JOURNAL OF CHEMOMETRICS
J. Chemometrics 2000; **14**: 335–349

Standardized Q-statistic for improved sensitivity in the monitoring of residuals in MSPC

Johan A. Westerhuis, Stephen P. Gurden and Age K. Smilde*

Process Analysis and Chemometrics, Department of Chemical Engineering, University of Amsterdam, Nieuwe Achtergracht 166, NL-1018 WV Amsterdam, The Netherlands

SUMMARY

This paper presents the standardized *Q*-statistic for monitoring residuals of latent variable models in multivariate statistical process control (MSPC). Before the summation of the squared residuals, they are scaled according to their expected variation obtained from normal operating conditions (NOC) data. Data from a simulated batch process and from an industrial batch process are used to show that this scaling improves the sensitivity of the *Q*-statistic considerably. The standardized *Q*-statistic is introduced for the off-line monitoring of batch processes, but it can also be used for the monitoring of continuous processes as well as for the on-line monitoring of batch processes. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: fault detection; process monitoring; Q-statistic

1. INTRODUCTION

In multivariate statistical process control (MSPC), empirical models are used to describe data available from chemical processes. These models are developed from process data obtained under normal operating conditions (NOC). The models divide the process data (**X**) into a systematic part ($\mathbf{TP}^{\mathrm{T}}$) and a residual part (**E**) which is not described by the model:

$$\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

The NOC process data only consist of common cause variation, i.e. variation in the process that is not due to a disturbance. Confidence limits can be developed around the common cause variation for both the systematic part of the variation and also the residual part. If new data from the same process are obtained, they should fall within these limits. In the case of a process disturbance the limits will be violated.

The approach of using empirical latent variable models for the monitoring of chemical processes was introduced by Kresta *et al.* [1] for continuous processes and by Nomikos and MacGregor [2] for batch processes. Several successful applications have been reported in the literature [1–8]. The

* Correspondence to: A. K. Smilde, Process Analysis and Chemometrics, Department of Chemical Engineering, University of Amsterdam, Nieuwe Achtergracht 166, NL-1018 WV Amsterdam, The Netherlands.
E-mail: asmilde@its.chem.uva.nl
Contract/grant sponsor: Netherlands Technology Foundation (STW).

Received 31 August 1999
Accepted 29 November 1999

systematic part of the process data, which is described by the process model, is monitored using a *D*-statistic chart. The *D*-statistic, which is similar to the Hotelling T^2 statistic, is a Mahalanobis distance between the new data and the center of the NOC data. Control limits of this statistic were given by Tracy *et al.* [9]. For monitoring the size of the residuals, the *Q*-statistic is used, which is a summation of the squared residuals of a specific batch. Confidence limits for the *Q*-statistic were estimated by Box [10] and Jackson and Mudholkar [11]. A detailed description of control limits for the *Q*-statistic can be found in Reference [2]. However, Nomikos [12] concluded that for the *Q*-statistic, confidence limits of 99·9% should be used in order not to obtain too many false alarms. Therefore Louwerse and Smilde [13] suggested projecting each of the NOC batches onto a model developed using the remaining NOC batches. Residuals from such projections better mimic the residuals of new batches that are projected onto the final model, and confidence limits of 99% are sufficient.

In the original approach of the *Q*-statistic, all squared residuals are summed without taking into account the relative size of the residuals with respect to their expected variation. In this paper it will be shown that the sensitivity of the *Q*-statistic can be improved considerably by considering the residuals relative to their expected variation. This is comparable to the use of standardized residuals for diagnostic purposes [14]. The new *Q*-statistic is therefore called the standardized *Q*-statistic. The expected variation can be obtained from the residuals of the NOC data. The improvement of the *Q*-statistic by scaling the residuals according to their expected variation is introduced for the off-line monitoring of batch processes. Two different batch data sets will be used to show that the *Q*-statistic becomes more sensitive when the residuals are standardized. The standardized *Q*-statistic can easily be extended to work also in the on-line monitoring of batch processes and in the monitoring of continuous processes. This is described at the end of Section 2.

2. THEORY

In MSPC, empirical models are used to describe data available from chemical processes. In a batch process, *J* process variables are measured for *K* time periods of *I* different batch runs. The measurements can be arranged in a three-way array $\mathbf{X}$ ($I \times J \times K$). For convenience, in this paper the three-way array of batch processes will always be considered matricized to a matrix $\mathbf{X}$ ($I \times JK$) where the batch direction is maintained [15].

Using a set of *I* different batches obtained under normal operating conditions (NOC), an empirical model is developed to describe the data as well as possible. The general form of this model was already shown in Equation (1), where $\mathbf{X}$ contains the process data, $\mathbf{TP}^{\mathrm{T}}$ is the model that contains the systematic part of the common variation within the NOC data, and the residuals $\mathbf{E}$ contain the part not described by the model. $\mathbf{T}$ ($I \times R$) describes the differences between the batch runs, and $\mathbf{P}$ ($JK \times R$), describes the similarities among the batch runs. Any structure may be applied to $\mathbf{P}$, e.g. a PARAFAC, Tucker3 or Tucker1 structure. The number of components *R* is usually much smaller than *I* and *JK*.

From the process model, two types of statistics with known distributions are calculated. These are the *D*-statistic for the systematic part of the process variation and the *Q*-statistic for the residual part of the process variation. These statistics are used for statistical process control, to monitor whether new batches are still in statistical control. Using the distributions, confidence limits for the two statistics can be obtained. For monitoring new batches, the process data of the new batch, $\mathbf{x}_{\mathrm{new}}$ ($JK \times 1$), are projected onto the model:

$$\mathbf{x}_{\mathrm{new}}^{\mathrm{T}} = \mathbf{t}_{\mathrm{new}}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}} + \mathbf{e}_{\mathrm{new}}^{\mathrm{T}}, \qquad \mathbf{t}_{\mathrm{new}}^{\mathrm{T}} = \mathbf{x}_{\mathrm{new}}^{\mathrm{T}}\mathbf{P}(\mathbf{P}^{\mathrm{T}}\mathbf{P})^{-1}, \qquad \mathbf{e}_{\mathrm{new}}^{\mathrm{T}} = \mathbf{x}_{\mathrm{new}}^{\mathrm{T}} - \mathbf{t}_{\mathrm{new}}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}} \tag{2}$$

The new scores $\mathbf{t}_{\mathrm{new}}$ ($R \times 1$) and the new residuals $\mathbf{e}_{\mathrm{new}}$ ($JK \times 1$) are used to calculate a new *D*-statistic value and a new *Q*-statistic value respectively for monitoring the new batch.

Both the *D*-statistic and the *Q*-statistic have a different distribution for the batches in the training set than for the batches in the test set. The process model is developed to optimally fit the batches in the training set, where the batches in the test set are projected onto this model. Therefore the *D*-statistic and *Q*-statistic are usually larger for batches in the test set than for batches in the training set.

In order to correct for this difference, Tracy *et al.* [9] used different distributions for the confidence limits of the *D*-statistic. For the batches in the training set a β-distribution was used for the confidence limits of the *D*-statistic, which are lower than the confidence limits of the *F*-distribution that is used for the test set. The *D*-statistic $D = (\mathbf{t} - \bar{\mathbf{t}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{t} - \bar{\mathbf{t}})$ follows an *F*-distribution if the scores $\mathbf{t}$ are independent of the mean scores $\bar{\mathbf{t}}$ and the covariance matrix $\mathbf{S}$ of the scores. This is true for batches from the test set but not for batches of the training set. Therefore a β-distribution is used for the confidence limits of the *D*-statistic of the batches in the training set.

For the *Q*-statistic the difference between training set and test set is not taken into account. Nomikos and MacGregor [2] used the same χ^2 distribution for the confidence limits of the *Q*-statistic of batches from the training set and also of batches from the test set. However, the confidence limits for the *Q*-statistic obtained using the training set do not represent the *Q*-statistic of batches in the test set. Industrial experiences show that a significance level of 99·9% should be used in order not to have too many false alarms [12]. Instead of using different distributions for the *Q*-statistic of the NOC batches in the training set and the *Q*-statistic of new batches in the test set, a different approach can be used.

In order to better mimic the *Q*-statistic of new batches, Louwerse and Smilde [13] suggest projecting each NOC batch in the training set onto a model developed from the remaining $I - 1$ NOC batches in the training set. The projected *Q*-statistics are usually larger than the original *Q*-statistics that are obtained from batches that are also used in the model. Therefore, when using the projection approach, confidence limits of 99% are sufficient. There is however a small difference between the NOC residuals and new residuals. The NOC batches are projected on a model developed from $I - 1$ NOC batches, whereas new batches are projected on a model developed from I NOC batches. This difference can be ignored if I is large. The projection approach to determine *Q*-statistics from NOC batches and confidence limits for the *Q*-statistic will be called the ordinary *Q*-statistic and will be considered the common approach for the remainder of this paper. The ordinary *Q*-statistic will be compared to the standardized *Q*-statistic.

2.1. Q-statistic

The *Q*-statistic is used to compare residuals of new batches to a control limit defined using a set of residuals obtained from batches that were run under normal operating conditions. If $\mathbf{e}_{\mathrm{new}}$ is a vector of residuals of a new batch run, these residuals are usually well described by a multinormal distribution with mean zero and covariance matrix $\mathbf{\Sigma}$, $\mathbf{e}_{\mathrm{new}} \sim \mathbf{N}(0,\mathbf{\Sigma})$. The quadratic form of these residuals is well approximated by a weighted χ^2 distribution $g\chi^2_h$, where the weight g and the degrees of freedom h are both functions of the eigenvalues of $\mathbf{\Sigma}$ [10]. Sometimes g and h are called the matching moments of the distribution. The *Q*-statistic for the residuals of a new batch $\mathbf{e}_{\mathrm{new}}$ is thus defined as

$$Q_{\mathrm{new}} = \sum_{jk=1}^{JK} (e_{\mathrm{new},jk})^2 \sim g\chi^2_h \tag{3}$$

Nomikos and MacGregor [2] describe several ways to determine the confidence limits for the *Q*-statistic from the residuals $\mathbf{E}$ ($I \times JK$) of the batch runs obtained under NOC. In the present paper the Jackson and Mudholkar [11] approximation is used. This approach uses a normal distribution to

approximate the χ^2 distribution of the squared residuals:

$$Q_{\lim,\alpha} = \theta_1 \left[1 - \theta_2 h_0 \left(\frac{1 - h_0}{\theta_1^2} \right) + \frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} \right]^{1/h_0} \tag{4}$$

where $h_0 = 1 - 2\theta_1\theta_3/3\theta_2^2, \theta_1 = \mathrm{tr}(\mathbf{V})$, $\theta_2 = \mathrm{tr}(\mathbf{V}^2)$ and $\theta_3 = \mathrm{tr}(\mathbf{V}^3)$. $\mathbf{V}$ is the covariance matrix of $\mathbf{E}$, and z_α is the standard normal variable with $1 - \alpha$ confidence limit, having the same sign as h_0.

2.2. Scaled residuals

In the ordinary *Q*-statistic as defined in Equation (3), each residual $e_{\mathrm{new},jk}$ is given equal weight, i.e. in the summation of all squared residuals, leading to the *Q*-statistic, the expected value of $(e_{\mathrm{new},jk})^2$ is assumed equal, $E(e^2_{\mathrm{new},jk}) = \sigma^2$ for all pairs of *j*, *k*. However, in many cases this will not be true. Figure 1 shows control limits based on the standard deviation of the residuals of the process data. These data are obtained from a simulated semi-batch emulsion copolymerization of styrene and butadiene. These batch process data will be described further in Section 3. The residuals are from 48

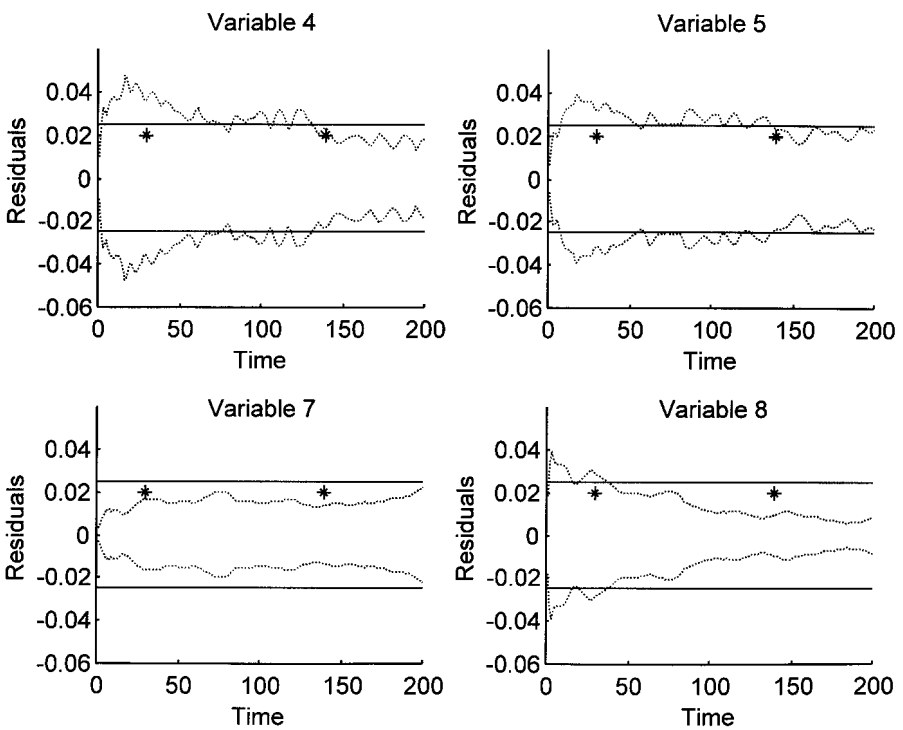

Figure 1. Control limits for residuals of simulated batch data. Dotted curves represent 2·57 times standard deviation of residuals of each separate process variable at each time period. Solid lines represent 2·57 times standard deviation of all residuals. Asterisks represent simulated residuals of size 0·02 at times 30 and 140 for each process variable.

batch runs obtained under normal operating conditions. The residuals of four process variables (4, 5, 7 and 8) of the simulated semi-batch process are shown after a three-component PARAFAC model was fitted to the process data. These four process variables, which describe the temperature of the reactor and the cooling system as well as the density of the latex and the conversion rate, are presented because they are the ones in which the disturbances of erroneous batches are detected (see Section 3·1). The dotted curves represent control limits based on 2·57 times the standard deviation of the residuals of each separate process variable at each time period, which corresponds to a confidence value of 99%. This shows that the standard deviation of the residuals is not the same for each process variable or for each time period. The solid lines represent control limits based on 2·57 times the pooled standard deviation of all residuals of the NOC process data. These limits are implicitly used by the ordinary *Q*-statistic by summing all residuals without taking into account their expected variation. According to the dotted curves, variables 7 and 8 have much smaller residuals than variables 4 and 5. Furthermore, in variable 8 the residuals decrease during the batch run. The difference in the size of the residuals is caused by the fact that some of the process variables are described better by the process model, and thus have smaller residuals, than other process variables. This also holds for different periods during the batch run.

The difference in size of residuals can also be caused by different scaling procedures. After autoscaling, each column of the **X** matrix has equal sum of squares, whereas scaling each process variable to equal sum of squares, as is done in variable slab scaling, leads to unequal sum of squares for each column in **X**. In the last case the residuals of each column are also expected to have unequal sum of squares.

In every subplot of Figure 1 a residual of size 0·02 at times 30 and 140 is represented by an asterisk. All these eight residuals are considered equally important in the ordinary *Q*-statistic, i.e. they all have the same contribution of $(0{\cdot}02)^2$. However, such residuals of 0·02 are well expected in variables 4 and 5 but not in variable 7 and at the end of variable 8. Residuals of 0·02 in variable 7 and at the end of variable 8 should give a high contribution to the *Q*-statistic because they are rare, whereas the same residuals in variables 4 and 5 should give a moderate contribution to the *Q*-statistic. By summing the squared residuals, as is done in the ordinary *Q*-statistic, the information on a residual being rare or common for a specific variable at a specific time period is lost. In order to keep this information, the residual of a new batch should be compared to the *I* NOC residuals from the same process variable at the same time period. This correct comparison can easily be performed by scaling the residuals by the spread expected for a specific process variable at a specific time period:

$$\tilde{e}_{ijk} = \frac{e_{ijk}}{s_{jk}} \tag{5}$$

Here the residuals of the NOC batches are divided by the standard deviation of the residuals of the *I* batches (s_{jk}) for each process variable *j* at time period *k*. The scaled residuals $\tilde{\mathbf{e}}_i$ are well approximated by a multinormal distribution with mean zero and covariance matrix $\tilde{\boldsymbol{\Sigma}}, \tilde{\mathbf{e}}_i \sim \mathbf{N}(0, \tilde{\boldsymbol{\Sigma}})$. Therefore, to estimate confidence limits for the standardized *Q*-statistic, the approximation of Jackson and Mudholkar [11] can be used for the scaled residuals $\tilde{\mathbf{E}}$.

When a new batch $\mathbf{x}_{\text{new}}$ ($JK\times1$) is monitored, first the residuals $\mathbf{e}_{\text{new}}$ ($JK\times1$) are calculated as described in Equation (2). Then these residuals are scaled using the standard deviation of the NOC residuals (s_{jk}):

$$\tilde{e}_{\text{new},jk} = \frac{e_{\text{new},jk}}{s_{jk}} \tag{6}$$

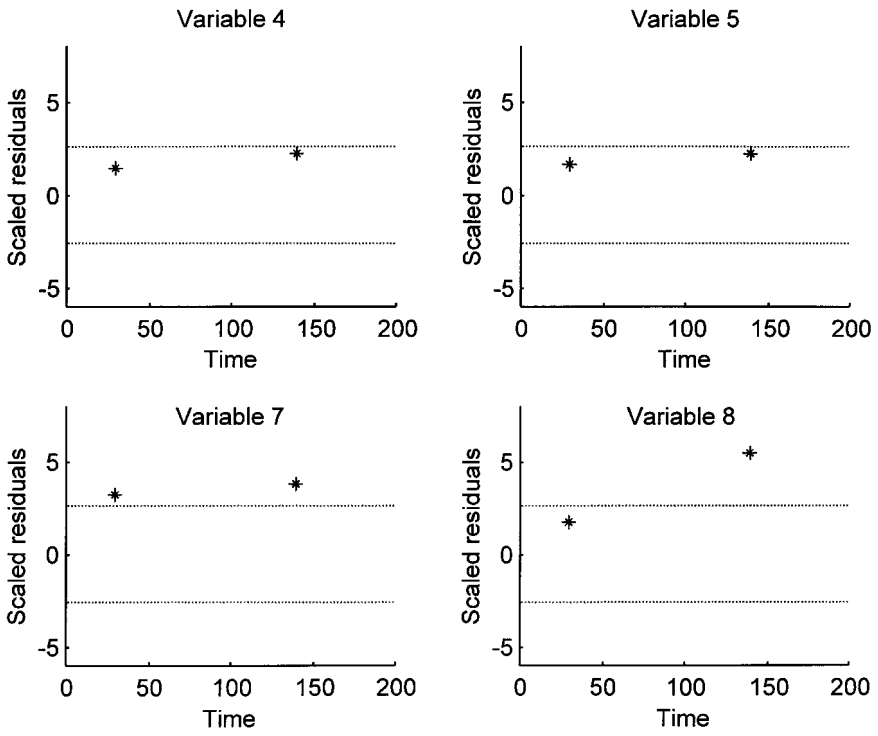

Figure 2. Control limits for scaled residuals of simulated batch data. Dotted lines represent 2·57 times standard deviation of residuals of each separate process variable at each time period, which equals the standard deviation of all scaled residuals. Dotted lines represent the same control limits as the dotted curves in Figure 1.

By using this scaling for new residuals, each scaled residual is compared to its expected spread according to the NOC residuals. In cases where the residuals of the NOC data show higher variation than average, the residuals of new batches should also be allowed to have high residuals. However, for periods with small residuals of the NOC data the residuals of the new batch must also have small residuals to be considered 'in control'. The standardized Q-statistic for the scaled new residuals is

$$Q_{\text{new}} = \sum_{jk=1}^{JK} (\tilde{e}_{\text{new},jk})^2 \sim \tilde{g}\chi^2_{\tilde{h}} \tag{7}$$

where $\tilde{g}$ and $\tilde{h}$ are the corresponding moments of the χ^2 distribution of the scaled residuals.

Figure 2 shows the effect of the scaling of the residuals. Again the same four process variables are shown. The dotted lines represent the same control limits as the dotted curves in Figure 1. Because of the scaling, the spread of the residuals is the same for all process variables at every time period, and thus the control limits have become straight lines. The asterisks added, which represented residuals of 0·02 at times 30 and 140, are also scaled. Now the residuals indicated with the asterisks have different contributions in the standardized Q-statistic. The residual in variable 8 at time 140 now has a high contribution to the standardized Q-statistic, whereas before the scaling its contribution was the same as the other residuals represented by the asterisks.

The idea of the standardized *Q*-statistic can be extended by scaling the residuals by the covariance structure of the residuals instead of scaling them by the variance as is done in the present paper. This may lead to an even more improved sensitivity of the *Q*-statistic.

2.3. On-line monitoring of batch processes and continuous processes

In this paper the *Q*-statistic of scaled residuals is introduced for the off-line monitoring of residuals of batch process data. However, this can easily be extended to the on-line monitoring of batch processes and to the monitoring of continuous processes.

For both the on-line monitoring of batch processes and the monitoring of continuous processes, data of a new measurement $\mathbf{x}_{\text{new}}^{\text{T}}$ ($1 \times J$) are compared to their expected variation obtained from NOC data $\mathbf{X}$ ($I \times J$). The residuals of the new measurement $\mathbf{e}_{\text{new}}^{\text{T}}$ ($1 \times J$) will be compared to their expected variation obtained from the NOC residuals $\mathbf{E}$ ($I \times J$). Because usually the process variables will not be explained equally well, the standard deviation of the I NOC residuals will be different for each of the J process variables. By scaling these residuals with the standard deviation of the I measurements of the corresponding process variables, the contribution of each residual to the *Q*-statistic (or in on-line monitoring often called SPE) will be corrected for the expected size of the residuals for that specific process variable. For on-line monitoring of batch processes, only the measurements of a specific time period t_k are considered. This means that the NOC set $\mathbf{X}_t$ ($I \times J$) changes for each different time period. However, the same ideas can be applied for each new set of $\mathbf{X}_t$ and $\mathbf{x}_{\text{new}}^{\text{T}}$.

3. RESULTS

3.1. Simulated batch process data

To show the effect of scaling the residuals with the corresponding standard deviation from the residuals of the NOC data, a benchmark data set of a simulated semi-batch emulsion polymerization of styrene and butadiene will be used [16]. Meaningful disturbances such as impurities in the initial charge of the organic phase and in the butadiene feed to the reactor were added. Measurements were taken from flow rates, temperatures, density, estimates of the conversion and energy release. A detailed description can be found in the literature [3]. Fifty batches were simulated to construct the NOC data, by introducing typical variations. Three additional batches were simulated, one with normal conditions and two with product that was out of the specification range. One of the erroneous batches had an initial organic impurity contamination in the butadiene feed. The other erroneous batch had the same problem, but the contamination was higher and started halfway through the batch operation.

The NOC data were arranged in a three-way array $\underline{\mathbf{X}}$ ($I \times J \times K$) of $I = 50$ batches, $J = 9$ process variables and $K = 200$ times points. To describe the variation of the process variables around their average trajectories, each column of $\underline{\mathbf{X}}$ was centered to mean zero. Furthermore, each process variable was scaled to unit sum of squares. This type of scaling is called slab scaling [17, R Bro, A K Smilde, submitted manuscript].

The NOC data were modeled with a PARAFAC model [18,19]. Three components were found to best fit the data. Although the 50 batches were simulated to come from normal operating conditions, two batches were fitted rather badly by the PARAFAC model and were both outside the 99% control limit of the *D*-statistic. These two batches were removed from the NOC set, and the final $\underline{\mathbf{X}}$ ($48 \times 9 \times 200$) was used to develop the MSPC model. The PARAFAC model with three components describes 21% of the total variance in the NOC data. This amount is small; however, such low percentages are often seen in the modeling of batch process data. To determine the effect of scaling on the *Q*-statistic, residuals of process variables 4, 5, 7 and 8 were studied in detail. These four process

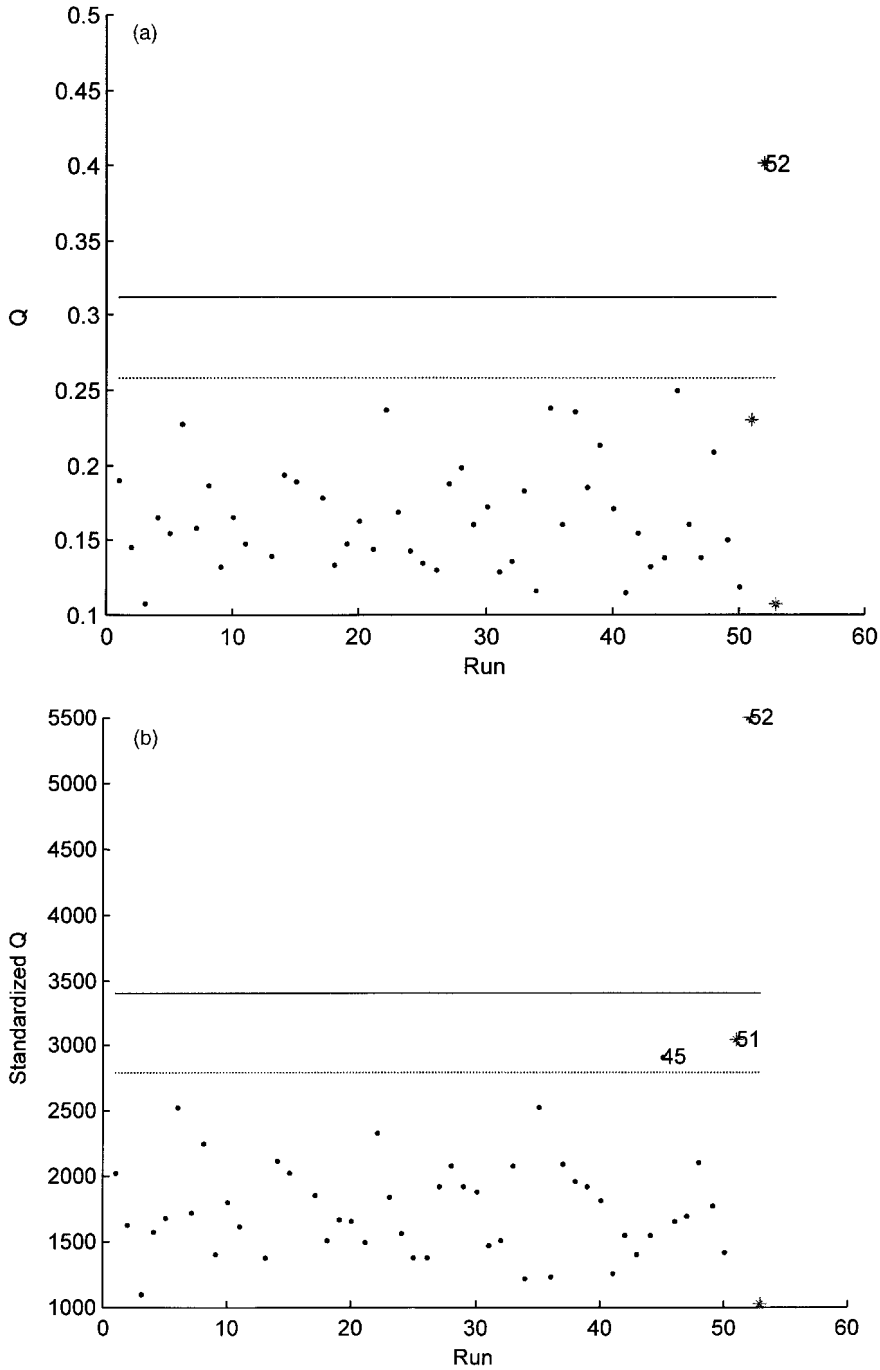

Figure 3. (a) Ordinary *Q*-statistic chart and (b) standardized *Q*-statistic chart with 95% and 99% control limits of the simulated SBR batch data. Points represent the NOC batches and asterisks represent the new batches.

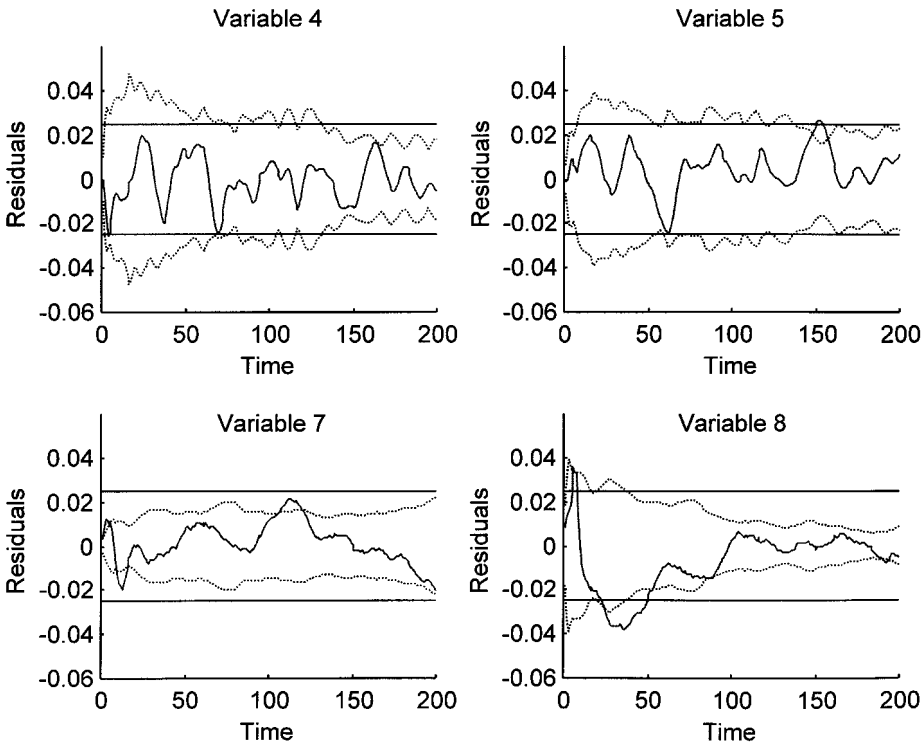

Figure 4. Residuals of batch 51 (solid curves). Solid lines represent control limits based on 2·57 times the standard deviation of all residuals of the four process variables of all time periods of the NOC batches. Dotted curves represent control limits based on 2·57 times the standard deviation of residuals of each specific process variable for each specific time period of the NOC batches.

variables describe the temperature of the reactor and the cooling system as well as the density of the latex and the conversion rate of the polymerization. The disturbances of the two erroneous batches are detected mainly in these four process variables.

Figure 3 shows the *Q*-statistic charts for this specific data set. Figure 3(a) shows the ordinary *Q*-statistic calculated using the standard residuals, and Figure 3(b) shows the standardized *Q*-statistic chart using the scaled residuals. The *Q*-statistic is calculated using the residuals of all process variables available. The first 48 batches (1–50 without 12 and 16) are the NOC batches obtained without any disturbance. Batches 51–53 are the new batches, of which 51 and 52 have a disturbance of impurities entering the reactor. In both cases all NOC batches and batch 53 are within the 95% confidence limit, which shows that these batches are all in control, except for batch 45 which is just above the 95% limit in the standardized *Q*-statistic chart. In Figure 3(a), batch 52 is above the limits and batch 51 is well below the 95% confidence limit. In Figure 3(b), batch 52 is out further than in Figure 3(a), while batch 51 is now above the 95% confidence limit.

Figures 4 and 5 show the effect of scaling the residuals of batch 51 with the standard deviation known from the NOC data. Figure 4 shows the ordinary residuals of process variables 4, 5, 7 and 8 of this batch where impurities entered the reactor from the start of the process. Just as in Figure 1, the dotted curves are the limits of 2·57 times the standard deviation of the *I* residuals of each separate process variable at each separate time period. The solid lines represent the limits of 2·57 times the

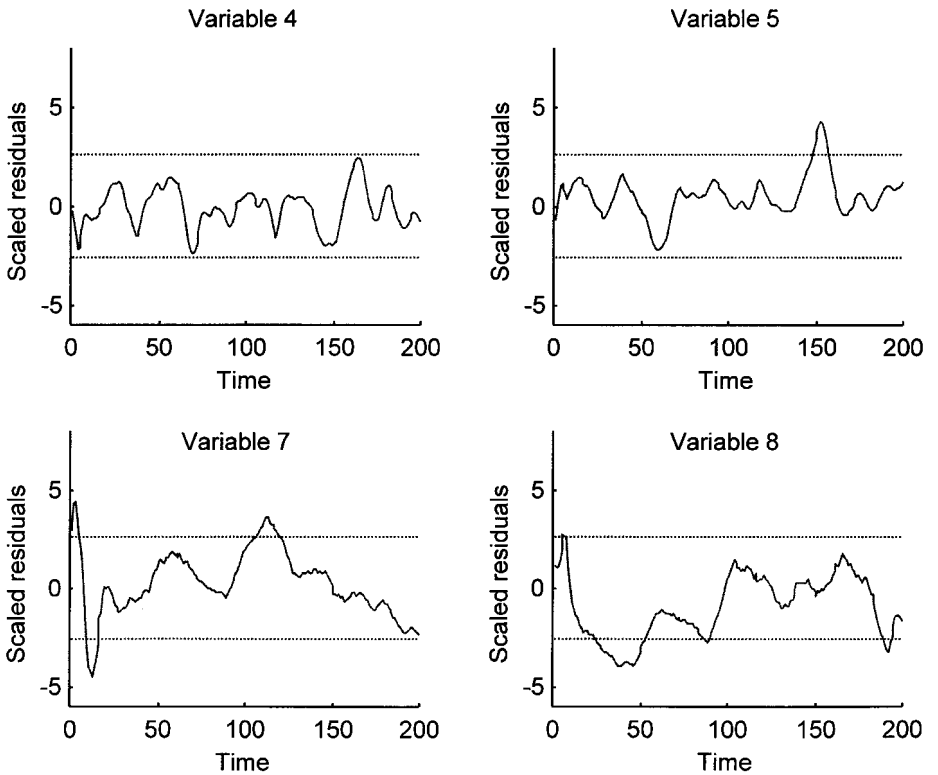

Figure 5. Residuals of batch 51 scaled by the standard deviation of the corresponding process variable and time period of the NOC batches (solid curves). Dotted lines represent the same control limits as the dotted curves in Figure 4.

pooled standard deviation of all $I \times JK$ residuals of the process data. These lines are the limits considered when the residuals are summed as in the ordinary Q-statistic. Figure 5 shows the residuals scaled by the standard deviation of the residuals of the corresponding process variable and time period. This means that after scaling, all residuals of each process variable and time period are assumed to have variance equal to one. Thus the residuals are scaled according to their expected variation. The dotted lines are the new confidence limits, which are the same limits as the dotted curves in Figure 4.

It is clearly seen that the scaled residuals are more sensitive to deviations from normal operating behavior than the ordinary residuals. Before scaling, small residuals have a small contribution to the ordinary Q-statistic, even if these residuals are larger than expected. After scaling, these residuals will have a high contribution to the standardized Q-statistic because they are larger than expected. This can be seen for process variable 7 at times 20 and 105 and for process variable 8 at time 190. If residuals are high, then before scaling, they have a large contribution in the ordinary Q-statistic, even if such a high residual was expected according to the NOC residuals. After scaling, such a residual has a lower contribution to the standardized Q-statistic, because such a residual is not rare for that process variable at that specific time during the batch run. This can be seen for process variable 8 at times 7–10.

This shows that standardizing the residuals of batch 51 allows them to be compared to the expected variation at each specific process variable and time period. Some periods therefore gave higher

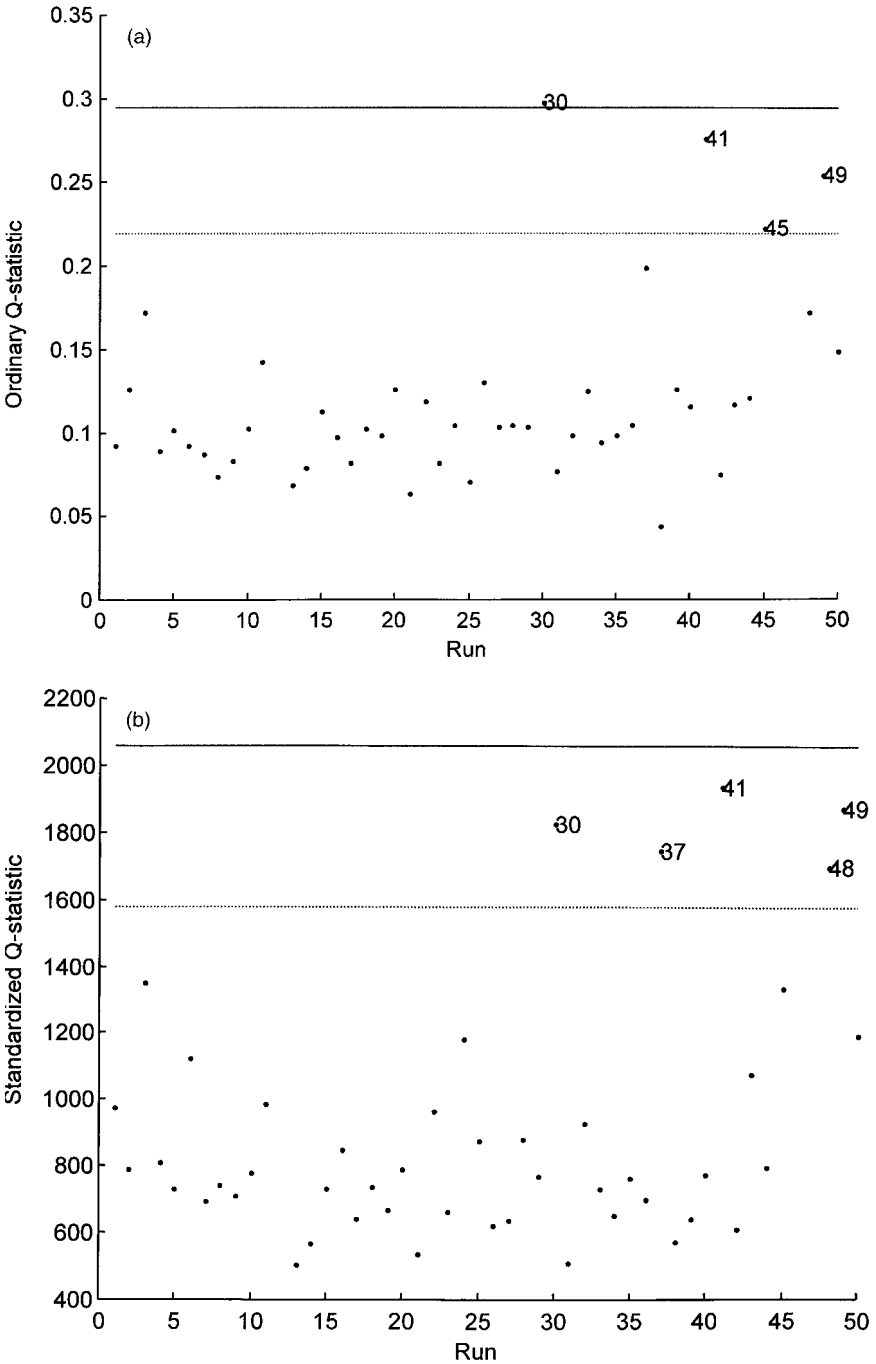

Figure 6. (a) Ordinary *Q*-statistic chart and (b) standardized *Q*-statistic chart with 95% and 99% control limits of the industrial polymerization batch process data.

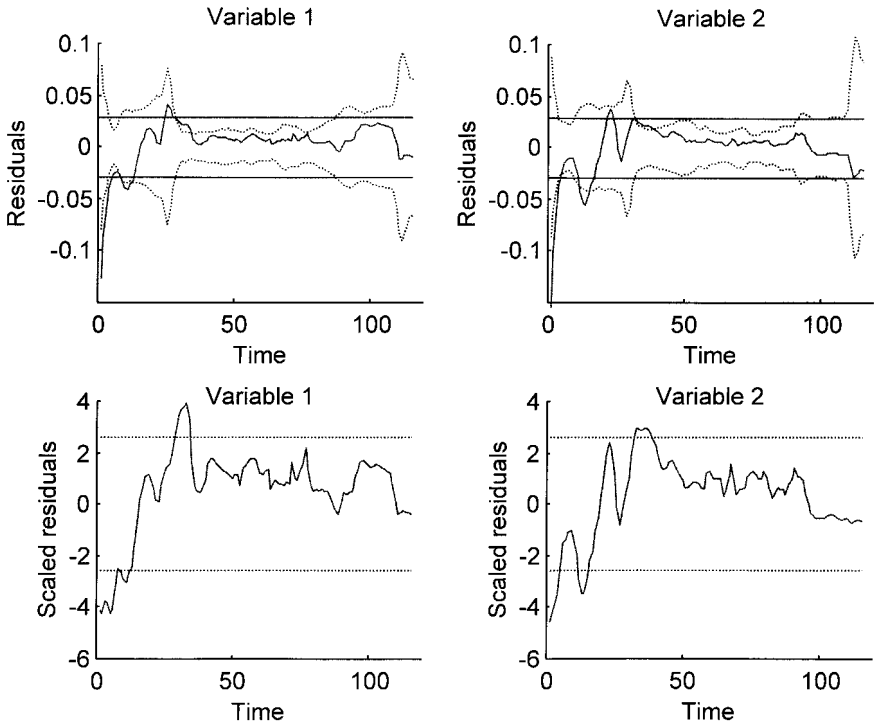

Figure 7. (Top plots) Residuals of batch 30 of the industrial application (solid curves). Solid lines represent control limits based on 2·57 times the standard deviation of all residuals of the four process variables of all time periods of the NOC batches. Dotted curves represent control limits based on 2·57 times the standard deviation of residuals of each specific process variable for each specific time period of the NOC batches. (Bottom plots) Standardized residuals of batch 30 (full curves). Dotted lines represent the same control limits as the dotted curves in the top plots.

contributions to the *Q*-statistic because residuals were higher than expected even if the absolute value of the residuals was not high. This forced the *Q*-statistic of batch 51 to go outside the 95% confidence limit, and thus this batch becomes suspicious and should be examined in more detail.

3.2. Industrial batch process data

The second batch process data set was provided by DuPont and was described previously in the literature [2,20,21]. Fifty batches were obtained from a polymerization process, and eight process variables were obtained over 116 time intervals. A Tucker3 model with three components in the batch mode, two components in the variable mode and three components in the time mode was used to model the three-way array. Three batches were found to be outside the control limits and were removed from the NOC data. The remaining 47 batches were modeled using a Tucker3 model with the same number of components as used above.

Figures 6(a) and 6(b) show respectively the ordinary *Q*-statistic and the standardized *Q*-statistic for the 47 NOC batches (batches 1–50 without 12, 46 and 47). The main difference between the two *Q*-statistic plots is that after standardizing the residuals, batch 30 has come below the 99% confidence limit and batches 37 and 48 have gone above the 95% confidence limit.

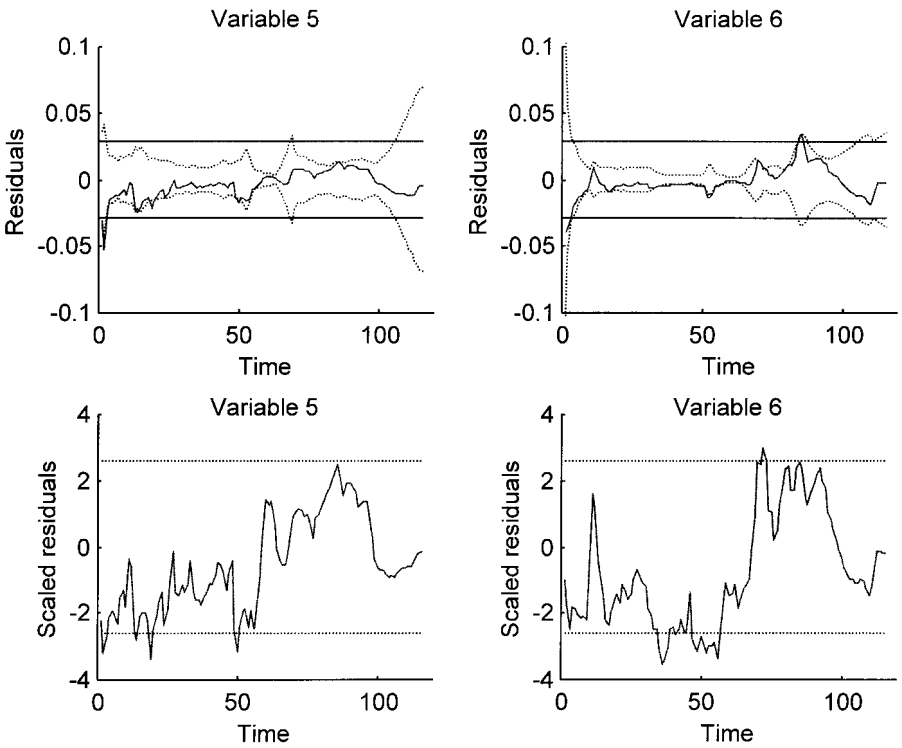

Figure 8. (Top plots) Residuals of batch 37 of the industrial application (solid curves). Solid lines represent control limits based on 2·57 times the standard deviation of all residuals of the four process variables of all time periods of the NOC batches. Dotted curves represent control limits based on 2·57 times the standard deviation of residuals of each specific process variable for each specific time period of the NOC batches. (Bottom plots) Standardized residuals of batch 37 (full curves). Dotted lines represent the same control limits as the dotted curves in the top plots.

Batch 30 might be considered an erroneous batch because it is just outside the 99% confidence limit in the ordinary *Q*-statistic plot. However, after standardizing the residuals, it is only just outside the 95% confidence limit. Figure 7 shows the residuals of batch 30 for process variables 1 and 2. In the top plots of Figure 7 the solid lines represent control limits based on 2·57 times the pooled standard deviation of all residuals of the NOC process data. These limits are implicitly used by the ordinary *Q*-statistic by summing all residuals without taking into account their expected variation. The dotted curves represent control limits based on 2·57 times the standard deviation of each separate process variable at each time period. This shows that the standard deviation of the residuals is not the same for each process variable and for each time period. The residuals of batch 30 are represented by the solid curves. It is clear that in the first few time periods, batch 30 is far outside the solid-line confidence limits. This is the main reason that the ordinary *Q*-statistic of batch 30 is outside the 99% confidence limit. In the bottom plots the residuals are standardized and the dotted-line control limits are the same as the dotted-curve confidence limits of the top plots. Batch 30 is still outside the confidence limits at the beginning of the run, but not as far as without the standardization. In other words, the high residuals at the beginning of batch 30 for process variables 1 and 2 are not that abnormal, because high residuals are expected in this region. Therefore batch 30 is not as deviating as is indicated by the ordinary *Q*-statistic.

Batch 37 went outside the 95% confidence because of the standardization; however, it is still not considered an erroneous batch. Figure 8 shows the residuals of this batch for process variables 5 and 6. Just as in Figure 7, here the ordinary residuals are presented in the top plots and the standardized residuals are presented in the bottom plots of Figure 8. At 20 min and also around 50 min the standardized residuals are outside the confidence limits, as can be seen in the bottom plots. Although the ordinary residuals are rather small and fall well inside the ordinary confidence limits, as can be seen in the top plots, they are higher than expected for that specific region.

Concluding, standardizing residuals by scaling them by their corresponding standard deviation improves the sensitivity of the *Q*-statistic. This may force some batches to go outside confidence limits, but it can also force batches to go inside confidence limits. Batches that seem erroneous because of the ordinary *Q*-statistic may be good after all.

4. CONCLUSION

In this paper the sensitivity of the *Q*-statistic is improved by scaling the residuals of new batches by their expected variation. This expected variation is determined as the standard deviation of a set of residuals obtained from process data obtained under normal operating conditions. By this scaling, the relative contribution of a residual to the *Q*-statistic, as compared to the spread of residuals of the NOC data, is considered instead of its ordinary contribution. The use of scaled residuals instead of ordinary residuals improves the power of the *Q*-statistic considerably, as shown using a simulated batch process and an industrial polymerization batch process. The newly defined *Q*-statistic is called the standardized *Q*-statistic.

ACKNOWLEDGEMENTS

These investigations were supported by the Council for Chemical Sciences of the Netherlands Organization for Scientific Research (NWO-CW) with financial aid from the Netherlands Technology Foundation (STW). The authors would like to thank John MacGregor and Dora Kourti of McMaster University, Hamilton, ON, Canada for providing the simulated batch data, and Kenneth Dahl of DuPont, Wilmington, DE, USA for providing the industrial process data.

REFERENCES

1. Kresta J, MacGregor JF, Marlin TE. Multivariate statistical monitoring of process operating performance. *Can. J. Chem. Engng.* 1991; **69**: 35–47.
2. Nomikos P, MacGregor JF. Multivariate SPC charts for monitoring batch processes. *Technometrics* 1995; **37**: 41–59.
3. Nomikos P, MacGregor JF. Monitoring of batch processes using multiway principal component analysis. *AIChE J.* 1994; **40**: 1361–1375.
4. Kourti T, MacGregor JF. Multivariate SPC methods for process and product monitoring. *J. Qual. Technol.* 1996; **28**: 409–428.
5. Kourti T, MacGregor JF. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics Intell. Lab. Syst.* 1995; **28**: 3–21.
6. Kourti T, Lee J, MacGregor JF. Experiences with industrial applications of projection methods for multivariate statistical process control. *Comput. Chem. Engng.* 1996; **20**: S745–S750.
7. Neogi D, Schlags CE. Multivariate statistical analysis of an emulsion batch process. *Ind. Engng. Chem. Res.* 1998; **37**: 3971–3979.
8. Wise BM, Gallagher NB. The process chemometrics approach to process monitoring and fault detection. *J. Process Control* 1996; **6**: 329–348.
9. Tracy ND, Young JC, Mason RL. Multivariate control charts for individual observations. *J. Qual. Technol.* 1992; **24**: 88–95.

10. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems: effect of inequality of variance in one-way classification. *Ann. Math. Statist.* 1954; **25**: 290–302.
11. Jackson JE, Mudholkar GS. Control procedures for residuals associated with principal component analysis. *Technometrics* 1979; **21**: 341–349.
12. Nomikos P. Detection and diagnosis of abnormal batch operations based on multi-way principal component analysis. *ISA Trans.* 1996; **35**: 259–266.
13. Louwerse DJ, Smilde AK. Multivariate statistical process control of batch processes using three-way models. *Chem. Engng. Sci.* 2000; **50**: 1225–1235.
14. Belsley DA, Kuh E, Welsh RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. J Wiley: New York, 1980.
15. Kiers HAL. Towards a standardized notation and terminology in multiway analysis. *J. Chemometrics* 2000; **14**: 105–122.
16. Broadhead TO, Hamielec AE, MacGregor JF. Dynamic modeling of the batch, semi-batch and continuous production of styrene–butadiene copolymers by emulsion polymerization. *Makromol. Chem. Suppl.* 1985; **10**: 105–128.
17. Harshman RA, Lundy ME. Data preprocessing and the extended PARAFAC model. In *Research Methods for Multimode Data Analysis*. Law HG, Snyder CW, Hattie JA, MacDonald RP (eds). Praeger: New York, 1984; 216–284.
18. Harshman RA. Foundations of the PARAFAC procedure: models and conditions for exploratory multi-mode factor analysis. *UCLA Working Papers Phonet.* 1970; **16**: 1–84.
19. Bro R. PARAFAC, tutorial and applications. *Chemometrics Intell. Lab. Syst.* 1997; **38**: 149–171.
20. Boque R, Smilde AK. Monitoring and diagnosis batch processes with multiway covariates regression models. *AIChE J.* 1999; **45**: 1504–1520.
21. Kosanovic KA, Dahl KS, Piovoso MJ. Improved process understanding using multiway principal component analysis. *Ind. Engng. Chem. Res.* 1996; **35**: 138.

