ELSEVIER

Analytica Chimica Acta 384 (1999) 261–269

ANALYTICA CHIMICA ACTA

Artificial neural networks (ANNs) in the analysis of polycyclic aromatic hydrocarbons in water samples by synchronous fluorescence

R. Ferrer*, J. Guiteras, J.L. Beltrán

Departament de Química Analítica, Universitat de Barcelona, Diagonal 647, Barcelona E-08028, Spain

Received 5 June 1998; received in revised form 14 August 1998; accepted 19 October 1998

Abstract

Backpropagation artificial neural networks, principal component regression and partial least squares have been compared in order to establish the best multivariate calibration models for the analysis of mixtures of polycyclic aromatic hydrocarbons containing 10 of these compounds (anthracene, benz[a]anthracene, benzo[a]pyrene, chrysene, fluoranthene, fluorene, naphthalene, perylene, phenanthrene and pyrene). The synchronous fluorescence spectra (recorded at wavelength increments of 50 and 100 nm) of 85 standards, with concentrations ranging from 0 to 20 ng ml^{-1}, have been used for this purpose.

Keywords: PAHs; Artificial neural networks; Partial least squares; Principal component regression; Synchronous spectrofluorimetry; Multivariate calibration

1. Introduction

The determination of several compounds in a mixture can be a difficult problem, especially if their analytical characteristics are not very different. In order to avoid any previous chromatographic step, new procedures which take full advantage of the possibilities offered by chemometrics and by the wide availability of powerful and relatively inexpensive computers may become an alternative.

Among the available multivariate calibration procedures, partial least squares regression (PLSR) and principal component regression (PCR) are the most suited for multicomponent spectral analysis [1–6]. PLSR is based on the regression of chemical concentrations on latent variables (factors), and therefore, it differs from other calibration procedures (such as PCR) in that it uses the concentration data from the training set and the spectral data for modelling, whereas PCR only uses the spectral data [7]. This enables PLS to reduce the influence of dominant, but irrelevant, factors and to yield models of lower dimensionality. The main advantage of multicomponent analysis lies in the speed with which results can be obtained, because the previous separation step, usually a lengthy process, can be avoided.

A newer data processing system, that has been used for the resolution of spectral data [8–11], are the artificial neural networks (ANNs), which try to simu-

*Corresponding author. Tel.: +34-93-4021286; fax: +34-93-4021233; e-mail: raul@zeus.qui.bu.es

0003-2670/99/$ – see front matter
PII: S0003-2670(98)00778-8

late the nervous human system. A characteristic of ANN systems used for data processing is the requirement of intense computation during the training process (fitting process). This however, results in the advantage of the relatively fast prediction process of unknown samples offered by trained ANNs.

Among the different kinds of neural networks available, one of the most widely used for spectroscopic data is the backpropagation network, which learns by being exposed to sample inputs and outputs from a training database.

Of all spectroscopic techniques, spectrofluorimetry is well suited to be used in combination with multicomponent analysis, because there are relatively few compounds having intrinsic fluorescence. This fact, combined with the possibility of changing either the excitation wavelength or the emission wavelength, or both simultaneously, greatly increases the selectivity of spectrofluorimetric methods. Additionally, detection limits can often be improved by simply increasing the intensity of the radiation used for excitation.

Among the compounds having intrinsic fluorescence, polycyclic aromatic hydrocarbons (PAHs) are particularly important. These substances, whose mutagenic and/or carcinogenic effects are well-known, can be originated by natural and anthropogenic processes, and they can be found in many different kinds of samples, such as biological (e.g. meat, fish) [12–14] or environmental (e.g. soils, sediments, airborne particulate, water) [15–20]. For this reason, their detection and monitoring has become an important problem and this has led to the development of new and faster analytical methods, offering improved selectivity and sensitivity [12,14,21].

In this paper, backpropagation ANNs have been compared with PLS and PCR procedures for the prediction of 10 PAHs (anthracene, benz[a]anthracene, benzo[a]pyrene, chrysene, fluoranthene, fluorene, naphthalene, perylene, phenanthrene and pyrene) using the synchronous fluorescence spectra, recorded at two different wavelength increments. The synchronous spectra from 70 mixtures of the compounds to be determined were used as calibration set in order to build the model, and a further set of 15 mixtures was used as external validation set to test the predictive ability of the method. An aqueous micellar medium was used as solvent, in order to increase the fluorescence of the compounds.

2. Quantification algorithms

2.1. Artificial neural networks (ANNs)

A neural network performs a non-linear iterative process to fit the data. The principal computational elements are the nodes, which are arranged in layers to set the architecture of the network. In this paper, feedforward neural networks trained with backpropagation algorithms have been used. The architecture of these neural networks comprises three kinds of node layers: an input layer (which in this case contains the spectral data), one or more hidden layers, and an output layer (which contains the concentrations related to the spectral data).

2.2. Partial least squares (PLS) and principal component regression (PCR)

Both methodologies have been widely described in the bibliography [1–7]. Both combine principal component analysis (PCA) with inverse least-squares (ILS) regression. In the case of partial least squares regression, the mode in which all components are determined simultaneously (PLS-2) has been used.

3. Experimental

3.1. Reagents

Stock standard solutions (about 200 $\mu g\ ml^{-1}$) of PAHs were prepared by dissolving the pure solid (Supelco) in either methanol (benzo[a]pyrene, fluoranthene, fluorene, naphthalene, phenanthrene and pyrene) or acetonitrile (anthracene, benz[a]anthracene, chrysene and perylene), depending on its solubility.

Acetonitrile and methanol were of analytical reagent quality (Merck). Doubly distilled water (Milli-Q+, Millipore) and Brij-35 (Polyoxyethylenlauryl-ether, Merck) were used in the surfactant solutions.

3.2. Apparatus

Synchronous fluorescence spectra were recorded on an Amino Bowman Series 2 spectrofluorimeter, equipped with a quartz cell (1 cm pathlength) and slit widths of 16 nm in both monochromators.

3.3. Procedures

Standards were prepared by addition of known amounts of the stock solutions to 25 ml of 3.6×10^{-3} mol l^{-1} solution of Brij-35 (40 times its critical micellar concentration). Eighty five standards obtained in this way were divided into two sets, the first one composed by 70 standards was the calibration set, while the other 15 standards were used to perform the validation set. The range of concentrations was set between 0 and 20 ng ml^{-1}, so that each standard was linearly independent of the other standards (that is, two standards do not have the same ratio in the concentration of the 10 PAHs). The synchronous spectra of each solution were recorded at wavelength increments ($\Delta\lambda$) of 50 and 100 nm. The excitation range was from 200 to 550 nm, so the emission range varied with $\Delta\lambda$ (e.g. for $\Delta\lambda$=50 nm, the emission scan range was from 250 to 600 nm). In all cases emission readings were taken each nanometer (351 data points for each spectrum).

Before being used in the calculations, the two synchronous spectra recorded for each standard or sample were concatenated together to obtain a global spectrum (Fig. 1).

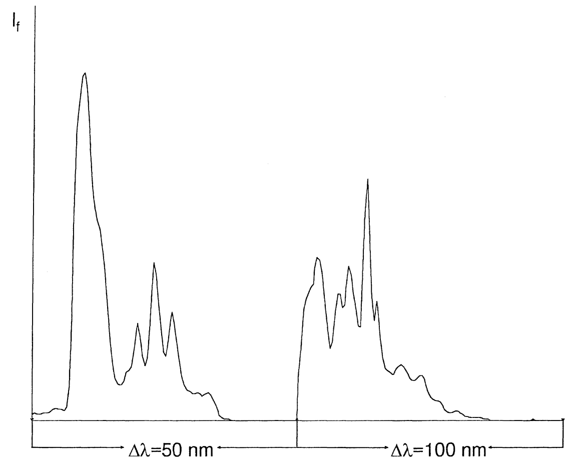

Fig. 1. Global spectrum obtained by the concatenation of synchronous spectra recorded at wavelengths increments of 50 and 100 nm.

3.4. Data processing

Data were processed on a 133 MHz Pentium PC-compatible computer. The artificial neural networks were designed and trained with the QNet97 software (32-bit Neural Network Modelling, Vesta Services). For principal component analysis and partial least squares (PLS) regression, the algorithms from the PLS_ToolBox [22], written in MATLAB language (MathWorks) were used. Models for principal component regression (PCR) were obtained with a program written by the authors in MATLAB language. In the three cases (ANN, PCR and PLS) the input and output data were scaled.

Two kinds of data were used for the training of the backpropagation neural networks. First, the full synchronous spectra (which consisted in the two synchronous spectra linked) were used. In this case, and in order to accelerate the calculations, the number of data points in each spectra was reduced by taking one of each three points, thus reducing the number of points from 702 to 234. The second type of data was the scores obtained by the principal component analysis (PCA) of the spectral data. In this case, the number of factors chosen to explain the spectral data was 14.

The learning data matrix thus obtained had a dimension of 70×234 data points in the case of the synchronous spectra and 70×14 data points in the case of the scores.

4. Results and discussion

There are two important parameters to be optimised in the artificial neural networks: the architecture of the network (number of hidden layers) and the number of nodes of each layer. Of all the different architectures tested, the best results were obtained, both in cases of full spectra and of score values, with single hidden layer architectures, as the addition of more layers did not improve the results and increased the time of calculation.

As shown in Fig. 2, the time of calculation is linearly dependent of the number of neurons contained in the hidden layer. If the values of the scores are used instead the synchronous spectra, the time of calculation is reduced up to six times.

A secondary parameter is the learning rate, which controls the size of the node weight adjustments during training (controls the rate at which the ANN attempts to learn). Fig. 3 shows the learning rate effect

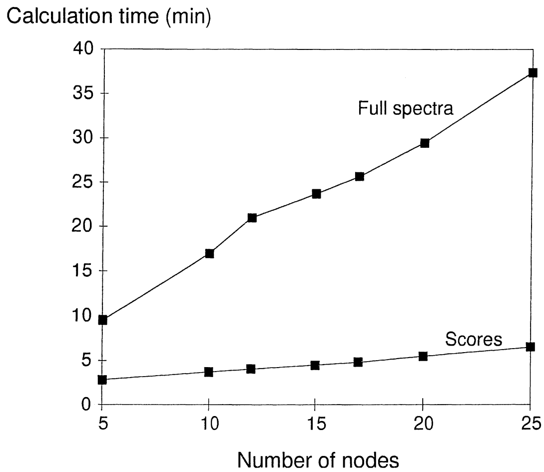

Fig. 2. Plot of time of calculation versus the number of nodes in the hidden layer.

over the training RMS error. The first plot (a) shows the RMS error with a learning rate of 0.01, with a learning rate of 0.05 the divergences begins to appear (b), and this is shown clearly in the case of learning rates of 0.07 (c), 0.10 (d) and 0.20 (e). The last plot (f) shows the effect of the automatic learning rate adjust, which is allowed by the Qnet97 program.

Another parameter studied was the function used for the connection between nodes (for the non-linear transformation of the input data inside nodes). Sigmoid, hyperbolic tangent and hyperbolic secant non-linear functions have been tested, but better results have been obtained with the sigmoid function:

$$f(x) = \frac{1}{1+\mathrm{e}^{-x}}.$$

To obtain the best results, different architectures of ANNs were tested, where the number of nodes in the hidden layer and the number of compounds in the output layer were studied. In order to compare the different models, the relative root mean squared difference (RRMSD) was calculated:

$$\mathrm{RRMSD}(\%) = \frac{100}{\overline{C}} \sqrt{\frac{1}{J} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\hat{c}_{i,j} - c_{i,j}\right)^2},$$

where $\hat{c}_{i,j}$ and $c_{i,j}$ are the predicted and the real concentration for compound i in sample j (I and J are the number of compounds and the number of samples, respectively). $\overline{C}$ is the mean of the concentrations of all the compounds in all the samples. These predictions were performed in an independent set of 15 standards. The equation for the RRMSD shown corresponds to the global error (for all the compounds in all the 15 standards); for the RRMSD values of each independent compound, the first summatory is omitted.

In Table 1 the global RRMSD values for the architectures tested are shown. In all cases, either when only the 10 PAHs were determined or when the surfactant was added as the 11th compound (because it is also fluorescent) the lower RRMSD values (up to three times) were obtained when the scores were used instead of the full synchronous spectra. When scores were used, slightly better results were obtained when the surfactant was added as 11th compound. The architecture chosen for all further calculations (Fig. 4) consisted of an input layer with 14 nodes (scores), a hidden layer with 10 nodes and an output layer with 11 nodes (for the 10 PAHs and the surfactant).

Table 1
RRMSD(%) values for the different artificial neural networks assayed

Nodes in the hidden layer	Synchronous spectra		Scores	
	10 PAHs	10 PAHs+ Brij-35	1 PAHs	10 PAHs+ Brij-35
5	53.03	51.96	45.98	46.38
9	35.56	32.98	20.63	19.48
10	27.99	26.20	11.64	9.63
12	26.00	27.64	11.23	10.13
15	29.41	28.40	9.84	9.82
17	27.41	27.78	11.01	9.91
20	26.41	27.98	10.31	9.33
25	28.74	28.93	10.73	9.46

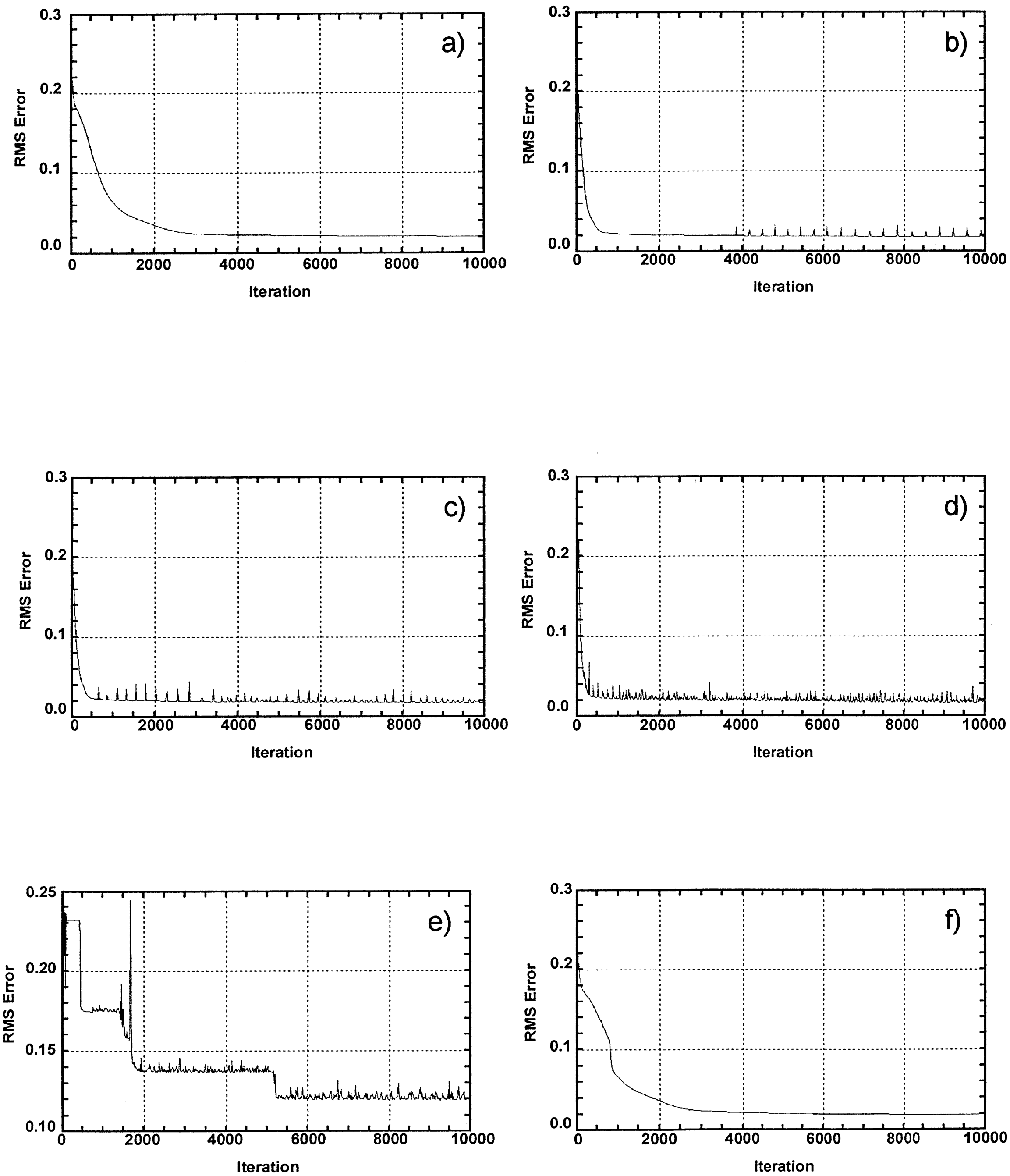

Fig. 3. Effect of the learning rate on the RMS error.

The results obtained with ANNs were compared with those obtained by other multivariate calibration procedures, such as PCR and PLS.

For PLS and PCR, which are factor analysis methods, the first step was the determination of the number of factors that allowed the system to be modelled

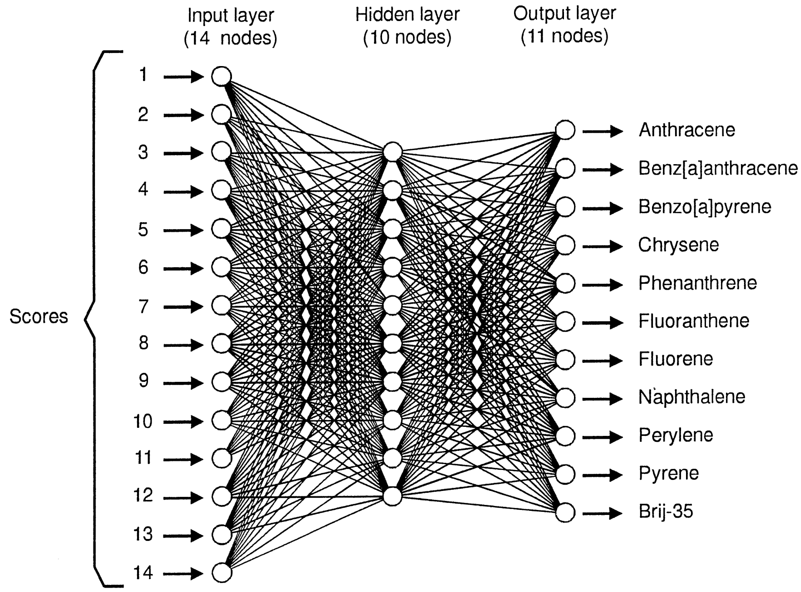

Fig. 4. Representation of the architecture of the optimal neural network chosen.

without overfitting the concentration data. For this purpose, a cross-validation method, leaving one sample at time, was used [23]. This means that, for a calibration set with n standards, the PLS and PCR calibrations were carried out with $n-1$ samples, and data thus obtained were used to calculate the concentrations of the sample left out. This process was repeated for the n samples. For each compound, the prediction error sum of squares (PRESS), defined as

$$\mathrm{PRESS} = \sum_{j=1}^{J} \left(\hat{c}_j - c_j\right)^2,$$

where $\hat{c}_j$ and c_j are the predicted and the real concentration for the sample j and J is the number of samples, was calculated. At the end of the process, the cumulative PRESS (sum of the PRESS for each sample of the calibration set) was obtained as a function of the number of factors. The optimum model was selected taking the minimum number of factors that yield a cumulative PRESS (CUMPRESS) that did not have any significant differences with the minimum CUMPRESS. For this purpose, the statistical F was used [7]. The value of F was calculated as:

$$F_k = \frac{\mathrm{CUMPRESS}(k)}{\mathrm{CUMPRESS}(\mathrm{min})},$$

where F_k is the calculated value, CUMPRESS(k) is the CUMPRESS value obtained in a model with k factors, and CUMPRESS(min) is the minimum value of CUMPRESS obtained. The minimum number of factors obtained was 15 for the PLS-2 model and 13 for the PCR model.

The calibration models for PCR and PLS were validated with the independent set of 15 standards previously used for ANN. In Table 2, the RRMSD values for each PAH are shown for the three models. These values indicate that the three methods have similar prediction ability, as they have similar RRMSD values, except in the case of anthracene, fluoranthene and naphthalene, which have higher values in the case of the PCR model. About the time required for calculation, the ANN chosen used about 3.75 min, while PCR needed 9 min and PLS needed 12.5 min. This means that ANN requires a shorter time for calculation than the other procedures (about 25–30% of PLS).

Finally, the three models were used for the analysis of water samples, which had been spikde with the 10 PAHs at four concentration levels (4, 9, 12 and 20 ng ml^{-1}). The results obtained by each procedure are shown in Table 3.

From a comparison of these results, it can be deduced that, overall, the best method is partial least

Table 2
RRMSD(%) values for each PAH obtained with the three procedures assayed

Compound	ANN (14–10–11)[a]	PCR (13 factors)	PLS (15 factors)
Anthracene	8.67	11.27	8.13
Benz[a]anthracene	8.06	6.72	8.10
Benzo[a]pyrene	5.83	5.41	5.02
Chrysene	12.49	12.46	13.04
Phenanthrene	8.88	8.85	7.54
Fluoranthene	10.35	17.21	10.34
Fluorene	13.16	14.89	12.55
Naphthalene	13.69	17.77	11.52
Perylene	5.70	3.80	3.60
Pyrene	9.07	11.07	11.85

[a] 14–10–11: 14 nodes in the input layer, 10 nodes in the hidden layer and 11 nodes in the output layer.

Table 3
Determination of PAHs in spiked natural water samples

Compound	Found (ng ml^{-1})			
	Added (ng ml^{-1})	ANN	PCR	PLS
Anthracene	4.2	3.4±0.4	6.1±0.3	5.2±0.2
	8.9	6.6±1.5	9.3±0.2	8.4±0.6
	12.9	7.3±1.9	10.4±0.4	10.2±0.7
	17.9	6.7±2.7	15.2±0.9	14.3±1.2
Benz[a]anthracene	3.9	4.4±0.4	4.3±0.3	3.7±0.1
	8.2	9.8±0.4	8.7±0.6	8.0±0.3
	11.9	11.0±0.3	11.1±0.6	10.6±0.4
	16.4	12.0±0.2	15.2±0.9	16.1±0.6
Benzo[a]pyrene	4.7	3.7±0.2	5.5±0.4	4.9±0.2
	10.0	9.5±0.3	11.0±0.8	10.2±0.6
	14.3	11.0±0.1	13.9±0.8	13.2±0.6
	19.3	11.9±0.1	21.2±1.2	20.2±1.0
Chrysene	3.8	3.0±0.2	2.3±0.2	4.4±0.2
	8.1	8.8±0.3	6.5±0.2	9.0±0.4
	11.7	10.7±0.5	9.1±0.3	11.8±0.6
	16.2	13.1±0.2	14.9±0.2	18.5±0.4
Phenanthrene	4.3	4.0±0.6	2.1±0.4	4.7±0.3
	9.1	8.8±0.6	5.3±0.3	8.5±0.3
	13.1	9.4±0.5	7.5±0.2	10.3±0.5
	18.1	10.1±0.1	10.9±0.2	15.1±0.6
Fluoranthene	4.7	5.9±0.1	9.5±0.6	7.0±0.1
	10.0	10.8±0.3	15.3±0.4	11.9±0.3
	13.1	12.1±0.2	17.5±0.6	14.5±0.3
	20.0	13.3±0.1	26.0±0.7	21.6±0.3
Fluorene	4.5	−0.5±0.1	−0.2±0.3	0.3±0.2
	9.7	0.8±0.5	2.8±0.2	3.3±0.1
	13.9	1.4±0.4	5.2±0.2	5.6±0.1
	19.3	0.8±0.3	8.8±0.3	9.0±0.1
Naphthalene	5.2	3.2±0.3	1.5±0.4	3.6±0.2
	11.0	9.2±1.0	5.5±0.2	7.9±0.4
	13.9	11.7±1.1	9.3±0.4	10.9±0.7
	21.9	14.2±0.5	14.6±0.4	17.6±0.7
Perylene	3.8	5.2±0.4	5.1±0.3	4.5±0.1
	8.0	10.5±0.4	9.9±0.6	9.1±0.4
	11.5	11.9±0.2	12.5±0.5	11.8±0.3
	16.0	13.4±0.1	19.2±0.8	18.2±0.5
Pyrene	4.1	3.0±0.2	5.7±0.6	4.2±0.2
	8.7	8.7±0.6	11.2±1.3	9.2±0.7
	12.6	10.4±0.1	13.5±1.0	12.0±0.5
	17.4	11.7±0.2	20.6±1.9	18.3±1.3

squares regression, although, in some cases, the differences with the other two procedures are not great. Overall results obtained by either artificial neural networks or principal component regression are roughly equivalent, although the prediction ability of each system depends on the individual compound to be determined. For example, ANNs have a prediction ability for chrysene, phenanthrene, fluoranthene and naphthalene, while PCR is better to predict anthracene, benz[a]anthracene and benzo[a]pyrene, while both procedures give similar predictions for perylene and pyrene.

The three calibration procedures were unable to give good predictions for the concentration of fluorene, which can be attributed to interferences from other substances present in natural water samples.

Fig. 5 shows the representation of the first principal component verses the second principal component for the calibration set, the validation set and the water samples. The difference in the predictions between PLS and ANN could be explained by taking into account of this representation. As is shown, the samples M1 and M2 are inside the calibration set, while M3 and M4 are clearly displaced from this set. This is

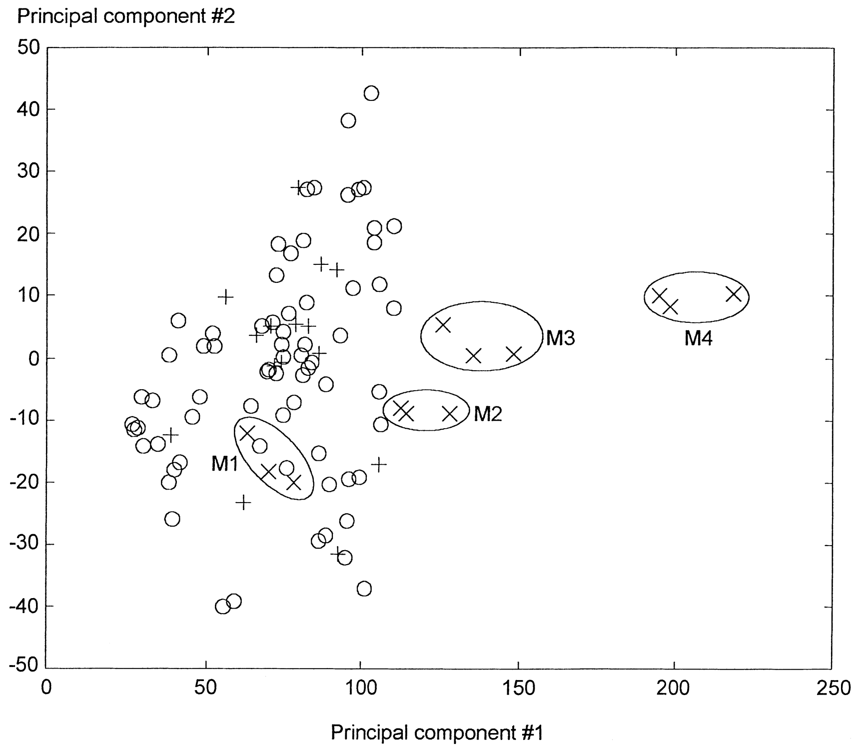

Fig. 5. Representation of the first principal component versus the second principal component of the PCA decomposition of the calibration set (○), validation set (+) and natural water (×) synchronous spectra.

probably due to the fact that, while the standards may contain high concentrations of some analytes but low of zero-concentrations of other analytes, the samples contains the same concentrations of all PAHs, and therefore, in some cases the spectra of the samples may more intense than those of the standards.

This difficulty can be solved by linear procedures, such as PLS, but not by the non-linear approach used by ANN.

5. Conclusions

Calibration with backpropagation neural networks allows the determination of the contents of PAHs in multicomponent samples with results similar to those obtained by the use of other multivariate calibration procedures (principal component regression and partial least squares regression), but in a much shorter time.

The scores obtained with principal component analysis of the spectral data should be used, instead of full synchronous spectra, in order to reduce the time required for calculation (up to 6 instead of full synchronous spectra, in order to reduce the time required for calculation (up to six times) and the error in the predictions (about three times).

The results obtained by ANNs in the simultaneous determination of 10 PAHs in spiked natural water samples show that artificial neural networks may become a valid alternative to other multivariate calibration procedures for the analysis of complex mixtures.

Acknowledgements

The authors thank the Comisión Interministerial de Ciencia y Tecnología (AMB98-0327) for supporting this study.

References

[1] F. Navarro-Villoslada, L.V. Pérez-Arribas, M.E. León-González, L.M. Polo-Diez, Anal. Chim. Acta 313 (1995) 93–101.

[2] P.X. Zhang, D. Littlejohn, Chem. Intel. Lab. Sys. 34 (1996) 203.
[3] A. Espinosa Mansilla, F. Salinas, M. Del Olmo, I.P. Paya, Appl. Spectrosc. 50 (1996) 449.
[4] R.D.B. Jiménez, A.I.J. Abizanda, F.J. Moreno, J.J.A. León, Clin. Chim. Acta. 249 (1996) 21.
[5] F.R. Vandevoort, K.P. Memon, J. Sedman, A.A. Ismail, J. Am. Oil Chem. Soc. 73 (1996) 411.
[6] R. Bro, J. Chemometr. 10 (1997) 47.
[7] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1193.
[8] P.J. Gemperline, J.R. Long, V.G. Gregoriou, Anal. Chem. 63 (1991) 2313.
[9] T.J. McAvoy, H.T. Su, N.S. Wang, M. He, J. Horvath, H. Semerjian, Biotechnol. Bioeng. 40 (1992) 53.
[10] C. Borggard, H.H. Thordberg, Anal. Chem. 64 (1992) 545.
[11] T.B. Blank, S.D. Brown, Anal. Chem. 65 (1993) 3081.
[12] A. Bjorseth (Ed.), Handbook of Polycyclic Aromatic Hydrocarbons, Marcel Dekker, New York, 1983.
[13] S.A. Wise, B.A. Benner Jr., R.G. Christensen, B.J. Koster, J. Kurz, M.M. Shantz, R. Zeizier, Environ. Sci. Technol. 25 (1991) 1595.
[14] C.A. Menzie, B.B. Potocki, J. Santodonato, Environ. Sci. Technol. 26 (1992) 1278.
[15] H.G. Kicinski, S. Adamek, A. Kettrup, Chromatographia 28 (1989) 203.
[16] C. Escrivá, E. Viana, J.C. Moltó, Y. Picó, J. Mañes, J. Chromatogr. 676 (1994) 375.
[17] E.R. Brouwer, A.N.J. Hermans, H. Lingeman, U.A.Th. Brinkman, J. Chromatogr. A. 669 (1994) 45.
[18] G. Codina, M.T. Vaquero, L. Comellas, F.B. Puig, J. Chromatogr. A. 673 (1994) 21.
[19] S.A. Wise, M.M. Shantz, B.A. Benner Jr., M.J. Hays, S.B. Shiller, Anal. Chem. 67 (1995) 1171.
[20] A.I. Krilov, I. Kostyuk, N.F. Volynets, J. Anal. Chem. 50 (1995) 494.
[21] T. Vo-Dinh (Ed.), Chemical Analysis of Polycyclic Aromatic Compounds, Wiley, New York, 1989.
[22] B.M. Wise, PLS Toolbox for use with MATLAB™, CPAC, Washington, 1992.
[23] A. Garrido Frenich, M. Martínez Galera, J.L. Martínez Vidal, M.D. Gil García, J. Chromatogr. A. 727 (1996) 27.