Available online at www.sciencedirect.com
SCIENCE DIRECT®
Talanta 66 (2005) 1350–1354
Talanta
www.elsevier.com/locate/talanta

Classification of aniseed drinks by means of cluster, linear discriminant analysis and soft independent modelling of class analogy based on their Zn, B, Fe, Mg, Ca, Na and Si content

J.M. Jurado, A. Alcázar, F. Pablos *, M.J. Martín, A.G. González

Department of Analytical Chemistry, Faculty of Chemistry, University of Seville, E-41012 Seville, Spain

Received 21 October 2004; received in revised form 18 January 2005; accepted 28 January 2005
Available online 19 February 2005

Abstract

Zinc, boron, iron, magnesium, calcium, sodium and silicon were determined in aniseed drinks by inductively coupled plasma-atomic emission spectrometry (ICP-AES). These elements were considered as chemical descriptors to characterise Spanish-certified aniseed drinks brands of origin. Different pattern recognition (PR) procedures were applied for these purposes. This chemometric study have included methods like principal component analysis (PCA), non-supervised PR techniques such as cluster analysis (CA), supervised PR methods of hard modelling like linear discriminant analysis (LDA) and soft independent modelling of class analogy (SIMCA).

Keywords: Boron; Iron; Magnesium; Calcium; Sodium; Silicon; Zinc; Chemometrics; Linear discriminant analysis; Cluster analysis; SIMCA; Multivariate analysis; Aniseed drinks; Inductively coupled plasma-atomic emission spectrometry

1. Introduction

Trace elements can be very useful as markers for identification of the product's geographical origin and authenticity [1–6]. Assessment of food sample's origin has been mostly conducted through multivariate analysis in combination with pattern recognition techniques [7–10]. Aniseed drinks are obtained by distillation of pressed fermented grapes, dregs and other fermented saccharate raw materials, flavoured with star aniseed (*Illicium verum*), green aniseed (*Pimpinella anisum*), fennel (*Foeniculum vulgare*) or some other plants. This kind of beverage is very popular in Spain and it has been produced for centuries. It is produced in many locations around the country and several of them are elaborated under controlled conditions and can be considered quality products. The composition and manufacture of these drinks in the European Union is ruled by the Community Directive 1576/89/EEC [11]. In this Directive, three certified brands of origin are considered for Spanish aniseed drinks, namely Cazalla, Chinchón and Rute. As a consequence, it is of great importance to have analytical methods to differentiate between authentic products and others with different origin that could pass as authentic ones.

In this paper, Zn, B, Fe, Mg, Ca, Na and Si have been determined by inductively coupled plasma-atomic emission spectrometry (ICP-AES) in samples of Spanish aniseed drinks from the three certified brands of origin and others not belonging to these classes. Those elements were used as chemical descriptors to apply methods like principal component analysis (PCA), non-supervised pattern recognition techniques, such as cluster analysis (CA) and supervised techniques like linear discriminant analysis (LDA) and soft independent modelling of class analogy (SIMCA). These chemometric techniques have been used to classify samples belonging to the brands of origin and to differentiate them from samples with other origin.

* Corresponding author. Tel.: +34 954557173; fax: +34 954557168.
E-mail address: fpablos@us.es (F. Pablos).

0039-9140/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2005.01.052

Table 1
Classification and characteristics of the aniseed drinks samples

Code	Number of samples	Certified brand of origin
C	22	Cazalla
Ch	8	Chinchón
R	6	Rute
NC[a]	38	

[a] Samples not included in any of the certified brands of origin.

2. Materials and methods

2.1. Apparatus and software

A Fisons-ARL 3410 inductively coupled plasma-atomic emission spectrometer (FISONS Instruments, Valencia, CA) equipped with a Minitorch and a Meinhard nebulizer was used for metal determinations. The operating conditions were described in a previous work [12]. The statistical package Statistica 5.5 from Statsoft [13] was used for PCA, CA and LDA calculations. SIMCA calculations were performed using SIMCA P software for multivariate modelling and analysis from Umetrics [14].

2.2. Reagents and samples

Titrisol (Merck) stock solutions (1000 mg/l) were used to prepare working standards. Nitric acid (65%) and hydrogen peroxide (30%) of analytical grade, used in the mineralisation procedure, were from Merck. Ultrapure Milli-Q water was always used.

Samples ($N = 74$) were obtained from local stores, belonging to the three typical Spanish-certified brands of origin and others of different origins. All samples were contained in glass bottles and stored at 4 °C until analysis. Table 1 resumes the codes used to identify the samples indicating their geographic origin.

2.3. Determination of boron, iron, magnesium, calcium, sodium, silicon, zinc by ICP-AES

Samples were previously mineralised by treating them with a mixture of HNO_3/H_2O_2 (10:1) according to the conditions established in a previous work [12]. The obtained solutions were suitably diluted and filtered before the ICP-AES determinations.

3. Discussion and results

3.1. Determination of Zn, B, Fe, Mg, Ca, Na and Si in aniseed drinks by ICP-AES

Table 2 shows the obtained contents of above-cited elements in the analysed samples. It can be observed that Na was the metal with higher contents ranging between 215.322 and 1.267 mg/l. Ca, Mg and Si presented values in the intervals 9.518–0.195, 11.911–0.012, 5.096–0.169 mg/l, respectively. Zn, B and Fe were present at lower concentrations, with values less than 0.178 mg/l. Considering the average values in each group of samples maximum differences were obtained for the contents in Ca, Mg, Na, Zn and Si. In a first approach, these elements could be considered as good chemical descriptors to differentiate samples from different origins. In this way, a chemometric study was carried out to discriminate between the considered certified brands of origin.

3.2. Classification of aniseed drinks

In order to construct the classification model, samples belonging to the known certified brands of origin (classes C, R and Ch) were considered. Consequently, for chemometric calculations a data matrix with seven columns, corresponding to the determined elements and 36 rows, corresponding to the analysed aniseed drinks with certified brand of origin was built. Auto scaled data were used in all calculations. The remaining samples, labelled as NC, have a very heterogeneous origin and do not correspond to any of the previous classes. These samples were used to check the efficiency of the classification model.

3.2.1. Principal component analysis

In order to find possible tendencies in the samples and the discriminant power of the variables, PCA was applied. By using PCA, the *n*-dimensional data set can be plotted in a smaller number of dimensions, usually 2 or 3. This allows the observation of groupings of cases, which can define the structure of the data set. PCA finds the maximum variations in the data set and forms new variables known as principal components (PCs). Each successive PC accounts for as much of the remaining variability as possible and each new variable must be totally independent of all other variables [15]. After applying PCA to our data set, three PCs were extracted. The percentage of variance explained by each PC is 44.80, 27.11 and 16.79%, respectively. According to the loadings of the variables in the first PC (Table 3), the most contributing descriptors were Si, Mg, Ca and Zn. On the other hand, the correlation matrix shows strong correlations between Zn, Ca and Mg and also between Na and Si. Thus, Mg and Si explain, by themselves, the observed variance and can be considered the most discriminant variables. When representing the scores of the samples in the three-dimensional space defined by the calculated PCs (Fig. 1), several groups of samples appear. Samples of the same certified brand of origin show certain trends to be grouped thought this fact do not happen in all cases.

3.2.2. Cluster analysis

In order to assess this tendency, a hierarchical agglomerative cluster analysis of samples was performed [16]. This procedure finds natural groupings of the data set. According to PCA findings, Mg and Si were found to be the descriptors with more contribution. Consequently, cluster analysis was

Table 2
B, Ca, Fe, Mg, Na, Si, Zn average concentrations (mg/l) and ranges found in aniseed drinks

Origin	C	Ch	R	NC
B	0.032 ± 0.038 (0.001–0.112)	0.005 ± 0.002 (0.002–0.006)	0.004 ± 0.001 (0.003–0.005)	0.030 ± 0.032 (0.003–0.104)
Ca	0.989 ± 0.364 (0.557–1.748)	0.568 ± 0.197 (0.302–0.860)	2.862 ± 0.153 (2.634–3.030)	4.399 ± 2.660 (0.195–9.518)
Fe	0.026 ± 0.019 (0.004–0.087)	0.059 ± 0.035 (0.025–0.115)	0.030 ± 0.009 (0.017–0.039)	0.049 ± 0.043 (0.003–0.178)
Mg	0.393 ± 0.175 (0.119–0.670)	0.029 ± 0.011 (0.012–0.045)	2.246 ± 0.106 (2.163–2.447)	5.551 ± 3.522 (0.133–11.911)
Na	40.041 ± 12.908 (23.350–64.497)	1.609 ± 0.163 (1.267–1.765)	17.076 ± 0.917 (16.048–18.488)	46.783 ± 70.374 (8.384–215.322)
Si	2.378 ± 0.569 (1.225–3.354)	0.292 ± 0.059 (0.180–0.340)	0.189 ± 0.014 (0.169–0.221)	2.353 ± 1.263 (0.639–5.096)
Zn	0.015 ± 0.012 (0.003–0.05)	0.019 ± 0.002 (0.017–0.023)	0.051 ± 0.029 (0.033–0.103)	0.031 ± 0.026 (0.002–0.107)

Average ± standard deviation. Concentration ranges in brackets.

Table 3
Factor loadings in the three first principal components

Variable	Factor 1	Factor 2	Factor 3
Zn	0.724086	−0.453204	−0.321403
B	−0.376071	−0.630008	−0.526243
Fe	0.187933	0.373333	−0.832462
Mg	0.819599	−0.416525	0.311755
Ca	0.762437	−0.603528	−0.016428
Na	−0.681718	−0.640812	0.068043
Si	−0.846836	−0.455521	0.006301

applied by using Mg and Si as variables. Taking the Euclidean distance as similarity measurement and the Ward's method as amalgamation rule, three clusters were obtained. Fig. 2 shows the corresponding dendrogram, where C, R and Ch samples are grouped in three different clusters. In the amalgamation schedule, the samples were grouped in clusters that correspond to their brand of origin membership. As cluster analysis was carried out using two variables, Mg and Si, similar information could be obtained from a variable–variable plot, as can be seen in Fig. 3.

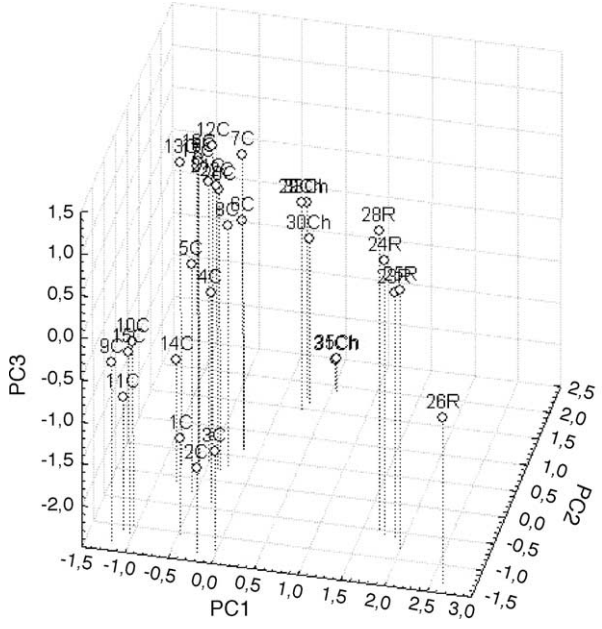

Fig. 1. Scores plot of the aniseed drinks samples in the three-dimensional space of the first PCs. C, Cazalla; R, Rute; Ch, Chinchón.

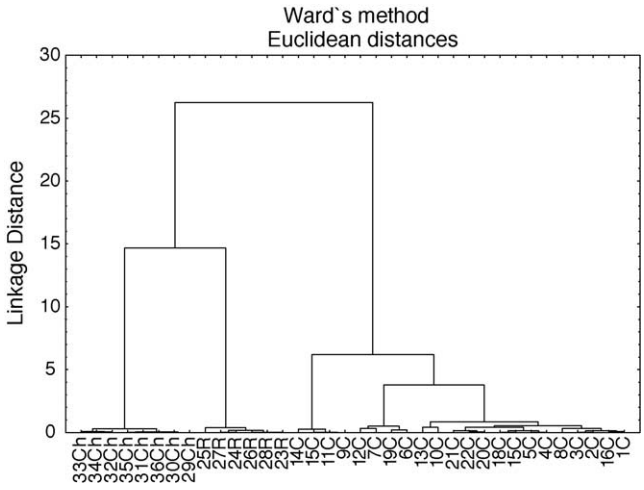

Fig. 2. Dendrogram of aniseed drinks samples using Mg and Si as variables. C, Cazalla; R, Rute; Ch, Chinchón.

3.2.3. Linear discriminant analysis

Because of the a priori knowledge of the class membership of the samples, it is possible to apply supervised PR methods to the data. In this case, the 36 samples of aniseed drinks belong to three different classes, corresponding to the brands of origin: Cazalla (C), Rute (R) and Chinchón (Ch). LDA enables to calculate discriminant functions as linear combinations of the selected descriptors, which maximize the ratio

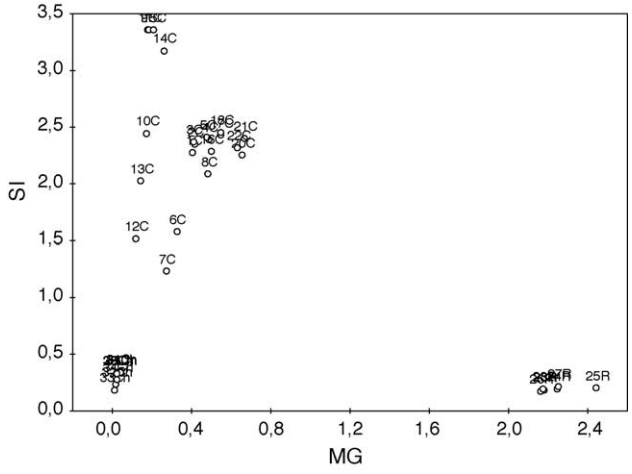

Fig. 3. Mg–Si plot.

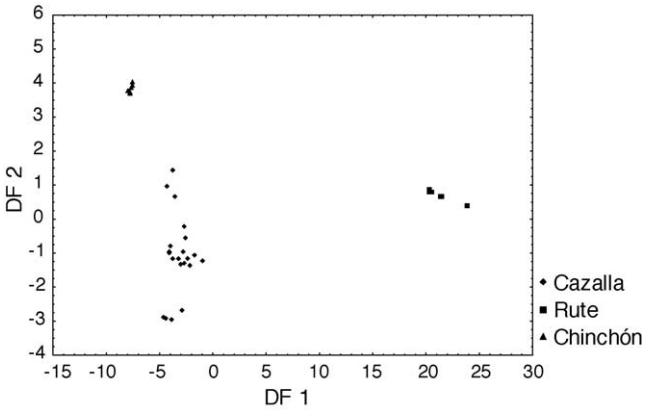

Fig. 4. Scatter plot of the aniseed drinks samples in the space of the two discriminant functions.

between-classes sum of squares and the within-classes sum of squares [17]. After applying LDA, two discriminant functions were obtained. Fig. 4 shows the distribution of the samples in the plane of the two calculated discriminant functions. As it can be observed a complete separation of the three considered classes was achieved. The recognition ability, according to the a posteriori probabilities were 100% for each class. The leave one out method [18] was used as cross-validation procedure to evaluate the classification performance. The prediction abilities were of 100% for all classes. Though an excellent classification was achieved, false-positive or false-negative can be present but not detected. A soft modelling method can be applied to detect them.

3.2.4. Soft independent modelling of class analogy

SIMCA [19] constructs a PC model for each class separately in the training set. The perpendicular distance of any object to the hyperplane defined by the first PCs is used for classification purposes. Every considered sample is assigned to one class according to its distance from the class model. In our case, each class was fitted to two PCs, leading to two-dimensional hyperboxes with a critical normal distance for assignation purposes. SIMCA, being a soft modelling procedure, enables us to detect the number of false-positive/negative for each class. For these purposes, two parameters sensitivity (SENS) and specificity (SPEC) [20,21] were used to validate the classification procedure. SENS of a class is referred to the number of objects belonging to this class that are correctly classified. SPEC of a class corresponds to the number of objects not belonging to this class that are correctly considered as belonging to different classes. These features can be easily transformed to the number of false-positive (FP) and of false-negative (FN) according to the following rules for an ideal class A:

$$\%\mathrm{FN} = 100\left(1 - \frac{\langle n_\mathrm{A}\rangle}{n_\mathrm{A}}\right), \qquad \%\mathrm{FP} = 100\left(1 - \frac{\langle n'_\mathrm{A}\rangle}{n'_\mathrm{A}}\right)$$

where n_A is the number of objects belonging to class A, n'_A the number of objects not belonging to class A, $\langle n_\mathrm{A}\rangle$ the number of objects belonging to class A and correctly classified as of A, and $\langle n'_\mathrm{A}\rangle$ the number of objects not belonging to class A and correctly classified as non-A. Accordingly, in our study, the percentages of FN and FP for each studied class were calculated, being the number of false-positive/negative less than 5%, even zero in the case of R and Ch samples. In Table 4, confusion matrix of the model indicating the correct/incorrect predictions for each class and also for the NC group of samples is presented.

Table 4
Confusion matrix for the three considered classes

	Cazalla		Rute		Chinchón	
Predicted/actual	–	+	–	+	–	+
–	52	0	66	0	68	0
+	1	21	0	8	0	6

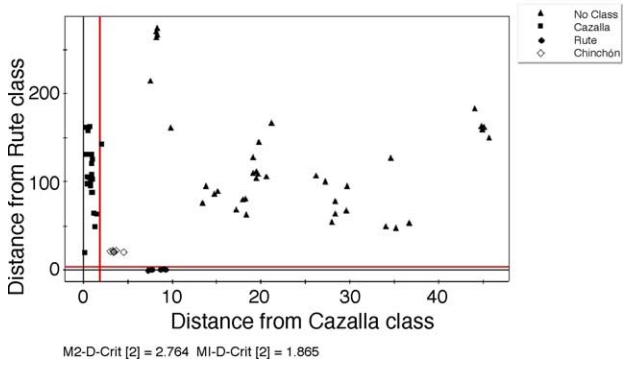

Fig. 5. Cooman's plot.

The separation of the studied classes can be easily shown by using the Cooman's plot, as it can be seen in Fig. 5. Moreover, due to the soft-modelling features of SIMCA, new samples (belonging or not to one of the studied classes) can be tested for class memberships. If we consider the set of samples without certified brand of origin (NC), none of them are assigned to the three classes according to SIMCA. In Fig. 5, these no class (NC) samples appear far from the class boxes. This fact confirms the efficiency of the constructed model for the classification of aniseed drinks of the Cazalla, Rute and Chinchón brands of origin.

4. Conclusions

From these results, it can be concluded that the contents of Zn, B, Fe, Mg, Ca, Na and Si are good chemical parameters to carry out the authentication of the three brands of origin of Spanish aniseed drinks. Mg and Si are the most discriminant variables for this purpose.

References

[1] E.A. Hernández-Caraballo, R.M. Avila-Gómez, T. Capote, F. Rivas, A.G. Pérez, Talanta 60 (2003) 1259.
[2] A. Alcázar, F. Pablos, M.J. Martín, A.G. González, Talanta 57 (2002) 45.

[3] M.J. Martín, F. Pablos, A.G. González, Anal. Chim. Acta 358 (1998) 177.
[4] P.L. Fernández, M.J. Martín, F. Pablos, A.G. González, J. Agric. Food Chem. 49 (2001) 4775.
[5] S. Frias, J.E. Conde, J.J. Rodríguez, F. García, J.P. Pérez, Talanta 59 (2003) 335.
[6] R. Kokkinofta, P.V. Petrakis, T. Mavromoustakos, C.R. Theocharis, J. Agric. Food Chem. 51 (2003) 6233.
[7] J.R. Piggot (Ed.), Statistical Procedures in Food Research, Elsevier, London, 1986.
[8] M.S. Valdenebro, M. León-Camacho, F. Pablos, A.G. González, M.J. Martin, Analyst 124 (1999) 999.
[9] D. González-Arjona, V. González-Gallero, F. Pablos, A.G. González, Anal. Chim. Acta 381 (1999) 257.
[10] A. Terrab, A.G. González, M.J. Diez, F. Heredia, J. Sci. Food Agric. 83 (2003) 637.
[11] Directive (CEE) No. 1576/89, Off. J. Eur. Commun. L160 (12-6-1989). http://europa.eu.int/eur-lex/en/index.html.
[12] J.M. Jurado, A. Alcázar, F. Pablos, M.J. Martín, A.G. González, Talanta 63 (2004) 297.
[13] Statsoft, Statistica for Windows, Computer Program Manual, Tulsa, OK, 1999. http://www.statsoft.com.
[14] Umetrics AB, SIMCA-P 9, User Guide and Tutorial, Umeà, 2001. http://www.umetrics.com.
[15] C. Chatfield, A.J. Collins, Introduction to Multivariate Analysis, Chapman & Hall, London, 1980.
[16] D.L. Massart, L. Kauffman, Interpretation of Analytical Data by Use of Cluster Analysis, Wiley, New York, 1992.
[17] D. Coomans, D.L. Massart, L. Kaufman, Anal. Chim. Acta 112 (1979) 97.
[18] R. Henrion, G. Henrion, Überwachte klassifikation in multivariate datenanalysen, Springer-Verlag, Berlin, 1995.
[19] S. Wold, C. Albano, W.J. Dunn, U. Edlunb, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, M. Sjöström, in: B.R. Kowalski (Ed.), Chemometrics, Mathematics and Statistics in Chemistry, Reidel, Dordrecht, 1984.
[20] D. González-Arjona, G. López-Pérez, A.G. González, Chemom. Intell. Lab. Syst. 57 (2001) 133.
[21] M. Forina, C. Armanino, R. Leardi, G. Drava, J. Chemom. 5 (1991) 435.