INVESTIGACION

Estimating triacylglycerols from fatty acids by chemometrics. An application in Spanish virgin olive oil.

By **J. García Pulido** and **B. Aparicio López**
Instituto de la Grasa y sus Derivados
Avda. Padre García Tejero, 4 - 41012 Sevilla

RESUMEN

Métodos quimiométricos para el cálculo de triglicéridos mediante ácidos grasos en aceites de oliva vírgenes españoles.

Se ha analizado el nivel de aceptación del patrón "1, 3-al azar 2-al azar" de la distribución de los grupos acil dentro de los glicéridos del aceite de oliva. Para ello se han comparado los valores de los triglicéridos obtenidos por HPLC con los deducidos mediante esa teoría con 36 muestras de aceites de oliva españoles. Hotelling's T^2, Componentes Principales y Correlación Canónica han sido los procedimientos estadísticos usados para estudiar ambos grupos de datos. Previamente se habían analizado la normalidad de la distribución de los datos y las correlaciones intra-intergrupo de triglicéridos.

Se propone un camino alternativo para estimar algunos triglicéridos mediante los valores de los ácidos grasos. Se ha empleado un procedimiento de regresión múltiple por pasos, obteniéndose valores de R que fluctúan entre 0,75 y 0,94 excepto para el triglicérido SOS.

PALABRAS-CLAVE: Aceite de oliva virgen - Acido graso - Estudio quimiométrico - Triglicéridos (composición).

SUMMARY

Estimating triacylglycerols from fatty acids by chemometrics. An application in Spanish virgin olive oil.

The level of acceptation of "1, 3-random 2-random" pattern of acyl groups distribution in virgin olive oil glycerides has been analysed in 36 samples, by comparing the values of triacylglycerols obtained by HPLC with those reported by the cited theory. Hotelling's T^2, Principal Components and Canonical Correlation Analysis have been used to study both data sets. The normality of the data distributions, the intra-and intergroup correlations of the triacylglycerols were studied before these multivariate statistical algorithms were applied.

An alternative way of estimating triacylglycerols from only total fatty acids has also been studied. A Stepwise Multiple Regression procedure has been employed, and Multiple R coefficients fluctuate between 0,75 and 0,94 except for the variable SOS.

KEY-WORDS: Chemometrics study - Fatty acid - Triglycerides (composition) - Virgin olive oil.

1. INTRODUCTION

Vander Wal, Coleman and Fulton, and Gunstone (1967) suggested independently that the composition of the triacylglycerols could be determined by applying the "1,3-random 2-random" distribution theory to the results of quantifying the total and beta-position fatty acids. At their epoch there was no method to quantify triacylglycerols, and this was the only way to estimate them. We have tested this theory experimentally with virgin olive oil, following the perfecting of quantification methods using HPLC.

However, this theory has less utility today because we need two analysis to obtain total and beta-position fatty acids. Moreover, if an alternative way of estimating triacylglycerols using only total fatty acids can be got by Multiple Regression. Furthermore, Gas Chromatography (GC) offers the possibility of identifying triacylglycerols more easyly than HPLC. So, we have applied Regression Analysis to obtain equations relating triacylglycerols quantified by GC and fatty acids.

This paper analyses the possibilities of both methodologies.

2. MATERIALS AND METHODS

Chemical apparatus

Graciani (1988) described the apparatus and methodology for quantifying the triacylglycerols by means of HPLC. In this paper, the following have been considered: L2O, LnO2, L2P, O2L, LOP, P2L, O3, O2P, OP2, SO2. The number signifies that this triacylglycerol is the sum of all the isomers of glycerine position so formulated.

For triacylglycerols quantification by Gas Chromatography a Chrompack CP900 fitted with a FID detector was employed. A fused silica column (25 m x 0,25 mm I.D.) coated with phenylmethyl-silicone TAP (0,1 μM thickness) was used. The carrier gas was H_2. 0,2 microlitres samples of 0,05% Spanish olive oil in hexane were held at 33°C for one minute and programmed at 1° C/min to the final temperature of 344°C. Following triacylglycerols were quantified: O2L, LOP+PoO2, L2O+LnO2, SO2, O3, O2P, POS, L2P+PoLO, P2L+PPoO, OP2 and S2O. The sum signifies that more than one triacylglycerol has been quantified in those peaks.

The total and beta-position fatty acids were quantified by means of gas chromatography following the methods described by UNE 55037 and UNE 55079 standards.

Data set

In the verification of "1,3-random 2-random" theory two data sets have been used, each one holding triacylglycerols quantified by HPLC and triacylglycerols calculated by applying these equations to total and beta-position fatty acids.

36 samples were chosen on the premise that the data set represented different majority varieties, zones of production, and more than one olive oil-production season. In addition, the set had to represent oils with different triacylglycerols contents.

In the Regression Analysis there were used two other data sets, one with triacylglycerols quantified by Gas Chromatography and the other containing triacylclycerols calculated with the regression equations. On the same selection basis, 125 samples were chosen, 75 of which were used as data set to build regression equations and 50 were used as test set in order to verify them.

Computer systems

A 80386 minicomputer was used to run BMDP and SPSS+ X statistical procedures.

3. RESULTS AND DISCUSSION

3.1 Estimating triacylglycerols by "1,3-random 2-random" distribution theory

Following the theory explained by Gunstone (1967), the values of total and beta-position fatty acids were used to estimate triacylglycerols. This group of data and triacylglycerols quantified by HPLC make up the two sets of data to be studied statistically.

Basic statistics were firstly analysed for the data distribution of each triacylglycerol in each set (henceforth, class). Class 1 is the set of data obtained by HPLC and class 2 that calculated using the theory cited.

Distributions of each triacylglycerol (henceforth, variable) were studied, and mathematical transformations were applied to those variables with skewness and kurtosis values very different to zero. Then, a test of normality (Tabachnick & Fidell, 1983) was applied, rejecting the group of minority triacylglycerols (L2O, LnO2, L2P and P2L), so that the authors decided not to use them.

Once the significant variables had been chosen, the data sets were re-analysed to detect the existence of outliers. As a statistical criterion for their identification, the ratio CHISQ/DF (quotient of Chi-squared to degrees of freedom) was used (Tabachnic and Fidell, 1983). In this way, two outliers were found, one for each class, which were not eliminated given their proximity to the limiting value of the criterion (2.806).

Analysis of the correlation matrix

The correlation matrix in which the variables of both classes took part was studied. The existence of a positive correlation between identical variables of the two classes can be seen in table I. This in general corresponds to the maximum value of that column or row, following the logic that the correlation between two identical variables –although obtained by different methods– should be positive and stronger than that between two which relate to different chemical compounds.

Table I
Correlations between variables of the same and different class. The variables of class 2 have been identified with an X.

	O2L	LOP	O3	O2P	OP2	SO2	O2LX	LOPX	O3X	O2PX	OP2X	SO2X
O2L	1,000											
LOP	0,812	1,000										
O3	-0,601	-0,870	1,000									
O2P	-0,225	0,199	-0,465	1,000								
OP2	-0,038	0,299	-0,582	0,506	1,000							
SO2	-0,623	-0,574	0,452	-0,264	0,012	1,000						
O2LX	0,835	0,802	-0,667	0,026	0,018	-0,680	1,000					
LOPX	0,824	0,879	-0,804	0,207	0,172	-0,716	0,876	1,000				
O3X	-0,735	-0,842	0,798	-0,203	-0,175	0,564	-0,761	-0,947	1,000			
O2PX	-0,048	0,207	-0,379	0,618	0,439	-0,174	-0,164	0,266	-0,344	1,000		
OP2X	0,330	0,541	-0,646	0,520	0,373	-0,418	0,242	0,651	-0,726	0,869	1,000	
SO2X	-0,712	-0,747	0,685	-0,281	-0,199	0,782	-0,821	-0,816	0,646	-0,154	-0,444	1,000

However, this general theory has one important exception: the little correlation existing between the variables OP2 of both classes (0.373).

On the other hand, the high level of correlation within class 2 drove us to analyse the existence of multicollinearity. Tabachnick and Fidell (1983) recommend to ana-

lyse the SMC values (Squared Multiple Correlations) of each variable with the other variables of its set. These authors are of the opinion that multicollinearity does exist if the SMC value is higher than 0,99. At that order of values, no variable was detected in either of the two classes. However, four of them, belonging to class 2 (O2L, LOP, O3 and OP2), have values higher than 0,96 which can explain the strong correlation within this class.

Hotelling's T^2

Next step, after all this preliminary study, was to analyse whether the two classes presented significant differences. The statistical procedure chosen was Hotelling's T^2 (Bisquerra, 1989), which can be considered as a generalization of T-Student test in the case of more than one dependent variable (triacylglycerols).

The test rejected the null hypothesis ($F=125,95$, $p<0,00005$). This is, the two classes show significant differences. Analysing each one of the dependent variables considered individually, it was determined that only the variables O2P ($p<0,00005$) and OP2 ($p=0,0015$) have got significant differences, against the other variables: O2L ($p=0,3131$), LOP ($p=0,0058$), O3 ($p=0,0133$) and SO2 ($p=0,0459$).

Having rejected the possibility that all the variables accepted the null hypothesis using Hotelling's test, it was opted to look for other ways of comparing both sets of data, permitting to reaffirm these results as well as to gain other usefull information about them. Principal Components and Canonical Correlation Analysis (henceforth PCA and CCA, respectively) were applied to this objective (Bisquerra, 1989).

Principal Components Analysis

PCA is aimed to describe data structure in each class so that we can compare them and decide about their similitude.

Table II shows the coefficients for the first two varivectors of each class. In both classes, the first two varivectors explain more than 80% of the variance, with the first varivector able to explain more than 50%, 53% in class 1 and 63% in class 2.

Table II
Varivectors and explained variance in each class.

Chemical Parameter	Class 1 Factor 1	Class 1 Factor 2	Class 2 Factor 1	Class 2 Factor 2
O2L	0,950	0,000	0,972	0,000
LOP	0,915	0,265	0,925	0,336
O3	-0,737	-0,611	-0,826	-0,447
O2P	0,000	0,854	0,000	0,977
OP2	0,000	0,866	0,372	0,922
SO2	-0,761	0,000	-0,882	0,000
Variance				
In data space	0,529	0,277	0,651	0,276
In factor space	0,656	0,344	0,703	0,297

As it can be seen, in both classes the variance assigned to the first varivector is almost totally explained by the variables O2L, LOP, O3 and SO2, whilts the triacylglycerols O2P and OP2 only play a little or no part in it. Two pairs have been formed by the variables O3-SO2 and O2L-LOP, which have opposing effects explaining the variance of this first factor. On the other hand, the second factor is explained mainly by O2P and OP2, while O2L and SO2 play no part.

Up to here, we can conclude there is the same structure in both data sets: The first factor is explained by the same variables in both classes, as well as the second factor.

A Regression Analysis between the first varivector of both classes has also been applied, R coefficient being 0,91 for $p<0,0001$, in agreement with the idea of factors of both groups being very similar. But if a Regression Analysis between the second varivectors is performed, R coefficient is 0,526 for $p<0,001$, indicating that these factors are not the same, the triacylglycerols they explain being not either.

So, PCA tell us about the same conclusion than Hotelling's T^2: OP2 and O2P obtained by "1,3-random 2-random" theory do not fit to their real values quantified by HPLC.

Canonical Correlation Analysis

In order to know the number of canonical variables necessary for expression of the dependence between the two classes, Bartlett's test was applied, selecting canonical variables until the probability of the remaining autovectors is non-significant ($p>0,01$). One significant canonical variable was found with a coefficient of correlation of 0,95, explaining 91% of data variance. So, CCA permits also conclude there is a great relationship between both classes.

But CCA also provides us with an analysis of correlation of each variable of a class with all the variables of the other class. This allows knowing the explained variance for the first using the variables of the second. Tables III and IV show the quadratic coefficient of correlation and the explained variance, with their level of probability. We can see that all the variables, except O2P and OP2, which have $p>0,001$, can ben explained with their homologous in the other class. Once again it is confirmed that the theory does not fit for these variables.

Table III
Quadratic multiple correlation of each variable of the first class with all the variables of the second.

Variab.	R-Squared	R-Squared adjusted	Statistic F	Degrees of freedom		Tail prob.
O2L	0,7599	0,7103	15,30	6	29	<0,0001
LOP	0,7959	0,7536	18,85	6	29	<0,0001
O3	0,7175	0,6590	12,27	6	29	<0,0001
O2P	0,5111	0,4100	5,05	6	29	0,0012
OP2	0,2349	0,0767	1,48	6	29	0,2184
SO2	0,6658	0,5967	9,63	6	29	0,0000

Table IV
Quadratic multiple correlation of each variable of the second class with all the variables of the first.

Variab.	R-Squared	R-Squared adjusted	Statistic F	Degrees of freedom		Tail prob.
O2L	0,7975	0,7555	19,03	6	29	<0,0001
LOP	0,8875	0,8643	38,14	6	29	<0,0001
O3	0,7888	0,7451	18,05	6	29	<0,0001
O2P	0,4175	0,2970	3,46	6	29	0,0105
OP2	0,5021	0,3990	4,87	6	29	0,0015
SO2	0,7727	0,7257	16,43	6	29	<0,0001

3.2. Estimating triacylglycerols by regression analysis

"1,3-random 2-random" theory has been verified analysing the most important triacylglycerols estimated from total and beta position fatty acids, but from a practical point of view, no enough advantages would be gained using this theory against the triacylglycerols quantification by gas chromatography techniques. Mathematical procedures are justified if they can actain optimal results with the lowest chemical cost, for example, if they only need one chemical analysis to estimate triacylglycerols properly.

Stepwise Multivariate Regression Analysis (Bisquerra, 1989) has been the statistical algorithm used to study whether only total fatty acids allow estimating the most important triacylglycerols.

With this analysis we obtain equations that relate each triacylglycerol (dependent variable) with fatty acids (independent variables). It is well-known that fatty acids values are always represented in percentages, so they are not independent from each other because there is a lineal equation that links them. However, a satisfactory way of dealing with this problem was found removing those fatty acids which did not relate to any triacylglycerol chemical composition: Margaric, Margaroleic, Arachidic, Gadoleic and Behenic.

In relation to the number of samples, Peña (1986) suggests that regression works on the basis of a number of samples four or five times the number of independent variables (fatty acids). In this case, Regression has been applied on 75 samples, using Palmitic, Palmitoleic, Stearic, Oleic, Linoleic and Linolenic as total number of independent variables. Thus, data set agrees with that suggestion.

Before applying Regression, we had to build a decision about what fatty acids should be included in each analysis. It was decided to use only those fatty acids involved in the chemical composition of each triacylglycerol, so that the possible equations had a chemical background. After that, statistical procedures would be able to select the variables with most significance to equation.

Table V shows variables in the equations, their multiple R and adjusted R-Square. These results show that the majority of triacylglycerols can be estimated by only total fatty acids with Multiple R coefficients greater than 0,75. However, the adjusted R-squared coefficients are good enough only for the variables O2L, LOP+PoO2, L2O+LnO2, SO2, O3 and O2P, these being the most important triacylglycerols in virgin olive oil. The other triacylglycerols obtained by GC represent a very little percentage of total, and so its estimation is very difficult.

Table V
Formulating triacylglycerols by total fatty acids: Multiple R, Adjusted R-Square and variables in the equation.

Triacylglycerol	Multiple R	Adjusted R-Square	Variables in the equation
O2L	.94	.89	Linoleic.
LOP+PoO2	.91	.83	Palmitic, Linoleic
L2O+LnO2	.91	.82	Linoleic, Linolenic.
SO2	.88	.78	Stearic, Oleic.
O3	.88	.77	Oleic.
O2P	.82	.67	Palmitic, Oleic.
POS	.78	.59	Stearic, Palmitic.
L2P+PoLO	.76	.58	Oleic.
P2L+PPoO	.76	.57	Palmitic, Linoleic, Palmitoleic.
OP2	.75	.56	Palmitic.
S2O	.54	.28	Stearic.

Table VI provides the equations obtained in the analysis by using standardized regression coefficients (beta).

Table VI
Regression equations using beta coefficients.

O2L = .94777*Linoleic.
LOP+PoO2 = .37914*Palmitic + .80540*Linoleic.
L2O+LnO2 = .96312*Linoleic - .14476*Linolenic.
SO2 = .70798*Stearic + .35853*Oleic.
O3 = .88088*Oleic.
O2P = .73053*Palmitic + .68349*Oleic.
POS = .77737*Stearic + .21991*Palmitic.
L2P+PoLO = .76888*Oleic.
P2L+PPoO = .47412*Palmitic + .32170*Linoleic + .31596*Palmitoleic.
OP2 = .75464*Palmitic.
S2O = .54157*Stearic.

Verifying regression equations by a test set

A problem one may find using Regression Analysis is that good equations can be obtained with a data set, but when these equations are applied to a test set, results are not good enough.

A pragmatic suggestion for verifying a function is to divide the data set in two, using one half for analysis and the other for validation (Bisquerra, 1989). So, it was made a computer program to split our 125 samples in two random groups, one with 75 samples and the other with 50. The first group was used to build equations discussed above. Verification was carried out applying those equations to the second group and then determining the number

of samples whose triacylglycerols were within the confidence limits calculated for each equation.

Table VII shows the confidence limits and the percentages of samples correctly classified within them. T-Student was calculated for p=0,05. As can be observed, equations are not influenced by the set of data with which they were obtained.

Table VII
Verification of regression equations. Confidence boundaries and percentages of correct classification for a=0,05.

Triacylglycerol	Confidence boundaries	Correct classifications
O2L	t*0,916	90%
LOP+PoO2	t*0,577	96%
L2O+LnO2	t*0,309	96%
SO2	t*0,572	94%
O3	t*1,997	92%
O2P	t*1,032	92%
POS	t*0,181	96%
L2P+PoLO	t*0,213	86%
P2L+PPoO	t*0,203	98%
OP2	t*0,338	99%
S2O	t*0,131	88%

So, Regression Analysis seems a good technique to estimate triacylglycerols from only total fatty acids, and it can allow economizing one chemical analysis.

ACKNOWLEDGEMENT

The authors wish to thank Drs. Enrique Graciani, Head of the Quality and Characterization Department, and Arturo Cert, Manager of the Chemical Analysis Laboratory, for providing HPLC and Gas Cromatography data sets and suggestions about chemical interpretation of results. D. Manuel Rodríguez Aguilar and Dña. Carmen Arévalo for their collaboration in the quantification of the samples. This work was supported by CYCIT-SPAIN ALI88-0187-CO202.

REFERENCES

Bisquerra, R. (1989).- "Introducción conceptual al análisis multivariable".- PPU.

Graciani, E. (1988).- "Caracterización del aceite de oliva virgen español. IV. Comparación de su composición en triacilgliceroles con otros aceites y grasas".- Grasas y Aceites **49**, 163-173.

Gunstone, F.D. (1967).- "An introduction to the chemistry and Biochemistry of fatty acids and their glycerides".- 168-171. Chapman and Hall.

Peña, D. (1986).- "Estadística. Modelos y Métodos".- Alianza Editorial, Madrid.

Tabachnick, B.G. and Fidell, L.S. (1983).- "Using Multivariate Statistics".- Harper and Row, Publishers, New York.

UNE 55037 Norm.- "Determination of fatty acids by Gas Chromatography".

UNE 55079 Norm.- "Qualitative and quantitative determination of beta-position fatty acids".

(Recibido: Noviembre 1991).