Chemometric methods as applied to the acoustic characterisation of zooplanktonic communities

C. M. Martínez

Oceanology Section, Facultad de Ciencias, Iguá 4225, 11400 Montevideo, Uruguay.
carmar@glaucus.fcien.edu.uy

Abstract

The reliability of Principal Components Regression (PCR) methods in the processing of multifrequency acoustic data is discussed. The method does not require the knowledge of individual scattering responses, can be applied to complex populations, is less sensitive to noise, and is effective for small organisms. The statistical model allows for the characterisation of the zooplanktonic community in terms of biological groups, as well as for the inclusion of non-acoustical variables, correlated with the biological distributions.

1. Introduction

Multifrequency acoustics is a powerful technique for the detection and biomass estimation of zooplankton assemblages. Normally, inversion methods require information about single individual scattering responses, which is normally difficult to estimate. The Non-Negative Least Squares (NNLS) [1] algorithm is the most widely used, as it converges more satisfactorily than simple regression methods. Using high frequencies, on the order of MHz, it is possible to detect small particles, but in this case range is severely diminished, and consequently remote sensing capability is lost.

Multivariate methods could be an alternative to NNLS and other least squares algorithm. In previous papers [2, 3], we have explored the PCR (Principal Components Regression) calibration method, on simulated and real data, which performs better than the NNLS method. In this paper we introduce some theoretical aspects related with the development of a general theory of calibration modelling. These considerations could improve the multifrequency technology for the estimation of biomass distribution.

2. Multivariate methods

The mathematical model for Multifrequency Acoustics is simple. Here, in matrix form:

$$\mathbf{I} = \mathbf{N}\,\mathbf{R}, \tag{1}$$

where **I** represents the instrumental data matrix (acoustic backscattering at several frequencies), **N** the matrix of abundances, and **R** the individual scattering responses for each group. The inversion problem, that is, the estimation of **N** knowing **I** and **R**, is a classical ill-posed problem. The associated measurement errors and the collinearity of the **I** matrix poses problems of convergence in the matrix inversion. In terms of the inversion model, **I** represents the independent variables and **N** the dependent ones.

To solve the problem, the Non-Negative Least Squares method (NNLS) has been chosen as the standard procedure. This method requires knowledge of individual scattering responses for all significant size classes present, and of course a precise calibration of the instrument itself. However, NNLS does not perform well when populations are complex, and/or for small class sizes. An instrumental approach is to increase the insonification frequency, in order to obtain a better detection of small organisms [4, 5]. However, increasing frequency introduces a loss in the remote sensing capability, as a result of wave attenuation.

The increasing contributions in the field of data processing applied to Analytical Chemistry have been guiding the development of a new discipline called "Chemometry". One of the most active research areas is the development of new methodologies for the computation of calibration models, that is, to solve for the concentrations of the different analytes in complex mixtures. However, a general theoretical framework is relatively recent, limiting a more systematic approach to the problem of inversion. Booksh and Kowalski [6] discussed an interesting approach, and although the examples presented in his paper come from the Analytical Chemistry field, the generalisation for other analogous problems is possible. The general idea is to discuss the development of calibration models (that is, the set of coefficients that solve the inverse problem), by considering the type of data available, which in turn depends on the kind of analytical instrument. From that, an instrument that gives a single datum per sample is a zero-order instrument, as a single number is a zero-order tensor. Examples of this kind of instrument are ion-selective electrodes or, for the acoustic case, a single measure of

TG Leighton, GJ Heald, HD Griffiths, G Griffiths, (eds.), 'Acoustical Oceanography', Proc. Institute of Acoustics Vol. 23 Part 2, 2001.

acoustic backscattering using a single frequency.

First-order instruments can be considered as arrays of zero-order instruments. They give multiple measurements per sample at the same time, and the data are ordered as a vector, that is, a first-order tensor. Well-known examples are spectrophotometers or chromatographs, and in the acoustic case, a multifrequency sonar. Increasing the order of the instrument and consequently, the order of the experimental data, hase as the main consequence an increase in the quality of the analysis. Following this reasoning, a second-order instrument provides a matrix (a second-order tensor) of data per sample. In the Analytical Chemistry field, these instruments are called "hyphenated" (meaning "combined" or "coupled"), e.g., GC-MS (Gas Chromatography-Mass Spectrometry).

In principle, there is no theoretical limit to the order of the instrument. Following Booksh and Kowalski, Table 1 summarises the main advantages and disadvantages of the different instruments. The main point here is related to the statistics behind each class of measurement. Actually, precise descriptions of the statistical properties of zero- and first-order models are available, but this is not the case for second-order models, for which there is still a lot of work to do. However, we note that second-order instruments have the capacity, in principle, to make more accurate estimations in the presence of interference.

Calibration Order	Interferences	Statistics
Zero order	Cannot detect, analysis biased	Simple, well defined
First order	Can detect, analysis biased	Complex, defined
Second order	Can detect, analysis accurate	Complex, not fully investigated

Table 1. Basic characteristics of different calibration models. Adapted from Booksh and Kowalski, 1994 [6].

A multifrequency sonar is a first-order instrument, so first-order calibration models can be developed using statistical techniques and linear algebra.

Multivariate calibration methods in Analytical Chemistry have been developed strongly since the 1980's, evolving from Classical Linear Regression (CLS), to Multiple Linear Regression (MLR), Principal Components Regression (PCR) and Partial Least Squares (PLS). Depending on the system under study, some methods perform better than others, but in general PCR and PLS are now the most widely used. In addition, each class of methods has variants, developed to suit specific needs. Until now, PCR is the only method in the list that has been tested in the processing of multifrequency acoustic data. In this method, the inversion problem is solved through the representation of the instrumental data matrix in terms of their orthogonal functions, which define the abstract space in which the original variance is expressed. The inversion is then facilitated because the original data matrix, normally highly collinear, is substituted by an orthogonal representation. Of course, the calculation of model coefficients needs some independent values for the dependent variables (in our case, size classes or groups). However, this is not a major problem, as plankton samples are normally taken during field trips.

In PCR calibration, the objective is the construction of a linear model where the value for each dependent variable (i.e., abundance of each size class or group), is related to the independent variables (the multifrequency data matrix) through a matrix **K** of model coefficients:

$$\mathbf{N} = \mathbf{I}\,\mathbf{K} \tag{2}$$

The **K** matrix is calculated as the multiplication of two matrices:

$$\mathbf{K} = \mathbf{V}\,\mathbf{B} \tag{3}$$

where **V** is the reduced matrix of eigenvectors of the covariance matrix of **I**, and **B** is the matrix of partial regression coefficients between **N** (the matrix of abundances in the calibration or learning set) and the empirical orthogonal functions of the covariance matrix of **I** (the principal components). The calculation of these partial regression coefficients is facilitated, from a mathematical point of view, as the data matrix **I**, (which is normally highly collinear), is substituted by its orthogonal representation in a reduced factor space. This substitution allows for improved cleaning of noise (random error, represented by the least-significant factors, not retained).

On the other hand, the most widely used method, Non-Negative Least Squares (NNLS), is in fact a Classical Least Squares (CLS) with a restriction. In CLS, the basis of the model is a "hard" or causal relationship between the instrumental response and the analyte level (or abundance level) [7]. This relationship is the individual scattering response. In PCR however, the relationship is statistical. From this, it is possible to include additional variables, related to the concentration (abundances), providing that the covariance matrix is substituted by the correlation one (to normalise for the utilisation of different physical units).

Another important difference between a hard-based model (CLS, NNLS) and soft-based model (like MLR, PCR and PLS), is that the latter does not require a precise knowledge of the levels of interference. In NNLS, for instance, it is necessary to known the individual scattering responses for *all* the size classes present in the biological community.

3. PCR performance

We have tested the PCR method using two approaches. First, we have simulated complex populations of particles, and analyzed the results of the inversion as a function of the individual scattering model and the Signal-to-Noise ratio (SNR). In these cases, it is possible to compare the PCR results with those obtained using the NNLS algorithm. Mixed populations of particles and its scattering response at seven frequencies (10, 40, 80, 120, 200, 300 and 500 kHz) were simulated. Second, we have applied the method and different combinations of parameters for data processing, to real data obtained in 1988 in the Atlantic Ocean, near the Glenans Islands (South Brittany, France). The information consists on multifrequency data at five strata, together with the identification and quantification of net samples (vertical hauls with Nansen Standard nets). Three profiles were used, so the amount of information is limited, but in any case the results were very acceptable. The population consisted of 26 zoological groups, some of them identified at the level of genera, with sizes (in terms of Equivalent Sphere Radius, ESR) ranging from 0.1 mm to 0.8 mm (eight size classes). The frequencies used were 38 kHz (Simrad EK 40), and 75, 90, 110 and 120 kHz (prototype sonar from IFREMER, [8]). So, the dimension of the *complete* acoustic data matrix was 15 observations × 5 frequencies, and those of the abundance data matrix was 15 observations × 26 groups. Using separately each profile for purposes of calibration and validation decreases the observations to 5. The paper of Martínez and David [3] gives a complete description of the ancillary data and the observation conditions. Several error estimates can be defined, but in the simulations a Standard Error of Prediction (SEP) was chosen, defined as the root-mean-square error between real and predicted values.

3.1 Sensitivity to noise

Comparisons using simulated populations show that PCR is less sensitive to noise than NNLS [2]. This characteristic is based on two aspects of the method. First, the regression does not possess matrix inversion problems, as it is well conditioned. Second, the exclusion of the orthogonal functions represented by the less-significant eigenvalues has the effect of filtering the non-correlated part of the signal. Figure 1 shows the results on simulated data. A complex population of five different size classes has been simulated, using the Anderson model [9] and the simplified Johnson model [10] for the calculation of individual scattering responses. The graphics illustrate the behaviour of the Standard Error of Prediction (SEP) ratio, defined as SEP_{NNLS}/SEP_{PCR}, as a function of the Signal-to-Noise ratio introduced in the scattering data. Using the simplified model of Johnson for the calculation of scattering responses, NNLS errors arrive up to 25 times greater than PCR ones. Of course, in this case the "hard" model can be applied because the individual scattering responses are known. NNLS errors are greater than PCR ones, depending on a) the information contained in the scattering response (i.e., the more complex is), and b) the correlation of the instrumental data matrix. Both causes are reflected in the graphics. The Anderson model is more complex, so the SEP ratios decrease, showing that at SNR values approaching 35dB, the prediction errors are comparable between the two inversion algorithms. Of course, the specific values depend on the simulation conditions, but the comparison illustrates a general feature, as the only difference is the scattering model chosen for the simulations.

3.2 Analysis of complex populations

The application of PCR to real data demonstrates its capability to calibrate for complex populations, in the cases when NNLS is precluded owing to the ignorance of all individual scattering responses. Table 2 shows the correlation coefficients between real and predicted values for the real data tested, using different variations of the analysis. Depending on the target group, the improvement on the prediction can be explained by its correlation with the independent variables added (it is convenient here to remind that the "independent" variables are defined as the set of instrumental data, whose corresponding data matrix is normally highly correlated).

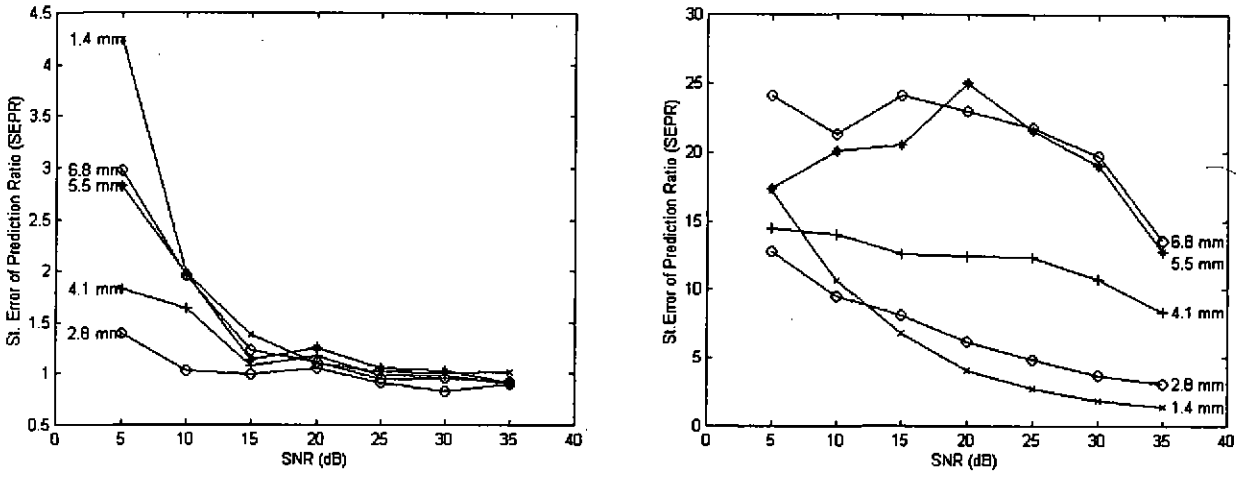

Figure 1. Standard Error of Prediction Ratio between the results obtained with the NNLS algorithm and those obtained with PCR modeling, as a function of the Signal-to-Noise-Ratio. Left: simulation using the Anderson model. Right: simulation using the simplified Johnson model.

GROUP	I	II	III	IV	ESR
Cirripedia	0.18	0.43	0.17	0.43	0.1
Naupli	0.25	0.67	0.25	0.69	0.1
Gasteropoda	0.68	0.60	0.66	0.75	0.1
Pelecypoda	0.74	0.41	0.71	0.41	0.1
Echinoidea	0.54	0.75	0.62	0.75	0.1
Small Copepoda	0.07	0.75	0.40	0.82	0.2
Cladocera	0.65	0.71	0.65	0.71	0.2
Ostracoda	0.46	0.64	0.46	0.64	0.2
Small Medusae	0.58	0.65	0.83	0.63	0.2
Doliolela	0.68	0.48	0.63	0.48	0.2
Medium Copepoda	0.54	0.78	0.68	0.78	0.3
Amphipoda	0.64	0.62	0.77	0.62	0.3
Polychaeta	0.48	0.42	0.48	0.40	0.3
Appendicularia	0.51	0.88	0.68	0.89	0.3
Medium Medusae	0.10	0.70	0.08	0.71	0.3
Chaetognata	0.16	0.51	0.17	0.52	0.4
Euphausiacea	0.54	0.89	0.87	0.86	0.5
Brachuyra	0.47	0.59	0.47	0.60	0.5
Anomura	0.55	0.70	0.47	0.69	0.5
Big Copepoda.	0.16	0.77	0.17	0.76	0.6
Siphonophora	0.31	0.53	0.27	0.71	0.7
Big Medusae	0.50	0.48	0.51	0.48	0.8
Total Copepoda	0.30	0.85	0.48	0.90	
Total Crustacea	0.20	0.88	0.48	0.89	
Total Medusae	0.48	0.65	0.49	0.63	
Total Non Crustacea	0.45	0.87	0.73	0.79	

Table 2. Maximal correlations for differents sets of independent variables.

I = acoustical data only;
II = acoustical data and depth of catch;
III = acoustical data and temperature;
IV = acoustical data, depth of catch and temperature.
ESR = Equivalent sphere radius in mm.

3.3 Analysis of organisms with weak acoustic responses.

Calibration for small sizes is also possible. By "small" we understand organisms with an ESR less than 1 mm. The main problem here is that, at the frequencies employed for detection, the scattering response for these organisms lies in the Rayleigh region, giving high collinearity to the acoustical data matrix. However, the PCR method can solve effectively the matrix inversion. Then, abundance estimation is possible at frequencies in the order of a few hundreds of kHz. This represents a general feature, but in addition, in some cases is also the correlation between the target group and other components that allows for this estimation. These two features has a particular relevance: it is not necessary to increase the frequency in order to characterise small organisms, and consequently, the remote sensing capability is preserved, as the attenuation of the sound is not a limiting factor. Figure 2 illustrates the mean results obtained for some groups of small sizes, and Figure 3 shows results for two sizes of medusae, together with medium- and big-sized Copepoda. These mean results are compared with the mean distributions (considering the three profiles of data).

3.4 Improvement of performance when other independent variables are included

Correlation coefficients improve in almost all the cases when another independent variable (depth of catch and temperature in the example) are included in the model (Table 2). This improvement is due to two main causes. First, the addition of new independent variables allows for improved cleaning of noise, as the number of eigenvalues increase. The identification of significant and non-significant factors is thus facilitated. Second, a certain amount of information is added [11].

3.5 Calibration for biological groups

Among the important advantages of a "soft" calibration model, like PCR, is the possibility to calibrate for biological groups, and not only for size classes. From an ecological point of view, a quantitative description of the community in terms of the species or groups gives more information than a size spectrum [12]. The description of a biological community only in terms of size implies a loss of useful information, in particular concerning the possible biological relationships. Nevertheless, the reconstruction of a size spectrum from biological information is possible, so both descriptions are available. It is fundamentally the statistical nature of the relationships between the abundances and the instrumental data used by the algorithm, which allows for the calibration in terms of biological groups. Of course, this implies that the acoustic survey needs to be complemented with an appropriate biological sampling, but this extra effort is compensated by an increase in useful information for the biologist.

4. Perspectives

On the basis of our experience working on simulation models, we have attempted to characterise additional applications. One of the most interesting ones is to explore the capabilities of this calibration method to give information about the time and spatial dynamics of zooplankton populations. We will briefly outline the conceptual basis of this possibility.

We have seen that the addition of non-acoustical variables to the acoustic data set improves the prediction error. The additional variables, selected for their correlation with the organism's abundance, add information and also allow for the resolution of more "acoustical" species. When the additional data is hydrological (for instance, salinity, temperature, stability) the calibration model represents this information in terms of their orthogonal representation. In other words, the calculated principal components represent the variability of the acoustic data *and* those of the hydrological field. In Oceanography, the decomposition in orthogonal functions has a solid theoretical and mathematical basis, and constitutes one of the most widely used methods for the parameterisation of the vertical (physical) structure. Moreover, the temporal extrapolation of the vertical structure can be accomplished, under certain assumptions, by extrapolating the orthogonal functions [13]. From this it would be possible to predict, at appropriated scales, the temporal variability of biomass. However, as the correlation between the physical and biological fields is not simple, this approach needs to be tested first on the basis of simulation and the analysis of relatively simple communities.

We are exploring the capabilities of PCR and similar algorithms more to gather information about the structure and dynamics of pelagic populations, especially zooplankton, than to obtain a size spectrum. The inclusion of non-acoustic variables in the model allows for a characterisation of the population structure not only in terms of its acoustic response, but also considering the physical structure of the water column. Then, additional information concerning the coupling between the physical and biological fields can be assessed.

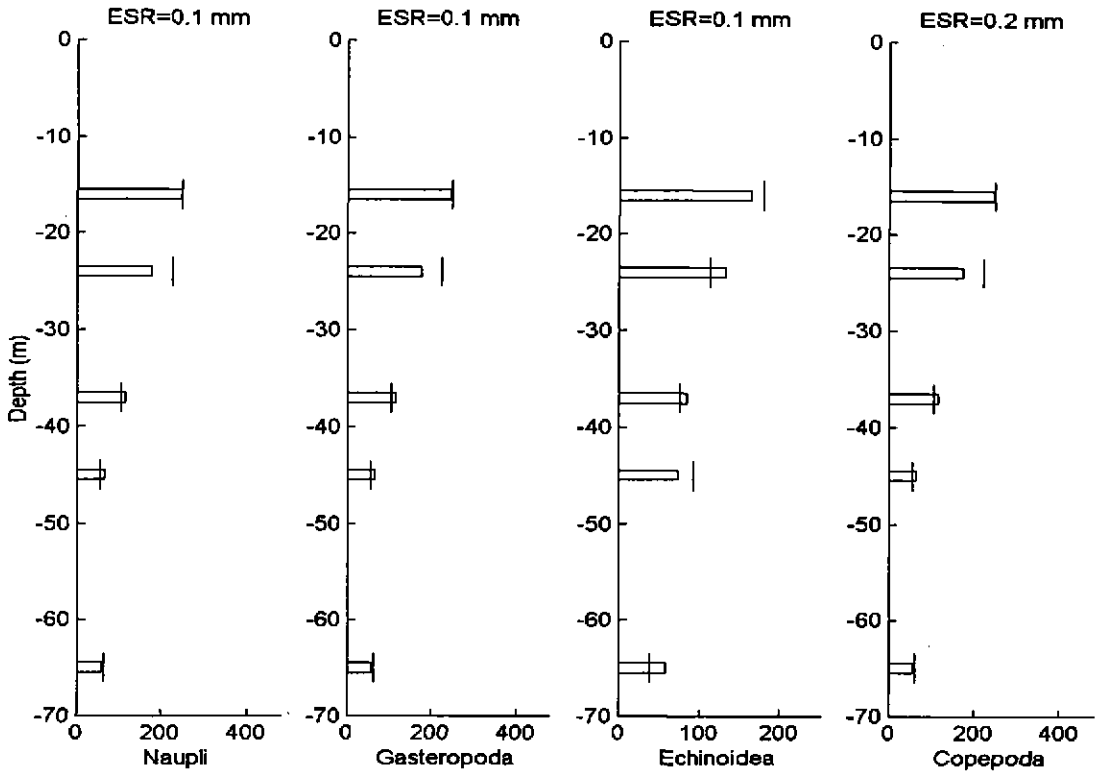

Figure 2. Mean distributions (individuals m^{-3}) for four groups of 0.1 and 0.2 mm Equivalent Sphere Radius (ESR). Bars represent estimations and the vertical line the real value. Prediction made using acoustic data and depth.

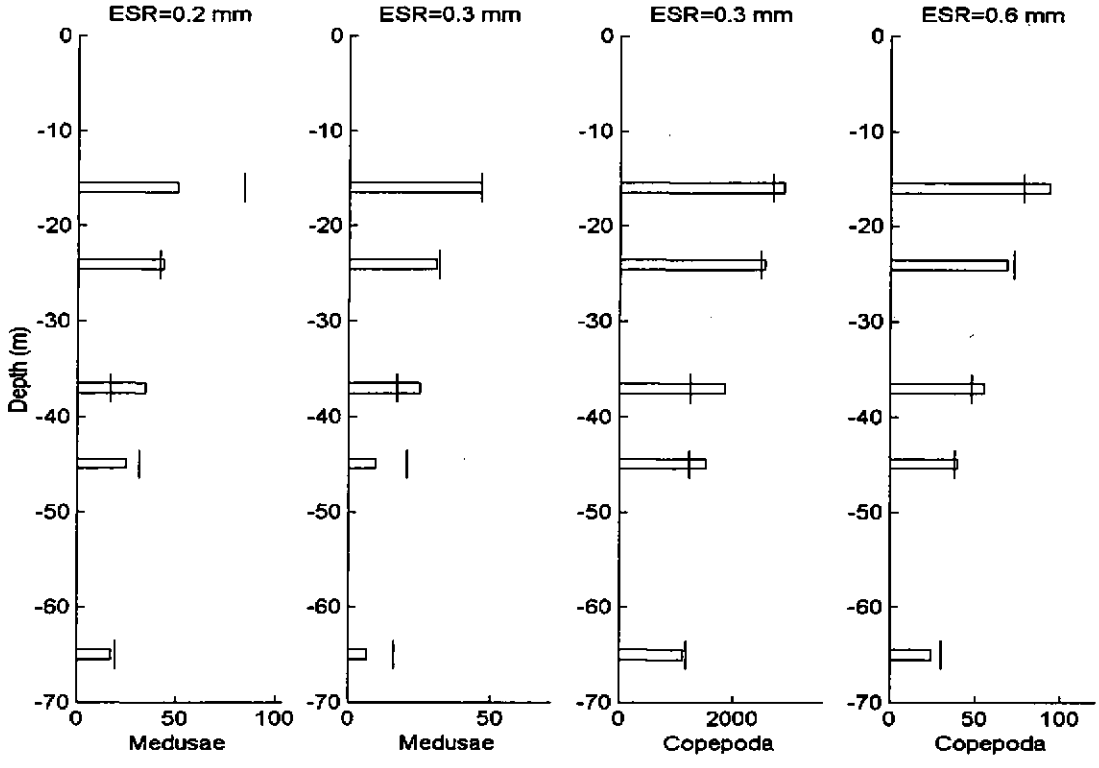

Figure 3. Representation as for Figure 2, for four groups of 0.2, 0.3 and 0.6 mm ESR.

Another aspect that could constitute a useful axis for future research is the exploration of other similar methods for calibration, well developed in Chemometry. For instance, Partial Least Squares (PLS), All Possible Combinations (APC) coupled with Multiple Regression or PCR, and finally, Neural Networks, wait to be tested. PLS is based on the representation of both the instrumental data matrix and the dependent data matrix by its orthogonal equivalent [14]. For some problems in Analytical Chemistry PLS gives better predictions than PCR. APC is a procedure to select the best set of coefficients calculated using MLR or PCR. Neural Network methods does not give additional information as do PCR or PLS, but it could constitute a practical approach for the design of dedicated signal-processing instruments.

The development of an instrument (or combination of instruments) giving second-order data could enhance the characterisation of complex populations of particles, solving for some of the problems encountered. Among these problems are the matrix effects, and baseline problems. Some experimental work (simulation and hard experiments) will provide the basis for an effective coupling of methodologies. From the instrumental point of view, the measurement of acoustic scattering at several frequencies and at the same time, at different angles between the emission and the reception could constitute the basis of a profiler that gives a second-order data matrix (number of observations × number of frequencies × number of angles).

Finally, an interesting point is related to the possibility of re-analysis of field data using multivariate methodologies, such as PCR. Existing multifrequency data is normally gathered together with physical, chemical and biological information. The acoustic data was mostly analysed using the NNLS method, but PCR can improve the estimation by taking into consideration this additional information. A re-analysis program will gather more information about community structure and possibly some dynamical aspects, and contribute to the establishment of the limits of application of the methodology and its further development.

5. Conclusions

We have briefly discussed how a theoretical framework for calibration modelling can contribute to the development of new instruments and procedures in the field of multifrequency acoustics. Among the most significant conclusions we have arrived in our work, the first is the demonstration that often, the solutions come from other fields of research, providing that the systems are analogous (that is, expressed by the same type of mathematical model). In this sense, there is a lot of experience in the Chemometric field that could be applied. Planktonic organisms, and particles in general, behave analogous to analytes in a chemical sample from the analytical viewpoint.

The second conclusion is that multifrequency acoustics, as other spectrometric approaches, has a enormous potential to describe not only the composition of a given system, but also some dynamical aspects. In this sense, coupled techniques for instrumentation and signal processing are on the table to be adapted. In particular, the addition of non-acoustical variables can improve the estimations of abundance. This addition can be done, from the instrumental side, in order to give second-order data matrices.

The third conclusion is important from the biological viewpoint. The utilisation of multivariate calibration models can solve the characterisation of the populations in biological terms. Even if a size spectrum is a practical description of a given population, the estimation of the number of individuals by biological groups is more relevant.

References

[1] Lawson CL and Hanson RJ. Solving least squares problems. Prentice-Hall Inc., Englewood Cliffs, N.J., 1974.

[2] Martínez CM and David PM. Principal component calibration models in the acoustic evaluation of zooplankton size spectra. I. Simulation studies. *ICES 79th Statutory Meeting*, Paper L:20, Sess. X. 1991.

[3] Martínez CM and David PM. Principal component calibration models in the acoustic evaluation of zooplankton size spectra. *J. Acoust. Soc. Am.* 1992; **92**(3): 1428-1439.

[4] Holliday DV and Pieper RE. Volume scattering strenghts and zooplankton distributions at acoustic frequencies between 0.5 and 3 MHz. *J. Acoust. Soc. Am.* 1980; **67**(1): 135-146.

[5] Holliday DV, Pieper RE and Kleppel GS. Determination of zooplankton size and distributions with multifrequency acoustic technology. *J. Cons. int. Explor. Mer* 1989; **46**: 52-61.

[6] Brooksh KS and Kowalski BR. Theory of Analytical Chemistry. *Anal. Chem.* 1994; **66**(15):782A-791A.

[7] Thomas EV. A primer on multivariate calibration. *Anal. Chem.* 1994; **66**(15): 795A-804A

[8] Alais P, Challande P, Diner N and Person R. Développement d'un sonar multifaisceaux. Colloque de

Physique, 1er. Congres Francais d'Acoustique 1990. *J. Acoustique*, 1990, C2, No 2 (suppl.), Tome 51,:321-324.

[9] Anderson VC. Sound scattering from a fluid sphere. *J. Acoust. Soc. Soc. Am.* 1950; **22**(4):426-431.

[10] Johnson RJ. Sound scattering from a fluid sphere revisited. *J. Acoust. Soc. Am.* 1977; **61**(2):375-377.

[11] Lorber A and Kowalski BR. The effect of interferences and calibration design on accuracy: implications for sensor and sample selection, *J. Chemom.* 1988; **2**: 67-79.

[12] Slobodkin LB, Botkin DB, Maguire B, Moore B and Morowitz H. On the epistemology of ecosystem analysis. In: Estuarine Perspectives, ed. V.S. Kennedy, Academic Press, 1980, pp. 497-507.

[13] Vasilenko VM and Mirabel AP. Parametrization of the vertical structure of currents in the Tropical Atlantic by means of statistically orthogonal functions (SOF). *Oceanology USSR* 1976; **16**(2): 124-127.

[14] Geladi P and Kowalski BR. Partial Least-Sqaures: a tutorial. *Anal. Chim. Acta* 1986; **185**:1-17.