PLS: A versatile tool for industrial process improvement and optimization

Alberto Ferrer[1], Daniel Aguado[2, *, †], Santiago Vidal-Puig[1], José Manuel Prats[1] and Manuel Zarzo[1]

[1]*Department of Applied Statistics, Operations Research and Quality, Technical University of Valencia, Camino de Vera s/n, 46022 Valencia, Spain*

[2]*Department of Hydraulic Engineering and Environment, Technical University of Valencia, Camino de Vera s/n, 46022 Valencia, Spain*

SUMMARY

Modern industrial processes are characterized by acquiring massive amounts of highly collinear data. In this context, partial least-squares (PLS) regression, if wisely used, can become a strategic tool for process improvement and optimization. In this paper we illustrate the versatility of this technique through several real case studies that basically differ in the structure of the **X** matrix (process variables) and **Y** matrix (response parameters). By using the PLS approach, the results show that it is possible to build predictive models (soft sensors) for monitoring the performance of a wastewater treatment plant, to help in the diagnosis of a complex batch polymerization process, to develop an automatic classifier based on image data, or to assist in the empirical model building of a continuous polymerization process.

KEY WORDS: classification; fault diagnosis; monitoring; multivariate image analysis; PLS time series; soft sensor

1. INTRODUCTION

Massive amounts of data are routinely collected from processes in modern highly automated industries. Extracting useful information from these data is essential for making sound decisions

*Correspondence to: Daniel Aguado, Department of Hydraulic Engineering and Environment, Technical University of Valencia, Camino de Vera s/n, 46022 Valencia, Spain.
†E-mail: daaggar@hma.upv.es

Contract/grant sponsor: Spanish Government (MICYT)
Contract/grant sponsor: European Union (RDE funds); contract/grant number: CTM2005-06919-C03-03/TECNO

for process improvement and optimization. Nowadays this is a strategic issue for industrial success in the tremendous competitive global market.

Partial-least squares (PLS) regression [1] is a versatile tool with many desirable properties (i) it is able to cope with highly collinear and low rank data, which is not the case of multiple linear regression [2]; PLS allows analysing data with more variables than observations; (ii) PLS provides models with high stability of predictions because the risk of overfitting is minimized; (iii) PLS is very efficient in handling missing data and therefore it provides inferential models extremely robust to sensor failure and (iv) with the aid of careful data analysis and easy-to-use charts, PLS is able to detect outliers, which improves the quality of the fitted models and reduces the risk of extrapolation when new observations are projected over the model. All this can be obtained with low computational requirements.

As commented by Martens and Naes [3], PLS is a term for multivariate modelling methods derived from Herman Wold's basic concepts of iterative fitting of bilinear models in several blocks of variables. These concepts arose around 1975 as a practical solution to specific data-analytic problems in econometrics and social sciences. The most common implementation in econometrics has been one-factor path modelling of multiblock relationships [4].

PLS is a projection method that models the relationship between a response matrix $\mathbf{Y}$ and a predictor matrix $\mathbf{X}$. Both matrices are decomposed into smaller ones as follows:

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} + \mathbf{E} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}$$

$$\mathbf{Y} = \sum_{a=1}^{A} \mathbf{u}_a \mathbf{c}_a^{\mathrm{T}} + \mathbf{F} = \mathbf{U}\mathbf{C}^{\mathrm{T}} + \mathbf{F}$$

where $\mathbf{T}$ and $\mathbf{U}$ are the score matrices, $\mathbf{P}$ and $\mathbf{C}$ are the loading matrices, and $\mathbf{E}$ and $\mathbf{F}$ are the residual matrices for $\mathbf{X}$ and $\mathbf{Y}$, respectively, for a model with A latent variables determined by cross-validation. The x-scores $\mathbf{t}_a$ are linear combinations of the $\mathbf{X}$ matrix (in the first PLS latent variable) or $\mathbf{X}$-residual matrix ($\mathbf{X}_a$) (in the ath latent variable)

$$\mathbf{t}_a = \mathbf{X}_{a-1}\mathbf{w}_a, \quad \mathbf{X}_a = \mathbf{X}_{a-1} - \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}}$$

$\mathbf{w}_a$ being the weight vector for the ath latent variable.

This is performed in a way to maximize the covariance between $\mathbf{T}$ and $\mathbf{U}$, both related by the inner relationship

$$\mathbf{U} = \mathbf{T}\mathbf{B} + \mathbf{H}$$

where $\mathbf{B}$ is a diagonal matrix and $\mathbf{H}$ is a residual matrix. This allows PLS to be expressed as a predictive model

$$\mathbf{Y} = \mathbf{T}\mathbf{B}\mathbf{C}^{\mathrm{T}} + \mathbf{F}^* = \mathbf{X}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{B}\mathbf{C}^{\mathrm{T}} + \mathbf{F}^*$$

where $\mathbf{F}^*$ is a residual matrix.

PLS models can be fitted using all the responses simultaneously, as shown before (PLS2), or by building one model for each single response (PLS1). In the PLS1 model, the latent subspace is found by maximizing the covariance between the x-scores $\mathbf{t}_a$ and one response $\mathbf{y}$ (column vector of the response matrix $\mathbf{Y}$). In general, the more correlated the response variables are, the better the performance of PLS2 with respect to PLS1 in terms of predictive accuracy and interpretation.

Different algorithms have been proposed for PLS models. The key idea behind them is to replace the maximum likelihood principles (statistically well sounded but suffering from lack of applicability in data-rich environments due to ill-conditioning or singularity problems) by sequential algorithmic approaches, based on a series of local least-squares fits. For a more detailed explanation of the different algorithms used and the mathematical and statistical structures of PLS, see, for example, [3, 5, 6].

PLS can be considered as a prediction/regression method useful for near collinear data. There exist a large number of regression methods that have been proposed in the literature for the same purpose. Some of them treat each response separately (e.g. principal component regression or ridge regression), while others combine multiple responses by taking advantage of the structure of matrix **Y** (e.g. reduced rank regression or shrunken canonical correlation models, C&W-GCV). Several authors have performed comparative studies between all of these competitors via simulation studies using prediction errors as performance index [7, 8]. The main findings from these simulation studies are (i) they often give fairly similar results and (ii) the results are extremely dependent on the simulations done and, thus, they may not be representative for real data encountered in particular fields. In the discussion of [8], some authors have criticized these simulation studies because they do not consider other criteria for evaluating the success of a model, such as parsimony, bias, interpretability and diagnosis for outliers, and other data inhomogeneities.

Nevertheless, apart from predictive ability, one of the most appreciated properties of PLS models from a practitioner's point of view is *model reduction*. This PLS property has changed the statistical paradigm of *variable selection* to the practitioner paradigm of *variable compression*. This ability to discover latent structures seems to function fairly well in practice (especially in data-rich environments) because this uncovers hidden information and improves model understanding.

PLS was further developed in the field of chemometrics, mainly dealing with problems in handling spectroscopic data. These techniques produce hundreds of variables (the light absorbance at different wavelengths) for a set of calibration samples. In this context, PLS provides good predictive models by finding those latent variables that present a good correlation with the response variables and at the same time explain a fair amount of the spectral variance. However, the results are poor if the **X** matrix is characterized by dominant latent components that explain a high proportion of the data variance, but are orthogonal to the response variables. Attempting to overcome this limitation, different methods of orthogonal signal correction have been proposed [9]. In the context of multivariate statistical process control, the **X** matrix often presents a multiblock structure (e.g. a block of variables measuring temperature, another block for pressure, a few variables providing information about initial conditions, etc.). In these situations, the traditional PLS method is sometimes unable to produce good predictive models, and different multiblock algorithms have been developed [10].

Most successful PLS applications have been reported in chemistry and related disciplines. But this tool is now expanding to many other areas such as business, finance and marketing. A handbook of PLS recently published [11] illustrates the application of this tool in studies of business performance, brand preference, employee behaviour, customer satisfaction and loyalty. Other PLS applications to improve business strategy can be found in [12–17].

When analysing the large number of variables collected nowadays from modern industrial processes, it is usual to find that these variables have different numerical ranges. The numerical range of a given variable influences its variance (i.e. a large numerical range implies a large variance) and PLS is variance dependent. Therefore, scaling the raw data is usually required to make it more suitable for the analysis. Note that the pre-processing can make the difference between

a useful model and no model at all [18]. In the applications that will be presented throughout this paper, the raw data were mean centred and scaled to unit variance. In this manner, different measurement units of the collected variables can be handled, thus, giving equal importance *a priori* to each variable.

By properly arranging the predictor and the response data structure, PLS becomes a strategic tool able to adapt to very different scenarios. This is illustrated through several real industrial case studies in the following sections.

2. PLS: SOFT SENSOR

Nowadays, a large number of process variables can be collected at modern wastewater treatment plants (WWTPs). On the one hand, these variables can be measured online by means of inexpensive, robust and low-maintenance sensors, but they do not directly provide information on process performance. On the other hand, the process output quality variables that clearly reflect the WWTP performance are usually measured less frequently in a quality control laboratory. There are special probes that allow an online measurement of some key quality variables, but they are not usually employed in small WWTPs because they involve high investments and require important maintenance costs.

Analysing samples in a quality control laboratory has several drawbacks: the analyses are expensive, slow and do not allow an early detection of problems that might appear in the process. Thus, there is a strong interest in taking advantage of the information contained in the process variables to build empirical predictive models (soft sensors) for monitoring the performance of a WWTP. In this manner, chemical analyses could be replaced or at least reduced. This is an appealing use of PLS, which has been successfully applied in many different contexts [19–25]. This wide variety of examples shows that considerable effort has been placed on applying this multivariate versatile tool for making the most of the operational data available.

In this case study, data from a sequencing batch reactor (SBR) operated for enhanced biological phosphorus removal (EBPR) from wastewater have been analysed to develop a PLS soft sensor. It is a batch process that consists of three main stages per cycle taking place in the same reactor. During the first stage, which lasts 1.5 h, the reactor is kept in anaerobic conditions and phosphorus is released by polyphosphate-accumulating organisms (PAOs). Afterwards, the reactor is aerated for 3 h allowing PAOs to uptake phosphorus and store it intracellularly as polyphosphate. As the phosphate uptake is higher than the release, a net phosphate uptake is achieved. Finally, the activated sludge is settled for 1.5 h, thus, producing a clarified effluent of treated wastewater. In the SBR, five process variables are registered online by means of inexpensive sensors: electric conductivity (Cond), redox potential (ORP), dissolved oxygen concentration (DO), pH and temperature (Temp). The SBR was operated under constant conditions until steady state was reached. Then it was extensively sampled to characterize the EBPR performance. For this purpose, samples were withdrawn at regular time intervals (every 15 min) and analysed in the quality control laboratory for several pollutants. In this particular investigation, the trajectories (i.e. time evolution) of phosphorus, potassium and magnesium were used as response variables.

Data from the batch process were arranged in two three-way matrices: $\underline{\mathbf{X}}$ (20 batches $\times$ 5 process variables $\times$ 340 time points) and $\underline{\mathbf{Y}}$ (20 batches $\times$ 3 quality responses $\times$ 18 time points). These matrices were unfolded batchwise, resulting in a two-way matrix $\mathbf{X}$ (20×1700) and a two-way matrix $\mathbf{Y}$ (20×54) that were analysed applying standard bilinear PLS.

The available data set consisted of 20 batches: 15 were used for model fitting and the remaining for validation. Based on a preliminary study, it was decided to transform the original variables using the difference (Δ) between each variable value and its value at the beginning of the batch.

A PLS-2 model was built using the transformed trajectories of all variables (process and quality). Analysing the weights of the PLS model, it was found that the trajectories with more predictive contribution were ΔpH and ΔCond. Moreover, the weights of the three quality trajectories were similar, thus, indicating a highly positive correlation among them. This result matches the hypothesis of electroneutrality postulated by Comeau *et al.* [26]: in the EBPR process, both potassium and magnesium are released and taken up simultaneously with phosphorus and act as counter ions to maintain electroneutrality inside the bacterial cell.

In order to obtain a more parsimonious inferential model, a new PLS-2 model was formulated using only ΔpH and ΔCond as explicative trajectories. The good performance of the fitted PLS model to predict the trajectory of the phosphorus concentration is shown in Figure 1. Similar results were obtained for potassium and magnesium (not shown).

It should be noted that the developed predictive models take into consideration the auto- and cross-correlation of the explicative variables (i.e. incorporate process dynamics) during the entire batch because of the unfolding method used [27].

Despite the good prediction ability of the model, it could be argued that an important drawback is that it requires the entire trajectories of the explicative variables to be known, that is, the predictions are done when the batch has been completed. However, this inconvenience can be overcome and, therefore, the model applied online, by estimating the future observations via the PLS model (i.e. by using missing data imputation methods) as shown in [20]. As pointed out by these authors, the stability property of the estimated scores becomes especially important when developing a predictive model, because if these estimations are smooth and nearly constant even at the beginning of the batch, relatively accurate predictions for the response variable will be obtained from the beginning of the batch. As an example, Figure 2 displays the predicted values from the model at different time points of a validation batch together with the experimental values. The scores were estimated using the trimmed score regression method proposed by Arteaga and Ferrer [28].

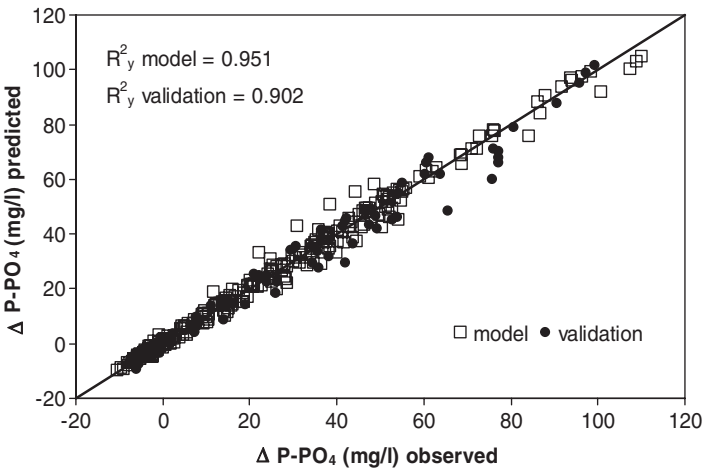

Figure 1. Observed *versus* predicted phosphorus concentration by the PLS model.

DOI: 10.1002/asmb

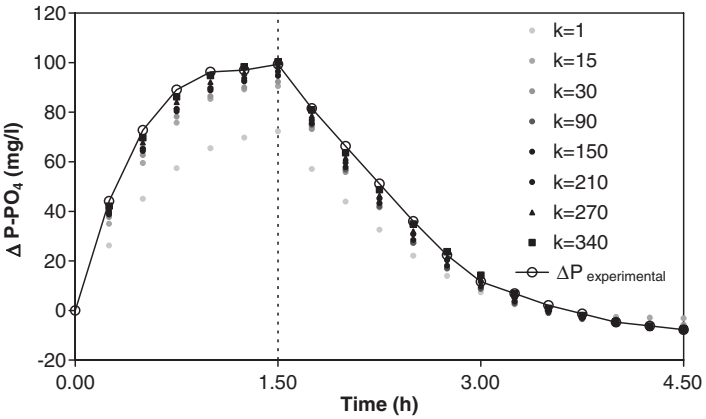

Figure 2. Experimental and online predicted values of transformed trajectory of phosphorus concentration (ΔP) in a validation batch. At each time point, the future unknown part of the batch was estimated using missing data imputation methods [28].

These results successfully illustrate that it is possible to build an efficient soft sensor for monitoring the performance of the SBR. As only data from the inexpensive and low-maintenance sensors installed in the SBR are used as input, the soft sensor can be considered as a cost-effective tool. Moreover, monitoring the residuals of this model can be useful in detecting outliers and assessing whether the model needs to be updated. These issues are of special relevance in the area of PLS research and practice as highlighted in different contributions [29–32].

3. BLOCKWISE PLS: BATCH PROCESS DIAGNOSIS

Batch chemical processes are difficult to control. Quality parameters are determined by analytical methods once the process finishes, and some of them are often out of quality specification limits. As a consequence, the product has to undergo further processing or standardization steps to achieve a proper commercial quality, or maybe has to be sold under a lower-quality category, resulting in a considerable economic impact. To avoid these problems, a common approach is to reduce as much as possible the deviations of process variables from the target trajectories. However, in industrial conditions this is complicated and not always effective, as quite often process engineers do not know which are the key critical points of the process that require a more accurate control to reduce the variability of quality parameters around the nominal value to avoid batches out of specifications.

Data from a petrochemical batch polymerization process have been analysed to diagnose the causes of variability of one of the final quality parameters: the hydroxyl index (I_{OH}), also referred to in the literature as *hydroxyl number*. This process takes place in four batch stages and there are 52 electronic sensors that record online different kinds of information such as temperature, pressure, flow, pH, etc. These data are used online for the engineering process control and are routinely stored in databases (one datum recorded every minute from each sensor). Once the batch finishes, the hydroxyl index of the polymer (polypropylene oxide) is determined in the laboratory,

and the problem is that about 15% of the batches produced are out of quality specifications. To provide a solution, data from 69 historical batches have been analysed in order to identify the critical points of the process.

In industrial conditions, most stages have a different duration from batch to batch, which results in serious problems for most statistical data analysis methods. To overcome this limitation, different alignment techniques can be used to synchronize the trajectories. In this case, we applied the indicator variable approach [33]. With this methodology, data are arranged in a three-way structure: batches × process variables × time. This matrix has been unfolded according to the methodology proposed by Nomikos and MacGregor [34], obtaining a large matrix with 69 batches and 9100 variables. This matrix is structured in 52 blocks, each containing the data registered by one sensor along the development of the batch. For each batch, the evolution *versus* time of one process variable is often called a trajectory. Data have been mean centred to get rid of the main non-linearities as well as scaled to unit variance to give the same *a priori* weight to all variables. The target is to identify those variables that in a certain part of the process are correlated with the hydroxyl index. Considering it as the response variable, a PLS-1 regression has been applied to this matrix, resulting in a model with one relevant component with a goodness of prediction by cross-validation $Q^2 = 0.33$. This value is obviously not high enough to allow a reasonable prediction of the hydroxyl index for a new batch. In fact, working with observational data (no experimental design has been run) it is usual to obtain predictive models with low goodness of prediction Q^2. Nevertheless, this value is significantly greater than zero according to cross-validation, suggesting that those variables with highest contribution in the first component might lead to the identification of critical points of the process.

If a PLS-1 regression is conducted with centred variables and scaled to unit variance, the weights of the first latent vector are proportional to the linear correlation coefficient between the process variables and the quality parameter [35]. Hence, high loadings in absolute value correspond to process variables that in certain time points are significantly correlated with the hydroxyl index. However, the diagnosis becomes difficult from the analysis of the weight plot (Figure 3), as the weights are scattered and there are no outstanding trajectories especially with high values.

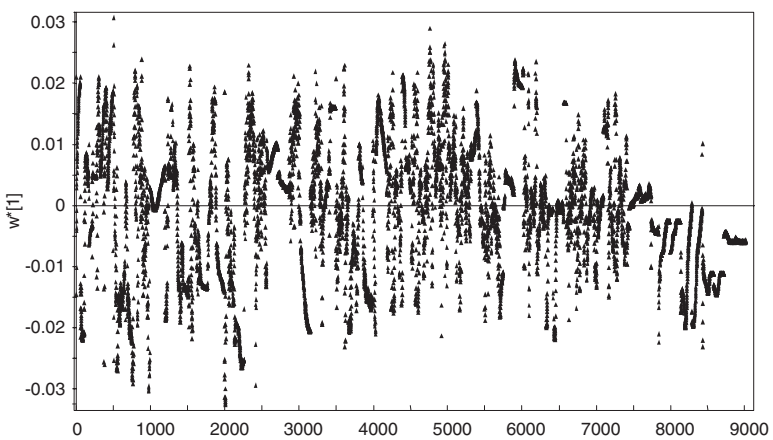

Figure 3. Weight plot of the first PLS component fitted with all process trajectories. Positive weights correspond to positive correlation with the hydroxyl index (response variable).

As the diagnosis is not clear using the standard methodology of batch processes [36, 37] other alternatives were tried. Blockwise PLS is a new approach for process diagnosis, following the idea proposed in a previous paper [35]. Aguado *et al.* [38] successfully applied this methodology for process understanding of a wastewater batch reactor. Starting from the unfolded matrix, a PLS regression is conducted for every block of variables and, for the first component, the associated latent variable and its Q^2 are calculated. This procedure produces a new matrix of 52 latent variables that contains the main information in order to predict the I_{OH}. A further analysis of this matrix can be carried out to identify stages more correlated with the final quality parameters, outliers, shifts in the process, etc. However, in this case just the analysis of the 52 Q^2 values highlights the key information.

If the 52 values of Q^2 are charted on a normal probability plot, a linear trend is observed (Figure 4), but the highest seven values are slightly separated from the straight line, highlighting those trajectories most important from a statistical point of view. This result simplifies the diagnosis. The software SIMCA-P used for the analysis considers as threshold of significance for Q^2 a value of about 0.1, but in this case there are 15 values higher than 0.1 that follow the straight line in the normal probability plot and do not seem to provide relevant information. The pressure during the second stage (2PR), the temperature during the first stage ($1T^a$), the derivative trajectory of this temperature ($1T^ad$) and the derivative trajectory of pressure (1PRd) are the latent variables with highest Q^2 values.

Regarding the latent variable of pressure (2PR), the highest weights correspond to the beginning of the second stage. During a period of about 20 min, the pressure of the batches with highest hydroxyl index (out of specifications) describes a trajectory that tends to be lower than the mean trajectory. The opposite occurs for batches with lowest I_{OH}. Thus, during the beginning of the second stage, the pressure has a negative correlation with I_{OH}. The pressure just at the third minute of this stage is significantly correlated with the hydroxyl index (Figure 5), and similar results are obtained considering other time points. Actually, if a further PLS is conducted with the trajectory

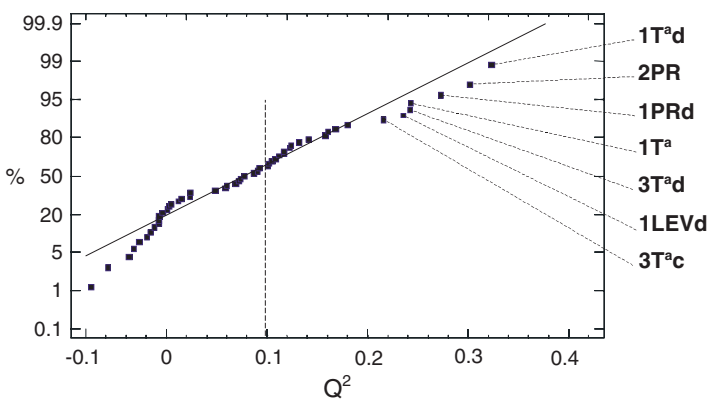

Figure 4. Normal probability plot of 52 Q^2 values. Each one corresponds to the first component of a PLS model fitted with the block of recorded variables from one of the 52 sensors installed in the process, and considering $y = I_{OH}$. Variable codes are shown for the highest Q^2 values that depart from the straight line. The first number corresponds to the stage. $1T^ad$, 1PRd, 1LEVd are derivative trajectories. The vertical dashed line is the significance limit considered by the software SIMCA-P.

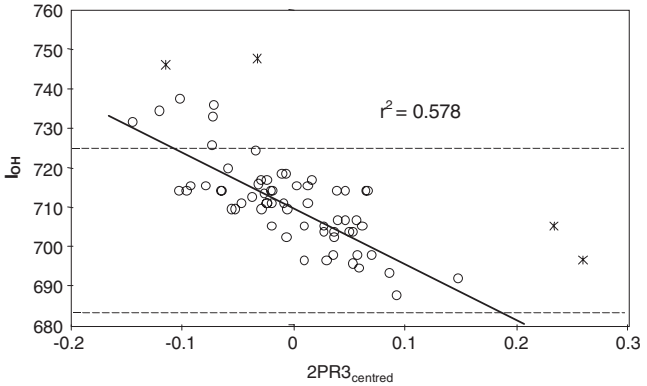

Figure 5. Scatterplot of the hydroxyl index *versus* the pressure (centred values) in the third minute of the second stage (2PR3) for all the 69 batches. The regression line and squared correlation coefficient have been calculated discarding the four data with abnormal residuals (shown in asterisks). Horizontal dashed lines define the tolerance quality interval.

of temperature during this period, it results in $Q^2 = 0.6$ and hence this multivariate model could be used for online monitoring. Furthermore, these results reveal information regarding the diagnosis: as the correlation appears from the first minutes of the second stage, it seems that the first stage is critical.

With respect to the temperature during the first stage, the highest weights of this latent variable ($1T^a$) correspond to the final period of the addition of polyalcohol and during the addition of alkaline solution. Actually, if the original trajectories are checked out, during these periods the batches with highest I_{OH} show a trajectory of temperature lower than average. Hence, the temperature in this stage is correlated with I_{OH}, as it is the pressure in the second stage.

It is important to point out that caution has to be taken when interpreting these correlations. Only if correlation is due to a cause–effect relationship, we can consider those variables as critical points or key process variables whose variability should be reduced to produce a reduction in the variability of the I_{OH}.

In this case, further studies conducted have pointed out that the flowmeter that controls the addition of polyalcohol in the first stage seems to be the critical point, and the hypothesis is that an excessive variability of the mass of polyalcohol used as a reagent seems to be the main cause of variability of I_{OH}. Hence, the advice is to improve the control of this flowmeter in order to achieve a more accurate measure of the mass really added to the tank. From a statistical point of view, the only way to identify a causal correlation and verify this hypothesis is conducting a design of experiments (DOE). Blockwise PLS can be an efficient methodology to select the likely key process variables to run a DOE out of the large amount of original process variables.

4. PLS-DISCRIMINANT ANALYSIS: AUTOMATIC CLASSIFIER BASED ON IMAGE DATA

The classification of parts or objects according to shape, colour or size is common in industrial environments. In the case of agricultural products such as fruits, there is a strong interest

in developing automatic systems to classify fruit pieces based on the presence of defects (i.e. phytopathologies and physiopathies that may affect the fruit).

Our example is related to the classification of orange fruit into quality categories based on the separation of sound pieces from those affected by different defects. In the citric industry, there exist some machines that classify oranges basically in terms of colour and size. These machines do not achieve very good results when trying to discriminate between phytopathologies and physiopathies, probably because colour-based inspection machines try to discriminate between different problems dealing just with the spectral values of the pixels, but not with the spatial structure of the orange peel. Despite the big efforts made to improve the machine classification results during the last years, still an important percentage of the fruit introduced in the agricultural cooperatives is inspected and classified by the human eye, mainly when there is a need to distinguish between some diseases. However, human inspection is not feasible due to its low productivity and lack of reliability for the classification.

Traditional image analysis classification techniques are based on the extraction of colour or texture statistics from the images, turning them into just one vector of characteristics that is used to compare new images with those used to build the models [39]. This is equivalent to consider the image as a sample from which some variables are measured. The introduction of PLS in the multivariate image analysis (MIA) field started with the works of Prof. Esbensen [40, 41]. Outstanding recognition deserves the work of Prof. MacGregor [42, 43]. Eriksson *et al*. [44] have also applied a discriminant version of PLS in a recent work.

Here we propose to integrate both spectral and textural information into the same data structure related to different types of orange images available (sound or affected by different diseases), and to analyse it using PLS-discriminant analysis (PLS-DA), in a similar manner as proposed by Prats-Montalbán and Ferrer [45]. We will consider the image as a sample of pixels instead of a statistics vector worked out from the image to characterize it. Hence, we will be able to deal in a better way with the variability present in the pixels. This is achieved by the application of the MIA strategy [46] on RGB images coming from orange fruits, incorporating for each colour channel the spatial information (texture) following the methodology proposed by Bharati and MacGregor [47].

This is a four-step procedure: first of all, the spatial information is extracted through spatially shifting, for each colour channel, the RGB image in adjacent directions, and then stacking the shifted images on top of each other to form a four-way pixel array, as displayed in Figure 6(a). The amount of neighbouring pixel intensities to save depends on the texture to analyse. In this manner, we obtain a four-way data structure, with the first two dimensions ($n_1 \times n_2$) linked to the pixels that spatially form the image, the third linked to the J pixels constituting the neighbouring window, and the fourth related to the $K = 3$ colour channels (RGB) of the images. Once the shifting process has been applied on each of the R, G and B colour channels, in order to analyse the pixels it is necessary to unfold the image according to the MIA strategy, i.e. locating the pixels in the sample mode by means of stacking each column of the images one below the other, until one vector is obtained for each of the J shifted images analysed, for the three R, G and B channels. In this manner, pixels become the samples in the three-way internal structure displayed in Figure 6(b). However, in order to apply a PLS model, we have to carry out a second unfolding of this three-way data structure, as displayed in Figure 6(c), obtaining a two-way data structure with the first mode linked to the $n_1 \times n_2$ pixels constituting the samples, and with the second mode integrating both the spatial and colour channels (variables).

As we are going to apply a PLS-DA approach, it is necessary to repeat this procedure for each one of the images belonging to different types of oranges. Once all the two-way data matrices are

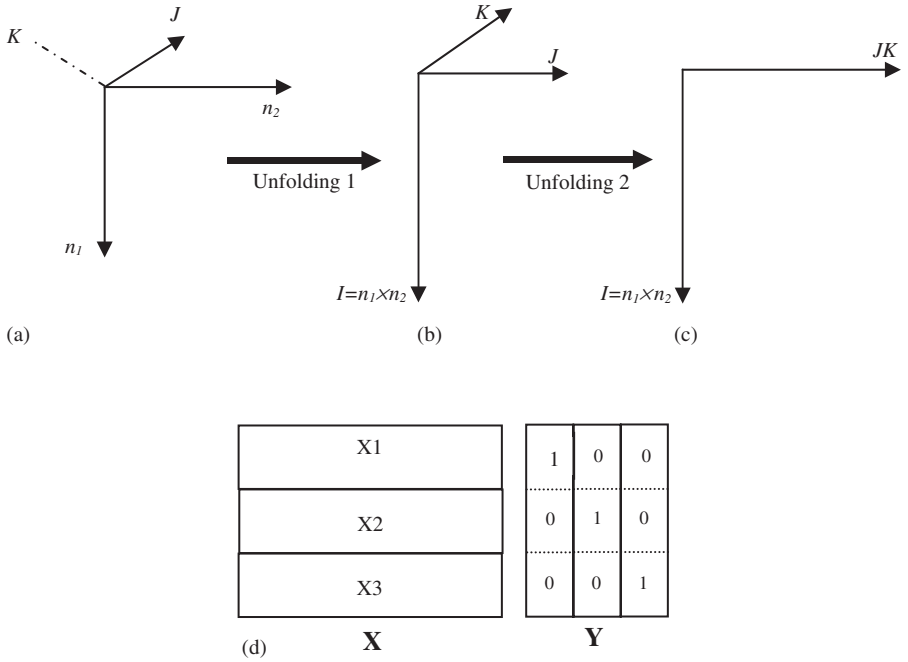

Figure 6. Scheme of the data structure building procedure for the spectro-textural multivariate image: n_1 and n_2 define the spatial dimensions of the image, which after first unfolding turn into the I sample dimension; $K = 3$ (R, G and B spectral channels); and J is the number of neighbouring pixels saved to include the spatial information (texture). (a) Four-way dimensional data structure; (b) three-way internal data structure; (c) two-way data structure; and (d) stacking process for a three-class PLS-DA model and creation of the dummy variables.

obtained for each of the orange images to analyse, by stacking them on top of each other, the final two-way data matrix **X** is created. The final step is to create as many dummy variables as classes of oranges that are to be fitted by the PLS-DA model. This step is accomplished by creating a **Y** matrix that contains as many columns as different classes of oranges. For each of these columns, the value 1 is assigned to the pixels belonging to the class linked to that column, and 0 for the rest, as shown in Figure 6(d). Usually, one training image per defect is enough for calibrating the classification model.

Once the data structures have been created, it is possible to build a PLS model. The number of relevant components should be determined by means of a cross-validation procedure. Finally, new RGB orange images can be classified by projecting them onto the fitted PLS-DA model, working out the percentage of pixels over the residual sum of squares (RSS) limit for the fitted model. If this percentage of pixels of the new projected image is higher than the percentage of pixels over the RSS limit for the training image data set, then the image is classified as not belonging to any of the modelled classes. On the contrary, if this percentage is lower than the RSS limit, then the new image is classified as belonging to the class showing an average predicted value for the pixels closest to '1'.

The application of this procedure for the classification of three types of diseases on a validation image data set with 67 oranges with several types of diseases gave a 78% success rate in the

classification. As these results were not completely satisfactory, it was decided to project only the images corresponding to oranges of the same classes as the ones used to build the models. In this manner, it is not necessary to establish a maximum percentage of pixels over the limit for the RSS statistic. When considering this alternative, the PLS-DA approach reached up to a 93% success rate in the classification, which is a very good result, always taking into account the limitations of this alternative.

5. PLS-TIME SERIES: MODEL BUILDING

The objective of this case study is to present the capabilities of PLS time series (PLS-TS) methodology [18] for the estimate of a transfer function (TF) model of an industrial polymerization process between two input variables (X): reactor temperature (T) and ethylene flow (E); and two output variables (Y): a quality property of the polymer, *Melt Index* (MI), and a measure of the process throughput (APRE). The model can be used for designing predictive controllers. Real data from three manufacturing periods provided by a petrochemical company have been investigated.

As the two output variables were slightly correlated, we proceeded to build a TF model for each output variable separately. From the different TF models that can be used to represent the process, the finite impulse response (FIR) model was chosen. This is a simple but non-parsimonious TF model, where each output variable at time t, $Y_{j,t}$ $(j=1,2)$, is expressed as a linear combination of values at time t and past values of the input variables $X_{i,t-k}$ $(i=1,2,\ k=0,\ldots,L)$:

$$Y_{j,t}=\sum_{i=1}^{2}(\beta_{i0}X_{i,t}+\beta_{i1}X_{i,t-1}+\beta_{i2}X_{i,t-2}+\cdots+\beta_{iL}X_{i,t-L_i})+\varepsilon_{j,t}$$

where L_i is the number of lags for input variable X_i related to the inertial properties of the system; the residual part, $\varepsilon_{j,t}$, is assumed to be white noise; and $Y_{j,t}$ and $X_{i,t-k}$ measure the deviations from steady state. Other TF models such as, e.g. autoregressive with exogenous variables (ARX) models, which incorporate past measured outputs $Y_{j,t-k}$ as inputs, can also be fitted by PLS-TS [18, 48].

The phases of model building by using PLS-TS methodology are:

- *Initial exploratory analysis of data*: Study of the nature of the series and their dynamics (auto- and cross-correlation functions); determination of number of differences to obtain stationary series and number of lags to consider in the formulation of the model; process fault detection (residual and score plots from preliminary PLS models).
- *Pre-treatment of data*: The original matrix of input variables **X** is expanded with new lagged variables for every input according to Figure 7. The number of lags to take into account depends on the results of the initial exploratory analysis about the dynamic behaviour of the series. Variables are centred and scaled to unit variance.
- *TF model identification*: A PLS-TS model is estimated separately for every output variable and expressed as a FIR model by using the regression PLS coefficients plots. As every variable is scaled by its standard deviation, the coefficients β will approximately determine the importance of the variables in the model. To find a parsimonious model, we proceed to select the most influential lags by picking up the variables with greatest value of β that, at the same time, are consistent in the three manufacturing periods under study. Figure 8 shows the PLS regression coefficients plot for the differenced model for predicting MI.

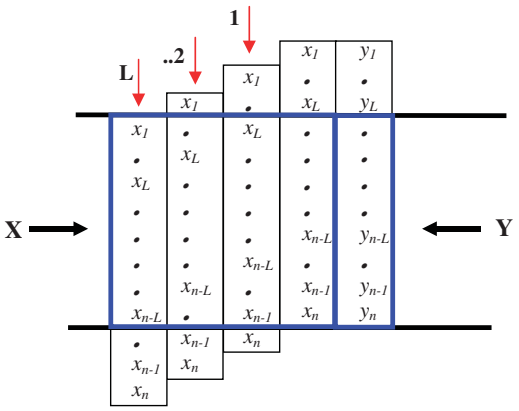

Figure 7. Overview of the lagging process to create the expanded **X** matrix. Dropping rows from the top and bottom is carried out to regularize the data structure.

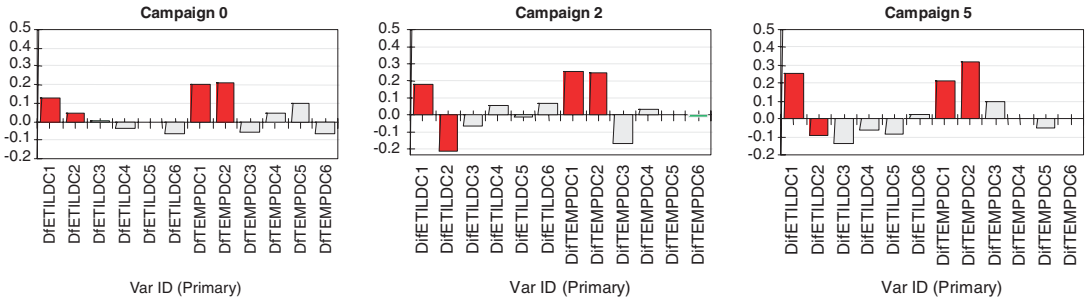

Figure 8. PLS coefficient plot of reactor temperature and ethylene flow (differenced data) as a function of lag number.

- In this case the first two lags for both input variables (DIFETILDC1, DIFETILDC2 and DIFTEMPDC1, DIFTEMPDC2) were included in the model for MI. Following a similar procedure (not shown), the first two lags for reactor temperature and only the first lag for ethylene flow were considered for the APRE model (differenced data).
- If the model contains a high number of highly correlated input variables, a more parsimonious model can be obtained by pruning the original one. The weight plot of the PLS-TS model will help to identify the variables that provide redundant information. Owing to the small number of input variables in our study, pruning was not considered.
- *Final model estimation*: In this case, the estimated model was

$$\nabla \mathrm{MI}_t = \beta_1 \nabla T_{t-1} + \beta_2 \nabla T_{t-2} + \beta_3 \nabla E_{t-1} + \beta_4 \nabla E_{t-2} + a_{1t}$$

$$\nabla \mathrm{APRE}_t = \beta_1 \nabla T_{t-1} + \beta_2 \nabla T_{t-2} + \beta_3 \nabla E_{t-1} + a_{2t}$$

where ∇ is the differential operator ($\nabla X_t = X_t - X_{t-1}$).

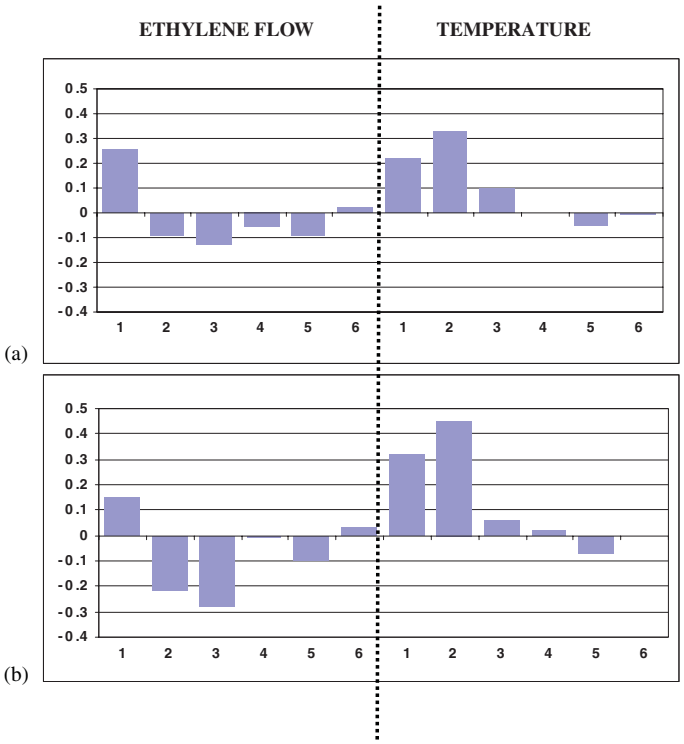

Figure 9. (a) PLS coefficient plot for ∇MI model and (b) cross-correlation functions between ∇MI and ∇E, and between ∇MI and ∇T. Campaign 5.

$$Y_{j,t} = \sum_{i=1}^{2}\left(\beta_{i0}X_{i,t} + \beta_{i1}X_{i,t-1}\cdots + \beta_{iL}X_{i,t-L_i}\right) + \varepsilon_{j,t}$$

$$Y_{j,t} = \sum_{i=1}^{2}\frac{\omega(B)}{\delta(B)}B^{b}X_{i,t} + \boxed{\frac{\theta(B)}{\varphi(B)}a_t}$$

Figure 10. FIR model (up) *versus* BJ model (down). Noise $\varepsilon_{j,t}$ is partially modelled in Box–Jenkins methodology.

From this case study, we concluded that:

- PLS-TS can be successfully applied to estimate process dynamics using a variety of TF models (FIR, ARX, etc.). The results were consistent with those obtained from using the statistically well-sounded Box–Jenkins (BJ) methodology [49, 50].
- PLS-TS method provides a pack of very useful graphic tools for the descriptive study of the data, allowing an easy identification of abnormal periods in the data set. The regression PLS

coefficients plot helps to identify the TF model. This plot is similar to the cross-correlation function when the input variables are not correlated. For instance, Figure 9 shows how the PLS coefficient plot for ∇MI model is similar to the cross-correlation functions between ∇MI and ∇E, as well as between ∇MI and ∇T in campaign 5.

- Comparing the BJ methodology to FIR and ARX models fitted by PLS-TS, the former leads to more parsimonious models because noise $\varepsilon_{j,t}$ can also be modelled (see Figure 10). Nonetheless, PLS-TS may serve as an exploratory tool complementary to the BJ methodology.
- PLS-TS turns out to be a good choice for TF model building in the case of complex multi-input multi-output systems, with inputs and outputs highly correlated, where BJ methodology becomes practically unfeasible.

6. CONCLUSIONS

Specialized tools for specific problems are easily available in the literature of statistical data analysis. Nevertheless, very few techniques are versatile enough to be applied in a very wide range of problems (e.g. discrimination, classification, process modelling, process diagnosis and fault detection) dealing with so many different data structure scenarios (e.g. collinearity, rank deficiency, missing data, etc.), which are typical in complex modern processes.

This is the great virtue of PLS models: the same tool can be used in very different contexts by properly arranging the data structure. This makes PLS an easily transferable tool: a key property to be successfully applied for problem solving in highly competitive and time-demanding environments.

Although the examples illustrated here come from chemometrics environments, where PLS has been widely used in the last years, there is a tremendous potential impact of PLS in other areas such as marketing, business strategy and the application of statistics in business analysis.

ACKNOWLEDGEMENTS

This research was partially supported by the Spanish Government (MICYT) and the European Union (RDE funds) under grant CTM2005-06919-C03-03/TECNO. The authors would like to thank E. Moltó (Instituto Valenciano de Investigaciones Agrarias-IVIA, Spain) for the orange images data set. The authors are also grateful to CALAGUA research group (University of Valencia and Technical University of Valencia, Spain) for providing them with the SBR data set. Finally, the authors also wish to acknowledge the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

1. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. *Analytica Chimica Acta* 1986; **185**:1–17.
2. Wold S, Wold H, Dunn WJ, Ruhe A. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing* 1984; **5**:735–743.
3. Martens H, Naes T. *Multivariate Calibration*. Wiley: New York, 2001.
4. Wold H. Soft modelling: the basic design and some extensions. In *Systems under Indirect Observation*, *Causality–Structure–Prediction*, Jöreskog KG, Wold H (eds). North-Holland: Amsterdam, 1981.
5. Helland IS. On the structure of partial least squares. *Communication in Statistics—Simulation and Computation* 1988; **17**(2):581–607.
6. Höskuldsson A. PLS regression methods. *Journal of Chemometrics* 1988; **2**:211–228.
7. Frank IE, Friedman JH. A statistical view of some chemometrics regression tools (with Discussion). *Technometrics* 1993; **35**:109–148.

8. Breiman L, Friedman JH. Predicting multivariate responses in multiple linear regression (with Discussion). *Journal of the Royal Statistical Society*, *Series B* 1997; **59**(1):3–54.
9. Svensson O, Kourti T, MacGregor JF. An investigation of orthogonal signal correction algorithms and their characteristics. *Journal of Chemometrics* 2002; **16**:176–188.
10. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics* 1998; **12**:301–321.
11. Esposito-Vinzi V, Chin WW, Henseler J, Wang H (eds). *Handbook of Partial Least Squares. Concepts*, *Methods and Applications in Marketing and Related Fields*. Springer: Berlin, 2007.
12. Croteau AM, Bergeron F. An information technology trilogy: business strategy, technological deployment and organizational performance. *Journal of Strategic Information Systems* 2001; **10**(2):77–99.
13. Guenzi P, Pardo C, Georges L. Relational selling strategy and key account managers' relational behaviors: an exploratory study. *Industrial Marketing Management* 2007; **36**(1):121–133.
14. Sohn SY, Joo YG, Han HK. Structural equation model for the evaluation of national funding on R&D project of SMEs in consideration with MBNQA criteria. *Evaluation and Program Planning* 2007; **30**(1):10–20.
15. Joo YG, Sohn SY. Structural equation model for effective CRM of digital content industry. *Expert Systems with Applications* 2008; **34**(1):63–71.
16. Fornell C, Cha J. In *Partial Least Squares in Advanced Methods of Marketing Research*, Bagozzi RP (ed.). Blackwell: Cambridge, MA, 1994; 52–78.
17. Hulland J. Use of partial least squares (PLS) in strategic management research: a review of four recent studies. *Strategic Management Journal* 1999; **20**:195–204.
18. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. *Multi- and Megavariate Data Analysis* (2nd edn). Umetrics AB: Umea, Sweden, 2006.
19. Henriksen H, Naes T, Segtnan V, Aastveit A. Using near infrared spectroscopy for predicting process conditions. A laboratory study from pulp production. *Journal of Near Infrared Spectroscopy* 2005; **13**(5):265–276.
20. Aguado D, Ferrer A, Seco A, Ferrer J. Comparison of different predictive models for nutrient estimation in a sequencing batch reactor for wastewater treatment. *Chemometrics and Intelligent Laboratory Systems* 2006; **84**:75–81.
21. Zhang J. Offset-free inferential feedback control of distillation compositions based on PCR and PLS models. *Chemical Engineering and Technology* 2006; **29**(5):560–566.
22. Sharmin R, Sundararaj U, Shah S, Griend LV, Sun YJ. Inferential sensors for estimation of polymer quality parameters: industrial application of a PLS-based soft sensor for a LDPE plant. *Chemical Engineering Science* 2006; **61**(19):6372–6384.
23. Mahani MK, Chaloosi M, Maragheh MG, Khanchi AR, Afzali D. Prediction of acute in vivo toxicity of some amine and amide drugs to rats by multiple linear regression, partial least squares and an artificial neural network. *Analytical Sciences* 2007; **23**(9):1091–1095.
24. Bruwer MJ, MacGregor JF, Bourg WM. Soft sensor for snack food textural properties using on-line vibrational measurements. *Industrial and Engineering Chemistry Research* 2007; **46**(3):864–870.
25. Lee HW, Lee MW, Park JM. Robust adaptive partial least squares modeling of a full-scale industrial wastewater treatment process. *Industrial and Engineering Chemistry Research* 2007; **46**(3):955–964.
26. Comeau Y, Rabinowitz B, Hall KJ, Oldham KW. Phosphate release and uptake in enhanced biological phosphorus removal from wastewater. *Journal of Water Pollution Control Federation* 1987; **59**:707–715.
27. Camacho J, Picó J, Ferrer A. Bilinear modelling of batch processes. Part I: theoretical discussion. *Journal of Chemometrics* 2008; DOI: 10.1002/cem.1113.
28. Arteaga F, Ferrer A. Dealing with missing data in MSPC: several methods, different interpretations, some examples. *Journal of Chemometrics* 2002; **16**(8–10):408–418.
29. Moller SF, von Frese J, Bro R. Robust methods for multivariate data analysis. *Journal of Chemometrics* 2005; **19**(10):549–563.
30. Rousseeuw PJ, Debruyne M, Engelen S, Hubert M. Robustness and outlier detection in chemometrics. *Critical Reviews in Analytical Chemistry* 2006; **36**(3–4):221–242.
31. Furusjo E, Svenson A, Rahmberg M, Andersson M. The importance of outlier detection and training set selection for reliable environmental QSAR predictions. *Chemosphere* 2006; **63**(1):99–108.
32. Lee HW, Lee MW, Park JM. Robust adaptive partial least squares modeling of a full-scale industrial wastewater treatment process. *Industrial and Engineering Chemistry Research* 2007; **46**(3):955–964.
33. Westerhuis JA, Kourti T, MacGregor JF. Comparing alternative approaches for multivariate statistical analysis of batch process data. *Journal of Chemometrics* 1999; **13**:397–413.

.

34. Nomikos P, MacGregor JF. Multivariate SPC charts for monitoring batch processes. *Technometrics* 1995; **37**(1): 41–59.
35. Zarzo M, Ferrer A. Batch process diagnosis: PLS with variable selection versus block-wise PCR. *Chemometrics and Intelligent Laboratory Systems* 2004; **73**(1):15–27.
36. Kourti T, MacGregor JF. Process analysis, monitoring and diagnosis using multivariate projection methods—a tutorial. *Chemometrics and Intelligent Laboratory Systems* 1995; **28**:3–21.
37. Kourti T, Nomikos P, MacGregor JF. Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *Journal of Process Control* 1995; **5**(4):277–284.
38. Aguado D, Zarzo M, Seco A, Ferrer A. Process understanding of a wastewater batch reactor with block-wise PLS. *Environmetrics* 2007; **18**:551–560.
39. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2000; **22**(1):4–37.
40. Lied TT, Geladi P, Esbensen K. Multivariate image regression (MIR): implementation of image PLSR-first forays. *Journal of Chemometrics* 2000; **14**:585–598.
41. Huang J, Esbensen KH. Applications of the angle measure technique (AMT) in image analysis. Part II: prediction of powder functional properties and mixing components using multivariate AMT regression (MAR). *Chemometrics and Intelligent Laboratory Systems* 2000; **54**:1–19.
42. Yu H, MacGregor JF. Multivariate image analysis and regression for prediction of coating and distribution in the production of snack foods. *Chemometrics and Intelligent Laboratory Systems* 2003; **72**:57–71.
43. Liu JJ, MagGregor JF. Modeling and optimization of product appearance: application to injection-molded plastic models. *Industrial Chemical Engineering Research* 2005; **44**:4687–4696.
44. Eriksson L, Wold S, Trygg J. Multivariate analysis of congruent images (MACI). *Journal of Chemometrics* 2005; **19**:393–403.
45. Prats-Montalbán JM, Ferrer A. Integration of colour and textural information in multivariate image analysis: defect detection and classification issues. *Journal of Chemometrics* 2007; **21**:10–23.
46. Geladi P, Grahn H. *Multivariate Image Analysis*. Wiley: Chichester, U.K., 1996.
47. Bharati MH, MacGregor JF. Texture analysis of images using principal component analysis. *SPIE/Photonics Conference on Process Imaging for Automatic Control*, Boston, MA, U.S.A., 2000.
48. Dayal BS, MacGregor JF. Identification of finite impulse response models: methods and robustness issues. *Industrial and Engineering Chemistry Research* 1996; **35**:4078–4090.
49. Barceló S, Vidal S, Ferrer A. Case study of comparison of multivariate statistical methods for process modelling. *Third Annual Meeting of European Network for Business & Industrial Statistics* (ENBIS), Barcelona, Spain, 2003.
50. Box GEP, Jenkins GM, Reinsel GC. *Time Series Analysis—Forecasting and Control* (3rd edn). Prentice-Hall: Englewood Cliffs, NJ, U.S.A., 1994.