ELSEVIER

Chemometrics and Intelligent Laboratory Systems 44 (1998) 331–340

Chemometrics and intelligent laboratory systems

Modelling and diagnostics of batch processes and analogous kinetic experiments

Svante Wold [a,*], Nouna Kettaneh [b], Håkan Fridén [c], Andrea Holmberg [d]

[a] *RG Chemometrics, Umeå University, S-901 87 Umeå, Sweden*
[b] *MDS, 17 Birch Road, Kinnelon, NJ 07405, USA*
[c] *Umetri, Box 7960, S-907 19 Umeå, Sweden*
[d] *Software Point, Valkjärventie 1, Fin-02130 Espoo, Finland*

Received 8 September 1997; revised 7 August 1998; accepted 7 September 1998

Abstract

In chemical kinetics and batch processes K variables are measured on the batches at regular time intervals. This gives a $J \times K$ matrix for each batch (J time points times K variables). Consequently, a set of N normal batches gives a three-way matrix of dimension ($N \times J \times K$). The case when batches have different length is also discussed. In a typical industrial application of batch modelling, the purpose is to diagnose an evolving batch as normal or not, and to obtain indications of variables that together behave abnormally in batch process upsets. Other applications giving the same form of data include pharmaco-kinetics, clinical and pharmacological trials where patients (or mice) are followed over time, material stability testing and other kinetic investigations. A new approach to the multivariate modelling of three-way kinetic and batch process data is presented. This approach is based on an initial PLS analysis of the $((N \times J) \times K)$ unfolded matrix ((batch $\times$ time) $\times$ variables) with 'local time' used as a single y-variable. This is followed by a simple statistical analysis of the resulting scores and results in multivariate control charts suitable for monitoring the kinetics of new experiments or batches. 'Upsets' are effectively diagnosed in these charts, and variables contributing to the upsets are indicated in contribution plots. In addition, the degree of 'maturity' of the batch can be as predicted vs. observed local time. The analysis of batch data with respect to various questions is discussed with respect to typical objectives, overview and summary, classification, and quantitative modelling. This is illustrated by an industrial example of yeast production.

Keywords: Batch modelling; Multivariate modelling; PLS; PCA; Multivariate control charts

1. Introduction

Batch-wise manufacturing processes are common in all industry, including the chemical, pharmaceutical, bio-technical, and semi-conductor industries. Typical examples include beer brewing, car painting, molding, emulsion polymerization, spray-drying, spray-coating, fermentation, yeast production, and wafer etching. In addition, very similar modelling and data-analytic problems occur in pharmaco-kinetics, clinical and pharmacological trials where patients or mice are followed over time, in material stability testing, and in other kinetic investigations. We shall henceforth refer to the trials in all these applications as 'batches'.

* Corresponding author. Fax: +46-90-13-8835

0169-7439/98/$ - see front matter
PII: S0169-7439(98)00162-2

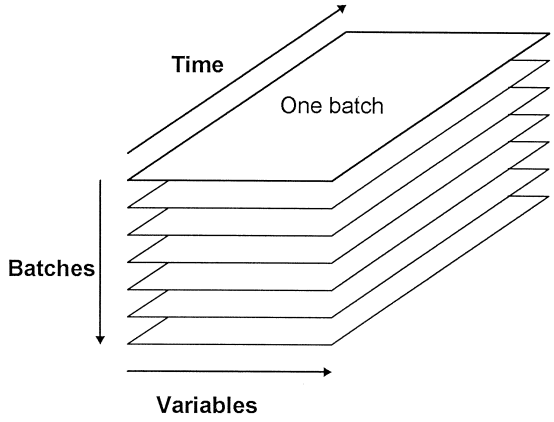

Fig. 1. Three-way table of historical data of, for example, N batches of yeast production, J time points, and K variables. In example 1, there are $N = 23$ normal batches of yeast, $J = 83$ time points (every 10 min for 14 h), and $K = 7$ variables.

A batch starts with the charging of the reactor with starting material (initial conditions are described by the optional vector $\boldsymbol{z}$). The batch is initiated, and then the observation of the first process point is made (a vector $\boldsymbol{x}_{i1}$ with K elements, $\boldsymbol{x}_{i1k}$) The evolution of the new batch (index i) is then followed by the measurements of the same K variables at time 2, 3, ..., until time J, when the batch is terminated, the product collected, the reactor cleaned, etc., followed later by the start of another batch.

Hence, in a batch process where K variables are measured at frequent time intervals ($j = 1, 2, \ldots, J$), each trial results in a data matrix with dimension $J \times K$. A set of N normal batches hence gives a three-way matrix of dimension $(N \times J \times K)$, see Fig. 1. In the case when batches have different length in time and the number of time points of the batches varies, a special alignment of the data is needed as discussed below.

Like with continuous processes, great interest has recently arised to use the multivariate data measured on the process for diagnostic purposes, so called multivariate statistical process control (MSPC) [1,2]. Statistical methods for the analysis of batch process data have recently been used successfully for monitoring and diagnostics [2,3]. These methods are based on multi-way PCA and PLS [4], and are now being widely adopted for multivariate SPC of batch processes in the chemical and semiconductor industries.

The approach of Nomikos and MacGregor [2] is very powerful for the analysis of historical production data, and also shown to be very effective for the monitoring of new evolving batches. However, this approach assumes that data for the complete batch is available, and hence one must use some kind of imputation of the missing data of a new evolving batch. Nomikos and MacGregor described several ways to do this imputation, all of which work well in practice. Alternatively, a recursive approach of updating multi-block PCA of PLS models may be used, as recently described by Rännar et al. [5].

In the present paper we introduce yet another variant of multivariate batch modelling for monitoring and diagnostic purposes. It is based on the unfolding of the three-way training matrix of historical data in the batch direction, giving an $(N \times J) \times K$

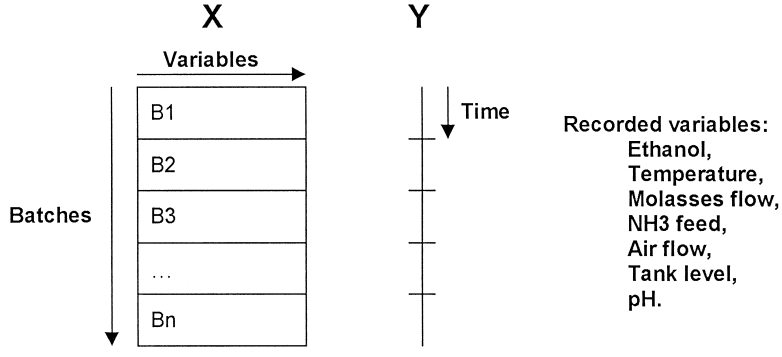

Fig. 2. The three-way matrix of Fig. 1 unfolded along the batch direction to give a two-way matrix with $N \times J$ rows and K columns. Each row has the data ($\boldsymbol{x}_{ijk}$) from a single batch observation (batch i, time j, variable k). In the yeast example, the $K = 7$ variables are $\mathbf{X}1$ = ethanol concentration, $\mathbf{X}2$ = temperature, $\mathbf{X}3$ = Molasses feed, $\mathbf{X}4 = NH_3$, $\mathbf{X}5$ = air flow, $\mathbf{X}6$ = tank level, and $\mathbf{X}7$ = pH.

matrix (Fig. 2). The subsequent multivariate analysis of this matrix provides a natural and straight-forward scheme for the monitoring of new evolving batches by means of ordinary Shewhart charts of their multivariate scores. These are complemented by charts indicating the similarity of the evolving batch to the multivariate model (DModX charts or SPE charts), and charts displaying the maturity of the evolving batch in terms of its predicted value of chronological time. For process points that show deviations from the model, contribution plots indicate which variables that are related to these deviations.

The approach is illustrated by an example of baker's yeast production at Jästbolaget in Solna, Sweden.

2. Multivariate batch monitoring

2.1. Data and objectives

As in all empirical modelling, there are *two phases* in the data analysis. In the first phase (*the learning or training phase*) one uses a training set of existing historical data to develop the model. Here the training set consists of data from a representative set of normal ('good') batches. Hence, the result is a model of normal batches showing the normal pattern of evolution, which variables are well modelled, their correlation patterns, the presence of outliers, etc.

In the second phase (*prediction (monitoring) phase*), the model is used to diagnose new evolving batches as 'normal' or 'not normal', and further to indicate which combinations of variables that are related to the 'non-normality' of a deviating new batch.

2.2. Existing approaches

Multivariate batch modelling was pioneered by Nomikos and MacGregor [2,3], who developed an approach based on unfolding the three-way training matrix (Fig. 1) so that the batch direction is preserved. This gives a two-way matrix of dimension $(N \times (J \times K))$, i.e., with N rows and $(J \times K)$ columns. This matrix is then modelled by PCA or PLS, giving a multivariate model of 'normal' batches. The three-way data is centered and scaled before the unfolding, and hence the multivariate analysis models the variation around the average trajectory of each variable.

A new batch can then be fitted to this model, and the standard multivariate process diagnostics calculated (scores, Hotelling's T^2, and the residual standard deviation or variance). If the new batch is to be diagnosed before its completion, however, one must in some way fill in the values of the data that are not yet measured, i.e., from the present batch time to the end [2,3].

A second approach based on recursive updating of a hierarchical PCA model was recently published [5]. Here each time slice of the three-way matrix is treated as a block in a hierarchical model [4,6]. This model is calculated by a recursive updating, starting from the first time slice. The result is a PCA model calculated in such a way that for an incomplete (evolving) new batch no filling in of later missing values is needed. A limited comparison between the two approaches indicates that the hierarchical approach works as well, but needs fewer model dimensions.

2.2.1. Present approach

The present approach is based on unfolding the three-way training matrix to preserve the direction of the variables (see Fig. 2). The resulting matrix has $n_{\text{obs}} = N \times J$ observations (rows) and K columns. Hence this matrix, **X**, has the individual observation as a unit in contrast to the unfolded matrix of previous approaches which have the whole batches as units (matrix rows). A dummy y-variable is constructed which is either the 'local batch time' or the 'batch maturity index'.

The data are centered and scaled, usually to unit variance. We note that this gives a different centering and scaling than used in the previously described approaches, where the average trace was removed by the centering [2,3,5]. This average trace will now instead be described by the score values of the first component as shown below.

The complete analysis of the batch data are made in a number of steps (levels), relating to (a) the individual observations, (b) the evolution of the batches, and (c) the whole batches.

2.2.1.1. Modelling the individual observations (level 1). A PLS model [8,9,11] is developed for the centered and scaled **X** and 'local batch time', y. Here

care is necessary to include sufficiently many components (A) so that much of the **X**-matrix is 'explained'. This results in an ($n_{obs} \times A$) score matrix, **T**, plus PLS weight and loading matrices **W** and **P** (of dimensions $K \times A$). The PLS model is:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \quad (1a)$$

$$y = \mathbf{T}c + \boldsymbol{f} \quad (1b)$$

Here the matrix **E** and the vector $\boldsymbol{f}$ contain the residuals, 'left-overs', after the modelling. The scores **T** are calculated to both well approximate (summarize) **X** according to Eq. (1a), and predict y according to Eq. (1b).

The score matrix **T** is a good summary of **X**, and PLS focuses this summary on y (local time). Hence the first column of $\boldsymbol{t}$ ($\boldsymbol{t}_1$) will contain strong contributions of those **X**-variables that vary monotonously with y, i.e., either increase or decrease with time. The second component ($\boldsymbol{t}_2$) is an aggregate mainly of those variables that change 'quadratically' with time, i.e., first go up to a maximum, and then decrease, or first go down, and then up. The third component catches the variables having a cubic behaviour, etc.

The value of 'local time' predicted by the PLS model, y_{pred}, is very suitable as a 'maturity index' that can be used to indicate how far the batch has evolved, if it is ready for termination, etc. In the case that a maturity variable is available, y_{pred} will contain a strong contribution from this variable, but also from other variables correlated with this maturity variable. Hence y_{pred} may still be a better alternative for a maturity index since it is smoother and has contributions from a wider spectrum of variables.

In the case that the batches are of widely different length (different number of time points), the resulting y_{pred} can now be used to re-express each batch in terms of interpolated data so that each batch thereafter has the same number of rows. Thus the data are re-digitized at regular intervals of y_{pred} for instance at 0%, 5%, 10%, ... of completion $= \boldsymbol{t}_{max}$ instead of previously at regular time intervals.

A second analysis of the same type is then done with the aligned data, y now being the value of the maturity index, typically the 'aligned' y_{pred} from the first analysis.

Instead of $y =$ 'local time', one can use a continuously measured maturity variable such as the alcohol content of beer, the average molecular weight in a polymerization, or the lignin content (kappa number) in a pulp cooking batch. This makes the present approach flexible and adaptable to most batch processes.

2.2.1.2. Batch trace level (level 2). The purpose at this level is to calculate the typical evolution trace of a normal, 'good' batch. This is done as follows. First the scores of the level 1 model (the observation level) are chopped up and reorganized so that the scores of one batch form one row vector ($\boldsymbol{t}_1$ followed by $\boldsymbol{t}_2$, followed by $\boldsymbol{t}_3$, etc.) in a matrix $\mathbf{X}_T$. This matrix has N rows (one per batch) and $A \times J$ (AJ) columns from the A score vectors and the J 'time points' per batch.

Now the matrix $\mathbf{X}_T$ is used to derive minimum and maximum tolerated values at each time point (j) of $\boldsymbol{t}_{j1}$, $\boldsymbol{t}_{j2}$, $\boldsymbol{t}_{j3}, \ldots,$ $\boldsymbol{t}_{jA}$ from the column averages and standard deviations of $\mathbf{X}_T$, as indicated in Fig. 3. In MSPC one would typically use tolerance intervals of the average $\pm$ 3 SD at each time point. Hence, we now have for each score vector an average trace with

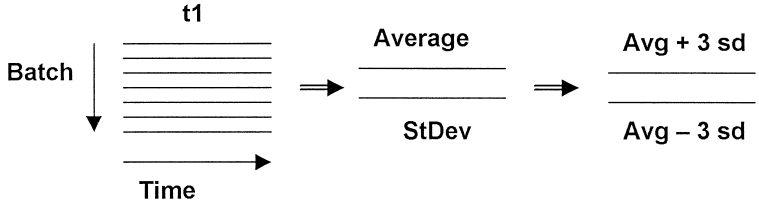

Fig. 3. The score values for each batch are arranged as row vectors under each other, giving an $N \times (J \times A)$ matrix, $\mathbf{X}_T$. Above, we see only the first part of $\mathbf{X}_T$ corresponding to the 1st component (t_1). From this matrix on calculates the averages and standard deviations (SDs) of the matrix columns, and then control intervals as the averages $\pm$ 3 SD.

upper and lower tolerance limits (Fig. 4). This specifies the normal evolution traces, one for each component score, of new batches. Indeed we see that all normal batches closely follow these normal traces (Fig. 5).

The matrix $\mathbf{X}_T$ can also be used to derive sharper bounds for new batches taking the auto-correlation structure of the scores into account. This can be made by subjecting $\mathbf{X}_T$ to a principal components analysis (PCA), giving a model of the normal variation of the combined scores. For diagnosis of new batches, one can develop several models, e.g., one for the first 10% of the batch evolution, one model for the first 20%, etc., until a final model for the whole 100%. Alternatively, the adaptive approach of Rännar et al., can be applied to the score matrix $\mathbf{X}_T$, giving an adaptively updating PCA model.

An additional trace can be constructed from the values of predicted y (local time or maturity) for each time point of each batch. These predicted values result from the PLS model when observations of a new batch are plugged into the model. The predicted y-values should be fairly close to the 'real' y-value for a batch evolving at the 'normal' rate. Forming a matrix $\mathbf{X}_Y$ with the traces of predicted y of each batch as rows, and then analyzing it in the same way as $\mathbf{X}_T$ above, gives tolerance intervals of y_{pred} for each time point (Fig. 4).

2.2.1.3. Batch level (top level = level 3). The objective is to make a model of the whole batch. This model will be based on (when available) the initial conditions (Z), the evolution trace matrix $\mathbf{X}_T$, and the properties and quality of the completed batch ($\mathbf{Y}$). We

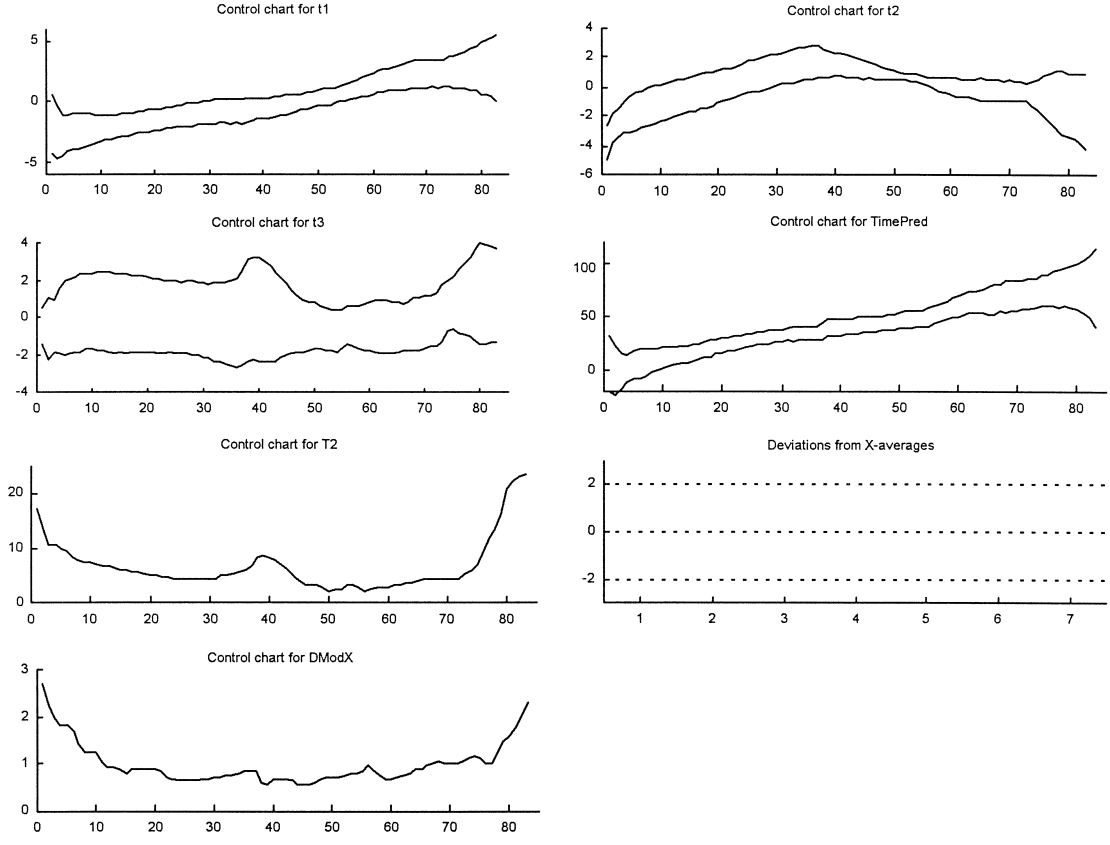

Fig. 4. Resulting control intervals for the yeast data of t_1, t_2, t_3, y = time, Hotelling's T^2, and the residual standard deviation (DModX in Simca-P).

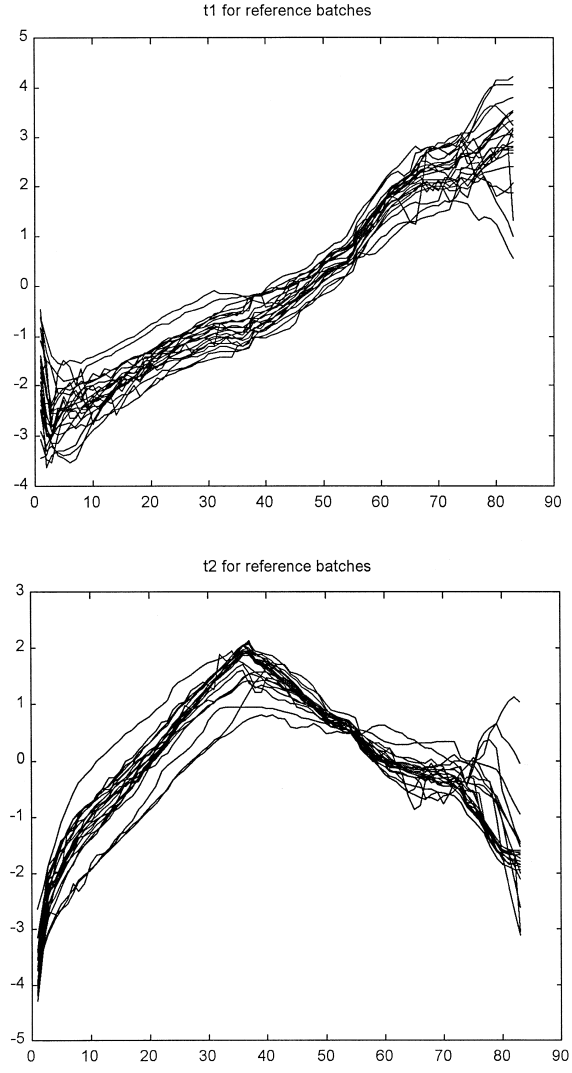

Fig. 5. The traces of the two first scores, t_1 and t_2, of the 23 normal yeast batches.

wish to understand how **Y** is influenced by the combination of initial conditions and the batch evolution. Either one can use ordinary PLS with Z and $\mathbf{X}_T$ together forming the predictor matrix, or a hierarchical PLS approach [4].

The resulting model can be used to predict **Y** for new batches from their initial conditions and their evolution traces. This may be advantageous in the case that the measurement of **Y** is laborious and time consuming, and one wishes to make an initial fast quality assessment before the real **Y** is available. This is often the case in biotechnical production, where extensive analysis and testing of the final products is made before acceptance (or rejection).

The level 3 model also allows the full interpretation of the batch data. Groups of batches may be discovered, outliers will be found, etc. Critical time periods will be indicated by 'periods' of large weights, loadings, and VIP values. Analogously, important factors in the initial conditions will stand out. The 'ideal' evolutionary trace corresponding to a desirable **Y**-profile can be deduced, and so on.

In the case that **Y**-values do not exist (as in the case of the yeast batches), the **X**-matrix can still be used to develop a PC model [10,11]. Thus, the scaled and centered **X** is modelled as (**E** is the matrix of residuals).

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \tag{2}$$

We see that this is the same type of model as the **X**-part of PLS (Eq. (1a)). The difference is that in PCA the scores (**T**) are calculated to give an optimal summary of **X**, while in PLS this optimality is somewhat relaxed to make **T** better predictors of **Y**.

This PC model can then be used to classify new batches as normal (similar = well fitting to the PC model) or non-normal (far from the model). This corresponds to the asymmetric SIMCA classification described by Dunn and Wold [7]. Such classification may be advantageous if the acceptance/rejection testing is laborious and time consuming, as with wafers in the semi-conductor manufacturing. For process diagnostic purposes a quick but imprecise classification may be sufficient to judge if a new batch should be started, or the process should be halted for maintenance. This simpler diagnostic is, of course, less ambitious than the final classification of a finished wafer as acceptable or not.

The preliminary classification of a new batch is made as follows. One first subtracts the same centering vector as was used for **X** from the batch vector $\boldsymbol{x}_{\text{new}}$ Then the resulting vector is multiplied by the same scaling weights as for **X**, and then one calculates the score values for the new vector as (**P** contains the loadings of model 2):

$$\boldsymbol{t}_{\text{new}} = \boldsymbol{x}_{\text{new}}\mathbf{P} \tag{3}$$

Then residuals for the new vector are calculated as:

$$\boldsymbol{e}_{\text{new}} = \boldsymbol{x}_{\text{new}} - \boldsymbol{t}_{\text{new}}\mathbf{P}' \tag{4}$$

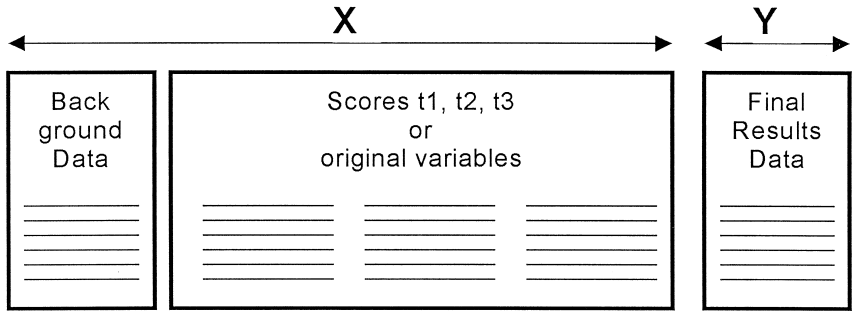

Fig. 6. The data for level 3 (batch level) modelling. Each row has the data of one batch. The left matrix contains data describing initial conditions. The middle matrix contains the unfolded scores which describe the evolution of each batch. The optional right matrix (**Y**) contains the responses, the properties of the complete batch such as yield, purity, activity, etc.

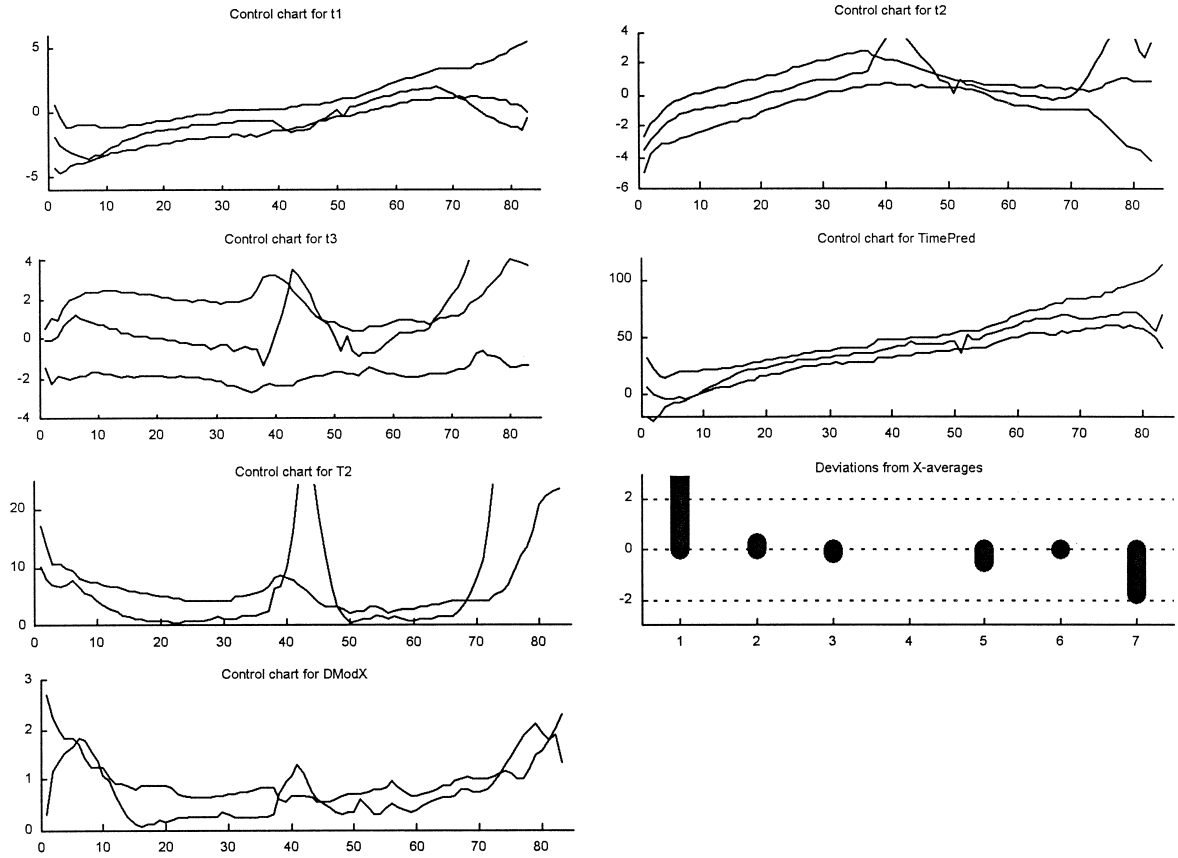

Fig. 7. The traces of a bad yeast batch going out of control at time point 6 (see DModX, lower left) due to a pH problem. It also went outside the control limits around time 38 and time 67 due to too much ethanol. The contribution plot (lower right) is for the last time point (no. 83) and indicates too high ethanol (**X**1) and too low pH (**X**7).

The standard deviation of these residuals (RSD) is then finally calculated (here A is the number of components in the PC model and M is the number of columns in $\mathbf{X}$)

$$\text{RSD} = \text{DModX} = \boldsymbol{e}_{\text{new}}\boldsymbol{e}'_{\text{new}}/(M-A) \quad (5)$$

This RSD (distance to the model, DModX) is a measure of the dissimilarity of the new batch to the PC model, and thereby to the training set of normal batches. Its square is approximately F-distributed, and can be used to calculate the probability that the new batch belongs to the same population as the previous set of normal batches [7,10,12]. Hence a critical level of DModX corresponding to, say, 5% probability can be calculated, and batches having residuals giving a DModX larger than this value are considered as non-normal, non-acceptable.

The PC model of $\mathbf{X}_{\text{T}}$ of the 23 normal yeast batches gave three significant components, modelling together $R^2 = 0.69$ of the variation in $\mathbf{X}_{\text{T}}$, and with a $Q^2 = 0.53$. The individual residual standard deviations (RSD) of the batches with respect to this PC model are shown in Fig. 8 together with the RSD values of the four 'bad' batches.

2.2.2. Phase 2. Monitoring of new batches

The monitoring of a new batch starts on the lower observation level. A new batch is started. From the vector of its initial conditions ($\boldsymbol{z}_{\text{new}}$), and the 0% level 2 model, an initial estimate can be obtained of how 'normal' this initial vector is.

Then the batch is started, and after one time unit the first $\boldsymbol{x}$-vector is recorded. It is inserted into the level 1 PLS model, giving predicted values of the score values of this observation, $\boldsymbol{t}_1$ to $\boldsymbol{t}_a$. With more than two model dimensions, these score values can be combined into a Hotelling's T^2. In addition, the resulting $\boldsymbol{x}$-residuals can be used to calculate a measure of closeness to the model (DModX = residual SD, or SPE = residual sum of squares). These results can now be plotted in the appropriate control charts (Fig. 6), with limits derived as described above (trace level). These charts indicate whether the batch is starting normally or not. If the values are outside the normal ranges, contribution plots based on the $\boldsymbol{x}$-values or the residuals indicate which variables together are related to the deviations.

In addition, the PLS model gives a predicted value of y (local time or maturity index). Plotting y_{pred} vs. y_{obs} gives a very interesting indication whether the batches are developing too fast (over-mature, $y_{\text{pred}} > y_{\text{obs}}$) or too slowly (under-mature, $y_{\text{pred}} < y_{\text{obs}}$).

The same procedure is repeated after the second time points, then the third, etc., and one is able to follow the evolution of the batch in the trace plots (Fig. 7).

At regular intervals, say 10%, 20%, etc., of the evolution, the resulting score vectors can be rearranged and inserted into the partial level 3 models to infer whether the evolution pattern is consistent with normal behavior, and also get very preliminary estimates of the final quality values, $\mathbf{Y}$.

2.2.3. Level 3, complete batch data

Once the batch is finished its initial condition vector ($\boldsymbol{z}$) and score vectors can be combined to form an $\boldsymbol{x}$-vector that describes the background and evolution of the completed batch. This vector can be inserted in the level 3 model to predict values of the response variables ($\mathbf{Y}$), or—as in the present example when $\mathbf{Y}$-variables are not available—to indicate the similarity of the new batch to the training set of normal batches.

In the yeast example, the four test batches (classified as not acceptable) indeed were far from the

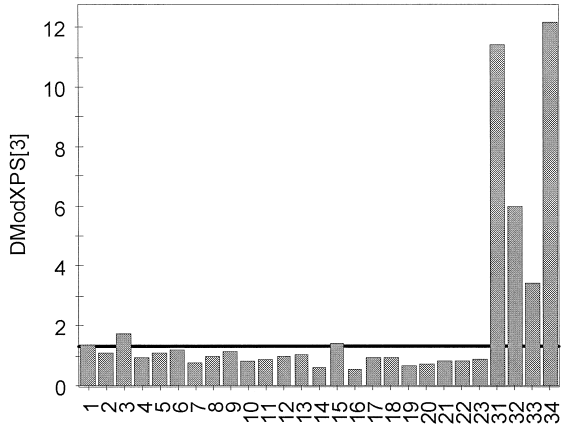

Fig. 8. The resulting residual standard deviation (DModX, Eq. (5)) from the level 3 PC model of the 23 normal and the four 'bad' yeast batches (nos. 31–34). The horizontal line at 1.4 indicates the critical level (5%) for 'non-normality'.

level 3 PC model. This is seen in Fig. 8, which shows the resulting residual standard deviation (DModX) from the level 3 PC model of the 23 normal and the four 'bad' batches. The latter are indeed far outside the 'normal' values and hence clearly are flagged as 'bad'.

3. Conclusions and discussion

We here describe a new approach to multivariate batch process modelling and monitoring. On its primary level it is more oriented to the monitoring of the individual time points, and thereby more focused on following the evolution of the batch, than is the approach of Nomikos and MacGregor. The latter is more focused on classifying the whole completed batch as normal or not. With the resulting scores of a completed batch, the present approach also gives a model of the whole batch, but using a compressed version of the data (the scores) instead of using all the data as Nomikos and MacGregor. This saves computer time and space when very many variables are monitored and hence may be advantageous for complicated batch processes. With few monitored variables, the complete batch analysis can be based on the on these original variables, thereby making the 'upper level' of the present approach equivalent to the approach of Nomikos and MacGregor.

The multivariate approaches to batch monitoring bring the methods of statistical control charting to apply also to the area of batch processes. This provides tools for improving quality and lowering costs due to earlier detection of faults and upsets. A majority of existing processes still are run in batch mode and these methods hence have wide applicability.

4. List of symbols

The table shows the notation. This is similar to the one used by Nomikos and MacGregor [2], but slightly changed to more closely correspond to standard notation used in chemometrics.

Index	Limits	Meaning
i	$1, \ldots, N$	Batch index
j	$1, \ldots, J$	Time index
k	$1, \ldots, K$	Variable index of $\mathbf{X}$
l	$1, \ldots, L$	Index of background variables ($\mathbf{Z}$)
m	$1, \ldots, M$	Index of response variables ($\mathbf{Y}$)
n_{obs}		Total number of observations (rows) in $\mathbf{X}$, $N \times J$
a	$1, \ldots, A$	Component index

Matrix	Elements	Meaning
$\mathbf{X}$	x_{ijk}	Value of variable k at time j in batch i
$\mathbf{Y}$	y_{im}	Value of response variable m of batch i
$\mathbf{Z}$	z_{il}	Value of background variable l of batch i
$\mathbf{T}$	t_{ija}	Scores (PLS or PCA), at time j in batch i
$\mathbf{W}$	w_{ka}	Weights (PLS), variable k, component a
$\mathbf{P}$	p_{ka}	Loadings (PLS or PCA), variable k, component a

Acronym	Meaning
MSPC	Multivariate statistical process control (Checking)
PCA	Principal components analysis
PLS	Projection to latent structures using partial least squares

Acknowledgements

SW gratefully acknowledges support from the Swedish Natural Science Research Council (NFR).

References

[1] J.V. Kresta, J.F. MacGregor, T.E. Marlin, Multivariate statistical monitoring of process operating performance, Can. J. Chem. Eng. 69 (1991) 35–47.
[2] P. Nomikos, J.F. MacGregor, Multivariate SPC charts for monitoring batch processes, Technometrics 37 (1995) 41–59.

[3] P. Nomikos, J.F. MacGregor, Multi-way partial least squares in monitoring batch processes, Chemometrics Intell. Lab. Syst. 30 (1995) 97–108.
[4] S. Wold, P. Geladi, K. Esbensen, J. Öhman, Multi-way principal components and PLS-analysis, J. Chemometrics 1 (1987) 47–56.
[5] S. Rännar, J. MacGregor, S. Wold, Adaptive Batch Monitoring using Hierarchical PCA, Submitted to Chemometrics and Intell. Lab. Systems, 1997.
[6] S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multi-block PLS and PC models, for easier interpretation, and as an alternative to variable selection, J. Chemometrics 10 (1996) 463–482.
[7] W.J. Dunn III, S. Wold, Structure–activity analyzed by pattern recognition: the asymmetric case, J. Med. Chem. 23 (1980) 595–599.
[8] S. Wold, A. Ruhe, H. Wold, W.J. Dunn III, The collinearity problem in linear regression. The partial least squares approach to generalized inverses, SIAM J. Sci. Stat. Comput. 5 (1984) 735–743.
[9] S. Wold, E. Johansson, M. Cocchi, PLS—Partial least-squares projections to latent structures, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design; Theory, Methods and Applications, ESCOM Science Publishers, Leiden, Holland, 1993.
[10] J.E. Jackson, A User's guide to principal components, Wiley, New York, 1991.
[11] A. Höskuldsson, Prediction Methods in Science and Technology, Vol. 1, Thor Publishing, Copenhagen, 1996, ISBN 87-985941-0-9.
[12] Simca-P, version 3.0, Multivariate process modelling software (MS Windows), Umetri, Box 7960, S-907 19 Umeå, Sweden.