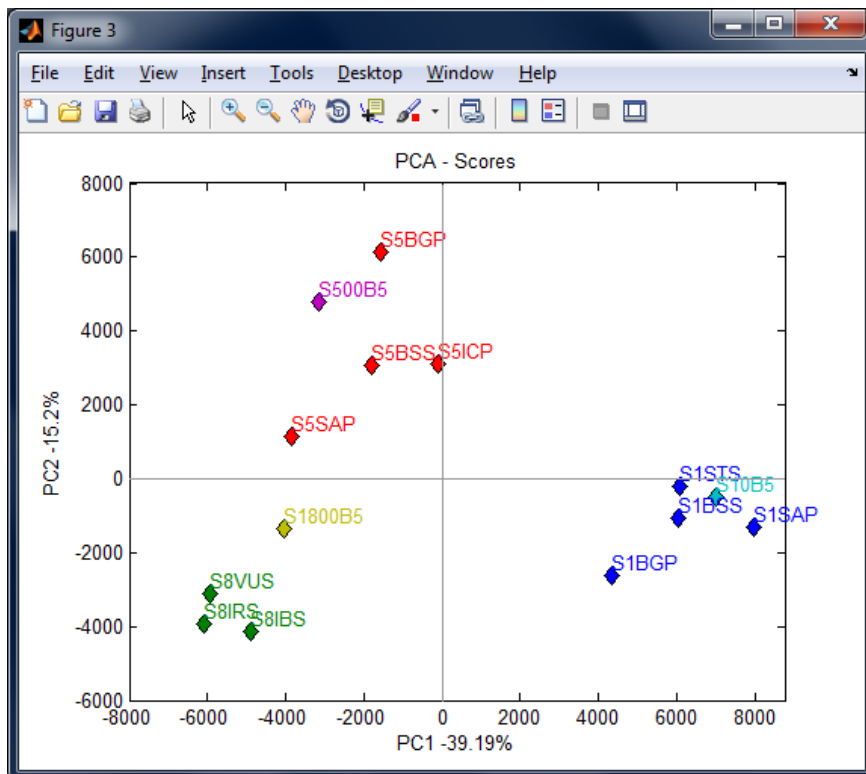


Figura 84. Gráfico de scores PC1 x PC2 de histogramas R, G e B das médias das replicatas de óleos diesel escaneados – Matlab®.  
 Fonte: Autor, extraído do software Matlab®, 2014.

Já os gráficos dos *loadings*, Figuras 85 e 86, empregando ChemoStat e Matlab®, respectivamente, demonstram os pesos exercidos pelos histogramas R, G e B, para cada variável, na PC1, que corresponde a 39,19% dos dados.

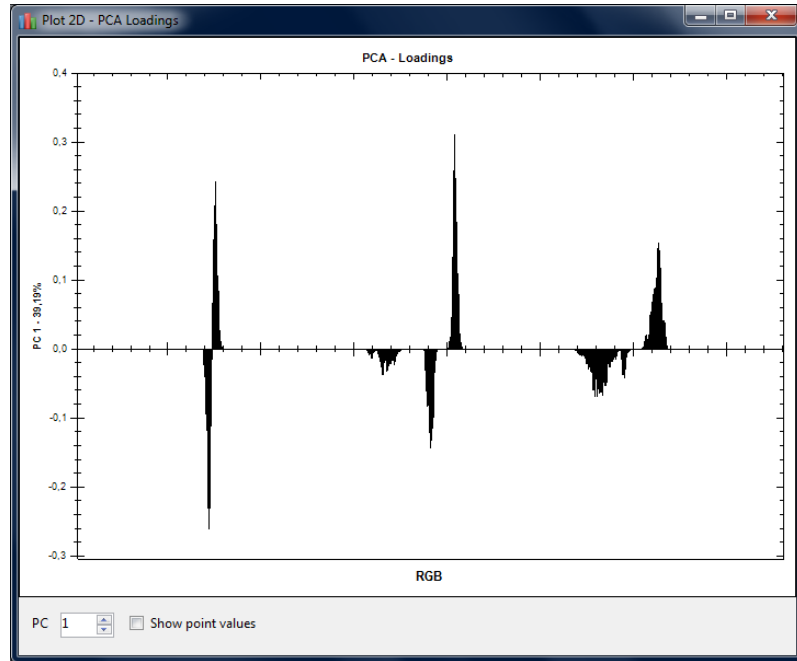


Figura 85. Gráfico de *loadings* da PC1 de histogramas R, G e B das médias das replicatas de óleos diesel escaneados – ChemoStat.

Fonte: Autor, extraído do *software* ChemoStat, 2014.

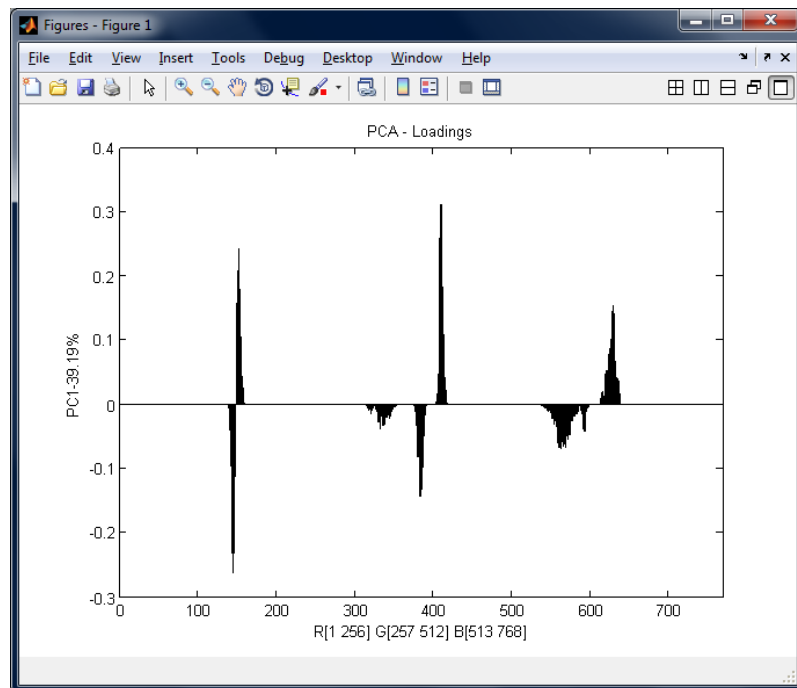


Figura 86. Gráfico de *loadings* da PC1 de histogramas R, G e B das médias das replicatas de óleos diesel escaneados – Matlab®.

Fonte: Autor, extraído do *software* Matlab®, 2014.



Os resultados obtidos na análise de componentes principais (PCA) no *software* ChemoStat estão em completa concordância com os resultados obtidos no Matlab®.

### 5.3.6.2 Modelos de scores

A PCA executada a partir das médias dos pixels dos modelos de cores R, G, B, R relativo (r%), G relativo (g%), B relativo (b%), S, V, L e I, de cada amostra separadamente, nos *softwares* ChemoStat e Matlab® demonstraram que com duas componentes principais é possível descrever 94,19% da variância dos dados, conforme apresentado na Tabela 6.

Tabela 6. Valores para variância e variância acumulada das seis primeiras componentes principais dos dados das imagens de óleo diesel (modelos de cores) - ChemoStat e Matlab®.

PC	ChemoStat		Matlab®	
	var (%)	var acum (%)	var (%)	var acum (%)
1	74,25	74,25	74,25	74,25
2	19,94	94,19	19,94	94,19
3	5,54	99,73	5,54	99,73
4	0,18	99,91	0,18	99,91
5	0,06	99,97	0,06	99,97
6	0,03	100,00	0,03	100,00

Fonte: Autor, 2014.

Assim como no histograma, analisando o gráfico dos *scores* da PC1 x PC2 das médias das variáveis R, G, B, r%, g%, b%, S, V, L, I autoescalados, como ilustram as Figuras 87 e 88, empregando ChemoStat e Matlab®, respectivamente, é possível também observar a separação das amostras S10, S500 e S1800 e o agrupamento das amostras comerciais em seus respectivos tipos, com base nas amostras-padrão.

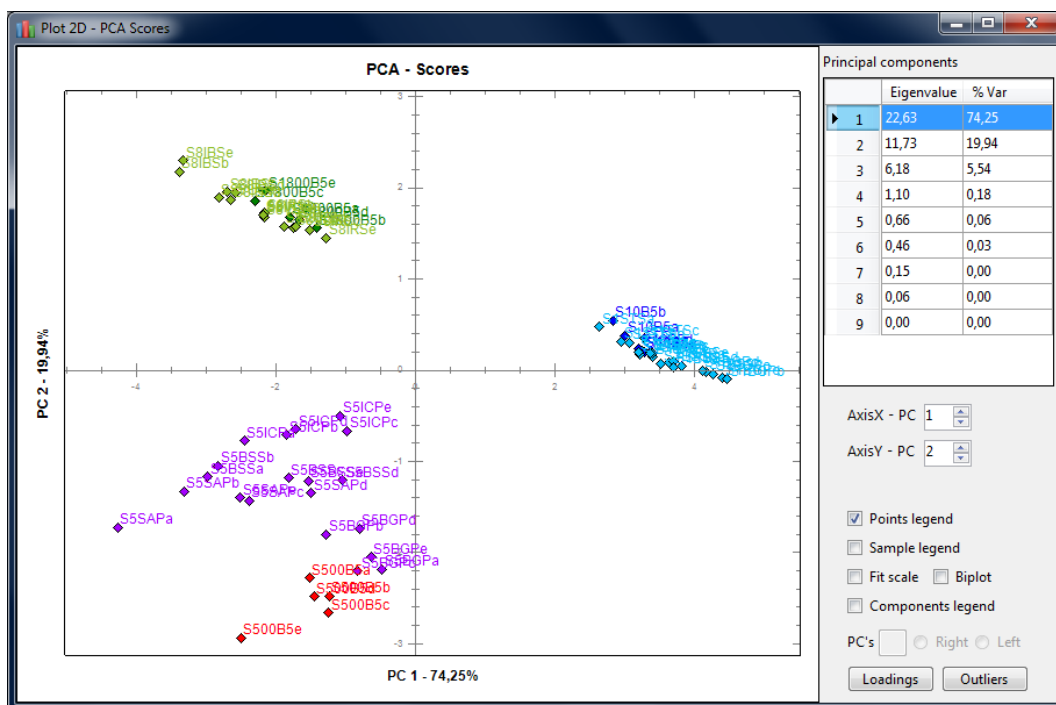


Figura 87. Gráfico de scores PC1 x PC2 dos modelos de cores R, G, B, r%, g%, b%, S, V, L, I de óleos diesel escaneados – ChemoStat.  
Fonte: Autor, extraído do software ChemoStat, 2014.

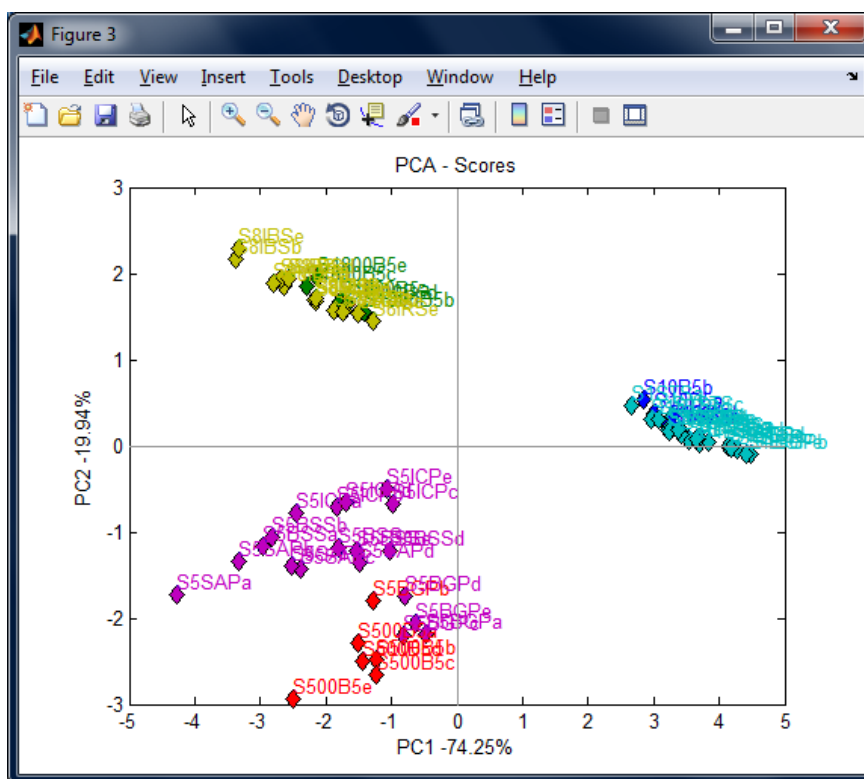


Figura 88. Gráfico de scores PC1 x PC2 dos modelos de cores R, G, B, r%, g%, b%, S, V, L, I de óleos diesel escaneados – Matlab®.  
Fonte: Autor, extraído do software Matlab®, 2014.

Já os gráficos dos *loadings*, exibidos pelas Figuras 89 e 90, empregando ChemoStat e Matlab®, respectivamente, demonstram os pesos exercidos por cada variável nos modelos de cor, que corresponde a 74,25% na PC1.

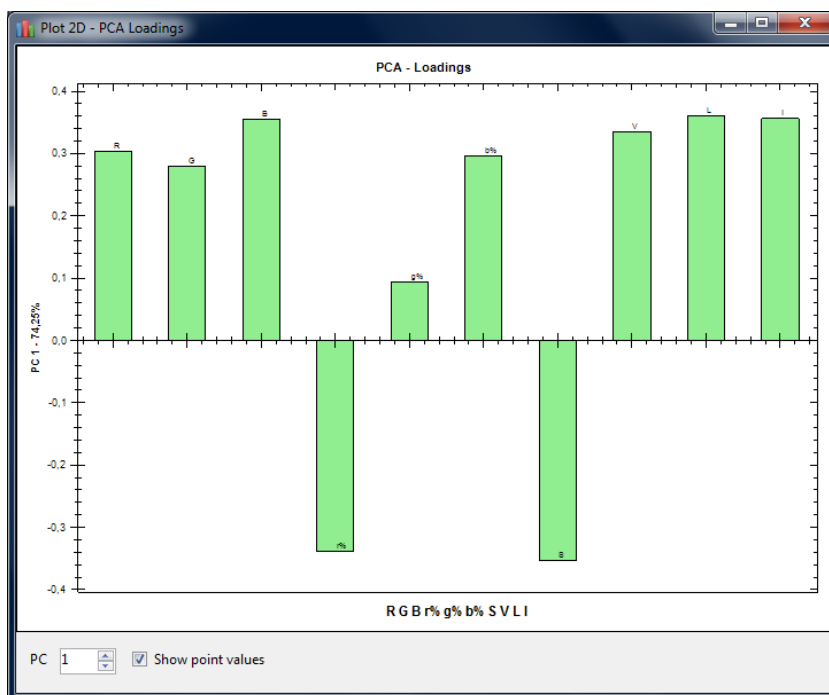


Figura 89. Gráfico de *loadings* da PC1 dos modelos de cores R, G, B, r%, g%, b%, S, V, L, I de óleos diesel escaneados – ChemoStat.

Fonte: Autor, extraído do software ChemoStat, 2014.

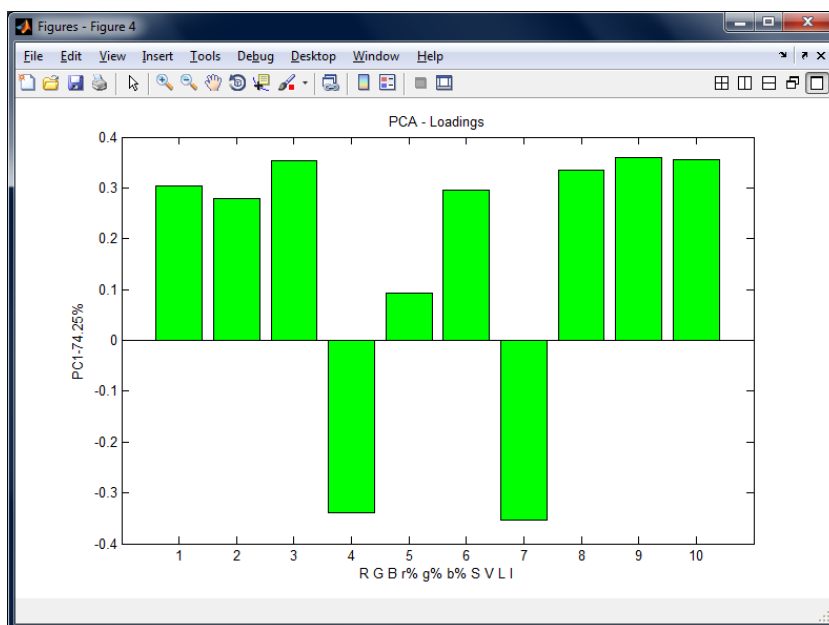


Figura 90. Gráfico de *loadings* da PC1 dos modelos de cores R, G, B, r%, g%, b%, S, V, L, I de óleos diesel escaneados – Matlab®.

Fonte: Autor, extraído do software Matlab®, 2014.



Conforme se observa, apesar de não se apresentarem na mesma ordem, não há ocorrência de amostras anômalas pois o  $T^2$  de cada uma das imagens ficou abaixo do limite superior de controle (“UCL”).

Os resultados obtidos na análise de componentes principais (PCA) no *software* ChemoStat estão em completa concordância com os resultados obtidos no Matlab®.

### 5.3.7 Algoritmo HCA

A opção “HCA” quando aplicada na grade de dados, exibe uma nova janela apresentando o resultado da Análise por Agrupamento Hierárquico, apresentando as mesmas janelas de gráficos e opções disponíveis para os espectros.

Para realizar a validação da HCA com ligação completa (*complete linkage*), utilizaram-se as mesmas amostras analisadas pela PCA (Figura 83), utilizando o histograma RGB das médias das imagens de óleo diesel, como mostram as Figuras 93 e 94, empregando ChemoStat e Matlab®, respectivamente.

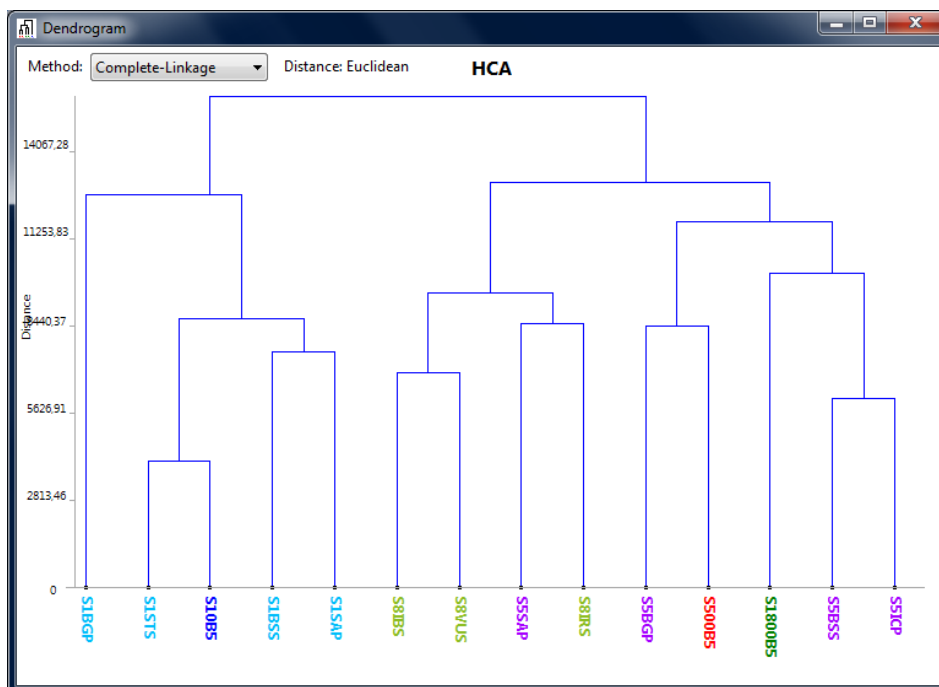


Figura 93. Dendrograma HCA – ligação completa – das amostras de óleos diesel escaneadas – Chemostat.

Fonte: Autor, extraído do *software* ChemoStat, 2014.

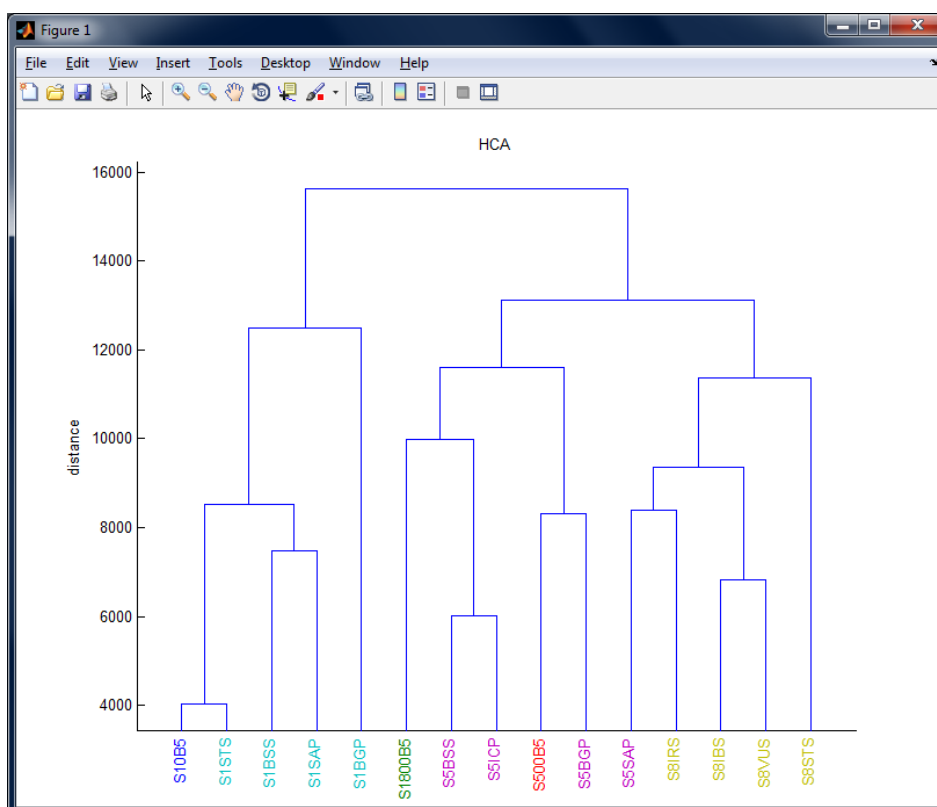


Figura 94. Dendrograma HCA - ligação completadas - amostras de óleos diesel escaneadas – Matlab®.

Fonte: Autor, extraído do *software* Matlab®, 2014.

Observa-se que as tendências de separação foram confirmadas através do dendrograma obtido pela HCA (Figura 93). Neste dendrograma verifica-se a presença de dois agrupamentos, um maior, que corresponde as amostras de diesel S500 e S1800, que possuem cor mais escura, variando de vermelho à tons de laranja, respectivamente, e um grupo menor formado pelas amostras de diesel S10, tons amarelados.

Os resultados obtidos na análise de componentes principais (PCA) no *software* ChemoStat estão em completa concordância com os resultados obtidos no Matlab®.

#### 5.4 Solução *online*

A solução *online*, chamada de ChemoStat *web version*, ou versão web, foi desenvolvida a partir dos mesmos algoritmos utilizados na versão *desktop*. A interface teve que ser adaptada para a entrada de dados, apresentando, desta maneira, apenas alguns dos principais recursos como normalização entre os limites

de zero e um, SNV, MSC, Média móvel, primeira e segunda derivada, Savitzky-Golay, identificação de amostras por classe, PCA e HCA.

A seguir serão discutidas as interfaces da versão *online*.

#### 5.4.1 Tela de acesso e registro

O acesso ao site se dá pelos endereços “http://www.chemostat.com.br” ou “http://www.chemostat.net”. Aparecerá a tela de entrada com opções de registro de usuário (link “*here*”), campos para *login* (e-mail de entrada previamente registrado), “*password*” (senha de entrada previamente registrada), “*can't access your account*” (recuperação de senha) e botão de confirmação, indicado em vermelho na Figura 95, para entrada no sistema.

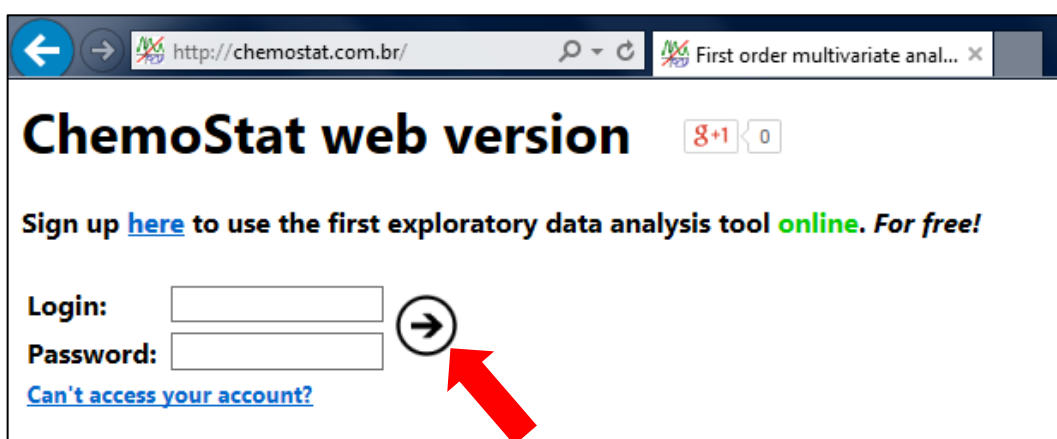


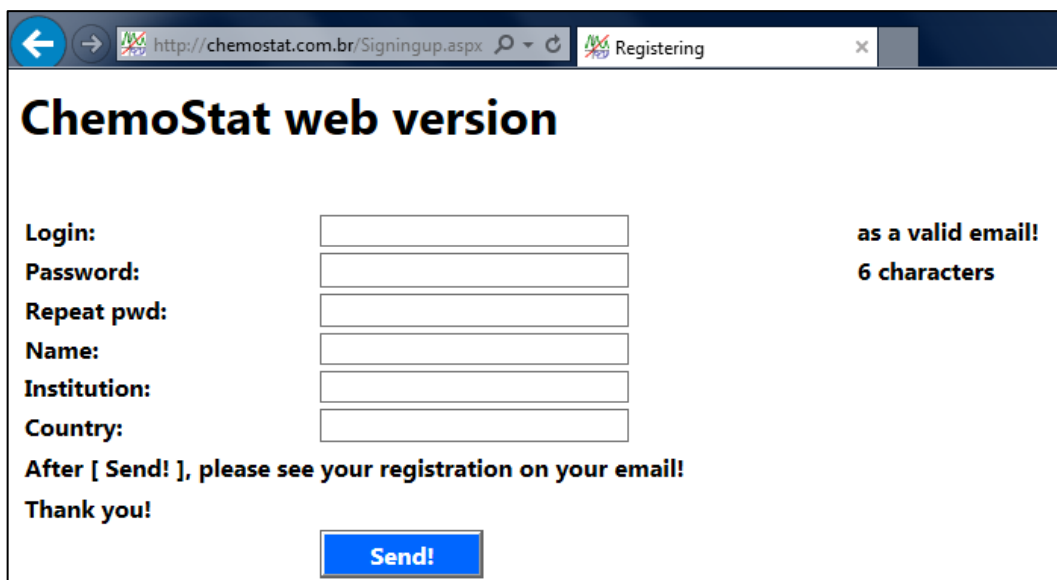
Figura 95. Tela de entrada - ChemoStat versão web.  
Fonte: Autor, extraído do ChemoStat versão web, 2014.

#### 5.4.2 Registro de usuários novos

Para utilização do sistema, o usuário deverá ser registrado. Para tanto na tela principal deverá ser acionado o link “*here*” da frase “*Sign up here to use the first exploratory data analysis. For free!*”, que significa, “Registre-se aqui para utilizar a primeira ferramenta de análise exploratória *online*. Gratuitamente”.

Acessada a tela, aparecerão campos obrigatórios para serem preenchidos como “*login*” (será utilizado o e-mail), “*password*” e “*repeat pwd*” (senha de seis caracteres), “*Name*” (nome do usuário), “*Institution*” (nome da instituição de estudo/pesquisa), “*Country*” (país de origem). Ao clicar no botão “*Send*”, o registro

será enviado ao servidor e ao e-mail do usuário avisando que está apto para utilização do sistema.



**ChemoStat web version**

**Login:**  **as a valid email!**

**Password:**  **6 characters**

**Repeat pwd:**

**Name:**

**Institution:**

**Country:**

**After [ Send! ], please see your registration on your email!**

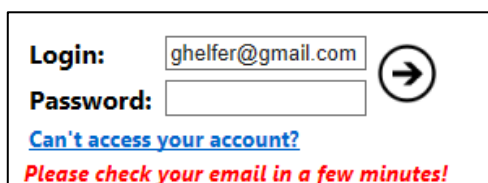
**Thank you!**


**Send!**

Figura 96. Tela de registro de usuários - ChemoStat versão web  
Fonte: Autor, extraído do ChemoStat versão web, 2014.

#### 5.4.3 Perda da senha de acesso

O sistema permite resgatar a senha de acesso via email, para tanto, na tela principal o usuário deverá preencher seu “login” (e-mail) e clicar no link “Can’t access your account?”. Se a informação de “login” (e-mail) estiver correta, aparecerá a mensagem “Please, check your email in a few minutes!” (“Por favor, verifique sua conta de e-mail em alguns minutos”) e o usuário receberá a senha pelo e-mail cadastrado, conforme ilustra a Figura 97.



**Login:**  

**Password:**

[Can't access your account?](#)

**Please check your email in a few minutes!**

Figura 97. Detalhe da mensagem de recuperação de senha - ChemoStat versão web.  
Fonte: Autor, extraído do ChemoStat versão web, 2014.



#### 5.4.4 Tela principal para análise exploratória

Tão logo sejam preenchidos o usuário e senha e acessado o sistema via botão de confirmação, aparecerá a tela principal do sistema com alguns comandos básicos como importação de dados via Excel/ texto, tratamento de dados, identificação das amostras, técnicas PCA e HCA, de acordo com Figura 98.

**ChemoStat web version**

Welcome *admin* ,

**Copy data from excel ( [example](#) )**

**Paste** First column and first row will be considered as headers.

Value	Sample
1	0.123
2	0.789

**Plot 2D**

**Choose corrections, transformations, preprocessing methods:**

Normalize 1-0

SNV

MSC

Moving Average Points  Ex.: 5 (odd values)

1st Derivative Points  Ex.: 7 (odd values)

2nd Derivative Points  Ex.: 9 (odd values)

Savitky-golay Deriv  Polynm  Points  Ex.: 1-1-17, 1-2-21

Autoscale

Meancenter

**Sample ID:**

Classify by number of sample characters  0=Auto

**Principal Component Analysis**

Scores:  x  PC's

**Autoscale** **Meancenter** **None**

Loadings:  PC

**Autoscale** **Meancenter** **None**

**Hierarchical Clustering Analysis**

Euclidean distance

**Single Linkage** **Complete Link.** **Average Link.**

Figura 98. Tela principal para análise exploratória de dados - ChemoStat versão web. Fonte: Autor, extraído do ChemoStat versão web, 2014.

O sistema permite a entrada de dados somente via Excel ou texto utilizando o recurso da área de transferência. Para tanto, os dados deverão ser selecionados, “copiados” - do Excel ou algum editor de texto (neste caso, desde que números sejam separados por tabulação) - e “colados” no site através do botão “*Paste*”, conforme ilustra a Figura 88. Em alguns casos aparecerá uma mensagem solicitando a permissão do navegador em acessar a área de transferência (Figura 99). A opção “*Allow access*” ou “Permitir acesso” realiza a importação dos dados corretamente.

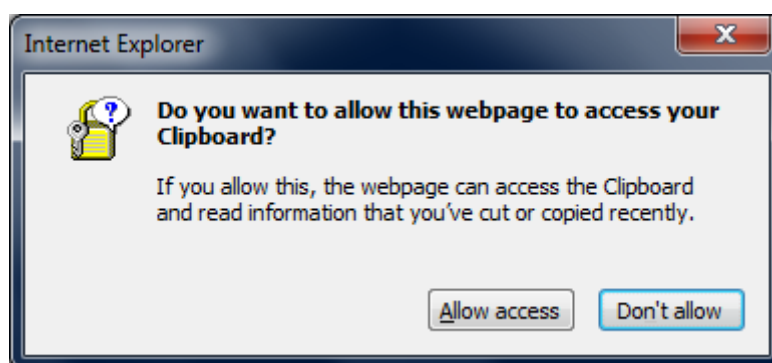


Figura 99. Caixa de diálogo para permissão de acesso à área de transferência.  
Fonte: Autor, extraído do ChemoStat versão web, 2014.

Uma vez carregados na tela, os dados podem ser plotados num gráfico. Para isso, basta acionar o botão “*Plot 2D*” e uma nova janela abrirá com os dados plotados num gráfico de linhas.

Além disso, casos os dados forem de espectroscopia, podem ser aplicados os tratamentos de sinais pela marcação nas caixas checagem. A coluna de marcação “*Run order*” significa a ordem em que serão aplicados os tratamentos, caso mais de um seja selecionado. Os tratamentos “*Moving average*”, “*1st derivative*” e “*2nd derivative*” necessitam que sejam informados a quantidade de pontos (“*Points*”), ou número de ondas, nos quais será aplicado o algoritmo. O sistema ainda permite a identificação a partir de cores pré-estabelecidas através da análise sintática do nome da amostra. Para que isto ocorra, deve ser marcada a caixa de checagem “*Classify by number of sample characters*” e informado o número de caracteres que são semelhantes entre as replicatas, sendo que o valor zero o sistema calcula de forma automática.

A técnica PCA para “Scores” é executada através dos botões “Autoscale”, cujos dados serão previamente autoescalados; “Meancenter”, centrados na média; e “None”, para nenhum pré-processamento prévio. Em todos os casos, quando pressionados, uma nova janela abrirá com o gráfico de scores conforme os campos “PC’s” preenchidos. A tela padrão foi configurada para PC1 x PC2. Já a técnica PCA para “Loadings” ocorre de forma semelhante à opção de “Scores”, no que diz respeito aos botões “Autoscale”, “Meancenter” e “None”. Em todos os casos, quando pressionados, uma nova janela abrirá com o gráfico de loadings de acordo com o campo “PC” preenchido. A tela padrão foi configurada para PC1.

A técnica HCA, calculada pela distância euclidiana, é executada através dos botões “Single-Linkage”, “Complete-Linkage” e “Average-Linkage”, denominado pelo método de ligação. Em todos os casos, quando pressionado o botão, uma nova janela abrirá com o respectivo dendrograma.

Na Figura 100 foi realizada a mesma PCA (scores) nos espectros (FT-MIR) de óleos vegetais com dados autoescalados, previamente normalizados de zero e um, aplicadas SNV e primeira derivada com janela de 5 pontos na região entre 1066 e 1169  $\text{cm}^{-1}$  (Figura 64).

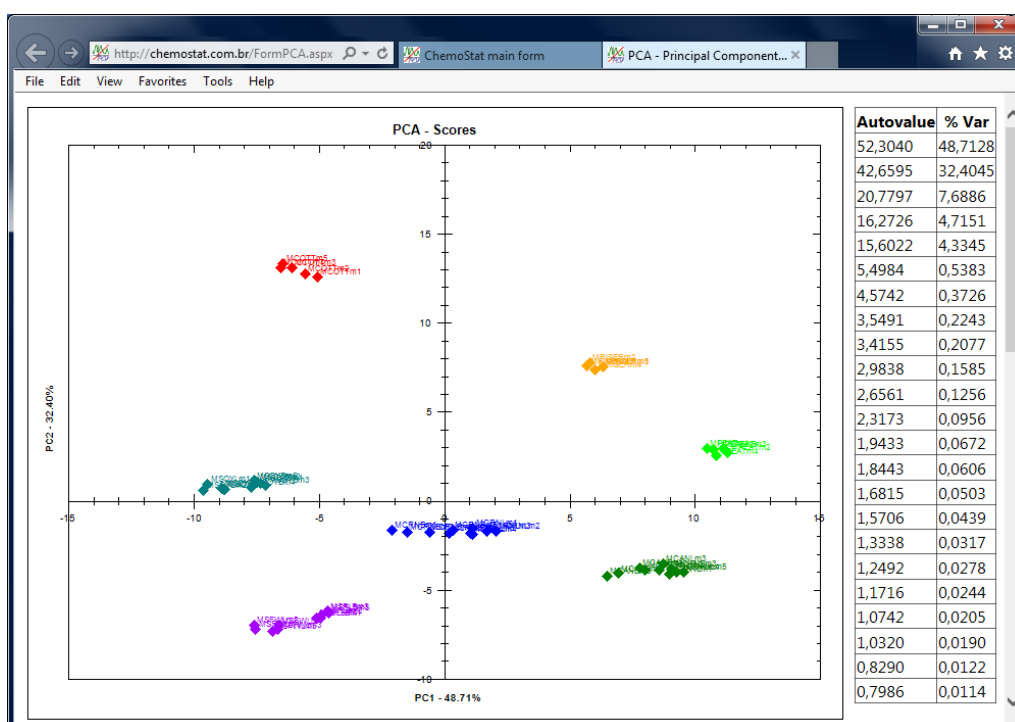


Figura 100. Gráfico de scores (PC1 x PC2) para espectros de óleos vegetais (FT-MIR), na região entre 1066 e 1169  $\text{cm}^{-1}$  - ChemoStat versão web.

Fonte: Autor, extraído do ChemoStat versão web, 2014.

Já na Figura 101 foi realizada a mesma PCA (*loadings*) nos espectros (FT-NIR) de óleos vegetais com dados centrados na média, previamente normalizados de zero e um, aplicadas SNV e primeira derivada com janela de 5 pontos na região entre 5500 e 6000  $\text{cm}^{-1}$  (Figura 47).

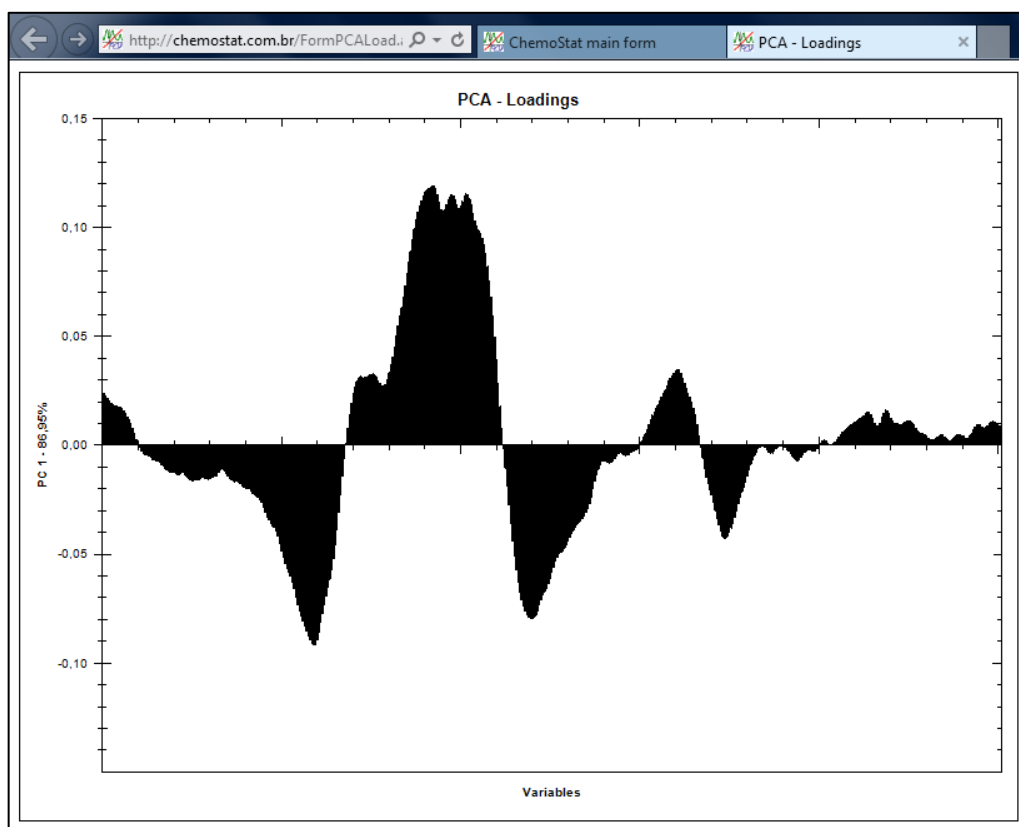


Figura 101. Gráfico de *loadings* (PC1) para espectros de óleos vegetais (FT-NIR), na região entre 5500 e 6000  $\text{cm}^{-1}$  - ChemoStat versão web.

Fonte: Autor, extraído do ChemoStat versão web, 2014.

Posteriormente foi realizada a HCA com método “*Complete-Linkage*” (Figura 102) sobre o mesmo conjunto de dados discutidos na Figura 71, ou seja, espectros (FT-MIR) de óleos vegetais com dados autoescalados, previamente normalizados de zero e um, aplicadas SNV e primeira derivada com janela de 5 pontos na região entre 1066 e 1169  $\text{cm}^{-1}$ .

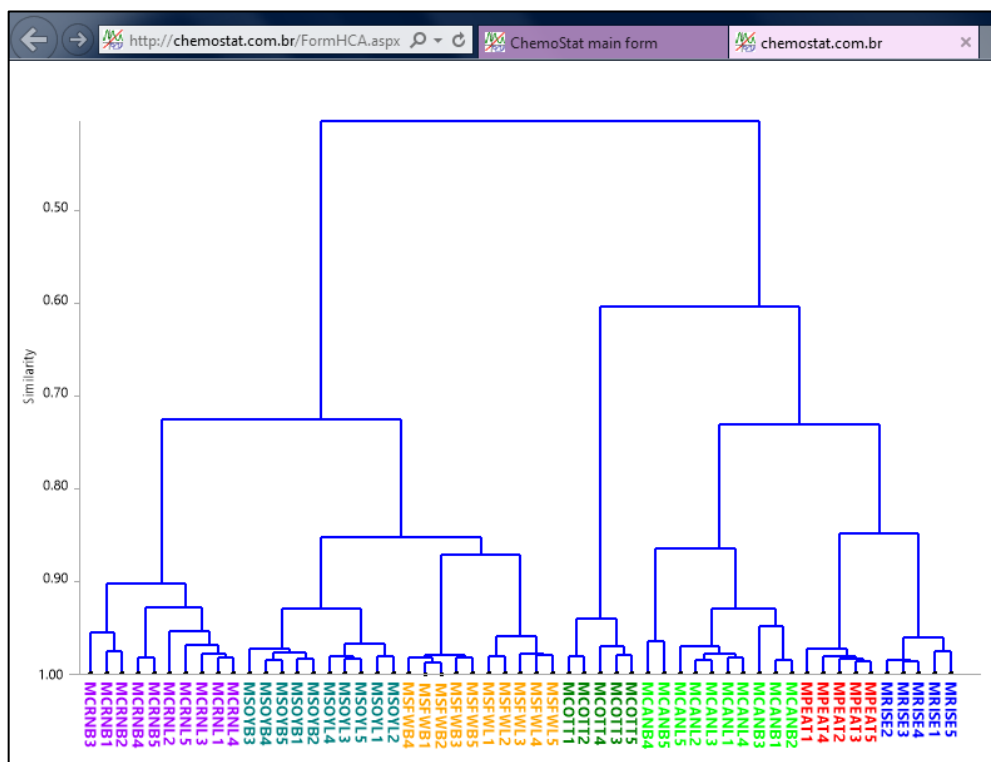


Figura 102. Dendrograma HCA – ligação completa – para espectros de óleos vegetais (FT-MIR), na região entre 1066 e 1169  $\text{cm}^{-1}$  - ChemoStat versão web.

Fonte: Autor, extraído do ChemoStat versão web, 2014.

Ambos os resultados da PCA (*scores* e *loadings*) e da HCA, estão em completa concordância com os resultados obtidos no *software* ChemoStat e, conseqüentemente, com o aplicativo Matlab®.

## 5.5 Dados de outras origens

O *software desktop* e a versão *online* permitem a entrada de dados de qualquer natureza, não se restringindo apenas a espectros e imagens. Para tanto, o *software* deve estar no modo de operação “Imagem” e os dados devem ter sua origem a partir do Excel ou texto, utilizando a área de transferência através dos atalhos das teclas (*Control+V*, ou colar), ou botão “Paste”, no caso da versão *online*.

## 5.6 Perspectivas futuras

Como forma de dar continuidade ao trabalho e aumentar as funcionalidades do *software* desenvolvido, sugere-se a implementação de técnicas de classificação,

como SIMCA (Soft independent modelling of class analogies, ou Modelagem Independente Flexível por Analogias de Classes) e de análise quantitativa, como o CLS (Classical Least Squares, ou Mínimos Quadrados Clássicos), o PLS (Partial Least Squares, ou Mínimos Quadrados Parciais) e o iPLS (interval Partial Least Squares, ou Mínimos Quadrados Parciais em intervalos). Além disso, promover outras funções à solução online, como importação de arquivos de extensão “.asc” e “.sp” e iPCA.

## 6 CONCLUSÕES

O desenvolvimento deste trabalho permitiu criar um *software* gratuito contemplando as técnicas de análise de agrupamento hierárquico (HCA), análise de componentes principais (PCA), análise de componentes principais por intervalos (iPCA), assim como técnicas de correção, transformação dos dados e detecção de amostras anômalas, com as seguintes características:

- Fácil instalação e manuseio. O *software* consiste em 3 arquivos, totalizando um espaço físico menor que 10Mb para sua execução e não requer uma instalação “formal” no Windows<sup>®</sup>, o que possibilita seu uso, sem necessidade de privilégios de administrador.
- Interface gráfica amigável, apresentando janelas, botões e menus autoexplicativos o que permite seu uso sem necessidade de conhecimento de programação de rotinas em nível de usuário.
- Múltiplas entradas de dados: de espectroscopia no infravermelho (arquivos adquiridos em espectrômetro Perkin-Elmer com extensão “.sp” e .ASCII), de imagens digitais e a partir da área de transferência (dados de outras naturezas).
- Gráficos e tabelas com recursos de cores para identificação das amostras, possibilitando uma melhor interpretação dos dados.
- Múltiplas saídas de dados: nos formatos de texto (Excel, “.txt” e ASCII) e figuras (“.bmp”, “.png”, “.jpg”, entre outros).

As principais funcionalidades do *software* foram exploradas utilizando espectros no infravermelho médio e próximo de óleos vegetais e imagens digitais de diferentes tipos de óleo diesel. Como forma de validar os resultados do *software*, os mesmos conjuntos de dados foram analisados no Matlab<sup>®</sup> e os resultados em ambas as ferramentas coincidiu nas mais diversas combinações.

Além da versão *desktop*, o reuso dos algoritmos permitiu disponibilizar uma versão *online* com alguns recursos básicos de tratamento de dados além das técnicas de análise de agrupamento hierárquico (HCA) e análise de componentes principais (PCA).

## REFERÊNCIAS

AGOSTON, M. K. *Computer Graphics and Geometric Modeling: Implementation and Algorithms*. London: Springer, 2005.

ALISKE, Marcelo Adriano. *Medidas de espectroscopia no infravermelho médio para a determinação do teor de biodiesel em óleo diesel*. 2010. 100 f. Dissertação. (Programa de Pós-Graduação em Engenharia e Ciências dos Materiais – Mestrado) – Universidade Federal do Paraná, Curitiba, 2010.

ANTONELLI, A.; COCCHIB, M.; FAVAA, P.; FOCAB, G.; FRANCHINIA, G., MANZINIB, D.; ULRICIA, A. Automated evaluation of food colour by means of multivariate image analysis coupled to a wavelet-based classification algorithm. *Analytica Chimica Acta*, v. 515, p. 3-13, jul. 2004.

BARBOSA, L. C. *Espectroscopia no Infravermelho na Caracterização de Compostos Orgânicos*. Viçosa: Editora da UFV, 2007.

BICUDO, A. J. A.; PINTO, L. F. B.; CYRINO, J. E. P. Clustering of ingredients with amino acid composition similar to the nutritional requirement of Nile tilapia. *Scientia Agricola*, v. 67, n. 5, p. 517-523, set./out. 2010.

BRASIL. Resolução Nº 6, de 16 de setembro de 2009. *Conselho Nacional de Política Energética - CNPE*. Disponível em: <[http://www.mme.gov.br/mme/galerias/arquivos/conselhos\\_comite/CNPE/resolucao\\_2009/Resoluxo\\_6\\_CNPE.pdf](http://www.mme.gov.br/mme/galerias/arquivos/conselhos_comite/CNPE/resolucao_2009/Resoluxo_6_CNPE.pdf)>. Acesso em: 15 de jan. 2014.

BRASIL. Resolução Nº 65, de 09 de dezembro de 2011. *Agência Nacional do Petróleo*. Disponível em: <[http://nxt.anp.gov.br/nxt/gateway.dll/leg/resolucoes\\_anp/2011/dezembro/ranp%2065%20-%202011.xml](http://nxt.anp.gov.br/nxt/gateway.dll/leg/resolucoes_anp/2011/dezembro/ranp%2065%20-%202011.xml)>. Acesso em: 15 de jan. 2014.

BRERETON, Richard G. *Applied Chemometrics for Scientists*. University of Bristol, UK. Chichester: Wiley, 2007.

BRERETON, Richard G. *Chemometrics for Pattern Recognition*. University of Bristol, UK. Chichester: Wiley, 2009.

BRO, R. ; SMILDE, A. K. Centering and scaling in component analysis. *Journal of Chemometrics*, v. 17, p. 16-33, 2001.

BURGER, W.; BURGE, M. *Principles of digital image processing - Fundamental techniques*. London: Springer-Verlag, 2009.

CAMO Unscrambler Software Inc.: Woodbridge, NJ, USA, 2006

CHAMPION, J.; CHAMPION, C.; SULLIVAN, R. ZedGraph - A flexible charting library for .NET. 2012. Disponível em: <<http://sourceforge.net/projects/zedgraph/>>. Acesso em 23 jun. 2012.



COSTA FILHO, P. A.; POPPI, R. J. Aplicação de algoritmos genéticos na seleção de variáveis em espectroscopia no infravermelho médio. Determinação simultânea de glicose, maltose e frutose. *Química Nova*, v. 25, n. 1, 2002.

FERRÃO, Marco Flôres. *Aplicação de técnicas espectroscópicas de reflexão no infravermelho no controle de qualidade de farinha de trigo*. 2000. 219 f Tese (Doutorado) - Universidade Estadual de Campinas, Campinas, 2000.

FERRÃO, Marco Flôres. Técnicas de reflexão no infravermelho aplicadas na análise de alimentos. *Tecno-lógica*, Santa Cruz do Sul, v. 5, n. 1, p. 63-85, 2001.

FERREIRA, M. C.; ANTUNES, A.; MELGO, M.; VOLPE, P. Quimiometria I: calibração multivariada, um tutorial. *Química Nova*, São Paulo, v. 22, n. 5, 1999.

FURTADO, J. C.; FERRÃO, M.; KONZEN, P.; MOLZ, R.; BASSANI, I. Otimização via algoritmo genético e busca tabu na determinação de proteína em farinha de trigo por reflexão no infravermelho. *Tecno-Lógica*, Santa Cruz do Sul, v. 6, n. 2, p. 41-71, 2002.

GELADI, P.; GRAHN, H.; ESBENSEN, K; BENGTSSON, E. Multivariate Image Analysis. *Trends in analytical chemistry*, v. 11, n. 3, 1992.

GLASBEY, C. A; HORGAN, G. W. *Image Analysis for the Biological Sciences*. Edinburgh: Wiley, 1995. Disponível em: <<http://www.bioss.ac.uk/people/chris.html>>. Acesso em: 11 dez. 2013.

GODINHO, M. S.; PEREIRA, R.; RIBEIRO, K.; SCHIMIDT, F.; OLIVEIRA, A.; OLIVEIRA, S. Classificação de refrigerantes através de análise de imagens e análise de componentes principais (PCA). *Química Nova*, v. 31, n. 6, p. 1485-1489, 2008.

GORAYEB, Sâmia Rodrigues. *Ferramenta computacional para geração do gráfico de controle multivariado de  $T^2$  de Hotelling*. 2010. 58 f. Dissertação (Programa de Pós-graduação em Matemática e Estatística) – Universidade Federal do Pará, Belém, 2010.

GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing*. New Jersey: Pearson Education, 2008.

GUILLÉN, M. D.; CABO, N. Infrared spectroscopy in the study of edible oils and fats. *Journal of the Science of Food and Agriculture*. v. 75, p. 1–11, 1997.

HANBURY, Allan. Constructing cylindrical coordinate colour spaces. *Pattern Recognition Letters*, v. 29, n. 4, p. 494–500, mar. 2008.

INFOMETRIX Inc.: *Pirouette User Guide*. Version 4.5, Bothell, WA, 2011.

JARVIS, R. M.; BROADHURST, D.; JOHNSON, H.; O'BOYLE, N.; GOODACRE, R. PyChem - a multivariate analysis package for Python. *Bioinformatics*, v. 22, n. 20, p. 2565-2566, jul. 2006.

JONSSON, F. A stand-alone implementation of the Savitzky–Golay smoothing filter. *Tutorial*. Disponível em: <<http://jonsson.eu/programs/cweb/sgfilter/>>. Acesso em: 03 nov. 2013.

LAQQA. Desenvolvido pelo Laboratório de quimiometria em química analítica. UNICAMP, Campinas, 2006. Disponível em: <<http://laqqa.iqm.unicamp.br/>>. Acesso em: 26 nov. 2012.

LEARDI, R.; NORGAARD, L. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *Journal of Chemometrics*, v.18, n. 11, p. 486–497, 2004.

MARQUES FILHO, O.; VIEIRA NETO, H. *Processamento Digital de Imagens*, Rio de Janeiro: Brasport, 1999.

MATLAB<sup>®</sup>. The Mathworks, Inc.: Natick, MA, USA.

MATOS, G.; PEREIRA-FILHO, E.; POPPI, R.; ARRUDA, M. Análise exploratória em química analítica com emprego de quimiometria: PCA e PCA de imagens. *Revista Analytica*, n. 6, p. 38-46, 2003.

MICROSOFT VISUAL STUDIO 2010<sup>®</sup>. Microsoft Corporation: Redmond, WA, USA.

MILLER, James M.; MILLER, Jane C. *Statistics and Chemometrics for Analytical Chemistry*. Harlow: Pearson Education, 2010

MINGOTTI, S. A. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Belo Horizonte: Editora UFMG. 2005.

MORETTO, E.; FETT, R.. *Tecnologia de óleos e gorduras vegetais na indústria de alimentos*. São Paulo: Varela, 1998.

NORGAARD, L. iToolbox for Matlab<sup>®</sup>. 2004. Disponível em: <<http://www.models.life.ku.dk/itoolbox>>. Acesso em: 26 abr. 2012.

MUELLER, D.; HESSE, H.; COSTA, A. B; MARDER, L.; FERRÃO, M. Aplicação da espectroscopia no infravermelho associada à análise de componentes principais (PCA) na classificação e diferenciação de óleos vegetais. In: WORKSHOP EM SISTEMAS E PROCESSOS INDUSTRIAIS. 1., 2011, Santa Cruz do Sul, *Anais...* Santa Cruz do Sul: EDUNISC, 2011.

NUNES, C. A.; FREITAS, M.; PINHEIRO, A.; BASTOS, S. Chemoface: a novel free user-friendly interface for chemometrics. *Journal of the Brazilian Chemical Society*, v. 23, n. 11, p. 2003-2010, 2012.

PANERO, F. S.; VIEIRA, M.; CRUZ, A.; MOURA, M.; SILVA, H. Aplicação da análise exploratória de dados na discriminação geográfica do quiabo do Rio Grande do Norte e Pernambuco. *Eclética Química*, v. 34, n. 3, p.33-40, 2009.

PASQUINI, C. Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society*, v. 14, p. 198-219, 2003.

PAVIA, D. L.; LAMPMAN, G.; KRIZ, G. *Introdução à espectroscopia*. São Paulo: Cengage Learning, 2010.

PEDRINI, H.; SCHWARTZ, W. R. *Análise de imagens digitais - princípios, algoritmos e aplicações*. São Paulo: Thomson Learning, 2008.

PEREIRA, A. C. F.; PONTES, M.; GAMBARRA NETO, F.; SANTOS, S.; GALVÃO, R.; ARAÚJO, M. NIR Spectrometric determination of quality parameters in vegetable oils using iPLS and variable selection. *Food Research International*, v. 41, p. 341-348, 2008.

PERKIN-ELMER Inc. *User Guide*. Massachusetts: USA, 2010.

PETROBRÁS DISTRIBUIDORA. Desenvolvido pela Petrobrás Distribuidora. 2014. Apresenta informações gerais sobre a empresa. Disponível em: <<http://www.br.com.br/wps/portal/portalconteudo/produtos/paraindustriasetermeletricas/oleodiesel>>. Acesso em: 15 de jan. 2014.

SANCHEZ, Jordi Cruz. *Desarrollo de nuevas metodologías de espectroscopía NIR en aplicaciones industriales. estudio y aplicación de técnicas de imagen química (NIR-CI)*. 2010. 247 f. Tese. (Programa de Doctorado de Química), Universitat Autònoma de Barcelona, Barcelona, 2010.

SANTOS, A. R. *Metodologia científica: a construção do conhecimento*. 4. Rio de Janeiro: DP&A, 2001.

SMITH, Alvy Ray. Color gamut transform pairs. *Computer Graphics*, v. 12, n. 3, p. 12-19, ago. 1978, Disponível em: <<http://www.icst.pku.edu.cn/course/ImageProcessing/2013/resource/Color78.pdf>> Acesso em: 12 set. 2013.

SMITH, L. I. Tutorial on Principal Component Analysis. 2002. Disponível em: <[http://www.sccg.sk/~haladova/principal\\_components.pdf](http://www.sccg.sk/~haladova/principal_components.pdf)> Acesso em: 14 set. 2012.

SOARES, I. P.; REZENDE, T.; PEREIRA, R.; SANTOS, C.; FORTES, I. Determination of biodiesel adulteration with raw vegetable oil from ATR-FTIR data using chemometric tools. *Journal of Brazilian Chemical Society*, v. 22, n. 7, p.1229-1235, 2011.

SOUZA, C. R. The Accord.NET Framework. 2012. Disponível em <<http://accord.googlecode.com>> Acesso em: 10 abr. 2012.

TAVARES, Patrícia Silva. *O gráfico de controle multivariado  $T^2$  de Hotelling como instrumento de análise da qualidade numa indústria de alumínio*. 2003. 80 f. Dissertação. (Programa de Pós-Graduação em Engenharia de Produção), Universidade Federal de Santa Catarina, Florianópolis, 2003.

TEÓFILO, R. F. *Métodos Quimiométricos: Uma Visão Geral - Conceitos básicos de quimiometria*, Viçosa: Universidade Federal de Viçosa, 2013. Disponível em: <<http://www.deq.ufv.br/acoes/Download.php?id=23>>. Acesso em: 15 de jan. 2014.

TRINDADE, M. M. et al. Espectroscopia no infravermelho por reflexão total atenuada horizontal (HATR) aplicada na identificação de óleos vegetais comerciais. *Tecnológica*, Santa Cruz do Sul, v. 9, n. 1, p. 59-74, jan./jun. 2005.

UNIVERSIDADE FEDERAL DE SANTA CATARINA. Desenvolvido pelo curso de Arquitetura e Urbanismo da UFSC. Curso de iluminação. Disponível em: <[http://www.arq.ufsc.br/labcon/arq5656/Curso\\_Iluminacao/07\\_cores/luz\\_01.htm](http://www.arq.ufsc.br/labcon/arq5656/Curso_Iluminacao/07_cores/luz_01.htm)> Acesso em: 15 de jan. 2014.

VIERA, M. S.; FRANCESQUETT, J.; FACHINI, D.; GERBASE, A.; FERRÃO, M. Avaliação de adulteração de misturas biodiesel diesel empregando espectroscopia no infravermelho e análise por componentes principais. In: XX ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO. 10., 2010. São Carlos, *Anais...* São Carlos: ABEPRO, 2010.

VISENTAINER, J. V.; FRANCO, M. R. B. *Ácidos graxos em óleos e gorduras: identificação e quantificação*. São Paulo: Varela, 2006.

XIAOBO, Z.; JIEWENA, Z.; POVEYB, M.; HOLMESB, M.; HANPINA, M. Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*, v. 667, p. 14-32, 2010.

YANG, H.; IRUDAYARAJ, J.; PARADKAR, M. Discriminant analysis of edible oils and fats by FTIR, FT-NIR and FT-Raman spectroscopy. *Food Chemistry*, v. 93, p. 25-32, 2005.

WEHRENS, Ron. *Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences*. Berlin: Springer, 2011.

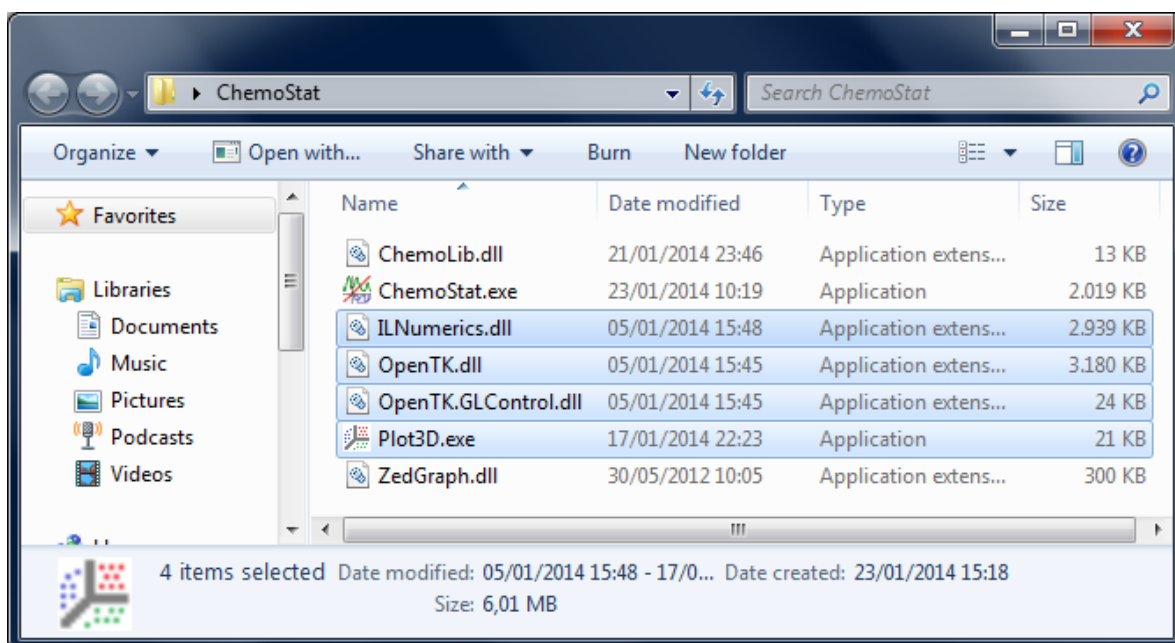
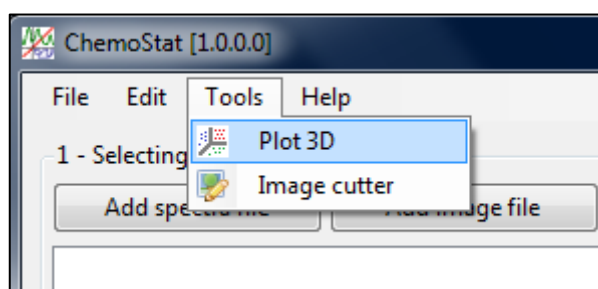
ZAMBIAZI, R. C; PRZYBYLSKI, R.; ZAMBIAZI, M.; MENDONÇA, C. Fatty acid composition of vegetable oils and fats. *Boletim do Centro de Pesquisa de Processamento de Alimentos*, Curitiba, v. 25, n. 1, p. 111-120, jan./ jun. 2007.

## ANEXO A: Programa “Plot 3D”

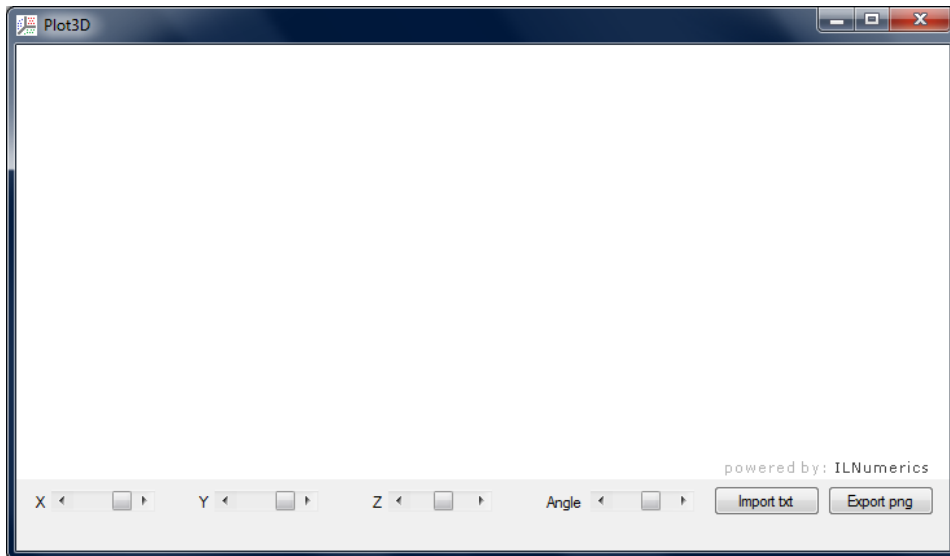
O programa Plot3D foi desenvolvido como uma ferramenta auxiliar para plotagem de scores de PCA em três dimensões. É composto por 4 arquivos:

- Plot3D.exe
- ILNumerics.dll
- OpenTK.dll
- OpenTK.GLControl.dll

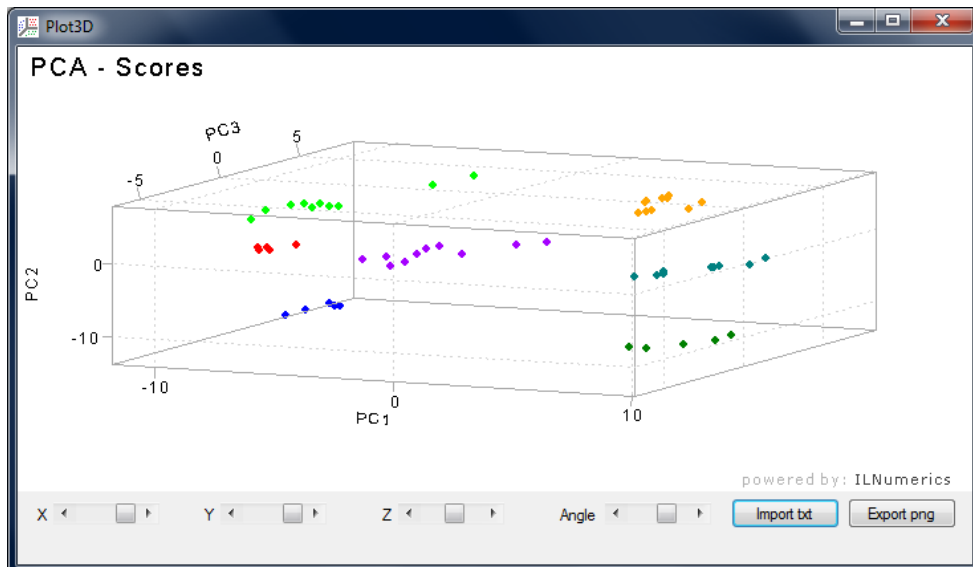
Para sua utilização, basta executar o arquivo “Plot3D.exe” diretamente, ou através da barra de menu “*Tools >> Plot 3D*” do *software* ChemoStat, caso seus quatro arquivos estiverem no mesmo diretório do programa ChemoStat, conforme demonstram as figuras abaixo.



A tela principal do programa Plot3D possui 4 botões de rolagem do gráfico na barra inferior, além dos botões de inserir arquivo texto “*Import txt*” (proveniente do gráfico de *scores* do ChemoStat), além do botão de exportação “*Export png*” no formato de imagem “png”, como demonstra a figura a baixo.



Uma vez importado o gráfico pode ser manipulado pelos botões “X”, “Y”, “Z”, “Angle” ou até mesmo pelo deslocamento do *mouse* ao clicar num ponto do gráfico. Para isso, basta ficar pressionado o botão esquerdo do *mouse* durante o seu movimento.



## ANEXO B: Fórum de discussão sobre algoritmos da PCA.

### Infometrix Software

A place to exchange information about Pirouette and other Infometrix products.

## Signs of PCA scores are inverted in 2 nearly identical sets

Infometrix Software Forum Index -> Pirouette Algorithms

View previous topic :: View next topic	
Author	Message
<b>Isramos</b> Moderator  Joined: 08 Oct 2003 Posts: 87 Location: Infometrix, Inc., Bothell, WA	<p>Posted: Mon Nov 14, 2005 1:06 pm    Post subject: signs of PCA scores are inverted in 2 nearly identical sets</p> <hr/> <p>A user writes:</p> <p><b>Quote:</b></p> <p>Would you or one of your colleagues be willing to provide us with some assistance to help us explain an unexpected result when using Pirouette 3.11 for PCA and understand its cause?</p> <p>The Study: The goal of this study was to compare 2 methods. The experiment consists of 11 samples each analyzed for 255 variables using 2 methods (the first method and the second method). The absolute intensities of the data collected for each sample differ between methods, however the relative intensities do not (i.e. "scaling differences" are observed between methods) These "scaling differences" between the 2 methods may vary from sample to sample A sample-by sample visual inspection of the data from 2 methods revealed nearly no differences between the two methods once scaled (i.e. once scaled there is virtually no difference between the results from two methods)</p> <p>The Analyses: A PCA for 9 factors was performed on the data collected with the first method in Pirouette 3.11 using mean-center preprocessing and a divide by (Sample 1-norm) transform A PCA for 9 factors was performed on the data collected with the second method in Pirouette 3.11 using mean-center preprocessing and a divide by (Sample 1-norm) transform</p> <p>The Results: The PC1/PC2/PC3 scores plot from the first method PCA was nearly the "mirror image" of the PC1/PC2/PC3 scores plot from the second method PCA</p>

	<p>The PC1/PC2/PC3 loadings plot from the first method PCA was nearly the "mirror image" of the PC1/PC2/PC3 loadings plot from the second method PCA</p> <p>Comparison of the PC1/PC2/PC3 scores from the first method PCA to those of the second method PCA revealed:          For all 11 samples, the signs of the Factor 1 scores in the first method PCA were inverted when compared to those of the second method PCA, although the magnitudes were nearly identical.          For all 11 samples, Factor 2 scores in the first method PCA were nearly identical to those of the second method PCA (identical in sign, nearly identical in magnitude).          For all 11 samples, the signs of the Factor 3 scores in the first method PCA were inverted when compared to those of the second method PCA, although the magnitudes were nearly identical.</p> <p>Comparison of the PC1/PC2/PC3 loadings from the first method PCA to those of the second method PCA revealed the same sign inversion in Factors 1 &amp; 3.</p> <p>We do not understand why the signs of the scores and loadings of Factors 1 &amp; 3 are inverted when comparing the first method PCA to that of the second.          Could the "scaling differences" of the initial data be responsible for this result?          What conventions does Pirouette use to determine the sign of the scores &amp; loadings (the orientation of the PCs)?</p>
<p><b>Back to top</b></p>	
<p><b>Isramos</b> Moderator</p> <p>Joined: 08 Oct 2003          Posts: 87          Location: Infometrix, Inc., Bothell, WA</p>	<p>Posted: Mon Nov 14, 2005 1:17 pm Post subject:</p> <hr/> <p>Did you not get the smoke and mirrors add-on to Pirouette? You have stumbled upon a case where the latter is used.</p> <p>Seriously, this is not an unexpected result. Without getting into too much detail, let me explain how PCA works. The core algorithm behind the data decomposition step in PCA is usually called something like NIPALS (based on the Power Method, if you needed to know). This algorithm does the decomposition one eigenvector at a time. To derive the first eigenvector, the algorithm forms a working vector that is then optimized through a series of iterations. That initial working vector can be a vector of ones, or ones and zeros, or random numbers, or even a vector extracted from the data set. You will see all of these in the literature.</p> <p>Once the iterations have converged and the first eigenvector produced, the information represented in this first factor is "removed" from the data set, and this new data matrix is then processed to extract the second eigenvector, etc. If this process is repeated until as many eigenvectors as the rank of the matrix are computed, you could then reconstruct the original data matrix exactly by multiplying the scores (complete rank) by these eigenvectors (or, loadings, also complete rank). If you flip the sign on all</p>



values in any pair of eigenvectors, the reconstructed matrix will be the same. Thus, the sign is not critical, but the magnitudes are.

The data sets you are working with are similar but not identical. Therefore, it is not unexpected that the sign on the converged 1st eigenvector from the 2nd data set could be opposite that derived from the 1st data set. The reconstructed data will be appropriate, however, if an additional eigenvector also has its sign inverted. This is what you observed in the sign on the 3rd eigenvector.

In your case, as you described, the scores and loadings plot look like mirror images, but, and this is key, the relationships among samples in the score plot or among variables in the loadings plot will be unchanged with sign inversion.

But, how to compare the results from the two data sets? If you only want to make plots, then you could rotate the scores of one data set to be 180 degrees from that of the other. Remember, the sign is not critical, just the magnitudes.

Another way would be to project all data (i.e., from both sets combined) into the PCA space of just one of the two sets (i.e., do a PCA prediction). The resulting PCA prediction scores should show just how similar are the two sets, but in one plot rather than two, and without sign inversion between the two sets.

---

Scott

[Back to top](#)

**Isramos**  
Moderator

Joined: 08 Oct 2003  
Posts: 87  
Location: Infometrix,  
Inc., Bothell, WA

Posted: Mon Nov 14, 2005 1:35 pm Post subject:

The user replies:

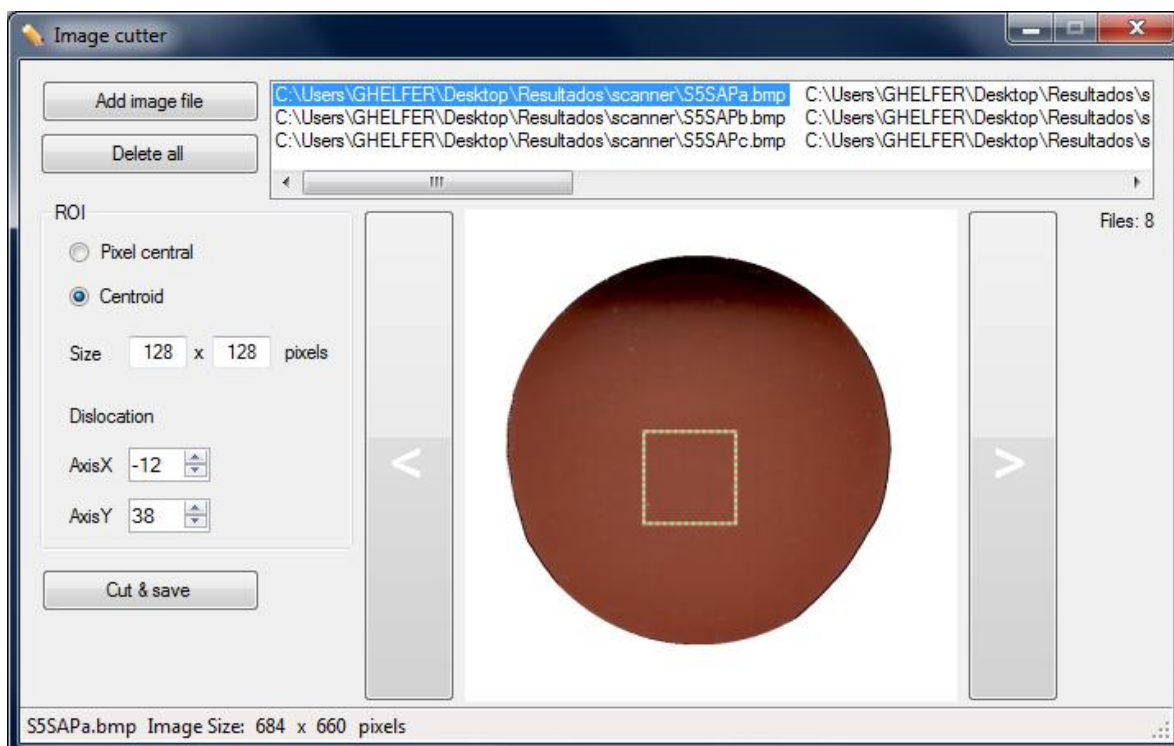
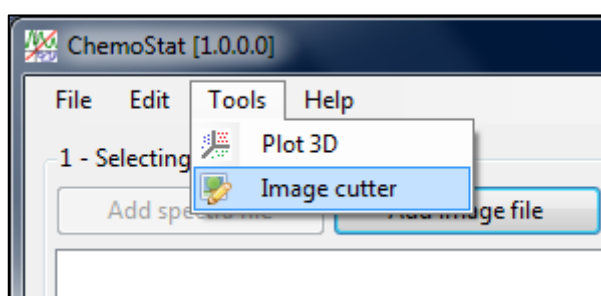
**Quote:**

Thank you for the speedy response and the thorough explanation. They are both greatly appreciated. Your explanation has provided us with the reassurance that these two data sets are comparable, regardless of the signs selected for the PC factors. It is comforting to know that we can count on your sound advice.

## ANEXO C: Editor de imagens

A ferramenta de edição “*Image cutter*” foi desenvolvida para recortar uma janela ou região de pixels menor em relação a imagem inteira. Sua utilização destaca-se nos escaneamentos devido ao uso de uma maior resolução na aquisição de imagens

Para sua utilização, basta acessá-lo pela barra de menu “*Tools >> Image cutter*”, conforme demonstram as figuras abaixo.



O botão “*Add image file*” adiciona as imagens na área de trabalho, enquanto que o botão “*Delete all*” exclui todas as mesmas.

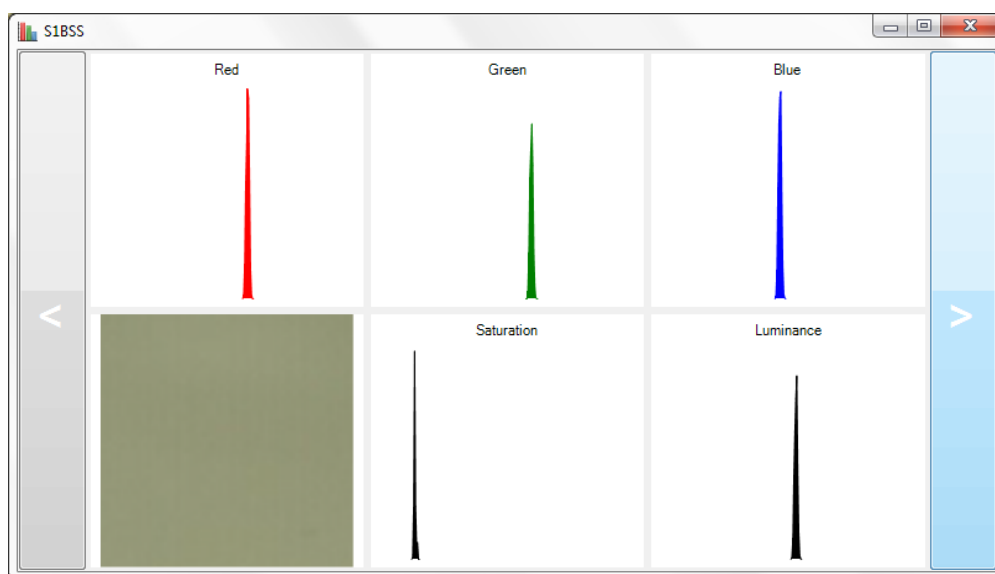
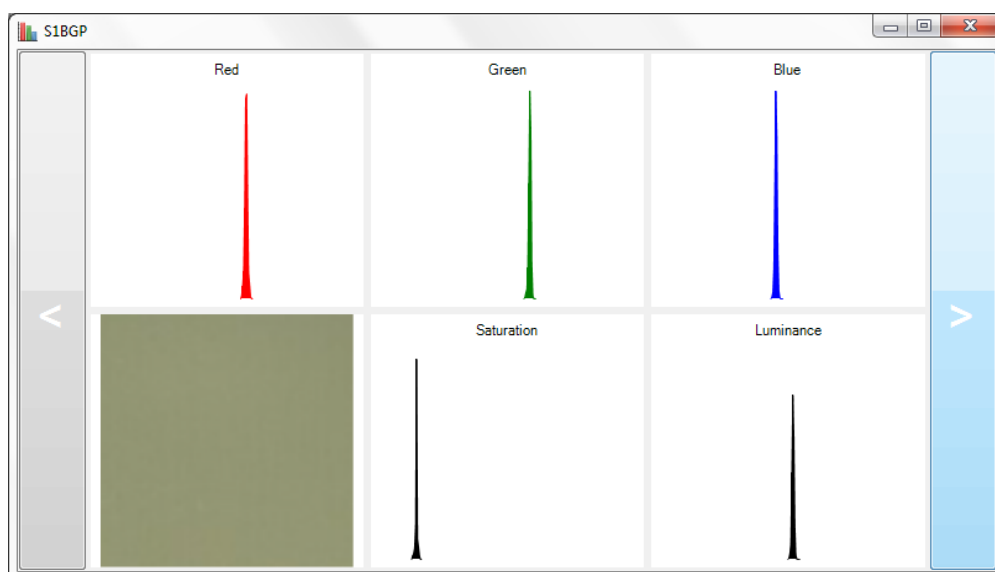
O campo “ROI” (“*region of interest*”, do inglês, “região de interesse”) controla a área demarcada na imagem. O método “*Pixel central*” busca o centro da imagem e aplica uma janela de acordo com o tamanho informado no campo “*Size*”. Já o método “*Centroid*” realiza a busca do centro das massas (centróide) para imagens circulares ou quadrangulares e aplica, de forma similar, uma janela de acordo com o campo “*Size*”.

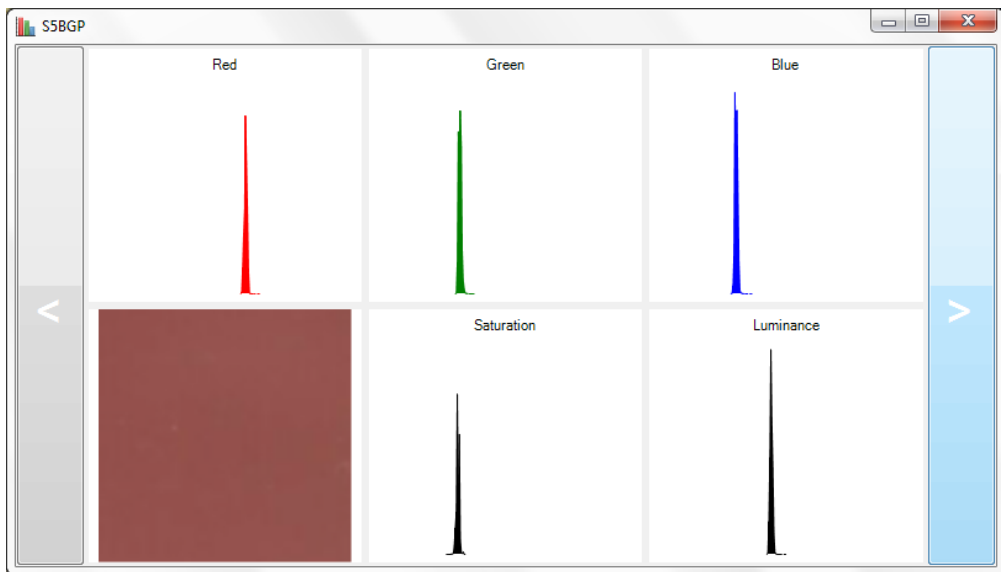
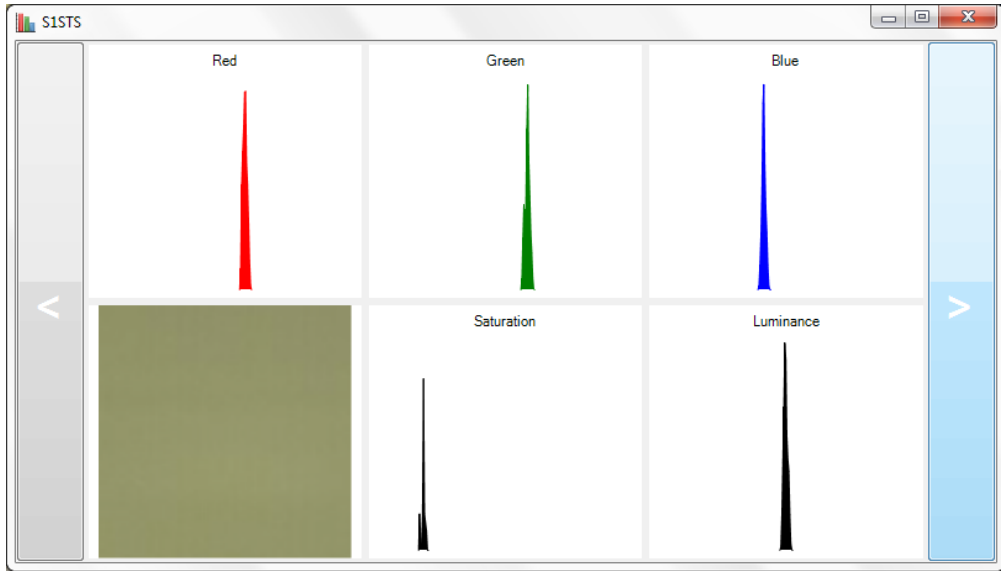
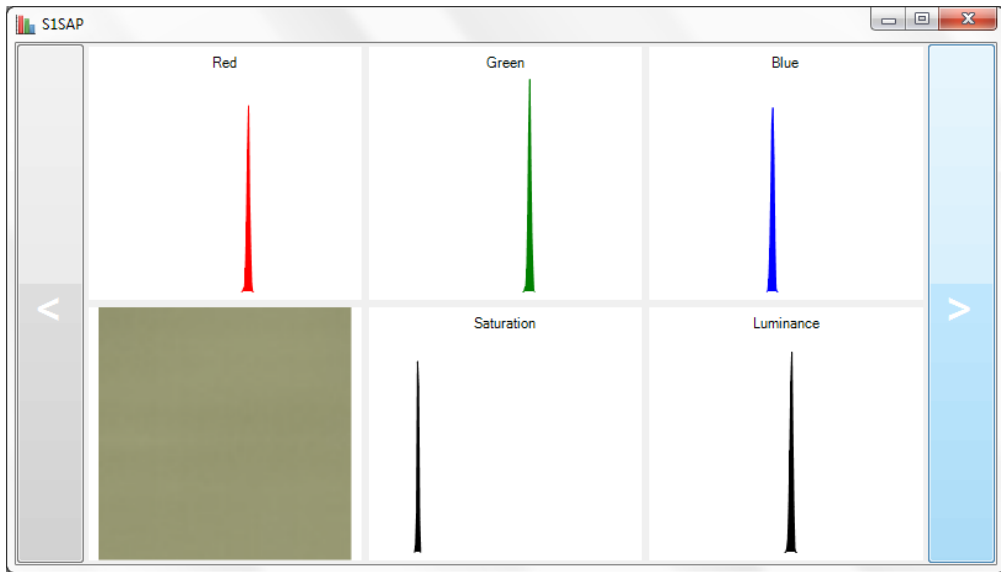
Os valores de “*AxisX*” e “*AxisY*” promovem o deslocamento da área demarcada para direita ou esquerda e para cima ou baixo, respectivamente. O valor inicial é o ponto (0,0), para tanto são aceitos valores positivos ou negativos para seu deslocamento.

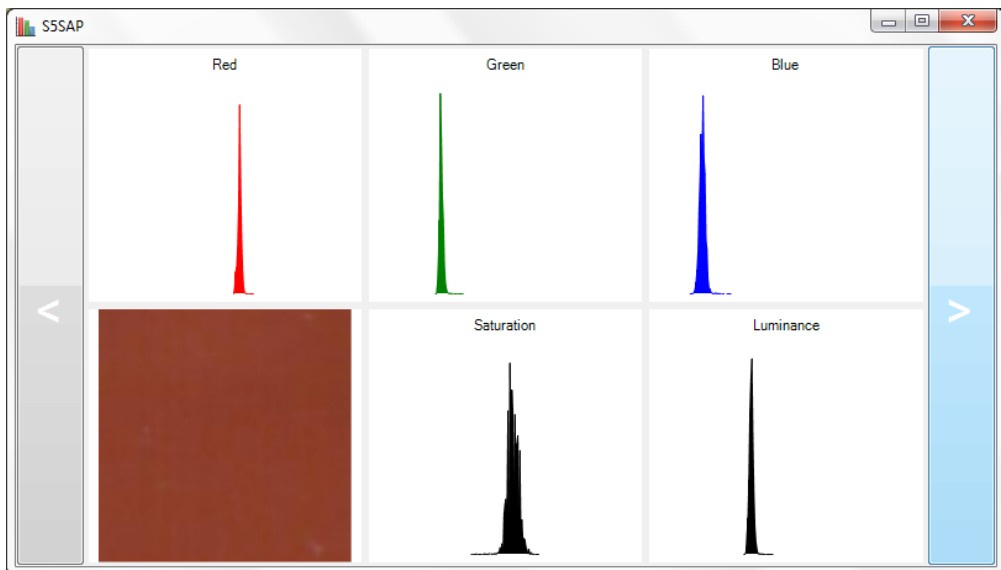
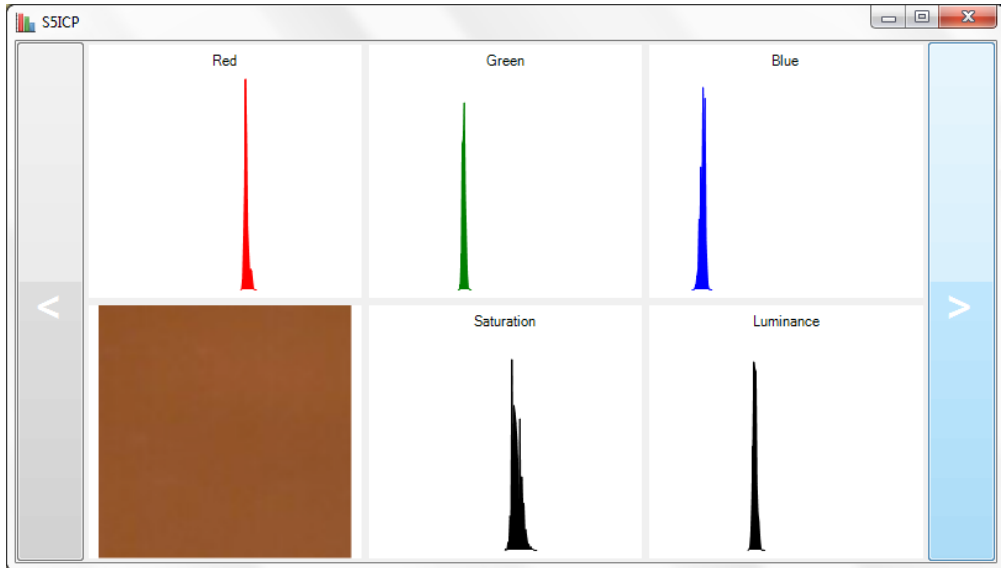
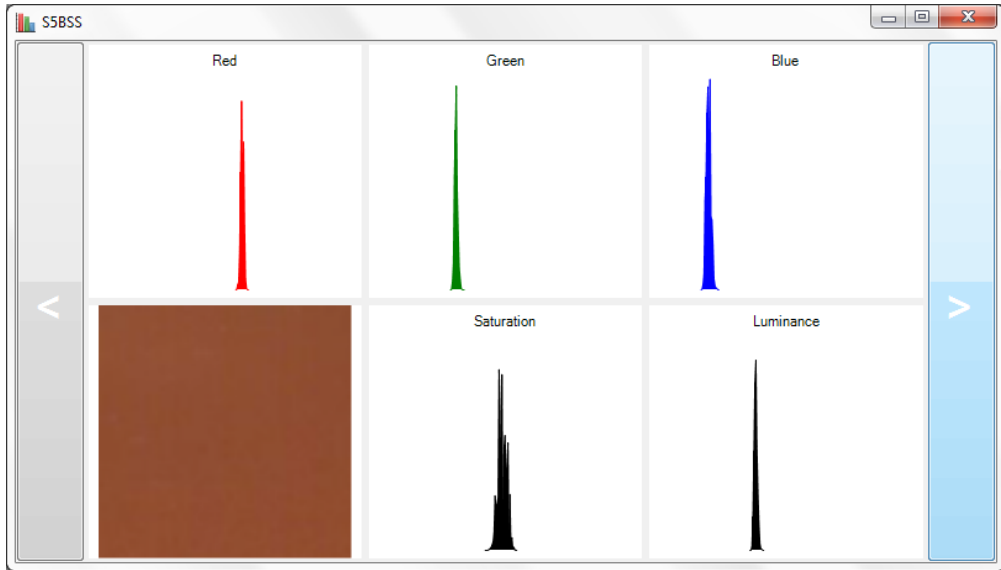
Os botões laterais “<” e “>” permitem a visualização de todas as imagens importadas, avançando ou retrocedendo uma a uma, verificando em cada uma delas a ROI demarcada.

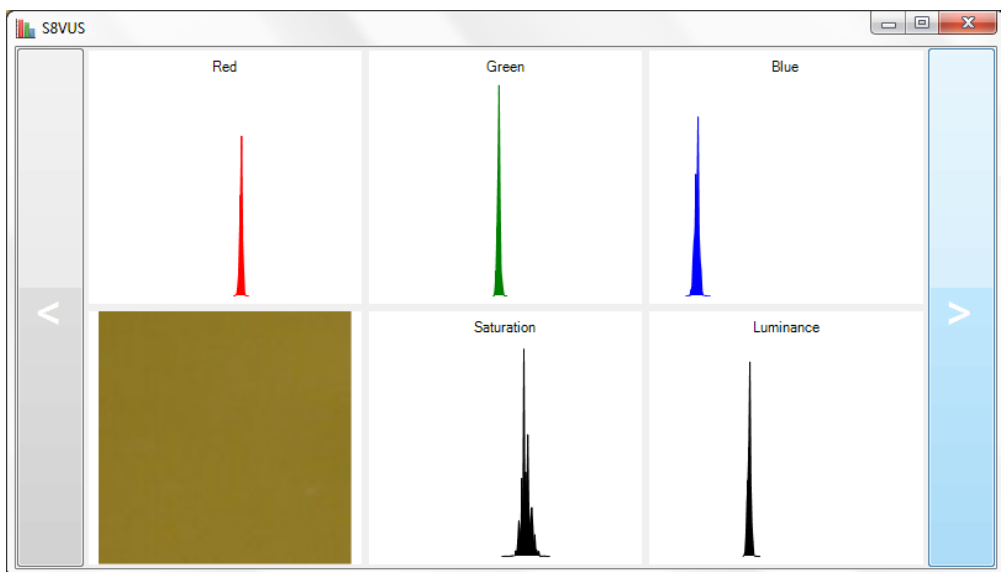
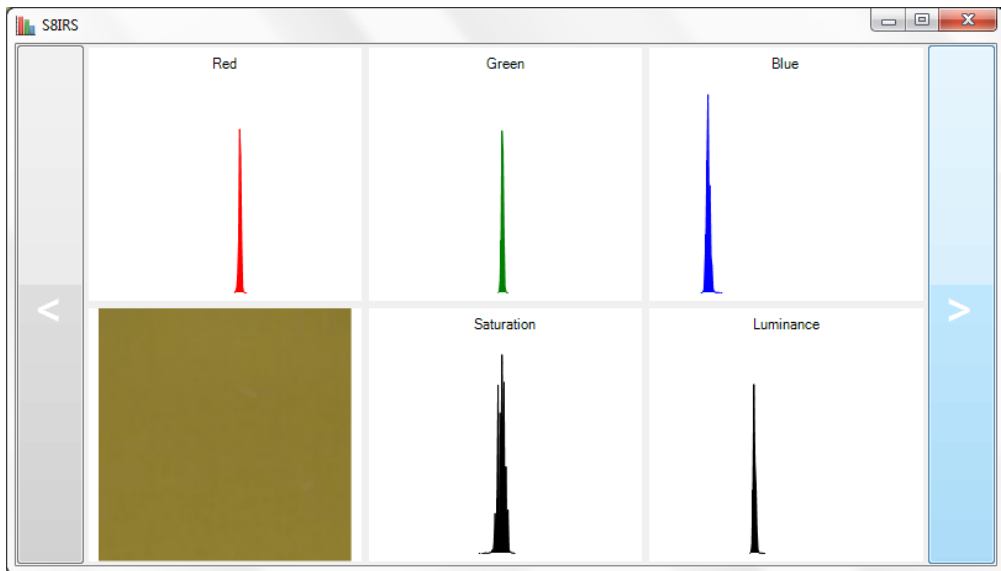
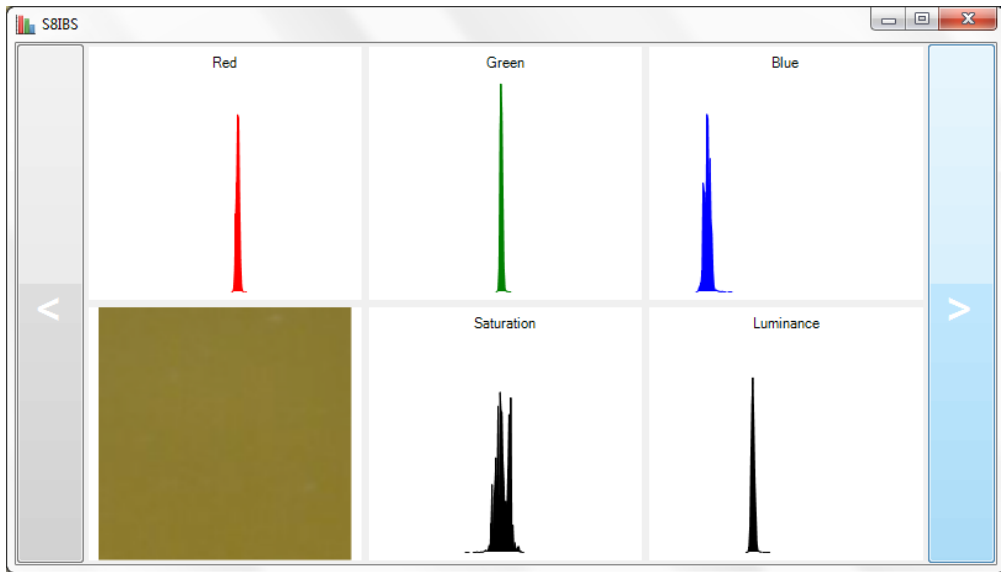
O botão “*Cut & save*” abre uma janela solicitando o diretório de destino e gera imagens novas a partir da ROI selecionada.

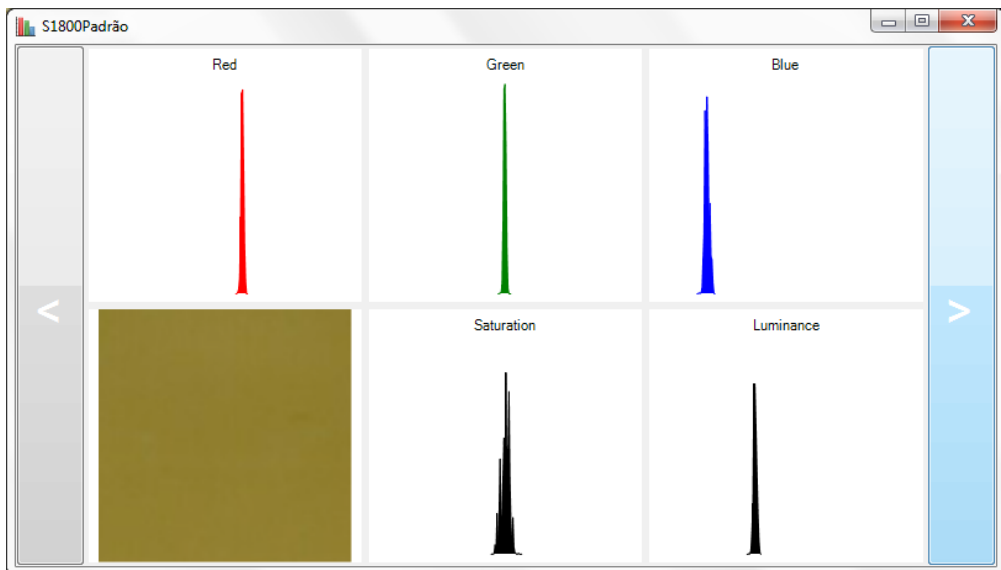
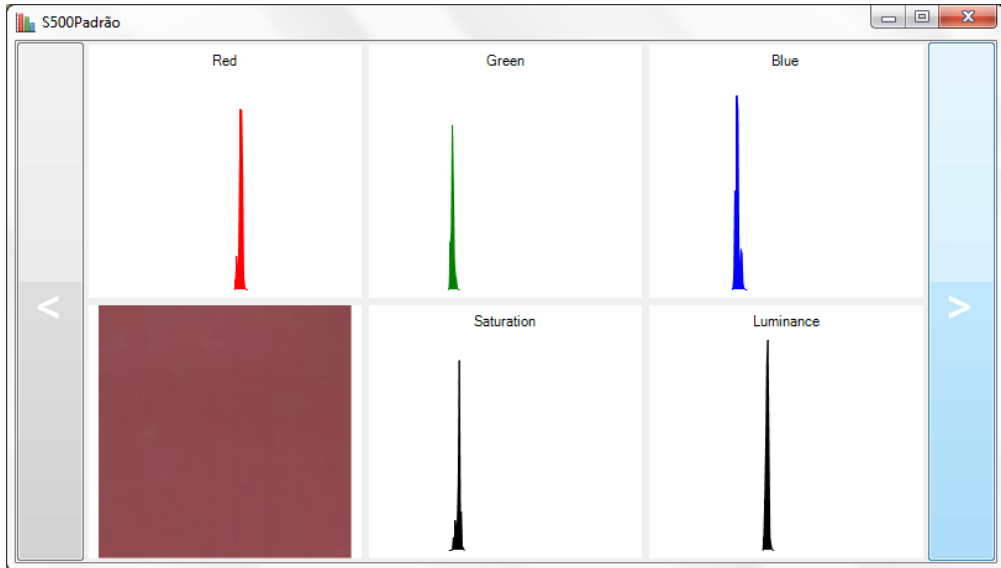
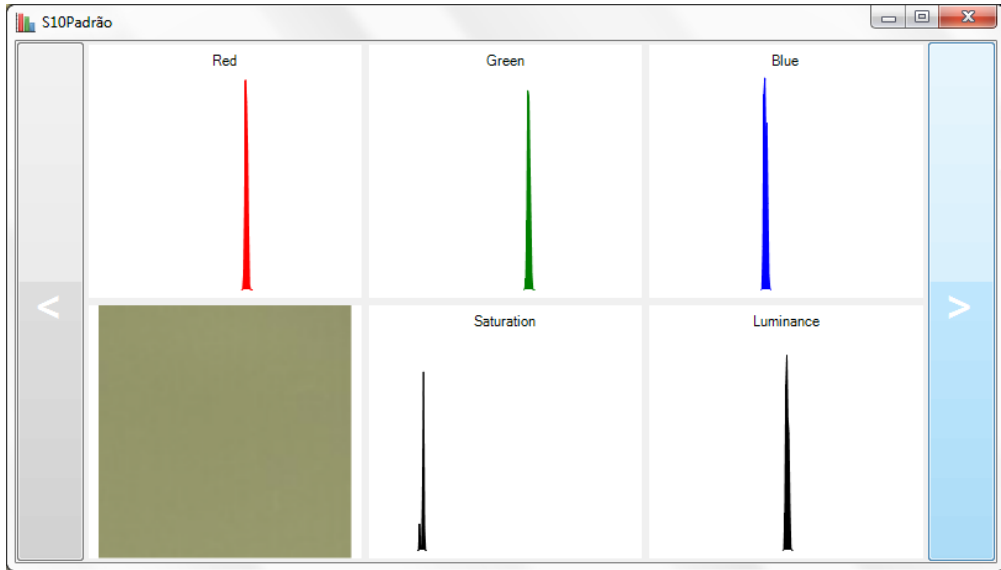
## ANEXO D: Imagens escaneadas das amostras de óleo diesel comercial













## ANEXO E: Apêndices de publicações

Publicações em congressos	Títulos	Observações
 <p><b>WSPi</b> 2013 WORKSHOP EM SISTEMAS E PROCESSOS INDUSTRIAIS</p>	<p>Análise de Componentes Principais por Intervalos (iPCA) como Método de Seleção de Região Espectral no Infravermelho Próximo para Discriminação de Óleos Vegetais</p>	<p>Santa Cruz do Sul, RS, 8 a 10 de maio de 2013.</p>
 <p><i>I Escola de Inverno de Quimiometria</i> 1<sup>st</sup> Chemometric Winter School</p>	<p>Identificação da Degradação de Erva Mate por Infravermelho Próximo</p>	<p>São Carlos, SP, 26 a 30 de agosto de 2013.</p>
 <p><b>XX Encontro de Química da Região Sul</b></p>	<p>Uso do Histograma de Imagens Digitais e Análise por Componentes Principais na Diferenciação de Óleo Diesel Comercial.</p>	<p>Lajeado, RS, 14 a 16 de novembro de 2013.</p>
Publicações em periódicos	Títulos	Observações
 <p><b>TECNO-LÓGICA</b> Revista do depto. de Química e Física, do depto. de Engenharia, Arquitetura e Ciências Agrárias e do Mestrado em Tecnologia Ambiental Mestrado em Sistemas e Processos Industriais</p>	<p>Análise de Componentes Principais por Intervalos (iPCA) como Método de Seleção de Região Espectral no Infravermelho Médio e Próximo para Discriminação de Óleos Vegetais</p>	<p>Santa Cruz do Sul, v. 17, n. 2, p. 108-116, jul./dez. 2013</p>