JEAP

ISSN: 2525-815X

Journal of Environmental Analysis and Progress

Journal homepage: www.jeap.ufrpe.br/

10.24221/jeap.6.1.2021.2815.035-043

Use of Multivariate Statistical Methods for Classification of Olive Oil

Moacyr Cunha Filho[a], Renisson Neponuceno de Araújo Filho[b], Ana Luiza Xavier Cunha[c], Victor Casimiro Piscoya[d], Guilherme Rocha Moreira[a], Iloane dos Santos Lima[a], Ronaldo Dionísio da Silva[e], Rejane Magalhães de Mendonça Pimentel[f], Dayane de Souza Lima[g], Josue Luiz Marinho Junior[g], Ricardo Oliveira Silva[h], Tatijana Stosic[a], Milton Marques Fernandes[i], João Lucas Aires Dias[b]

[a] Universidade Federal Rural de Pernambuco-UFRPE, Programa de Pós-Graduação em Biometria e Estatística Aplicada, Rua Dom Manoel de Medeiros, s/n, Dois Irmãos, Recife, Pernambuco, Brasil. CEP: 52171-900. E-mail: guirocham@gmail.com, iloane.lima@hormail.com, tastosic@gmail.com *Corresponding author: moacyr2006@gmail.com

[b] Universidade Federal do Tocantins-UFT, Curso de Engenharia Florestal. Rua Badejos, Lote 7, s/n, Chácara 69-72, Jardim Sevilha, Gurupi, Tocantins, Brasil. CEP: 77404970. E-mail: renisson@uft.edu.br, jlucas.florestal@gmail.com.

[c] UFRPE, Programa de Pós-Graduação em Engenharia Agrícola. E-mail: analuizaxcunha@gmail.com.

[d] UFRPE, Programa de Pós-graduação em Engenharia Ambiental. E-mail: victor.piscoya@ufrpe.br.

[e] Instituto Federal de Pernambuco-IFPE, Curso de Agronomia. Rua Propriedade Terra Preta Zona Rural, s/n, Vitória de Santo Antão, Pernambuco, Brasil. CEP: 55600-000. E-mail: ronaldo.dionisio@vitoria.ifpe.edu.br.

[f] UFRPE, Departamento de Biologia, Área de Botânica. E-mail: rejanemmpimentel@gmail.com.

[g]UFT, Programa de Pós-Graduação em Ciências Florestais e Ambientais. E-mail: daythi16@gmail.com, josue.marinho@hotmail.com.

[h] Universidade Federal de Pernambuco-UFPE, Departamento de Quimica Fundamental. Av. Professor Moraes Rego, s/n, Recife, Pernambuco, Brasil. CEP: 50740-540. E-mail: rosilvaufpe@gmail.com.

[i] Universidade Federal de Sergipe-UFS, Departamento de Departamento de Engenharia Florestal. Av. Marechal Rondon, s/n, Jardim Rosa Elze, São Cristóvão, Sergipe, Brasil. CEP:49100-000. E-mail miltonmf@gmail.com.

ARTICLE INFO

Received 17 Oct 2019
Accepted 11 Sep 2020
Published 11 Feb 2021

ABSTRACT

Multivariate statistical methods can contribute significantly to classification studies of extra virgin and common olive oil groups. Therefore, nuclear magnetic resonance (NMR) was used to discriminate olive oil samples, multivariate statistical techniques Principal Component Analysis - PCA, Fuzzy Cluster, Silhouette Validation Method to describe and classify. The groups' distinction into organic and common was observed by applying the non-hierarchical Fuzzy grouping with a distinction between the two groups with a 65% confidence interval. The validation was performed by the silhouette index that presented S (i) of 0.73, which showed that the adopted grouping presented adequate strength and distinction criterion. However, PCA only analyzed the behaviors of data from extra virgin olive oil. Thus, the Fuzzy clustering method was the most suitable for classifying extra virgin olive oil.

Keywords: *Olea europaea*, NMR, Principal Component Analysis, Fuzzy, Silhouette method.

Introduction

Olive oil is an oil of plant origin, derived from the fruit of a typical Mediterranean tree species, *Olea europaea* L., also known as an olive tree (Morais et al., 2016). This oil is known and used worldwide in many ways for its health benefits (Vogel et al., 2015). The increasing use of olive oil increases market demand and contributes to the recognition of its use (Wang et al., 2019). Thus, making it necessary to create an international agreement for the marketing of the product, regulated by the International Olive Oil Council (Alvarenga et al., 2017).

In Brazil, the regulation of olive oil is carried out by the Ministry of Agriculture, Livestock, and Supply (Gonçalves et al., 2015), through normative instructions. They aim to classify the product based on identity and quality requirements while defining the sampling, mode of presentation, and marking or labeling of packaging according to a product classification (Gonçalves et al., 2015). Besides these, another commonly analyzed parameter is the classification of common and organic olive oil (Ferreira et al., 2009).

This classification is done in the laboratory by adopting the nuclear magnetic resonance spectroscopy (NMR) technique (Zhu et al., 2017). Several overlapping information, even though it is very efficient for structural clarification of the analyzed sample. In general, they can make data interpretation confusing, and in some cases, lead to misclassification (Guellaoui et al., 2019). The use of multivariate statistics tools with the application of metabonomic strategies may be a way to solve such a problem in NMR analysis (Xu et al., 2012a).

In this context, the application of methods can improve quantitative research quality, making data interpretation more understandable and minimizing the possibility of misclassification of components. There are several multivariate statistical techniques; it is up to the researcher to define their objective and the nature of their data because there are grouping and classification techniques, each with a distinct function. Therefore, the present work aimed to use multivariate statistical methods to classify common or organic olive oil.

Material and Methods

The study was conducted at the Departamento de Química Fundamental (DQF) of the Universidade Federal de Pernambuco (UFPE), where 40 samples of olive oil were analyzed. The extra virgin olive oil samples were dissolved 60 µL of the sample in 640 µL CDCl3 in a 435 mm diameter NMR tube. NMR spectra were obtained using the VNMRS400 spectrometer, operating at 399.99 MHz, for the 1H core, with a 6.4 kHz spectral window, hold time (d1) equal to 1 s, acquisition time equal to 2.556 s, 90º radiofrequency (RF) pulse, 64 repetitions, and 26ºC temperature. The spectra were processed with 0.3 Hz line broadening. After spectroscopic analysis, the samples were stored in a clean amber container for future proper disposal. The ^{1}H NMR spectra have their phases adjusted, baseline correction, and bin construction manually using the Mestre Nova 9.0 software. Bins were defined at 0.03 ppm intervals between δ 0.00 and 6.80 ppm. Data were arranged in a matrix for chemometric treatment.

The data were arranged in a matrix containing information regarding the nature of the sample, spectral data in a format of bins, and information related to the class to which a given sample belongs. The preprocessing techniques used in this work were: sum normalization and line self-scaling, and models were built for each preprocessing type.

Sum normalization was obtained by dividing each bin by the respective sum of each sample's total integration area (Equation 1). This preprocessing aims to obtain data that can be compared with each other without changing the information contained in the variables.

$$A_{ij}^{ns} = \frac{A_{ij}}{\sum_i^j A_{ij}} \qquad \text{Eq.(1)}$$

A_{ij}^{NS} = bin normalized by the sum, $A_{\,ij}$ = original bin, $\sum_j^i A_{ij}$ = sum of integration areas for each sample.

As a principle for the calculation of the PCA, it considered a random vector $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{p})$, containing p components, with a mean vector $\mu = \mathrm{E}(\mathrm{X}) = (\mu_{1,}\, \mu_2, \ldots, \mu_\mathrm{p})$. The covariance matrix of the random vector X, square of dimension p, is denoted by $Cov(X) = \Sigma_{p\times p}$. The covariance matrix is symmetric, nonnegative matrix, that is, $a^t\Sigma_a > 0$ for every vector of constants $a \in R^p$. This condition implies that the eigenvalues of the matrix $\Sigma_{p\times p}$ denoted by $\lambda_1, \lambda_2, \ldots, \lambda_{p,}$ (p,) are nonnegative, ie $\lambda_i \geq 0$, for any $i = 1,2,\ldots,p$ (Graybill, 1983). By the Spectral Decomposition Theorem (Lay, 2007), where $\Sigma_{p\times p}$ is a covariance matrix, there is an orthogonal matrix $\mathrm{P}_{p\times p}$,, i.e., $\mathrm{P}^T\mathrm{P} = \mathrm{P}\mathrm{P}^T = I$ (Equation 2):

$$P^T \sum P = \theta \qquad \text{Eq.(2)}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_{p,}$, are the eigenvalues of the matrix $\Sigma_{p\times p}$ ordered in descending order. In this case, we say that the matrix $\Sigma_{p\times p}$ is similar to the matrix θ.

The 1st column of matrix θ is the normalized auto vector, and i corresponding to the auto vector $\lambda_{i,}$ with $i = 1,2,\ldots,p$; which is denoted by $e_i = (e_1, e_2, \ldots, e_p)^T$. Then the matrix θ is given by $\theta = [e_1, e_2, \ldots, e_p]$ and the spectral decomposition theorem gives the following valid equality (Equation 3):

$$\Sigma_{p\times p} = \sum_{i=1}^{p} \lambda e_i e_i^T = \mathrm{P}\theta\mathrm{P}^T \qquad \text{Eq.(3)}$$

Since $\theta_1, \theta_2, \ldots, \theta_p$, form a basis of R^p, the vector can be written as $\sum_{i=1}^{p} \alpha_i \mathrm{P}_i = \alpha^T\mathrm{P}$ for some $\alpha_i = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$.

Being θ orthogonal, $\alpha^T\alpha = 1$ and the variance of $\alpha^T X$ is less than or equal to λ_1 and taking $\alpha = \mathrm{O}_1$,, one has to $var(\mathrm{P}_1 X) = \mathrm{P}_1\Sigma\mathrm{P}_1 = \lambda_1$, and define a random variable $U_1 = \mathrm{P}_1^T X$ as the first major component of X. To obtain other major components, a non-correlation constraint of the next component U_i with the previously obtained components is made $(U_1, \dots, U_{i-1})$. Thus the components are defined as random vectors $U = (U_1, \dots, U_p) = \mathrm{P}^T$, where the columns of P are the auto vectors of Σ. Notably, the covariance matrix of the new matrix U is diagonal, where the elements are the eigenvalues λ_i.

In the application of dimensionality reduction, PCA has the property of minimizing the mean square error between the reconstructed data and the original data. For example, it is assumed that you have input data X of dimensionality m and output data Y of dimensionality m, where $m_1 < m$.

The cluster development line becomes critical when you want to classify a dataset according to its characteristics or measured variables. The term class is pertinent, given the information on how many partitions and which partitions are in a data set, as each observation or group of olive oil belonging to such samples. Thus, classification is called being the analysis performed in specific databases; the data analysis work is called clustering and aims to study the similarity relationships between the data or olive oil samples, determining which data form which groups. Groups are formed so that the similarity between the samples of one group is maximized (intra-group similarity) and the similarity between samples of different groups (intergroup similarity) is minimized. Then, formally, given an input data set ($\vec{X} \in \mathbb{R}^p$), a function is found (Equation 4):

$$f\colon \mathbb{R}^p \times \mathrm{W} \longrightarrow \mathrm{G} \qquad \text{Eq.(4)}$$

where, W is an adjustable parameter vector by means of a supervised or unsupervised learning algorithm, which determines c-groups from the original data matrix X, and, according to Xu et al. (2012), we have: $\mathrm{G} = \mathrm{G}_1, \mathrm{G}_2, \dots, \mathrm{G}_c$ (c ≤ n) where, $\mathrm{G}_i \neq \emptyset, \mathrm{i} = 1, \dots, \mathrm{c}$; $\bigcup_{i=1}^{c} \mathrm{G}_i = \mathrm{X}$; $\mathrm{G}_i \cap \mathrm{G}_j = \emptyset, \mathrm{i}, \mathrm{j} = 1, \dots, \mathrm{c}$, and $\mathrm{i} \neq$, assuming the classical approach of classification or grouping.

The silhouette index was proposed by Rousseeuw (1987) to evaluate partitioning methods. In this case, each object (olive oil sample) is represented by a value s (i) called a silhouette, which is based on the comparison of homogeneity and the "separation" of each group. Thus, for object i, the value of the silhouette is given by (Equation 5):

s(i)=(b(i)-a(i)/max 〚(a(i), b(i))〛 Eq.(5)

where $-1 \leq s(i) \leq 1$ and a(i) is the average distance from object i to objects in your group; b (i) is the average distance from object i to objects in other groups.

Negative values of negative s(i) suggest that the individual i is similar to individuals of other classes. Values of s (i) in the vicinity of 1 give evidence that i is well ranked.

Results and Discussion

The application of 1H NMR spectroscopy in analyzing the data obtained for samples of common and organic olive oil was superimposed (Figure 1). The olive oil sample profiles do not differ in the presence or absence of any characteristic signal, so they required preprocessing to eliminate possible effects caused by sample dilution. For that, we adopted the sum normalization technique performed on the line.

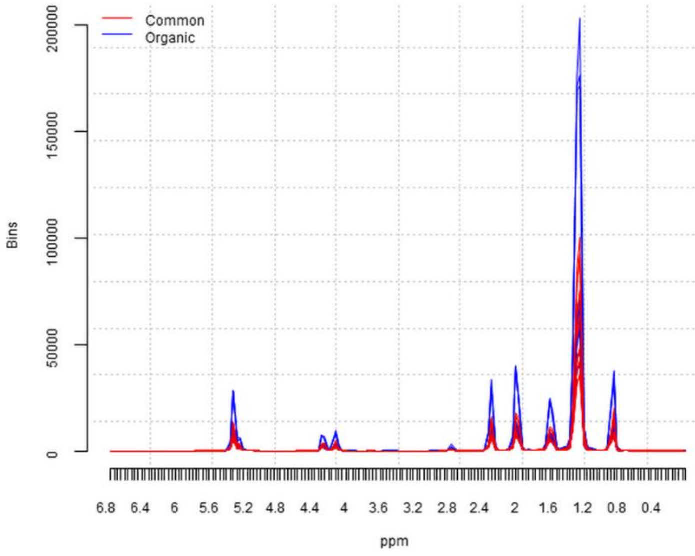

Figure 1. Bins obtained through ^{1}H NMR spectroscopy (400 MHz) of all common and organic olive oil samples. Font: Cunha Filho, M. (2919).

After obtaining the self-scaled bins on the line, the spectra were superimposed (Figure 2). It is possible to verify the effect of the preprocessing on the samples' classification because it is noted the presence of characteristic points that differentiate the processed samples.

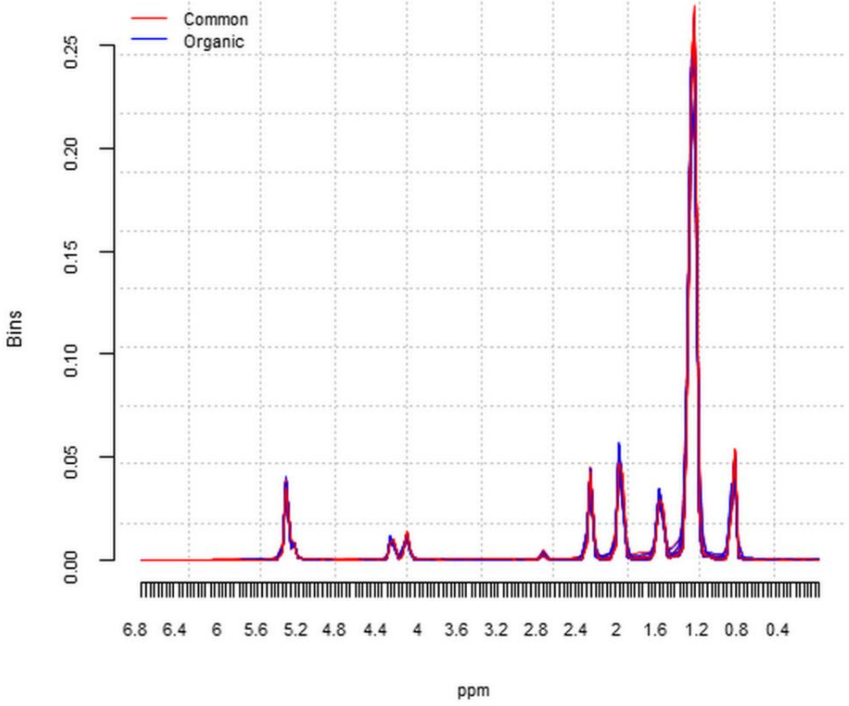

Figure 2. Auto-scaled data in a row of all common and organic olive oil samples. Font: Cunha Filho, M. (2919).

Self-scaling of the data contributed to adding equivalent weights to the samples' spectral points, enabling them to be classified later by statistical methods. It was observed that the two main components PC1 and PC2 showed that it is possible to describe 99.9% of the data variance, and PC1 contributes about 99.75% of this variance. According to Meira et al. (2011), working with biofluids, it took three principal components (PC) to explain the behavior of their data and understand the contribution of each substance in the final product structure. Meira et al. (2011) found that three main components are responsible for 95.39% of the variance, attributing 55.98% of the variance to PC1, 33.62% for PC2, and 5.79% for PC3. In this context, it was found that the first significant component (PCA) results for extra virgin olive oil data were significant enough to describe the behavior of the analyzes. Figure 3 shows the data behavior for the first five principal components.

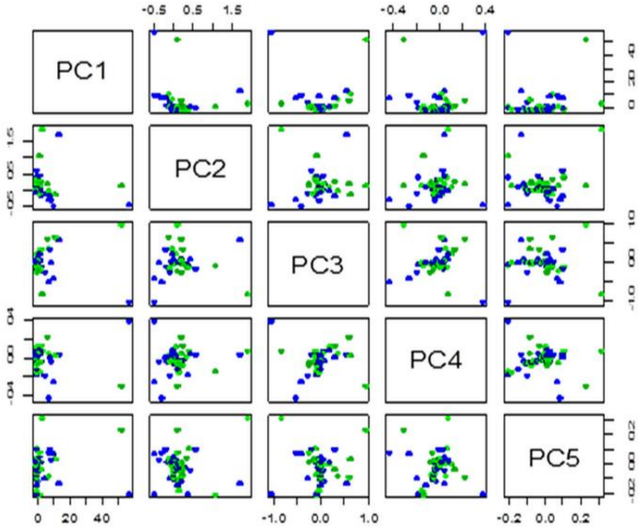

Figure 3. Correlation between each principal component and organic and common oils. Font: Cunha Filho, M. (2919).

The cumulative percentage of the total variation of the first two components (99.9%) satisfactorily explains the variability in the samples evaluated. According to Mardia et al. (1979), when in a Principal Component Analysis, the first two or three components accumulate a relatively high percentage of the total variation (usually above 70%), they can satisfactorily explain the variability manifested between the evaluated samples. Here we have that the first two principal components present a high explanatory power among the studied groups. Analyzing the score graph (Figure 4), it was observed that the accumulated percentages of variance are explained by the first two main components, highlighting that the first Principal Component has high significance.

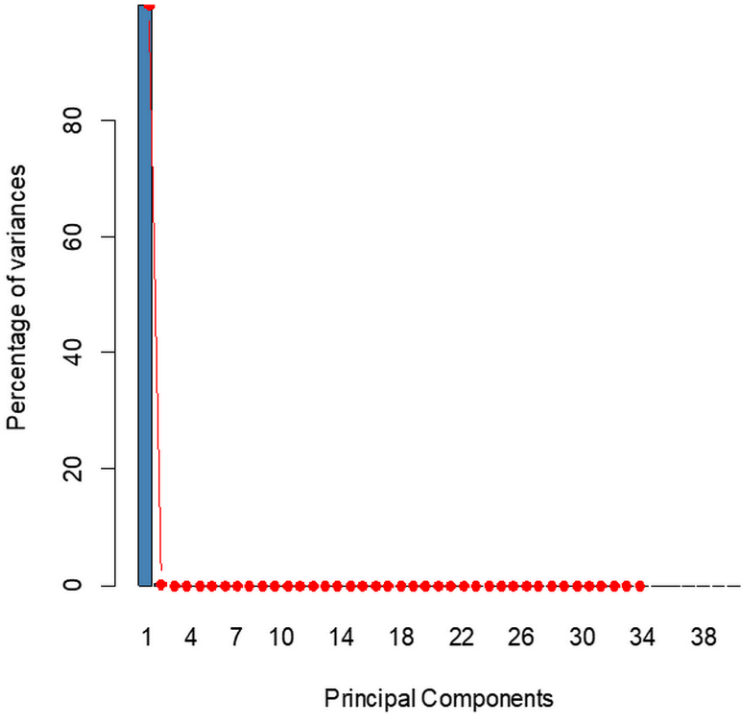

Figure 4. Cumulative percentage graph of variances principal component (PC) to each principal component. Font: Cunha Filho, M. (2919).

Figure 4 shows a two-dimensional representation of olive oil variables, commonly known as "biplot". The variables are grouped according to their correlation coefficients, with each major axis corresponding to a set of correlated variables. Since the correlations between the variables result from the olive oils measurements, each major axis represents a direction of space along which the variance (or difference) between the olive oils is maximized. In the biplot (Figure 5), the graph of PC1 versus PC2 shows that the samples of the organic olive oil have a larger grouping among their similar ones than the data of the common olive oil, thus representing an incredibly significant distinction between some samples. It is also found that there was no clear separation between the groups since samples of both organic and common olive oil have similar characteristics at the same point.

The PCA shows no clear separation between the samples of common and organic olive oil, thus making it impossible to use this method for the separation of this type of product. This result may be explained by the similarity of the compounds in both cases. Because of this result, it is necessary to use the Fuzzy clustering technique to verify the separation between groups of olive oil samples, as shown in Figure 6.

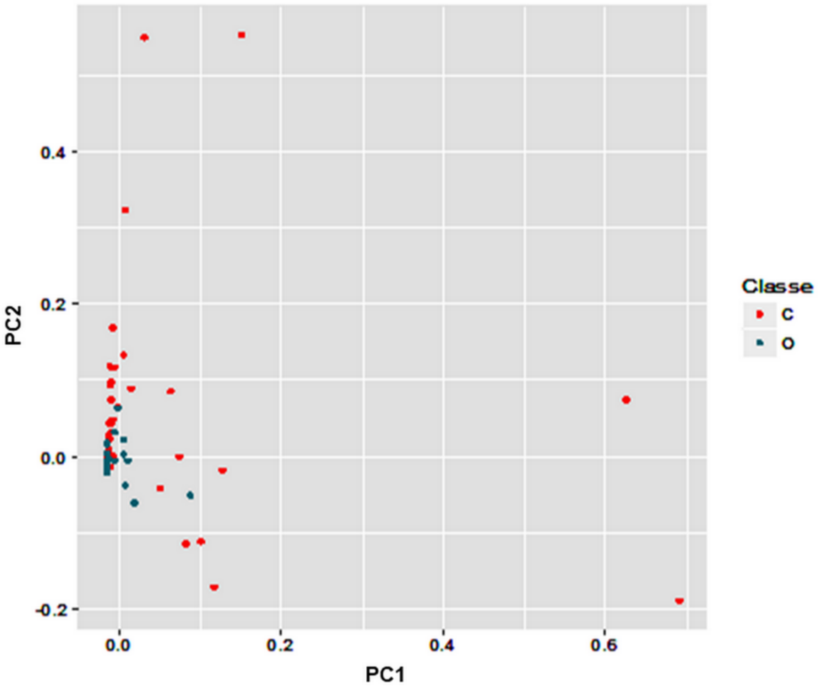

Figure 5. Graph of PC1 versus PC2 with 99.9% of explained data variance. Font: Cunha Filho, M. (2919).

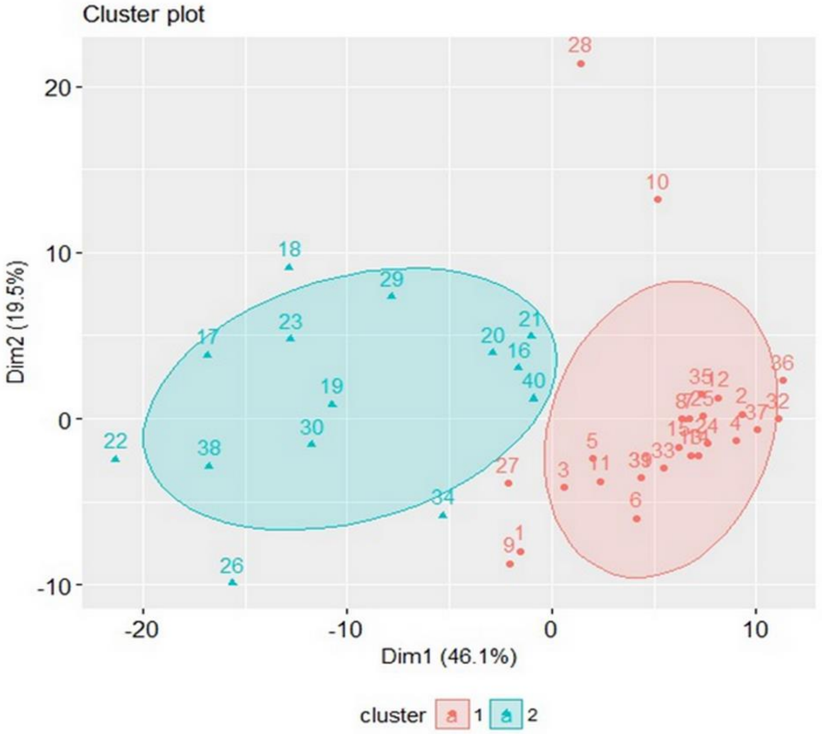

Figure 6. Fuzzy grouping graph of all common and organic olive oil samples. Font: Cunha Filho, M. (2919).

The Fuzzy grouping graph (Figure 6) is 75% normal and promotes group distinction. Samples of common olive oil have a substantial similarity within their group, whereas organic olive oil samples have little similarity within their group. Some samples are not part of the confidence ellipses constructed at 65% confidence (samples 10 and 28) and may have occurred due to improper handling of some samples.

It is also noted that there are samples considered as common, but they present more remarkable similarity with the organic group. In the study by Oliveira et al. (2016), it was found that the Fuzzy clustering method allocated 100 cisterns located in the Pajeú backlands to the group to which the analysis obtained the highest relevance. Thus, the group determinations were useful in the cluster analysis of plates of the Pajeú region. This study found that the Fuzzy cluster analysis distinguished the organic and common olive oil groups noticeably. The sensitivity of the technique was observed concerning the analyzed samples since we were able to identify some mistakes made in handling the technique for obtaining data from olive oil.

The average silhouette statistic obtained by the Fuzzy cluster method was 0.73 (Figure 7), a value that does not raise evidence of inadequacy regarding the classification of olive oils in the respective groups. The grouping performed is adequate, since according to Vale (2005), S (i) between 0.71 and 1.00 is considered a distinct and robust structure.

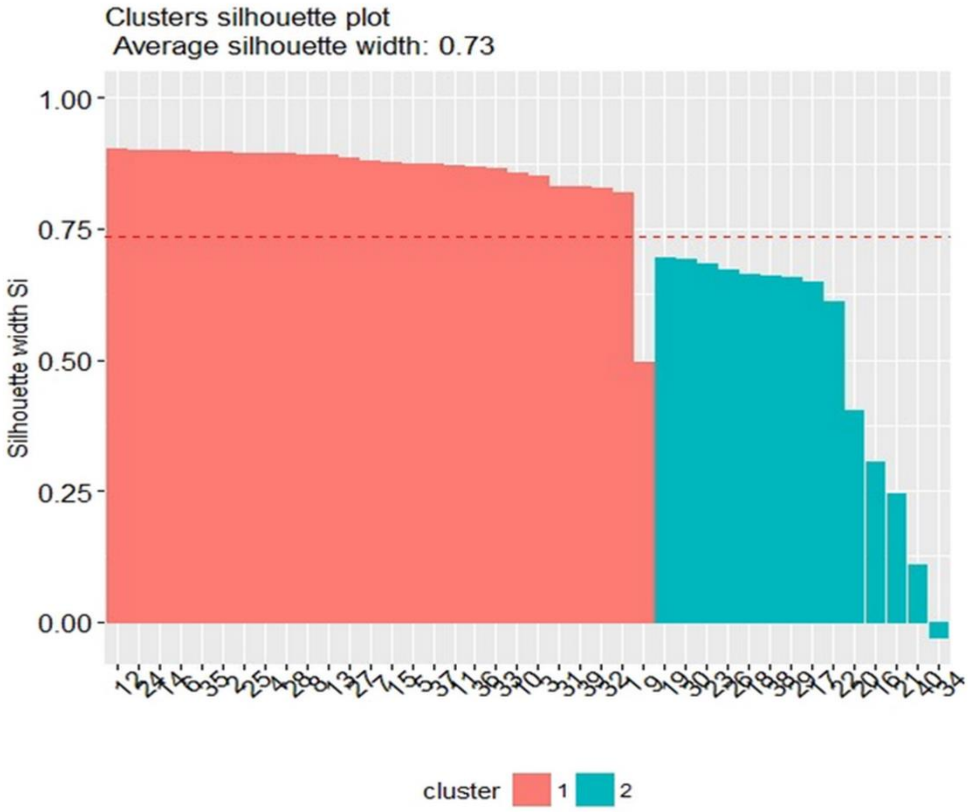

Figure 7. Silhouette graph of all common and organic olive oil samples. Font: Cunha Filho, M. (2919).

The observations are well grouped into their respective groups. We note that the group of common oils are the ones that are best grouped because their values are all positive. It was evidenced that the silhouette statistics verified the analyses in the organic grouping, but they are seen like samples coming from the common olive oil. The samples' behavior within each cluster, one to sample 40 in cluster 2, has a negative coefficient indicating that it is not well allocated within the cluster of organic oils.

According to Table 1, used for comparison of the groups, it consistently analyzed the olive oil samples, corroborating the verification of the relevance and similarity of each sample to its particular group.

Table 1. Group comparison similarity of samples with all common and organic olive oil samples. Font: Cunha Filho, M. (2919).

Samples	Common	Organic
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%

Considering Table 1 it was confirmed the similarity of the samples with the groups and between groups, we note that some samples have marked characteristics for both groups, for example, samples 40 and 34, by the Fuzzy grouping both are separated with more remarkable similarity for the group. However, we can see from the group comparison table that the sample's similarity between the groups is quite pronounced. This fact may be directly correlated with the way the samples were diluted for the analysis.

Conclusions

Fuzzy's non-hierarchical clustering technique distinguishes between extra virgin olive oil groups, with about 65% confidence. The clusters' quality was attributed through the silhouette statistics index with 0.73 s (i), indicating the strength and power of distinction in the clusters. Using the PCA technique can verify data from extra virgin olive oil but does not explicitly separate the oils into organic and common.

References

Alvarenga, A. A.; Cruz, J. L.; Oliveira, A. F.; Silva, L. F. D. O.; Gonçalves, E. D.; Norberto, P. M. 2017. Nutritional Quality of Olives and Olive oil Produced in the Serra da Mantiqueira from Brazil. Agricultural Sciences, 8, (7), 518.

Ferreira, E. D. S.; Silveira, C. D. S., Lucien, V. G., Amaral, A. S. 2009. Caracterização físico-química da amêndoa, torta e composição dos ácidos graxos majoritários do óleo bruto da castanha-do-brasil (*Bertholletia excelsa* HBK). Alimentos e Nutrição Araraquara, 17, (2), 203-208.

Gonçalves, R. P.; Março, P. H.; Valderrama, P. 2015. Thermal degradation of tocopherol and oxidation products in different olive oil classes using UV-Vis spectroscopy and MCR-ALS. Química Nova, 38, (6), 864-867.

Graybill, F. 1983. Matrices with applications in statistics. Principles and procedures of statistics, 2. ed., 220p.

Lay, D. C. 2007. Álgebra Linear e suas Aplicações. 2. ed. Rio de Janeiro: LTC. 250p.

Mardia, K. V.; Kent, J. T.; Bibby, J. M. 1979. Multivariate analysis. London: Academic, 512p.

Meira, M.; Quintella, C. M.; Ferrer, T. M.; Silva, H. G.; Guimarães, A. K; Santos, M A; Costa Neto, P.R; Pepe, I. M. 2011. Identificação de adulteração de biocombustível por adição de óleo residual ao diesel por espectrofluorimetria total 3D e análise das componentes principais. Quimica Nova, 34, (1), 621-624.

Morais, J. T.; Esquerre, K. P.; Kiperstok, A.; Queiroz, L. M. 2016. Prediction of organic matter removal from pulp and paper mill wastewater using an artificial neural network. Desalination and Water Treatment, 57, (57), 27969-27977.

Oliveira, M. R. G.; Cruz, D.V.; Cunha Filho, M. 2016. Mapping plaques Cisterns by Fuzzy grouping analysis. IEEE Latin America Transactions, 14, (10), 4367-4372.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal Computational Applied Mathematics, 20, (1), 53-65.

Vale, M. N. 2005. Agrupamentos de dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos. Rio de Janeiro, Dissertação (Mestrado), Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro. 120p.

Vogel, P.; Machado, I. K.; Garavaglia, J.; Zani, V. T.; Souza, D.; Dal Bosco, S. M. 2015. Polyphenols benefits of olive leaf (*Olea europaea* L) to human health. Nutrición Hospitalaria, 31, (3), 1427-1433.

Xu, J.; Cai, S.; Li, X.; Dong, J.; Ding, J.; Chen, Z. 2012a. Statistical two-dimensional correlation spectroscopy of urine and serum from metabolomics data. Chemometrics and Intelligent Laboratory Systems, 112, (1), 33-40.

Xu, R.; Xu, J.; Wunsch, D. C. 2012b. A comparison study of validity indices on swarm-intelligence-based clustering. IEEE Transactions on Neural Networks, 16, (3), 645-678.

Zhu, W.; Su, S.; Shou, Z. 2017. Social ties and firm performance: The mediating effect of adaptive capability and supplier opportunism. Journal of Business Research, 78, (1), 226-232.