Multivariate analysis applied to near infra red spectra of milk. Anal. Chem., 59, 2187-91

ARTICLE *in* ANALYTICAL CHEMISTRY · OCTOBER 1987

Impact Factor: 5.64 · DOI: 10.1021/ac00144a038 · Source: PubMed

4 AUTHORS, INCLUDING:

Dominique Bertrand
data_frame
190 PUBLICATIONS **2,504** CITATIONS
SEE PROFILE

Marie-Françoise Devaux
French National Institute for Agricultural R…
95 PUBLICATIONS **2,037** CITATIONS
SEE PROFILE

Reprinted from Analytical Chemistry, 1987, 59, 2187.

Multivariate Analysis Applied to Near-Infrared Spectra of Milk

Paul Robert,* Dominique Bertrand, and Marie Francoise Devaux

National Institute of Agronomic Research, Laboratory of Feed Animal Technology, Rue de la Géraudière, 44072 Nantes Cedex 03, France

Rene Grappin

National Institute of Agronomic Research, Milk Experimentation Station, 38800 Poligny, France

The application of near-infrared spectroscopy to the study of milk samples is hindered by two difficulties. The water absorption is very large in comparison with fat, protein, and lactose absorption. Moreover, the fat globules of milk are scattering particles which produce spectral deformations. Different mathematical treatments were tested to reduce the influence of the particle size and to identify characteristic wavelengths of fat, protein, and lactose. Spectral variations due to light scattering particles could be reduced by centering the absorbances. Principal components analyses and correspondence factorial analyses, applied to near-infrared spectra, allowed the assignment of some wavelengths to fat and protein. The wavelengths at 1724, 1752, and 2308 and 2344 nm were found to be indicative of the fat content. Two wavelengths at 2050 and 2180 nm could be assigned to protein. The large absorption band of the lactose at 2094 nm seemed to discriminate the samples according to their lactose content.

Rapid measurement of the chemical composition of foods and food products by diffuse reflectance spectroscopy in the near-infrared region (1100–2500 nm) has been widely used (1–5). One of the main difficulties when studying milk is the high absorption of water in this spectral region. The characteristic absorption bands of fat, protein, and lactose are very weak in comparison with the water bands and are masked. Milk has a near-infrared spectrum very similar to that of water (6). A study of the absorption bands of the constituents cannot be simply achieved from a collection of milk spectra by direct observation. To resolve this problem, it is possible to chemically extract each constituent and to examine their spectra. This procedure does not take into account spectral variations due to interactions between constituents: modifications of the spectra of protein or lactose depending on the hydration may be observed. Moreover the constituents may be partially denatured during their extraction. Multivariate analysis is another approach that can be attempted when the samples have a high variation in their composition.

Factorial analyses allow the splitting of a collection of data into a sum of characteristic phenomena. Some authors have applied these mathematical treatments to near-infrared

0003-2700/87/0359-2187$01.50/0

spectra. The partial least squares regression was developed to obtain precise predictions of chemical constituents, from a minimum number of calibration samples (*7*). An evaluation of the bread baking quality of wheats was obtained by Devaux et al., using the multiple discriminant analysis (*8*). Cowe and McNicol used principal components analysis to study near-infrared spectra of wheat flour and milled barley (*9*). This mathematical method was tested by Robert et al. to predict forage quality (*10*).

The purpose of this work is to use multivariate statistical analysis, and in particular principal components analysis (PCA) and correspondences factorial analysis (CFA), to study near-infrared spectra of milk. These mathematical treatments applied to spectra give a method of finding wavelengths that are indicative of the different constituents.

EXPERIMENTAL SECTION

Milk Collection. Samples of milk were selected so as to present high variations of fat, protein, and lactose contents. To increase the range of the chemical variations between the samples, milk from different species was chosen. A total of 38 samples were obtained from the experimental station of INRA (Institut National de la Recherche Agronomique) at Poligny. The number and nature of the samples were as follows: 10 cow milk samples; 10 goat milk samples; 10 ewe milk samples; 4 cow colostrum and 4 mastitis cow milk samples.

Crude protein content was measured according to the Kjeldahl method (N × 6.38) (*11*). Fat content was estimated by using the Gerber method (*12*). The determination of the lactose content was evaluated according to an enzymatic method (*13*).

Near-Infrared Measurements. Near-infrared spectra were obtained with a Technicon (InfraAlyzer 500) spectrometer equipped with an integrating sphere, from 1100 to 2500 nm in 4-nm increments (351 data points for each spectrum). For each milk, the absorbances (log $1/R$) were recorded twice. The liquid cell utilized worked in "transflectance". The light successively passed through a glass window and through the sample. It was then reflected from a diffusely reflecting ceramic and passed back again through the sample and through the window. The depth of the cell was 100 μM. The samples were thermostated (40 °C) and homogenized in a Technicon device before their spectra were collected.

Mathematical Treatments. The proposed mathematical treatments were tested on an IBM-PC microcomputer. The software has been developed at the INRA (Nantes). The routines allowed centering the spectral data, performing principal components analyses, and carrying out correspondence factorial analyses.

PCA and CFA programs were developed according to Foucard (*14*).

Correction of the Spectra Data. Milk contains light scattering particles in the form of fat globules and protein micelles. These particles can introduce erroneous information. An increase of the absorption bands due to an increase of the optical path length can occur when the near-infrared beam passes through such a light scattering medium (*15*). Variation in the mean particle size or in the particle size distribution produces variations of the path length and thus in the general intensity of the spectra. The values of the absorbances increase with the mean particle size. In the case of powders, various mathematical treatments have been tested to reduce the effect of granularity (*16–18*). Preliminary studies have shown the efficiency of the mathematical treatment which consists in centering the spectral data according to the equation

$$C_{i,j} = (A_{i,j}) - \frac{1}{n}(\sum_i A_{i,j}) - \frac{1}{m}(\sum_j A_{i,j}) + \frac{1}{nm}(\sum_i \sum_j A_{i,j}) \quad (1)$$

where $C_{i,j}$ is the centered data for the sample i at the wavelength j, $A_{i,j}$ is the absorbance for the sample i at the wavelength j, n is the number of samples, and m is the number of wavelengths.

From each spectrum, the mean of its absorbances and the mean spectrum of the studied collection are subtracted. The mean of all the original data is then added to each new spectrum in order to make the sum of the absorbances in each spectrum equal to zero and to make the sum of the absorbances at each wavelength equal to zero. Centering the data in factorial analyses generally avoids giving too large a weight to wavelengths that have particularly intense absorbances.

The proposed mathematical treatment gives tables of data in which the mean of each row and the mean of each column are zero. Thus, there is correspondence between PCA achieved on such a table and PCA realized on the transposed table (*19*). This property presents some interest as it allows a comparative study of the similarity maps of the samples obtained by PCA and of the spectral patterns extracted from a collection of spectra.

Factorial Analyses. PCA was used on the one hand to carry out morphological analysis and on the other hand to draw graphical representations showing the resemblances and the differences between spectra. PCA works on rectangular tables: the rows are called "observations" and the columns "variables". Morphological analysis is obtained by PCA on rows of wavelengths and columns of sample spectra. PCA on rows of sample spectra and columns of wavelengths gives similarity maps of spectra. PCA is a multivariate statistical treatment that creates principal components made up of synthetic variables called scores and loadings. The scores are linear combinations of the original variables. The loadings quantify the influence of original variables on the principal components. The scores present the advantage of being orthogonal, i.e., independent of each other. They are determined so as to split up the total variance of the information in decreasing order. The first principal components take into account phenomena with maximal variances. Generally, the number of principal components necessary to describe the information is much smaller than the number of the original variables. Graphical representations showing the relations between the variables or between the observations are obtained by PCA.

PCA performed with the near-infrared spectra as variables and the wavelengths as observations are called morphological analyses by Le Nouvel (*20*). In this case, the principal components obtained are linear combinations of the spectra. This procedure breaks up a spectrum into a sum of spectral patterns that are independent of each other. The spectral pattern corresponding to a principal component can be drawn by associating each wavelength to its factorial coordinate on that principal component. The shapes of the spectral patterns often show resemblances to known spectral features of some constituents. In this study, extraction of spectral patterns from a collection of near-infrared spectra of milk was achieved both on the absorbances and on the centered data.

To carry out PCA with spectra as observations and wavelengths as variables, it was necessary to make a selection among the available wavelenghts, as the program cannot handle more than 80 variables. This choice was based on the results of the morphological analysis. The spectral region that showed the most variation between centered spectral patterns was selected.

The 20 milk samples that were retained as calibration samples were called principal observations and used for the creation of the principal components. The 18 prediction samples (supplementary observations) were only projected in the PCA space. The contributions of the wavelengths to the creation of principal components can be evaluated from the formula

$$C_i^\lambda = 1000(X_i^\lambda)^2/l_i \quad (2)$$

where C_i^λ is the contribution of the wavelength λ to the creation of the principal component i, X_i^λ is the coordinate of the wavelength λ on the principal component i, and l_i is the eigenvalue associated to the principal component i.

PCA with spectra as observations was performed on centered data to draw similarity maps and to analyze the resemblances between spectra.

CFA is a mathematical treatment that was developed by Benzecri (*21*). This treatment is similar in its principle to PCA. Initially designed to describe frequency tables, its field of application has been widened to tables containing positive data. In a first step, the spectral data table is transformed to give the same role to the rows and the columns. So, the synthetic variables calculated by CFA and called principal factors are similar using either the spectra or the wavelengths as observations. The reprocity of the analysis on the rows and the columns allows representations of spectra and wavelengths on a same graph. An interpretation of the proximity between a given wavelength and

Table I. Fat Protein and Lactose Contents (g/L) of Milk

constituent	range	mean	std dev
Principal Observations (N = 20)[a]			
fat	19.00–69.00	41.85	14.37
protein	25.85–67.04	42.87	11.78
lactose	38.80–54.30	49.56	4.02
Supplementary Observations (N = 18)			
fat	24.50–64.00	37.78	10.92
protein	28.68–50.63	36.39	7.11
lactose	44.90–53.30	49.97	1.89

[a] N, number of milk samples

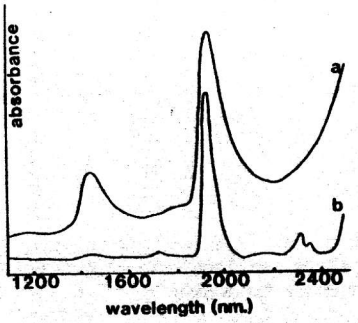

Figure 1. Mean spectrum (a) and variance spectrum (b) of the 76 recordings.

the spectrum of a particular sample is possible. Moreover, as spectra are continuous signals, the wavelengths can be linked in increasing order to give a line that corresponds to a "parametric curve". Changes of direction in the parametric curve usually point out significant wavelengths. CFA is less sensitive than PCA to variations of the general intensity of the spectra. Thus in this study CFA was carried out on the absorbances. For practical reasons, the 351 wavelengths were conidered as observations.

RESULTS AND DISCUSSION

Chemical results for crude protein, fat, and lactose contents are given in Table I. The correlation matrix calculated from the 38 milk samples showed that only crude protein content and fat content present a positive correlation (r = 0.59).

The mean and variance spectra of the 76 recordings (Figure 1) were obtained by calculating the mean and the variance of the absorbances at each wavelength. The mean spectrum was similar to that of water and had two intense absorption bands at 1450 and 1940 nm. The variance spectrum featured absorption bands at 1940, 2306, and 2345 nm. Characteristic shapes of the water spectrum (1940 nm) and of the fat spectrum (2306 and 2345 nm) were observed.

Morphological Analyses. Results of the morphological analysis performed on the absorbances of the 20 calibration samples are presented Figure 2 and Figure 3. The coefficients of correlation between milk spectra were high (r = 0.99). The spectral patterns associated to the first three principal components took into account almost the total variance. The first principal component, which accounted for 99.98% of the variance, was characteristic of water (Figure 2). The spectral pattern defined by the principal component 2 exhibited peaks at 1214, 1721, 1752, and 2304 and 2343 nm. A shoulder at 2379 nm was also observed. These wavelengths could be assigned to the near-infrared spectrum of fat. The negative peak at 1939 nm indicated that for principal component 2 an inverse relationship took place between water and fat content. The interpretation of principal component 3 was more difficult as it only accounted for 0.001% of the variance. Nevertheless two peaks at 2050 and 2179 nm might be characteristic of protein content.

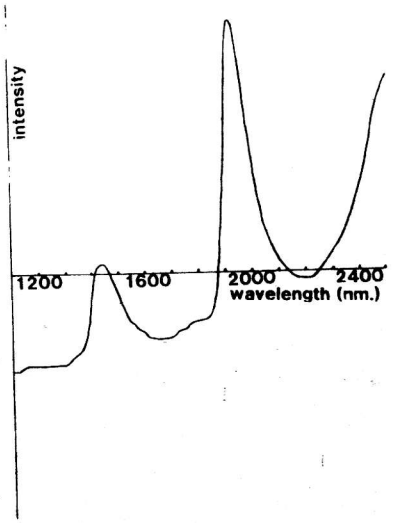

Figure 2. Morphological analysis on absorbances. Spectral pattern associated with principal component 1.

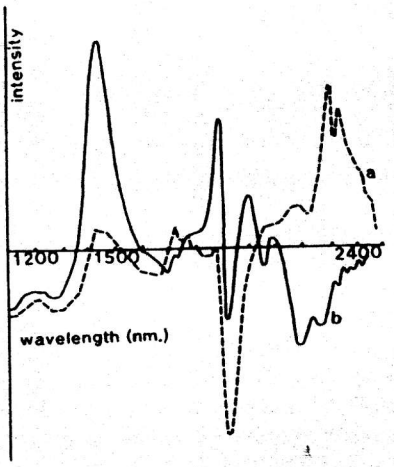

Figure 3. Morphological analysis on absorbances. Spectral patterns associated with principal components 2 (a) and 3 (b).

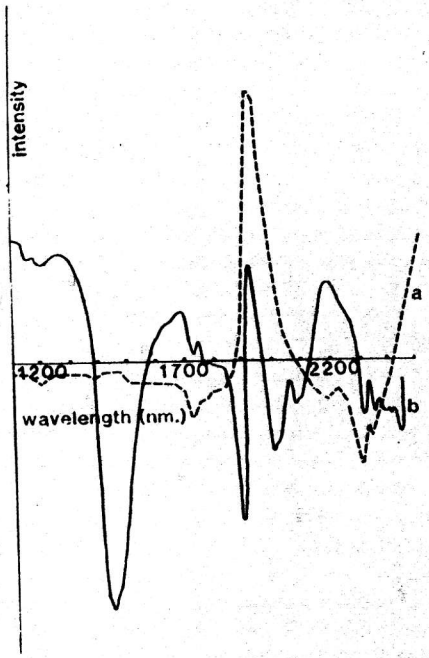

Figure 4. Morphological analysis on centered data. Spectral patterns associated with principal components 1 (a) and 2 (b).

The morphological analysis performed on the centered data is given Figure 4. The first principal component (94.45% of

Table II. Contributions of Wavelengths to the Creation of the First Four Principal Components[a,b]

wavelengths, nm	C1	R1	C2	R2	C3	R3	C4	R4
2036	3.80	0.92	3.00	0.04	51.30	0.34	88.20	0.17
2044	1.90	0.86	1.50	0.14	44.80	0.42	87.70	0.23
2064	0.10	0.42	1.74	0.30	40.70	0.80	10.80	0.16
2084	0.00	0.23	0.90	0.23	48.10	0.91	4.90	0.11
2092	0.10	0.49	0.50	0.14	44.10	0.80	24.80	0.23
2100	0.40	0.66	1.00	0.18	44.60	0.66	37.20	0.23
2180	7.20	0.92	35.90	0.38	2.10	0.05	0.00	0.00
2308	61.10	0.99	40.70	0.15	0.00	0.00	0.30	0.00

[a] Ci, contribution to the creation of principal component i. [b] Ri, coefficients of correlation between wavelengths and principal component i.

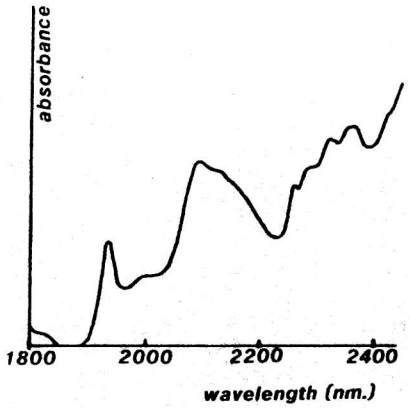

Figure 5. Near-infrared spectrum of the lactose monohydrate.

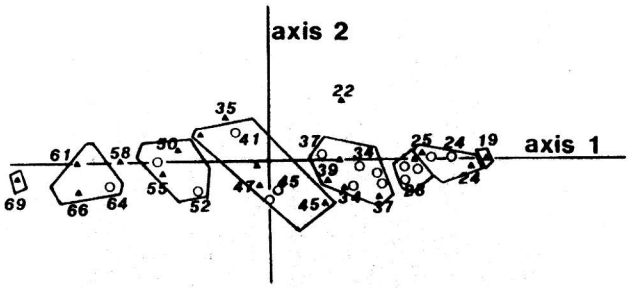

Figure 6. Similarity map defined by principal components 1 and 2: principal observations (▲), supplementary observations (O). The numbers correspond to fat content (g/L).

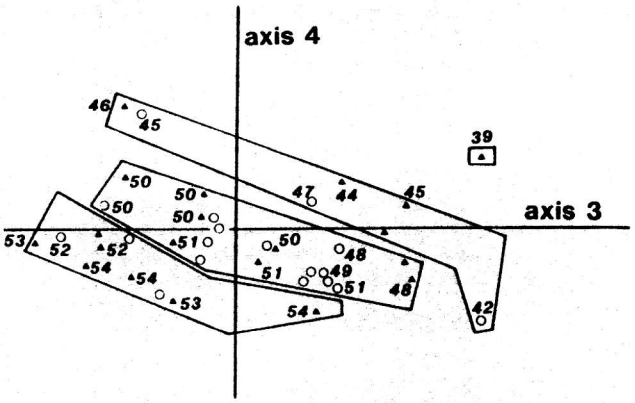

Figure 7. Similarity map defined by principal components 3 and 4: principal observations (▲); supplementary observations (O). The numbers correspond to lactose content (g/L).

the variance) opposed the wavelength of water at 1939 nm to the wavelengths of fat; the second one showed two peaks at 2050 and 2179 nm (3.28% of the variance). The spectral patterns were similar to those previously observed for principal components 2 and 3. The effect of centering the data was principally to eliminate the first factor which corresponded to spectral variations due to scattering. In this case, a reduction of the useless information seems possible by centering the data.

Principal Components Analysis. The 80 wavelengths within the limits 2030–2350 nm were selected as variables. This spectral region contains wavelengths of protein and fat, which were exhibited by spectral patterns of the morphological analyses, and no intense absorption band of water. Moreover, the spectrum of the lactose monohydrate (Figure 5) presents in this region absorption bands at 2094, 2260, 2284, and 2320 and 2350 nm. The first four principal components, which are linear combinations of wavelengths, took into account 99.41% of the cumulated variance. An interpreation of the principal components was attempted by calculating the correlation coefficients with the chemical data. Fat was correlated to principal component 1 (r = 0.93) and principal component 2 (r = 0.35). These principal components had coefficients of correlation of 0.77 and 0.57 with the protein content. The lactose was correlated to principal components 3 (r = 0.70) and 4 (r = 0.64). Principal components 1 and 2 (respectively 95.16% and 3.21% of the variance) defined the similarity map given in Figure 6. A discrimination of the spectra according to the fat content was observed for both the principal and supplementary observations. Two cow colostrum samples with fat contents of 35 and 22 g/L seemed to be uncorrectly plotted on this similarity map. They had high protein contents (67.04 and 57.74 g/L) and were correlated to the principal component 2 (r = 0.68 and 0.64) which was representative of proteins. A discrimination of the spectra according to the lactose content was observed on the graphical representation (Figure 7) defined by the principal components 3 (0.97% of the variance) and 4 (0.14% of the variance). The most important contributions of the wavelengths to the creation for the principal components are given Table II. The wavelength at 2308 nm, which could be assigned to fat content, was important in the creatin of the two first principal components. The principal component 2 was associated to a characteristic wavelength of protein at 2180 nm. Wavelengths of lactose (2036, 2044, 2064, 2084, and 2093 and 2100 nm) had high contributions to the principal components 3 and 4.

Correspondence Factorial Analysis. A graphical representation of both the wavelengths and the spectra is given Figure 8. The two first axes took into account of 97.4% of the cumulated variance. Axis 1 discriminates spectra of milk according to the fat content. Close to the milks that had high fat content, the parametric curve showed changes of direction at 1724 and 2308 and 2344 nm. The samples with low fat content were near the wavelength at 1924 nm. An opposition between a spectral information related to the fat content and a wavelength assigned to water was observed. Change of direction of the parametric curve toward spectra with high protein content was seen at 2180 nm.

The factorial analyses appear to be efficient methods to describe the spectral data. They allow the extraction of information from near-infrared spectra that are very similar.

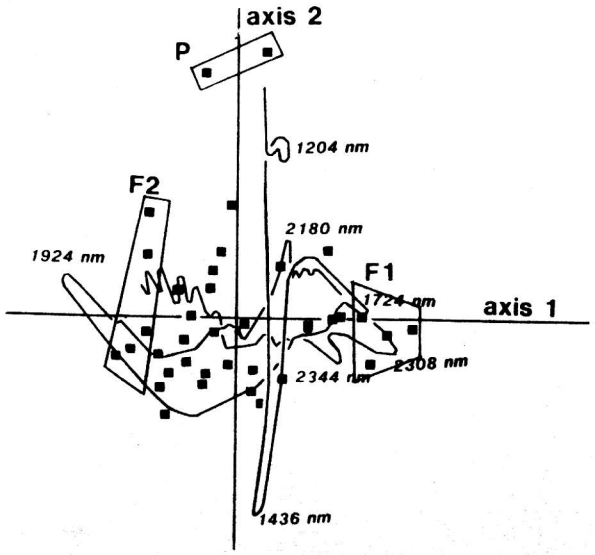

Figure 8. CFA on the near-infrared spectra of milk: parametric curve (——), spectra of milk (■); P, samples with high protein content (67.7 and 67.0 g/L); F1, samples with high fat content (range 58.0–69.0 g/L); F2, samples with low fat content (range 19.0–24.5 g/L).

Spectral patterns that are characteristic of the principal constituents are obtained by morphological analyses, and the resemblances between spectra are visualized on the similarity maps of PCA. An assignment of the wavelengths can be realized by using CFA.

Registry No. Lactose, 63-42-3.

LITERATURE CITED

(1) Williams, P. C. *Cereal Chem.* **1975**, *52*, 561.
(2) Norris, K. H.; Barnes, F. E.; Moore, J. E.; Shenk, J. S. *J. Anim. Sci.* **1976**, *43*, 889.
(3) Barnes, F. F.; Marten, G. C. *J. Anim. Sci.* **1979**, *48*, 1554.
(4) Frank, J. F.; Birth, G. *J. Dairy Sci.* **1981**, *65*, 1110.
(5) Baer, R. J.; Frank, J. F.; Loewenstein, J. *J. Assoc. Off. Anal. Chem.* **1983**, *66*, 858.
(6) Goulden, J. D. S. *J. Dairy Sci.* **1958**, *25*, 228.
(7) Martens, H.; Jensen, S. A. *Proceedings of the 7th World Cereal and Bread Congress, Prague*, Elsevier: Amsterdam, 1982; pp 607–647.
(8) Devaux, M. F.; Bertrand, D.; Martin, G. *Cereal Chem.* **1986**, *63*, 151.
(9) Cowe, I. A.; McNicol, J. W. *Appl. Spectrosc.* **1985**, *39*, 257.
(10) Robert, P.; Bertrand, D.; Demarquilly, C. *Anim. Feed Sci. Technol.* **1986**, *16*, 215.
(11) Norme AFNOR V. 04.210, Association Française de Normalisation 1975, Paris.
(12) Norme AFNOR V.04.211, Association Française de Normalisation, 1971, Paris.
(13) Document 174, 1983; Fédération Internationale de Laiterie, International Dairy Federation: Square Vergote 41, Bruxelles.
(14) Foucart, T. *Analyse Factorielle—Programmation sur microordinateurs*; Masson: Paris, 1982.
(15) Ben-Gera, I.; Norris, K. H. *Isr. J. Agric. Res.* **1968**, *18*, 117.
(16) Norris, K. H.; Williams, P. C. *Cereal Chem.* **1984**, *61*, 158.
(17) Mc Clure, W. F.; Hamid, A.; Giebnecht, F. G.; Weeks, W. W. *Appl. Spectrosc.* **1984**, *38*, 322.
(18) Robert, P.; Bertrand, D. *Sci. Aliments* **1985**, *5*, 501.
(19) Lefevre, J. *Introduction aux analyses statistiques multidimensionelles*, 3rd ed.; Masson: Paris, 1983.
(20) Le Nouvel, J. Thesis, Université Rennes I, 1981.
(21) Benzecri, J. P.; Benzecri, F. *Pratique* de l'analyse des données; Dunod: Paris, 1980.

RECEIVED for review October 27, 1986. Accepted May 11, 1987.