

ANÁLISE DE COMPONENTES PRINCIPAIS POR INTERVALOS (*i*PCA) COMO MÉTODO DE SELEÇÃO DE REGIÃO ESPECTRAL NO INFRAVERMELHO PRÓXIMO PARA DISCRIMINAÇÃO DE ÓLEOS VEGETAIS

Gilson Augusto Helfer¹, Lilian de Fátima Ferreira da Silva², Fernanda Carla Bock², Luciano Marder^{1,2*}

¹Programa de Pós-graduação em Sistemas e Processos Industriais, Universidade de Santa Cruz do Sul (UNISC), Santa Cruz do Sul – Brasil

²Departamento de Química e Física, Universidade de Santa Cruz do Sul (UNISC), Santa Cruz do Sul – Brasil

*E-mail: lucianomarder@unisc.br

ÁREA

Sistemas Baseados em Conhecimento para melhoria de processos industriais

RESUMO

Neste trabalho foi empregada a análise de componentes principais por intervalo (*i*PCA) como método de seleção de variáveis, associada à espectroscopia no infravermelho próximo, na discriminação de óleos vegetais. Os seguintes óleos vegetais foram analisados: óleo de soja, canola, girassol, arroz, milho, algodão e amendoim. Para tal análise foi utilizado um espectrômetro Perkin Elmer Spectrum 400, com esfera de integração (NIRA), empregando-se refletância difusa e especular no intervalo entre 4.000 e 10.000 cm^{-1} , resolução de 4 cm^{-1} e 32 varreduras. Os espectros foram adquiridos em quintuplicata, foram normalizados e posteriormente corrigidos com as transformadas SNV (variação normal padrão) e primeira derivada com uma janela de 5 pontos. Os modelos foram desenvolvidos através do aplicativo Matlab versão 7.0 empregando análise de componentes principais por intervalos (*i*PCA) disponível no pacote *i*Toolbox. Foram realizados modelos de 8, 16 e 32 intervalos, com os dados centrados na média. O presente trabalho indica que é possível discriminar diferentes óleos vegetais através da espectroscopia de infravermelho associada à análise de componentes principais por intervalo (*i*PCA) como método de seleção de variáveis.

Palavras-chave: Análise multivariada, óleos vegetais, quimiometria, *i*PCA

1 INTRODUÇÃO

Autenticidade é um critério de qualidade muito importante para óleos e gorduras comestíveis, devido a uma grande diferença de preços entre os diferentes tipos de derivados de óleo e de gordura. No entanto, a determinação dessa autenticidade é tradicionalmente um processo demorado e trabalhoso, geralmente usando métodos cromatográficos. Combinada com métodos quimiométricos, a espectroscopia vibracional na região do infravermelho, é uma técnica analítica que vem sendo utilizada para verificar a autenticidade dos óleos e gorduras comestíveis, devido à sua simplicidade, rapidez e facilidade de preparação da amostra¹.

Ao utilizarmos esta técnica associada à quimiometria, nos deparamos com um grande número de variáveis que implicam desde problemas no desempenho computacional até nos modelos gerados pelas características potenciais dos interferentes, principalmente devido as não linearidades².

Para resolver este impasse utilizam-se métodos de seleção de variáveis de inúmeros tipos, a partir da literatura molecular vibracional até subtração de espectros, redes neurais, exame de partículas e heurísticas das mais diversas formas. Os métodos de seleção são baseados no princípio da escolha de um pequeno número de variáveis segregadas a partir do espectro original, onde a remoção de variáveis não informativas vão produzir modelos mais robustos.

Xiabo³ demonstrou que a complexidade de um modelo pode ser reduzida através de uma pré-seleção criteriosa de comprimentos de onda. O simples fato de dividirmos o espectro em regiões proporcionais em tamanho e analisarmos cada um separadamente, já caracteriza um método de seleção. Neste sentido o presente trabalho sugere o emprego da análise das componentes principais por intervalos (iPCA) como um método de seleção de variáveis na discriminação de óleos vegetais.

2 REVISÃO DA LITERATURA

2.1 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 química⁴. As unidades fundamentais da maioria dos lipídios são os ácidos graxos, que podem ser saturados (ligações simples) e insaturados (duplas ligações). Os óleos vegetais têm como principal componente em proporção cadeias insaturadas⁵. Zambiasi⁶ via cromatografia gasosa (GC-FID) identificou a composição dos ácidos graxos em diversos óleos vegetais comercializados no Brasil, sendo os resultados apresentados na Tabela 1.

Tabela 1. Composição das amostras de óleos vegetais utilizadas no estudo

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

2.2 Espectroscopia no infravermelho

É uma técnica que nos permite identificar uma amostra usando radiação infravermelha através da vibração molecular da mesma. 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 infravermelho⁷.

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 frequências (energias) selecionadas de radiação, ocorrendo sua vibração por deformações axiais e/ou angulares⁸.

De acordo com Pasquini⁹, a região do infravermelho é absorvida pela maioria dos compostos orgânicos e inorgânicos que possuem ligações covalentes. A região do infravermelho próximo (4.000 à 14.290 cm^{-1}) tem como características picos largos e de baixa intensidade, enquanto no infravermelho médio (200 à 4.000 cm^{-1}) aparecem picos muito intensos e geralmente estreitos¹.

2.3 Espectroscopia Difusa e Especular

A espectroscopia no infravermelho próximo com esfera de integração (NIRA) apresenta medida de refletância absoluta e eficiente combinando especular, detecção uniforme mesmo em amostras heterogêneas e redução dos efeitos de polarização oriundos do feixe de iluminação e da amostra. É uma excelente ferramenta para análises qualitativas e quantitativas, possui detector Índio-Gálio-Arsênio (InGaAs) proporcionando uma alta relação sinal/ruído¹¹.

Ferrão¹² afirma que, em um procedimento de refletância difusa, parte do feixe sofre também um fenômeno especular, sendo a mais importante àquela que ocorre na interface da superfície da amostra. A reflexão especular é de maior intensidade na região onde a amostra apresenta forte absorção e, neste caso, podendo gerar severas distorções no espectro obtidos.

2.4 Quimiometria

Quimiometria é a aplicação de técnicas matemáticas e estatísticas à análise de dados relativos aos processos químicos tanto na área espectroscópica como cromatográfica. Sua importância é

caracterizada pela complexidade dos dados que podem ser obtidos via análise de rotina. A necessidade de analisar um conjunto extenso de variáveis demonstra um aspecto multivariado¹³.

Os dados multivariados, geralmente correspondem a uma matriz X de valores, correspondendo a m variáveis para n amostras, conforme ilustra Figura 1.

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

Figura 1. Representação de uma matriz de dados

2.5 Análise por Componentes Principais - PCA

Segundo Ferreira¹⁴, a base fundamental da maioria dos métodos modernos para tratamento de dados multivariados é o PCA (Principal Component Analysis), que 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.

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), resultando na Equação 1:

$$A = U \times S \times V' \quad (\text{Eq. 1})$$

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 matemática¹⁵.

2.6 Análise por Componentes Principais em intervalos - *i*PCA

A análise *i*PCA divide o conjunto de dados (variáveis) em um número de intervalos equidistantes. Para cada intervalo realiza-se uma PCA, que descreve a variação (ou dispersão) de um determinado conjunto de dados em intervalos pré-estabelecidos, apresentando os resultados em gráfico de escores. Este método objetiva dar uma visão geral dos dados e pode ser útil em estudos exploratórios e interpretativos em relação aos agrupamentos entre as amostras¹⁶.

3METODOLOGIA

3.1 Amostragem

Para o desenvolvimento desse trabalho foram utilizados oito diferentes óleos vegetais disponíveis no comércio local representada na Tabela 2.

Tabela 2. Amostras de óleos vegetais utilizadas no estudo

Tipo de óleo	Código	Marca	Lote	Fabricação	Validade
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	-

3.2 Aquisição dos espectros no infravermelho

O espectrômetro Perkin-Elmer 400 Series equipado com um detector InGaAs (Índio-Gálio-Arsênio) foi utilizado para as análises de espectroscopia no infravermelho próximo (FT-NIR). Os espectros adquiridos na região entre 4.000 e 10.000 cm^{-1} com uma resolução de 4 cm^{-1} e 32 varreduras, foram corrigidos com um espectro de branco (background) para cada replicata. Cada amostra foi analisada 5 vezes e, na sequência, os dados obtidos do espectro foram normalizados. O conjunto de espectro médios das amostras encontra-se na Figura 2.

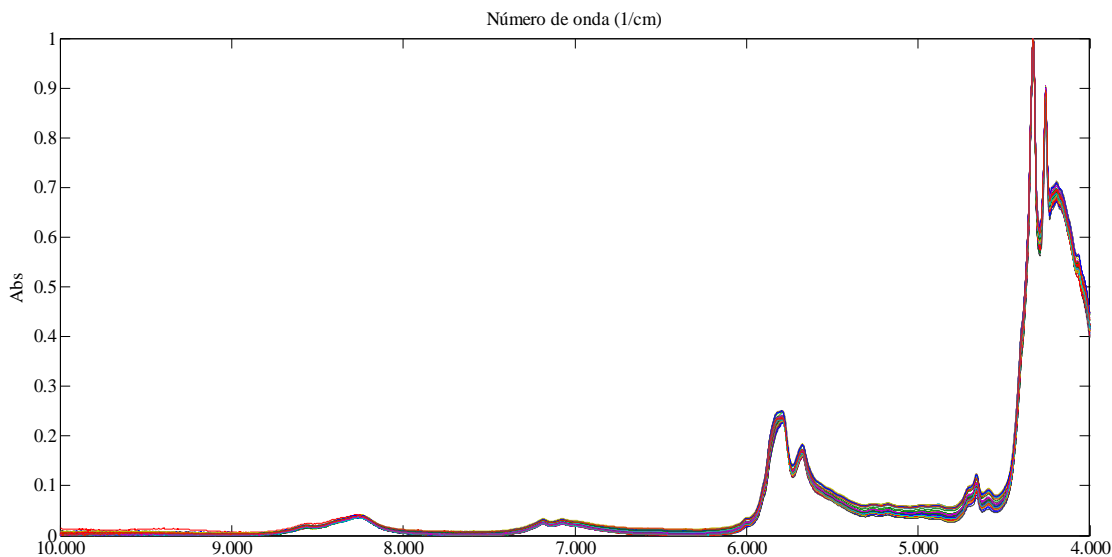


Figura 2. Espectro normalizado das amostras de óleo na faixa de 10.000 à 4.000 cm^{-1}

3.3 Análise dos dados

A matriz de espectros foi analisada por *iPCA*, através do software Matlab 7.0, empregando-se o pacote *iToolbox* desenvolvido por Norgaard¹⁷ com o objetivo de identificar a região espectral que caracteriza a separação dos diferentes tipos de óleo vegetal. Para esta análise os espectros foram divididos em 8, 16 e 32 regiões equidistantes. Além disso, foram aplicados métodos de correção de sinais, como a SNV (variação normal padrão) e a primeira derivada (janela de 5 pontos). Finalmente, visando melhorar o desempenho quimiométrico, os dados foram centrados na média, tornando o valor médio dos dados o ponto central de cada espectro. Na Figura 3 podem ser observados os espectros das amostras de óleos com a aplicação da SNV e primeira derivada.

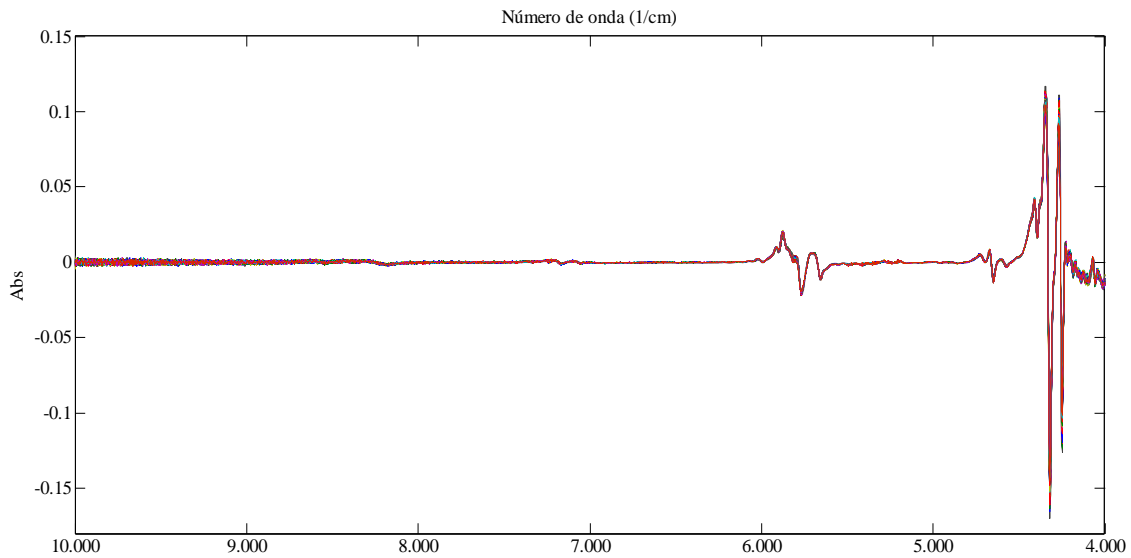


Figura 3. Espectros das amostras de óleos com aplicação da SNV e primeira derivada

4 RESULTADOS E DISCUSSÕES

Primeiramente foi realizada uma PCA com toda a faixa espectral, ou seja, de 4.000 a 10.000 cm^{-1} . Como pode ser observado na Figura 4, utilizando toda essa faixa, não se obteve uma boa separação dos óleos vegetais.

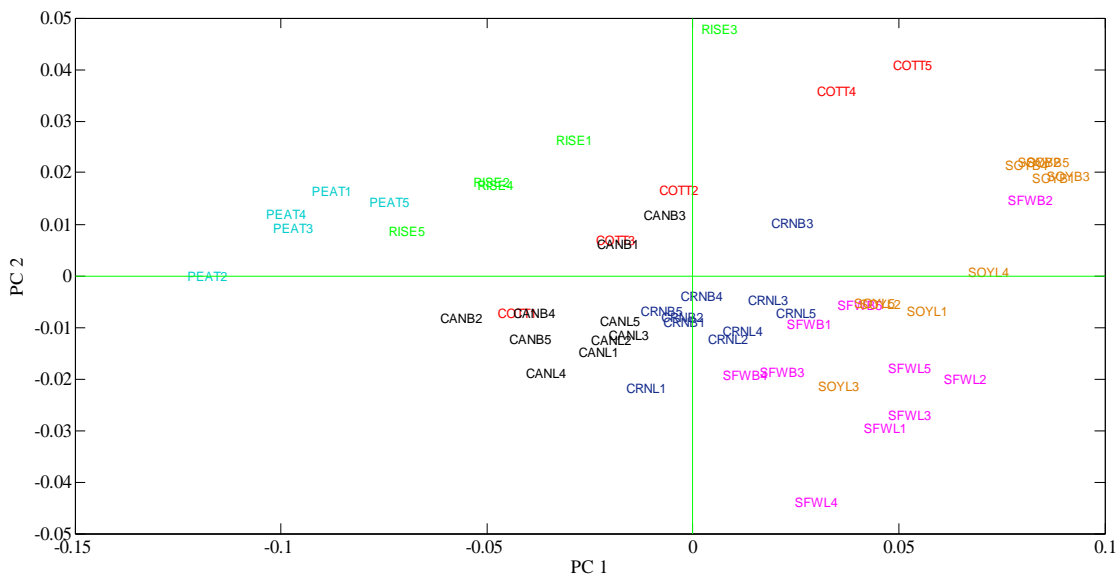


Figura 4. Gráfico de escores PC1 x PC2 - modelo utilizando toda faixa espectral

A seguir foram gerados modelos iPCa com 8, 16 e 32 intervalos. Os melhores resultados foram obtidos dividindo o espectro em 32 intervalos. A Figura 5 apresenta o gráfico da variância percentual para cada região estudada. Neste gráfico, para cada intervalo, ou seja, para cada região do espectro, foi calculada a variância, em percentual, em cada componente principal. O intervalo 23 acumula 94,70% das informações nas duas primeiras componentes principais.

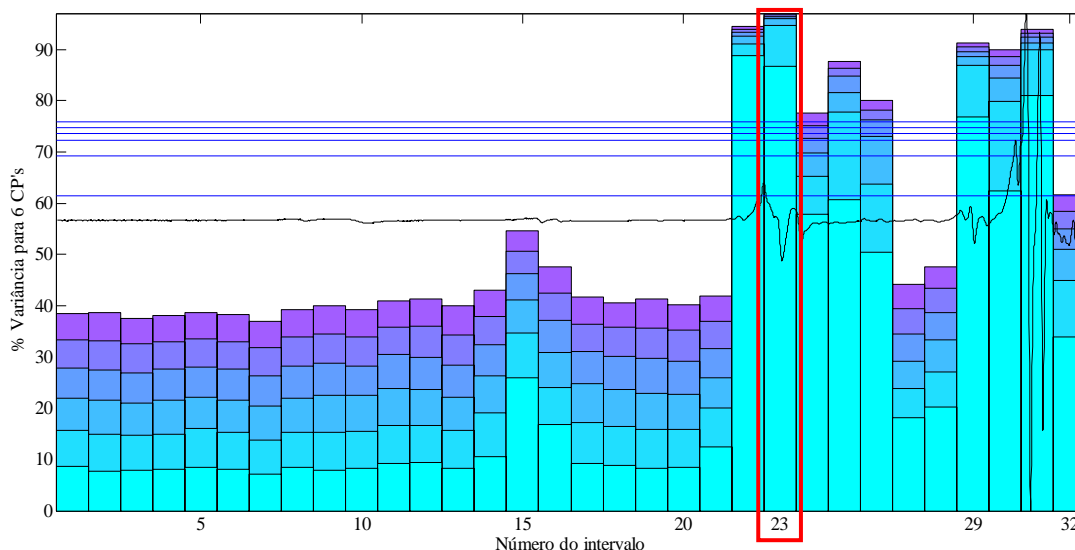


Figure 5. Variação percentual das componentes principais divididos em 32 intervalos

As alturas das barras representam, em forma percentual, a variância contida em cada componente principal para cada intervalo. Cabe ressaltar que a linha traçada 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 23 são apresentados na Tabela 3.

Tabela 3. Valores para variância e % acumulado da variância (PCA)

CP	%Var	%Acum
1	86,75	86,75
2	7,94	94,70
3	1,27	95,97
4	0,44	96,42
5	0,33	96,75
6	0,30	97,05

A região espectral onde se obteve uma melhor discriminação dos óleos vegetais correspondeu à faixa entre 5.683 e 5.869 cm^{-1} (intervalo 23). Segundo Yang¹ a região entre 5.500 e 6.000 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. A separação dos óleos vegetais estudados está associada ao fato de que cada óleo/gordura difere na composição, comprimento e grau de insaturação de ácidos graxos nas cadeias

dos triglicerídeos. A Figura 6 apresenta o gráfico dos escores para PC1 x PC2, obtido para o intervalo 23.

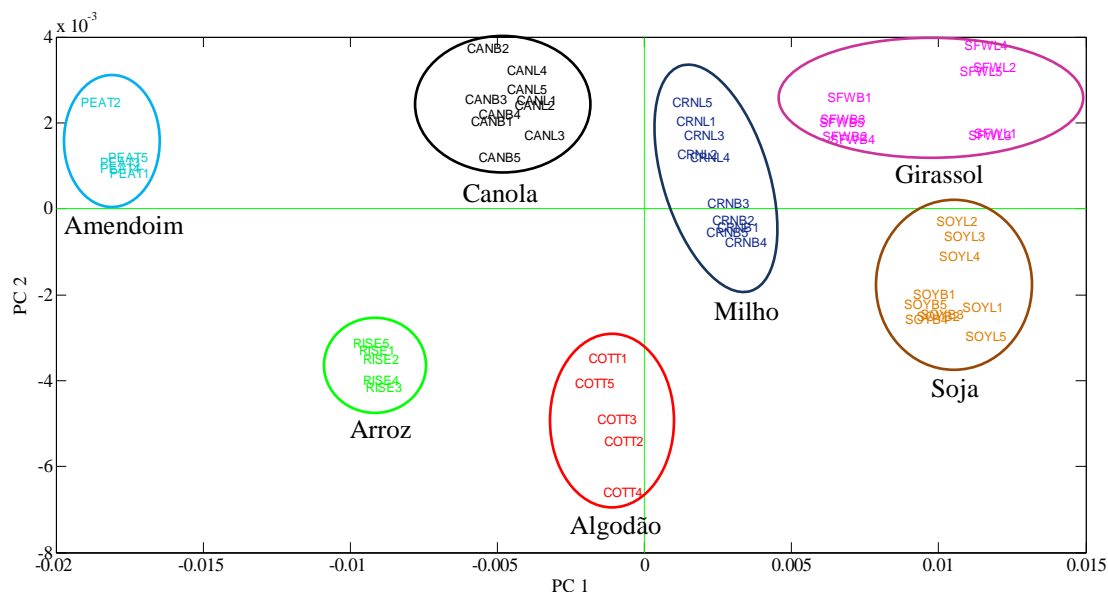


Figura 6. Gráfico de escores PC1 x PC2 – intervalo 23

5 CONSIDERAÇÕES FINAIS

O presente trabalho indica que a análise de componentes principais por intervalos (*i*PCA), como ferramenta de seleção de variáveis, possibilita encontrar regiões representativas visando a discriminação de diferentes óleos vegetais. A medida que a região foi sendo dividida e explorada por mais intervalos, foi se obtendo modelos mais robustos e de melhor distinção dos óleos de acordo com seus componentes de saturação e insaturação. A região espectral onde se obteve uma melhor discriminação dos óleos vegetais correspondeu à faixa entre $5.683 \text{e} 5.869 \text{ cm}^{-1}$.

6 REFERÊNCIAS

1. YANG, H.; IRUDAYARAJ, J.; PARADKAR, M. M. Discriminant analysis of edible oils and fats by FTIR, FT-NIR and FT-Raman spectroscopy. *Food Chemistry*, n. 93, p. 25-32, 2005.
2. 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. COBRA, M. Marketing essencial: conceitos, estratégias, controle. São Paulo: Atlas, 1988.
3. XIABO, Z. et al. Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*, n. 667, p. 14-32, 2010.
4. MORETTO, E.; FETT, R. Tecnologia de óleos e gorduras vegetais na indústria de alimentos. São Paulo: Varela, 1998.
5. VISENTAINER, J. V.; FRANCO, M. R. B. Ácidos graxos em óleos e gorduras: identificação e quantificação. São Paulo: Varela, 2006

6. ZAMBIAZI, R. C. et al. 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.
7. BARBOSA, L.C. Espectroscopia no Infravermelho na Caracterização de Compostos Orgânicos. Ed. da UFV: Viçosa. 189 p., 2007.
8. PAVIA, Donald L. et al. Introdução à espectroscopia. São Paulo: Cengage Learning, 2010.
9. PASQUINI, C. Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications. J. Braz. Chem.Soc., V.14. 198-219 p. 2003.
10. SILVERSTEIN, Robert M.; WEBSTER, Francis X.; KIEMLE, David J. Identificação espectrométrica de compostos orgânicos. 7. ed. Rio de Janeiro: LTC, 2007.
11. PERKIN ELMER, User Guide. Inc. Massachusetts: 2010.
12. FERRÃO, M.F. Técnicas de reflexão no infravermelho aplicadas na análise de alimentos. Tecno-lógica., Santa Cruz do Sul, v.5, n.1, p.63-85, 2001.
13. TRINDADE, M. M. et al. Espectroscopia no infravermelho por reflexão total atenuada horizontal (HATR) aplicada na identificação de óleos vegetais comerciais. Tecno-Lógic. Vol. 9, n. 1, p. 59-74, jan./jun. 2005.
14. FERREIRA, M. M. C. et al. Quimiometria I: calibração multivariada, um tutorial. Química Nova. Vol. 22, n. 5, set. 1999.
15. FERRÃO, M. F. Aplicação de técnicas espectroscópicas de reflexão no infravermelho no controle de qualidade de farinha de trigo. 2000. Tese (Doutorado) - Universidade Estadual de Campinas, 2000.
16. 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. Vol.18, n. 11, p. 486-497, 2004.
17. NORGAARD, Lars. Itoolbox for MATLAB. University of Copenhagen. Disponível em: <<http://www.models.life.ku.dk/itoolbox>>. Acesso em: 07 mar. 2012.