Available online at www.sciencedirect.com

SCIENCE DIRECT

Chemometrics and Intelligent Laboratory Systems 78 (2005) 96–102

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

X-ray scattering processes and chemometrics for differentiating complex samples using conventional EDXRF equipment☆

Maria Izabel Maretti Silveira Bueno[a,*], Martha T.P.O. Castro[b], Aline Moreira de Souza[a], Erica Borges Santana de Oliveira[b], Alete Paixão Teixeira[b]

[a]*Instituto de Química, Universidade Estadual de Campinas, CP 6154, Campinas 13084-971, SP, Brazil*
[b]*Instituto de Química, Universidade Federal da Bahia, Salvador 40170-290, BA, Brazil*

Received 14 August 2004; received in revised form 22 December 2004; accepted 4 January 2005
Available online 3 March 2005

Abstract

Mild variations in organic matrices, which are investigated in this work, are caused by alterations in X-ray Raman scattering. The multivariate approaches, principal component analysis (PCA) and hierarchical cluster analysis (HCA), are applied to visualize these effects. Conventional energy-dispersive X-ray fluorescence equipment is used, where organic compounds produce intense scattering of the X-ray source. X-ray Raman processes, before obtained only for solid samples using synchrotron radiation, are indirectly visualized here through PCA scores and HCA cluster analysis, since they alter the Compton and Rayleigh scattering. As a result, their influences can be seen in known sample characteristics, as those associated with gender and melanin in dog hairs, and the differentiation in coconut varieties. Chemometrics has shown that, despite their complexity, natural samples can be easily classified.

Keywords: Principal component analysis (PCA); Hierarchical cluster analysis (HCA); Natural sample differentiation; X-ray Raman scatter spectrometry (XRSS); Complex organic mixtures

1. Introduction

X-ray fluorescence spectrometry (XRF) is well known [1–3] for its ability to identify and quantify inorganic species in a fast, simple, low-cost, and non-destructive way, specially using its energy-dispersive variant, energy-dispersive X-ray fluorescence (EDXRF). Now, it is also possible to differentiate complex organic matrices using conventional EDXRF equipment, specifically observing the X-ray source scatter region. The method is being developed in conjunction with chemometric tools, such as principal component analysis (PCA) and hierarchical cluster analysis (HCA), which promptly offer this possibility, as demonstrated by the results reported herein.

XRF is based on the photoelectron phenomenon [1–4], which consists in exciting atoms with an energy source (X-ray emitting radioisotopes, X-ray tubes, accelerated particles, or synchrotron radiation) able to remove electrons from inner orbitals close to the nucleus. The technique has been applied in many areas; a few relevant applications include non-destructive quantification of metals in biological tissues [5] and of inorganic impurities in microchips [6] and in Mars soil, through remote control by proton-alpha induced X-ray fluorescence [7].

XRF determination of organic content in samples is considered a very difficult task since the fluorescence cross sections are very low for low-Z elements [1–4]. The few alternatives to directly measure characteristic lines of very low-Z elements involve sophisticated instrumentation [8], such as synchrotron radiation with specialized detector windows (or even no window under vacuum) and specialized focusing optics [8–10]. These methods have low precision and low accuracy and, even with

☆ Patent pending.

* Corresponding author. Tel.: +55 19 37883128; fax: +55 19 37883023.
E-mail address: bell@iqm.unicamp.br (M.I. Maretti Silveira Bueno).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.01.001

sophisticated apparatus, long measurement periods are required to reach satisfactory signal to background ratios [9–11].

Besides the photoelectron phenomenon, which is a specific elemental interaction, there are also concomitant interactions between X-rays and the sample. Among them, it is worthwhile to mention the Compton and Rayleigh effects that are caused by radiation scattering instead of absorption/emission [1], with higher intensities for low-*Z* elements. The Compton effect is related to inelastic (incoherent) scattering with some energy loss caused by *momentum* transfers between photons and electrons, and the Rayleigh effect to elastic (coherent) scattering, without any energy variation. More frequently, when working with XRF, people try to avoid this region since it contributes strongly to spectrum background [9].

Another weak process occurs, Raman X-ray scattering [10,11]. Specific lines of X-ray Raman processes are only observable under very special conditions. In this work, the intense Rayleigh and Compton scatterings produced by a rhodium X-ray tube of conventional EDXRF equipment, in contact with organic matrices, serve as the source for the mild and multi-occurrence Raman processes.

On the other hand, chemometrics has been used successfully with other spectroscopic methods [12]. A good example is the now well-disseminated use of near-infrared spectroscopy (NIR), conjugated with multivariate calibrations for analysis and visualization of several types of samples. It is well known that NIR does not produce specific transitions, but unresolved bands of concomitant events, caused by overtones and conjugation bands of CH, OH, and NH vibrational transitions. Thus, it is generally not possible to apply this technique without also using chemometric tools [13–16].

This paper shows that the combination of principal component analysis (PCA) or hierarchical cluster analysis (HCA) with common EDXRF procedures can reveal important hidden information. Properties related directly to organic species, which mildly alter the scatter profile in their respective spectra, can be revealed. This new method is called as X-ray Raman scatter spectrometry (XRSS). Two interesting applications are described. The first XRSS application used hairs from poodle dogs and, the second, used coconut water extracted from three varieties of this fruit.

2. Experimental

Thirty-four hair samples of poodle dogs (of known age, hair color, gender, health status, and living environment) were taken from veterinary clinics. Table 1 shows a compilation of these dog characteristics.

The samples were carefully washed, dried, and cut into small lengths, following recommended procedures [17]. The measuring procedure was very simple, consisting of weighting 200 mg of each sample directly into appropriate cells and submitting them to blank rhodium X-ray tube radiation, in triplicate. A common laboratory energy-dispersive XRF instrument (EDX 700, Shimadzu) was used under the same irradiation conditions: 50 kV of applied voltage in the tube, 25% of dead time, 10 mm of beam collimation, and 100 s of irradiation time.

Table 1
Physical and habit characteristics of dogs under analysis

Sample	Healthy?	Raised indoor?	Gender	Age (years)	Hair color
1	Yes	Yes	Female	7	White
2	Yes	Yes	Female	11	White
3	Yes	Yes	Female	8	White
4	Yes	Yes	Male	2	White
5	Yes	Yes	Female	1.25	White
6	No	Yes	Male	7	Black
7	Yes	Yes	Male	2	Black
8	Breast cancer	Yes	Female	7.75	Black
9	Yes	Yes	Male	4.5	Light brown
10	Yes	Yes	Male	2	Light brown
11	Yes	Yes	Female	2.5	Light brown
12	Yes	No	Male	3	White
13	No	Yes	Male	12	Black
14	Yes	Yes	Female	4	Light brown
15	Yes	Yes	Male	5	Light brown
16	No	No	Male	1	Light brown
17	Yes	Yes	Male	5	White
18	Yes	Yes	Male	3	Light brown
19	Yes	Yes	Female	3.75	Light brown
20	Yes	No	Male	4	White
21	Yes	Yes	Male	7	Light brown
22	Yes	Yes	Male	2	Light brown
23	Yes	Yes	Male	7	Light brown
24	Breast cancer	Yes	Female	5	Light brown
25	*No*	Yes	Male	5.7	Black
26	Yes	Yes	Male	1.5	Light brown
27	Yes	Yes	Male	0.4	White
28	Yes	Yes	Female	0.7	Light brown
29	Yes	Yes	Female	3	Light brown
30	Yes	Yes	Male	–	Light brown
31	Breast cancer	Yes	Female	9	Black
32	Yes	Yes	Male	13	Black
33	Yes	Yes	Female	2	Light brown
34	No	Yes	Female	0.7	Light brown

Eighteen natural samples of three coconut varieties were obtained in local markets, and their waters were extracted and immediately frozen. Before irradiation under room temperature (around 25 °C), the samples were filtered to remove solid particles. 2 mL of each sample was irradiated in quadruplicate, using exactly the same conditions for the dog hair analyses.

The chemometric methods [12] used for data analysis are PCA and HCA, from "PLS Toolbox," 2.0 version (Eigenvector Technologies, Manson, USA). MATLAB software, 6.1 version (The Mathworks, Natick, USA) was used to run them.

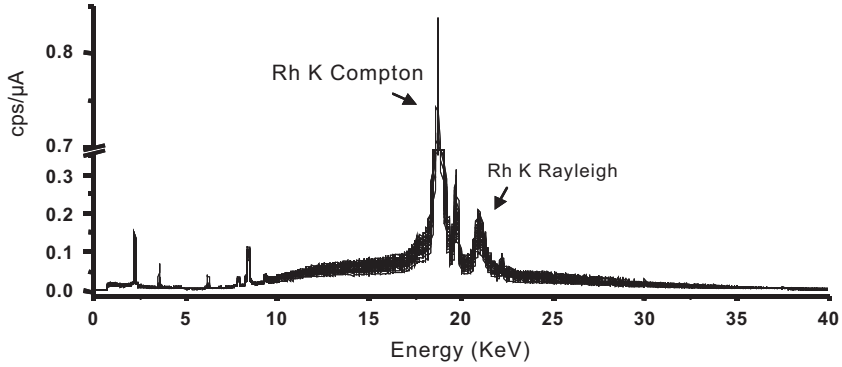

Fig. 1. Superimposed spectra of 34 hair samples of poodle dogs. Irradiation conditions are mentioned in the text.

The spectra were mathematically processed in the following order: (i) one data matrix was constructed in such a way that each row corresponded to the spectrum of a sample and each column to their respective energy values; (ii) spectra data were average-centered; and (iii) the unsupervised visualization methods, PCA and HCA, were applied [12].

3. Chemometric methods

One of the most important multivariate methods of data analysis is principal component analysis (PCA) [12,19], based on the correlation between variables. It aims at reducing a large set of correlated variables to a small number of uncorrelated principal components, PCs, onto which the data are projected. These PCs are built as linear combinations of original variables in such a way that the first component accounts for as much of the total variance in the original variables as possible. The first new axis, PC1, is chosen in such a direction that the variance is maximized along that axis; the second axis must be chosen orthogonal to the first one and in the direction to describe as much remaining variance as possible, and so on.

The raw data matrix, represented by **X**, is decomposed into two matrices, **T** and **V**, where $\mathbf{X}=\mathbf{T}\mathbf{V}^{\mathbf{T}}$.

The matrix **T**, known as 'score' matrix, represents the position of the samples in the new coordinate system. The second matrix, **V**, is the loading matrix and describes how the new axis (i.e., the PC) is built from the original variables.

Hierarchical cluster analysis (HCA) is another important multivariate method of analysis. In HCA, similarity or dissimilarity measures distances or correlation coefficients among pairs of samples. These coefficients are calculated in order to cluster the data and, depending on these distances, the samples are considered similar or not. Dissimilar samples will be positioned at larger or shorter distances relative to each other. HCA is able to group data into clusters having similar attributes. The results are presented in the form of dendograms to facilitate the visualization of sample relationships. Several cluster analysis methods are available, and the main classifications are the *agglomerative* and *divisive* methods. In the first case, called also "bottom–up," the hierarchical process starts at the bottom and merges a selected pair of subclusters into a single cluster. Then a pair to merge is selected as having the smallest intergroup dissimilarity. This cluster type can be subclassified in: (a) *single linkage*, when the intergroup dissimilarity is the closest (least dissimilar) pair; (b) *complete linkage*, when the intergroup dissimilarity is the furthest (most dissimilar) pair; and (c) *average linkage*, when the average dissimilarity between groups is considered. In the second main classification, the *divisive*, called also "top–down," the hierarchical process starts at the top and splits one cluster into two new clusters. The split is chosen to produce two new groups with largest between-group dissimilarity.

More elaborate descriptions of PCA and HCA can be found in textbooks and reviews, leading to multivariate analysis [12,17].

4. Results and discussion

4.1. Dog hair analysis

The first application was initiated to try to solve the controversy that accompanies the mineral analysis of hair in

Table 2
Percent variance captured by PCA model applied to dog hair spectra

Principal component number	Percentage of variance captured
1	98.39
2	0.90
3	0.60
4	0.01
5	0.01
6	0.01
7	0.01
8	0.01
9	0.01
10	0.01
11	0.01
12	0.01

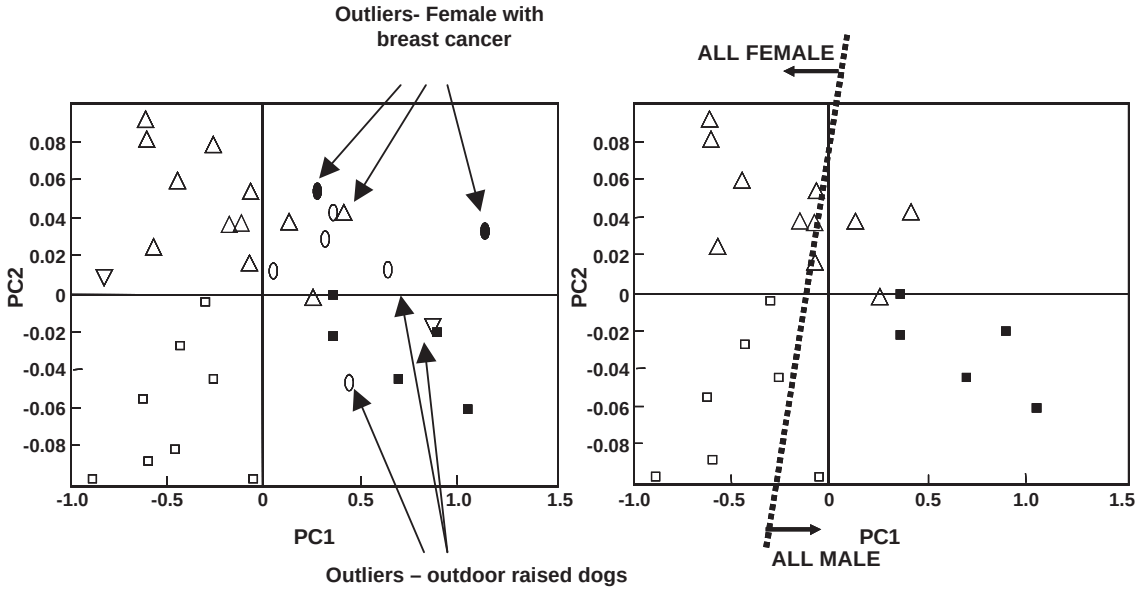

Fig. 2. Scores plots of dog hair samples after spectra processing by PCA. Left: PC2 versus PC1 for all dogs [(△) light brown hair; (■) black hair; (□) white hair]. Right: PC2 versus PC1 without considering the outliers [(○) white hair from sick dogs; (●) black hair from sick dogs; (▽) light brown hair from sick dogs].

orthomolecular medicine [18], often carried out by atomic absorption or emission spectrometries [7]. An easy-to-obtain sample was chosen, namely the hairs of poodle dogs, taken from veterinary clinics. The set of spectra was superimposed after X-ray irradiation (Fig. 1).

PCA and HCA results after the treatments using the software MATLAB were, to say the least, surprising. Since the samples were all of organic origin, the chosen spectral region was that of X-ray source scattering (from 18.8 to 22.0 keV). PC variance values and the plot of PC1 and PC2 score values are shown in Table 2 and Fig. 2, respectively. Fig. 2 reveals that PC1 visualizes hairs in accordance with gender and, PC2, by the color of their hairs (white, black, and light brown). The outliers revealed on the projection are sick (three of them were suffering from breast cancer), outside raised dogs (see Table 1 and Fig. 2). Fig. 3 shows the loading plot of PC1×PC2. Most information is really in the scatter region, which is an indication that organic variations among the samples do separate them. The maximum PC1 loading is located at 19.2 keV (Compton scatter Rh K X-ray peak) and is responsible for separating the genders. PC loadings show that shifts in energy are causing this effect, depending on the chemical environment around each atom. PC2 is able to visualize the dogs with respect to the presence of distinct melanins that reflect Raman shifts observed for samples.

As mentioned, the separation shown by PC1 was possible due to modifications caused in Rh K X-ray

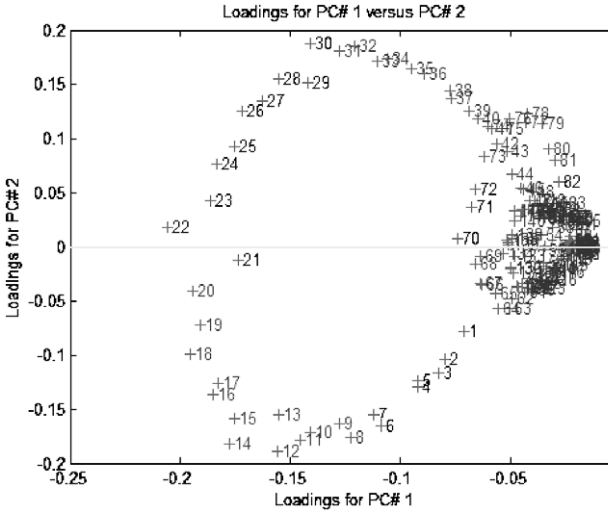

Fig. 3. Loading plots of dog hair samples after having their spectra processed by PCA. Each number corresponds to an energy variation of 0.02 keV, with the beginning at 18.8 keV. So the maximum in PC1, variable 22, is equal to 19.20 keV. In PC2, the first maximum, 30, is equal to 19.40 keV and the second maximum, 12, is equal to 19.04 keV.

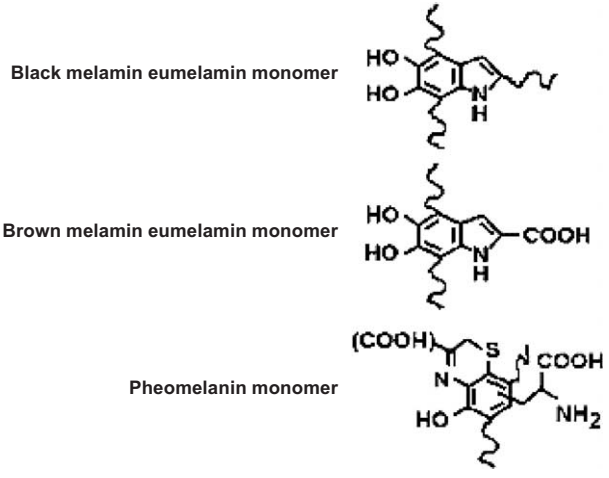

Fig. 4. Structures of different melanins, responsible for alterations in hair color [21].

Table 3
Percent variance captured by PCA model applied to coconut water spectra

Principal component number	Percentage of variance captured
1	69.70
2	4.74
3	3.23
4	2.72
5	2.54
6	2.14
7	2.08
8	1.97
9	1.70
10	1.67
11	1.53
12	1.26
13	1.18
14	1.10
15	0.92
16	0.80
17	0.73

scattering, depending on the sample characteristics associated with the gender. Fig. 2 shows this separation, but three samples were removed since they produced interference in the separation, certainly due to kinship among them. It is well known that this factor is of higher relevance than gender factors.

PC2 depends on the presence, concentration, or type of melanin in each sample. The results from white hair samples lie well away from the tendency shown by the colored hairs. The absence of melanin in these samples can account for this effect, as melanin significantly alters the hair structure and, thus, radiation scattering. The structures of melanins attributed to different hair colors are shown in Fig. 4 and suggest strong differentiation in the overall X-ray scattering [20,21].

The variance values (Table 2) and the plot of loadings (Fig. 3) show that the most important information is now observed around 19.2 keV (Compton Rh Kα line), not exactly in it. In spite of producing low variance levels (0.9%) (Table 2), PC2 can be considered of great importance, since it permits the separation shown in Fig. 2, besides producing appreciable loading values (Fig. 3).

When the animal is exposed to polluted environments (e.g., being raised outdoors or when there are metabolism alterations (sicknesses)), its color visualization is altered, producing outliers in the score plot (Fig. 2), as well as when there are kinships among the dogs.

4.2. Coconut differentiation

The other application was related to visualization of coconuts. The *Cocus* genus is constituted only by the *Cocus nucifera* L. species, which includes some varieties, such as the *Typica* (Gigante is the popular name) and the *Nana* (Anã is the popular name). The hybrids (as the Índia coconut) mostly result from mixing these varieties [22].

The *Nana* variety, with higher water content, is the most adequate as to have its "coconut water" commercially explored. Índia and Gigante have thick albumen pulps, and Gigante water tastes salty. These latter ones are then more used as raw materials in food and soap factories.

PCA and HCA were used again as an unsupervised visualization method in the same spectrum region (i.e., from 18.8 to 22.0 keV), and the data were also average-centered, as was done in the dog hair analyses.

The two first PCs accounted for 74.4% (Table 3) and it was possible to visualize distinct groups in accordance with each coconut variety (Gigante, Anã, and Índia), as is shown in Fig. 5. Fig. 6 shows the corresponding loading plots, after having the coconut water spectra processed by PCA.

It is clear from Fig. 5 that a differentiation is produced, depending on the coconut variety. Besides that, the grouping of the hybrid species is halfway between the groupings of their parents. Samples 16, 17, and 18 are of Índia hybrid variety and presented visual alterations in the fruit peels, presumably caused by microorganism deterio-

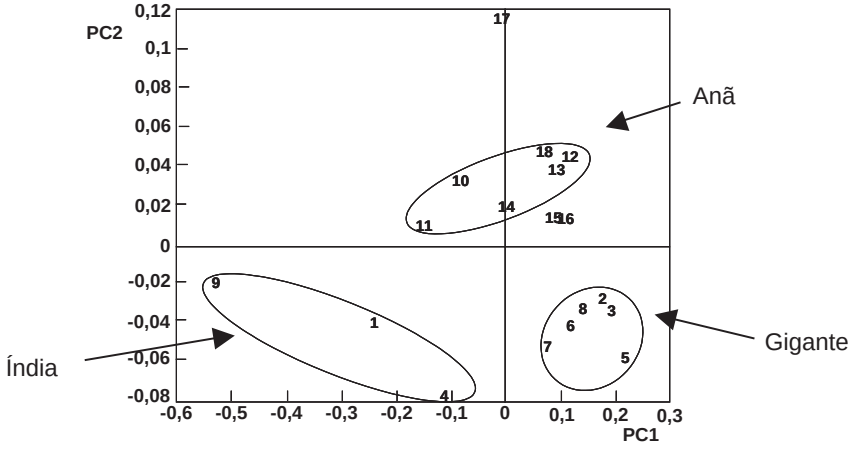

Fig. 5. Score plot of coconut samples after spectra processing by PCA.

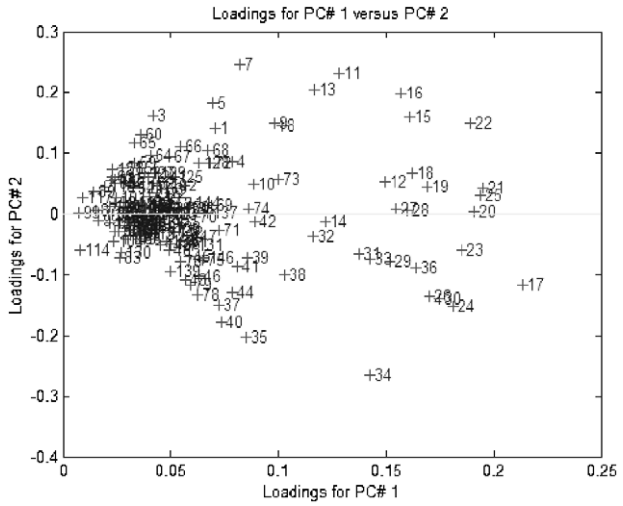

Fig. 6. Loading plot of coconut waters after spectra processing by PCA. Each number corresponds to an energy variation of 0.02 keV, with the beginning at 18.8 keV. So the maximum in PC1, variable 17, is equal to 19.14 keV. In PC2, the first maximum, 34, is equal to 19.48 keV and the second maximum, 11, is equal to 19.02 keV.

ration. This fact can explain their outlier positions in the scores graphic, in the same manner that outliers in dog hair analysis were related to those animals presenting biological alterations.

Analyzing the loading plot (Fig. 6), it can be observed that the same scatter region affects the multivariate separations of coconut varieties as was observed for dog hairs. The difference here is that variance values (Table 3) are more spread out among the first PCs. This multitude of variance values can be translated as due to a great variety of constituents: water, sugars, oils, lipids, and proteins, occurring in distinct concentrations in the three varieties. From the literature, water, sugar, and oil concentrations can be the major contributors for the variance, since their contents are spread out, depending on the coconut variety [22].

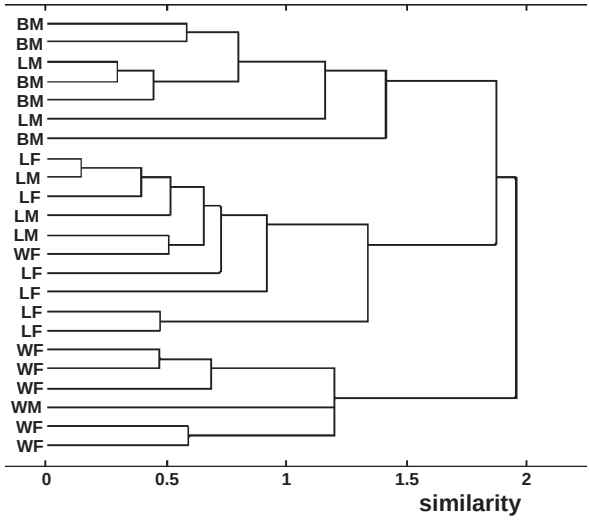

Fig. 7. Hierarchical cluster analysis for dog hair spectra (B=black hair; L=light brown hair; W=white hair; M=male; F=female).

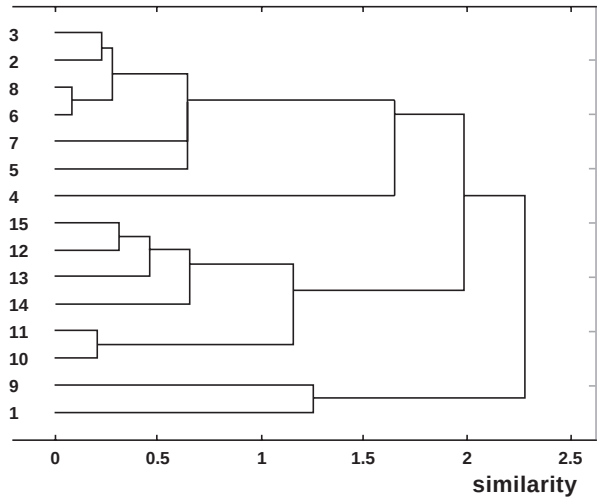

Fig. 8. Hierarchical cluster analysis for coconut water spectra.

4.3. HCA dendograms

To check the results obtained by PCA, the same spectra were submitted to HCA analysis. The agglomerative complete linkage method was applied in both cases. Figs. 7 and 8 show HCA plots for dog hairs and coconut waters, respectively. In Fig. 7, a clear cluster separation of samples is noted in accordance with gender and hair color, as was observed in PCA (Figs. 2 and 3). In Fig. 8, the same discrimination, now in the form of clusters for the coconut varieties can be seen, as observed with PCA in Figs. 5 and 6.

5. Conclusions

This paper shows that the multivariate approach is an efficient tool for the characterization of organic samples using X-ray. This is the first time that PCA has been applied to this technique. This methodology can be used in environmental biomonitoring, as well as in the indication of an animal's health status, helping in diagnoses of sickness. The spectral region that provides the most information in PCA analysis can be attributed to Compton rhodium K scattering. This spectral region is where organic matrices have an influence, either by concentration or by having distinct structures exposed to X-rays. Raman weak X-ray alterations are responsible for this, being able to differentiate the samples. From dog characteristics, the alterations in visualization can also be caused by administrations of different drugs, as has been noted elsewhere [21]. Since both human hair and dog hair have similar chemical structures [20], it is expected that this methodology can also be applied to human hair, where it can contribute to hair analysis based on orthomolecular medicine, which nowadays presents significant controversy [18]. XRSS techniques allied to PCA are able to discriminate many kinds of organic compounds, since they involve internal atomic processes. These organic variations include

sample characteristics associated with gender, hair color, or fruit varieties. XRSS can be used in orthomolecular medicine based on analysis of the organic content of hair, producing easily obtained information. Numerous other applications can be predicted.

In both cases studied here, a few samples were considered outliers, but this can be explored as a method for detecting metabolism problems or even for parentage tests.

XRSS does not need an intense source, like synchrotron radiation, or special high-resolution detectors; chemometrics aids in visualizing samples differentiated by this process. Many other organic–biological (and even low-Z inorganic) applications can be developed if the analyst considers X-ray tube scatterings as valuable processes in EDXRF spectra and not as a drawback.

Acknowledgements

The authors thank FAPESP and FAPESB (Brazilian Agencies) for financial support, and Carol H. Collins for language evaluation.

References

[1] R. Jenkins, X-ray Fluorescence Spectrometry, 2nd ed., Wiley-Interscience, New York, 1999.
[2] J. Despujols, J. Phys., IV 6 (1996) 611–618.
[3] N. Broll, J. Phys., IV 6 (1996) 583–597.
[4] R. Cesareo, A.L. Hanson, G.E. Gigante, L.J. Pedraza, S.Q.G. Mahtaboally, Phys. Rep. 213 (1992) 117–178.
[5] J. Borjesson, M. Isaksson, S. Mattsson, Acta Diabetol. 40 (2003) S39–S40.
[6] N.L. Misra, K.D.S. Mudher, Prog. Cryst. Growth Charact. Mater. 45 (2002) 65–74.
[7] T. Economou, Radiat. Phys. Chem. 61 (2001) 191–197.
[8] P.J. Potts, A.T. Ellis, P. Kregsamer, J. Marshall, C. Streli, M. West, P. Wobrauschek, J. Anal. At. Spectrom. 18 (2003) 1297–1316.
[9] M. Alvarez, V. Mazogray, X-ray Spectrom. 20 (1991) 67–71.
[10] T. Kazuyuki, Y. Udagawa, Phys. Rev., B 36 (1987) 9410–9412.
[11] T. Kazuyuki, Y. Udagawa, Phys. Rev., B 39 (1989) 7590–7594.
[12] O. Mathias, Chemometrics: Statistics and Computation in Analytical Chemistry, Wiley-VCH, Weinheim, 1999.
[13] H. Sato, M. Shimoyama, T. Kamiya, T. Amari, S. Sasic, T. Ninomiya, H.W. Siesler, Y. Ozaki, J. Near Infrared Spectrosc. 11 (2003) 309–321.
[14] B. Yuan, K. Murayama, Y.Q. Wu, R. Tsenkova, X.M. Dou, S. Era, Y. Ozaki, Appl. Spectrosc. 57 (2003) 1223–1229.
[15] R.K. Lauridsen, H. Everland, L.F. Nielsen, S.B. Engelsen, L. Norgaard, Skin Res. Technol. 9 (2003) 137–146.
[16] L. Pillonel, W. Luginbuhl, D. Picque, E. Schaller, R. Tabacchi, J.O. Bosset, Eur. Food Res. Technol. 216 (2003) 174–178.
[17] P. Geladi, Spectrochim. Acta, Part B 58 (2003) 767–782.
[18] N. Miekeley, M.T.W.D. Carneiro, C.L.P. da Silveira, Sci. Total Environ. 218 (1998) 9–17.
[19] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1–17.
[20] The Agency for Toxic Substances and Disease Registry Division of Health Assessment and Consultation and Division of Health Education and Promotion, Atlanta, Georgia, Eastern Research Group, Lexington, "Hair Analysis Panel Discussion: Exploring the State of the Science," found at http://www.atsdr.cdc.gov/HAC/hair_analysis/index.html, 2004.
[21] S. Jacques, Oregon Medical Laser Center, "Extinction coefficient of melanin," found at http://www.omlc.ogi.edu/spectra/melanin/extcoeff.html, 2004.
[22] Embrapa, "Coco pós-colheita," Embrapa Informação Tecnológica, Brasília, 2002.