MULTIVARIATE STATISTICAL CONTROL OF UNSYNCHRONIZED BATCH PROCESSES

Flavio S. Fogliatto* & Ndèye Niang**

* Département d'Ingénierie Industriel
Universidade Federal do Rio Grande do Sul
Av. Paulo Gama, 110 – Porto Alegre, RS 90040-060, Brésil
ffogliatto@producao.ufrgs.br

** Chaire de Statistique Appliquée et CEDRIC
Conservatoire National des Arts et Métiers
292, rue Saint Martin, 75141 Paris Cedex 03, France
niang@cnam.fr

RESUME

Les procédés par lots sont largement utilisés dans le secteur industriel notamment dans l'industrie agroalimentaire ou pharmaceutique. Dans ces procédés, les matières premières sont introduites dans un ordre spécifique et subissent une série de transformations donnant le produit final. La performance du procédé est mesurée par un ensemble de variables au fur et à mesure de son déroulement. Les données issues de tels procédés sont fortement auto-corrélées et sont en général contrôlées en utilisant des cartes de contrôle basées sur l'analyse des données fonctionnelles (CCPs). Dans cette communication nous étudions le cas particulier (et assez fréquent) de procédés par lots à temps variable que les CCPs ne permettent pas de contrôler directement. Nous proposons une nouvelle approche dans laquelle plutôt que d'aligner les lots ou d'utiliser des techniques de *time warping*, nous complétons les données en utilisant un modèle approprié de sorte que toute la variabilité temporelle est préservée. Ensuite, nous construisons des cartes de contrôles non paramétriques à partir des plans factoriels issus de l'application de la méthode STATIS aux données complétées. La méthode proposée est illustrée sur des données simulées.

MOTS-CLES: procédés par lots à temps variable, contrôle de qualité multivarié, méthode Statis

SUMMARY

Batch processes are widely used in several industrial sectors, such as food and pharmaceutical manufacturing. In a typical batch, raw materials are loaded in the processing unit and submitted to a series of transformations, yielding the final product. Process performance is described by variables which are monitored as the batch progresses. Data arising from such processes are likely to display a strong correlation-autocorrelation structure, and are usually monitored using control charts based on multiway principal components analysis (CCPs). In this paper we investigate the special (and rather frequent) case of batches with varying duration, which cannot be directly monitored using CCPs. We propose a new quality control strategy for monitoring such batches. In our proposition, batches are not aligned or time warped with respect to their trajectories, but are rather completed using a straightforward scheme. Thus all information on the variability in batch profiles along the time axis is preserved. The completed data set is reduced using the Statis method and monitoring of batch performance is accomplished directly on principal plane graphs, from which non-parametric control charts are derived. A simulated example ilustrates the proposed method.

KEYWORDS: Unsynchronized batches, Multivariate quality control, Statis method.

1. Introduction

Multivariate control charts (CCs) are indicated for simultaneously monitoring quality characteristics in a process (or product). The most commonly used multivariate CC is the Hotelling (or T^2) chart. Other multivariate CCs are reviewed by Lowry & Montgomery (1995) and Harris *et al.* (1999). Traditional CCs are based on the independence and multinormality assumptions which are not

always verifiable in practice. In addition, traditional multivariate CCs do not offer an efficient monitoring when the nominal behavior of variables is described by profiles. In those cases, variables may present a significant correlation and autocorrelation structure, and process monitoring is usually accomplished using multivariate charts based on multiway principal components analysis (MPCA). These charts are denoted here by CCPs.

Applications of CCPs to batch processes assume firstly that all batches considered in the analysis have the same length and are aligned with respect to each process stage. When that is not verified, CCPs must be adapted to handle variable batch duration. However, propositions found in the literature for that matter are not always satisfactory, as discussed below. Secondly, process variables monitored through CCPs are supposed to be multinormaly distributed, which enables the use of a T^2 CC to monitor the reduced data set; when that is not the case, the use of non-parametric CCs should be considered.

In this paper we propose a new quality control strategy for monitoring batch processes with varying duration. In our proposition, batches are not aligned or time warped with respect to their trajectories, but are rather completed using a straightforward scheme. Thus all information on the variability in batch profiles along the time axis is preserved. The completed data set is reduced using the Statis method (Escoufier, 1987), extending the approach in Scepi (2002). Two summarized representations of the batches become available: one that allows verifying batch progression compliance to the expected trajectories of its monitored variables, and other that summarizes trajectories of variables in a batch in each time instant, and allows detecting significant departures from their expected behavior in the time axis. Monitoring of batch performance with respect to the two summarized representations above is accomplished directly on principal plane graphs, from which non-parametric CCs are derived. The proposed method is illustrated using a simulated data set from Rosa (2005).

Our work extends the approach in Scepi (2002), where the use of the Statis method in multivariate quality control was initially proposed. However, at least two contributions separates our proposition from the one in Scepi (2002). The first concerns dealing with batches of variable duration, not discussed by the author. The second concerns the development of control charts and data completion strategies suitable for on line monitoring of batch processes. Scepi (2002)'s methods do not directly apply to on line monitoring of processes.

2. Background

The work here is closely related to the literature on quality control of batch processes. The application of CCPs to monitor batch processes of fixed length was initially proposed by Jackson & Mudhokar (1979), being further investigated by Nomikos & MacGregor (1995), Kourti & MacGregor (1996) and MacGregor (1997). Applications of CCPs in the monitoring of batch processes may be found in Kourti (2003), among others. In short, batch process monitoring using CCPs is carried out verifying the outputs of two CCs. The first is a T^2-chart for the scores obtained projecting future batches on the q PCs retained in the reference distribution. Such distribution is obtained running an MPCA on data coming from good batches that emerged from the process. The second is a Q chart for the residuals from the reference model. The first CC monitors the behavior of known process variability sources; the second CC detects any atypical events that disturb the process variables correlation-autocorrelation structure.

In the CCP monitoring scheme above batches are assumed to be synchronized and to have the same duration, i.e. all data vectors in the reference distribution as well as those arising from future batches have the same dimension. Otherwise, we have a situation where batches are not aligned in time, and the scheme above cannot be applied directly to the process data. Such special instance was first identified by Nomikos (1995), who suggested the use of an indexing variable, other than time, to monitor batch progression. Such variable should be measurable in the process, be monotonically increasing in time, and present the same initial and final values in all batches. In spite of being somewhat

restrictive, the requirements above were satisfied in the applications reported by Kourti *et al*. (1996) and Neogi & Schlags (1998).

Kassidas *et al*. (1998) proposed a different approach to the varying time batch problem, where batches were aligned using dynamic time warping (DTW) algorithms. Once aligned, batch data were treated using the CCP quality control strategy. However, DTW algorithms present some intrinsic limitations. Most notably, trajectories are aligned pairwise either (*i*) leading to an expanded common trajectory particular to each pair of trajectories (in the case of symmetric algorithms), or (*ii*) being aligned to match a reference trajectory common to all pairs of trajectories (in the case of asymmetric algorithms). In case (*i*) the resulting dataset will still be comprised of trajectories with different lengths; in case (*ii*) data points are allowed to be dismissed during the alignment of certain batches such that the length of pairs of trajectories coincide. The authors propose a compromise solution to overcome such limitations, where data points are averaged rather than dismissed during the alignment of trajectories. However, the greatest limitation in their approach seems to be related to the representation of batch variation along the time axis, which is altered when stages in the batch process are synchronized.

Kaistha *et al*. (2004) proposed the use of an *ad hoc* time warping technique to align batches of different durations. In their approach batch progression is supposed to be divisible in stages. Landmark points in stages (such as beginning and end) are used as anchor points in a simplified interpolation procedure, leading to batches that are synchronized stagewise. Once aligned, batch data are treated using a slightly altered version of the CCP strategy.

Finally, Rosa (2005) approached the varying time batch problem using a different analytical framework, where no dimensionality reduction techniques or procedures to align unequal batches are used. The author proposes the use of the Hausdorf distance as a measure of dissimilarity between a given batch and an average nominal batch. Such distance corresponds to the median of the minimum squared Euclidian distances between points in a given trajectory and all points in a reference trajectory. The reference distribution used for process monitoring is based on the Hausdorf distances. Despite its simplicity and the promising results obtained applying the method in simulated scenarios, there is no evidence that the Hausdorf distance captures the correlation-autocorrelation structure present in the original variables. In addition, the author did not propose diagnosis tools to use in combination with the CCs.

3. Method

Let N denote the total number of batches $(b = 1, \ldots, N)$ used to form the reference sample, from which a reference distribution will be determined and used to monitor future batches. For batches that are highly unaligned (i.e., the largest batch length differs substantially from the smallest batch length), N should be large enough such that all batch durations are well represented in the reference sample.

3.1 Completion of batches

Let X_{it} $(i = 1, \ldots, p;\ t = 1, \ldots, T)$ denote a sequence of observations gathered from process variable i as the batch progresses in time. Let $T_{\min}$ and $T_{\max}$ be the minimum and maximum batch duration, respectively, in the reference sample. Let $\mathbf{s}_i$ denote a $(T_{\min} \times 1)$ vector containing the sample standard deviations associated with the i-th variable, with element s_{it} giving the standard deviation of X_{it} estimated from all batches. Adjust a time series model to the observations in $\mathbf{s}_i$ and use the model to forecast s_{it} from $t = T_{\min+1}$ to $t = T_{\max}$.

Let $\bar{x}_{it}$ be the sample average of variable i observed at time t, measured from all batches, for $t > T_{\min}$. Pair the values of $\bar{x}_{it}$ and $\hat{s}_{it}$ (forecasted values of s_{it}), for periods $t = T_{\min+1}, \ldots, T_{\max}$, and use the pair of parameters to generate $(T_{\max} - T_{\min+1})$ normally distributed random variables. Complete the

batches with duration smaller than $T_{\max}$ with realizations from the simulated random variables at each time instant.

3.2 Quality control charts

We propose applying the Statis method on the completed reference data base. Following the method's nomenclature, each batch will correspond to a data table and each time period to an individual. There will be N tables of dimension $(T_{\max} \times p)$. Further, we propose the construction of two control charts using principal factor projections: the IS (InterStructure) and the CO_t (COmpromise) control charts. The IS CC (also used by Scepi, 2002) gives an overall assessment of the batches such that each batch (i.e., each table) is represented by a point in the chart. The CO_t CC gives an assessment of the individuals, which are the time periods, over all batches at time *t*. [The collection of CO_t CCs may also be joined in a unique CO CC, displaying the mean trajectories of the individuals.] The construction of these CCs and their use in quality control of future batches is detailed next.

Let $\mathbf{X}_b$ denote a matrix with process data from the *b*-th batch in the reference sample (columns in $\mathbf{X}_b$ correspond to sequences X_{it}). The *N* matrices $\mathbf{X}_b$ are organized in a three-dimensional array $\underline{\mathbf{X}}$ from which an interstructure (**IS**) matrix is derived. The general element of **IS** is given by (Escoufier, 1987):

$$Rv_{bb'} = tr(\mathbf{W}_b \mathbf{W}_{b'}) \Big/ \sqrt{tr(\mathbf{W}_b)^2 tr(\mathbf{W}_{b'})^2} \qquad (1)$$

with $\mathbf{W}_b = \mathbf{X}_b \mathbf{X}_b'$. Note that (*i*) b and b' are any pair of batches (such that $b \neq b'$), and (*ii*) the data matrices are usually normed to remove scale effects in the original variables.

Diagonalizing the **IS** matrix and projecting the batches from the reference data set in the first factorial plan, it is possible to obtain a reference data cloud. The IS chart may be obtained after establishing a $(1-\alpha)$ control region in the resulting biplot. For that matter, we use the proposition in Zani *et al.* (1998), comprised of the following three steps. First, a robust centroid is determined in the factorial plan corresponding to the chart. Next, an inner region is defined in the plan such that 50% of the points in the graph fall within its boundaries. Such region is defined by a *B*-spline curve that smoothes the contours of a convex hull containing the points. Finally, the control region is established in the chart by defining a multiple of the distance *l* between the centroid and the boundary of the 50%-hull that corresponds to the desired probability of false alarm (α). For an $\alpha = 0.01$, $l = 1.68$. Since the inner region spreads irregularly from the centroid, the control region will also be irregular in shape.

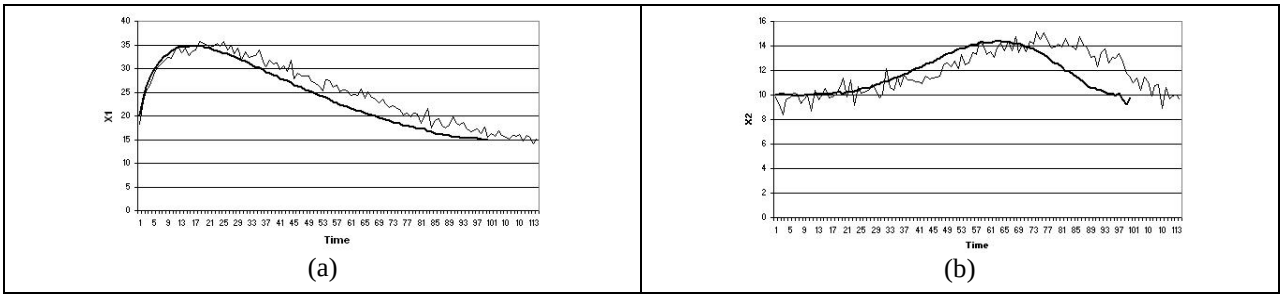
To obtain the CO_t CCs, start by diagonalizing a compromise matrix **CO**, defined as the linear combination $\sum_b \beta_b \mathbf{W}_b$, where β_b is a function of the *b*-th element of the largest eigenvector obtained from diagonalizing the **IS** matrix. Projecting the compromised individuals (i.e., rows of **CO**) in the first factorial plan, we obtain a summarized representation of batch trajectory over all batches as time progresses. The CO_t CC is obtained by projecting the individuals observed at time *t* (gathered from the *t*-th row of $\underline{\mathbf{X}}$) in the first factorial plan obtained from diagonalizing **CO**. Once again, we apply convex hull peeling and *B*-spline smoothing on the resulting bivariate plot to obtain a chart with an empirical α confidence level determined by the peeling of points in the data cloud.

Off-line process control of a future batch $b = N+1$ takes place initially by projecting $\mathbf{X}_{N+1}$ in the IS chart. In case the projection yields an out-of-control signal, the CO_t CCs are used to identify in which time instant the batch departed from the reference behavior. On-line process control takes place while the batch progresses in time and $t = t' \leq T$, where t' denotes the most recent time at which variables were sampled from the process. Thus, only a fraction of the process data table is available.

Since the idea is to verify the behavior of variables up to time t', their behavior in future time periods may be assumed to follow the average pattern. Missing data may be thus replaced in the table by the averages of process variables obtained from the sample of reference batches (other alternatives to complete the data table may also be considered). The new batch realized up to time t' may be monitored using the CO_t CCs. Note that there will be one CO_t CC for every time period up to t', allowing precise identification of out-of-control states in the process. The new batch may also be monitored using the IS CC. However, due to the missing data in $\mathbf{X}_{N+1}$, the smaller the value of t' the higher the probability of false alarms in the CC.

4. Example

We exemplify the method proposed in section 3 using simulated data available in Rosa (2005). The reference data set is comprised of 39 batches of varying duration $(T_{\min} = 90; T_{\max} = 100)$; three process variables (X_1, X_2, X_3) are considered. All batches were completed to $T = 100$ using the scheme proposed in section 3. The average trajectories of variables X_1 and X_2 in the reference distribution are depicted in Figure 1 (bolded lines). An additional batch (b = 40) is also considered in the analysis. The batch, abnormally long, is considered as out-of-control by Rosa (2005), even though it reproduces the process variables' basic profiles (trajectories of X_1 and X_2 in b = 40 are also given in Figure 1). Since our method was conceived to capture abnormal behaviors in the process time axis, we expect it to signalize b = 40 as an out-of-control batch.



(a) (b)

Figure 1. Average trajectories (bolded lines) of (a) X_1, and (b) X_2, and trajectories of batch b=40

We analyzed the completed three-dimensional data matrix using Statis. We projected the reference batches in the first factorial plan (which accounted for 86,5% of the total variance in $\underline{\mathbf{X}}$) derived from diagonalizing **IS** and determined the 99% convex hull, obtaining the IS CC in Figure 2. Note that b = 8 is positioned outside the hull. In Rosa (2005)'s CCs, b = 8 is characterized as in-control although basically coinciding with the upper control limit of the charts. We chose to leave b = 8 as part of the reference data set. Projecting b = 40 in the IS CC yields the coordinates $(f_1, f_2) = (0.138, -0.061)$, positioned far in the left of the chart in Figure 2 (justifying the discontinuity in the representation of Factor 1's axis). The IS CC clearly signalizes b = 40 as out-of-control, as we expected. Figure 2 also depicts the CO CC. For shortness, we do not present the CO_t CCs.

5. Conclusion

In this paper we propose a quality control method to monitor batch processes of varying duration. In our method, batches are completed to the maximum duration time in the data base and reduced in dimensionality using the Statis method (Escoufier, 1987). Process monitoring is accomplished through the use of two CCs : in the IS CC, batch progression compliance to the expected trajectories of its monitored variables is verified ; in the CO_t CC, trajectories of variables in a batch in each time instant are summarized and significant departures from their expected behavior in the time axis are detected. A simulated example, where three variables are monitored in a batch process, illustrates the method.

A natural extension of the work presented here would include the comparison of results obtained using the proposed method and other methods in the literature, most notably the dynamic time warping strategy suggested by Kassidas *et al.* (1998).

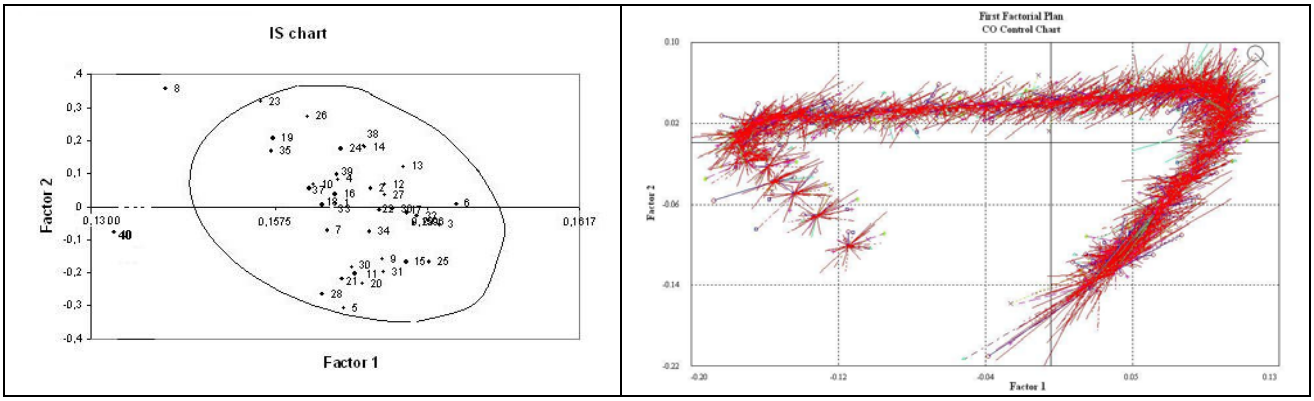


Figure 2. IS CC (left) and CO CC in the numerical example

Bibliographie

[1] Lowry, C.A. and Montgomery, D.C. (1995) A review of multivariate control charts. *IIE Transactions*, 27 (6), 800–810.

[2] Harris, T.C., Seppala, C.T. and Desborough, L.D. (1999) A review of performance monitoring and assessment techniques for univariate and multivariate control systems. *Journal of Process Control*, 9 (1), 1–17.

[3] Escoufier, Y. (1987) Three-mode data anlysis: the STATIS method. In *Methods for multidimensional data analysis*, ECAS, Fichet B. and Lauro N.C. (eds), 259–272.

[4] Scepi, G. (2002) Parametric and non parametric multivariate quality control charts. In *Multivariate Total Quality Control*, Physica-Verlag , Lauro C. *et al*. (eds), 163–189.

[5] Rosa, A.F.P. (2005) *Maîtrise statistique de procédés par lots à temps variable*. Thèse de Doctorat de l'Université de Nantes, Nantes (France).

[6] Jackson, J.E. and Mudholkar, G.S. (1979) Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics*, 21 (3), 341–34.

[7] Nomikos, P. and MacGregor, J.F. (1995) Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37 (1), p.41–59.

[8] Kourti, T. and MacGregor, J.F. (1996) Multivariate SPC Methods for Process and Product Monitoring. *Journal of Quality Technology*, 28 (4), 409–428.

[9] MaCGregor, J.F. (1997) Using on-line process data to improve quality: challenges for statisticians. *International Statistical Review*. 65 (3), 309–323.

[10] Kourti, T. (2003) Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *Journal of Chemometrics*, 17, 93–109.

[11] Nomikos, P. (1995) *Statistical process control of batch processes*. Ph.D. Thesis, McMaster University, Hamilton (Canada).

[12] Kourti, T., Lee, J. and MaCGregor, J.F. (1996) Experiences with industrial applications of projection methods for multivariate statistical process control. *Computers and Chemical Engineering*, 20, S745–S750.

[13] Neogi, D. and Schlags, C.E. (1998) Multivariate statistical analysis of an emulsion batch process. *Ind. Eng. Chem. Res.*, 37, 3971–3979.

[14] Kassidas, A., MaCGregor, J.F. and Taylor, P.A. (1998) Synchronization of batch trajectories using dynamic time warping. *AIChE Journal*, 44, 864–875.

[15] Kaistha, N., Moore, C.F. and Leitnaker, M.G. (2004) A statistical process control framework for the characterization of variation in batch profiles. *Technometrics*, 46 (1), 53–68.

[16] Zani, S., Riani, M. And Corbellini, A. (1998) Robust bivariate boxplots and multiple outlier detection. *Computational Statistics & Data Analysis*, 28, 257-270.