Pattern Recognition Analysis Applied to Classification of Honeys from Two Geographic Origins

R. Peña Crecente and C. Herrero Latorre*

Departamento Qulmica Analltica, Nutrici6n y Bromatologfa, Facultad de Ciencias de Lugo, Augas Férreas s/n,
27002 Lugo, Spain

Eleven legal parameters of quality control of honey were determined in 67 honey samples from Galicia (northwestern Spain) which were obtained from two production areas: Lugo and Orense. Classification of these honeys according to their geographic origin was achieved by pattern recognition techniques to the chemical data. Humidity and free acidity were found to be the most important features for the classification. In this case the use of pollen data to achieve a correct geographic classification of the honey samples is not necessary.

INTRODUCTION

Pattern recognition techniques are widely applied to food chemistry problems (Forina and Lanteri, 1984). These include the geographic classification of olive oils (Derde et al., 1984), concentrated orange juices (Bayer et al., 1980), and wines (Kwan and Kowalski, 1978; 1980; Maarse et al., 1987; Etievant et al., 1988; Vasconcelos and Chaves, 1989; Herrero and Mèdina, 1990; Herrero et al., 1992), authentication of Scotch whiskies (Saxberg et al., 1978) and sherry wines (Van der Schee et al., 1989) and analysis of sensory evaluations (Kwan and Kowalski, 1980). In this work, the application of principal component analysis (PCA), linear discriminant analysis (LDA), *K* nearest neighbor (KNN), and soft independent modeling of class analogy (SIMCA) to the chemical data obtained for legal quality control of honeys before commercialization permits the differentiation between two geographic origins of honey samples: Lugo and Orense.

EXPERIMENTAL PROCEDURES

Honey Samples. Description of the origin of the 67 honey samples is given in Table I. Samples from Lugo were provided by the local association of beekeepers. Only 12 samples from Orense with guaranteed geographic origin were available for this work. All samples examined (Lugo and Orense) were honeys of random (mixed) floral type from the 1990 harvest.

Quality Control Analysis. The parameters determined in each honey sample are listed in Table II. All analyses were performed according to the official methods of Spanish legislation (AOAC, 1990; BOE, 1986). The results of these determinations are summarized in Table III; the levels were similar to those found by other authors (Huidobro, 1983, 1984; Sancho, 1990). The only exception was the free acidity, which we found slightly higher in the samples from the Orense area.

Pattern Recognition Analysis. Each honey sample (object) was considered as an assembly of variables represented by the chemical data. These variables, called "features", formed a "data vector" which represented a honey sample. Data vectors belonging to the same group, such as geographic origin, were analyzed; the group was then termed a "category". Pattern recognition techniques used in this work were as follows.

PCA. This procedure (Mardia et al., 1979) was mainly used to achieve a reduction of dimension, i.e., to fit a K-dimensional subspace to the original p-variate ($p > K$) objects and permit a primary evaluation of the between-category similarity.

Table I. Local Origin of 67 Honey Samples

local area	no. of samples
Honey Samples from Lugo	
Serras Orientias	6
Lemos	5
Terra Cha	11
Municipio de Lugo	20
Mariña	13
Honey Samples from Orense	
Bande	12

Table II. Determined Quality Control Parameters

parameter	method	parameter	method
humidity water content	refractometric[a]	free acidity	volumetric[a]
ash	gravimetric[a]	lactonic acidity	volumetric[a]
insoluble matter	gravimetric[b]	total acidity	volumetric[a]
reducing sugars	volumetric[a]	hydroxymethylfurfural	spectrophotometric[b]
sucrose	volumetric[a]	pH	potentiometric[b]
conductivity	conductimetric[b]		

[a] AOAC (1990). [b] BOE (1986).

LDA. This classification procedure (Romeder, 1973) maximizes the variance between categories and maximizes the variance within categories. The method renders a number of orthogonal linear discriminant functions, equal to the number of categories minus 1. In this case of two types of geographic origins, a one-dimensional figure was obtained that was easily interpretable.

KNN. This classification method, which utilizes the distance between objects in the p-space as its criterion (Cover and Hart, 1967), is used to classify an object in the category which contributes to the greatest number of the k nearest known objects. The Euclidean n-space distance and its inverse square were used in this work. Only the closest k objects are used in making any given classification. The importance of a given feature in making the decisions is proportional to its contribution to the distance calculation.

SIMCA. This classification procedure uses linear discriminant functions derived from disjointed principal component analysis of the data (Wold, 1976). One set of functions is derived for each category studied by computing the category mean and a specified number of the principal components. Objects are classified into the category whose principal component model best reproduces the data. Only data points which are members of a given category are used in determining the model functions for that category. The importance of each feature in classification is determined by its contribution to the category covariance matrices.

Pattern recognition analyses were performed by means of statistical software packages STATGRAPHICS (Statistical Graphics Corp., 1987) and PARVUS (Forina et al., 1988) in a IBM PS/2 Model 80 computer using a HP Laserjet II as graphic output.

0021-8561/93/1441-0560$04.00/0

Table III. Summary of the Results Obtained for the Quality Control Parameters Determined[a]

variable	samples from Lugo				samples from Orense			
	mean	SD	max	min	mean	SD	max	min
humidity, %	17.5	1.1	20.6	15.7	16.8	1.0	18.3	15.2
ash, %	0.37	0.20	1.01	0.03	0.57	0.08	0.74	0.43
insoluble matter, %	0.028	0.030	0.140	0.001	0.026	0.016	0.060	0.010
reducing sugars, %	69.85	2.62	78.30	61.65	67.11	1.86	69.60	64.70
sucrose, %	1.68	1.16	6.29	0.20	1.89	0.86	3.80	0.60
conductivity, μS/cm	277	106	644	66	278	31	340	221
free acidity, mequiv/kg	30.96	5.81	42.80	16.94	48.83	6.55	58.30	39.33
lactonic acidity, mequiv/kg	4.46	2.87	12.80	0.10	3.58	1.01	5.49	2.24
total acidity, mequiv/kg	35.42	6.76	50.10	20.00	52.41	6.56	61.30	42.56
HMF, mg/kg	8.9	6.4	49.7	0.8	7.5	3.7	13.4	1.7
pH	4.08	0.34	4.77	3.46	4.43	0.16	4.64	4.10

[a] SD, standard deviation. Max, maximum. Min, minimum.

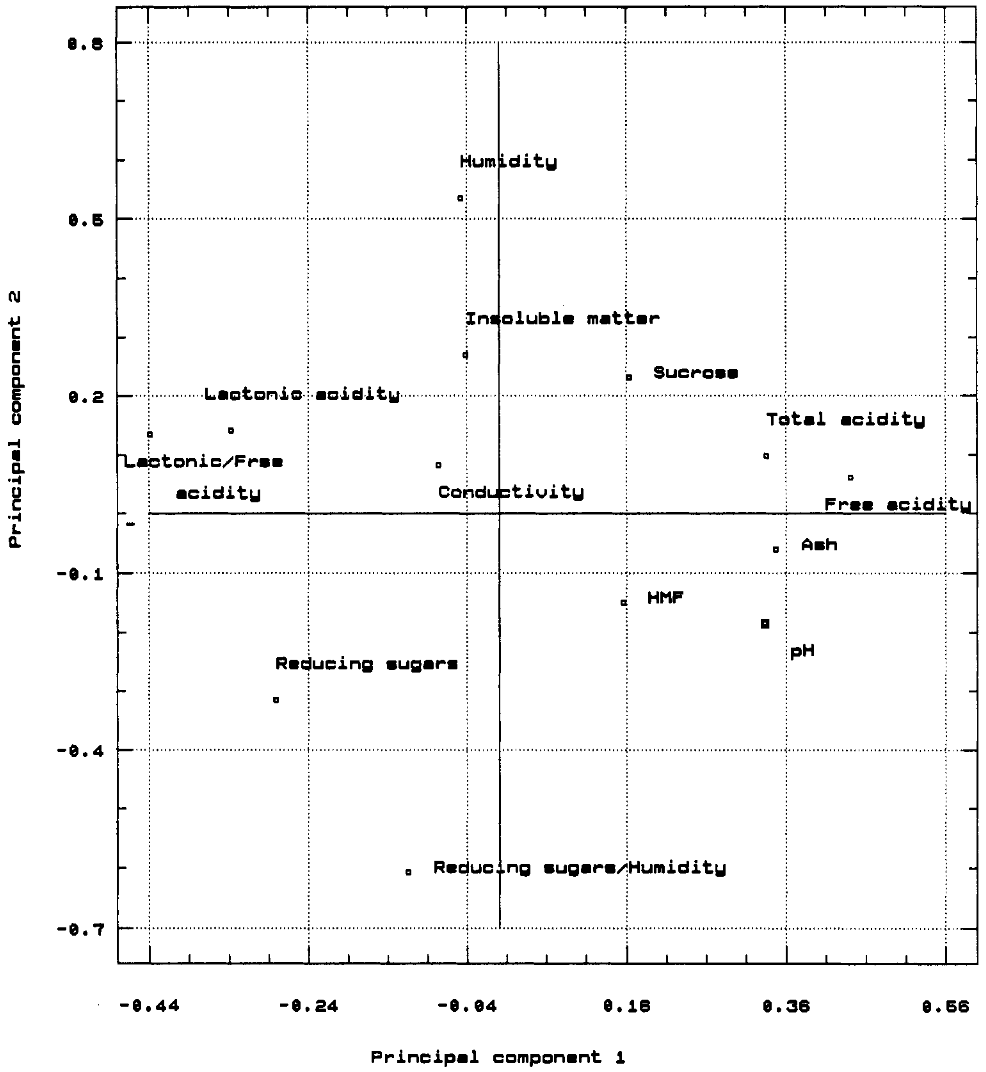

Figure 1. Plot of component weights in first principal component vs component weights in second principal component from PCA of 13 features.

RESULTS AND DISCUSSION

Classification based on geographic origin was performed in the data set. There were 13 features for each data vector, the 11 parameters listed in Table II plus the free acidity/total acidity ratio and the humidity/reducing sugars ratio.

Principal components were calculated by using a routine of STATGRAPHICS. A scatter plot was obtained which correlates the weighting factors of features in the first principal component vs the weighting factors in the second principal component. It can be seen from Figure 1 that free acidity and the lactonic acidity/total acidity ratio are the dominant features in the first principal component, while humidity and reducing sugars/humidity ratio strongly dominate the second principal component. Both components account for 53% of the total variability. Chemically the first principal component can be associated with the acidity of the honey, just as the second can be associated with its water content.

Group classification of PCA provided interesting results. When a three-dimensional plot of the principal components 1–3 was drawn, a clear separation of the objects (honey samples) into two groups corresponding to two geographic origins was achieved (Figure 2).

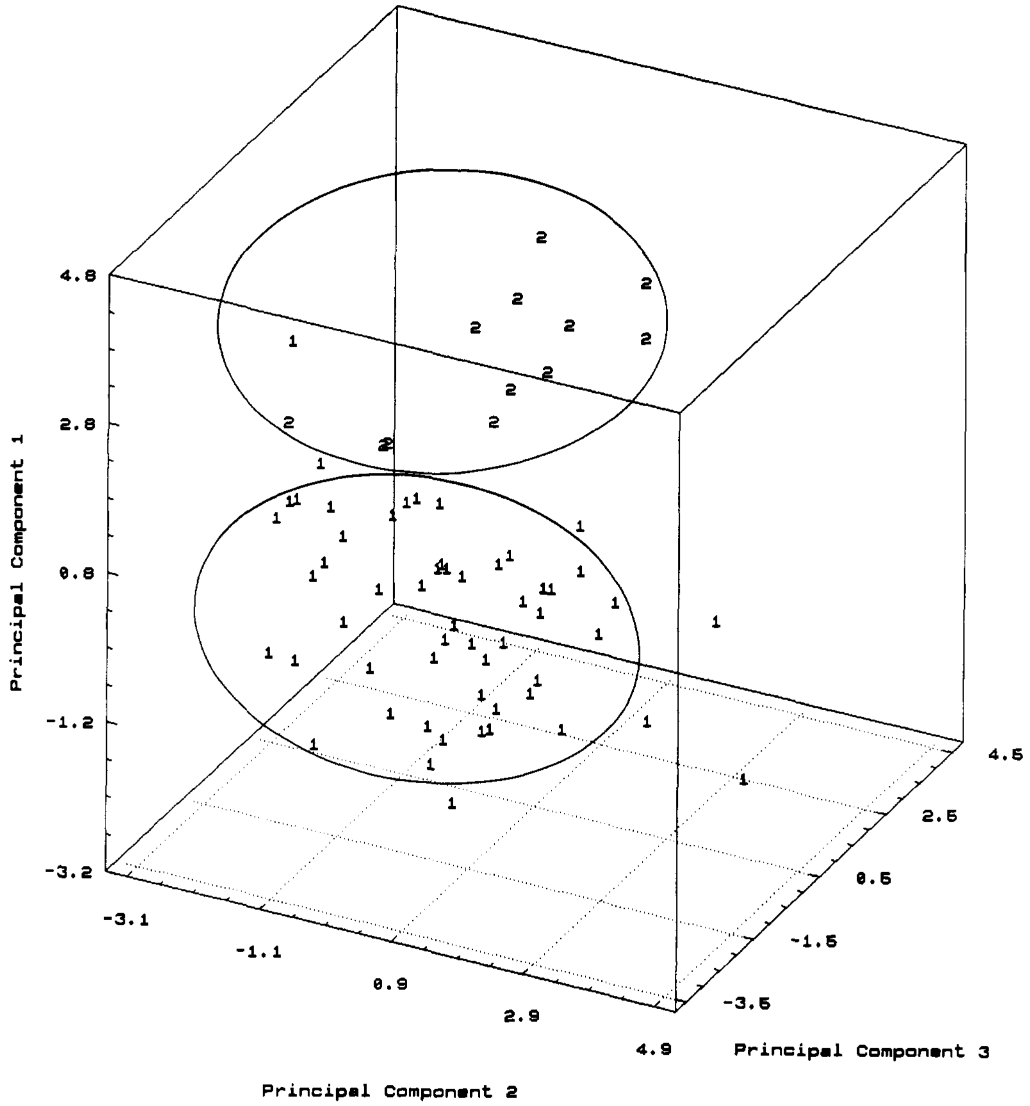

Figure 2. Plot of the first three principal components in PCA of honeys from two geographic origins (1, Lugo; 2, Orense).

Usually, it would be ideal to maintain a data vector/feature ratio at a value higher than 3. This could only be achieved in this case (only 12 unsuspected samples from Orense were obtained) by choosing the features which contained the most discriminant information for the classification. Stepida routine available from PARVUS package was designed for the task. This subprogram selects the relevant variables for classification purposes according to the pooled model of linear discriminant analysis, choosing the variables producing the greatest Mahalanobis distance between two categories. Selection of a small number of key features offered other advantages besides increasing the reliability of mathematical classification results; two-dimensional plots of the key features, one vs the other, allowed visual examination of the data set. The first feature chosen by Stepida to be most important was free acidity, while the second selected feature was humidity. These results agree with the conclusions obtained by PCA, where free acidity and humidity were the dominant features in each principal component.

Three classification methods, LDA, KNN, and SIMCA, were applied to the data set using the selected features of free acidity and humidity. The 67 objects were randomly divided between training (or learning) set and evaluation (or prediction) set; the percentage of objects placed in the evaluation set was 25% for each category. To obtain a good evaluation of the predictive ability of each method, the previous division procedure was repeated 10 times for different constitutions on the two sets. Table IV shows the results of the three applied methods. LDA and SIMCA attained results with, respectively, 3 and 5 misclassifications of a total of 67 honey samples. KNN using Euclidean distance presents three objects misclassified; however, using inverse square Euclidean distance, only one object was incorrectly classified. Therefore, it was concluded that by using only the first two selected legal quality control parameters, satisfactory results in classification of honey samples were obtained; however, a number of additional selected features were required before classification results could be improved to 100%.

Table IV. LDA, KNN, and SIMCA Results for Geographic Classification

classification method	% correct assignation (Lugo/Orense)	misclassified samples
LDA (two selected features)	95/100	3
KNN (two selected features)	100/75[a]	3
$K = 5$	100/92[b]	1
SIMCA (two selected features)	91/100	5

[a] Using Euclidean distance. [b] Using inverse square of Euclidean distance.

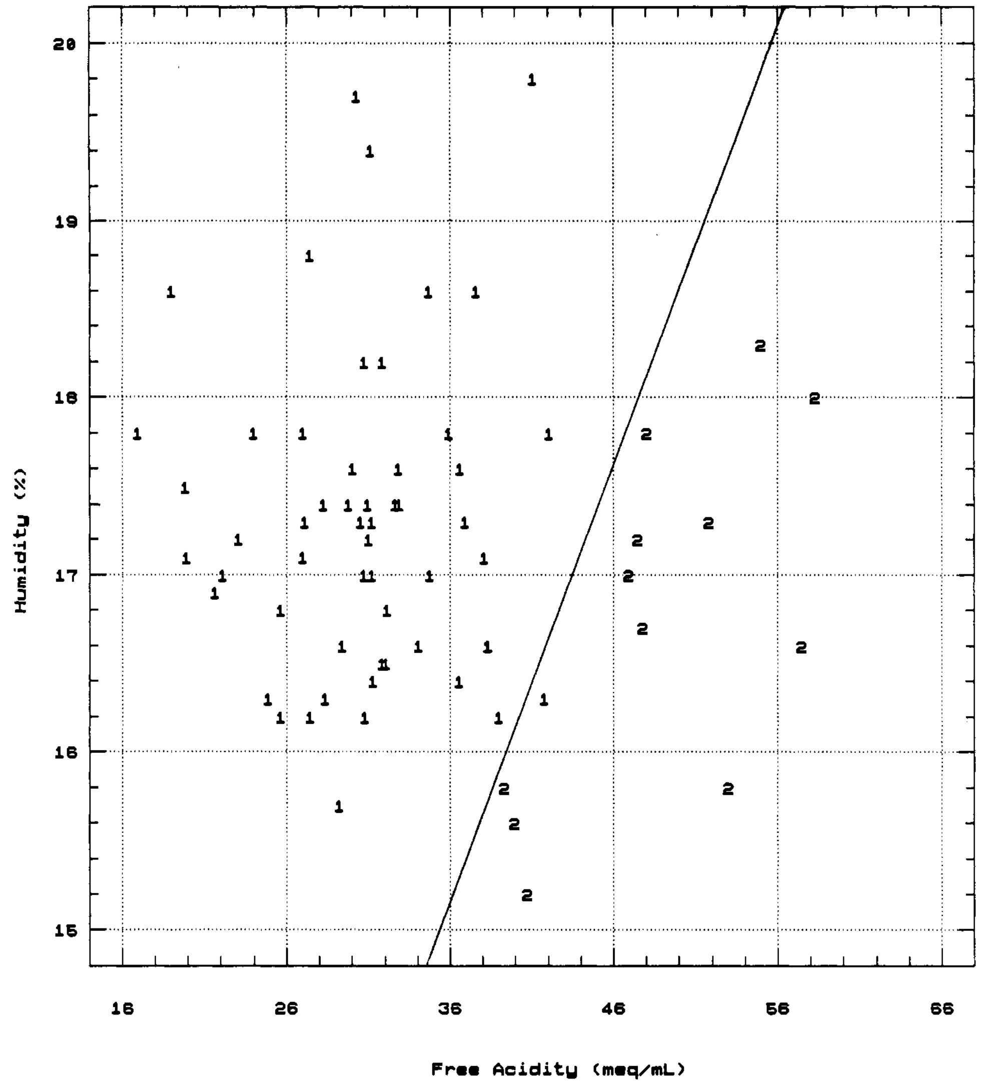

Figure 3. Plot of the first selected feature, free acidity, vs second selected feature, humidity, for classification of honeys from two geographic origins (1, Lugo; 2, Orense).

The data were then examined by a two-dimensional plot of humidity vs free acidity (Figure 3). One Orense sample was misclassified, confirming the results of KNN.

CONCLUSION

Pattern recognition is able to extract useful information from a massive amount of data. Information was used to relate chemical compositions of honeys and their geographic origin. Classification of honeys from Lugo and Orense was made possible by using only two legal quality control features, free acidity and humidity, and pollen studies were not necessary to achieve this objective. A two-dimensional plot of these legal quality control parameters was sufficient to solve the classification problem and may be used to authenticate honeys from two geographic origins.

ACKNOWLEDGMENT

We express our gratitude to the Asociación Lucense de Apicultura for providing honey samples from Lugo. We are grateful to A. Chouciño (The University of Kansas) for her help. This work was partly financed by the fund for scientific research of the University of Santiago, Spain.

LITERATURE CITED

AOAC. *Official Methods of Analysis*, 15th ed.; Helrich, K., Ed.; Association of Official Analytical Chemists: Arlington, VA, 1990.

Bayer, S.; McHard, J. A.; Winefordner J. Determination of geographic origins of frozen concentrated orange juices via pattern recognition. *J. Agric. Food Chem.* **1980**, *28*, 1306.

BOE. "Orden de 12 de Junio de 1986 por la que se aprueban los métodos oficiales de análisis para la miel"; Boletín Oficial del Estado 18 Junio: Madrid, 1986.

Cover, T. M.; Hart, P. E. *IEEE Trans. Inf. Theory* **1967**, *IT13*, 21.

Derde, M. P.; Coomans, D.; Massart, D. L. SIMCA Demonstrated with Characterization and Classification of Italian Olive Oils. *J. Assoc. Off. Anal. Chem. Symp. Ser.* **1984**, *18*, 49.

Etievant, P.; Schilich, P.; Bouvier, J.; Symonds, P.; Bertrand, A. Varietal geographic classification of french red wines in terms of elements, amino acids and aromatic alcohols. *J. Sci. Food Agric.* **1988**, *45*, 21.

Forina, M.; Lanteri, S. Data analysis in food chemistry. In *Chemometrics, Mathematics and Statistics in Chemistry*; Kowalski,, B. R., Ed.; Riedel: Dordrecht, Holland, 1984.

Forina, M.; Laerdi, R.; Armanino, C.; Lanteri, S. *PARVUS, an extendable package of programs for exploration, classification and correlation*; Elsevier: Amsterdam, 1988.

Herrero, C.; Mèdina, B. Classification of wines from Galicia and other regions of Spain using trace elements. *Connaiss. Vigne Vin* **1990**, *24*, 147.

Herrero, C.; Mèdina, B.; Latorre, M. Characterization of wines from Galicia using trace elements. *Connaiss. Vigne Vin* **1992**, *26*, 185.

Huidobro, J. F. Ph.D. Dissertation, Universidad de Santiago de Compostela, Spain, 1983.

Huidobro, J. F. *Bol. Inf. Agrar.* **1984**, January, 86–96.

Kwan, W. O.; Kowalski, B. R. Classification of wines applying pattern recognition to chemical composition data. *J. Food Sci.* **1978**, *13*, 1320.

Kwan, W. O.; Kowalski, B. R. Correlation of objective chemical measurements and subjective sensory evaluations. Wines of vitis vinifera Pinot Noir fron France and the United States. *Anal. Chim. Acta* **1980a**, *122*, 215.

Kwan, W. O.; Kowalski, B. R. Pattern recognition analysis of gas chromatographic data. Geographic classification of Vitis Vinifera cv. Pinot Noir from France and the United States. *J. Agric. Food Chem.* **1980b**, *28*, 356.

Maarse, H.; Slump, P.; Tas, A.; Schaefer, J. Classification of wines according to type and region based on their composition. *Z. Lebensm. Unters. Forsch.* **1987**, *184*, 198.

Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic Press: New York, 1979.

Romeder, J. M. *Mèthodes et programmes d'analyse discriminante*; Bordas: Paris, 1973.

Sancho, M. T. Ph.D. Dissertation, Universidad de Santiago de Compostela, Spain, 1990.

Saxberg, B. F.; Deuwer, D. L.; Booker, J. L.; Kowalski, B. R. Pattern recognition and blind assay techniques applied to forensic separation of whiskies. *Anal. Chim. Acta* **1978**, *103*, 201.

STATGRAPHICS User's Guide; Statistical Graphics Corp.: Rockville, MD, 1987.

Van der Schee, H. A.; Bouwknegt, J. P.; Tas, A.; Maarse, H.; Sarneel, M. M. The authentication of sherry wines using pattern recognition: an interlaboratory study. *Z. Lebensm. Unters. Forsch.* **1989**, *188*, 324.

Vasconcelos, P.; Chaves, H. Characterization of elementary wines of vitis vinifera varieties by pattern recognition of free amino acid profiles. *J. Agric. Food. Chem.* **1989**, *37*, 931.

Wold, S. Pattern recognition by means of disjoint principal components. *Pattern Recognit.* **1976**, *8*, 127.

Received for review October 30, 1992. Accepted December 28, 1992.