Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 73 (2004) 15–27

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Batch process diagnosis: PLS with variable selection versus *block-wise* PCR

Manuel Zarzo, Alberto Ferrer*

Department of Applied Statistics, Operations Research and Quality, Polytechnic University of Valencia, Camino de vera s/n, 46022 Valencia, Spain

Received 17 June 2003; accepted 12 November 2003
Available online 8 March 2004

Abstract

Data from a batch chemical process have been analysed in order to diagnose the causes of variability of a final quality parameter. The trajectories of 47 process variables from 37 batches have been arranged in a matrix by using alignment methods. Two different approaches are compared to diagnose the key process variables: PLS with variable selection and *block-wise* PCR. The application of Unfold Partial Least Squares Regression (U-PLS) leads to one significant component. By means of weight plots, the variables most correlated with the final quality are identified. Nevertheless, with observed data, it is not possible to know if correlation is due to causality (and hence related to a critical point) or is due to other causes. Pruning PLS models by using variable selection methods and technical information of the process has allowed the process variables most correlated with the final quality to be revealed. The application of Principal Component Regression to the trajectories of the process variables (*block-wise* PCR) has given straightforward results without requiring a deep knowledge of the process. The results obtained have been used to propose several hypotheses about the likely key process variables that require a better control, as a previous step to conducting further studies for process diagnosis and optimisation, like experimental designs.

Keywords: PLS; PCR; Variable selection; Multiblock models

1. Introduction

Chemical processes can operate in continuous or batch stages. Batch processes are more difficult to control, because, often, the duration of the different stages is not constant from batch to batch, and it can have an effect on the final quality. The problem is that only variables like pressures or temperatures are known from the process. The quality is usually determined in the laboratory once the batch finishes after several hours. Very often, the final quality of the product is not right, and the causes remain unknown. From a statistical point of view, causes may be due to out-of-control situations (undesirable events or faults to be avoided), or due to an excessive variability in certain critical points that operate under control but would require a better adjustment to reduce the variability.

Two different approaches can be considered in order to diagnose the factors that affect the final quality of a batch process. One approach is the design of experiments (DOE), based on ‘‘forcing variability’’, that is, forcing the process to operate in certain predefined conditions according to a properly designed set of experiments. The advantage of this approach is that it allows the identification of the factors that significantly affect the final quality, and also allows the establishment of the optimal operative conditions. The main problem is the need for conducting the experiments, which is expensive and in the course of the experiments there is a risk of obtaining a low-quality product or threaten the security of the process. Besides, in very complex processes, the number of potential factors can be really very high, which would make unapproachable any experimental proposal. If finally some factors are chosen to conduct a DOE, the probability to leave out of the experiments unknown factors that significantly affect the final quality can be very high.

A different approach, yet complementary, is the analysis of the ‘‘observed variability’’ of the process. With the

* Corresponding author. Tel.: +34-963877000x74932; fax: +34-963877499.

E-mail addresses: mazarcas@eio.upv.es (M. Zarzo), aferrer@eio.upv.es (A. Ferrer).

0169-7439/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2003.11.009

current development of sensors and controllers for chemical processes, quite often, huge data bases from process variables are available for historical batches. These data bases contain valuable information, and if they are conveniently analysed off-line, predictive models of the final quality can be obtained in order to detect faults, critical points or to implement models for on-line monitoring. The main advantage of this approach versus DOE is that it does not require any experimentation with the process, yet, the statistical analysis of process data is more complex than the data analysis from DOE. Nomikos and MacGregor [1] developed a methodology to analyse historical data of batch processes aimed at monitoring, which has also been applied for diagnosis [2]. This methodology unfolds three-way data matrices (batches × process variables × time) into a two-way matrix suitable for the application of Principal Component Analysis (PCA) and Partial Least Squares Regression (PLS). Both are methods based on projection to latent structures, which provide a way to handle the highly correlated data registered by electronic sensors from chemical processes. Moreover, they deal effectively with multiple quality and productivity variables and with missing data, and they provide a good tool to extract and highlight the systematic variation of data, reducing the dimension of the problem to a space of few components.

When analysing observed variability, care has to be taken in the diagnosis of the variables significantly correlated with the final quality. On the one hand, the observed correlation can be at random. But in the case that correlation is not at random, it does not necessarily imply that the variables cause the variation of the quality. Correlation can be due to different reasons, and not always due to a cause–effect relationship. When data from a process are analysed for diagnosis, the faults or critical points will be related to variables with causal correlation regarding the final quality. But in the analysis of observed data with statistical methods, it is not possible to know the reasons of the observed correlation, unless empirical models are supplemented with external information of the process. For example, if a change occurs in a process that affects 10 process variables and also the final quality parameter, all of them will be correlated with the quality, but it does not mean that they are all responsible for the variation in quality. Aimed at developing an efficient methodology for diagnosis able to cope with the drawbacks described, two approaches are presented: The use of PLS models with a procedure for variable selection using external information of the process, and a simple method based on the application of Principal Component Regression to the different trajectories that has been denominated ‘‘*block-wise* PCR’’, which has turned out to be more efficient than the previous one. Both approaches have been applied to diagnose a batch chemical process in order to detect the critical points that require a better control to reduce the variability of the final quality and hence to avoid batches out of specifications. Wold et al. [3] have presented a discussion on the use of hierarchical models and variable selection.

The proposal of this paper is to merge DOE and analysis of observed data: The use of multivariate statistical tools to reduce the number of potential key process variables and select the likely critical factors, as a previous step to conducting a further DOE with them in order to finally achieve the diagnosis of the process. A similar methodology to deal with data from a chemical batch process is presented by Yacoub and MacGregor [4].

2. Methodology for data pre-treatment

2.1. Description of the process and variables

The process studied is the elaboration at industrial scale of the polymer polypropylene oxide (PPOX), used as a raw material for the fabrication of flexible polyurethane foams. Data have been supplied by a chemical company located in Spain. It is a batch process in four consecutive stages, using propylene oxide and a polyalcohol as reagents. The average total duration of the process is 22 h, and the average duration of each stage is presented in Table 1. Every stage takes place in a different tank or reactor, and in some of them, several consecutive operations (sub-stages) are carried out. When the stage finishes, all the content of the batch is transferred to the next tank. In total, 47 parameters are registered from the process by electronic sensors, whose code (in brackets) and description is presented below.

Stage 1 is the preparation of the initiating solution and takes place in three sub-stages. The first one begins with the addition of a predetermined amount of polyalcohol to the empty tank, and the instant (1FP) and cumulated flow (1FPcu) are registered. The second sub-stage is the addition of an alkali in watery solution, and also the instant (1FAL) and cumulated (1FALcu) flow are collected. In the third sub-stage, a vacuum is carried out to dehydrate the solution. The tank is heated by means of a heating circuit controlled by a valve (1VAL). The information collected inside the tank along the stage is: pressure (1PR), temperature (1T[a]), pH (1pH) and level (1LEV).

Table 1
Process variables considered in the analysis

Stage	N.ac.[a]	N.dis.[b]	N.t.[c]	N.cu.[d]	N.de.[e]	N.tr.[f]	N.al.[g]	Hours[h]
1	9	1	2	0	3	13	1635	5.6
2	12	6	1	5	0	12	1655	6.3
3	12	3	1	5	0	15	2709	6.7
4	14	5	1	2	0	12	2986	3.4
Total	47	15	5	12	3	52	8985	22

[a] Number of process variables acquired.
[b] Number of process variables discarded.
[c] Number of variables of time from the alignment.
[d] Number of cumulated oscillating variables.
[e] Number of derivative transformed variables.
[f] Total number of trajectories ($a-b+c+d+e$).
[g] Total number of aligned variables.
[h] Average duration of the stage (hours).

When this stage finishes, the content of the batch is transferred to a new reactor where stage 2 takes place: the elaboration of the pre-polymer. A certain amount of exactly measured propylene oxide is added, and the instant (2FO) and cumulated (2FOcu) flows are registered. Inside the tank, different information is acquired: pressure (2PR), stirring intensity (2STI) and temperature measured at two different points (2T^{a}a and 2T^{a}b). As the reaction is exothermic, the temperature is controlled by means of a cooling circuit, from which 6 variables are collected: the temperature at three different points (2T^{a}c, 2T^{a}d and 2T^{a}e), the pressure (2PC) and two valve openings (2VALa, 2VALb).

When the reaction finishes, the content of the batch is transferred to a new tank where the pre-polymer turns into the polymer (stage 3). A second addition of propylene oxide is conducted, registering the instant (3FO) and cumulated (3FOcu) flows and the valve that controls the addition (3Oval). Inside the reactor, four process variables are collected: pressure (3PR), intensity of the stirrer (3STI) and temperature measured at two different points (3T^{a}a, 3T^{a}b). A cooling circuit controls the temperature, and several variables are collected from the circuit: temperature at two different points (3T^{a}c, 3T^{a}d), pressure (3PC) and two valve openings (3VALa, 3VALb).

When the polymerisation reaction finishes, the content of the batch is transferred to a new tank to launch stage 4, which consists of three sub-stages. First, a short vacuum is conducted to eliminate rests of propylene oxide that have not reacted in the previous stage. Afterwards, an acid in watery solution is added, and the instant (4FAC) and cumulated (4FACcu) flow are registered. The acid is added from a tank from which 3 variables are collected: temperature (4Taci), pressure (4PRaci) and a valve that regulates a heating circuit (4VALaci). The third sub-stage is a vacuum dehydration, that takes place at high temperature, regulated by a heating circuit controlled by a valve (4VAL). The tank is stirred, and three parameters are registered from the stirrer (4STIa, 4STIb, 4STIc). During this stage, temperature (4T^a) and pH (4pH) are collected inside the tank, and also the pressure, measured by two sensors of differing accuracy and measuring range (4PRa, 4PRb). When the dehydration finishes, the tank is emptied and the pressure in the emptying pipe (4PC) is collected.

Once the batch is elaborated, a sample is taken and brought to the laboratory to analyse the quality. One of the quality parameters measured is the hydroxyl index (that will be referred as I_{OH}). It is proportional to the ratio of the number of hydroxyl groups per molecule and the molecular weight, and basically depends on the type and proportion of the reagents. This parameter is positively correlated with the rigidity of the polyurethane foam obtained from the PPOX. Thus, there is a strong motive to produce the polymer by reducing the variability of the hydroxyl index around the nominal value as much as possible. The analytical results are available after several hours. The problem detected in the factory is that the variability of this quality parameter is too high, and as a consequence, a certain percentage of batches are out of specifications. There is a keen interest to identify the critical points of the process that require a better control to reduce the variability of the final quality, and hence to avoid batches out of specifications. The company has provided data from 37 batches chosen randomly from the period between December 3, 2000 and January 18, 2001. For every batch, the 47 process variables mentioned are available, which are registered on-line from the process by means of electronic sensors with a sampling frequency of 1 min.

2.2. Variables discarded and transformed

Two of the 47 process variables have been discarded because they supply nearly the same information as other variables: 2T^{a}e (its values nearly coincide with 2T^{a}c) and 3Oval (values highly correlated with 3FO). The variables 4PRaci and 4VALaci have also been removed, as they provide information about the tank of acid, and according to technical knowledge, this information is not important. The values of the variables 4PRa and 4PRb have been combined into a single variable.

In the case of the additions, taking into account that the instant flow is the derivative of the cumulated flow, both variables are related and supply a similar information. The main information is provided by the cumulated flows that have been used as indicator variables for the alignment, producing new variables of time for the addition of different reagents: polyalcohol (1t-p), alkaline solution (1t-al), propylene oxide in stage 2 (2t-ox) and 3 (3t-ox), and acid solution (4t-aci). These new variables contain redundant information regarding the cumulated flows, and for that reason, some of them have been discarded (1FALcu, 2FOcu, 4FACcu), and also some instant flows (2FO, 3FO, 4FAC).

A way to incorporate external information specific to the process is with transformed variables generated from process variables. Ramaker et al. [5] describe several examples about how to incorporate external information. The trajectories of the variables from the cooling circuits oscillate with high amplitude and low period (a few minutes). Thus, the information will not be at the value that one of these variables takes at any given moment, but in the general pattern of the variable along the stage of the batch. For that reason, cumulated values (integrated over time) that are more sensitive in detecting changes of trends have been calculated for these variables. The valve opening variables contain values comprised from 0 (valve closed) to 100 (totally opened). The cumulated values have been calculated respect zero, obtaining a new trajectory that increases monotonically, and the new variables 2VALa_cu, 2VALb_cu and 4VAL_cu have been created. These transformed variables will be more powerful at detecting those valves that in average have been opened or closed for longer. Some variables of temperature and pressure have also shown an oscillating evolution, and the cumulated values have been calculated with respect to the mean value, producing a non-monotonic trajectory, and the following new variables have been created: 2PR_cu, 2T^{a}c_cu, 2PC_cu, 3PR_cu, 3T^{a}a_cu, 3VALa_cu, 3VALb_cu, 3PC_cu and 4PR_cu. The most oscillating original variables (2VALa, 2VALb, 2PC and 3PC) have been discarded for the analysis, as their information relies on the corresponding cumulated variables created.

Another type of transformation used is the derivative of the trajectory. According to technical information, the degree of dehydration in stage 1 can be related with the slope of the trajectory of the variables that supply information inside the tank. For that reason, new variables have been created: the derivative of pressure, temperature and level in stage 1 (coded as 1PRd, 1T^{a}d and 1LEVd).

2.3. Alignment

Starting with the variables provided by the company, discarding and adding new variables as indicated, 52 trajectories are available according to Table 1. In the process, all stages and sub-stages have a different duration from batch to batch, what requires the application of alignment methods to correct, synchronise or align the trajectories of the variables in order to handle comparable data among batches. In those cases where a reagent is added, the indicator variable approach has been used, as proposed by Nomikos and MacGregor [6]. The application of this method has been possible because the cumulated flows are monotonically increasing variables that are accurately measured and present the same values at the beginning and the end for all batches (as the total amount added is constant). Considering the cumulated flow as indicator variable, a new pseudo-temporal variable used is the "time needed to fill the 0%, 1%, 2%,. . ., 100% of the total amount of the reagent" (variables 1t-p, 1t-al, 2t-ox, 3t-ox, 4t-aci previously cited). These times are obtained from the trajectory of the cumulated flows using linear interpolation, and at the different times, the values of the other process variables of the sub-stage are calculated. Thus, the time needed to fill 0% corresponds to all batches with the beginning of the sub-stage, and the time to fill 100% is the end of the sub-stage, resulting in a synchronisation or alignment of the trajectory in a scale of pseudo-time.

This method could not be applied to the sub-stages of dehydration. In these cases, the sub-stage starts with a lowering of the pressure, and ends when the vacuum breaks and turns into atmospheric pressure. The duration is not constant but it is not possible to define an indicator variable to settle the end of the sub-stage. The criterion used is the one described by Louwerse et al. [7]: For every batch, the duration of the vacuum is divided into 100 parts, obtaining the instants of time that account for 0%, 1%, 2%,. . ., 100% of the duration of the sub-stage, and at those instants of time, the values of the other process variables are calculated.

2.4. Unfolding

The application of the alignment methods produces for every stage a new matrix of three-way data, which consists of J process variables measured in K instants of corrected time, in I batches. The application of PLS to matrices unfolded into a two-way structure is called "unfold PLS" (U-PLS) [8], a term that is currently preferred instead of "multiway PLS" proposed by Wold et al. [9]. In the unfolding, one of the directions remains unaltered, while the other is the rearrangement of the two other directions slice by slice. Six ways of unfolding the matrix are possible, as indicated in Table 2. Matrices B and C are equivalent, just reordering the rows; matrices D and E are also equivalent by reordering the columns. The F matrix is the transpose of A. More details about multiway methods and their unfolding for analysis have been described by Smilde [10].

Wold et al. [11] use type A, which, according to those authors, is motivated when on-line monitoring of the batch process is wanted. The unfolding procedure used by Nomikos and MacGregor [1] corresponds to type D. This is more common for the analysis of historical data (as in the present case) but is also quite widely applied for on-line monitoring. Westerhuis et al. [12] and Kourti [13] compare several ways to unfold the data matrix and discuss their effects for the monitoring of batch processes.

In this paper, unfolding type E is going to be used, maintaining the direction of batches and arranging the trajectory of the first process variable, afterwards the next trajectory, and so on according to Fig. 1. As the matrices unfolded according to types D and E are equivalent, the application to PLS to both matrices produces exactly the same results. Nevertheless, the interpretation of the results from the graphs can lead to different conclusions according to the type of unfolding. The advantage of type E in the diagnosis of PLS models, as described later on, is that it allows the comparison of the trajectory of the weights of the PLS model with the trajectory of the original variables, making it easier to diagnose the underlying causes of variability.

The total number of aligned variables is 8985 (see Table 1). It is a very high number of variables, compared with

Table 2
Types of unfolding a three-way matrix, according to Westerhuis et al. [12]

Type	Structure[a]	Direction[b]
A	$KI \times J$	variables
B	$JI \times K$	time
C	$IJ \times K$	time
D	$I \times KJ$	batches
E	$I \times JK$	batches
F	$J \times IK$	variables

[a] Structure of the unfolded matrix.

[b] Direction that remains unaltered.

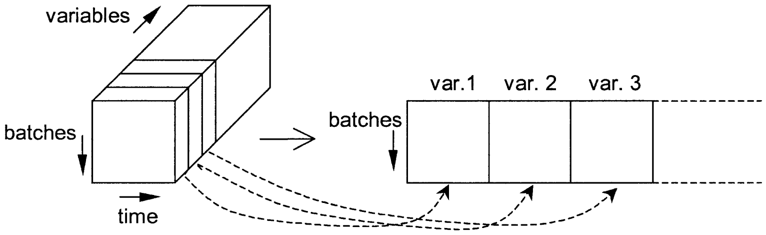

Fig. 1. Unfolding scheme corresponding to type E.

other industrial examples of batch processes described in the literature.

3. Batch process diagnosis: results and discussion

In order to help in the diagnosis of the key process variables that affect the final quality, two different approaches are going to be compared: PLS with variable selection methods and *block-wise* PCR.

3.1. PLS with variable selection methods

A PLS regression has been applied to the unfolded matrix with the software SIMCA-P 8.0 of Umetrics. Due to the small number of batches available, the data set has not been split into a training and prediction set. Thus, all available batches have been used to build the different models, using cross-validation to determine significant components. In all PLS models conducted, data have been centred and scaled to unit variance. The first component is significant, and explains 15% of the variation in the **X** matrix (R_X^2), and 54% the variation of the quality variable (R_Y^2), with a cross-validated (Q^2) value equal to 30%. Fig. 2 shows the weights of the 8985 variables in the first PLS component. Many periods with high values of weights in absolute value can be seen, but no trajectory stands out more than the rest, which would help the diagnosis.

The analysis of the weights allows the identification of the variables most correlated with the final quality, but it is not possible to know if that correlation is causal. To try to overcome this drawback, PLS models have been simplified with variable selection criteria by incorporating external information of the process (technical knowledge). The methodology applied justifies the type of pre-processing used for the variables: data centred and scaled to unit variance. One revision of pre-processing methods for multi-way data has been presented by Harshman and Lundy [14].

As a row of the unfolded matrix (one observation) contains all the information of a batch, when data are centred, the average trajectory for the 37 batches is subtracted from every process variable, eliminating the main non-linearity due to the dynamic behaviour of the process. Thus, the multivariate analysis with centred data is a study of the systematic variation of the trajectories with respect to the average trajectory. The critical points will be those variables whose deviation from the average trajectory causes the variability of the final quality.

The scaling of the variables is an important matter, as it affects the multivariate models. When process data are analysed, the simplest approach is to scale all aligned variables to unit variance. One drawback of this approach is that possibly the weight of periods of low variability in the trajectory of a process variable (that might be associated with noise) can be overestimated in comparison with the rest of the trajectory with consistent information. To avoid this,

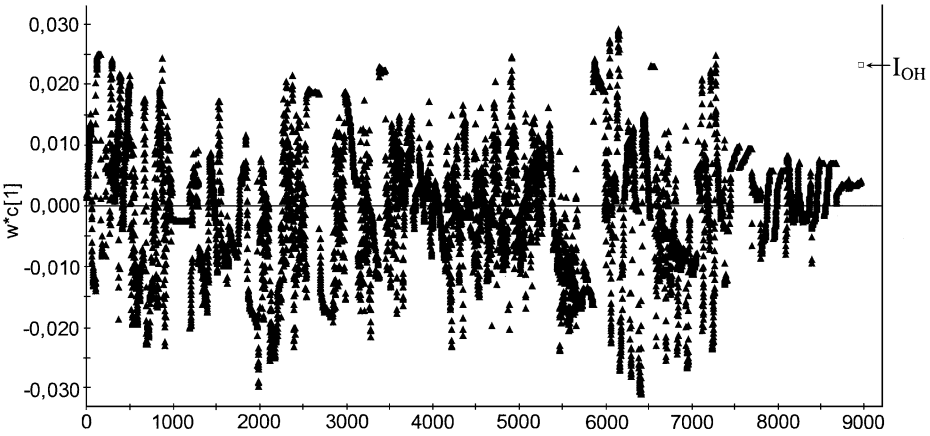

Fig. 2. Plot of the weights of the 8985 aligned variables in the first PLS component.

different criteria of scaling can be used. One advantage of using several scaling criteria is the possibility to introduce external information of the process variables. If it is known a priori that a certain period is critical, it is recommended to weight that period in order to exert more influence over the model. In this case, although according to technical knowledge, it is known that the importance of all the variables is not the same, the scaling to unit variance has been used because it facilitates the diagnosis of PLS models, according to the methodology described later, as this type of scaling satisfies the following property:

> Working with data centred and scaled to unit variance and with a single response variable, the weights of variables in the first PLS component are proportional to the linear correlation coefficients between the variables and the response variable.

Brown [15] demonstrates this property as follows. When PLS is applied to a data matrix **X** of I observations by N aligned variables (being N the product of J process variables and K instants of corrected time), the observations are projected over a certain direction, obtaining a score vector $\boldsymbol{t}$ that contains the projections of those observations in the direction defined by the unit vector of weights $\boldsymbol{w}$. When a single response variable is used (then the PLS is usually referred as PLS-1), that direction is the one that maximises the covariance between the score vector and the response variable ($\boldsymbol{y}$ vector). From the definition of covariance and taking into account that values are centred, this expression follows:

$$cov(\boldsymbol{t},\boldsymbol{y}) = \frac{\sum_{i=1}^{I}(t_i-\bar{t})\cdot(y_i-\bar{y})}{I-1} = \frac{\sum_{i=1}^{I}(t_i\cdot y_i)}{I-1} = \frac{\boldsymbol{t}^{\mathrm{T}}\cdot\boldsymbol{y}}{I-1} \tag{1}$$

When the **X** matrix is projected over the direction defined by $\boldsymbol{w}$, the $\boldsymbol{t}$ vector is obtained.

$$\mathbf{X}\,\boldsymbol{w}=\boldsymbol{t} \quad \rightarrow \quad \boldsymbol{w}^{\mathrm{T}}\,\mathbf{X}^{\mathrm{T}}=\boldsymbol{t}^{\mathrm{T}} \tag{2}$$

As PLS maximises the covariance between vectors $\boldsymbol{t}$ and $\boldsymbol{y}$, the following expressions are obtained:

$$\max\ cov(\boldsymbol{t},\boldsymbol{y}) \propto \max(\boldsymbol{t}^{\mathrm{T}}\boldsymbol{y}) = \max(\boldsymbol{w}^{\mathrm{T}}\ \mathbf{X}^{\mathrm{T}}\boldsymbol{y}) = \max[\boldsymbol{w}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{y})] \tag{3}$$

The last equivalence is the scalar product of two vectors: the unitary vector $\boldsymbol{w}$ and $(\mathbf{X}^{\mathrm{T}}\boldsymbol{y})$. The modulus of this last vector is a constant for a given data, as it contains the covariates between the N variables $\boldsymbol{x}_n$ and $\boldsymbol{y}$. Thus, the scalar product is maximum if both vectors are parallel. As $\boldsymbol{w}$ has to be parallel to $(\mathbf{X}^{\mathrm{T}}\boldsymbol{y})$, and as all variables have unit variance, it leads finally to expression (4), demonstrating that the elements of the weight vector $\boldsymbol{w}$ (that is, the weights of the variables in the first PLS component) are proportional to the linear correlation coefficients between the variables and the response variable.

$$\boldsymbol{w}=\frac{\mathbf{X}^{\mathrm{T}}\cdot\boldsymbol{y}}{||\mathbf{X}^{\mathrm{T}}\cdot y||}\propto\begin{bmatrix}\boldsymbol{x}_1^{\mathrm{T}}\cdot\boldsymbol{y}\\ \boldsymbol{x}_2^{\mathrm{T}}\cdot\boldsymbol{y}\\ \cdots\\ \boldsymbol{x}_N^{\mathrm{T}}\cdot\boldsymbol{y}\end{bmatrix}\propto\begin{bmatrix}cov(\boldsymbol{x}_1,\boldsymbol{y})\\ cov(\boldsymbol{x}_2,\boldsymbol{y})\\ \cdots\\ cov(\boldsymbol{x}_N,\boldsymbol{y})\end{bmatrix}=\begin{bmatrix}r(\boldsymbol{x}_1,\boldsymbol{y})\\ r(\boldsymbol{x}_2,\boldsymbol{y})\\ \cdots\\ r(\boldsymbol{x}_N,\boldsymbol{y})\end{bmatrix} \tag{4}$$

According to the type of unfolding used, Fig. 2 represents the trajectory of the weights for the juxtaposition of the 52 trajectories. It can be observed that from the aligned variable number 7364, weights fluctuate in a band narrower than for the previous variables. These weights correspond to seven trajectories of stage 4 (4Taci, 4pH, 4t-aci, 4STIa, 4STIb, 4STIc and 4PC). To ensure that the observed weights (proportional to the correlation coefficients) do not differ significantly from zero, a PLS has been conducted with these trajectories, and the first component is not significant. The absence of correlation does not necessarily imply that these variables do not have any real effect over the hydroxyl index, but that the variability observed for these variables in the group of 37 batches does not produce any significant effect on the variability of the final quality. It is possible that one variable affects the final quality but is very well controlled, with a very reduced variability from batch to batch: This would provoke no observed correlation. Then, it can be ascertained that these process variables are not critical points that require a better control, and they have been removed from the model.

Working with thousands of variables, the probability of obtaining correlation at random is high. That obliges one to be careful during the diagnosis. Although the probability of randomly getting a variable with high weights at absolute value is high, the probability of randomly finding a group of aligned variables at consecutive moments in time is low. For that reason, emphasis should not be laid so much on particular values of high correlation, but on runs with high weights at absolute value (groups of consecutive aligned variables in the weight plot), which have been called “runs of correlation”. Taking a look at Fig. 2, the problem is that nearly all the trajectories not yet discarded have some runs of correlation. The diagnosis gets complicated, because the presence of correlation does not imply the identification of a critical point. Only when correlation is due to a cause–effect relationship (that is, causal correlation) will a critical point be identified. In observed data, however, this diagnosis is not possible if the model is not supplemented with external information.

A good review of different variable selection methods for PLS models is presented by Gauchi and Chagnon [16], who use real data from chemical batch processes and apply different methods based on statistical criteria related with the goodness of fit and prediction, but do not implement external information of the process. The methodology used in the present paper to try to distinguish which runs of correlation are associated with critical points is based on the application of technical information of the process according to the following procedure. For every trajectory of process variables, the evolution of the weights ***w*** versus time (selecting the range of the aligned variables in Fig. 2 corresponding to that trajectory) has been compared with the original trajectory of the process variable, with data aligned without centering nor scaling (where the trajectories of the batches with higher values of I_{OH} have been highlighted). Between both charts, there is a total correspondence, as for every aligned variable of the trajectory, the weight and original values for the 37 batches are known. Once the runs with correlation are identified, they are matched with the original trajectory, and external information of the process (technical knowledge) is applied in order to see if any kind of diagnosis is possible, or to find an explanation for the observed correlation. Of course, it requires a good knowledge of the process and of the likely causes responsible for the variability of the final quality. In the case that no reasonable explanation can be found and that the process variable in the period with correlation is not considered important, the whole trajectory has been removed.

The methodology used is now described with an example: the case of 3T^{a}d (trajectory of temperature at a point of the cooling circuit in stage 3). The top of Fig. 3 shows the weights of that trajectory (obtained from Fig. 2, selecting the range of the aligned variables from 4870 to 5067 that correspond to that trajectory), and at the bottom, the original aligned values. The vertical scale has been omitted for reasons of confidentiality. A run of correlation with weights greater than 0.020 can be observed at the top of the figure. Compared with Fig. 2, these values of weights are high, with respect to the rest of variables. If in the interval of time when this run occurs we observe the original values of the unscaled variables, it corresponds with a period where the temperature has been very well controlled, with a very low variability, in comparison with the subsequent moments with higher variability, although no correlation is observed. Using technical knowledge, it is clear that this observed correlation is not causal; thus, the trajectory can be discarded.

Operating with this methodology, trajectories have been removed progressively. If a trajectory was considered important, and any technical fundament could be formulated to support or explain the correlation, it was maintained in the final model. In most cases, however, it was not so easy to decide whether the trajectory was important or not. First, the trajectories considered less important have been discarded, producing an intermediate model where it was risky to remove more variables. However, to try to achieve the diagnosis, it is necessary to take the risk of discarding more trajectories, assuming that a trajectory with causal correlation may be removed. It should be pointed out that extensive variable selection is very risky and that this must be carried out cautiously. Finally, a model with only seven trajectories has been obtained. The first component is significant in the initial, intermediate and final models. The parameters R^2_X, R^2_Y and Q^2 of the three models are compared in Table 3. The methodology is slow and time consuming, as all of the trajectories have been carefully analysed, and it also requires a deep knowledge of the process.

It can be observed in Table 3 that in the intermediate model, when about two thirds of the aligned variables are discarded, the goodness of fit and the goodness of prediction are higher. When only seven trajectories are left in the final model, both parameters decrease, but are higher than the ones

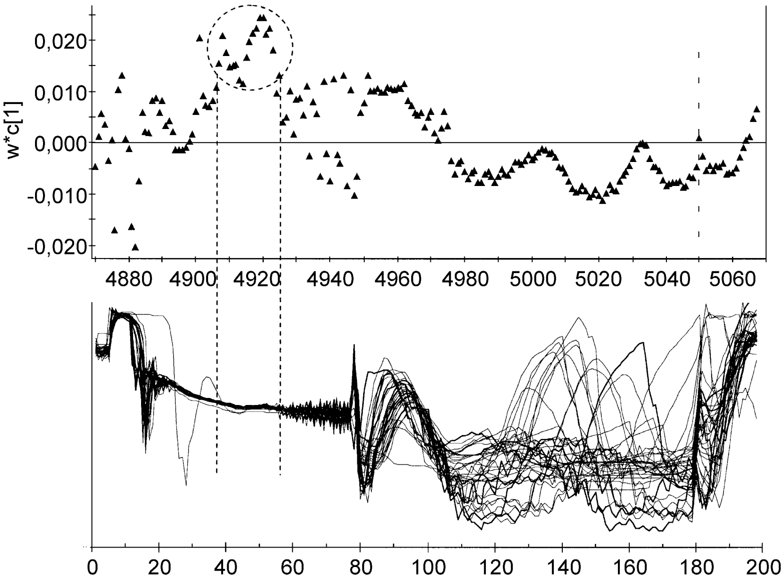

Fig. 3. Correspondence between the weights of the trajectory 3T^{a}d in the first component of the initial PLS model, and the aligned original values of 3T^{a}d (trajectories for the six batches with higher values of I_{OH} highlighted in thicker lines).

Table 3
Comparison of three PLS models

PLS model	N[a]	N[b]	R_X^{2c}	R_Y^{2d}	Q^{2e}
Initial	52	8985	0.15	0.53	0.30
Intermediate	19	3107	0.12	0.71	0.53
Final	7	1067	0.19	0.57	0.46

[a] Number of trajectories included in the model.
[b] Number of aligned variables included in the model.
[c] Variance explained by the first PLS component (R_X^2).
[d] Goodness of fit for the first PLS component (R_Y^2).
[e] Goodness of prediction for the first PLS component (Q^2).

for the initial model. Thus, this final model can be considered suitable for diagnosis purposes, as it retains the main variables contributing to the predictive capability of the model.

Fig. 4 shows the weights of the first component for the final PLS model. Vertical dashed lines separate the seven trajectories, that are the following (in the order appearing in the figure): 1FPcu, 1PR, $1T^a$, 2PR, 4PR, 4VAL_cu and $4T^a$.

It can be observed that all of them have runs of correlation with high absolute weights. These values are not comparable with the ones from Fig. 2, because when trajectories are removed, weights are recalculated to maintain the unit modulus of the weight vector, although the relative values among the aligned variables remain the same. The last three trajectories belong to stage 4 and present high absolute weights, similar to the rest of trajectories. Nevertheless, by applying technical knowledge, those trajectories can be discarded. The hydroxyl index is a quality parameter related with the chemical structure of the polymer. At the end of stage 3, it is already formed, and in stage 4, operations that do not alter the chemical structure take place. So, with this information, it is known that the observed correlation for the latent variables in stage 4 will not be causal, and hence not associated to critical points. Finally, the remaining four trajectories can be considered as hypothetical critical points: 1FPcu, 1PR, $1T^a$ and 2PR.

3.2. Block-wise PCR

In the previous methodology, a variable selection procedure has been carried out with the unfolded matrix of 37 batches by 8985 aligned variables. However, the reduction of variables to interpret PLS models often removes information. When the **X** matrix can be divided into meaningful blocks, different methods have been described with names like multiblock PLS, consensus PCA, hierarchical PCA or hierarchical PLS. Westerhuis et al. [17] review these algorithms and compare them from a theoretical point of view. All of them produce a model at two levels: the upper level where the relationships among blocks are modelled and the lower one showing the details of each block. The model is conducted by a single algorithm with a hierarchical structure that calculates weights and scores at both levels. Many applications of these multiblock models have been described in the chemometric literature [18].

In batch processes, the unfolded matrix is the arrangement slice by slice of the trajectories of process variables. In the present case, the **X** matrix could be subdivided in four blocks, corresponding to the different stages of the process, or even in 52 blocks (one per trajectory). The use of multiblock models can take advantage of this subdivision, and successfully applications have been described for monitoring and fault diagnosis [19]. Different methods like multiblock PLS or hierarchical PLS could be applied to diagnose the PPOX process. In this paper, however, a simpler procedure is presented, carrying out the analysis in two stages. First, for every one of the 52 trajectories that comprise the unfolded matrix, a PCA has been carried out with data centred and scaled to unit variance, and the significant components have been obtained, whose number ranged from 1 to 13. Due to the small number of batches available, all of them have been used to build the models. No batches have been considered as a prediction data set to validate the components. Nevertheless, the cross-validation procedure implemented in the software SIMCA-P has been used. From each component, the percentage of explained variance in **X** (R_X^2) has been worked out. The projection of the batches over every component generates a latent variable, which allows the transformation of every submatrix of trajectory into a new submatrix of scores, with a reduced number of new latent variables. For example, beginning with the trajectory of the temperature in stage 1, formed by 230 aligned variables, conducting a PCA, nine significant components are obtained, which account for 97.2% of the

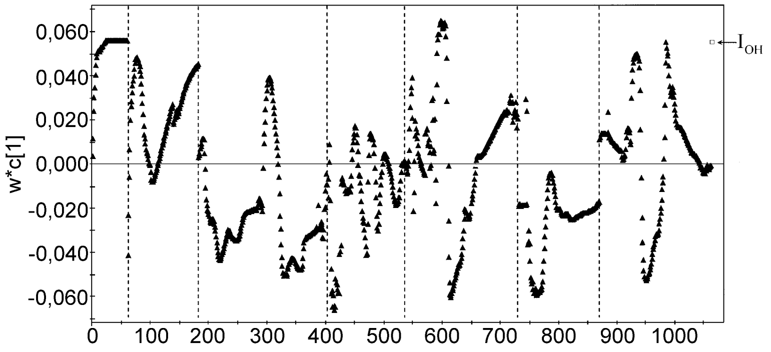

Fig. 4. Weights of the first component for the final PLS model, with seven trajectories.

Table 4
Latent variables selected from the Score Matrix

Stage	N[a]	N[b]	N[c]
1	67	6	6
2	65	13	13
3	115	15	10
4	65	9	9
Total	312	43	38

[a] Initial number of latent variables.

[b] Number of latent variables significantly correlated with I_{OH} (at $\alpha = 0.05$).

[c] Number of latent variables significantly correlated with I_{OH} and R^2_X>0.05 (explained variation of the trajectory).

total variance of the data. Carrying out 52 PCA (one per trajectory), the initial unfolded matrix of 8985 variables is transformed into a Principal Component score matrix with only 312 latent variables. Although the reduction in size is considerable with respect to the initial unfolded matrix, both matrices contain nearly the same information.

In a second stage, this PC score matrix has been analysed to obtain predictive models of the final quality by using simple linear regression. This methodology has been called "*block-wise* PCR", as it is a Principal Component Regression carried out in every block of aligned variables than comprise a trajectory. As one of the referees has suggested, and also according to Wold et al. [20], an alternative approach would have been to conduct PLS models with the blocks, given that a ***y*** variable is present.

For every one of the 312 latent variables, a simple linear regression analysis has been conducted, considering the I_{OH} as response variable, and two parameters have been obtained: the squared linear correlation coefficient (that will be referred as R^2_{IOH}) and the *p-value* (that shows if the correlation is significant). It results that from the 312 latent variables, only 43 are significantly correlated with the I_{OH}, at a confidence level of 95% (that is, with a *p-value* < 0.05). The maximum R^2_{IOH} found has been 34%.

It seems reasonable to suppose that those latent variables with causal correlation will explain a certain amount of the variance of the corresponding trajectory. If a latent variable presents a very low R^2_X, although it is significantly correlated with the I_{OH}, it is probable that the correlation is at random. The criteria adopted has been to remove those latent variables that explain less than 5% of the variance ($R^2_X < 5\%$). This happens in five latent variables from stage 3 whose $R^2_X < 3\%$ and with R^2_{IOH} ranging from 16.6% to 20.4%. Thus, 38 latent variables are remaining, as stated in Table 4.

It seems reasonable that the critical stage of the process will be linked to one of the latent variables with higher values of R^2_{IOH}. In this case, however, considering the nine latent variables with $R^2_{IOH} > 20\%$, three of them belong to stage 1 (whose values are 34.0%, 25.5% and 21.2%), two belong to stage 2 (20.2% and 29.4%), three belong to stage 3 (21.1%, 21.5% and 23.5%) and one to stage 4 (27.6%). It is not clear which is the critical stage, as no one presents higher values or a specially high value. Therefore, some kind of additional information is required to achieve a further selection among the 38 latent variables, aimed at identifying the critical stage.

Arranging the values of I_{OH} versus time, it results that the greater values of I_{OH} are obtained after batch number 20. To check if a swift in the mean value of the quality index has occurred, a chart of cumulative sums (CUSUM) has been plotted integrating over time the difference of the values with respect to the mean. This chart is included in Fig. 5 together with others of the same type that will be commented later.

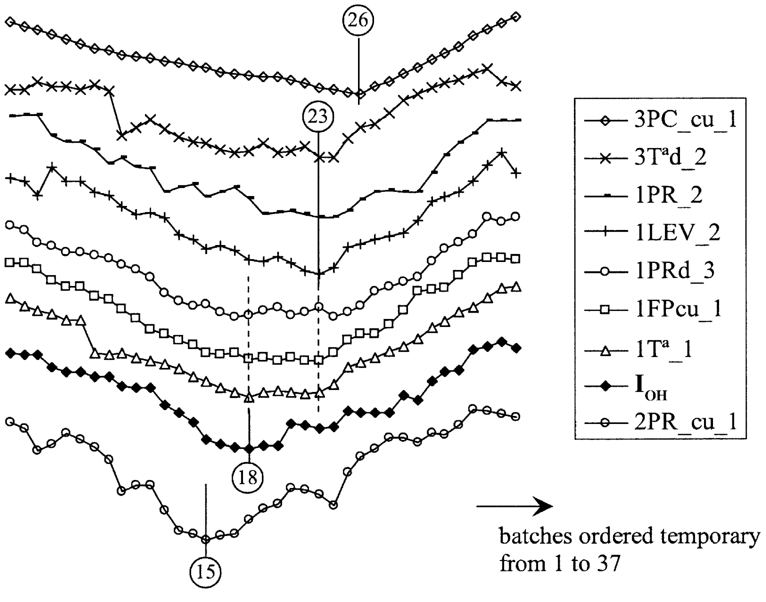

Fig. 5. Overlapping of the CUSUM chart of the hydroxyl index with the eight latent variables of Table 5 from stages 1, 2 and 3. Vertical lines mark the instant when the change of trend occurs, with the indication of the batch order.

A change of trend is detected, which occurs at about batch number 18 (produced on December 23, 2000). Something must have changed in the process around that date, which has provoked a change in the mean value. This result supplies an important information, because the likely latent variables with causal correlation associated to critical points should also suffer a change of trend around that date. With the same procedure used to obtain the CUSUM chart for the values of I_{OH}, the CUSUM values (according to the temporal order of the 37 batches) have been calculated for every one of the 38 latent variables selected, obtaining a matrix of CUSUM values of scores. The idea is to find those latent variables that produce a CUSUM chart most similar to the one for the hydroxyl index. For that purpose, a simple linear regression has been conducted to relate the CUSUM values of every one of the 38 variables, with the CUSUM values of I_{OH}, and the squared correlation coefficient of the regression has been calculated, which will be referred to as R^2_{CUSUM}. Fig. 6 represents the relationship between the coefficients R^2_{IOH} and R^2_{CUSUM} for the 38 latent variables.

We observe that the values of R^2_{CUSUM} are higher than R^2_{IOH}. Such high apparent correlation should not lead to confusion, and it should be remembered that the CUSUM values are calculated for a certain batch using the information of all the previous batches. So, the coefficient R^2_{CUSUM} does not account for the percentage of the variability of I_{OH} explained by the latent variable, as it is the case for the coefficient of determination R^2_{IOH}. It is just an index, and the greater the value, the more similar will be the CUSUM chart to the one for the I_{OH}.

A group of seven latent variables can be observed in Fig. 6, with R^2_{CUSUM} values lower than the rest. It means that the shape of the CUSUM chart for these variables is very different to the one for the I_{OH}, so they will not be associated to critical points. For the rest of latent variables, a certain positive correlation is observed: the latent variables with higher correlation with the I_{OH} also present, on average, more correlation working with CUSUM values. In Fig. 6, a line has been drawn that leaves 10 latent variables above. The latent variables with higher values for both correlation coefficients require a deeper analysis, and their characteristics are presented in Table 5. In Fig. 6, it can also be observed that the greater values correspond to stage 1, which seems to point to this stage as the likely critical one respect the quality parameter.

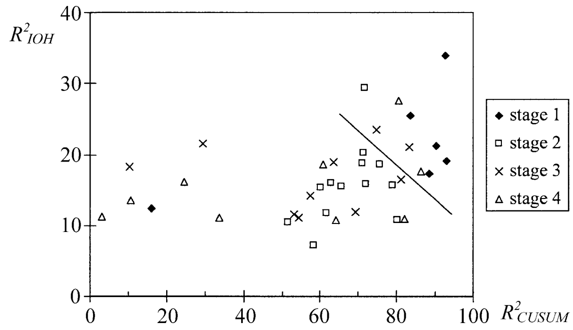

Fig. 6. For every one of the 38 latent variables indicated in Table 4, relationship between the squared correlation coefficient calculated with the hydroxyl index (R^2_{IOH}), versus that coefficient calculated with the cumulated values (R^2_{CUSUM}).

Table 5
Further analysis with the 10 latent variables selected from Fig. 6

Latent variable	Stage[a]	Nv[b]	PC[c]	R_X^{2d}	R^{2e}_{IOH}	*p-value*[f]	R^{2g}_{CUSUM}
1FPcu_1	1	61	1	0.90	0.34	0.0002	0.93
1T[a]_1	1	230	1	0.43	0.21	0.0048	0.90
1PR_2	1	227	2	0.15	0.25	0.0020	0.84
1PRd_3	1	49	3	0.15	0.19	0.0077	0.93
1LEV_2	1	225	2	0.18	0.17	0.0116	0.88
2PR_cu_1	2	134	1	0.62	0.29	0.0005	0.72
3PC_cu_1	3	128	1	0.96	0.23	0.0023	0.75
3T[a]d_2	3	198	2	0.14	0.21	0.0048	0.83
4PR_4	4	294	4	0.06	0.18	0.0096	0.86
4PR_cu_4	4	241	4	0.08	0.28	0.0008	0.81

[a] Batch stage of the trajectory that generated the latent variable.

[b] Number of aligned variables of the trajectory that generated the latent variable.

[c] Number of Principal Components of the trajectory that corresponds to the latent variable.

[d] Variation of the trajectory explained by the component (R^2_X).

[e] Squared correlation coefficient between the latent variable and I_{OH}.

[f] *p-value* of the linear regression between the latent variable and I_{OH}.

[g] Squared correlation coefficient between the latent variable and I_{OH} using cumulated values.

From the information in Table 5, by applying external information of the process it is possible to discard the two latent variables of stage 4, as the polymer is completely formed in that stage. Besides, those latent variables explain little variance, less than 10%. The eight remaining ones will require a further analysis to try to identify the critical points. For that purpose, the CUSUM charts of these eight latent variables, together with the one of the I_{OH}, have been overlapped in Fig. 5, in order to check which chart matches the change of trend with the one for the I_{OH}. Four different changes of trend can be observed, that occur approximately in batches number 15 (variable 2PR_cu_1), 18 (I_{OH}), 23 (3T[a]d_2, 1PR_2 and 1LEV_2) and 26 (3PC_cu_1). These latent variables cited can be discarded, as their change of trend does not coincide with the one of the I_{OH}. For the remaining three latent variables (1PRd_3, 1FPcu_1 and 1T[a]_1), it is not clear if the change of trend occurs in batch 18 or 23, and they will be discussed by applying technical knowledge of the process.

−**1FPcu_1**: First component of the cumulated flow of polyalcohol once the addition finishes. When the control valve closes, the integrator of the flowmeter slowly increases the value of total addition. The correlation detected indicates that the greater value of this cumulated flow, the greater I_{OH}. It is known by technical knowledge that if the amount added of polyalcohol increases, so does the I_{OH}, and that relationship is known. So, the correlation found could be a result of a fault in closing the addition valve. However, according to the process engineers, when the predefined setpoint is reached, the valve closes completely, and even assuming an escape of polyalcohol through the valve, the differential amount registered by the flowmeter would not justify the variation of the hydroxyl index, according to the relationship between both variables.

−**1PRd_3**: Third component of the derivative of pressure during the dehydration of stage 1. This component explains only 15% of the corresponding derivative trajectory. Analysing the weights of the variables in the 3rd component of this trajectory, it results that the variables with higher weights correspond to the first minutes of the vacuum ramp of the dehydration, when the descent of pressure is more steep. According to technical knowledge, it is known that the profile of pressure affects the residual content of water at the end of the dehydration, and that the variation of this content may also affect the hydroxyl index. However, according to the engineers, the residual water content is likely to be related with the pressure at the end of the dehydration, whose importance has not been outlined in the PLS analysis. Although it is not possible to determine if this observed correlation is causal, taking into account that the change of trend matches approximately the one of the I_{OH}, it will be a factor to be considered.

−**$1T^a$_1**: First component of temperature inside the tank during stage 1. This explains nearly half of the variance of the trajectory, and the shape of the CUSUM chart is quite similar to the equivalent chart of the I_{OH}. To obtain more information about this latent variable, the weights of the variables of temperature in stage 1 in the first Principal Component are presented in Fig. 7, and compared with the original values of the trajectory.

It can be observed in the upper part of Fig. 7 that the higher weights in absolute value correspond to the end of the sub-stage 1 (addition of polyalcohol) and all the sub-stage 2 (addition of alkaline solution). The lower part of Fig. 7 presents the trajectories of temperature during stage 1 in original aligned values, for the 37 batches, and those highlighted in thicker lines correspond to the batches with higher values of I_{OH}. The vertical scale has been omitted for confidentiality reasons. Comparing both charts, in the period with higher absolute weights, the batches with higher hydroxyl index present lower values of temperature. If the trajectory of weights in the PCA model (upper part of Fig. 7) is compared with the trajectory of the weights of $1T^a$ in the final PLS model (3rd trajectory of Fig. 4), a similar profile can be observed. The information from both charts is equivalent: The negative weights in the PLS model for the trajectory of temperature imply a negative correlation with the final quality, which can be observed in the original trajectories. The weights in the PCA model reflect the period that accounts for that latent variable correlated with the I_{OH}.

According to technical knowledge of the process, as the hydroxyl index is related with the chemical structure of the polymer, the main causes of its variability are factors affecting the thermodynamics of the reaction, mainly the

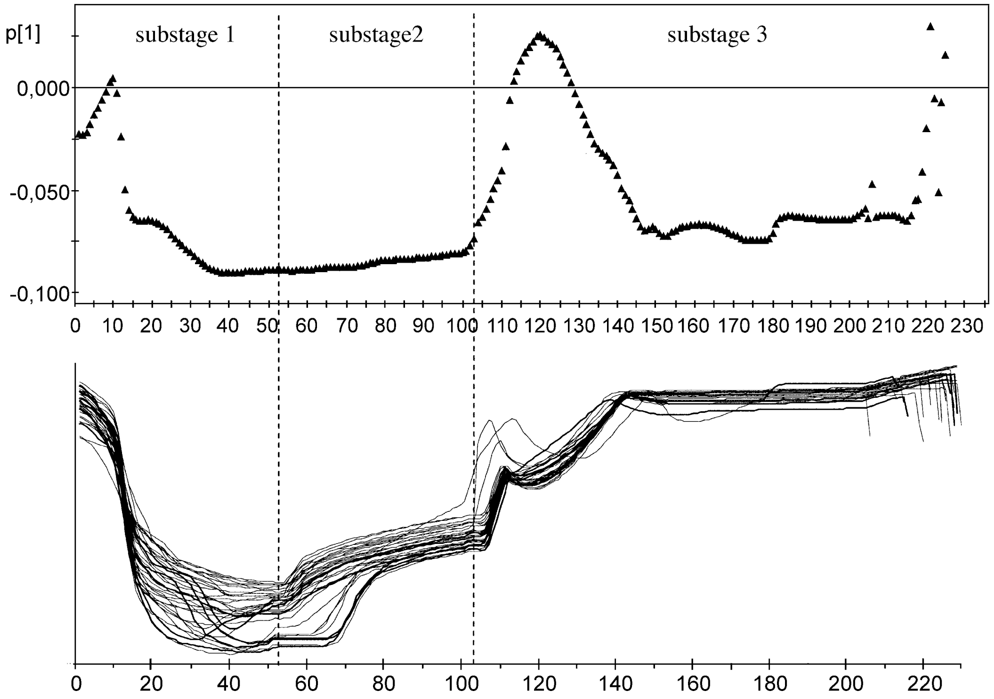

Fig. 7. Correspondence between the weights of the trajectory of $1T^a$ in the first Principal Component, and the aligned original values of $1T^a$ (trajectories for the six batches with higher values of I_{OH} highlighted in thicker lines). Vertical dashed lines show the three sub-stages that comprise this stage.

exact amount of reagents and the likely presence of residual water. The effect of the kinetics of the reaction is unknown, that is, how can the temperature of the alkaline solution affect the I_{OH}. Possibly, the temperature inside the tank during the second sub-stage is correlated with the temperature of the polyalcohol when it enters the tank (which is not registered). It could be that, for any reason, a variation in this temperature causes a measuring error in the flowmeter that controls the addition of polyalcohol, which would explain the variation of the hydroxyl index. However, it is not possible to know if the correlation found is causal, that is, if the variation of temperature during the addition of alkaline solution is responsible for the variability of the I_{OH}, or if there is another cause that provokes the variation of that temperature and the I_{OH} at the same time.

3.3. Further studies proposed

Although the results obtained have not achieved a complete diagnosis of the process, the methodology applied has identified some latent variables as potential critical points, supported by hypotheses based on external information of the process that would justify the variation of the hydroxyl index. To corroborate if the results obtained are consistent, a new set of batches corresponding to a different period of time could be analysed. If we suppose that the main underlying causes of the variability of the I_{OH} remain the same along the time, a consistency study with new batches should lead to similar results, and, at the same time, would allow the selection of the variables more clearly associated to critical points.

Despite the interest of this consistency study, the only way to develop a predictive model based on cause–effect correlation in order to find the optimum operative conditions of the process and to identify definitively the critical points is by means of an experimental design. A DOE planned without having carried out this analysis would not have made sense, due to the high amount of data collected from many process variables. However, now, there are suspicions about a reduced set of process variables associated to potential critical points. Although the DOE proposed has not been achieved yet in the factory, it should be conducted according to the factors selected by the methodology used. Temperature during stage 1 (in particular, during the addition of alkaline solution) has been identified by both approaches. So, it should be one of the factors for the DOE. However, according to technical knowledge, it is supposed to be related with the temperature of the polyalcohol while entering the tank, so this additional factor should also be considered. Another one is the cumulated values of the flow of polyalcohol (1FPcu), identified by both approaches. As commented in Section 2.2, it is known that a variation of the total amount of polyalcohol affects the hydroxyl index. Pressure in stage 1 is another key process variable that both approaches have identified: The original values appear to be important with PLS, and the derivative of the trajectory, with *block-wise* PCR. So, it is recommended to include both factors: the derivative of the pressure during the vacuum ramp and the pressure at the end of the dehydration. Finally, the pressure in stage 2 is another potential critical variable, detected only with the PLS approach. In can be seen in Fig. 4 that the highest weights of this trajectory are reached at the beginning. So, the factor to include in the DOE should be the pressure at the beginning of stage 2.

The implementation of any statistical process control approach comprises an off-line stage for modelling and an ulterior stage of on-line monitoring. This paper deals with the modelling stage, focused on the development of a predictive model of the I_{OH} aimed at diagnosis. Generally speaking, once identified the causal factors that affect the final quality, two situations are possible. In some cases, an action to improve the engineering process control of the critical points would be enough to avoid a wrong final quality. For example, if the causal factor is an inadequate mass of a reagent, the use of a more accurate flowmeter would improve the quality. In other cases, the engineering control does not solve the problem, and it can be supplemented with multivariate on-line models to monitor the critical points, in order to early detect out-of-control situations that can negatively affect the final quality. These models are carried out with the key variables that have causal correlation with the quality or show no correlation but are known to affect the quality if their control fails. In the first case models will be based on PLS, and in the second one, on PCA.

Nevertheless, in chemical processes, some authors [4] recommend the monitoring with PCA models of all the process variables, as potentially any out-of-control situation could affect the composition or chemical structure of the product and have an effect in the quality perceived by consumers, which is not always measured by the quality parameters analysed.

4. Conclusions

Aimed at diagnosing the causes of variability of the hydroxyl index (I_{OH}) of PPOX produced in a chemical factory, process variables from 37 batches have been analysed by applying the methodology of Nomikos and MacGregor [1]. Carrying out a PLS to the unfolded matrix of aligned variables, the first component is significant, but the identification of the key process variables that cause the variability of the I_{OH} (critical points of the process) is not easy, because when analysing observed variability not generated by means of an experimental design (DOE), it is not possible statistically to know if correlation is due to a cause–effect relationship associated to a critical point.

The methodology proposed in this paper to identify the critical points in complex processes with a very high number of variables potentially influencing the final quality, consists of two stages. First, the analysis of the observed data with multivariate statistical methods based on projection to latent structures, supplemented with technical information of the process, in order to screen and detect the potential key process variables. Afterwards, the execution of a DOE with the previously selected potential key variables to identify significant factors and set optimum operative conditions, achieving an improvement of the final quality of the product.

Two different approaches have been applied to analyse the observed data in order to screen the likely key process variables. The first one uses PLS regression with a progressive simplification of models applying a variable selection method based on comparing the trajectory of weights in the PLS model with the original trajectories of the batches and incorporating technical information of the process. This procedure focuses on the analysis of parts of the trajectory with high weights in absolute value (runs of correlation), in order to avoid random correlation. The main drawback of any selection method is the risk of removing variables with causal correlation and the requirement of a deep knowledge of the process. In the end, the final model for this process contains four potential key process variables.

Another approach presented is a simple procedure that has been denominated *block-wise* PCR. From every trajectory of a process variable, a PCA has been carried out to obtain the significant latent variables. Afterwards, those latent variables significantly correlated with the I_{OH} have been selected. After a second pruning process, by using CUSUM charts, three potential causal variables have been selected. They also present a change of trend that approximately matches the one that occurs with the I_{OH}. Although these likely key process variables are included in the group identified with the previous approach, this one has resulted much more straightforward and faster in conducting the analysis, without requiring so much technical information of the process.

Although the causes of variability of the I_{OH} have not been identified yet, some process variables have been pointed out as likely critical points. The only way to finally achieve this diagnosis is with a DOE conducted on these factors. Once those with a significant effect on the final quality have been identified, it will be possible to set the optimum levels and to improve their control reducing the variability around their optimum values or trajectories, in order to force the hydroxyl index to move around the nominal value with minimum variance, producing a quality improvement of the PPOX.

Acknowledgements

This research was supported by the Spanish Government (CICYT) and the European Union (RDE funds) under the projects 1FD 1997-2159 and DPI2001-2749-C02-01. Manuel Zarzo is grateful for a predoctoral fellowship from the *Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria* (INIA). The authors wish to thank Dr. John MacGregor and Dr. Theodora Kourti for their comments on unfolding methods.

References

[1] P. Nomikos, J.F. MacGregor, AIChE J. 40 (1994) 1361–1375.
[2] T. Kourti, J.F. MacGregor, Chemom. Intell. Lab. Syst. 28 (1995) 3–21.
[3] S. Wold, N. Kettaneh, K. Tjessem, J. Chemom. 10 (1996) 463–482.
[4] F. Yacoub, J.F. MacGregor, Chemom. Intell. Lab. Syst. 65 (2003) 17–33.
[5] H.J. Ramaker, E.N.M. Sprang, S.P. Gurden, J.A. Westerhuis, A.K. Smilde, J. Process Control 12 (2002) 569–576.
[6] P. Nomikos, J.F. MacGregor, Technometrics 37 (1995) 41–59.
[7] D.J. Louwerse, A.A. Tates, A.K. Smilde, G.L.M. Koot, H. Berndt, Chemom. Intell. Lab. Syst. 46 (1999) 197–206.
[8] H.A.L. Kiers, J. Chemom. 14 (2000) 105–122.
[9] S. Wold, P. Geladi, K. Esbensen, J. Ohman, J. Chemom. 1 (1987) 41–56.
[10] A.K. Smilde, Chemom. Intell. Lab. Syst. 15 (1992) 143–157.
[11] S. Wold, N. Kettaneh, H. Friden, A. Holmberg, Chemom. Intell. Lab. Syst. 44 (1998) 331–340.
[12] J.A. Westerhuis, T. Kourti, J.F. MacGregor, J. Chemom. 13 (1999) 397–413.
[13] T. Kourti, J. Chemom. 17 (2003) 93–109.
[14] R.A. Harshman, M.E. Lundy, in: H.G. Law, C.W. Snyder Jr., J.A. Hattie, R.P. McDonand (Eds.), Research Methods for Multimode Data Analysis, Praeger, New York, 1984, pp. 216–284.
[15] P.J. Brown, Measurements, Regression and Calibration, Clarendon Press, Oxford, 1993.
[16] J.P. Gauchi, P. Chagnon, Chemom. Intell. Lab. Syst. 58 (2001) 171–193.
[17] J.A. Westerhuis, T. Kourti, J.F. MacGregor, J. Chemom. 12 (1998) 301–321.
[18] L.E. Wangen, B.R. Kowalski, J. Chemom. 3 (1988) 3–20.
[19] T. Kourti, P. Nomikos, J.F. MacGregor, J. Process Control 5 (1995) 277–284.
[20] S. Wold, J. Trygg, A. Berglund, H. Antti, Chemom. Intell. Lab. Syst. 58 (2001) 131–150.