Volatile profiles of green Spanish-style table olives: application of compositional data analysis for the segregation of their cultivars and production areas.

A. Garrido-Fernández, A. Montaña, A. H. Sánchez-Gómez, A. Cortés-Delgado, A. López-López*

Departamento de Biotecnología de Alimentos, Instituto de la Grasa (CSIC), Campus Universitario Pablo de Olavide, Edificio 46, Ctra. Utrera km 1, 41013 Sevilla, Spain.

*Corresponding author:

Antonio López López

Tel.: +34 954 611550

e-mail: all@cica.es

Abstract

The work presents the application of compositional data methodology to analytical results, taking as an example the study of the volatile profiles of green Spanish-style table olive according to cultivars and production areas. For this purpose, the volatile compounds (analysed by GC-MS and expressed as percentages of the total area) were considered as a compositional data set in the Simplex space and, as a result, analyzed by their specific new statistical tools. Application of compositional exploratory tools (variation array, tertiary graphs, biplots, or coda-dendrogram) allowed differentiating cultivars and production areas based on their volatile profiles. Also, the application of Cluster and Principal Component analysis to the *ilr* transformed values (coordinates), following the new methodology, led to more realistic results than the formally incorrect implementation of the standard multivariate analysis (developed for data from the Euclidean space) to percentages (data in the Simplex). Therefore, the work presents a novel consideration of the volatile profiles of table olives as compositional data and shows their proper analysis by statistical tools specifically developed for them.

Keywords: Green table olives, Volatile profiles, Fermentation, Compositional data analysis, Multivariate analysis.

1. Introduction

Table olives are the largest fermented vegetable in western countries. The global production was 2,661,000 tonnes in 2013/2014 season [1]. The most popular are the green Spanish-style table olives; its fermentation is carried out in big containers (16,000 L volume), but the process remains spontaneous [2]. This circumstance leads to marked differences among processed olives from different cultivars, origins or even between factories within the same villages. The IOC has developed a methodology for the sensory characterization of products from this style [3, 4]. However, the evaluation is limited to the perception of some defects, related to volatile compounds, which only are used for the classification of the product into categories while the assessment of gustative and kinesthetic sensations, neither of them related to volatile compounds, do not play any specific role in table olive valorization. As a result, the use of sensory analysis to green table olives is still far from the decisive application that the sensory analysis has in olive oil [5].

Also, the processed olives may be characterized by their volatile profiles. Several works have been devoted to studying them in the case of Spanish-style processing [6-11]. However, these works have not led yet to an appropriate definition of the green Spanish-style table olive fermentation profile and the volatile compounds produced from diverse cultivars and factories can be quite different due to the spontaneous fermentation [2].

The application of multivariate techniques to segregate among volatile profiles of various types of table olives, cultivars, processing conditions or starter cultures is relatively frequent [11-17]. However, the multivariate techniques used in these cases were developed for data from the Euclidean space, and their application to volatile

compound data (usually expressed as a percentage of the total area of the peaks) causes problems due to the correlation among variables and the generation of singular matrices during statistical analysis. The first to be aware of these problems was Pearson [18], who pointed out the dangers that could be derived from the interpretation of correlations between ratios whose numerators and denominators have common parts. Then, the analysis of compositional data, with their concentrations based on ratios between the proportion of the same whole, is likely to be fraught with difficulty. The special and intrinsic feature of a compositional data is that the proportions of a composition are naturally subjected to a constant-sum (1, 100 or any other value). However, this constraint “has been widely ignored or wised away and inappropriate “standard” statistical methods, devised for and successfully applied to unconstraint data, have been applied with disastrous results” [19]. The need of new statistical procedures for the analysis of compositional data is then evident. Aitchison was the first to study in detail the problem and develop methodologies that could be applied properly to them [19]. Raw chromatographic data and data processed by central log ratio transformation (*clr*) were applied to differentiate and characterize Arabica and Robusta coffee types by application of cluster and PCA analysis to volatile compounds from 30 commercially available coffee samples [20]. Also, robust compositional biplot was used for the resolution of particular categories of wines based on eight phenolic acids [21].

The aim of this work was investigating the application of compositional data analysis to study the effect of cultivars and production area on the volatile profiles of green Spanish-style table olive fermentations. The results obtained with the new statistical tools are compared with those deduced by implementing the classical

multivariate statistical approach (developed for data from the Euclidean space) directly to percentages (Simplex space).

2. Material and Methods

2.1. Experimental design

The experiment was carried out with Manzanilla and Hojiblanca, the main cultivars used for green Spanish-style table olive processing [22]. Manzanilla olives were from Alcalá de Guadaira (Seville), Posadas (Córdoba), and Almendralejo (Badajoz) and were coded as MAL, MC, and MAM, respectively. The Hojiblanca olives were from Alameda (Málaga), Estepa (Seville), and Casariche (Seville) and were coded as HA, HE, and HCS, respectively. The process took place in polyethylene vessels, containing 5.2 kg fruits and 3.4 L successive liquid solutions. Initially, fresh olives were treated with a 2.0 % w/v NaOH solution, until the lye penetrated two-thirds of the flesh, with the objective of hydrolysing the olive bitter glucoside oleuropein [2]. Then, olives were washed to remove the NaOH excess and, finally, the olives were covered with brine (11.4% NaCl) and were let to follow the typical (spontaneous) green Spanish-style table olive fermentation for 5 months. The process took place at room temperature, in the Department's pilot plant. The treatments were conducted in duplicate, except for sample HCS (due to fruit shortage). In the case of Hojiblanca cultivar, the pH at the end of the fermentation was corrected to ≤ 4.0 with lactic acid, to assure the product stability during the storage.

2.2. Analysis of volatiles

The volatile composition was quantified in triplicate by headspace solid phase microextraction (HS-SPME) followed by gas chromatography coupled to mass spectrometry (GC-MS), according to the method of Cortés-Delgado et al. [11]. Briefly, extraction was carried out using a 1-cm-long divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) fiber (Supelco, Bellefonte, PA). SPME was performed in 15 mL vials, sealed with PTFE-silicone septa, containing 2.5 g of homogenized olive pulp and 7.5 mL of 30% (w/v) NaCl with magnetic stirring. Extraction was performed at 60 °C for 60 min. All GC-MS analyses were performed using an Agilent 7890A gas chromatograph coupled to an Agilent 5975C mass selective detector (Agilent Technologies, Santa Clara, CA). The carrier gas was helium at a constant flow of 1 mL min⁻¹. The components were separated on a VF-WAX MS capillary column (30 m x 0.25 mm x 0.25 µm film thickness) from Agilent. The injections were performed in the splitless mode. The initial oven temperature was set at 40 °C for 5 min; then, it was increased at 3 °C/min to 195 °C, at 10 °C/min to 240 °C and, finally, held there for 15 min. The injector temperature was set at 250 °C and the desorption time of the fiber in the injector was 15 min. For the mass selective detector conditions, the quadrupole, ion source, and transfer line temperatures were maintained at 150, 230, and 250 °C, respectively. GC-MS detection was carried out in the complete scanning mode in the 40-400 amu mass range. The GC peak area of each compound was obtained from the ion extraction chromatogram (IEC) by selecting target ions for each one. These ions corresponded to the base ion (m/z 100% intensity), molecular ion (M⁺) or another characteristic ion for each molecule. Hence, some peaks that could be co-eluted in TIC mode can be integrated with a value of resolution greater than 1. Results for each volatile were expressed as a percentage of its peak area on the total. Compounds were identified by comparing their mass spectra, retention times and linear

retention indices (RI) against those obtained from pure standards when available and comparison of the mass spectrum with those of the NIST 08 MS library. For the determination of the RI, a C7-C30 *n*-alkanes series was used, and the values were compared, when available, with those reported in the literature for similar chromatographic columns.

2.3. *Statistical analysis*

A data set is called compositional if it provides portions of a total (here, peak areas). The individual parts of the composition are called components, which amounts represent their importance within the whole. This consideration implies that compositions are multivariate by nature, a characteristic that represents a substantial difference between them and other multivariate datasets. This means that each compositional marginal variable has not meaning on itself; that is, isolated from the rest. Therefore, any previous selection of marginal variables, as usually most multivariate analyses begin, is not of application in compositional data analysis. Also, in this case, the standard descriptive statistics are not very informative because the arithmetic mean and the variance (or standard deviation) of individual components, as measures of central tendency and dispersion respectively, do not fit with the Aitchison geometry [23]. The alternatives for compositional data are found in the concepts of centre, variation matrix, and total variance [19]. Another essential exploratory tool is a particular biplot, developed for these data by Aitchison [24], which, due to the nature of the data, has a particular interpretation as will be commented later.

Aitchison [19], guided by the principle that the interest in compositional data lies on the relative proportions of the components, developed a series of transformations based on ratios with the idea of applying the classical statistics to them. Among the first

ones are the additive log-ratio (*alr*) and the centred log ratio (*clr*) transformations. The former does not preserve distances while the second leads to singular matrices. Therefore, neither of them can be directly associated with an orthogonal coordinate system. However, Egozcoue et al. [25] defined the isometric log ratio transformation (*ilr*). It allows the association of compositions with coordinates in an orthonormal system and their values may be obtained by a sequential binary partition (balance). Usually, based on either preceding descriptive analyses or priori knowledge, the use of a specific sequential binary partition may result particularly attractive for the selection of specific balances. Its graphical presentation is a peculiar coda-dendrogram, which represents the partitions in the form of a tree, including the sample mean and the variance of each balance as well as box-plots summarizing the order of their statistics [25, 26]. Besides, the main advantage of this transformation (*vs. alr* and *clr*) is that its associated orthonormal coordinates are vectors of real variables in the Euclidean space that can be treated with the existing battery of conventional descriptive analysis. In any case, all transformations (*alr*, *clr*, and *ilr*) are useful because of their particular convenience for specific tasks. In this work, the standard Cluster and Principal Component analysis have been applied to the *ilr* transformed values (coordinates) to segregate treatments (combinations of cultivars and production areas) according to their volatile profiles. Also, the results obtained by this compositional strategy have been compared to those based on the direct application of the standard multivariate techniques to the original data (percentages in the Simplex), as it is usually the case.

Statistical analyses were performed using CoDaPack v. 2.01.14 (Department of Computer Science and Applied Mathematics, University of Girona, Spain) and XLSTAT 2014 (Addinsoft, Paris, France).

3. Results and discussion

Each cell in the dataset was the average of three analysis from each replicate treatment. In a first overview, the volatile compounds present in very low proportion and sporadically found were discarded from the study (26 (isobutyl acetate), 28 (ethyl butanoate), 41 (isoamyl lactate), and 72 ((+)-cycloisositivene) although preserving the original number assignation. Despite that, the number of components retained (98 compounds) in the depurated dataset was large (Table 1S). In this work, their identification was mainly achieved by their order numbers, except for those playing marked roles in the statistical analysis that are identified by names. In any case, the correspondence between numbers and names is shown in Table 1.

Regardless of the depuration process, some compounds were still absent from one or several cells. In the standard multivariate analysis they were considered as absent (zero concentration); however, in compositional data analysis, missing values, which appear quite often, is one of the biggest problems to deal with. To prevent misinterpretations, the kind of missing values observed (completely at random (MAR), structural zero (SZ), below the detection limit (BDL), true zero (TZ), and missing not at random (MNAR)) should be clearly defined, with their own inner logic. Then, they should be treated properly. In analytical chemistry, the most common situation is observed when some compounds (variables) cannot be estimated because, apparently, the actual presence is below the detection limit of the measurement procedure. Therefore, the missing volatile values in this work can clearly be defined as BDL (that is, compounds which were in such so small amounts that the measurement method was not sensitive enough to distinguish the true value from zero). The suggested methodology to solve the problem, in this case, is replacing them by some fraction of

the detection limit. In this way, to prepare the dataset importation into CoDaPack, the zeros were substituted with <0.001 (detection limit). (Table S2). After importing the dataset, the program substitutes the <0.001 symbol by a figure which is usually 65% of detection limits; that is, the levels of the absent compounds are set at 0.00065. As the introduction of this value in the cells of BDT introduces a disturbance in the original matrix (in the sense that their rows no further sum 100), the program re-estimate the matrix data so that all rows sum again 100 (Table S3). After this pre-treatment, the data were ready to be analysed by the methods specifically developed for compositional data analysis (that is, working in the Simplex) or be transformed into coordinates in the Euclidean space (see later) for the application of the conventional multivariate methods.

Summary statistics in compositional data are usually related to centre values, the corresponding quartiles, and the variation array. The centre is defined as the geometric mean of values over columns (variables). The estimations would be done, according to groups (Table S4) or over groups, with only one value per variable (Table S5). In general, central data for variables split into groups were relatively low due to the presence in some samples of missing values, which were assigned 65% detection limit value (Table S4). Also, the boundaries for the different quartiles for such variables were rather narrow due to overall close values obtained in each replicate (Table S4). In the case of overall estimations (without segregating by samples) the centre values were also low, but limits for the quartiles were wider due to the relatively higher variability among treatments (Table S5). The central values and quartiles have limited roles in exploratory compositional data analysis, but the traditional average and standard deviations (or variances) have not particular meaning either as they were not developed for data from the Simplex [27].

In the case of application of the standard multivariate techniques to the original data, one common technique is the application of one-way ANOVA to remove from further analyses the variables without significant differences among treatments. However, in the compositional analysis, each variable has not any meaning individually, but the importance is relative to the presence of other compounds, as established by Aitchison [19]. Therefore, in this work, to make the studies comparable, the total numbers of compounds (variables) retained after the data pre-treatment were used for the statistical analysis of both compositional data and percentages (by standard techniques) [28].

3.1. Variation array

It is a method for the estimation of the variances of the log ratios of each variable on the others. Its definition and properties can be found elsewhere [27]. It is a symmetric matrix since $\log(a/b)=-\log(b/a)$ and $\text{var}(-c)=\text{var}(c)$ [28]. The variation array matrix may also be calculated according to groups (Table S4, variation arrays) or for the overall data matrix (Table S5, variation array). In sample MC, for example, the variance for $\log(2 \text{ (order number) or (ethanol)/1 (order number) or (isopropanol)})$, considering the two replicates, is 0.0003, which means that the values obtained in the two replicates of this treatment were fairly close. On the contrary, in sample MAL, the variance of $\log(4 \text{ (1-propanol)/1 (isopropanol)})$ is rather important (7.2879), indicating a great variability between their replicates (Table S4, variation array). When only one replicate was available (HSC), the estimation of the within treatment variance is not possible (the returns for the variation array are just zeros). In general, the within treatment variances were low and had a reduced interest *per se*, although they might also play an essential role for estimating differences between treatments. More interesting are the dispersions

over groups of each variable on each one of the rest because they allow identifying those that show high variances among treatments and such large log ratio variances could be useful for segregating among treatments. For example, it is observed that the variance of the log ratio (2 (ethanol)/1 (isopropanol)) over all treatments (and their replicates) is rather large (17.0304) (Table S5, variation array); on the contrary, the levels obtained within any of the six treatments (Table S4, variation arrays for the different treatments) is moderate. The high variance indicates that this log ratio may differ between cultivars and production areas and could be useful for finding significant differences between treatments. Furthermore, the variance of log (5 (isobutanol)/1 (isopropanol)) is still larger (33.7461) (Table S5, variation array) and could even be more interesting. In this way, just by observing the variances of the various log ratios one may deduce those that could be more interesting for differentiating among treatments. The great advantage of this variation array matrix is its detailed description of the variances among the several possible log ratios. In CoDaPack, the greater values are usually highlighted in progressive stronger tones to facilitate its localisation.

However, one may be more interested in the identification of those variables with great overall variances, summarised in just one figure. This is usually estimated as *clr* variances, a transformation which preserves distances among treatments. In this way, you may observe the variance of each variable over the rest and throughout treatments [29]. Those variables with the greatest overall variances will be more prone to segregate among treatments (Table S5). The 10 greatest *clr* variances for the variables studied are shown in Table 2. It may be observed that volatile compound 51 (propanoic acid), 4 (1-propanol), and 1 (isopropanol) have the highest overall variability and, consequently, they may be useful for segregating among cultivars and production areas. Contributions

of 64 (2-heptenal) and 23 (propyl acetate) are also important while those of 68 ((E)-2-decenal), 32 (methyl hexanoate), 15 (1-heptanol), 5 (isobutanol), and 6 (1-butanol) have progressive lower contributions. These variables will then play an essential role in studying the possible differences in the volatile profiles among treatments. Finally, a measure of the global dispersion is the total variance, (sum of the *clr* variances), which definition and properties may be found elsewhere [27]. This total variance summarises the variation matrix in a single quantity (Table S5, total *clr* variance), and it is possible to define it because all parts in a composition share a common scale. In this work, total variation is moderate and advances that the differences among treatments exist but are relatively reduced.

3.2. Graphical presentation of volatiles as compositional data

For a characterization of the sample space of compositional data, named after its definition as the Aitchison space, readers are referred to Aitchison [19] and Pawlowsky-Glahn et al. [27]. In practice, the representation may be achieved in the Simplex, which takes into consideration the constraint imposed by the constant sum of all parts. Obviously, the best segregation among green Spanish-style table olive volatile profiles, according to their origins, should be achieved as a function of those variables with the highest variances. As the maximum numbers of axes (variables) that can be visualized are four, the volatiles identified by 1 (isopropanol), 4 (1-propanol), 51 (propanoic acid), and 64 (2-heptenal) were chosen (because of their high *clr* variances). Usually, the samples are presented as a function of centred values, particularly when some samples are situated close to the vertexes or axes, as is the case. The tertiary graph shows an important dispersion among treatments (cultivars and production areas) (Figure 1). Then, differences in the volatile profiles among the fermentation processes of green

Spanish-style table olives carried out with fruits from diverse cultivars and production areas may be expected. It is possible to estimate the Principal Components based on these variables and represent the new axes in the Simplex (Figure 1). The estimation may be useful for detecting possible linear relationships among treatments, which will then be aligned along the corresponding Principal Component axes; although, apparently, such situation is not observed in the present experiment.

3.3. Exploring relationships between volatile compounds and treatments

The biplot adapted to compositional data by Aitchison and Greenacre [24] is a powerful exploratory tool. The covariance biplot is particularly appropriate to study the relationship among variables (volatile compounds) (Figure 2). As usual, the angle between the volatiles may give an overview of the correlation structure. In this way, variables 102 (1,4-dimethoxybenzene), 101 (butyrolactone), 6 (1-butanol), 84 (phenol) or 78 (α -muurolene) follow all of them a similar direction; so they are strongly related and give redundant information. Similarly, 43 (ethyl decanoate) and 5 (isobutanol), 68 ((E)-2-decenal) and 32 (methyl hexanoate), 15 (1-heptanol) and 62 (heptanal) form angle close to 0, which cosine is 1, and are consequently strongly related. The length of the rays can usually be interpreted with caution when the comparisons include sub-compositions because their values depend on each sub-composition throughout their geometric mean (which may differ between the full and any considered sub-composition) [27]. However, for a particular composition, the length of the rays may be related to their variances. Hence, the length of the rays (Figure 2) of the diverse variables are clearly linked to the corresponding estimated variance values. Those with the highest variances (Table 2) stand out from those with low values which are piled up around the centre (Figure 2). Furthermore, some of the variables with the highest

variability among treatments are, at least in part, redundant since the angle between them is close to 0 (cosine 1). This is the case of 1 (isopropanol) vs. 23 (propyl acetate); 4 (1-propanol) vs. 51 (propanoic acid); 32 (methyl hexanoate) vs. 68 ((E)-2-decenal); or 64 (2-heptenal) vs. 5 (isobutanol). Then, as far as for the segregation properties is concerned, they supply redundant information, and each coupled could be substituted just by one of them. However, as in compositional data analysis, no variable has a particular meaning by itself [28], the whole data set was retained for the analysis. Furthermore, their simultaneous presence may also imply important biochemical relationships.

Of particular interest in the interpretation of the compositional biplot are the links between variables, measured as the distance between the extremes of the respective rays, which may be related to the variances between their log ratios [27]. It is apparent (Figure 2), in agreement with the high correlation between them, the relatively low variability between 4 (1-propanol) and 51 (propanoic acid) (variance $\log(4/51)$ close to zero). This means that the ratio 4/51 is almost constant between treatments (that is, for all olive cultivars and production areas). Further studies on the pathways of formation of such compounds may help for deeper insights into the biochemistry of the fermentation process and the formation of volatile compounds. Similar situations are also observed for 1 (isopropanol) vs. 23 (propyl acetate), 32 (methyl hexanoate) vs. 68 ((E)-2-decenal), or 5 (isobutanol) vs. 64 (2-heptenal). Therefore, their log ratios may be constant regardless of olive origin and may be a characteristic of the process. On the contrary, those volatile compounds forming an angle close to 90° ($\cos 90^\circ = 0$) in the biplot should be considered unrelated. This condition is observed in 1 (isopropanol) and 23 (propyl acetate) vs. 32 (methyl hexanoate) and 68 ((E)-2-decenal) or these vs. 5

(isobutanol) and 64 (2-heptenal); then, the evolution of those non-related compounds may follow completely independent metabolic pathways. Finally, an opposed collinearity between vertices may also be observed, indicating a negative relationship among them. It may show that one of them may be transformed into the another during the fermentation. Therefore, confronting these relationships with the expected metabolic pathways of the different compounds could be useful for better understanding the fermentation process of the green Spanish-style table olives.

The biplot based on the projection of variables (or cases, as commented below) onto the plane of the first two Principal Components is habitual. It represents most of the variability but, sometimes, when there are not well explained (close to the origin) variables, the projection onto other components may be useful. However, in this analysis, the projections of the volatile profiles onto the PC2 (20.03% variance) *vs.* PC3 (9.53%) and PC1 (49.13%) *vs.* PC3 (9.53%) planes did not contribute to new relevant additional information of interest.

The scores of the samples can also be projected onto the covariance compositional biplot, and their relative positions observed (Figure 2). The samples appeared well distributed around the biplot with the replicate samples close together. The form biplot is usually suggested as the most appropriate presentation for the interpretation of the distance between samples (Figure 3). The treatments were also well distributed around the biplot, with HE in the first quadrant, HA and HCS in the second, MC and MAM in the third, and MAL in the fourth, However, in this analysis, the graphic showed a general trend for situating the replicates closer than in the covariance biplot and grouped around the centre, although their visibilities are reduced due to the cloud of variables in this zone. Therefore, regardless of the type of biplot, the treatments

are well segregated, indicating that the cultivar and the production areas of the fruits may lead to fermented products with different volatile profiles. No improvement on information could either be observed projecting the cases onto the other PCs axes.

3.4. Dendrogram and sequential binary partition

As commented above, a particular kind of coordinates are the balances, which are useful for modeling and interpretation [27]. The great interest of them is based on the fact that such coordinates may be subjected to the complete list of multivariate standard statistical tools already available. In general, the procedure for obtaining the coordinates is based on a sequential binary partition, which can be considered as a hierarchy of the parts of a composition. The balance for each order of partition is defined as $b = (rs/(r+s))^{0.5} \cdot \ln[(x_{i1} \cdot x_{i2} \cdot \dots \cdot x_{ir})^{1/r} / (x_{j1} \cdot x_{j2} \cdot \dots \cdot x_{js})^{1/s}]$. In the expression, r and s stand for the number of terms in the numerator and denominator, respectively, x_{i1} to x_{ir} the values of the variables in the numerator, and x_{j1} to x_{js} those in the denominator. The b value may then be considered as the normalized (by the first term of the formula) log ratio of the geometric mean of each group of parts. For a detailed definition, selection of orthonormal bases (axes), and estimations of coordinates, readers are referred to Pawlowsky-Glahn et al. [30].

When one has previous information on the problem, the sequence of balances may incorporate the variances of the variables in decreasing order and, usually, most of the variance is retained in the first balances. The variables with the highest variances among treatments can be selected from the variation array (Table S5). The first ten variables with the highest variances (Table 2) could then be good candidates for leading the successive balances. In this way, the first one will be established between 51 (propanoic acid) in the numerator and the rest of variables as the group in the

denominator: $b1=(1*97/(1+97))^{0.5} \cdot \ln (v51^{1/1}/(v1*v2..v102)^{1/97})$. In the expression, v51 and v1 to v102 (except v51) stand for the values of their respective variables. The second balance will then be: $b2=(1*96/(1+96))^{0.5} \cdot \ln (v4^{1/1}/(v1*v2..*v102)^{1/96})$. Therefore, b2 will be calculated in a similar way, using 4 (1-propanol) as the numerator and the rest of variables, except 51 (propanoic acid) already used in the first balance and v4 (1-propanol), in the denominator. Once established the sequence of binary partitions, the successive balances (and their variances among treatments) are estimated directly by any of the available programs (CoDaPack, in this case).

The interpretation of the balances (Table 2, ten first balances and variances) is based on their properties. The first two balances are positive (that is the numerator of the log ratio is greater than the denominator). As a result, overall, the values of 51(propanoic acid) and 4 (1-propanol) are higher than the geometric means of the respective groups over which the balance is estimated. On the contrary, the balance for 1 (isopropanol) is negative; that is, the isopropanol content value is lower than the geometric mean of the remaining (not used in the numerator) volatile compounds. In agreement with the criterion followed for the selection of variables, the first three balances also account for the decreasing greatest variances and continue declining as the volatile compounds placed in the numerator had lower variances among treatments (except in the case of alteration of the variable variance order). The sequential binary partition may be outlined by a matrix of 1, -1, and 0, respectively, where the 1 means that the variable involved belongs to the group of the numerator, the -1 that form part of the group placed as denominator, and 0 that the variable is not included in either of the two previous groups (Table S6).

In this case, the interpretation of the balances is of relative interest because the objective of the work is detecting differences in the volatile profiles, according to cultivars and production areas. However, the decrease/increase of a specific group of compounds initially present could be related to the formation/disappearance of another one/s. The compositional analysis allows the comparison between any groups of variables (sub-compositions) one may be interested.

The coda-dendrogram (Figure 4) represents a convenient graphical presentation of the balances and their variances. The intersection of the vertical bar (variances) with the horizontal line for each group indicates the balance value. When the balance is displaced to the right, the weight of the group in the numerator is greater than that in the denominator and vice-versa when is displaced towards the left. The length of the vertical line represents the variance of the balance, which lengths decline sharply as the sequence of the balances is progressing (increase the order) and the variable in the numerator has less variance (Figure 4). In the case of groups, each of them has its own vertical bar, although they may overlap because of the similarity of the balances for the groups (Figure 4). In addition, the box below each balance (for each group or overall) represents a whisker (median and 0, 25, 75, and 100 quartiles).

In this work, the main interest was focused on the values of the *ilr* transformed values (or coordinates in the Euclidean space) because they can be subjected to the classical multivariate techniques (Table S7). Interestingly, the coordinates are not systematically positive anymore but include both positive and negative values, as it is usual in the Euclidean space. Therefore, the *ilr* coordinates (number of compositional variables minus 1) for each sample have the properties of being coordinates in orthonormal axes and are, therefore, able to be subjected to the habitual multivariate

techniques. Besides, the transformation has the property of preserving distances, which is of particular convenience in this study, designed for segregating between the volatile profiles of green Spanish-style table olive fermentations according to cultivars and production areas. The *ilr* transformed data (coordinates in the Euclidean space) were then subjected to cluster and PCA analysis.

3.5. Working in coordinates

The *ilr* transformed data (Table S8) (coordinates on an orthonormal basis) can be submitted to the battery of standard multivariate statistical methods. A first approach may be achieved by clustering using the Euclidean distance for dissimilarity and the Ward agglomeration method. After the analysis, there was proper segregation between treatments with the only exception of one MAM sample, included into the MC treatment (Figure 5, panel A). However, when the same analysis was applied to the original data expressed in percentages, the samples were classified into three groups (Figure 5, panel B). Only HE and MAL were correctly segregated while the third group included HA, MC, MAM and HSC. This means that using percentages, most of the samples were improperly classified into the same cluster, regardless of their different cultivars or production areas. Therefore, the application of the cluster analysis to the *ilr* transformed data or coordinates (estimated according to the compositional data analysis methodology) led to a more realistic grouping of treatments than when the same multivariate technique was directly applied to the volatile profiles expressed as percentages.

A typical method for segregating treatments is by projecting the cases' scores onto the first two PCs. Following the compositional methodology, the PC analysis was also applied to coordinates (*ilr* transformed data), and its result presented graphically

(Figure 6, panels A and B). The samples were grouped according to the cultivars and production areas with clear differences between HE with respect to the other treatments (Figure 6, panel A). Only HSC was relatively close to HA replicates. Also, in general, there were limited distances between replicates. However, when the PCA analysis was applied to the original set of data (percentages), the projections of the cases' scores on the plane of the first two PCs (Figure 6, panel B), showed greater distances between replicates within treatments in the case of MAL and HA than using *ilr* transformed data. Besides, the sample HCS was rather close to one of the HA replicates, while the another HA replicate was relatively far from both. Therefore, the PCA application to the *ilr* transformed data or coordinates (following the compositional data methodology) produced better segregation among treatments (combinations of cultivars and production areas) than when used with the original data expressed in percentages.

4. Conclusions

The work presents a novel methodology for the study of analytical results as compositional data and the application of their specific statistical tools. Those volatile compounds with the highest variances, as shown in the variance array, were the most appropriate candidates for segregating among treatments. In fact, the ternary presentation of the samples in the Simplex (the natural space of the compositional data) as a function of only the first four variables with the highest variance, clearly separated the samples. Also, the covariance, form biplots, and coda-dendrogram, exploratory graphs in the new statistical methodology, segregated samples properly. The results obtained by cluster and PC analysis, following the compositional data methodology (using *ilr* transformed or coordinates), were more truthful than those achieved by the formally incorrect application of the same multivariate techniques (developed for the

Euclidean space) directly to the original data in percentages. Therefore, the work opens the possibility of implementing the complete set of new statistical tools developed for compositional data to properly treating not only the volatile profiles of green Spanish-style table olives but, in general, of any food.

Acknowledgement

This work was supported by the Ministry of Economy and Competitiveness from the Spanish government through Project AGL2014-54048-R, partially financed by the European Regional Development Fund (ERDF).

References

- [1] International Olive Council (IOC), Key figures on the world market for table olives. <http://www.internationaloliveoil.org/estaticos/view/134-approved-balances>, 2015 (accessed 15.04.16).
- [2] A. Garrido-Fernández, M. J. Fernández-Díez, R. M. Adams, Table olives. Production and processing, Chapman and Hall, London, 1997.
- [3] IOC (International Olive Council), Sensory analysis of table olives. COI/OT/Nº1/Rev 1 February 2010, IOOC, Madrid, 2010.
- [4] IOC (International Olive Council), Guidelines for tasters and panel leader training in the sensory assessment of table olives and panel management according to standard COI/OT/MO Doc. Nº1 2010. T.OT/GFMO November 2010, IOOC, Madrid, 2010.
- [5] International Olive Council (IOC), Collection of methods and standards adopted by the international olive council for sensory analysis of olive oils, olive-pomace oils

and table olives, <http://www.internationaloliveoil.org/estaticos/view/224-testing-methods>, 2015 (accessed 17.05. 16).

- [6] H.P. Fleming, J.L. Etchells, T.A. Bell, Vapor analysis of fermented Spanish-type green olives by gas chromatography, *J. Food Sci.* 34 (1969) 419-422.
- [7] A. Montaña, A.H. Sánchez, L. Rejano, Rapid quantitative analysis of headspace components of green olive brine, *J. Chromatogr.* 521 (1990) 153-157.
- [8] R. Iraqi, C. Vermeulen, A. Benzekri, A. Bouseta, S. Collin, Screening for key odorants in Moroccan green olives by gas chromatography-olfactometry/aroma extract dilution analysis, *J. Agric. Food Chem.* 53 (2005) 1179-1184.
- [9] N. Sabatini, V. Marsilio, Volatile compounds in table olives (*Olea Europaea* L., Nocellara del Belice cultivar), *Food Chem.* 107 (2008) 1522-1528.
- [10] M. Cano-Lamadrid, I.F. Giron, R. Pleite, F. Burlo, M. Corell, A. Moriana, A.A. Carbonell-Barrachina, *LWT- Food Sci. Technol.* 62 (2015) 19-26.
- [11] A. Cortés-Delgado, A.H. Sánchez, A. de Castro, A. López-López, V.M. Beato, A. Montaña, Volatile profile of Spanish-style Green table olives prepared from different cultivars grown at different locations, *Food Res. Int.* 83 (2016) 131-142.
- [12] R. Malheiro, P.G. de Pinho, S. Casal, A. Bento, J.A. Pereira, Determination of the volatile profile of stoned table olives from different varieties by using HS-SPME and GC/IT-MS, *J. Sci. Food Agric.* 91 (2011) 1693-1701.
- [13] S. Dabbou, M. Issaoui, F. Brahmi, A. Nakbi, H. Chehab, B. Mechri, M. Hammami, Changes in volatile compounds during processing of Tunisian-Style table olives, *J. Am. Oil Chemists' Soc.* 89 (2012) 347-354.

- [14] G. Bleve, M. Tufariello, M. Durante, E. Perbellini, F.A. Ramires, F. Grieco, M.S. Cappello, S. De Domenico, G. Mita, M. Tasioula-Margari, A.F. Logrieco, Physico-chemical and microbiological characterization of spontaneous fermentation of Cellina di Nardò and Leccino table olives, *Front. Microbiol.* 5 (2014) 570.
- [15] C.L. Randazzo, A. Todaro, A. Pino, I. Pitino, O. Corona, A. Mazzaglia, C. Caggia, Giarraffa and Grossa di Spagna naturally fermented table olives: Effect of starter and probiotic cultures on chemical, microbiological and sensory traits, *Food Res. Int.* 62 (2014) 1154-1164.
- [16] G. Bleve, M. Tufariello, M. Durante, F. Grieco, F.A. Ramires, G. Mita, M. Tasioula-Margari, A.F. Logrieco, Physico-chemical characterization of natural fermentation process of Conservolea and Kalamàta table olives and development of a protocol for the pre-selection of fermentation starters, *Food Microbiol.* 46 (2015) 368-382.
- [17] A. Martorana, A. Alfonzo, L. Settanni, O. Corona, F. La Croce, T. Caruso, G. Moschetti, N. Francesca, An innovative method to produce Green table olives based on “pied de cuve” technology, *Food Microbiol.* 50 (2015) 126-140.
- [18] K. Pearson, Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurements of organs, *Proceedings Royal Soc. London* 60 (1896) 489-498.
- [19] J. Aitchison, The statistical analysis of compositional data, Chapman & Hall, Ltd, London, 1986.

- [20] M. Korhoňová, K. Hron, D. Klimčíková, L. Müller, P. Bednář, P. Barták, Coffee aroma-Statistical analysis of compositional data, *Talanta* 80 (2009) 710-715.
- [21] K. Hron, M. Jelínková, P. Filzmoser, R. Kreuziger, P. Bednář, P. Barták, Statistical analysis of wines using a robust compositional biplot, *Talanta* 90 (2012) 46-50.
- [22] ASEMESA. <http://www.asesesa.es>, 2016 (accessed 16.11.2016).
- [23] V., Pawlowsky-Glahn, J.J. Egozcue, Geometric approach to statistical analysis on the simplex, *Stoch. Env. Res. Risk A.* 15 (2001) 384-398.
- [24] J. Aitchison, M. Greenacre, Biplots for compositional data, *J. R. Stat. Soc C-Appl.* 51 (2002) 375-392.
- [25] J.J. Egozcue, V. Pawlowsky-Glahn, Exploring compositional data with the coda-dendrogram, in: E. Pirard (Ed), *Proceedings of the XIth Annual Conference of the International Association for the Mathematical Geology (IAMG'06)*, 2006.
- [26] V. Pawlowsky-Glahn, J.J. Egozcue, Análisis de datos composicionales con el coda-dendrogram, in: J. Sicilia-Rodríguez, C. González Martín, M.A. González-Sierra, D. Alcaide (Eds.), *Actas del XXIX Congreso de la Sociedad de Estadística e Investigación Operativa (SEIO'06)*, Sociedad de Estadística e Investigación Operativa, Tenerife, España, 2006, pp. 39-40.
- [27] V. Pawlowsky-Glahn, A. Buccianti, *Compositional data analysis: Theory and applications*, John Wiley & Sons, Chichester, UK, 2011.
- [28] K.G. Van den Boogaart, R. Tolosana-Delgado, *Analyzing compositional data with R*, Springer, Heidelberg, Germany, 2013.
- [29] S. Thió-Henestrosa, M. Comas, *CoDaPack. V2 User's Guide*, Girona, Spain, 2011.

- [30] V. Pawlowsky-Glahn, J.J. Egozcue, R. Tolosana-Delgado, Modeling and analysis of compositional data, John Wiley & Sons, Chichester, UK, 2015.

Figure captions

Figure 1. Graphical presentation of samples in the Simplex space as a function of the first four variables with the highest overall clr variance. The figure also includes Principal Component Axes derived from the same variables. Symbols stand for MAL, Manzanilla from Alcalá de Guadaira (Sevilla); MC, Manzanilla from Posadas (Córdoba); MAM, Manzanilla from Almendralejo (Badajoz); HA, Hojiblanca from Alameda (Málaga); HE, Hojiblanca from Estepa (Sevilla); HSC, Hojiblanca from Casariche (Sevilla). z, is for values with imputed zeros, and c is for clousured data after zeros replacement.

Figure 2. Covariance biplot. Projection of variables and cases onto the two first Principal Components. Variance explained PC1, 49.13%; PC2, 23.06%. Symbols stand for MAL, Manzanilla from Alcalá de Guadaira (Sevilla); MC, Manzanilla from Posadas (Córdoba); MAM, Manzanilla from Almendralejo (Badajoz); HA, Hojiblanca from Alameda (Málaga); HE, Hojiblanca from Estepa (Sevilla); HSC, Hojiblanca from Casariche (Sevilla).

Figure 3. Form biplot. Projection of variables and cases onto the two first Principal Components. Variance explained PC1, 49.13%; PC2, 23.06%. Symbols stand for MAL, Manzanilla from Alcalá de Guadaira (Sevilla); MC, Manzanilla from Posadas (Córdoba); MAM, Manzanilla from Almendralejo (Badajoz); HA, Hojiblanca from Alameda (Málaga); HE, Hojiblanca from Estepa (Sevilla); HSC, Hojiblanca from Casariche (Sevilla).

Figure 4. Balance coda-dendrogram obtained by sequential binary partitions (Table S6) based on the progressive decreasing variance of the variables placed in the numerator. The balance

value for each group is indicated by the intersection of the vertical bar (denoting the corresponding variance) and the horizontal bar. Due to the number of variables their names in the legend below the graph overlap.

Figure 5. Cluster analysis based on the ilr transformed data (coordinates) (panel A) and original data (expressed in percentages) (Panel B). Symbols stand for MAL, Manzanilla from Alcalá de Guadaira (Sevilla); MC, Manzanilla from Posadas (Córdoba); MAM, Manzanilla from Almendralejo (Badajoz); HA, Hojiblanca from Alameda (Málaga); HE, Hojiblanca from Estepa (Sevilla); HSC, Hojiblanca from Casariche (Sevilla).

Figure 6. Projection of the treatments onto the plane of the first two Principal Coordinates, using ilr transformed data (coordinates) (Panel A) and original data expressed in percentages (Panel B). Symbols stand for MAL, Manzanilla from Alcalá de Guadaira (Sevilla); MC, Manzanilla from Posadas (Córdoba); MAM, Manzanilla from Almendralejo (Badajoz); HA, Hojiblanca from Alameda (Málaga); HE, Hojiblanca from Estepa (Sevilla); HSC, Hojiblanca from Casariche (Sevilla).

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Figure 6

Table 1. Correspondence between compounds and the number used for simplification throughout the supplementary material and figures.

Compound	Number	Compound	Number	Compound	Number	Compound	Number	Compound	Number	Compound	Number
Isopropanol	1	1-Nonanol	18	Hexyl acetate	35	Isobutanoic acid	52	Limonene	69	p-Cresol	86
Ethanol	2	Benzyl alcohol	19	Ethyl (E)-3-hexenoate	36	Butanoic acid	53	6-Methyl-5-hepten-2-one	70	p-Propyl guaiacol	87
2-Butanol	3	Phenylethyl alcohol	20	Methyl lactate	37	2-Methylbutanoic acid	54	Linalool oxide	71	Eugenol	88
1-Propanol	4	Ethyl acetate	21	Ethyl lactate	38	Hexanoic acid	55	(+)-Cycloisotativene	72	4-Ethyl phenol	89
Isobutanol	5	Methyl propanoate	22	Methyl octanoate	39	Heptanoic acid	56	Copaene	73	Vanillin	90
1-Butanol	6	Propyl acetate	23	Ethyl octanoate	40	(E)-3-Hexenoic acid	57	Dihydroedulan	74	Tyrosol	91
Isopentanol	7	Methyl butanoate	24	Isoamyl lactate	41	Octanoic acid	58	Linalool	75	Octane	92
3-Methyl-3-buten-1-ol	8	Methyl 2-methylbutanoate	25	Methyl decanoate	42	Decanoic acid	59	Caryophyllene	76	Decane	93
1-Pentanol	9	Isobutyl acetate	26	Ethyl decanoate	43	Benzoic acid	60	α -Terpineol	77	o-Xylene	94
3-Methyl-2-buten-1-ol	10	Methyl 3-methylbutanoate	27	Ethyl benzoate	44	Pentanal	61	α -Murolene	78	Styrene	95
3-Methyl-1-pentanol	11	Ethyl butanoate	28	Benzyl acetate	45	Heptanal	62	β -Damascenone	79	2-Bornene	96
1-Hexanol	12	Ethyl 2-methylbutanoate	29	Methyl salicylate	46	Octanal	63	Isogeraniol	80	Dimethyl sulfide	97
(Z)-3-Hexen-1-ol	13	Ethyl 3-methylbutanoate	30	Ethyl salicylate	47	2-Heptenal	64	Geraniol	81	Theaspirane A	98
1-Octen-3-ol	14	Isoamyl acetate	31	Methyl hydrocinnamate	48	Nonanal	65	Iridomyrmecin	82	Theaspirane B	99
1-Heptanol	15	Methyl hexanoate	32	Ethyl 3-cyclohexenecarboxylate	49	(E)-2-Octenal	66	p-Creosol	83	Dimethyl sulfoxide	100
2-Ethyl-1-hexanol	16	Ethyl hexanoate	33	Acetic acid	50	Benzaldehyde	67	Phenol	84	Butyrolactone	101
1-Octanol	17	Methyl (E)-3-hexenoate	34	Propanoic acid	51	(E)-2-Decenal	68	p-Ethyl guaiacol	85	1,4-Dimethoxybenzene	102

Table 2. The ten volatiles with the highest central log ratio variances (*clr* variances) as well as their corresponding balances and variances in the sequential binary partition deduced from them.

Order	Volatile	<i>clr</i> variance	Corresponding Balance	Balance Value	Balance Variances
1	51 (Propanoic acid)	13.4906	1	2.1005	13.6297
2	4 (1-Propanol)	13.0073	2	0.5093	13.3904
3	1 (Isopropanol)	12.7384	3	-1.4262	13.2697
4	64 (2-Heptenal)	8.1542	4	-2.9409	7.9114
5	23 (Propyl acetate)	7.9274	5	-0.4042	8.4113
6	68 ((E)- 2-Decenal)	6.3973	6	-1.1875	6.6598
7	32 (Methyl hexanoate)	6.3488	7	-1.8408	6.6904
8	15 (1-Heptanol)	5.9048	8	-1.1035	5.7431
9	5 (Isobutanol)	5.7241	9	-2.6820	5.2932
10	6 (1-Butanol)	4.9998	13	-2.2149	4.2676