Available online at www.sciencedirect.com
SCIENCE DIRECT®
Talanta 59 (2003) 735–749
Talanta
www.elsevier.com/locate/talanta

Multicomponent analysis of electrochemical signals in the wavelet domain

Marina Cocchi [a,*], J.L. Hidalgo-Hidalgo-de-Cisneros [b], I. Naranjo-Rodríguez [b], J.M. Palacios-Santander [b], Renato Seeber [a], Alessandro Ulrici [c]

[a] Dipartimento di Chimica, Università di Modena e Reggio Emilia, Via Campi 183, 41100 Modena, Italy
[b] Departamento de Química Analítica, Facultad de Ciencias, Universidad de Cádiz, Polígono Río San Pedro, Apartado 40, 11510, Puerto Real, Cádiz, Spain
[c] Dipartimento di Scienze Agrarie, Università di Modena e Reggio Emilia, Via Kennedy 17, 42100 Reggio Emilia, Italy

Received 5 August 2002; received in revised form 24 October 2002; accepted 13 November 2002

Abstract

Successful applications of multivariate calibration in the field of electrochemistry have been recently reported, using various approaches such as multilinear regression (MLR), continuum regression, partial least squares regression (PLS) and artificial neural networks (ANN). Despite the good performance of these methods, it is nowadays accepted that they can benefit from data transformations aiming at removing baseline effects, reducing noise and compressing the data. In this context the wavelet transform seems a very promising tool. Here, we propose a methodology, based on the fast wavelet transform, for feature selection prior to calibration. As a benchmark, a data set consisting of lead and thallium mixtures measured by differential pulse anodic stripping voltammetry and giving seriously overlapped responses has been used. Three regression techniques are compared: MLR, PLS and ANN. Good predictive and effective models are obtained. Through inspection of the reconstructed signals, identification and interpretation of significant regions in the voltammograms are possible.

Keywords: Differential pulse anodic stripping voltammetry; Multivariate calibration; Fast wavelet transform; Variables selection

1. Introduction

One of the main limitations to the application of electroanalytical techniques in the field of quantitative analysis is often due to lack of selectivity. In fact, it often happens that different species undergo oxidation or reduction at potential values that are very close to each other. In the case of differential pulse and square wave voltammetries, serious overlapping occurs when the difference in the peak potentials is less than 100 mV divided by the number of electrons involved in the electrode charge transfer. This situation is rather common in practice, since 100 mV represent an appreciable fraction of the accessible potential region. Besides experimental manipulations like changes of pH, of

* Corresponding author. Tel.: +39-59-2055029; fax: +39-59-373543.
E-mail address: cocchi@unimore.it (M. Cocchi).

0039-9140/02/$ - see front matter
PII: S0039-9140(02)00615-X

the supporting electrolyte, or the use of modified electrodes, chemometrics offers efficient alternatives to solve the problem of overlapping signals. The main approaches employed are deconvolution or semidifferential techniques coupled to curve fitting [1–5], multivariate curve resolution [6,7], and multivariate calibration [8–10]. Many successful applications of multivariate calibration in the field of electrochemistry have been recently reported [11–19], using different regression methods, i.e., multilinear regression (MLR) [15], principal component regression [14,16,19], continuum regression [12], partial least squares regression (PLS) [11,13,16,19] and artificial neural networks (ANN) [17,18].

The main advantage of using regression methods based on latent variables, such as PCR, PLS, etc., lies in their flexibility, which allows modelling of complex signals also in the presence of background noise. Despite the generally good performances of these methods, it is nowadays accepted that they can benefit from data transformations aiming at removing baseline effects, reducing noise, compressing the data [10,20]. The wavelet transform (WT) [21] is very efficient for all these purposes, since it offers the advantage of performing data reduction and denoising at the same time. The fast wavelet transform (FWT) has been applied as a pre-processing tool in multivariate calibration of NIR spectra [22–25], of fluorescence data [26], of X-ray powder diffraction spectra [27], while example of multivariate calibration of electroanalytical signals through FWT have not been reported so far.

The optimal wavelet filters, apart from few exceptions [22], are usually chosen empirically looking at the decomposition of the mean spectrum or at the shape of the signals. In the quoted papers, the level of decomposition is chosen either considering the features of the mean spectrum or simply as the maximum possible level of decomposition. As regards feature selection, mainly two approaches are proposed: (1) the wavelet coefficients are thresholded by using criteria based on the evaluation of PLS weights [24] or PLS regression coefficients [22]; (2) the wavelet coefficients are previously sorted by variance [23,26,27] or by correlation [25], and the subset giving stable or best performing regression models, is then selected.

In particular, Niemoller et al. [25] considered a fixed number, M, of the ranked (according to correlation with the y properties) coefficients, from which a starting population is derived and fed into a genetic algorithm (GA), which seeks the best combination of the M wavelet coefficients. The fitness function to be optimised contains the standard prediction errors for both calibration and internal validation sets, relative to MLR models. This approach seems particularly appealing because many different combinations of coefficients are tested. However, the use of GA is computationally heavy and the application of GA on a limited preselected number of coefficients further limits the search.

In the present work, we adopted a simplified approach where, instead of using GA, the selection of the wavelet coefficients to be used as the predictor variables is done by the recursive application of MLR models. Once the optimal wavelet coefficients are selected, different regression techniques can be employed for the calculation of the final calibration model. Furthermore all possible decomposition levels are considered. The proposed methodology goes through the following steps:

- the signals are decomposed into the wavelet domain by using the FWT at the maximum level of decomposition;
- for each level of decomposition the wavelet coefficients are sorted either according to their variance or to their squared correlation coefficient calculated with respect to the analyte concentrations;
- for each level of decomposition, the number of wavelet coefficients to be retained can be fixed or can be iteratively determined searching for the minimum of the standard deviation of error of predictions ($SDEP_{LOO}$, estimated by the Leave One Out procedure) by means of MLR;
- finally, the optimal decomposition level is considered the one giving the highest squared correlation coefficients.

The selected coefficients constitute a set of independent variables, which can be fed to differ-

ent regression techniques. For interpretative purposes, both the selected coefficients and the calculated regression coefficients can be reconstructed into the original domain by using the inverse FWT.

Further critical steps are the choice of the most suitable wavelet function and of the padding criterion [28–30]. In this work, 15 different wavelet functions and three kinds of padding were tested. The various sets of selected coefficients, corresponding to the different combinations of these options, were used as input to three regression techniques: MLR, PLS and ANN. The performance of the different regression models has been tested evaluating their predictive abilities on an external validation set.

The outlined WT-based feature selection procedure has been applied to a set of seriously overlapped voltammetric signals recorded on mixtures of thallium and lead in the concentration range 0.1–1 mg l^{-1}, which were recently collected by some of us; in a preliminary paper [31] this data set was analysed by ANN regression, coupled to Fourier Transform or WT compression. The results obtained were promising, even if WT was employed only for denoising purpose.

The systematic analysis conducted in the present work showed that quite satisfactory regression models can be obtained in the correspondence to different parameter combinations, suggesting that no general rules for the selection of optimal regression parameters (e.g. wavelet function, padding criterion, regression technique) can be drawn.

2. Methods

2.1. Wavelet analysis and feature selection

The WT is a powerful signal processing technique, whose peculiarity lies in the ability to map the frequency content of a signal as a function of the original domain, offering the possibility of (dual) time–frequency localisation. For a detailed description of the WT and of its properties reference is given to dedicated literature [21,28,32,33]; only a brief description is provided here. The discrete WT has been implemented through the Mallat's pyramidal algorithm also called FWT. It operates on an individual discrete signal of length 2^l by splitting it into 2^{l-1} long orthogonal subspaces, called approximations and details respectively. The decomposition is performed applying two filters (each wavelet being uniquely defined by a set of wavelet filter coefficients) to the original signal: a low-pass filter only retaining the low frequency content of the signal, i.e. the approximations, and a high-pass filter, collecting the high frequency content, i.e. the details. The procedure can be recursively applied (wavelet tree) by applying the same two filters to the approximation vector, until the length of the resulting vectors equals 1, as shown in Fig. 1. In this way, sharp and coarse properties of the signal are captured and disjointed into different sub-spaces, i.e. vectors or sets of wavelet coefficients, obtaining the so called 'signal multiresolution'. For each level of decomposition, j, it is possible to obtain a perfect reconstruction of the original signal by inverse FWT, using the approximations at level j and all the details from j to 1 level. In other words, the signal is represented in terms of a unique orthonormal basis [cAj cDj cDj−1...cDl]. On the contrary, when the goal is to remove noise or to perform data compression or feature selection, only a representative, i.e. informative for the given purposes, set of wavelet coefficients is retained by applying a suitable thresholding procedure.

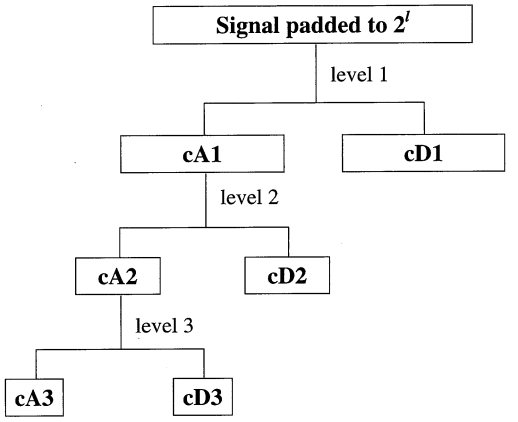
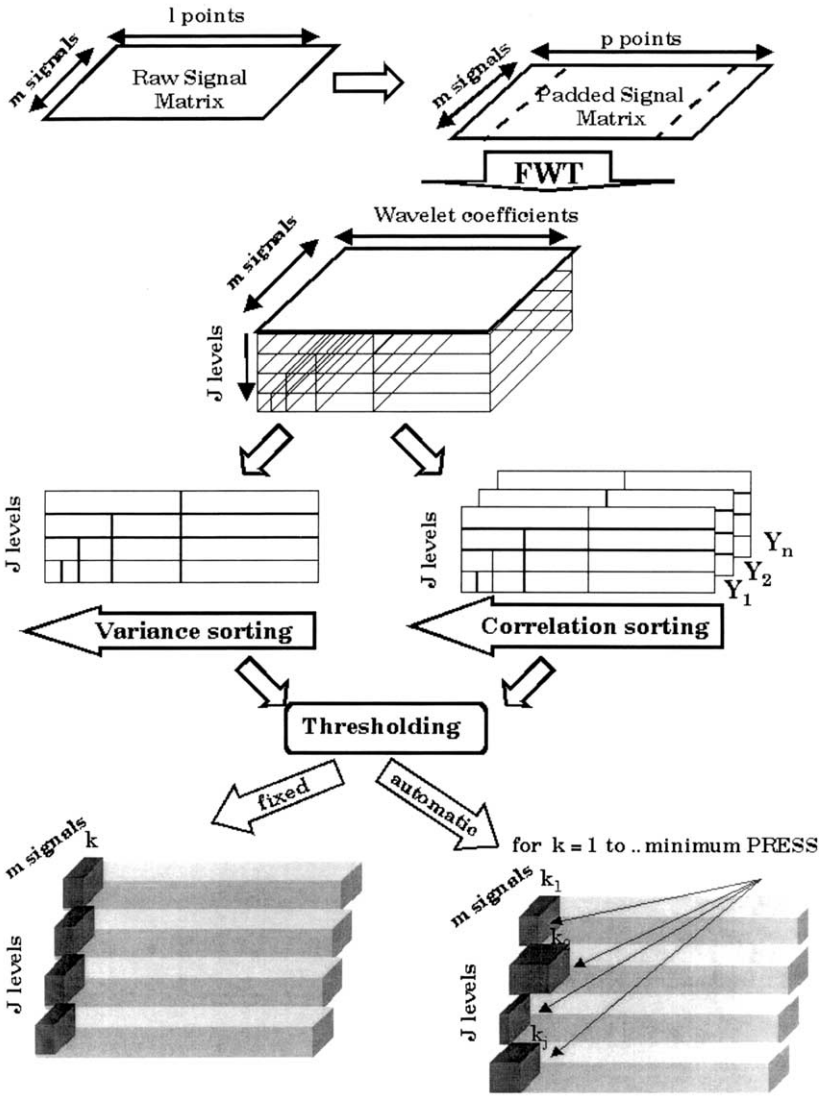


Fig. 1. Mallat pyramid algorithm. Approximation and details vectors are indicated by cA and cD respectively.



Scheme 1.

In the present work FWT is applied in order to accomplish feature selection prior to regression analysis. The procedure is outlined in Scheme 1. First, the signal matrix is padded to the next power of two, then each signal is decomposed in the wavelet domain until reaching the maximum decomposition level, obtaining a three-dimensional matrix with dimension $j_{level} \times p_{wavelet\ coefficients} \times m_{signals}$. In the subsequent step this three-dimensional matrix can be reduced to a two-dimensional matrix by calculating, alternatively: (i) the variance of the wavelet coefficients along the dimension of the signals (variance sorting), or (ii) the squared correlation coefficients with the y variables (correlation sorting). In the latter case, as many matrices as y variables are obtained. The elements of each row, i.e. each level of decomposition, of the variance matrix or of the correlation coefficient matrices respectively, are sorted in ascending order. Each slice of the wavelet

coefficients matrix ($p_{wavelet\ coefficients} \times m_{signals}$ at a given level of decomposition) is ordered accordingly.

Then, for each level of decomposition, the coefficients are thresholded according to two different criteria:

(1) A fixed number, k, of coefficients, defined by the user, is selected. In the case of variance sorting, the first k sorted coefficients are retained; in the case of correlation sorting, for each y variable a different number of coefficients may be chosen: $k1$ for $y1$, $k2$ for $y2$, and so on. These coefficients correspond to the first $k1$, $k2$, etc. sorted elements of the relevant correlation matrices. The coefficients to be selected for further analyses include all those chosen for each single y variable, without repetition, which means that if the same coefficient has been selected for more than one y variable, it is only considered once.

(2) An automatic selection criterion is implemented as follows. A first screening prunes the sorted wavelet coefficients by excluding those coefficients that show a pairwise correlation higher than 0.90 with at least one of the preceding coefficients. For each y variable, the number of retained coefficients is progressively increased from one to the rank of the wavelet coefficients matrix of the considered level ($p_{coefficients} \times m_{signals}$) and the corresponding MLR models are calculated. In order to obtain more stable regression models the pseudo-inverse matrix is used in the regression equation. For each y variable, the coefficients corresponding to the regression model attaining the minimum $SDEP_{LOO}$ are selected. The coefficients to be selected for further analyses include all those chosen for each single y variable, without any repetition.

In this way, a set of optimal wavelet coefficients is selected for each level of decomposition; the average squared correlation coefficient of each of these coefficients over the y variables is calculated. The optimal decomposition level is considered as that showing the highest mean squared correlation coefficient value.

The selected wavelet coefficients of the optimal decomposition level are then used as input variables for different regression methods.

The algorithm for performing selection procedure was written in MATLAB® 6.1 language by employing the Wavelet Toolbox ver. 2.1 [34].

2.2. Filters and values of the parameters

The decomposition into the FWT domain is essentially based on a simple scheme: convolution and downsampling. As usual, when a convolution is performed on finite-length signals, border distortions arise. Generally, to deal with this problem the signal is extended on the boundaries (signal padding) by computing few extra coefficients at each stage of the decomposition process, in order to get a perfect reconstruction.

The evaluation of the effects of different padding criteria on the resulting calibration models can be extremely important when the independent variables are wavelet coefficients deriving by the application of the FWT to a set of signals. In fact, the values of the wavelet coefficients that are calculated vary depending on the criterion that is adopted for signal extension. For this reason, three different padding criteria [29,30,34], that are available in the Wavelet Toolbox® for MATLAB, have been systematically compared in this work:

1) sym—symmetric padding: signals are recovered outside their original support by symmetric boundary value replication;
2) zpd—zero padding: signals are extended adding zeros outside the original support;
3) spd—smooth padding (order 1): signals are recovered outside their original support by a first-order derivative extrapolation: this is done using a linear extension fit to the first two and last two values.

Fifteen wavelets belonging to different families have been considered in the present study: 7 orthonormal wavelets from the Daubechies family (dbl, db2, db3, db4, db5, db10 and db20), 3 coiflets (coif1, coif2 and coif5) and 5 symlets (sym4, sym5, sym6, sym7 and sym8).

Both variance and correlation sorting of the coefficients were used. Both automatic and fixed, with four coefficients (two for each y variable in

the case of correlation sorting), selection criteria were employed.

All the combinations resulting from the above cited values of the parameters have been tested: this led to $3 \times 15 \times 2 \times 2 = 180$ cycles of calculations.

2.3. Partial least squares regression

The wavelet coefficients have always been mean-centred. The optimal number of PLS components has been chosen by cross validation. In order to determine the number of significant components, r^*, the value of the predicted residual error sum of squares, $PRESS_{LOO}$ (estimated by the Leave One Out procedure), obtained by adding a further component, is compared with the $PRESS_{LOO}$ value corresponding to the previous one. When the resulting ratio $[PRESS_{LOO}(r^*+1)/PRESS_{LOO}(r^*)]$ is higher than 1, r^* is reached.

The performance of each PLS model has been tested by the standard deviation of error of predictions, $SDEP_{TEST}$, estimated on a test set of 9 mixtures. For each combination of parameters (wavelet filter, padding, sorting and selection criteria) the best performing PLS models were selected and their predictive ability was further checked by an external validation set ($SDEP_{EXT}$) consisting of 8 mixtures.

For the calculations the PLS Toolbox ver. 2.1.1 [35] was employed and a MATLAB routine was written in order to calculate all the 180 PLS models automatically.

2.4. Multilinear regression

The wavelet coefficients have always been mean-centred. The pseudo-inverse has been used in the MLR equation with zero intercept. The performance of each MLR model has been tested by the standard deviation of error of predictions, $SDEP_{TEST}$, estimated on a test set of 9 mixtures. For each combination of parameters (wavelet filter, padding, sorting and selection criteria) the best performing MLR models were selected and their predictive ability was further checked by an external validation set ($SDEP_{EXT}$) consisting of 8 mixtures.

For the calculations of the MLR models a MATLAB routine was written in order to calculate all the 180 MLR models automatically.

2.5. Artificial neural network regression

Since the training set consists of 31 objects and the number of selected wavelet coefficients ranged from 3 to 12, only the n_i-2-2 topology, where n_i is the number of input coefficients, was considered in order to avoid overfitting [36,37].

The number of adjustable parameters (N) can be calculated by the formula [37]:

$$N = (\text{input nodes} \times \text{hidden nodes}) + (\text{hidden nodes} \times \text{output nodes}) + \text{hidden nodes} + \text{output nodes}.$$

It is clear that in order not to exhaust the degrees of freedom of our system not more than 11 input coefficients should be considered and that overfitting is likely to occur when the number of input coefficients exceed 6. For comparative purposes ANN regression models have been computed for all the 180 sets of coefficients but among the best performing models we selected those bearing a number of input coefficients not higher than 6. The configurations of the neural models tested were:

- training algorithm: improved back-propagation.
- activation functions: linear for the input layer and all possible combination of gaussian, sigmoid and hyperbolic tangent function for the hidden and output layers.

The program QNET® 2000 has been used for the ANN calculations.

The training of the net was stopped by minimising the standard deviation of error of predictions, $SDEP_{TEST}$, estimated on a test set of 9 mixtures. Since the starting weights are randomly generated for each set of coefficients five ANN runs were made and the resulting SDEPs were averaged.

For each combination of parameters (wavelet filter, padding, sorting and selection criteria) the

best performing ANN models corresponding to low $SDEP_{TEST}$ and to small number of coefficients, were selected. However, it is worth noticing that in the case of ANN the test set is used to stop the training of the network and thus it does not represent a true validation set, being rather a monitoring set. Accordingly the predictive ability of the chosen ANN models was checked by the external validation set ($SDEP_{EXT}$) of 8 mixtures.

3. Experimental section

3.1. Differential pulse anodic stripping voltammetry

The differential pulse anodic stripping voltammetry measurements were carried out at an Autolab/PGSTAT20 electrochemical system coupled to a Metrohm VA 663 Stand. An electrochemical three electrode cell, with a platinum auxiliary electrode, a silver/silver chloride, 3 M potassium chloride reference electrode and an Hanging Mercury Drop Electrode, from Metrohm, was employed.

Analytical reagent grade chemicals were used throughout the experiments. Voltammograms were recorded at room temperature. All solutions were de-aerated with nitrogen for at least 10 min prior to realising the experiments. A 2 M acetic acid/2 M ammonium acetate buffer solution was utilised as supporting electrolyte (pH 4.8–5.0). Lead and thallium solutions were prepared from nitrate salt stock solutions at 250 mg l^{-1} concentration.

The voltammetric parameters were as follows: deposition potential = −1.30 V; deposition time = 120 s; rest period = 20 s; initial potential = −1.30 V; end potential = 0.00 V; scan rate = 8.5 mV s^{-1}; pulse amplitude = 0.10 V; pulse time = 0.07 s; pulse repetition time = 0.6 s. The drop surface was approximately 0.52 mm^2.

A region of the full voltammogram of each sample corresponding to 80 points in the potential range from −0.30 to −0.70 V was used for the multivariate calibration analysis.

3.2. Sampling

Forty mixtures of thallium and lead at concentrations ranging from 0.1 to 1.0 mg l^{-1} were experimentally analysed. The whole experimental domain was spanned as shown in Table 1.

Nine out from these mixtures were used as internal test set (monitoring set for ANN), namely T2, T9, L3, L8, T1L6, T4L4, T10L5, T6L1, and T9L9, where L indicates lead, T indicates Thallium, 1 corresponds to a concentration of 0.1 mg l^{-1}, 2 corresponds to 0.2 mg l^{-1} and so on. After some time 8 additional mixtures were measured in order to obtain an external validation set ($TEST_{EXT}$); they are shown in Table 1 and correspond to T1L9, T2L10, T3L5, T5L3, T6L8, T8L6, T9L1, and T10L2, respectively.

4. Results and discussion

The results obtained with the automatic selection of the wavelet coefficients gives generally lower standard deviation error of calculations (SDEC) and of prediction of the validation set ($SDEP_{TEST}$) with respect to the fixed selection criterion.

The results obtained with the automatic selection criterion and different combinations of the other parameters for each regression technique are shown in Fig. 2a–d. In this figure, SDEC (Fig. 2a and c), and $SDEP_{TEST}$ (Fig. 2b and d) values are reported for each cycle of calculations. Each plot corresponds to a given sorting criterion of the wavelet coefficients, different symbols being used for the three padding criteria. On the abscissa for each regression technique, in the order MLR, PLS, and ANN, the wavelet filters are ordered as follow: db1, db2, db3, db4, db5, db10, db20, sym4, sym5, sym6, sym7, sym8, coif1, coif2, and coif5. Summarising each group of 15 points on the abscissa corresponds to a different regression method and within the 15 points the first 7 ones correspond to the daublets, the following 5 ones to the symlets and the last 3 ones to the coiflets wavelet family.

It can be seen that the variance sorting criterion performs generally better with respect to the

Table 1
Composition of the samples

[TI] (mg l^{-1})	[Pb] (mg l^{-1})										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0		L1[a]	L2[a]	L3[b]	L4[a]	L5[a]	L6[a]	L7[a]	L8[b]	L9[a]	L10[a]
0.1	T1[a]	T1L1[a]					T1L6[b]			T1L9[c]	
0.2	T2[b]		T2L2[a]					T2L7[a]			T2L10[c]
0.3	T3[a]			T3L3[a]		T3L5[c]			T3L8[a]		
0.4	T4[a]				T4L4[b]					T4L9[a]	
0.5	T5[a]			T5L3[c]		T5L5[a]					T5L10[a]
0.6	T6[a]	T6L1[b]					T6L6[a]		T6L8[c]		
0.7	T7[a]		T7L2[a]					T7L7[a]			
0.8	T8[a]			T8L3[a]			T8L6[c]		T8L8[a]		
0.9	T9[b]	T9L1[c]			T9L4[a]					T9L9[b]	
1.0	T10[a]		T10L2[c]			T10L5[b]					T10L10[a]

Columns, lead concentration; rows, thallium concentration. [a]Training set; [b]test set (monitoring set for ANN); [c]external test set.

regression sorting one, giving on average lower values of both SDEC and $SDEP_{TEST}$.

The selection of wavelet coefficients with the biggest variance has already given good results in regression tasks [23,26]. However, the lowest SDEC and $SDEP_{TEST}$ values obtainable by the different criteria are of similar magnitude.

The performance of the different models does not differ significantly by varying the padding criterion used, except for the wavelet filters of higher orders. This is somewhat to be expected since the extension of the signal on the boundaries (padding) requires the computation at each stage of decomposition, of few extra coefficients, whose number depends on the length of the filter. Different padding criteria lead to more and more different values of the wavelet coefficients with increasing levels of decomposition. However, the results obtained show that this problem does not constitute serious drawback if an effective criterion for the selection of the wavelet coefficients is adopted.

The linear (MLR and PLS) and the non linear (ANN) regression techniques furnish equivalent models with respect to fit and predictive capability, thus indicating that a linear equation is sufficient to explain the behaviour of the investigated system.

The best performing models for each combination of parameters are reported on Table 2 together with the $SDEP_{EXT}$ values. The $SDEP_{EXT}$ values are systematically worse than the corresponding $SDEP_{TEST}$ ones, resulting anyway within similar ranges. This is probably due to the fact that the mixtures belonging to the external test set were measured in a different time period and a calibration transfer procedure [13] has not been applied. In Table 2, the SDEC, $SDEP_{TEST}$ and $SDEP_{EXT}$ values for the PLS regression model calculated by using the whole voltammograms, each one consisting of 80 points, are also reported. The dimensionality of these PLS models was of 3 and 4 significant latent variables, according to leave one out cross validation, for thallium and lead respectively. The PLS models reported on Table 2, obtained after feature selection, show a dimensionality for both dependent variables (thallium and lead concentrations) that is lower than (2 significant PLS components) or equal to the dimensionality of the PLS models on the untreated voltamogramms. Only in the cases where the PLS models converged to the MLR, i.e. where the number of PLS components is equal to the number of selected coefficients, the dimensionality for the PLS model for thallium reached 4 latent variables (PLS models on rows 3, 5 and 10, Table 2). At a first sight, it may seem that there is no significant reduction of the model complexity after wavelet analysis. However, the number of PLS latent variables is influenced by the fact that the wavelet

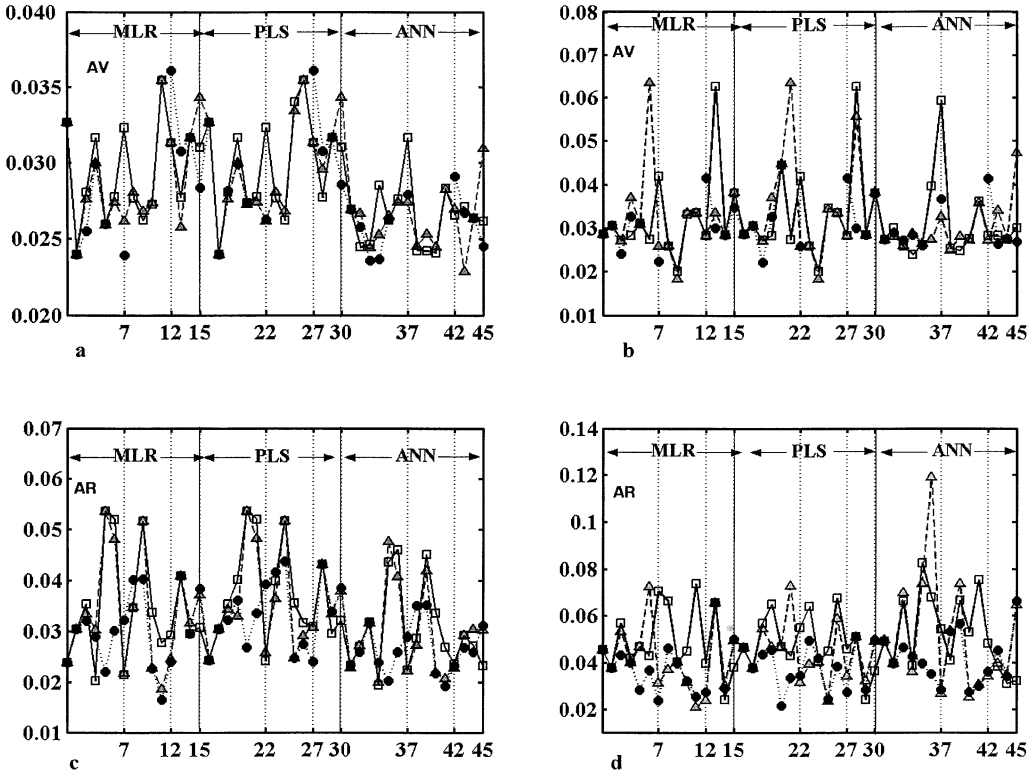


Fig. 2. (a) Automatic coefficient selection—Variance sorting (AV). SDEC values (training set) vs number of cycles for the different padding criteria: ▼, zero padding; □, symmetric padding; ●, smooth padding. (b) Automatic coefficient selection—Variance sorting (AV) $SDEP_{TEST}$ values (test/monitoring set) vs number of cycles for the different padding criteria: ▼, zero padding; □, symmetric padding; ●, smooth padding. In the case of smooth padding the points corresponding to db10, sym4, sym5, sym6 and sym7 were omitted from the plot (Fig. 2a and b) showing SDEP values between 0.15 and 0.28. These cases correspond to regression models where only one coefficient was selected; the following coefficients in order of variance being strongly correlated to this one were discarded. (c) Automatic coefficient selection—Correlation sorting (AR). SDEC values (training set) vs number of cycles for the different padding criteria: ▼, zero padding; □, symmetric padding; ●, smooth padding. (d) Automatic coefficient selection—Correlation sorting (AR) $SDEP_{TEST}$ values (test/monitoring set) vs number of cycles for the different padding criteria: ▼, zero padding; □, symmetric padding; ●, smooth padding.

coefficients were selected under the constrain of not being correlated, and the models are indeed more parsimonious being needed at maximum 9 wavelet coefficients as independent variables. A great benefit may come from this as well in data storage.

Despite the fact that the improvement (fit and prediction) of the regression models after wavelet compression is, in the present case, not particularly remarkable, these models perform better, suggesting that regression in the wavelet domain may be advantageous. Actually, the voltammograms relative to the studied mixtures do not show any significant instrumental noise and exhibit quite a smooth behaviour, and the advantages of wavelet analysis can be better appreciated in more complex matrices, where the signal to noise ratio cannot be enhanced to optimal extent.

Two of the best performing models in Table 2, i.e. the plot of residuals vs. experimental lead and thallium concentrations are reported in Fig. 3a and Fig. 4a and in Fig. 3b and Fig. 4b, respectively. The trends are in general satisfactory. The relative percent errors, for the training set and for the test set, result on average below 5% for the best MLR model (Fig. 3) considering both thallium

Table 2
Standard deviation of error of calculation (SDEC) and of prediction for test set ($SDEP_{TEST}$) and external test set ($SDEP_{EXT}$)

					SDEC	$SDEP_{TEST}$	$SDEP_{EXT}$
PLS considering the whole signal (80 variables)					0.0291	0.0347	0.0503
pad crit	wav	selcrit	lev	ncfs	SDEC	$SDEP_{TEST}$	$SDEP_{EXT}$
	MLR						
zpd	sym7	AR	3	10	**0.0186**	**0.0208**	**0.0350**
	sym5	AV	5	3	0.0268	0.0183	0.0423
	sym7	FR	4	4	0.0245	0.0229	0.0379
	coif2	FV	3	4	0.0315	0.0292	0.0521
sym	coif2	AR	6	4	0.0296	0.0240	0.0496
	sym5	AV	5	3	0.0262	0.0201	0.0454
	coif2	FR	5	4	0.0312	0.0242	0.0638
	sym8	FV	5	4	0.0275	0.0266	0.0633
spd	sym8	AR	3	6	0.0240	0.0274	0.0404
	db20	AV	4	6	0.0239	0.0224	0.0413
	coif5	FR	6	4	0.0275	0.0249	0.0983
	sym6	FV	3	4	0.0300	0.0250	0.0492
	PLS						
zpd	db20	AR	4	9	0.0302	0.0330	0.0440
	sym5	AV	5	3	0.0268	0.0183	0.0423
	sym7	FR	4	4	0.0260	0.0227	0.0365
	coif1	FV	5	4	0.0296	0.0306	0.0483
sym	coif2	AR	6	4	0.0296	0.0240	0.0496
	sym5	AV	5	3	0.0262	0.0201	0.0454
	coif2	FR	5	4	0.0322	0.0267	0.0633
	coif1	FV	5	4	0.0293	0.0305	0.0481
spd	sym8	AR	3	6	0.0267	0.0307	0.0477
	db3	AV	6	4	0.0282	0.0222	0.0406
	coif2	FR	3	4	0.0278	0.0329	0.0609
	coifl	FV	5	4	0.0292	0.0305	0.0480
	NN[a]						
zpd	sym6	AR	3	6	0.0215	0.0256	0.0717
	db3	AV	6	3	0.0237	0.0260	0.0486
	sym7	FR	4	4	0.0221	0.0281	0.0345
	coif5	FV	3	4	0.0263	0.0261	0.0477
sym	coif2	AR	6	4	0.0269	0.0298	0.0356
	sym5	AV	5	3	0.0237	0.0243	0.0436
	coif2	FR	5	4	0.0250	0.0275	0.0822
	sym4	FV	3	4	0.0250	0.0266	0.0438
spd	db20	AR	4	4	0.0272	0.0300	0.0533
	db3	AV	6	4	0.0230	0.0269	0.0465
	db2	FR	3	4	0.0243	0.0331	0.0432
	coif1	FV	5	4	**0.0258**	**0.0239**	**0.0300**

The reported values are: the average of SDEC and SDEP for $[Pb^{2+}]$ and $[TI]^{+}$, respectively; the padding criterion (pad_crit); the wavelet filter (wav); the criteria used in coefficient selection (selcrit): A = automatic, F = fixed; the criteria used in coefficient sorting: R = squared correlation coefficient; V = variance; the optimum decomposition level (lev); the number of selected wavelet coefficients (ncfs).

[a] The network topology is always ncfs-2-2. The transfer functions used in each level are for each row, respectively: lgg; lsg; lsg; lgg; lgg; lsg; lst; lgg; lsg; lgg; lgg; lgg. Where l stands for linear; g for gaussian; s for sigmoid and t for hyperbolic tangent; i.e. lgg: input = linear; hidden = gaussian; output = gaussian.

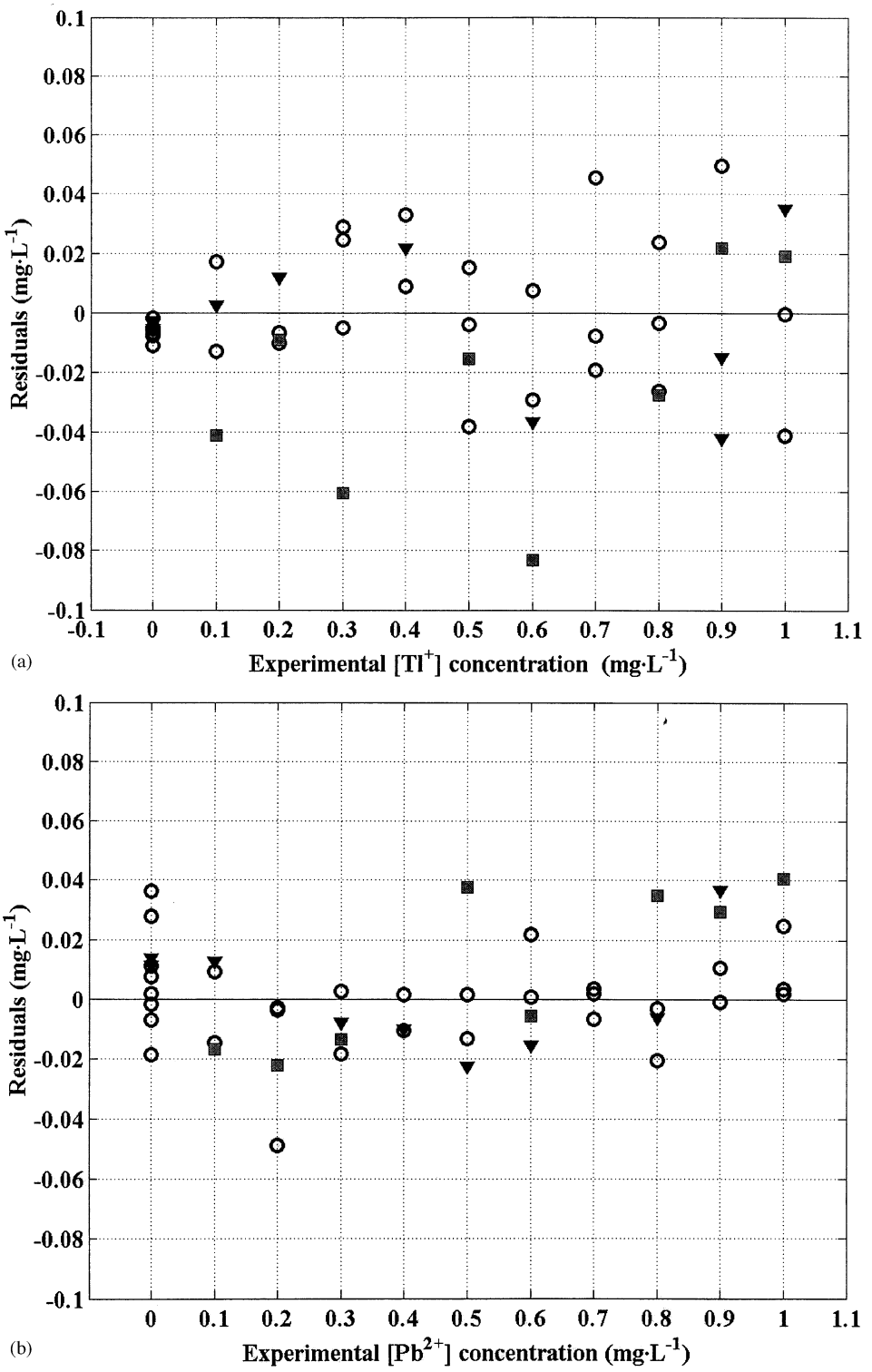


Fig. 3. (a) Residuals vs. experimental thallium concentrations, by using the MLR model calculated on the selected (sym7, automatic selection criterion, correlation sorting) wavelet coefficients: ○, training set; ▼, test/monitoring set; ■, external test set. (b) Residuals vs. experimental lead concentrations, by using the MLR model calculated on the selected (sym7, automatic selection criterion, correlation sorting) wavelet coefficients: ○, training set; ▼, test/monitoring set; ■, external test set.

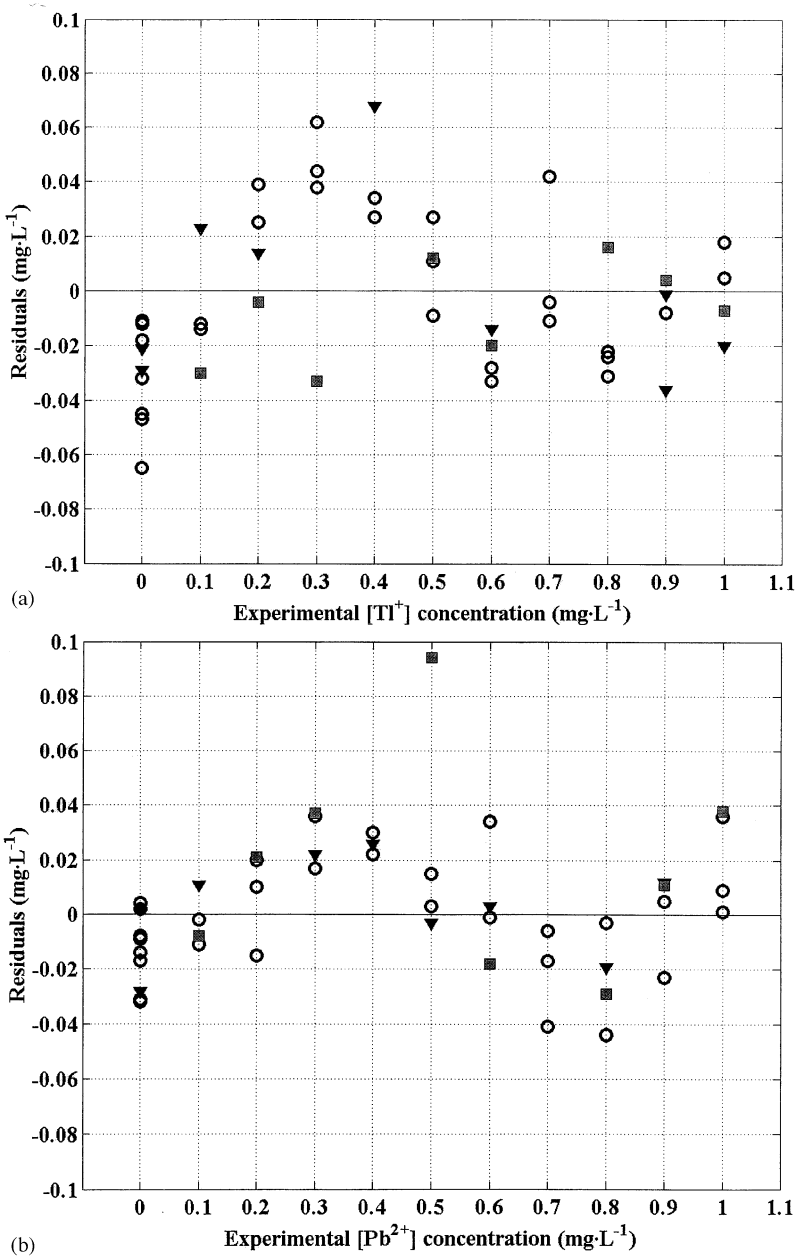


Fig. 4. (a) Residuals vs. experimental thallium concentrations, by using the NN model calculated on the selected (coif1, fixed selection criterion, variance sorting) wavelet coefficients: ○, training set; ▼, test/monitoring set; ■, external test set. (b) Residuals vs. experimental lead concentrations, by using the NN model calculated on the selected (coif1, fixed selection criterion, variance sorting) wavelet coefficients: ○, training set; ▼, test/monitoring set; ■, external test set.

and lead. The best ANN model (Fig. 4) behaves analogously for lead, while the average relative percent error for thallium, is of 7% for the training set and of 8% for the test set. In the case of the external test set, the average relative percent error, for both metals, is as well below 5%, considering

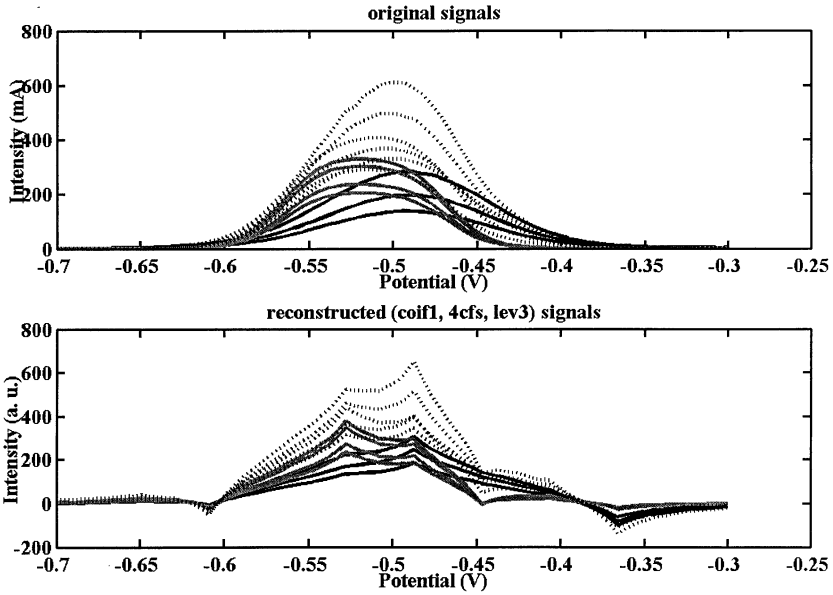


Fig. 5. Few representative original voltamogramms on top and the corresponding reconstructed signals (coif1 selected coefficients by using a fixed number, 4, of coefficients, with the variance sorting criterion) on bottom. Pure lead: solid grey lines; pure thallium: solid black lines; mixtures of the two metals: dotted grey lines.

either the MLR or the ANN model, except for two or three mixtures. In general, the lead content is better predicted than thallium. These errors are comparable with those reported in previously published studies [12,13], although it has to be taken into account that the experimental conditions are different and the thallium and lead peaks are less heavily overlapped in the case reported in the cited references.

In Figs. 5 and 6 a comparison between original and reconstructed signals for the two best performing models are reported. It is interesting to notice that in one case (coif1, Fig. 5) the selected wavelet coefficients highlight the position of the maximum corresponding to lead and thallium peaks, respectively. On the contrary, in the other case (sym7, Fig. 6), the wavelet coefficients focus on the regions where the lead and thallium peaks cross each other: these regions are thus supposed to capture the discontinuities due to the different slope directions of the signal corresponding to the two different peaks.

5. Conclusions

In this work we have shown that FWT can be effectively coupled to predictive feature selection criteria in order to find a minimum number of best performing wavelet coefficients. These coefficients constitute a new set of predictor variables that can be passed to any regression methods. The proposed procedure allowed us to calculate satisfactory multivariate calibration models for both $[TI^{+}]$ and $[Pb^{2+}]$ ions, whose voltammetric responses, were seriously overlapped under the studied experimental conditions. That of overlapped signals constitutes a well known and widely studied problem and many other chemometric approaches have been successfully applied, to thallium and lead mixtures as well; however, there are many advantages that can be envisaged when feature selection is accomplished in wavelet domain. These can be summarised as follows:

(1) Data reduction. Very few wavelet coefficients are able to model the relevant information

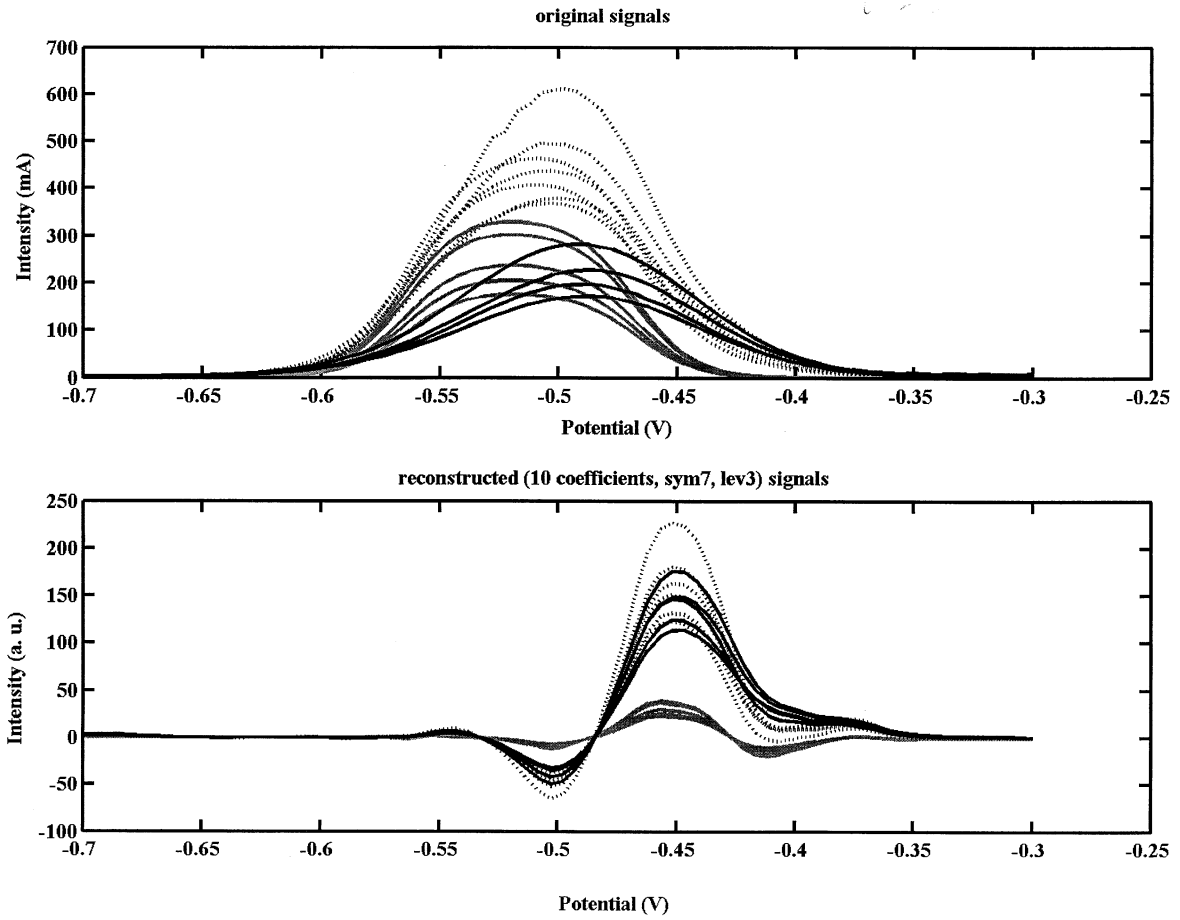


Fig. 6. Few representative original voltamogramms on top and the corresponding reconstructed signals (sym7 selected coefficients by using the automatic selection criterium, 10 coefficients, and the correlation sorting criterion) on bottom. Pure lead: solid grey lines; pure thallium: solid black lines; mixtures of the two metals: dotted grey lines; correlation sorting criterion) on bottom. Pure lead: solid grey lines; pure thallium: solid black lines; mixtures of the two metals: dotted grey lines.

contained in a whole signal. For the data examined by us, 3 or 4 coefficients have been sufficient to obtain predictive regression models. Thus, lowering to a significant extent the ratio between the number of variables and the number of objects, it is possible to use a wider pool of regression techniques in different experimental context;

(2) The possibility of doing simultaneously denoising, background removal, and feature selection;

(3) The selected wavelet coefficients correspond to contiguous regions of the signal; i.e. the order of the variables is implicitly taken into account, which is particularly helpful for interpretative purposes. In fact, once they are reconstructed in the original domain, it is not only possible to localise the spectral regions correlated to the dependent variables, but also to establish at which scale (frequency) the features of interest are located. In other words, the representation in the wavelet domain offers the possibility to use not only the single intensity values of the signal, but also peak widths, slopes of particular regions, degree of smoothness, and many other shape features, in order to predict the dependent variables.

Acknowledgements

Financial support from MURST (Rome) (Ricerche di Interesse Nazionale) and Junta de Andalucia are acknowledged. We also thank Ministerio de Educacion, Cultura y Deportes of Spain for the help given with a research grant.

References

[1] D.P. Binkley, R.E. Dessy, Anal. Chem. 52 (1980) 1335.
[2] T.F. Brown, S.D. Brown, Anal. Chem. 53 (1981) 1410.
[3] C.A. Scolari, S.D. Brown, Anal. Chim. Acta 166 (1985) 253.
[4] B. Raspor, I. Pizeta, M. Branica, Anal. Chim. Acta 285 (1994) 103.
[5] H.N.A. Hassan, M.E.M. Hassouna, I.H.I. Habib, Talanta 46 (1998) 1195.
[6] R. Tauler, A. Smilde, B.R. Kowalski, J. Chemometrics 9 (1995) 31.
[7] M. Esteban, C. Arino, J.M. Diaz-Cruz, M.S. Diaz-Cruz, R. Tauler, Trends Anal. Chem. 19 (2000) 49.
[8] H. Martens, T. Naes, Multivariate Calibration, Wiley, Chichester, 1989.
[9] R.G. Brereton, Analyst 125 (2000) 2125.
[10] P. Geladi, Chemom. Intell. Lab. Syst. 60 (2002) 211.
[11] A. Henrion, R. Henrion, G. Henrion, F. Sholz, Electroanalysis 2 (1990) 309.
[12] M.C. Ortiz, J. Arcos, L. Sarabia, Chemom. Intell. Lab. Syst. 34 (1996) 245.
[13] A. Herrero, M.C. Ortiz, Talanta 46 (1998) 129.
[14] J.M.G. Fraga, A.I.J. Abizanda, F.J. Moreno, J.J.A. Leon, Talanta 46 (1998) 75.
[15] K. Bessant, S. Saini, J. Electroanal. Chem. 489 (2000) 76.
[16] J. Saurina, S.H. Cassou, E. Fabregas, S. Alegret, Anal. Chim. Acta 405 (2000) 153.
[17] R.M. De carvalho, C. Mello, L.T. Kubota, Anal. Chim. Acta 420 (2000) 109.
[18] E. Cukrowska, L. Trnkova, R. Kizek, J. Havel, J. Electroanal. Chem. 503 (2001) 117.
[19] Y. Ni, L. Wang, S. Kokot, Anal. Chim. Acta 439 (2001) 159.
[20] V. Centner, J. Verdu-Andres, B. Walczak, D.J. Rimbaud, F. Despagne, L. Pasti, R Poppi, D.L. Massart, O.E. De Noord, Appl. Spectrosc. 54 (2000) 608.
[21] B. Walczak (Ed.), Wavelets in Chemistry, Elsevier Press, Amsterdam, NL, 2000.
[22] D.J. Rimbaud, B. Walczak, R.J. Poppi, O.E. De Noord, D.L. Massart, Anal. Chem. 69 (1997) 4317.
[23] J. Trygg, S. Wold, Chemom. Intell. Lab. Syst. 42 (1998) 209.
[24] B.K. Alsberg, A.M. Woodward, M.K. Winson, J.J. Rowland, D.B. Kell, Anal. Chim. Acta 368 (1998) 29.
[25] U. Depczynski, K. Jetter, K. Molt, A. Niemoller, Chemom. Intell. Lab. Syst. 47 (1999) 179.
[26] L. Eriksson, J. Trygg, R. Bro, S. Wold, Anal. Chim. Acta 420 (2000) 181.
[27] T. Artursson, A. Hagman, S. Bjork, J. Trygg, S. Wold, S.P. Jacobsson, Appl. Spectrosc. 54 (2000) 1222.
[28] G. Strang, in: G. Strang, T. Nguyen (Eds.), Wavelet and Filterbanks, Wellesey Cambridge Press, Wellesey, MA, 1996.
[29] F.T. Chau, T.M. Shih, J. Gao, C.K. Chan, Appl. Spectrosc. 50 (1996) 339.
[30] J.W. Hayes, D.E. Glover, D.E. Smith, M.W. Overton, Anal. Chem. 45 (1973) 277.
[31] J.M. Palacios-Santander, A. Jimenez-Jimenez, I. Naranjo-Rodriguez, L.M. CubillanaAguilera, J.L. Hidalgo-Hidalgo-de-Cisneros, Mikrochimica Acta, in press.
[32] I. Daubechies, Ten Lectures on Wavelets, SIAM Press, Philadelphia, USA, 1992.
[33] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, Oval Road, London, 1998.
[34] M. Misiti, Y. Misiti, G. Oppenheim, J.M. Poggi, Wavelet Toolbox User's Guide, MathWorks Inc, Natick, MA, 1999.
[35] B.M. Wise, N.B. Gallagher, PLS Toolbox 2.1.1, Eigenvector Research Inc, WA, USA, 1998.
[36] F. Despagne, D.L. Massart, Analyst 123 (1998) 157.
[37] J.W. Kauffman, P.C. Jurs, J. Chem. Inf. Comput. Sci. 41 (2001) 408.