**Iloane dos Santos Lima**

# Métodos Multivariados Aplicados para Classificação de Azeite de Oliva Extra Virgem

Recife

Agosto de 2017

**UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO**

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

# Métodos Multivariados Aplicados para Classificação de Azeite de Oliva Extra Virgem

Dissertação apresentada ao Programa de Pós-Graduação em Biometria e Estatística Aplicada como exigência parcial a obtenção do título de Mestre em Biometria e Estatística Aplicada.

**Área de Concentração: Biometria e Estatística Aplicada**

Orientador: Prof. Dr. Moacyr Cunha Filho

Co-orientador: Prof. Dr. Ronaldo Dionísio da Silva

Recife

Agosto de 2017

**UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO**

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

# Métodos Multivariados Aplicados para Classificação de Azeite de Oliva Extra Virgem

Iloane dos Santos Lima

Dissertação julgada adequada para a obtenção do título de Mestre em Biometria e Estatística Aplicada. Defendida e aprovada em 31/08/2017 pela comissão examinadora.

Presidente:

_____

Prof. Dr. Moacyr Cunha Filho
Universidade Federal Rural de Pernambuco

Banca examinadora:

_____

Prof. Dr. Ronaldo Dionísio da Silva
Instituto Federal de Educação, Ciência e Tecnologia - Campus Vitória – PE

_____

Prof. Dr. Manoel Rivelino Gomes de Oliveira
Universidade Federal da Paraíba

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca Central, Recife-PE, Brasil

L732m Lima, Iloane dos Santos
Métodos Multivariados para Classificação de Azeite de Oliva Extra Virgem / Iloane dos Santos Lima. – 2017. 48 f. : il.

Orientador: Moacyr Cunha Filho.
Coorientador: Ronaldo Dionísio da Silva.
Dissertação (Mestrado) – Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Biometria e Estatística Aplicada, Recife, BR-PE, 2017.
Inclui referências.

1. Metabonômica 2. Estatística da silhueta 3. Agrupamento Fuzzy I. Cunha Filho, Moacyr, orient. II. Silva, Ronaldo Dionísio da, coorient. III. Título

CDD 574.018

*Dedico este trabalho a Deus e a minha família.*

# Agradecimentos

Realizar um texto de agradecimento envolve uma emoção que não consigo descrever apenas em palavras, mas vamos nos esforçar para conseguir expressar essa emoção.

Agradeço primeiramente a Deus e a virgem Maria pela vida que me proporcionam a cada amanhecer.

Aos meus pais, em especial a minha mãe, Maria das Graças, pelo apoio diário que me dar forças para querer conseguir sempre o melhor. A minha linda irmã, Italine Lima, que me ensina a ter coragem sempre e aos meus sobrinhos, Ilo Lima e Athur Lima.

Ao meu Orientador Professor Doutor Moacyr Cunha Filho pela disponibilidade e pelos conhecimentos construídos que ultrapassam a vida acadêmica.

A minha eterna Orientadora Ana Patrícia Siqueira Tavares Falcão pela orientação realizada para os ensinamentos de conduta, o certo e o errado, obrigada pela disponibilidade e estímulo.

Ao meu amigo e Co-orientador Ronaldo Dionísio da Silva, sem sua disponibilidade e incentivo não estaria nesse momento realizando esses agradecimentos.

A minha amiga/irmã que esse programa me deu, Yara Lopes Abreu, criança muito obrigada por segurar na minha mão todas as vezes que eu pensei que não iria conseguir.

Aos meus amigos Jucarlos Rufino, Patrícia Ximenes, Sara Lúcia Castillo e David Avellaneda pelo apoio e disponibilidade diária.

Aos meus amigos de Luta David Venâncio, Manoel Rivelino e Erivaldo Neto sem vocês tudo iria ser mais complicado.

A minha irmã/amiga/prima Lívia Karoline Guerra Feitosa que me ergue sempre que me ver preste a desistir.

*“E o segundo mandamento, semelhante a este, é: Amarás o teu próximo como a ti mesmo. Não há outro mandamento maior do que estes. Marcos 12:31.”*

# Resumo

Metabonômica é uma estratégia que baseia- se na identificação de padrões de um determinado problema biológico, por meio da obtenção de dados espectroscópicos/espectrométricos de um dado biofluido, o uso da estatística para extração dessas informações contribui significativamente para realização de classificações de grupos. Desse modo, o presente trabalho objetivou-se ao uso da estratégia metabonômica, baseados em espectros de ressonância magnética nuclear de hidrogênio (RMN $^{1}$H) e técnicas estatísticas multivariadas de agrupamento (Análise de Componentes Principais (PCA), Agrupamento *Fuzzy*) de amostras de azeite de oliva extra virgem. Utilizou-se 40 amostras de azeite de oliva extra virgem para este estudo. A partir da matriz de dados espectrais, utilizou-se o pré-processamento normalização pela soma, nas amostras. A partir da PCA, 99,1% da variância explicada utilizando dois componentes apenas, não foi possível observar agrupamentos naturais dos dados. Com a aplicação do agrupamento *Fuzzy,* constatou-se que houve distinção dos grupos em orgânico e comum, obtendo 65% de confiança. A validação feita pelo índice da silhueta, que apresentou $S(i)$ de 0,73, demonstrado que o agrupamento adotado apresenta força e critério de distinção adequados. Desse modo, o método de agrupamento *Fuzzy* foi o mais indicado para a construção de um modelo de classificação de amostras de azeite extra virgem, distinguindo seus diferentes modos de produção, orgânico e comum.

Palavras-chave: Metabonômica, Estatística da Silhueta, Agrupamento *Fuzzy, RMN de $^{1}H$*.

# Abstract

Metabomics is a strategy that is based on the identification of patterns of a particular biological problem, by obtaining spectroscopic / spectrometric data of a given biofluid, the use of statistics to extract this information contributes significantly to the achievement of group classification. Thus, the present work aimed at the use of the meta-monetary strategy, based on nuclear magnetic resonance spectra of hydrogen and multivariate statistical techniques of grouping (principal component analysis (PCA), *Fuzzy* grouping) of samples of extra virgin olive oil. Were used 40 samples of extra virgin olive oil for this study. From the spectral data matrix, we used the pre-processing normalization by summation, in the samples. From the PCA, 99.1% of the variance explained using two components only, it was not possible to observe natural clusters of the data. with the application of the Fuzzy grouping, it was verified that there was distinction of the groups in organic and common, obtaining 65% confidence. The validation made by the silhouette index, which presented s (i) of 0.73, demonstrating that the adopted grouping presents adequate strength and criteria of distinction. Thus, the fuzzy grouping method was the most indicated in the construction of a classification model of samples of extra virgin olive oil, distinguishing their different modes of production, organic and common.



Key words: Metabolic, Silhouette statistic, *Fuzzy* grouping, *$^{1}$H-NMR*

# Lista de Figuras

# Lista de Tabela

# Lista de Abreviaturas

ANVISA - Agência Nacional de Vigilância Sanitária

MAPA- Ministério de Agricultura Pecuária e Abastecimento

IOOC - *International Olive Oil Council*

PCA - do inglês, *Principal Component Analysis*

ppm- Parte por milhão

RF- Radiofrequência

RMN $^{1}H$ - Ressonância Magnética Nuclear de Hidrogênio

SI- do inglês, *Silhouette Index*

# Sumário

# 1- Introdução

O azeite de oliva é um óleo de origem vegetal, proveniente do fruto de uma espécie de árvore típica de climas mediterrâneos, a *Olea europaea* L., também conhecida como oliveira (MANDARINO et al., 2005). Esse óleo é conhecido e utilizado mundialmente das mais diversas formas pelos benefícios que proporciona a saúde. O crescente uso do azeite de oliva aumenta sua demanda no mercado e contribui para o reconhecimento de seu uso. Tornando-o dessa forma, necessária a criação de um acordo internacional para a comercialização do produto, regulado pelo International Olive Oil Council (IOOC, 2003).

No Brasil, a regulamentação do azeite é realizada pelo Ministério da Agricultura, Pecuária e Abastecimento (MAPA), via instruções normativas. Das quais, visam classificar o produto com base em requisitos de identidade e qualidade, ao mesmo tempo em que define a amostragem, o modo de apresentação e a marcação ou rotulagem das embalagens, de acordo com a classificação do produto (MAPA, 2012). Além destes, outro parâmetro comumente analisado é a classificação do azeite em comum e orgânico (BRASIL, 2005).

Essa classificação, é feita em laboratório com a adoção da técnica de espectroscopia de ressonância magnética nuclear (RMN). Que mesmo sendo bastante eficiente para elucidação estrutural da amostra analisada, gera diversas informações sobrepostas. Que de modo geral podem tornar a interpretação dos dados confusa e em alguns casos promoverem erros na classificação. O uso de ferramentas de estatística multivariada, com a aplicação de estratégias metabonômicas pode ser uma maneira de solucionar tal problema nas análises RMN. (JURS, 1986; LINDON et al., 2000; WOLD et al., 2001; WESTERHUIS et al., 2008; GOODPASTER et al., 2010; XU et al., 2012).

Pode-se observar diversos trabalhos que envolvem estudos sobre azeite de oliva. Em trabalho pioneiro envolvendo a espectroscopia de RMN de $^{1}H$ e técnicas de análise multivariada não supervisionadas para caracterização e distinção entre amostras de azeite extra virgem na região centro-sul da Itália, os autores relatam discreta separação entre os grupos estudados, demonstrando certa dificuldade em

distinguir a variedade dos azeites em função da variabilidade do clima, no qual as oliveiras foram submetidas (SACCHI et al., 1998). Os estudos realizados por Vigli et al. (2003) trazem a possibilidade de combinar a espectroscopia RMN de $^{1}H$ e $^{31}P$ com análise multivariada para a distinção entre amostras pertencentes a 13 tipos de óleos vegetais na região da Grécia.

As técnicas de estatística multivariadas possuem grande variabilidade, o pesquisador torna-se o precursor para delimitar qual seu objetivo e qual a natureza de seus dados, nesse sentido a escolha da estatística multivariada corrobora com todas as características dos dados. Pois existem técnicas de agrupamento e de classificação, cada uma com uma função distinta. A análise por Componentes Principais é uma técnica de classificação que utiliza combinações lineares das variáveis originais e derivadas, em ordem decrescente de importância (JOHNSON; WICHERN, 1988). A transformação entre conjuntos de p variável não-correlacionada pode ser útil, pois esse novo conjunto vem a ter propriedades especiais em temos de variâncias.

As técnicas e os métodos multivariados, utilizam simultaneamente as informações de todas as variáveis respostas existentes na interpretação de uma base de dados, considerando as correlações existentes entre elas (FERREIRA, 2008). Nesse contexto, a aplicação de métodos como este podem melhorar a qualidade de pesquisas quantitativas, tornando a interpretação dos dados mais compreensível e minimizando a possibilidade de classificação errada de componentes.

Desse modo, propõe-se com o presente trabalho utilizar métodos estatísticos multivariados para classificação do azeite de oliva extra virgem, em comum ou orgânico.

# 2- Objetivos

## 2.1 Objetivo Geral

Aplicar modelos estatísticos multivariados, Agrupamento *Fuzzy* e Análise de Componentes Principais, para classificação de azeite de oliva extra virgem.

## 2.2 Objetivos Específicos

- Verificar se a técnica multivariada não hierárquica de agrupamento Fuzzy contribui na diferenciação entre azeite extra virgem orgânico e comum de diversas marcas comercializadas na Região metropolitana do Recife-PE;
- Analisar o comportamento dos dados por meio da técnica multivariada de PCA;
- Utilizar a estatística da Silhueta para verificar a qualidade do agrupamento Fuzzy para as amostras adotadas;

# 3- Fundamentação Teórica

## 3.1 Azeite de Oliva

O cultivo da *Olea europaea* L iniciou-se a cerca de quatro mil anos a.C. no norte do Mar Morto, expandindo-se para o Ocidente pelo Mediterrâneo e atualmente é cultivada em praticamente todos os continentes (RALLO et al., 2005). Trata-se de uma arvore de porte médio, sistema radicular caracterizado por uma raiz pivotante central em plantas originadas de sementes, e fasciculado para aquelas originadas de estacas. As folhas adultas são simples, elípticas ou lanceolada, com comprimento variando entre 5 e 7 cm e largura de 1,0 a 1,5 cm, sendo a região ventral de cor verde-escura e a região dorsal de cor esbranquiçada. Apresenta resistência a secas devido à presença tricomas ou placas foliares. A flor é constituída por quatro sépalas verdes e por quatro pétalas brancas, que formam a corola. O fruto, denominado azeitona, é uma drupa de tamanho pequeno e forma elipsoidal, cujas dimensões pode apresentar entre 1 a 4 cm de comprimento e diâmetro de 0,6 a 2 cm com acúmulo de azeite nas células do mesocarpo (RAPOPORT, 1998).

Os óleos vegetais são extraídos de grãos ou sementes e geralmente utilizam-se solventes durante a extração (Baccouri, et al., 2008). O azeite de oliva recebe essa classificação por ser oriundo do mesocarpo (parte mais suculenta do fruto) da azeitona. Vários tipos de azeite de oliva são encontrados e a seguinte classificação é feita pela ANVISA na resolução RDC nº 482, de 23/09/1999, com alguns valores atualizados na resolução RDC nº 270, de 22/09/2005 (ANVISA, 2005).

Em países da Comunidade Econômica Europeia localizada na região mediterrânea a produção de azeite de oliva corresponde a 79,8% em todo o mundo, destacando-a como maior produtora, sendo 42,9% obtidos na Espanha, 17,5% na Itália e 12,2% na Grécia, além de responder por quase 80% das exportações mundiais (MESQUITA et al., 2006).

O azeite de oliva representa um pequeno volume em termos de produção mundial (aproximadamente 2% do total de óleos produzidos), mesmo com a pequena porcentagem na produção, contribui com cerca de 15% do valor monetário da produção dos óleos (AUED; PIMENTEL et al., 2008). Dentre os diferentes tipos de alimentos, o azeite de oliva é considerado como a opção mais saudável entre os azeites comestíveis (EMBRAPA, 2011).

O Brasil é considerado um dos maiores importadores mundiais de azeitonas e derivados e o cultivo de oliveiras no país é uma atividade agrícola recente e em expansão (OLIVEIRA et al., 2009a). Em 2009, foram importadas aproximadamente 44 mil toneladas de azeite e 70 mil toneladas de azeitonas em conservas, movimentando mais de 1 bilhão de reais, no mercado nacional, com esses produtos (OLIVEIRA et al., 2012b).

Na Figura 1, destacamos as possíveis regiões no Brasil propicias ao plantio de oliveira.


**Figura 1**- Zonas com potencial para plantios de oliveiras na América do Sul.
Fonte: WREGE et al, 2015.

No Brasil, há regiões com grande potencial para o plantio de oliveiras, inclusive áreas localizadas na região nordeste do país. Com o intuito de viabilizar a expansão da plantação de oliveira no país foram introduzidas plantações em Minas Gerais e no Rio Grande do Sul (OLIVEIRA et al., 2010b; 2012a). Nesse sentido, recomenda- se

que os produtores recorram a informações técnicas para que ocorra a produção de frutos (VIEIRA NETO et al., 2008).

### 3.1.1 Classificação do Azeite de Oliva

Entre os constituintes químicos presentes na composição do azeite de oliva extra virgem, destacam-se: destacam-se: grau de acidez menor que 0,8%, tocoferóis, ou vitamina E (95% deste do tipo alfa- tocoferol), ácidos graxos monoinsaturados (55% a 88%), baixo conteúdo de ácidos graxos poliinsaturados (2% a 21%) e a presença de grande quantidade de compostos fenólicos antioxidantes que podem inibir a produção de hidroperóxidos (TENA et al., 2009).

Com uma quantidade minoritária, podem-se encontrar: flavonóides, rutina, luteonina e esqualeno. O teor dos compostos fenólicos no azeite de oliva é influenciado pela maturidade dos frutos no momento da colheita e pela forma de cultura utilizada (BECKER, 2004). Nesse contexto, o grau de acidez é um parâmetro físico-químico que classifica o azeite de oliva, pois o ácido oleico está intimamente relacionado com a natureza e a qualidade da matéria-prima.

A denominação “extra virgem” dos azeites se reservam àqueles obtidos a partir do fruto unicamente por procedimentos físicos, em condições sobre tudo térmicas, que não ocasionem à alteração do azeite e que não tenham sofrido tratamento algum, exceto a lavagem, a decantação, a centrifugação e a filtração (ARDOY, 2004). O azeite extra virgem deverá ter acidez, expressa em ácido oleico, de no máximo 0,8% e o virgem deverá ter sua acidez variando de 0,81% a 2%. (GONÇALVES, 2015)

O azeite “refinado” é obtido do refino do azeite virgem e tem no máximo 0,3 g/100 g em ácido oleico. Enquanto o azeite de oliva “puro” é composto da mistura de azeite refinado com azeite virgem e terá no máximo, 1,0 g/100 g em ácido oleico (ANVISA, 2005).

## 3.2 Análise Multivariada

O estudo investigativo surge da necessidade de novas descobertas, que são obtidas de pesquisas em várias áreas, de variáveis que são mensuradas em geral de maneira conjunta. Uma das ferramentas utilizadas em análise estatística de variáveis conjuntas é a estatística multivariada (ALBUQUERQUE, 2013). Os métodos

multivariados são um conjunto de técnicas que permitem ao investigador interpretar grandes conjuntos de dados, que podem ser referentes a indivíduos ou variáveis. Esses métodos buscam encontrar relações entre variáveis, entre indivíduos ou entre ambos (DOCAMPO et al., 2013).

### 3.2.1 Análise Componentes Principais (PCA)

Análise de componentes principais – PCA (do Inglês, Principal Componente Analysis) foi introduzida por Pearson em 1901 e desenvolvida de forma independentemente por Hotelling em 1933. É um método da estatística multivariada que tem por finalidade identificar a relação entre características extraídas dos dados visando sua redução, eliminação de sobreposições e a escolha das formas mais relevantes entre eles a partir de combinações lineares das variáveis originais (MOITA-NETO, 2009). Também denominada por Transformada de Hotelling, a PCA transforma variáveis discretas em coeficientes descorrelacionados através de uma transformação linear aplicada nos dados, de modo que os dados resultantes tenham suas componentes mais relevantes nas primeiras dimensões, denominadas de componentes principais (LAY, 2007).

A utilização dos componentes principais para definir um espaço de fatores que englobe os dados, não modifica os dados, apenas encontra um sistema de coordenadas mais conveniente, capaz de remover ruídos dos dados sem distorcê-los e de diminuir sua dimensionalidade e sem comprometer o conteúdo de informações. (NETO E MOITA, 1998). Em suma, a PCA tenta encontrar, simultaneamente no espaço dimensional transformado a direção ao longo do qual os pontos se encontrem espalhados com variabilidade máxima preservando a informação dos dados originais. (SILVA et al., 2005).

Num ponto de vista prático, isto é feito através dos seguintes passos: Obtenção de uma matriz que represente o conjunto de dados; Centralização dos dados em torno da média; Cálculo da matriz de covariância; Cálculo dos autovalores e autovetores; diagonalização da matriz de covariância (MORAIS, 2016).

O método PCA permite a eliminação da covariância entre as coordenadas de um vetor de variáveis aleatórias por meio de uma mudança de base. As bases formadas pelos auto-vetores da matriz de covariância permitem a eliminação da

covariância entre as coordenadas do vetor de entrada (LUDWIG JR. E MONTGOMERY, 2007).

O fluxo de informação dos passos descritos para obtenção das componentes principais é ilustrado na Figura 2.

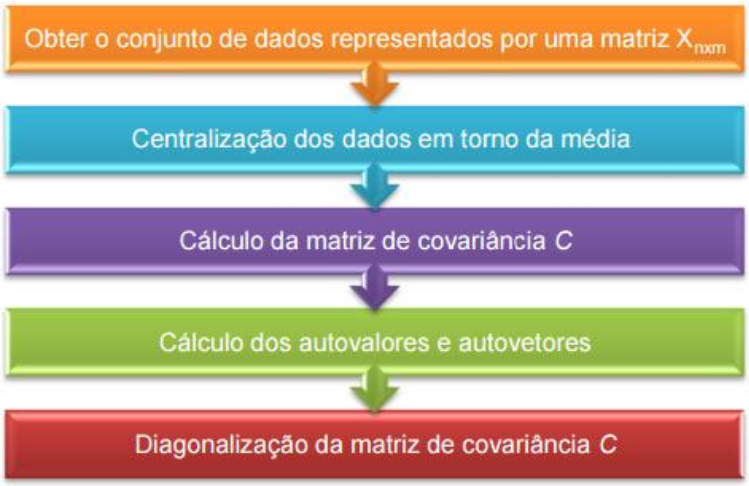


**Figura 2**- Fluxograma da obtenção das CPs em torno da média. Fonte: Morais, 2016.

**Primeiro passo:** os dados são organizados em uma matriz $X_{nxm}$, onde n representa o número de observações e m o número de variáveis independentes.
**Segundo passo:** Calcular a Média ou o Vetor Médio dos dados.

$$m_x = \frac{\sum_{i=1}^{M} xi}{M}, \tag{1}$$

Em que:

$m_x$ é o vetor médio;
$x_i$ são as amostras para $i = 1, 2, \ldots n$;
$\sum_{i=1}^{M} xi$ é média amostral da variável aleatória $x_i$;
M amostras de vetores em um conjunto qualquer.

**Terceiro passo:** consiste no cálculo da matriz de covariância $C_z$:

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \mu).\ (Y_i - \overline{Y})}{n} \tag{2}$$

Em que:

$X\ e\ Y$ são listas de dados, em que $X$, é a primeira e $Y$ é a segunda dimensão.

$\bar{Y}$ é a média das listas $Y$.

μ é a média das listas $X$.

$X_i$ e $Y_i$ são cada um dos elementos das listas nas duas direções $X$ e $Y$, na i-ésima posição.

A variável $n$ representa o número de itens de dados obtidos.

Quando os dados representam uma amostra (que inicia no índice 0), usa-se Qíno denominador e no somatório. Quando os dados representam o conjunto total da "população", usa-se simplesmente Q no denominador.

Se os dados tiverem mais de duas dimensões, é necessário ter a covariância entre cada par de dimensões. A partir dessa ideia, surge a matriz de covariância. A diagonal principal da matriz contém as variâncias e as demais posições a correlação entre as direções. Essa matriz é simétrica e real, de modo que é sempre possível encontrar um conjunto de autovetores ortonormais (ANTON et al.,2004).

**Quarto passo:** determinação dos autovalores λ e autovetores $V_n$ correspondentes da matriz $C_z$. Os autovetores são arranjados de modo decrescentes de acordo com os valores dos autovalores. Os autovetores $V_n$, formarão as colunas de uma matriz $P$:

$$P = \{v_1, v_2, \dots, v_n\} \tag{3}$$

**Quinto passo:** é a diagonalização. A matriz $P$ é empregada para mudar a base de $C_z$ obtendo uma matriz diagonal $D$ de autovalores de $C_z$.

$$D = P^{-1} C_z P \tag{4}$$

A matriz $D$ apresenta elementos iguais aos autovalores na diagonal principal, ou seja, não apresenta covariância, consequentemente não tem nenhuma informação redundante (LUDWIG JR. E MONTGOMERY, 2007).

### 3.2.2 Análise de Agrupamento

Análise de Agrupamento ou Análise de Clusters, é uma técnica criada a mais de setenta anos, que consiste na classificação de objetos em diferentes grupos, sendo que cada um dos grupos deve conter os objetos semelhantes, segundo alguma função de distância estatística (JOHNSON e WICHERN, 1998).

Quando se tem apenas duas variáveis de interesse (p = 2), um diagrama de dispersão entre elas permitirá uma visualização de possíveis agrupamentos entre os indivíduos. Quando a proporção de variáveis explicadas pelas primeiras componentes for significativa, isto é, aproximadamente 80% ou mais, resultar em um diagrama de dispersão das primeiras componentes principais para visualizar a existência de possíveis grupos (PICARD et al,. 2010).

O fluxo para uma análise de *clusters* compreende cinco etapas é ilustrado de forma resumida na Figura 3:

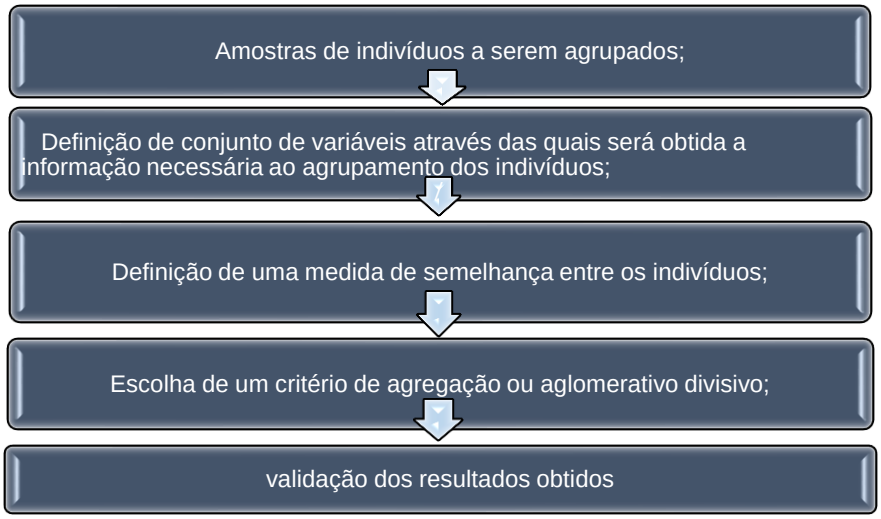


**Figura 3** - Fluxograma da análise de *cluster*

### 3.2.3 Agrupamento *Fuzzy*

Os conceitos de conjuntos *Fuzzy*, foram orginalmente propostos por Zadeh (1965) e modificados por Takagi e Sugeno (1985), em que os modelos são capazes de representar o comportamento de sistemas não-lineares graças a propriedade de serem aproximadores universais de funções em um espaço compacto. (WANG et al, 1992; KOSKO, 1994; WANG, 1998).

No agrupamento *Fuzzy*, cada objeto pertence a mais de um grupo com diferentes graus de pertinência, (lógica *Fuzzy),* em vez de pertencer apenas a um grupo. Um determinado objeto perto do centro de um grupo pertence a esse grupo com um grau mais elevado do que um objeto que está situado na extremidade desse grupo. Para cada objeto *Z*, o grau de pertinência descreve o quão forte esse objeto pertence a um determinado grupo (ZALIK, 2010).

O objetivo mais claro do agrupamento de *Fuzzy* é tratar os fenômenos naturais ou as situações reais, de modo que suas características mantenham- se originais, para que se obtenha uma modelagem mais próxima da realidade do fenômeno. Segundo Klir e Yuan (1995), a possibilidade de estudar a incerteza tende a reduzir a complexidade dos modelos e a aumentar sua credibilidade.

O pertencimento de uma dada amostra em um determinado grupo se dá pela função característica que a define, no caso de conjuntos crisp, atribuí-se os valores 0 ou 1 a cada elemento de um universo de discurso, de maneira a discrimina-lo como pertencente ou não ao grupo em questão. A função f que define um conjunto crisp Z fica do tipo:

$$f_Z = X \rightarrow \{0, 1\} \tag{5}$$

Em que:

$X$ é o universo de discurso (um conjunto de elementos);

$\{0,1\}$ define um conjunto de dois estados: 0 - não pertence; 1 - pertence.

Define-se a função de pertinência $\mu_Z$ de um conjunto *Fuzzy* Z como sendo do tipo:

$$\mu_Z = X \rightarrow [0, 1] \tag{6}$$

Em que:

$X$ é o universo de discurso (um conjunto de elementos crisp);

$[0,1]$ define um intervalo infinito de estados cujas extremidades significam a não pertinência de um elemento a um conjunto (0) e a pertinência total de um elemento a um conjunto (1).

O agrupamento *Fuzzy* leva vantagem em relação a outros métodos por particionamento, por fornecer informações mais detalhadas sobre a estrutura dos dados, pois apresenta os graus de associação de cada elemento a cada grupo e, consequentemente, não tem uma alocação clara de elementos para formar grupos (KAUFMAN, 1990). Desse modo, a utilização do agrupamento *Fuzzy* é pertinente em alguns estudos, no estudo realizado por Oliveira et al. (2016) verificou-se que o método de agrupamento *Fuzzy* alocou cem cisternas localizadas no sertão do Pajeú para o grupo ao qual a análise obteve a mais elevada pertinência. Assim, as determinações de grupos foram eficazes na análise de agrupamento de cisternas de placas da região do Pajeú. Neste trabalho verificamos que a análise de agrupamento *Fuzzy* conseguiu, de maneira eficaz, distinguir os grupos de azeite em orgânico e comum de maneira notória.

## 3.3 Estratégia Metabonômica

Existe uma área de estudo multifacetada, que procura identificar mudanças no perfil de metabólitos endógenos em biofluidos, e associa-las ao status bioquímico da fonte, essa área é denominada metabonômica. Para Nicholson et al. (1999) a metabonônica é a representação quantitativa da resposta metabólica e multiparamétrica. Que depende do tempo, estímulos fisiopatológicos ou modificações genéticas.

Os estudos metabonômicos envolvem geralmente dados que buscam identificar mudanças no perfil metabólitos endógenos em um dado biofluido, associando- a aos status bioquímico da fonte (NICHOLSON et al., 1989). A estratégia metabonômica utiliza biofluidos, que podem ser obtidos de forma não invasiva (urina) ou pouco invasiva (soro ou plasma) (ROBERTSON et al., 2000; GRIFFITHS et al., 2010). Além destes, existem estudos que mostram a utilização de biofluido não convencionais, como fluido cérebro-espinhal, bile, fluido seminal e biofluido de origem vegetal (SCHRIPSEMA, 2010).

Na busca por perfis metabonômicos, foram realizadas várias propostas de técnicas analíticas, sendo as mais frequentes a espectroscopia de RMN e a cromatografia, seja em fase líquida (HPLC) ou em fase gasosa (CG), acoplada à espectrometria de massas (EM). (NICHOLSON; WILSON, 1989; LI et al., 2011; WU;

ZHU; WANG, 2011; SMOLINSKA et al., 2012). Apesar de serem técnicas bastante eficientes para elucidação estrutural, RMN e MS geram diversas informações sobrepostas, o que pode tornar a interpretação desses dados uma ação difícil e confusa. Para solucionar este problema, a estratégia metabonômica faz uso de ferramentas de análise multivariada. Esse campo do conhecimento é denominado Quimiometria. (JURS, 1986; GOODPASTER et al., 2010; HENDRIKS et al., 2011 LINDON; NICHOLSON; WILSON, 2000; WESTERHUIS et al., 2008; WOLD; SJÖSTRÖM; ERIKSSON, 2001a; XU et al., 2012).

Com os avanços tecnológicos, torna-se mais comum a obtenção de dados por meio de experimentos elaborados em equipamentos sofisticados, gerando- se uma gama de dados. Nesse sentindo, tornou-se necessários estudos para verificar com objetividade essa gama de dados gerados por novos equeipamentos. Assim surgiu o ramo da Quimiometria, que analisa dados de multivariadas na área da Química. As ferramentas quimiométricas são veículos que podem auxiliar os químicos a caminharem mais eficientemente na direção do maior conhecimento (KOWALSKI E SEASHOLTZ, 1999).

A análise covariante ou Quimiometria ou análise multivariada consiste no uso de programas estatísticos em resultados de análises químicas (Patente Brasileira PI09059768 de 2009) visando um ou mais dos seguintes objetivos: análise exploratória, classificação dos dados, calibração multivariada, planejamento e otimização de um experimento.

### 3.3.1 Ressonância Magnética Nuclear (RMN)

A espectroscopia de RMN é uma técnica de análise estrutural, em termos moleculares, que pode ser medida através de interações associadas a radiofrequências oscilantes de pequenos campos eletromagnéticos vinculados a núcleos interagentes, imersos em um forte campo magnético externo (GRUTZNER, 2005; JUCHEM et al., 2014).

A espectroscopia de Ressonância Magnética Nuclear de Hidrogênio (RMN de $^{1}H$) mostra-se como uma ferramenta enérgica para a realização da tática metabonômica, pois a obtenção dos espectros é relativamente rápida e a preparação exige mínimas intervenções do analista, dando agilidade ao processo e minimizando

possíveis contaminações. A RMN de $^{1}H$ também possibilita a identificação e a quantificação relativa de diferentes metabólitos presentes no biofluido analisado, no entanto, em relação à espectrometria de massas, é uma técnica menos sensível (LINDON et al, 2001; PAN et al., 2007; GODOY et al., 2010).

# 4- Material e Métodos

## 4.1 Caracterização dos Dados

Este estudo foi realizado pelo Departamento de Química Fundamental (DQF) da UFPE, que analisaram 40 amostras de azeite de oliva extra virgem. Os dados foram obtidos por (SILVA, 2017), os cedendo para utilização no presente trabalho. As amostras avaliadas de azeite de oliva extra virgem, foram dissolvidas 60 µ L da amostra em 640 µ L de CDCl3, em um tubo de RMN de 5 mm de diâmetro. Os espectros de RMN foram obtidos utilizando o espectrômetro VNMRS400, operando a 399,99 MHz, para o núcleo de $^{1}H$, com janela espectral 6,4 kHz, tempo de espera (d1) igual a 1 s, tempo de aquisição igual a 2,556 s, pulso de radiofrequência (RF) de 90º, 64 repetições e temperatura de 26ºC. Os espectros foram processados com *line broadening* igual a 0,3 Hz. Após a análise espectroscópica, as amostras foram armazenadas em um recipiente limpo de âmbar, para futuro descarte adequado. Os espectros de RMN de $^{1}H$ tem suas fases ajustadas, correção de linha de base e construção dos bins de forma manual utilizando o software Mestre Nova 9.0. Os bins foram definidos em intervalos de 0,03 ppm entre os δ 0,00 e 6,80 ppm. Os dados foram dispostos numa matriz para tratamento quimiométrico.

## 4.2 Pré-processamento dos Dados

Os dados foram dispostos em uma matriz contendo as informações relativas a natureza da amostra, dados espectrais em formato de bins e informações relacionadas a classe à qual pertence uma dada amostra. As técnicas de pré-processamento utilizadas neste trabalho foram: normalização pela soma, sendo construídos modelos para cada tipo de pré-processamento.

A normalização pela soma foi obtida pela divisão de cada bin pela respectiva soma da área de integração total de cada amostra (Eq. 7). A proposta deste pré-processamento é obter dados que possam ser comparados entre si, sem alterar a informação contida nas variáveis.

$$A_{ij}^{ns} = \frac{A_{ij}}{\sum_{l}^{j} A_{ij}} \quad (7)$$

Em que:

$A_{ij}^{NS}$ = bin normalizado pela soma.

$A_{ij}$ = bin original.

$\sum_{j}^{i} A_{ij}$ = soma das áreas de integração para cada amostra.

## 4.3 Métodos

### 4.3.1 Análise Componentes Principais

Como princípio para o cálculo do PCA, considere um vetor aleatório $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{p})$, contendo $p$ componentes, com um vetor de médias $\mu = \mathrm{E}(\mathrm{X}) = (\mu_1, \mu_2, \ldots, \mu_\mathrm{p})$. A matriz de Covariâncias do vetor aleatório $X$, quadrada de dimensão p, é denotado por: $Cov(X) = \Sigma_{p\times p}$. A matriz de covariância é uma matriz simétrica, não negativa, ou seja, $a^t \Sigma_a > 0$ para todo vetor de constantes $a \in R^p$. Esta condição implica que os autovalores da matriz $\Sigma_{p\times p}$ denotados por $\lambda_1, \lambda_2, \ldots, \lambda_p$, são não-negativos, ou seja, $\lambda_i \geq 0$, para qualquer $i = 1,2, \ldots, p$ (GRAYBILL, 1983). Pelo teorema da Decomposição Espectral [LAY 2007], sendo $\Sigma_{p\times p}$ uma matriz de covariância, existe uma matriz ortogonal $\mathrm{P}_{p\times p}$, isto é, $\mathrm{P}^T\mathrm{P} = \mathrm{P}\mathrm{P}^T = I$, tal que:

$$\mathrm{P}^T \sum \mathrm{P} = \theta \quad (8)$$

Em que $\lambda_1, \lambda_2, \ldots, \lambda_p$, são os autovalores da matriz $\Sigma_{p\times p}$ ordenados em ordem decrescente. Nesse caso, dizemos que a matriz $\Sigma_{p\times p}$ é similar a matriz $\theta$.

A i-ésima coluna da matriz $\theta$ é o auto vetor normalizado e $i$ correspondente ao auto vetor $\lambda_i$ , com $i = 1,2, \ldots, p$; que é denotado por $e_i = (e_1, e_2, \ldots, e_p)^T$. Então a matriz $\theta$ é dada por $\theta = [e_1, e_2, \ldots, e_p]$ e pelo teorema da decomposição espectral tem-se a seguinte igualdade valida:

$$\Sigma_{p\times p} = \sum_{i=1}^{p} \lambda e_i e_i^T = \mathrm{P}\theta\mathrm{P}^T \quad (9)$$

Como $\theta_1, \theta_2, \ldots, \theta_p$, formam uma base de $R^p$, o vetor $a$ pode ser escrito como $\sum_{i=1}^{p} \alpha_i \mathrm{P}_i = \alpha^T \mathrm{P}$ para algum $\alpha_i = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$.

Sendo $\theta$ ortogonal, $\alpha^T \alpha = 1$ e a variância de a $\alpha^T X$ é menor ou igual a $\lambda_1$ e tomando $\alpha = \mathrm{O}_1$, tem-se que $var(\mathrm{P}_1 X) = \mathrm{P}_1 \Sigma \mathrm{P}_1 = \lambda_1$, e define-se a variável aleatória $U_1 = \mathrm{P}_1^T X$ como o primeiro componente principal de $X$. Para a obtenção de outros componentes principais é feita uma restrição de não correlação do próximo componente $U_i$ com os componentes anteriormente obtidos $(U_1, \ldots, U_{i-1})$. Desta forma as componentes são definidas como vetores aleatórios $U = (U_1, \ldots, U_p) = \mathrm{P}^T$, onde as colunas de $P$ são os auto vetores de $\Sigma$. É importante ressaltar que a matriz de covariância da nova matriz $U$ é diagonal, onde os elementos são os autovalores $\lambda_i$.

Na aplicação de redução de dimensionalidade, o PCA tem a propriedade de minimizar o erro quadrático médio entre os dados reconstruídos e os dados originais. Supõe-se, por exemplo, que se tem dados de entrada $X$ de dimensionalidade $m$ e dados de saída $Y$ de dimensionalidade $m$, em que $m_1 < m$.

### 4.3.2 Agrupamento não Hierárquico

#### 4.3.2.1 Agrupamento *Fuzzy*

A linha de desenvolvimento de agrupamento torna-se fundamental quando se deseja classificar um conjunto de dados de acordo com suas características ou variáveis mensuradas. Nesse intuito, o termo classe é pertinente, dada a informação de quantas partições e quais são essas partições em um conjunto de dados, bem como cada observação ou grupo de azeites pertence tais amostras. Desse modo, *classificação* é denominada sendo a análise realizada em determinados bancos de dados. O trabalho de análise de dados é denominado agrupamento e tem por objetivo estudar as relações de similaridade entre os dados ou amostras de azeite, determinando quais dados formam quais grupos. Os grupos são formados de maneira que se maximize a similaridade entre as amostras de um grupo (similaridade intra-grupo) e se minimize a similaridade entre amostras de grupos diferentes (similaridade inter-grupos). Então, formalmente, dado um conjunto de dados de entrada ($\vec{X} \in \mathbb{R}^p$), é encontrada uma função:

$$\mathrm{f}\colon \mathbb{R}^{p} \times \mathrm{W} \longrightarrow \mathrm{G} \tag{10}$$

Em que $W$ é um vetor de parâmetros ajustáveis, por meio de um algoritmo de aprendizado supervisionado ou não supervisionado, que determina c-grupos a partir da matriz de dados originais $X$, e, segundo Xu e Wunsch (2005), tem-se $\mathrm{G} = \mathrm{G}_{1}, \mathrm{G}_{2}, \ldots, \mathrm{G}_{c}$ (c ≤ n), tal que:

i) $\mathrm{G}_{i} \neq \emptyset, \mathrm{i} = 1, \ldots, \mathrm{c}$;

ii) $\bigcup_{i=1}^{c} \mathrm{G}_{i} = \mathrm{X}$;

iii) $\mathrm{G}_{i} \cap \mathrm{G}_{j} = \emptyset, \mathrm{i}, \mathrm{j} = 1, \ldots, \mathrm{c}$ e $\mathrm{i} \neq \mathrm{j}$ , supondo a abordagem clássica de classificação ou agrupamentos.

### 4.3.3 Distância Euclidiana

Segundo (Hair et al., 2010), o coeficiente de dissimilaridade mais conhecido e utilizado para indicar a proximidade entre objetos é o coeficiente da distância Euclidiana. É simplesmente a distância geométrica entre dois pontos em um espaço multidimensional. A distância entre duas observações (r e h) corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações ou pontos de r e h para todas as p variáveis (MCROBERTS et al., 2007).

$$\begin{aligned} d_{rh} &= \|x_r - x_h\| \\ &= \sqrt{(x_r} - x_h)^t (x_r - x_h) \\ &= \sqrt{\sum_{k=1}^{p} (x_{rk} - x_{hk})^2} \end{aligned} \tag{11}$$

Em que:

$x_r$ = é o vetor da r-ésima observação;
$x_h$ = é o vetor da h- ésima observação.

## 4.4 Validação do Método

### 4.4.1 Índice da Silhueta

O índice da silhueta foi proposto por Rousseeuw (1987), com o intuito de avaliar métodos de particionamento. Nesse caso, cada objeto (amostra de azeite) é representado por um valor $s(i)$ chamado de *silhueta*, que é baseado na comparação da homogeneidade e na "separação" de cada grupo. Com isso, para um objeto $i$, o valor da silhueta é dado por:

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \; onde -1 \leq \mathrm{s(i)} \leq 1 \qquad (12)$$

Em que:

$a(i)$ é a distância média do objeto i aos objetos do seu grupo;
$b(i)$ é a distância média do objeto i aos objetos dos outros grupos.

Valores negativos de $s(i)$ negativos sugerem que o indivíduo $i$ seja semelhante a indivíduos de outras classes. Valores de $s(i)$ na vizinhança de 1 dão indícios de que $i$ esteja bem classificado.

# 5- Resultados e Discussão

## 5.1 Azeite de Oliva Extra Virgem

A aplicação da espectroscopia de RMN de $^{1}H$ nas análises dos dados obtidos para amostras de azeite (comum e orgânico) foram sobrepostas (Figura 4). Os perfis das amostras de azeite de oliva não diferem quanto a presença ou ausência de algum sinal característico, por conta disso necessitaram de um pré-processamento que elimine possíveis efeitos causados pela diluição das amostras. Para tanto adotamos a técnica de normalização pela soma realizados na linha.

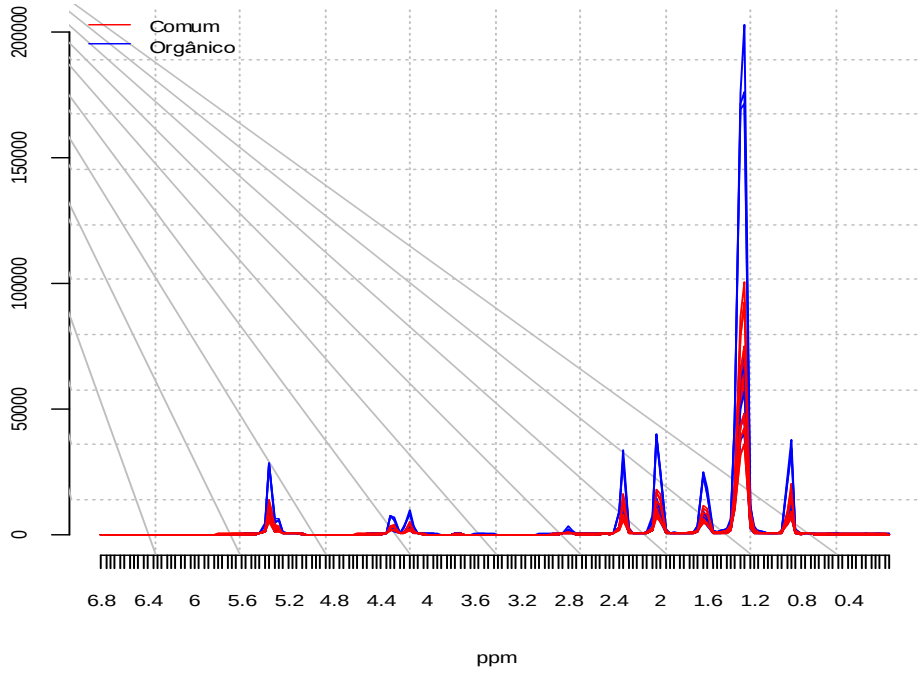


**Figura 4** – Espectros obtidos da espectroscopia de RMN de $^{1}H$ (400 MHz) para amostras de azeite, comum e orgânico.

### 5.1.1. Dados normalizados pela soma

Após a obtenção dos bins auto escalonados na linha, os espectros foram sobrepostos (Figura 5). É possível verificar o efeito do pré-processamento sobre a classificação das amostras, pois nota-se a presença de pontos característicos que diferenciam as amostras processadas.

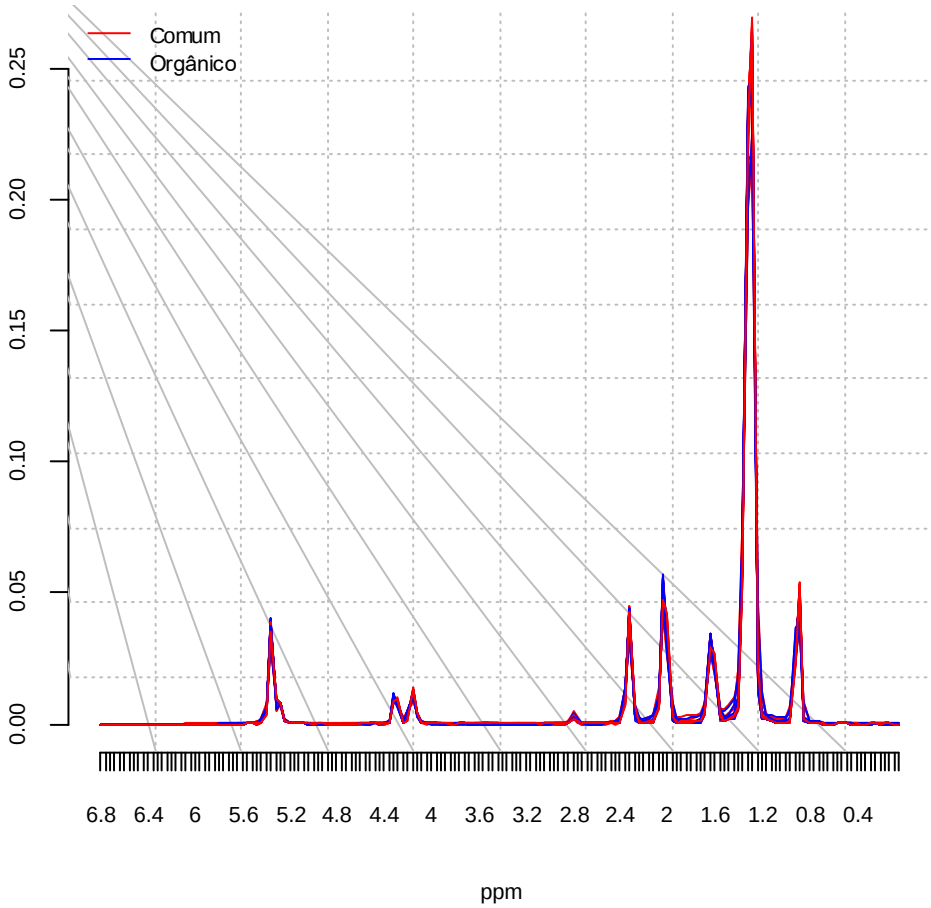


**Figura 5** – Dados normalizados pela soma.

Observa- se que o auto escalonamento dos dados contribuiu para adicionar pesos equivalentes aos pontos espectrais das amostras, possibilitando classifica-los, posteriormente, por meio de métodos estatístico.

## 5.2 Análise por componentes principais (PCA)

Após o pré-processamento dos dados, observamos que os dois componentes principais PC1 e PC2 mostraram que é possível descrever 99,9% da variância dos dados, e PC1 contribui com cerca de 99,75% dessa variância encontrada. De acordo estudos realizados por Meira et al. (2011) com biofluídos, foram necessários três componentes principais (PC) para explicar o comportamento de seus dados, bem como entender a contribuição de cada substância na estrutura final do produto. A mesma autora verificou que três componentes principais foram responsáveis por 95,39% da variância de seus dados, atribuindo 55,98% da variância ao PC1; 33,62% para PC2 e 5,79% para PC3. Nesse contexto, constatamos que os resultados obtidos pela primeira componente principal (PCA) para dados de azeite de oliva extra virgem foi suficientemente significativa para descrever o comportamento das análises. É possível verificar na Figura 6 o comportamento dos dados para as cinco primeiras componentes principais.

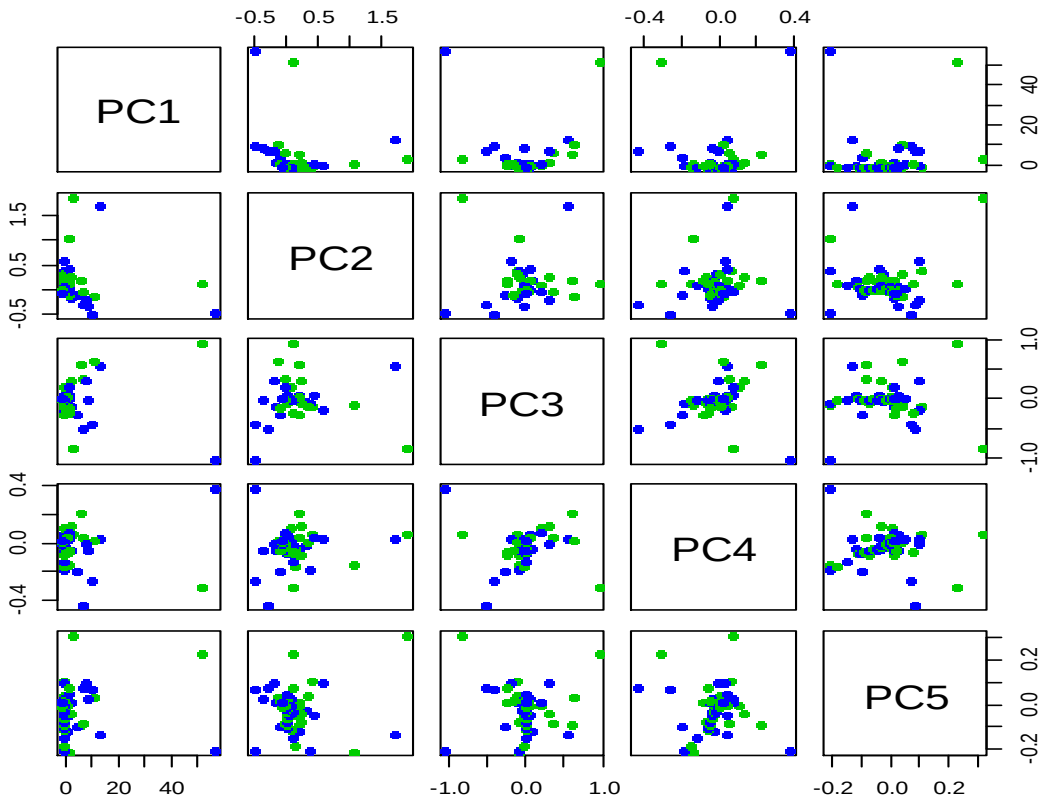


**Figura 6** - Panorama da Análise por Componentes Principais (PCA).

A porcentagem acumulada da variação total dos dois primeiros componentes (99,9%) explica satisfatoriamente a variabilidade nas amostras avaliadas. De acordo com Mardia et al. (1979), quando em uma Análise de Componentes Principais os dois ou três primeiros componentes acumularem uma porcentagem relativamente alta da variação total (em geral acima de 70%) eles explicarão satisfatoriamente a variabilidade manifestada entre as amostras avaliadas. Tem-se aqui que as duas primeiras componentes principais, apresentam um elevado poder de explicação entre os grupos estudados. Analisando o gráfico de escores (Figura 7) observamos que as porcentagens acumuladas da variância são explicadas pelas duas primeiras componentes principais, destacando que primeira Componente Principal possui alta significância.

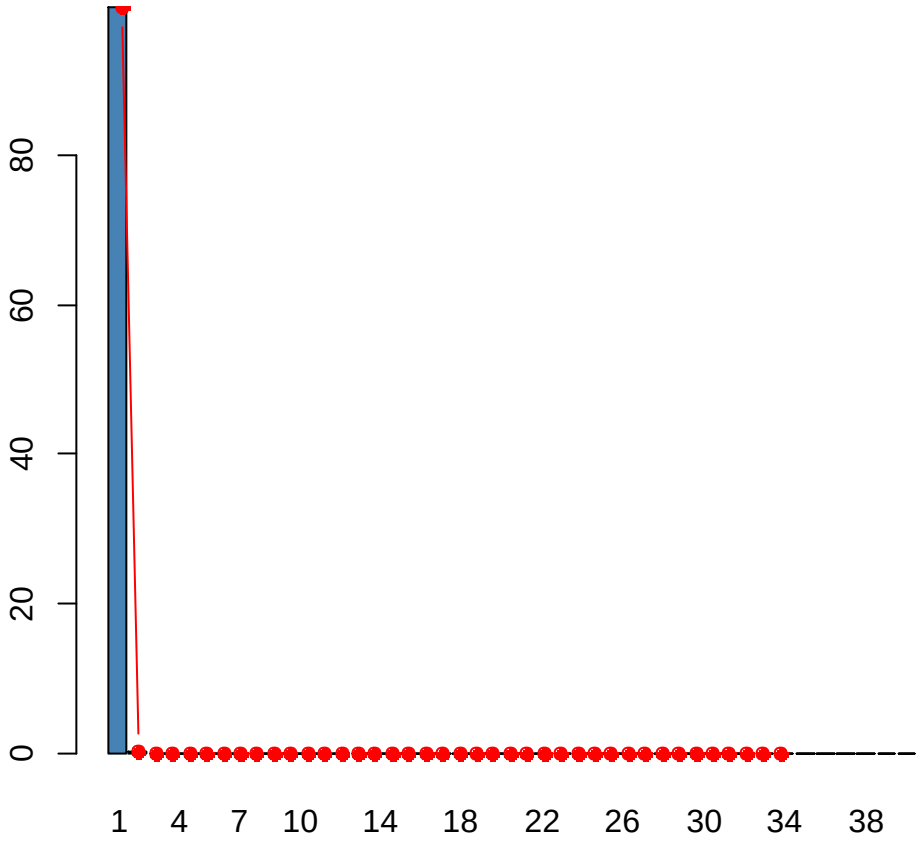


**Figura 7** - Gráfico de Porcentagem Acumulada das variâncias dos Componentes Principais (PC).

A Figura 7 apresenta uma representação bidimensional das variáveis dos azeites, que vulgarmente denomina-se “biplot”. As variáveis agrupam-se de acordo com os seus coeficientes de correlação, sendo que cada eixo principal corresponde a um conjunto de variáveis correlacionadas entre si. Como as correlações entre as variáveis resultam das medições efetuadas nos azeites, cada eixo principal representa uma direção do espaço ao longo da qual a variância (ou diferença) entre os azeites está maximizada. Observa-se no biplot (Figura 8) o gráfico da PC1 versus PC2, demonstrando que as amostras do azeite orgânico possuem um maior agrupamento entre os seus semelhantes do que os dados do azeite comum, representado, desse modo uma distinção bastante significativa entre algumas amostras.

Verifica-se ainda que não houve uma separação nítida entre os grupos, uma vez que amostras tanto do azeite orgânico quando do comum apresentam características semelhantes no mesmo ponto.

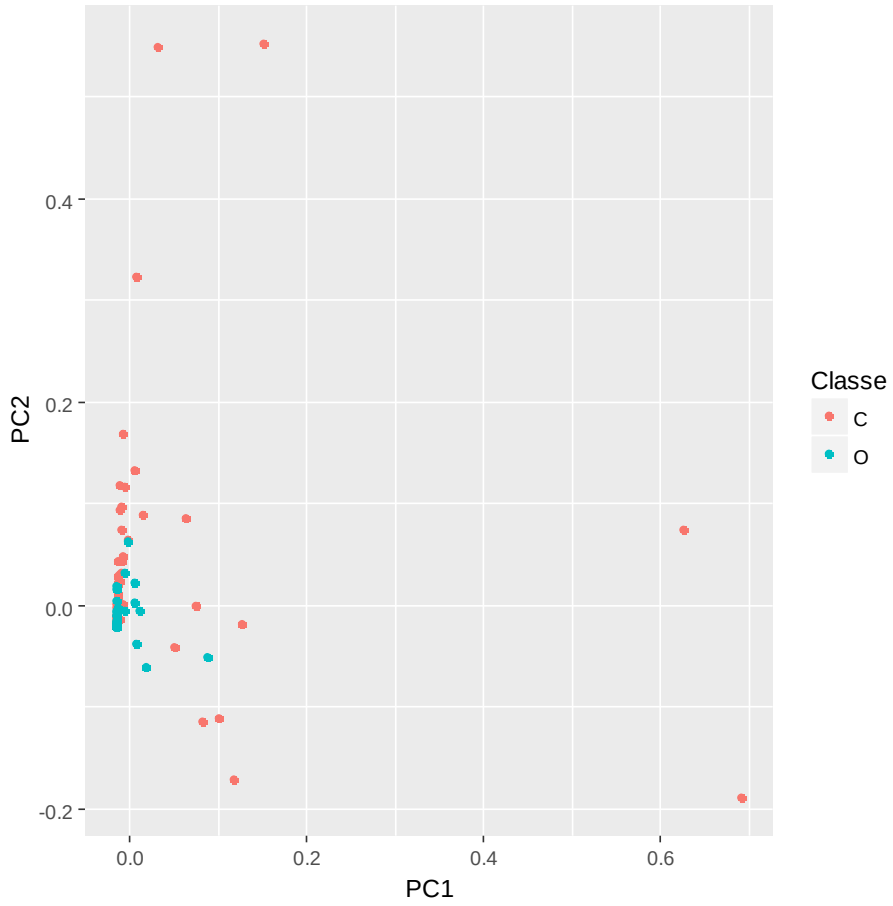


**Figura 8** - Gráfico da PC1 versus PC2 com 99,9% da variância explicada dos dados.

## 5.3 Agrupamento *Fuzzy*

Por meio da PCA constata-se que não houve uma separação nítida entre as amostras de azeite comum e orgânico, inviabilizando desta forma a utilização deste método para classificação deste tipo de produto. Este resultado pode ser explicado possivelmente pela semelhança dos compostos nos dois casos.

Por conta deste resultado, torna-se necessária a utilização da técnica de agrupamento *Fuzzy,* para verificarmos a separação entre os grupos das amostras de azeite de oliva, como mostra a Figura 9.

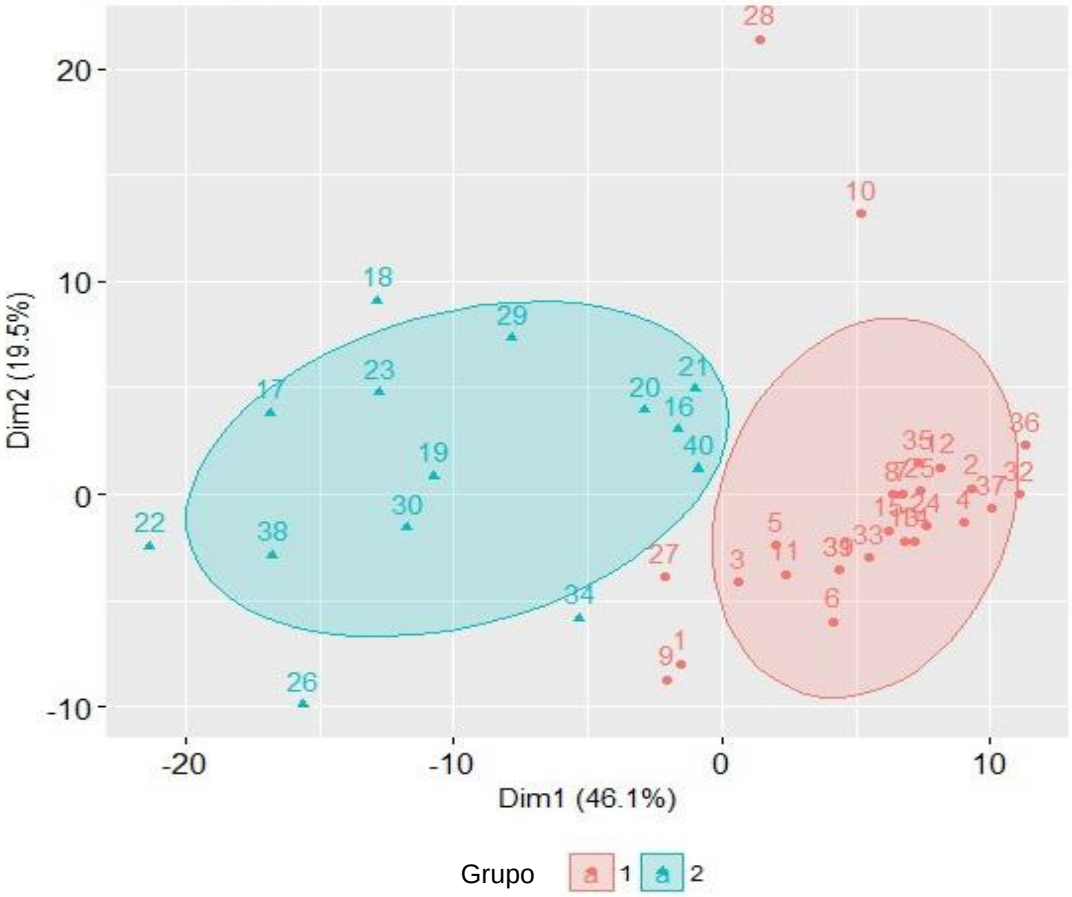


**Figura 9** - Gráfico dos grupos de azeite de oliva extra virgem por meio da técnica de agrupamento *Fuzzy.*

O gráfico do agrupamento *Fuzzy* (Figura 9), apresenta 65% de confiança promovendo uma distinção dos grupos. As amostras do azeite de oliva comum possuem uma similaridade forte dentro do seu grupo, já as amostras de azeite orgânico possuem pouca similaridade dentro do seu. Verifica-se que existem amostras que não fazem parte das elipses de confiança construídos a 65% de

confiança (amostras 10 e 28) do azeite de oliva extra virgem comum, no entanto o grupo é mais homogêneo do que o grupo do azeite de oliva extra vigem orgânico que é mais heterogêneo.

Nota-se ainda que existem amostras tidas como comuns, mas apresentam maior similaridade com o grupo orgânico. No estudo realizado por Oliveira et al. (2016) verificou-se que o método de agrupamento *Fuzzy* alocou cem cisternas localizadas no sertão do Pajeú para o grupo ao qual a análise obteve a mais elevada pertinência. Assim, as determinações de grupos foram eficazes na análise de agrupamento de cisternas de placas da região do Pajeú. Neste trabalho verificamos que a análise de agrupamento *Fuzzy* conseguiu, de maneira eficaz, distinguir os grupos de azeite em orgânico e comum de maneira notória.

## 5.4 Validação do Método

A estatística média da silhueta obtida pelo método do agrupamento *Fuzzy* foi 0,73 (Figura 10), valor que não levanta evidências da inadequação com relação a classificação dos azeites nos respectivos grupos. Isto é, o agrupamento realizado está adequado, pois de acordo com Vale (2005), $S(i)$ entre 0,71 e 1,00 é considerado estrutura forte e distinta.

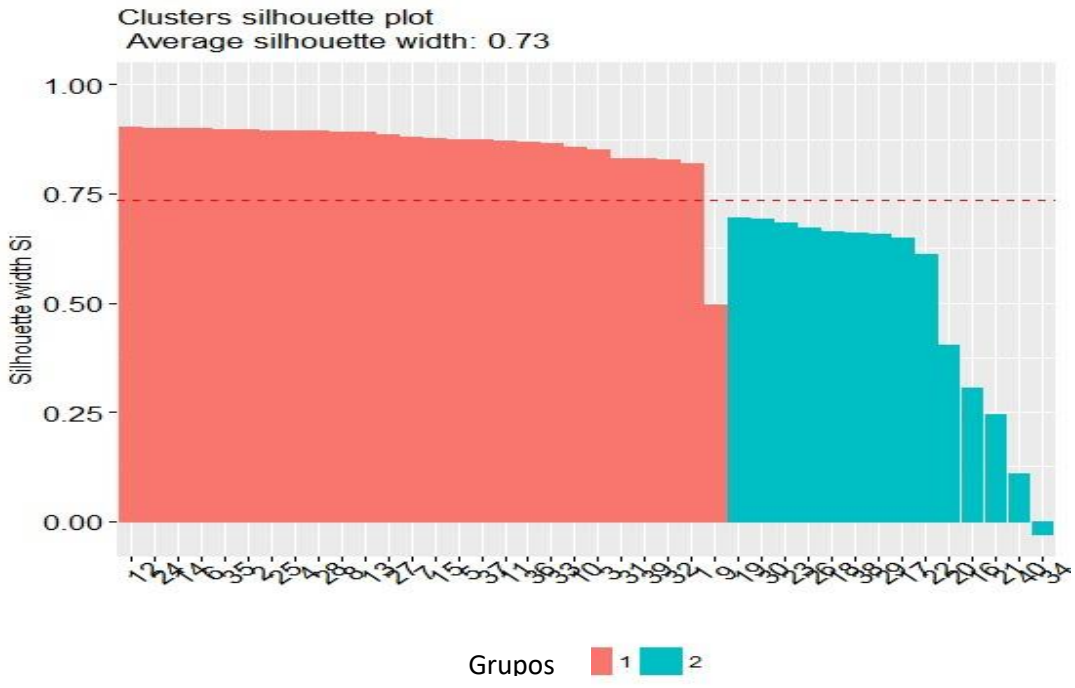


**Figura 10** - Gráfico da Estatística da Silhueta.

As observações estão bem agrupadas em seus respectivos grupos. Notamos que o grupo dos azeites comuns são os que estão melhor agrupados, pois os seus valores estão todos positivos. Evidenciou- se que a estatística da silhueta apesar de verificar algumas amostras como sendo classificadas no grupo dos azeites extra virgem comum as amostras possuem fortes características do azeite de oliva extra virgem orgânico. Com relação ao comportamento das amostras dentro de cada grupo, observa-se a amostra 40 no grupo 2, tem um coeficiente negativo indicando que a mesma não está bem alocada dentro do grupo dos azeites orgânicos. Sendo assim, a amostra 40 apesar de está presente no grupo 2 ela tem fortes características do grupo 1.

De acordo com a Tabela 1 utilizada para comparação dos grupos analisou de maneira coerente as amostras de azeite de oliva, corroborando com a verificação da pertinência e similaridade de cada amostra para com seu determinado grupo.

Tabela 1- Tabela de comparação dos grupos formados pelas amostras do azeite extra virgem.

| Amostras | Comum | Orgânico |
|---|---|---|
| **1** | 89% | 11% |
| **2** | 95% | 5% |
| **3** | 91% | 9% |
| **4** | 95% | 5% |
| **5** | 93% | 7% |
| **6** | 97% | 3% |
| **7** | 95% | 5% |
| **8** | 96% | 4% |
| **9** | 68% | 32% |
| **10** | 93% | 7% |
| **11** | 93% | 7% |
| **12** | 97% | 3% |
| **13** | 96% | 4% |
| **14** | 97% | 3% |
| **15** | 95% | 5% |
| **16** | 34% | 66% |
| **17** | 14% | 86% |
| **18** | 13% | 87% |
| **19** | 8% | 92% |
| **20** | 29% | 71% |
| **21** | 37% | 63% |
| **22** | 18% | 82% |
| **23** | 9% | 91% |
| **24** | 97% | 3% |
| **25** | 95% | 5% |
| **26** | 10% | 90% |
| **27** | 95% | 5% |

| | | |
|---|---|---|
| **28** | 95% | 5% |
| **29** | 10% | 90% |
| **30** | 8% | 92% |
| **31** | 91% | 9% |
| **32** | 89% | 11% |
| **33** | 94% | 6% |
| **34** | 47% | 53% |
| **35** | 96% | 4% |
| **36** | 92% | 8% |
| **37** | 91% | 9% |
| **38** | 13% | 87% |
| **39** | 91% | 9% |
| **40** | 42% | 58% |

Fonte: Análise realizada por RMN $H^1$ e analisadas por Agrupamento *Fuzzy.*
Nota: Dados trabalhados pelo autor.

Com relação a Tabela 1 confirma- se a similaridade das amostras com os grupos e entre os grupos, notamos que algumas amostras possuem características acentuadas para os dois grupos, por exemplo as amostras 40 e 34, pelo agrupamento *Fuzzy* ambas estão separadas com maior similaridade para o grupo do azeite orgânico, mas constatou- se pela tabela de comparação de grupos que a similaridade da amostra entre os grupos está bem acentuada, esse fato pode estar diretamente correlacionado com o modo que as amostras foram diluídas para a realização das análises.

# 6- Conclusões

A técnica de agrupamento não hierárquica de *Fuzzy* distingue os grupos de azeite de oliva extra virgem, com cerca de 65% de confiança. A qualidade dos agrupamentos foi atribuída por meio do índice da estatística da silhueta com s(*i*) 0,73 indicando força e poder de distinção no agrupamento. O agrupamento *Fuzzy* não apresentou homogeneidade total dos azeites alocados em cada grupo. A utilização da técnica de PCA consegue verificar o comportamento dos dados provindos do azeite de oliva extra virgem, mas não separa explicitamente os azeites em orgânicos e comuns.

# 7- Perspectivas

✓ Realizar aplicações de técnicas estatísticas de agrupamento hierárquico e rede neurais.

✓ Criar modelos metabônomicos desenvolvidos para utilização de outros biofluidos como urina, soro, sêmen para fins de classificação envolvendo diferentes patologias e estatísticas multivariadas.

✓ Iniciar coletas de azeite de oliva, para tornar os modelos metabônomicos desenvolvidos para classificação de azeite de oliva mais robustos.

# Referências Bibliográficas


ALBUQUERQUE, M. A. **análise de agrupamento hierárquica e incremental-estudo de caso em ciências florestais**. 2013. 160f. Tese (Doutorado em Biometria e Estatística Aplicada) – Universidade Federal Rural de Pernambuco, Departamento de Estatística e Informática, Recife, 2013.

Anton, H., Rorres C., **Álgebra Linear com Aplicações**, Bookman, Porto Alegra, 2004

AUED, Pimentel, S,; Takemoto, E.; Kumagai, E.E.; Cano, C.B. Determinação da diferença entre o valor real e teórico do triglicerídeo ECN 42 para a detecção de adulteração em azeites de oliva comercializados no Brasil. **Química Nova**, v. 31, p. 31- 34, 2008.

ARDOY, Zamora M.A., Báñez Sánchez F., Báñez Sánchez C., Alaminos García P. Aceite de oliva: influencia y beneficios sobre algunas patologías. **An Med Interna**,Madrid, 138-142,2004.

BACCOURI, O., Bendini, A., Cerretani, L., Guerfel, M., Baccouri, B., Lercker, G., Zarrouk, M., Miled, D. D. B., *Comparative study on volatile compounds from Tunisian and Sicilian monovarietal virgin olive oils.* **Food Chemistry**, v. 111, p. 322-328, *2008.*

BECKER, Denise Fabiana Silvestre. 2004. **Quantificação de fitoesteróis em azeite de oliva (Olea europaea) por cromatografia em fase gasosa**. Dissertação (Mestrado) – Universidade Estadual de Campinas. Faculdade de Engenharia de Alimentos, 2004.

BRASIL. Ministério da Saúde. Agência Nacional de Vigilância sanitária. Resolução RDC n° 270, de 22 setembro de 2005. Regulamento técnico para óleos vegetais, gorduras e creme vegetal. Diário Oficial da República Federativa do Brasil, Brasília – DF. 23 de setembro de 2005.

BRASIL. Agência Nacional de Vigilância Sanitária (ANVISA). Resolução RDC nº 482, de 23/09/1999, com alguns valores atualizados na resolução RDC nº 270, de 22/09/2005. “Alimentos com alegações visando a proteção à saúde da população”, 2005.

BRERETON, R.G. Introduction to multivariate calibration in analytical chemistry. **Analyst**, v. 125, p. 2125-2154, 2000.

BRERETON, R. G. Chemometrics and Statistics: Multivariate Classification Techniques. In: **Reference Module in Chemistry, Molecular Sciences and Chemical Engineering**. [s.l.] Elsevier, 2013a.

BRERETON, R. G., Chemometrics. John Wiley & Sons West Sussex, Inglaterra, 2003.

DIEHL, B. Chapter 1 - Principles in NMR Spectroscopy. In: DIEHL, U. H. W. (Ed.). **NMR Spectroscopy in Pharmaceutical Analysis**. Amsterdam: Elsevier, p. 1– 41, 2008.

DOCAMPO, E.; COLLADO, A.; ESCARAMIÍS, G.; CARBONELL, J.; RIVERA, J. et al. Cluster Analysis of Clinical Data Identifies Fibromyalgia Subgroups. **PLoS ONE**, Iran University of Medical Sciences, Iran (Republic of Islamic), p. 1 – 8, 2013.

DYSON, H. J.; PALMER III, A. G. 1.9 Introduction to Solution State NMR Spectroscopy. In: EGELMAN, E. H. (Ed.). **Comprehensive Biophysics**. Amsterdam: Elsevier, p. 136–159, 2012.

EMBRAPA. Cultivo de oliveira (Olea europaea L.) Dez. 2011. Disponível em: http://www.cpact.embrapa.br/publicacoes/catalogo/tipo/sistemas/sistemas16_novo_novo/11_mercados_e_comercializacao.htma. Acesso em: 10/ 12/2016.

FERREIRA, Daniel Furtado. Estatística multivariada. Lavras: Editora UFLA, 1° reimpressão. 2008. 662 p.

GRAYBILL, F. Matrices with applications in statistics. 2 ed. **Principles and procedures of statistics**, 1983.

GODOY, M. M. G. et al. Hepatitis C virus infection diagnosis using metabonomics. **Journal of Viral Hepatitis**, v. 17, n. 12, p. 854–858, 2010.

GOODPASTER, A. M.; ROMICK-ROSENDALE, L. E.; KENNEDY, M. A. Statistical significance analysis of nuclear magnetic resonance-based metabonomics data. **Analytical Biochemistry**, v. 401, n. 1, p. 134–143, 2010.

GONÇALVES, Rhayanna P.; MARÇO, Paulo H.; VALDERRAMA, Patrícia. DEGRADAÇÃO TÉRMICA DE TOCOFEROL E PRODUTOS DE OXIDAÇÃO EM DIFERENTES CLASSES DE AZEITE DE OLIVA UTILIZANDO ESPECTROSCOPIA UV-VIS E MCR-ALS. **Quim. Nova**, v. 38, n. 6, p. 864-867, 2015.

GRIFFITHS, W. J. et al. Targeted Metabolomics for Biomarker Discovery. **Angewandte Chemie International Edition**, v. 49, n. 32, p. 5426–5445, 2010.

GRUTZNER, J. B. NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY | Principles. In: POOLE, P. W. T. **Encyclopedia of Analytical Science (Second Edition)**. Oxford: Elsevier, p. 211–237, 2005.

HENDRIKS, M. M. W. B. et al. Data-processing strategies for metabolomics studies. **In-Vivo and On-Site Analysis II**, v. 30, n. 10, p. 1685–1698, 2011.

IOOC (International Olive Oil Council). Trade standard applying to olive oil and olive pomace oil. RES. COI/T.15/NC no, 2003.

JOHNSON, R. A.; WICHERN, D. W. Applied multivariate statistical analysis.4ed. New Jersey: Prentice Hall, 816 p, 1998.

JUCHEM, C.; ROTHMAN, D. L. Chapter 1.1 - Basis of Magnetic Resonance. In: ROTHMAN, C. S. **Magnetic Resonance Spectroscopy**. San Diego: Academic Press, p. 3–14, 2014.

JURS, P. C. Pattern recognition used to investigate multivariate data in analytical chemistry. **Science**, v. 232, n. 4755, p. 1219–1224, 1986.

KACHOURI, F., Hamdi, M., Enhancement of polyphenols in olive oil by contact with fermented olive mill wastewater by Lactobacillus plantarum. **Process Biochemistry**, v. 39, 841–845, 2004.

KAUFMANN, L.; ROUSSEEUW, P. J. Finding groups in data: an introduction to cluster analysis. New York: Jonh Wiley, 1990. 342 p.

KLIR, G. J. E YUAN, B. Fuzzy Sets and Fuzzy Logic: Theory and Applications. **Prentice-Hall**, 1995.

KOWALSKI B.; SEASHOLTZ M. B. The parsimony principle appliedto multivariate calibration. **Analytica Chimica Acta**. v. 277, n. 2, p. 165-177, 1993.

KOSKO, Bart. Fuzzy systems as universal approximators. **IEEE transactions on computers**, v. 43, n. 11, p. 1329-1333, 1994.

LAY, D. C. Álgebra Linear e Suas Aplicações, 2007.

LINDON, J. C.; HOLMES, E.; NICHOLSON, J. K. Pattern recognition methods and applications in biomedical magnetic resonance. **Progress in Nuclear Magnetic Resonance Spectroscopy**, v. 39, n. 1, p. 1–40, 2001.

LINDON, J. C.; NICHOLSON, J. K.; WILSON, I. D. Directly coupled HPLC– NMR and HPLC–NMR–MS in pharmaceutical research and development. **Hyphenated Techniques in LC and their input in Biosciences**, v. 748, n. 1, p. 233–258, 2000.

LUDWIG JR., O., MONTGOMERY, E. Redes Neurais: Fundamentos e Aplicações com Programas em C. 1 ed. Editora: Ciência moderna. 2007.

LUZ, E. R. Predição de propriedades de gasolinas usando espectroscopia FTIR e regressão por mínimos quadrados parciais. 2003. Dissertação (Mestre e Dissertação (Mestre em Química). Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2003.

MANDARINO, J. M. G; ROESSING, A. C.; BENASSI, V. de T. Óleos: alimentos funcionais. Londrina: EMBRAPA Soja, 2005, 91 p.

MARDIA, K.V.; KENT, J. T.; BIBBY, J. M. Multivariate analysis. London: Academic, 1979, 512 p.

MEIRA, Marilena et al. Identificação de adulteração de biocombustível por adição de óleo residual ao diesel por espectrofluorimetria total 3D e análise das componentes principais. **Quimica Nova**, v. 34, p. 621-624, 2011.

MESQUITA, D. L.; OLIVEIRA, A. F. de; MESQUITA, H. A. de. Aspectos econômicos da produção e comercialização do azeite de oliva e azeitona**. Informe Agropecuário**, Belo Horizonte, v. 27, n. 231, p. 7-12, 2006.

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO - MAPA. **Azeites de Oliva e Óleo de bagaço de oliva (Instrução Normativa)**. Brasília: MAPA, n. 1, 30 jan. 2012. Disponível em: Acesso em: Julho. 2016.

MOITA-NETO, J. M. Estatística Multivariada na Pesquisa. 5 ed. **Sapiência** (FAPEPI), 2009.

Morais, J. T. G. Análise de componentes principais integrada a redes neurais artificiais para predição de matéria orgânica, 2016.

NETO, J. M. M., MOITA, G.C.,. Uma introdução à análise exploratória de dados multivariados. **Química Nova**, v. 21, n. 4, p. 467- 469, 1998.

NAES, T.; ISAKSSON, T.; FEARN, T.; DAVIES, T. A User-friendly Multivariate Calibration and Classification. 1° ed., Chichester/UK: NIR Publications, p. 114-119, 2002.

NINFALI, P. et al. A 3-year Study on Quality, Nutritional and Organoleptic Evaluation of Organic and Conventional Extra-Virgin Olive Oils. **Journal of the American Oil Chemists' Society**, v. 85, n. 2, p. 151–158, 2008.

NICHOLSON, J. K.; WILSON, I. D. High resolution proton magnetic resonance spectroscopy of biological fluids. **Progress in Nuclear Magnetic Resonance Spectroscopy**, v. 21, n. 4–5, p. 449–501, 1989.

NICHOLSON, J. K.; LINDON, J. C.; HOLMES, E. "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. **Xenobiotica**, v. 29, n. 11, p. 1181–1189, 1999.

OLIVEIRA, A. F.; VIEIRA NETO, J.; GONÇALVES, E. D.; MESQUITA, D. L. Pioneirismo marca pesquisa sobre oliveira em Minas Gerais. Informe Agropecuário, Belo Horizonte, v. 30, p. 7-15, 2009a.

OLIVEIRA, A.F.; VIEIRA NETO, J.; GONÇALVES, E.D; VILLA, F.; SILVA, L.F.O. Parâmetros físico-químicos dos primeiros azeites de oliva brasileiros extraídos em Maria da Fé, Minas Gerais. Scientia Agraria, v.11, p. 255- 261, 2010a.

OLIVEIRA, M.C.; RAMOS, J.D.; PIO, R.; CARDOSO, M.G. Características fenológicas e físicas e perfil de ácidos graxos em oliveiras no sul de Minas Gerais. Pesquisa Agropecuária Brasileira, v.47, p. 30- 35, 2012b.

OLIVEIRA, M. R. G.; Cruz, D.V.; Cunha Filho, M. Mapping plaques Cisterns by Fuzzy grouping analysis. **IEEE Latin America Transactions**, v. 14, n. 10, p. 4367- 4372, 2016.

OTTAVIAN, M.; FACCO, P.; BAROLO, M.; BERZAGHI, P.; SEGATO, S.; NOVELLI, E.; BALZAN, S., J. Food Eng, v. 113 (2), p. 289- 298, 2012.

Ouni, Y., Taamalli, A., Gómez-Caravaca, A. M., Segura-Carretero, A., Fernández-Gutiérrez, A., Zarrouk, M., Characterisation and quantification of phenolic compounds of extra-virgin olive oils according to their geographical origin by a rapid and resolutive LC–ESI-TOF MS method. Food Chemistry, v. 127, p. 1263-1267, 2011.

PALMER, A. G. 1.13 NMR Spectroscopy: NMR Relaxation Methods. In: EGELMAN, E. H. (Ed.). **Comprehensive Biophysics**. Amsterdam: Elsevier, p. 216– 244, 2012.

PAN, Z. et al. Principal component analysis of urine metabolites detected by NMR and DESI–MS in patients with inborn errors of metabolism. **Analytical and Bioanalytical Chemistry**, v. 387, n. 2, p. 539– 549, 2007.

PICARD, N.; MORTIER, F.; ROSSI, V.; FLEURY, S. G. Clustering species using a model of population dynamics and aggregation theory**. Ecological Modelling**, p. 152 –160, 2010.

RAPOPORT, H. F. Botánica y morfologia. In: BARRANCO, D.; FERNÁNDEZ-ESCOBAR, R.; RALLO, L. El cultivo del olivo. 2. ed. Madri: Mundi- Prensa- Junta de Andalucia, 1998. p. 651.

RALLO, L. Variedades de olivo en España: una aproximación cronológica. In: RALLO, L.; BARRANCO, D.; CABALLERO, J. M.; DEL RÍO, C.; MARTÍN, A.; TOUS, J.; TRUJILLO, I. (Ed.). Variedades de olivo en España. Sevilla: Consejería de Agricultura y Pesca de la Junta de Andalucía; Madrid: Ministerio de Agricultura, Pesca y Alimentación/ Mundi-Prensa, cap. 1, p. 17-44, 2005

ROMERO, I. et al. Validation of SPME–GCMS method for the analysis of virgin olive oil volatiles responsible for sensory defects. **Talanta**, v. 134, p. 394– 401, 2015.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation

of cluster analysis. Journal Computational Applied Mathematics. v. 20, p. 53–65,

1987.

SACCHI, R. et al. Characterization of Italian Extra Virgin Olive Oils Using 1H-NMR Spectroscopy. **Journal of Agricultural and Food Chemistry**, v. 46, n. 10, p. 3947–3951,1998.

SCHRIPSEMA, J. Application of NMR in plant metabolomics: techniques, problems and prospects. **Phytochemical Analysis**, v. 21, n. 1, p. 14– 21, 2010.

SILVA, A.P.D. Efficient Variable Screening for Multivariate Analysis. Journal of Multivariate Analysis, v. 76, p. 35- 62, 2001.

TAKAGI, Tomohiro; SUGENO, Michio. Fuzzy identification of systems and its applications to modeling and control. **IEEE transactions on systems, man, and cybernetics**, n. 1, p. 116-132, 1985.

TENA, N.; Garcia-González, D. L.; Aparicio, R.; J. Agr. Food Chem, 2009, p. 57.

VALE, M. N. Agrupamentos de dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos. Rio de Janeiro, 2005. 120 f. Dissertação (Mestrado)

– Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, 2005.

VIGLI, G. et al. Classification of Edible Oils by Employing 31P and 1H NMR Spectroscopy in Combination with Multivariate Statistical Analysis. A Proposal for the Detection of Seed Oil Adulteration in Virgin Olive Oils. **Journal of Agricultural and Food Chemistry**, v. 51, n. 19, p. 5715– 5722, 2003.

VIEIRA NETO, Santiel Alves et al. Formas de aplicação de inoculante e seus efeitos sobre a nodulação da soja. **Revista Brasileira de Ciência do Solo**, v. 32, n. 2, 2008.

WESTERHUIS, J. et al. Assessment of PLSDA cross validation. **Metabolomics**, v. 4, n. 1, p. 81– 89, 2008.

Wang, L.X. Universal aproximation by hierarchical *Fuzzy* systems, *Fuzzy sets and Systems.* p. 223- 230, 1998.

Wang, L.-X; Mender, J.M. *Fuzzy* basis functions, universal aproximation, and orthogonal least-square e learning, *IEEE Transaction on Neural Netword*. p. 807-814, 1992.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **PLS Methods**, v. 58, n. 2, p. 109– 130, 2001.

XU, J. et al. Statistical two-dimensional correlation spectroscopy of urine and serum from metabolomics data. **Chemometrics and Intelligent Laboratory Systems**, v. 112, p. 33–40, 2012.

XU, R.; WUNSCH, D. II survey of clustering algorithms**. *IEEE Transactions on Neural Networks*,** v. 16, n. 3, p. 645 – 678, 2005.

ZALIK, K. R. Cluster validity index for estimation of Fuzzy clusters of different sizesand densities. **Pattern Recognition**, v. 43, p. 3374 – 3390, 2010.