A Chemometric Study of Genotypic Variation in Triacylglycerol Composition Among Selected Almond Cultivars

M.S. Prats-Moya*, N. Grané-Teruel, V. Berenguer-Navarro, and M.L. Martín-Carratalá

Department of Analytical Chemistry, University of Alicante, 03080-Alicante, Spain

ABSTRACT: Triacylglycerol composition of 10 almond cultivars grown in seven different environments was determined by high-performance liquid chromatography, and multivariate statistical techniques (principal component and cluster analysis) were used to detect and establish associations among almond cultivars. In particular, linear discriminant analysis showed that cultivars Desmayo Largueta, Masbovera, Texas, Non Pareil, Tuono, and Guara are different from each other and can be classified in different groups. This statistical approach also predicted the origin of eight hybrids produced from these cultivars. As a result, it is concluded that triacylglycerol composition in almonds can distinguish genetic from environmental effects on almond triacylglycerol composition.

Paper no. J8877 in *JAOCS 76*, 267–272 (February 1999).

KEY WORDS: Almond, characterization, chemometrics, high-performance liquid chromatography, triacylglycerols.

Chemometric techniques may reveal useful information from analytical data, including characterization of natural goods. However, caution must be taken to ensure that these techniques are used in an appropriate manner (1). We have applied some common statistical algorithms to fatty acids (2), free amino acids (3,4), and inorganic element composition of almond kernel (5). All these studies showed the existence of cultivar-dependent differences in chemical composition that were independent of growth conditions. To verify this hypothesis we evaluated triacylglycerol composition of the almond kernels in samples of cultivars that were grown under diverse conditions. A recent study on edible oils showed that triacylglycerol composition could be differentiated using a two-dimensional plot of principal component analysis (PCA) of mass spectrometric data (6). In this work, we determined the triacylglycerol composition of 10 almond cultivars (Mediterranean and American) from seven different localities and analyzed these data with multivariate techniques [principal component analysis (PCA), cluster analysis, and linear discriminant analysis (LDA)].

MATERIALS AND METHODS

The 10 almond cultivars used in this investigation were: Desmayo Largueta (DL), Marcona (MR), Guara (GU), and Masbovera (MS) from Spain; Texas (TE), Non Pareil (NP), and Titan (TI) from the United States of America; Tuono (TU) from Italy; Ferragnes (FE) from France; and Primorskyi (PR) from a Caucasian region. These cultivars were grown in six different geographical areas in Spain: María (Almeria, ALM); La Puebla de Don Fabrique (Granada, GR); Santomera and Cehegín (Murcia, MUR), Castalla and Bacarot (Alicante, ALIC), Mas Valero, Mas Bove, and Mora d'Ebre (Tarragona, TA); and Aula Dei (Zaragoza, ZAR); as well as Avignon (AVIG), in France. These locations differed significantly in climatological conditions. Granada and Zaragoza are in cold regions, while the rest of the localities have mild climates. A total of 107 samples was collected from cultivars at these locations in 1996. Samples were stored at 4°C until analysis.

Almond samples were blanched and ground in an electrical grinder. Oil was extracted in an all-glass apparatus, similar to the commercial Soxtec, using a mixture of diethyl ether/hexane (1:1 vol/vol). Samples of 0.1 to 0.2 mg of oil were dissolved in 4 mL of acetone and filtered through a 0.45 μm pore size filter. Aliquots were injected into a Waters Multisolvent high-performance liquid chromatography (HPLC) system (Milford, MA) equipped with a double piston pump, thermostatic control of column temperature, and a double refractive index detector model Waters 410. The column was a 4.6×250 m Symmetry C_{18}, supplied by Waters. The injection valve was a Rheodyne model 7125 with a 20 μL loop (Cotati, CA). HPLC-grade acetonitrile and acetone were obtained from Lab Scan Ltd. (Dublin, Ireland). Trilinolein, triolein, tripalmitin, and tristearin reference standards were from Sigma Chemical Co. (St. Louis, MO).

The optimal elution solvent found was acetone/acetonitrile 65:35 (vol/vol). Elution was carried out isocratically at a column temperature of 30°C and a flow rate of 1.5 mL/min for 30 min. The detector was held at 40°C and operated at medium sensitivity.

Identification of triacylglycerol molecular species (TGMS) was achieved by comparison with relative retention times obtained in a reference soy oil (7). The same response factor was assumed for all TGMS and the content of each TGMS was calculated relative to triolein.

Several statistical methods in an SPSS statistical package [SPSS, 1994 (8)] were used for data analysis. PCA was applied to autoscaled data. The number of the components to be retained was selected by using the Scree test and Kaiser crite-

*To whom correspondence should be addressed at Department of Analytical Chemistry, University of Alicante, P.O. Box 99, 03080-Alicante, Spain. E-mail: maria.prats@ua.es

ria (9). Cluster analysis was carried out by applying the average linkage method for agglomeration and the square of Euclidean distance as criterion of proximity (10). The LDA was conducted stepwise by employing the Wilks' lambda statistics (11) for variable selection. In all cases, the algorithms used were applied to the mean values from four replications for each sample.

RESULTS AND DISCUSSION

Nine TGMS were identified and determined in almond oil. In decreasing order of importance these were: OOO, OLO, POO, OLL, PLO, StOO, LLL, PLL, PLP, where O = 18:1, L = 18:2, P = 16:0, and St = 18:0. A summary of TGMS mean concentration values, along with their standard error (SE), from all samples among cultivars is given in Table 1. Individual data from each sample were used in the multivariate techniques.

The first step in this statistical approach is to determine correlations among variables. In the initial correlation matrix all triacylglycerols are highly correlated. However, this matrix was not an identity matrix according to the Bartlett Test of Sphericity (12), and has a determinant value of null. This requires a reduction of variables such as that provided by the PCA.

From the 107 samples studied, 91 were initially selected as the training set, with the remainder as the test set. A preliminary application of PCA to the data of the training set showed that two cultivars, Marcona and Ferrragnes (28 samples total), were so irregularly distributed that these data impeded the observation of the main emergent groups. Therefore, these two cultivars were removed from the training set of samples.

Nine new variables or principal components were deduced from this reduced set of samples. After application of the Kaiser criteria and the Scree test, only two principal components were retained. It was found that the first two principal components explained 84.4% of the total variance. The first principal component explained 72.3% of the variance, and the second accounted for 15.5% of the variance.

The model was accepted because reproduced communality values for all TGMS were near 1, and also because the lower percentage of residuals was only 30%. Table 2 contains in the lower left triangle the reproduced correlation matrix, in the diagonal the reproduced communality values, and in the upper right triangle the residuals between the observed correlations and the reproduced correlations.

Scores of the samples for the two first principal components (CP) were calculated as follows.

$$CP1 = 0.14\,LLL + 0.15\,LLO + 0.13\,OLO - 0.15\,OOO + 0.15\,PLL + 0.15\,PLO + 0.12\,PLP - 0.06\,POO - 0.08\,StOO \quad [1]$$

$$CP2 = -0.05\,LLL - 0.05\,LLO + 0.01\,OLO - 0.13\,OOO + 0.09\,PLL + 0.15\,PLO + 0.30\,PLP + 0.63\,POO + 0.45\,StOO \quad [2]$$

Figure 1 shows that three main associations are visualized by the two principal components. The first is formed by Masbovera samples, the second by Desmayo Largueta samples, and the third by Tuono and Guara. The remaining cultivars are all near to each other, but not clearly associated.

TABLE 1
Mean Triacylglycerol Molecular Species Percentage[a] Among Almond Cultivars and the SE for Each Mean

Cultivar	Number of samples	%LLL[b]	%LLO	%OLO	%OOO	%PLL	%PLO	%PLP	%POO	%StOO
DL	16	2.53	15.11	25.59	29.13	2.50	10.81	0.51	11.30	2.43
		(0.10)	(0.59)	(0.41)	(1.10)	(0.09)	(0.28)	(0.01)	(0.22)	(0.08)
FE	16	1.90	10.66	22.30	41.43	1.76	7.37	0.40	10.90	3.27
		(0.17)	(0.82)	(0.77)	(2.18)	(0.14)	(0.47)	(0.02)	(0.14)	(0.09)
GU	13	1.56	9.67	23.60	39.49	1.62	7.65	0.41	12.07	3.93
		(0.07)	(0.36)	(0.61)	(1.09)	(0.05)	(0.26)	(0.01)	(0.13)	(0.16)
MR	17	2.21	12.23	24.18	35.95	1.98	8.56	0.42	11.27	3.20
		(0.12)	(0.58)	(0.70)	(1.38)	(0.11)	(0.31)	(0.01)	(0.19)	(0.21)
MS	10	1.19	6.45	19.02	53.24	0.98	4.70	0.28	11.11	3.03
		(0.03)	(0.15)	(0.21)	(0.48)	(0.03)	(0.08)	(0.01)	(0.17)	(0.14)
NP	8	2.65	13.85	24.58	34.41	2.28	9.02	0.41	10.36	2.33
		(0.13)	(0.54)	(0.45)	(1.49)	(0.14)	(0.37)	(0.01)	(0.23)	(0.14)
PR	4	2.13	11.79	22.90	39.85	1.77	7.73	0.36	10.72	2.75
		(0.11)	(0.62)	(0.38)	(1.29)	(0.05)	(0.35)	(0.01)	(0.41)	(0.13)
TE	8	2.08	11.27	24.07	39.46	1.79	7.50	0.35	10.53	2.84
		(0.11)	(0.48)	(0.19)	(0.82)	(0.07)	(0.17)	(0.01)	(0.28)	(0.07)
TI	4	2.31	12.84	24.82	36.35	2.04	8.53	0.42	10.67	2.03
		(0.16)	(1.26)	(0.33)	(1.94)	(0.14)	(0.47)	(0.02)	(0.46)	(0.06)
TU	11	1.53	9.97	23.88	39.30	1.65	7.69	0.40	11.70	3.86
		(0.10)	(0.46)	(0.68)	(1.34)	(0.07)	(0.28)	(0.01)	(0.16)	(0.08)
Total	107	1.99	11.38	23.53	38.46	1.86	8.08	0.41	11.19	3.09

[a]Standard errors (SE) in parentheses.
[b]DL, Desmayo Largueta; FE, Ferragnes; GU, Guara; MR, Marcona; MS, Masbovera; NP, Non Pareil; PR, Primorskyi; TE, Texas; TI, Titan; TU, Tuono; L, 18:2; O, 18:1; P, 16:0; and St, 18:0.

TABLE 2
Reproduced Correlation Matrix, Residuals, and Communality Values

	LLL[a]	LLO	OLO	OOO	PLL	PLO	PLP	POO	StOO
LLL	*0.872	0.216	–0.109	0.003	0.064	–0.006	–0.054	0.012	0.091
LLO	0.916	*0.975	–0.210	–0.012	0.014	0.005	–0.032	–0.023	0.067
OLO	0.775	0.851	*0.783	–0.038	–0.076	–0.015	–0.039	–0.041	0.002
OOO	–0.846	–0.936	–0.875	*0.981	0.005	0.002	0.042	0.024	–0.055
PLL	0.877	0.951	0.858	–0.954	*0.948	–0.001	–0.024	–0.003	0.061
PLO	0.843	0.936	0.877	–0.985	0.955	*0.986	0.002	0.006	–0.012
PLP	0.668	0.773	0.775	–0.885	0.824	0.892	*0.861	0.021	–0.133
POO	–0.515	–0.451	–0.226	0.198	–0.324	–0.174	0.052	*0.879	–0.181
StOO	–0.597	–0.574	–0.394	0.402	–0.486	–0.393	–0.196	0.689	*0.606

[a]See Table 1 for abbreviations.
*Reproduced communality values. Below the diagonal line of these values is the reproduced correlation matrix, and above the diagonal line are the residuals between the observed correlations and the reproduced correlations.

Cluster analysis is an effective means of distinguishing intercorrelations among associations on the basis of nearness criteria between objects. We applied this algorithm to the scores of each sample as calculated in the PCA. The resultant dendrogram (Fig. 2) shows four distinct groups at a rescaled distance of 8 (i.e., 8/25 × 100 = 68% of similarity). Among these data Masbovera, Desmayo Largueta (the three exceptions were harvested in Murcia), and Guara and Tuono (with the exceptions of Guara grown in Avignon, Tuono grown in Avignon and Tuono grown in Almería) define distinct groups. The cultivars Titan, Primorskyi, and Texas are associated in one group, while the cultivar Non Pareil appears distributed between two groups—two samples are classified with Desmayo Largueta and four with Titan, Primorskyi, and Texas.

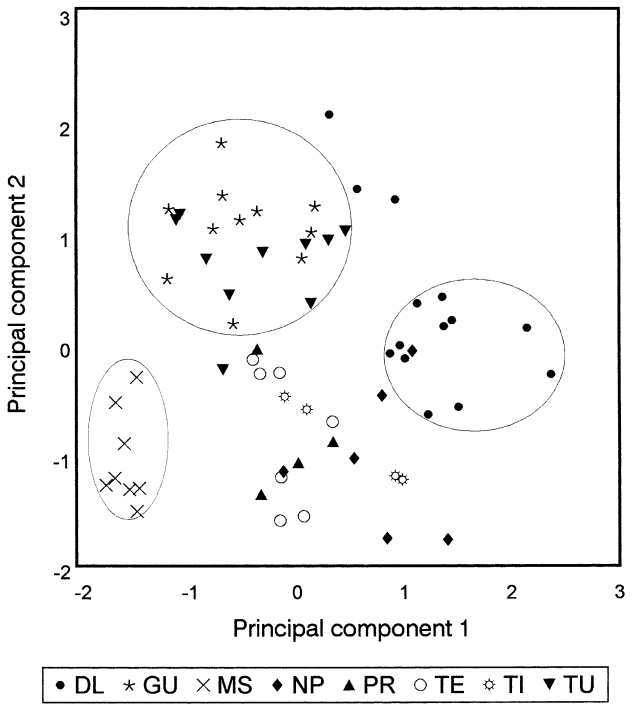

FIG. 1. Projections of the average scores for the training set samples on the reduced space of the two principal components. DL, Desmayo Largueta; GU, Guara; MS, Masbovera; NP, Non Pareil; PR, Primorskyi; TE, Texas; TI, Titan; TU, Tuono.

Based on this information, a descriptive discriminant analysis was conducted according to Wilks' stepwise method (13), assuming the following five defined groups: Desmayo Largueta (group 1); Masbovera (group 2); Guara + Tuono (group 3); Texas (group 4), and Non Pareil (group 5). This algorithm sequentially selected the variable (each TGMS mean value for a group) which minimized the Wilks' lambda value. The minimum tolerance level adopted for the retention of variables was 0.001, and the selection of variables was randomized. This stepwise algorithm effectively reduced variables and avoided intercorrelations.

All TGMS except PLP and POO were selected. Thereafter, linear combinations of retained variables were calculated, so that these functions maximized the differences between the established groups. These four discriminant functions retained 57.88, 35.47, 6.48, and 0.17% of the variance, respectively, with a canonical correlation of 0.959, 0.935, 0.749, and 0.179. The following coefficients for calculating these functions were deduced from the matrix of correlations between the variables and the functions:

$$FD1 = -1.39\,LLL + 0.03\,LLO + 2.24\,OLO + 1.97\,OOO + 1.94\,PLL - 0.51\,PLO + 1.45\,StOO \quad [3]$$

$$FD2 = 1.46\,LLL + 0.33\,LLO + 1.76\,OLO + 4.07\,OOO + 0.09\,PL + 0.54\,PLO - 0.21\,StOO \quad [4]$$

$$FD3 = -1.12\,LLL + 1.76\,LLO - 1.38\,OLO + 1.12\,OOO - 1.07\,PLL + 1.97\,PLO + 0.04\,StOO \quad [5]$$

$$FD4 = 1.00\,LLL + 0.16\,LLO + 1.28\,OLO + 1.73\,OOO - 2.02\,PLL + 1.84\,PLO + 0.12\,StOO \quad [6]$$

Finally, applying the rule of Bayes (14) to the numerical values or scores of these functions for all the samples, we calculated the probability of belonging to a defined group. All training set samples were correctly assigned to the five postulated groups (Fig. 3).

Validation of the discriminant function approach was

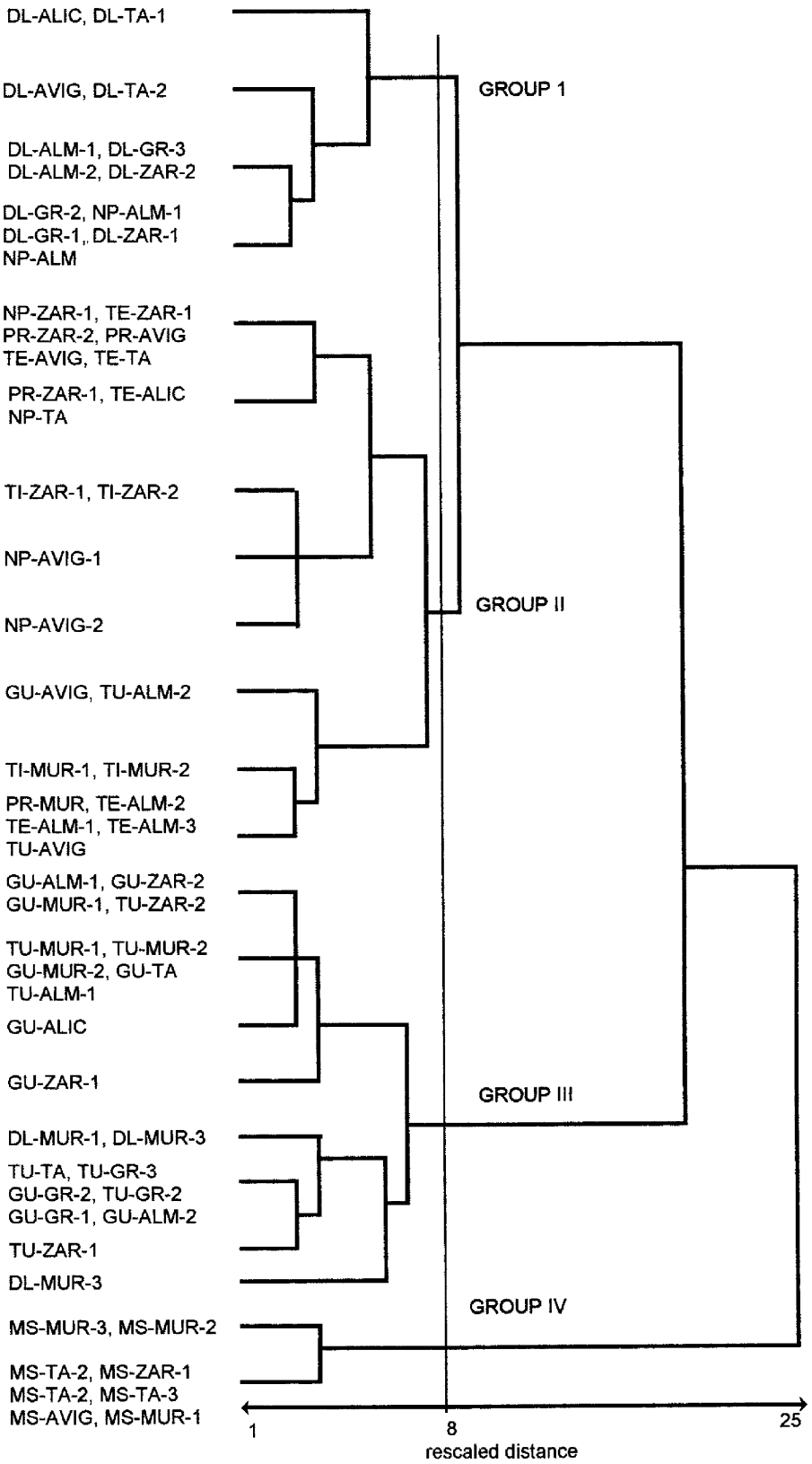

FIG. 2. Dendrogram from cluster analysis. Samples in a box are at a distance of less than one.

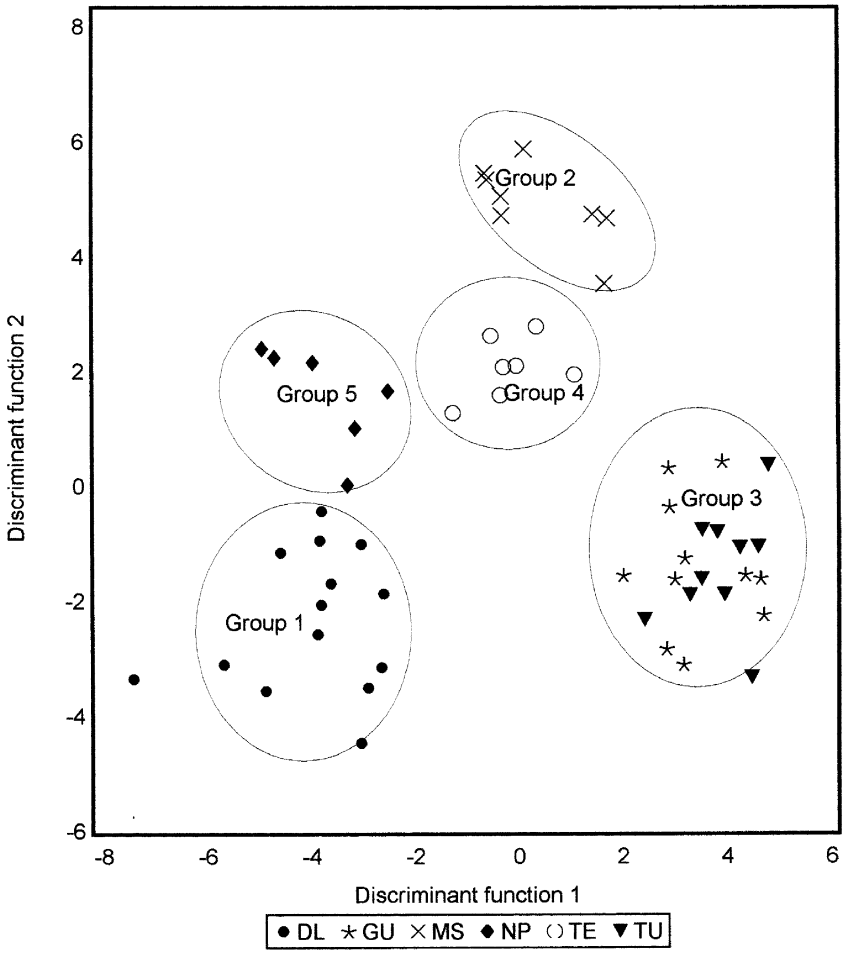

FIG. 3. Mean scores of almond cultivars projected on the reduced space of the three rotated discriminant functions. See Figure 1 for identification of cultivars.

achieved using a test set with the remaining 11 samples of the cultivars that were not included in the training set. Among this data, four samples of Titan and three samples of Primorskyi were classified within the Non Pareil group. However, Marcona and Ferragnes appeared in several groups, probably revealing a pronounced multiclonal character.

In addition, we applied the calculated functions to a set of hybrids obtained in the Agronomical Research Center CEBAS (Centro de Edafología y biología Aplicada del Segura, Murcia). This set consisted of the following samples: Peraleja × Tuono, Garrigues × Genco, four samples of Tuono × Ferragnes, four samples of Ferranges × Genco, Ferragnes × Tuono, two samples of Tuono × Genco, and Genco × Ferragnes. It was found that all except one hybrid from Tuono were assigned to group 3, which was the predicted group for Tuono and Guara.

As a conclusion, it appears that triacylglycerol composition of almond genotypes may be distinguished by genetic properties which are not obscured by environmental effects. Therefore, this approach may be used to differentiate almond cultivars from each other.

ACKNOWLEDGMENTS

We wish to thank the following for providing us with the almond samples: Federico Dicenta López-Higuera (CEBAS-Murcia), Rafael Socias i Company (SIA-Zaragoza), Philippe Froment (INRA-Avignon), Antonio Sanchez Navarro (La Puebla de Don Fabrique-Granada), Agustín Navarro Muñoz (Los Velez-Almería), and Francisco Vargas García (IRTA-Reus).

REFERENCES

1. Defernez, M., and E.K. Kemsley, The Use and Misuse of Chemometrics for Treating Classification Problems, *Trends Anal. Chem. 16*:216–221 (1997).
2. García-López, C., N. Grané-Teruel, V. Berenguer-Navarro, E. García-García, and M.L. Martín-Carratalá, Major Fatty Acid Composition of 19 Almond Cultivars of Different Origins. A Chemometric Approach, *J. Agric. Food Chem. 44*:1751–1755 (1996).
3. Prats, M.S., and V. Berenguer, Caracterización de Algunas Variedades de Almendra por su Composición en Aminoácidos Libres, *Rev. Esp. Cienc. Tecnol. Aliment. 34*:218–227 (1994).
4. Helena Seron, L., M.S. Prats-Moya, M.L. Martín-Carratalá, V. Berenguer-Navarro, and N. Grané-Teruel, Characterisation of 19 Almond Cultivars on the Basis of Their Free Amino Acids Composition, *Food Chem. 61*:455–459 (1998).

5. Prats-Moya, S., N. Grane Teruel, V.Berenguer Navarro, and M.L. Martín Carratalá, Inductively Coupled Plasma Application for the Classification of 19 Almond Cultivars Using Inorganic Element Composition, *J. Agric. Food Chem. 45*:2093–2097 (1997).
6. Lamberto, M., and M. Saitta, Principal Component Analysis in Fast-Atom-Bombardment Mass Spectrometry of Triacylglycerols in Edible Oils, *J. Am. Oil. Chem. Soc. 72*:867–87 (1995).
7. Wolff, F.P., F.X. Mordret, and A. Dieffenbacher, Determination of Triglycerides in Vegetable Oils in Terms of their Partition Numbers by High-Performance Liquid Chromatography, *Pure Appl. Chem. 63*:1173–1182 (1991).
8. *Statistical Package of Social Science*, Release 6.0.1, SPSS Inc., Chicago, 1994.
9. Cattel, R.B., The Scree Test for the Number of Factors, *Mult. Behav. Res. 1*, 245–276 (1966).
10. Afifi, A.A., and V. Clark, *Computer-Aided Multivariate Analysis*, 2nd edn., Van Nostrand Reinhold, New York, 1990, pp. 429–462.
11. Tabachnick, B., and L. Fidell, *Using Multivariate Statistics*, 12th edn., Harper Collins, New York, 1992, pp. 505–596.
12. Bartlett, M.S., The Statistical Significance of Canonical Correlations, *Biometrika 32*, 29–38 (1941).
13. Bisquerra Alzina, R., *Introduccion conceptual al analisis multivariante*, Promociones y Publicaciones Universitarias, Barcelona, 1989, pp. 251–253.
14. Bisquerra Alzina, R., Introduccion conceptual al analisis multivariante, Promociones y Publicaciones Universitarias, Barcelona, 1989, pp. 270–285.

[Received May 13, 1998; accepted August 14, 1998]