Non parametric on-line control of batch processes based on STATIS and clustering

Ndeye Niang Keita, Flavio Fogliatto, Gilbert Saporta

Journal de la Société Française de Statistique
Vol. 154 No. 3 (2013)

Non parametric on-line control of batch processes based on STATIS and clustering

Titre: Contrôle non paramétrique de procédés par lots basé sur STATIS et la classification

Ndèye Niang [1], Flavio S. Fogliatto [2] and Gilbert Saporta [1]

Abstract: Batch processes are widely used in several industrial sectors, e.g. food and pharmaceutical manufacturing. Process performance is described by variables which are monitored as the batch progresses. Data arising from such processes are usually monitored using control charts based on multiway principal components analysis. In this paper we propose a non parametric quality control strategy for monitoring batch processes with fixed as well as variable duration. In our proposition, data sets associated to batches are reduced using the STATIS method. Monitoring of batch performance is accomplished directly on principal plane graphs, from which non-parametric control regions are derived through convex hull peeling. This general approach allows off-line monitoring of batch processes as well as on-line monitoring after a constrained clustering step based on multivariate extension of W.D. Fisher's algorithm is carried out. A real example of batch process with fixed duration illustrates the proposed method.

Résumé : Les procédés par lots sont largement utilisés dans le secteur industriel notamment dans l'industrie agroalimentaire, chimique ou pharmaceutique. Le suivi de tels procédés est effectué à travers un ensemble de variables caractéristiques du procédé prélevées par un échantillonnage en ligne au fur et à mesure de son déroulement. Le procédé est contrôlé à travers des cartes multivariées basées sur une analyse en composantes principales particulière (multiway principal component analysis). Nous proposons une approche du contrôle de qualité des procédés par lots basée sur la méthode STATIS et des régions de contrôles non paramétriques obtenues à partir d'enveloppes convexes. Cette approche générale peut être utilisée pour le contrôle en fin de fabrication des procédés par lots ainsi que pour le contrôle en cours de fabrication après une étape de classification sous contrainte basée sur une extension multivariée de l'algorithme de W.D. Fisher. La méthode proposée est illustrée sur des données réelles issues d'un procédé par lots à temps fixe.

Keywords: Batch process, Clustering, Multivariate quality control, STATIS method,
Mots-clés : Procédés par lots, Classification, Contrôle de qualité multivarié, Méthode STATIS,
AMS 2000 subject classifications: 35L05, 35L70

1. Introduction

Batch processes are widely used in several industrial sectors, e.g. food and pharmaceutical manufacturing. In a typical batch, raw materials are loaded into 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.

Shewhart's univariate Control Charts (CCs) are usually applied in the monitoring of industrial processes (Montgomery, 2001). Such quality control strategy, added of a few assumptions, may

[1] CEDRIC, Conservatoire National des Arts et Métiers. 292, rue Saint Martin, 75141 Paris Cedex 03, France.
E-mail: [1]ndeye.niang_keita@cnam.fr, E-mail: [1]gilbert.saporta@cnam.fr
[2] Industrial Engineering Department, Universidade Federal do Rio Grande do Sul. Av. Paulo Gama, 110 Porto Alegre, RS 90040-060, Brazil
E-mail: [2]ffogliatto@producao.ufrgs.br

 ISSN: 2102-6238

be extended to the multivariate case. Multivariate CCs are indicated for monitoring multiple quality characteristics in a process (or product) simultaneously. Results obtained from univariate and multivariate CCs are particularly different when quality characteristics are correlated; in those cases the use of multivariate CCs is strongly recommended. The most commonly used multivariate CC is the Hotelling (or T^2) chart (Jackson, 1991). Other multivariate CCs are presented by Jackson (1991) and reviewed by Harris T.J. (1999); Lowry and Montgomery (1995); Wierda (1994).

Traditional multivariate CCs are based on independence and multinormality assumptions which are not always true in practice: samples collected from the process should be independent, which is rarely the case when data collection is automated and measurements are taken from the process on-line. CCs based on bootstrap methods (Liu and Tang, 1996) and non parametric control charts (Liu, 1995; Lombardo et al., 2008) should be considered to overcome the limitations in case of dependant measurements or non normal data. In addition, traditional multivariate CCs are not efficient when the variables' nominal behaviors are described by profiles as for data arising from batch processes, which are likely to display a strong correlation-autocorrelation structure. In those cases, process monitoring is usually accomplished using multivariate control charts based on multiway principal components analysis (MPCA). These charts are denoted here by MPCA-CCs. The application of MPCA-CCs to monitor batch processes of fixed length was initially proposed by Jackson and Mudholkar (1979), being further investigated by Kourti and MacGregor (1996); MacGregor (1997). Applications of MPCA-CCs in the monitoring of batch processes may be found in Flores-Cerrillo and MacGregor (2002); Kourti (2003), among others.

In this paper we focus on batches of constant duration. We propose a quality control strategy which enables off-line as well as on-line monitoring of batch processes. In our proposition, the data set is reduced using the STATIS method (Lavit et al., 1994). Two summarized representations of the batches become available. Monitoring of batch performance with respect to these two summarized representations is accomplished directly on principal plane graphs, from which non-parametric control regions based on convex hull peeling are derived. For off-line control, the proposed strategy is applied on the complete data set, whereas for on-line control it is applied on data sets obtained through a constrained clustering step based on a multivariate extension of Fisher's algorithm (Fisher, 1958). Such clustering step is one of the original contributions of this paper.

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 presented in the aforementioned work. The first is related to the strategy proposed to define the non-parametric control region. The second concerns on-line monitoring of batch processes; methods proposed in Scepi (2002) cannot be directly applied for that purpose.

The rest of the paper is organized as follows. In section 2 historical parametric MPCA-CCs are briefly presented. Section 3 presents the STATIS method in the context of our quality control strategy. Section 4 details the constrained clustering step, and in section 5 the determination of the non parametric control region is presented. Section 6 applies the method to a real data set from a batch polymerization reactor, available in the batch process literature (Nomikos and MacGregor, 1995; Eriksson et al., 2001). Section 7 gives the conclusion.

 ISSN: 2102-6238

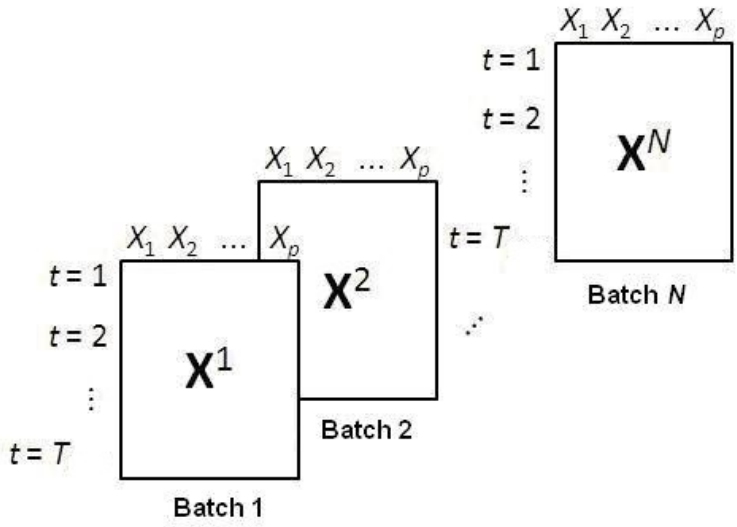

FIGURE 1. *Data matrices of batches in the reference sample*

2. Multiway principal components analysis control charts

To apply an MPCA-CC a reference sample comprised exclusively of data from batches that yielded products within specifications must be available. From these batches, a reference distribution will be determined and used to monitor future batches.

The following notations will be used in the rest of the paper: let N denote the total number of batches in the reference sample, t (t=1,..., T) is the time index, p (p=1,..., P) is the process variable index and i (i=1,..., N) is the batch index. Data from batch i are organized in a data matrix $\mathbf{X}^i$ where T outcomes of variables X^p are available. The N batches are organized in a three dimensional matrix $\underline{\mathbf{X}}$. Figure 1 illustrates the three-way data array (comprised of bidimensional matrices) in the context of the application proposed here, namely batch processes with constant duration.

In short, to implement an MPCA-CC scheme, a principal component analysis is performed on matrix $\underline{\mathbf{X}}$ unfolded into a bi-dimensional matrix $\mathbf{X}$ with rows corresponding to batches (see Figure 2). Batch process monitoring using MPCA-CCs is carried out verifying the outputs of classical multivariate CCs. A T^2 Hotelling chart for the scores is obtained projecting future batches on the first Q principal components (PCs) retained in the reference distribution. Future batches yielding a Hotelling statistic value $T^2_{Q,f}$ greater than the upper control limit UCL will be considered to be out-of-control, with

$$T^2_{Q,f} = \sum_{q=1}^{Q} \frac{c^2_{q,f}}{\sqrt{\lambda_q}}\,, \tag{1}$$

where $c^2_{q,f}$ is the coordinate of batch f in the q-th factorial axis, λ_q is the eigenvalue associated to the q-th eigenvector, the upper control limit is given by:

 ISSN: 2102-6238

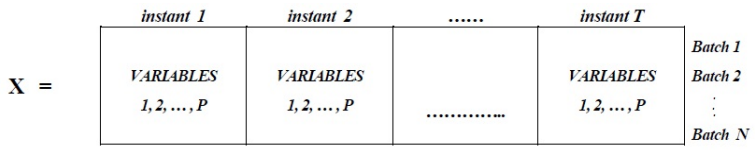

FIGURE 2. *Unfolded data matrix*

$$UCL = \frac{PT(N+1)(N-1)}{N(N-PT)} F_{1-\alpha,\, PT,\, N-PT}\ , \tag{2}$$

where $F_{1-\alpha,PT,N-PT}$ is the critical value (1-α quantile) of a Snedecor-Fisher distribution with PT, $N-PT$ degrees of freedom.

Additionnally, control charts based on remaining principal components (i.e. the residuals of the reference model) may be used to detect any atypical events that disturb the process variables' correlation-autocorrelation structure; on the other hand the first T^2 chart monitors the behavior of known process variability sources.

Such parametric control charts require process variables to be normally distributed. When this assumption is not verified, the use of non parametric control charts should be considered. Furthermore, in the MPCA-CC 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. Thus, the monitoring scheme above cannot be applied directly on-line as the new batch progresses in time. In this case data are available only up to time $t = t^* < T$, where t^* denotes the most recent time instant in which variables were sampled from the process. Thus the current batch length differs from the length of the reference batches. Notice that the MPCA-CCs can neither be used for off line monitoring of batch processes with variable duration.

For on-line monitoring, approaches can be found in Nomikos and MacGregor (1995) to complete missing data in the batches. Futhermore, several approaches to handle variable batch duration have been proposed (Doan and Srinivasan, 2008; Kaistha et al., 2004; Kassidas et al., 1998). However, propositions found in the literature for that matter are not always satisfactory. They generally consist of transforming process variables such that they present the same length and then applying the MPCA-CCs on process data. In the proposed approaches batches are aligned using dynamic time warping algorithms, which present some practical and theoretical intrinsic limitations; however, the greatest drawback in those approaches 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.

Castagliola and Ferreira (2006); Rosa (2005) approached the varying time batch problem using a different analytical framework, where no dimensionality reduction techniques or procedures to align unequal batches were used. The authors propose the use of the Hausdorff 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. Despite its simplicity and the promising results obtained applying the method in simulated scenarios, there is no evidence that the Hausdorff

 ISSN: 2102-6238

distance captures the correlation-autocorrelation structure present in the original variables.

Traditional control charts are one-dimensional graphical visualizations of repeated multivariate mean tests for independent batches. The non-parametric CCs proposed here are bi-dimensionnal graphs displaying a control contour rather than an upper control limit. Our method is based on STATIS factorial planes on which a convex hull peeling is applied. Since the resulting hull may not be smooth, in particular if the number of reference batches is small, a B-spline curve is adjusted to the hull to yield a smoother contour. For a brief introduction on B-splines, see Hastie et al. (2001). The B-spline we use is basically a cubic curve obtained by interpolation which is able to smooth a series of n points given in any order (Zani et al., 1998). In our applications, points are given by the factorial coordinates of the batches corresponding to the convex hull to smooth; each point corresponds to a knot of the B-spline.

3. STATIS for dimensionality reduction

In this section we briefly present the STATIS method, contextualized to the application proposed in this paper. In our proposition, each batch data set is considered as a matrix where the rows are time instants. This matrix is denoted by $\mathbf{X}^i \ \ (i = 1, ..., N)$ where T outcomes of P variables are available (see Figure 1). Prior to the analysis, each variable is usually centered (i.e. the mean of each variable is made equal to zero) and normalized to remove scale effects (i.e. the variance of each variable is made equal to one). We recall that N batches are used to form the reference sample, from which a reference distribution will be determined and used to monitor future batches. The reference sample should be comprised exclusively of data from batches that yielded products within specifications.

STATIS (Escoufier, 1987; Lavit et al., 1994) is a multivariate data analysis method aimed at exploring the structure of several data tables obtained under different circumstances. The multiple data tables are sets of variables measured on the same observations. The method consists of reducing the dimensionality of the data using a similarity measure based on Euclidean distances between configurations of points.

The method was originally proposed by des Plantes (1976); up to our knowledge, its use in quality control was first proposed by Scepi (2002). A recent review on the method and its extensions is available in Abdi et al. (2012). We summarize the method next.

Dimensionality reduction in STATIS is achieved through two main steps: the first step, named interstructure, is a global analysis of the between data table structure based on diagonalization such as in Principal Components Analysis (PCA), providing a visualization of their global proximity.

Next, from the interstructure analysis an optimal set of weights are derived and used to compute a linear combination of data tables that best represents the common information in the different data tables. The resulting matrix is denoted compromise. The second step, named intrastructure, is an analysis of the within data table structure based on diagonalization of the compromise matrix.

In the following subsections, operational steps of the interstructure and intrastructure analyses are presented.

 ISSN: 2102-6238

3.1. *Interstructure analysis*

The interstructure analysis is based on the definition of a similarity measure between pairs of data matrices $\mathbf{X}^i$, with operational steps illustrated in Figure 3. STATIS starts by transforming each $\mathbf{X}^i$ into a $(T \times T)$ matrix of scalar products $\mathbf{W}^i$ through the following expression: $\mathbf{W}^i = \mathbf{X}^i\,(\mathbf{X}^i)'$, where $(\mathbf{X}^i)'$ denotes the transpose of $\mathbf{X}^i$.This step is necessary to obtain square matrices of same dimension T. Then, in order to compare two tables $\mathbf{X}^i$ and $\mathbf{X}^{i'}$, STATIS uses a similarity measure derived from the Hilbert-Schmidt's scalar product defined as $S_{ii'} = trace(\mathbf{DW}^i\,\mathbf{DW}^{i'})$, where $\mathbf{D}$ is a matrix of importance weights for the time instants. Usually $\mathbf{D} = \frac{\mathbf{I}}{T}$, where $\mathbf{I}$ denotes a $(T \times T)$ identity matrix; i.e. uniform weights are assigned to all time instants.

In general, normalized matrices $\mathbf{W}^i$ are used in the $S_{ii'}$ equation; such matrices are obtained through the operation $\frac{\mathbf{W}^i}{\sqrt{trace[(\mathbf{DW}^i)^2]}}$. When that is the case, $S_{ii'}$ gives the *RV* coefficient (for vector correlation) between $\mathbf{W}^i$ and $\mathbf{W}^{i'}$ as the result:

$$RV_{ii'} = \frac{trace(\mathbf{DW}^i\mathbf{DW}^{i'})}{\sqrt{trace[(\mathbf{DW}^i)^2]trace[(\mathbf{DW}^{i'})^2]}} \tag{3}$$

In the developments to follow, we assume the use of normalized matrices $\mathbf{W}^i$.

RV is non negative and scaled between 0 and 1; the closer to 1, the more similar the matrices $\mathbf{W}^i$ and $\mathbf{W}^{i'}$.

Once coefficients between every pair of matrices are available, they are organized into a $(N \times N)$ square matrix $\mathbf{S}$, which may be multiplied by a matrix Δ [1] containing importance weights for the batches. The resulting matrix is then diagonalized such as in *PCA*; the corresponding eigenvectors are denoted $\mathbf{u}^k$ and their associated eigenvalues by λ^k.

Similarities between the N batches are visualized by projecting matrices $\mathbf{W}^i$ onto the factorial axes retained in the interstructure representation (denoted IS graph). Typically a good graphical representation of batches is obtained by projecting their respective matrices $\mathbf{W}^i$ onto the first factorial plan (associated with the two *PCs* with largest eigenvalues).

3.2. *Intrastructure analysis*

The so-called intrastructure analysis uses a compromise matrix, denoted by $\mathbf{W}^{CO}$, given by the weighted sum of normalized matrices $\mathbf{W}^i$. Matrix $\mathbf{W}^{CO}$ is obtained as follows:

$$\mathbf{W}^{CO} = \sum_{i=1}^{N} \alpha_i^1 \mathbf{W}^i \tag{4}$$

where weights α_i^1 are obtained from the first eigenvector $\mathbf{u}^1$, associated with eigenvalue λ^1, after standardization into $\alpha^1 = \frac{\mathbf{u}^1}{\sqrt{\lambda^1}}$. As all elements of $\mathbf{S}$ are non negative, the first eigenvector $\mathbf{u}^1$ is a

[1] Matrix Δ, with dimension $(N \times N)$, can be determined analyzing products emerging from each batch in terms of conformity to specifications; batches yielding products closest to specification targets are given the largest importance weights. In case product conformity information is not available, we recommend assigning the same importance weight to all batches, i.e. $\Delta = \frac{\mathbf{I}}{N}$.

 ISSN: 2102-6238

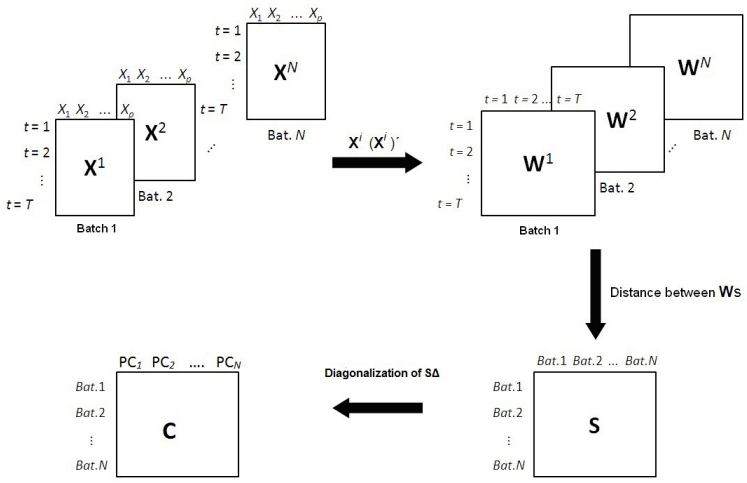

FIGURE 3. *Interstructure analysis-operational steps*

size factor and its elements will have the same sign. They reflect the overall similarity of a given batch with all other batches.

The optimal weights α_i^1 represent the agreement between data tables and the compromise (Abdi et al., 2012). If a batch highly differs from the others, it will have a weight close to 0 and less importance in the derivation of the compromise. Equation (4) leads to a compromise matrix that is robust to outliers, which is desirable in quality control applications, see Lavit et al. (1994).

Diagonalization of $\mathbf{W}^{CO}\mathbf{D}$ allows a visualization of artificial points B_t $(t = 1, ..., T)$ named compromise points. A summary of operational steps to perform this analysis is presented in Figure 4.

In opposition to matrix $\mathbf{S}$, $\mathbf{W}^{CO}$ preserves the information on the time index. Thus, it is possible to obtain graphical representations of the behavior of batches at each time instant. For that matter, each data matrix $\mathbf{W}^i$ is projected onto the factorial plans associated with the PCs retained in the PCA performed on $\mathbf{W}^{CO}\mathbf{D}$. Then selecting points corresponding to a given time instant t, we obtain a detailed representation of the variables joint behavior in each data matrix for observation t. Such plot, denoted CO graph, enables an easy interpretation of changes in the N matrices at each time instant.

4. Clustering for on line batch process control

The IS and CO graphs proposed in the former sections, are suitable for off-line monitoring of future batches. They are implemented after batch $N+1$ is finished, using the data matrix $\mathbf{X}^{N+1}$. On-line process control takes place as the new batch progresses in time. Thus, only a fraction of the new batch data matrix $\mathbf{X}^{N+1}$ will be available, and it is not possible to project the new batch on the graphs.

In this paper we introduce a new approach for on-line batch control based on the following assumption: the process reference behavior up to time instant t can be characterized by the first t rows in the reference data sets. Associated with each time instant t there will be a set of N partial

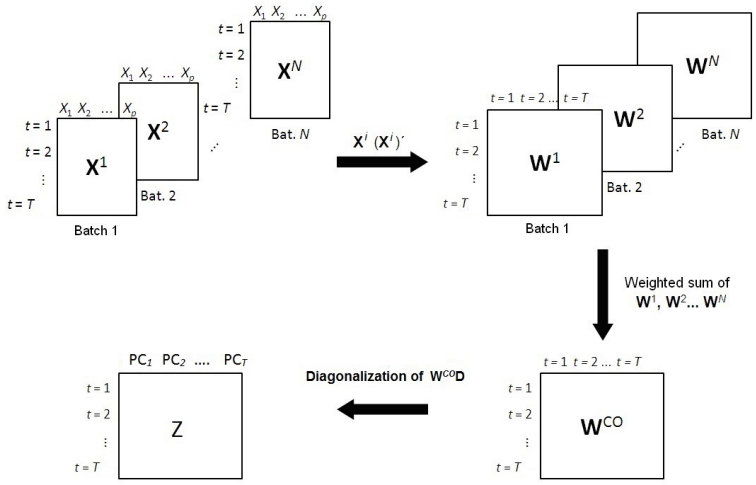

FIGURE 4. *Intrastucture analysis-operational steps*

reference tables of dimension $(t \times P)$, from which non-parametric control regions for the IS and CO charts may be derived.

However, such strategy could lead to a large number of control regions, some of which may be not informative. To overcome that we propose to determine interesting control time periods through clustering. A survey of clustering techniques may be found in Jain (2010).

In the context of a batch process as we have introduced it, the time instants are considered as items to cluster. As they are naturally ordered the desired clusters should be intervals of time instants. That is, batch duration should be split into intervals of time instants or periods. It is an optimal segmentation problem of multivariate series of ordered items. Thus temporally constrained clustering should be used.

Applying such clustering on a batch data set will provide a partition of the time instants into clusters, such that batch behavior will differ from one cluster to another. The cluster upper bounds provide the interesting time points to control. Our proposition is detailed next.

4.1. Temporally constrained clustering

In the univariate case, W.D. Fisher's method (Fisher, 1958) enables to optimally partition a set of T ordered objects O_t, $(t = 1, .., T)$, described by a variable $X = (x_1, ...x_t, ...x_T)$ into K clusters. The original method is based on the within class variance criterion. The problem is to find a partition $\underline{P}$ which minimizes the following expression:

$$W(\underline{P}) = \sum_{t=1}^{T} \sum_{k=1}^{K} p_t \delta_k^t (x_t - \bar{x}_k)^2 \tag{5}$$

 ISSN: 2102-6238

where $\bar{x}_k$ is the mean of x calculated in the *k-th* cluster, p_t is a weight assigned to object O_t , and $\delta_k^t = 1$ if O_t belongs to cluster k, otherwise $\delta_k^t = 0$. $W(\underline{P})$ can be rewritten as

$$W(\underline{P}) = \sum_{k=1}^{K} w(I_k) \tag{6}$$

where I_k $(k = 1, .., K)$ are clusters of $\underline{P}$ with $I_k = \{O_1, ..., O_t\}$ and $w(I_k)$ is the within class variance of X .

The W.D. Fisher's method is based on the fundamental following property, related to the additive nature of the within class variance criterion: if $\underline{P} = (\{O_1, ..., O_t\}, I_2, ..., I_K)$ optimally partitions the entire set of objects into K clusters, then $(I_2, ..., I_K)$ will optimally partition the set $\{O_{t+1}, ..., O_T\}$ into K-1 clusters. A detailed demonstration of this property may be found in Lechevallier (1990).

Using this relationship between optimal partitions into K clusters and optimal partitions into K- 1 clusters, Fisher's algorithm proceeds by successively computing optimal partitions into $2, 3, ..., K$ clusters. This is done using dynamic programming (Bellman, 1961) and enables to get an optimal partition of any set of ordered objects by exact optimization. The main steps of the algorithm are detailed in the appendix at the end of the paper.

From a geometric point of view, items can be considered as points on a straight line which have to be grouped in K clusters such that the sum of squared distances of the points to their centers of gravity is minimized. This allows a quite direct generalisation to the multivariate case.

We propose to use a multivariate extension presented in Lechevallier (1990) and used in Hébrail et al. (2010). In the extented approach, the ordered objects are described by a data matrix $\mathbf{X}$ comprised of P variables X^p $(p = 1, ..., P)$. The criterion in equation (6) is naturally extended to become a within class inertia criterion such that:

$$w(I_k) = \sum_{O_t \in I_k} \sum_{p=1}^{P} (x_t^p - \bar{x}_k^p)^2 \tag{7}$$

where $\bar{x}_k^p$ is the *p-th* coordinate of the *k-th* cluster center of gravity g_k. The criterion is additive and then the fundamental W.D. Fisher's property can be used to get an optimal partition of ordered objects described by several variables into K clusters.

The issue of choosing the number of clusters is, as for other direct clustering methods, a non trivial one. Some idea of the best number of clusters may be obtained by plotting the values of the criterion against K and looking for noticeable decrease.

4.2. Data tables for on-line batch process control

Applied to each batch data set $\mathbf{X}^i$, the multivariate extension of W.D. Fisher's method will provide a partition $\underline{P_i}$ of the time instants into K clusters of n_{ki} $(k = 1, ..., K)$ time points. Cluster sizes may vary slightly from one batch partition to another due to the variability that may exist between reference batches. Each partition provides a sequence of K instants $t_{ki} = \sum_{l=1}^{k} n_{li}$ corresponding to the number of rows in the cumulative partial reference data sets.

Combining results from the N partitions gives, for each k, a set of N values $t_{ki}, i = 1...N$, many of them being equal due to the slight variation in cluster sizes. These values determine critical

 ISSN: 2102-6238

TABLE 1. *Values of α and l used in the definition of the CC control region*

Probability α	l
0.01	1.68
0.05	1.13
0.10	0.86
0.25	0.43

periods (from $\min_i t_{ki}$ to $\sup_i t_{ki}$) during which the process has to be particulary monitored. The last instant t_K will be the one in the complete reference sample.

We propose to implement the on-line control regions at time instants $t_k = \sup_i t_{ki}$ to increase the probability to detect an out of control signal. Controling the process at the end of the critical period, i.e. at the latest instant will allow better detection of process shifts that occur before instant t_k, using the classical quality control assumption that a shift in process behavior will remain for the rest of the batch duration, unless corrective actions have been taken.

Alternatively other strategies may be considered, such as building a control chart for each different value of t_{ki}.

Finally this clustering step will provide $K-1$ samples of N subtables from the reference data tables. Each of these sets of subtables will be used for the on-line monitoring.

5. Non-parametric batch process monitoring

We now introduce the method to build the control regions from the factorial planes provided by the application of STATIS to the reference sample complete data sets for off-line monitoring, and to the K-1 samples of N subtables from the reference data tables obtained after the clustering step for on-line monitoring.

To obtain the control regions, we adapt a proposition in Zani et al. (1998) based on convex hull peeling, which was originally conceived for graphical detection of outliers in two-dimensional data sets. It consists of the following three steps:

step 1: a convex hull peeling procedure is performed on the factorial plane to get the convex hull such that a proportion π [2] of the points in the graph fall within its boundaries. A so-called inner region is obtained by adjusting a B-spline to the outermost points of the 50% convex hull. The B-spline provides a smooth contour to the hull; an alternative (and simpler) approach is to connect the outermost points in the hull with straight lines, but the resulting contour tends to be irregular in shape, in particular when the number of points is small.

step 2: a robust centroid is determined as the mean of observations inside the inner region. The smaller the proportion π, the more robust the centroid.

step 3 : Consider the distance δ between the robust centroid and a boundary point; a multiple l of this distance will be added to δ to establish the control region. The value of l is determined according to the desired false alarm probability α using Table 1 (Zani et al., 1998) (p. 267). When $\alpha = 0.01$, for example, l =1.68.

Alternatively, the CC control region may be determined defining a convex hull containing (1-α) of the data, identifying the boundary points, and adjusting a B-spline to these points. However,

[2] π is equal to 50% of the points in Zani et al. (1998), but a larger proportion may also be used.

 ISSN: 2102-6238

the procedure proposed here where the control region is expanded from the 50% hull tends to yield smooth contours, less sensitive to extreme points or outliers in the data set.

The procedure will be illustrated in section 6.

For off-line control, the application of STATIS will yield the IS (Interstructure) and CO (Compromise) factorial planes of reference batches. From these reference data clouds, both IS and CO control regions are obtained following the procedure described above.

The IS control region used for off-line monitoring of future batches is implemented after batch N+1 is finished, projecting the data matrix $\mathbf{X}^{N+1}$ onto the graph. When the new batch has its coordinates falling inside the IS control region it is considered as in control. Otherwise, the projection yields an out-of-control signal and the CO charts are analyzed to identify the point in time in which the process departed from the reference behavior.

For on-line control, the same scheme is successively used at each of the K-1 determined instants t_k, with STATIS applied to the K-1 samples of N subtables from the reference data tables obtained after the clustering step. There will be K-1 control regions available for on-line monitoring of the process: a future batch N+1 progresses up to a instant t_k, and the corresponding data matrix will be projected onto the associated IS control region. The batch will be deemed out-of-control in case its coordinates fall outside the control region.

6. Application

In what follows, we exemplify our propositions on a real data set used in the batch process literature (Nomikos and MacGregor, 1995; Eriksson et al., 2001). Data come from a batch polymerization reactor: 18 reference batches are selected to represent the process normal behavior. Additionally, a set of 11 batches is available to test the performance of the methods. This set contains 4 good batches and 7 bad batches. Since our method was conceived to capture abnormal behaviors of batches along the time axis, we expect bad batches to appear as out-of-control points in the proposed control charts.

For each batch, 10 variables are recorded at 100 time instants. Variables X_1, X_2, X_3, X_6 and X_7 are temperature measurements, variables X_4, X_8 and X_9 are pressure measurements, and variables X_5 and X_{10} represent flow rates of materials added to the reactor.

6.1. Off-line control of batch process

Analyzing the 18 reference data sets using the STATIS method, we obtained graphical representations of its interstructure and compromise behavior, which allow us to build IS and CO control charts with a 99% control region. Methods were implemented using the SAS software.

For the interstructure analysis, the two first eigen-pairs were retained representing 99.6% of the total variabilitity in the data. Figure 5 shows the 18 batches in the reference sample projected in the retained first factorial plan on which the convex hull peeling is applied.The second convex hull containning 7 points (little less than 50% of the points) determines the inner region. Figure 6 shows the result of adjusting a B-spline to the five outermost points of the inner region yielding the contour displayed in the graph. Figure 7 illustrates the IS control chart building: according to the second step of the procedure described in section 5, the centroid (point C in the graph) of the inner region is given by the average of the observations within it, and we define a false alarm

 ISSN: 2102-6238

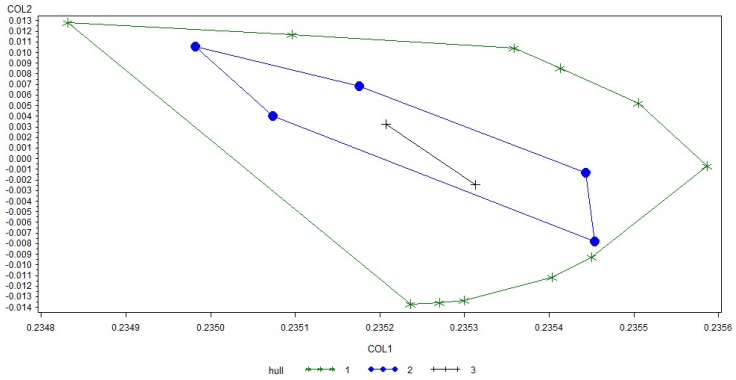

FIGURE 5. *Convex hull peeling on the first interstructure factorial plane of 18 reference batches: the second most extreme hull with dot symbols determines the inner region contour*

TABLE 2. *Compromise coefficients* α_i^1 *for each table*

Table	1	2	3	4	5	6	7	8	9
α	0.0557	0.0557	0.0556	0.0556	0.0557	0.0557	0.0557	0.0556	0.0557
Table	10	11	12	13	14	15	16	17	18
α	0.0557	0.0557	0.0557	0.0557	0.0557	0.0555	0.0556	0.0556	0.0556

probability $\alpha = 0.01$ with corresponding value l =1.68. Then, applying the third step, we get the second more external contour shown in Figure 7. This region determines the off-line IS control chart.

The 18 batches in the reference sample (points R, in the graph) were projected in the first factorial plan and two of them are positioned slightly outside the control region (see Figure 8). These false alarms may be due to the fact that we use only five out of eighteen points to build the control region.

Figure 9 shows the off-line control results for the supplementary batches. All 7 bad batches (points B, in the graph) were detected as out-of-control, with stronger signal in 6 of them. The out-of-control batch close to the boundaries was diagnosed as bad by experts and as having different behavior from other bad batches; it is generally not detected to be out of control in other monitoring approaches (Eriksson et al., 2001). The good batches (points G, in the graph) also fall within the control region, as expected, but with a false alarm for one of them. These results are similar to those obtained in Eriksson et al. (2001), but the authors combined several charts (based on Hotelling's T^2 distribution and tolerance regions on PCA score plots) while our method is based on a single non parametric control chart.

The first eigenvector from the IS analysis gives the compromise coefficients α_i^1 for table i= 1 to 18; see Table 2.

All tables have very similar weights in the compromise linear combination. That means the compromise reflects well the information in each of reference data sets i.e. their common structure.

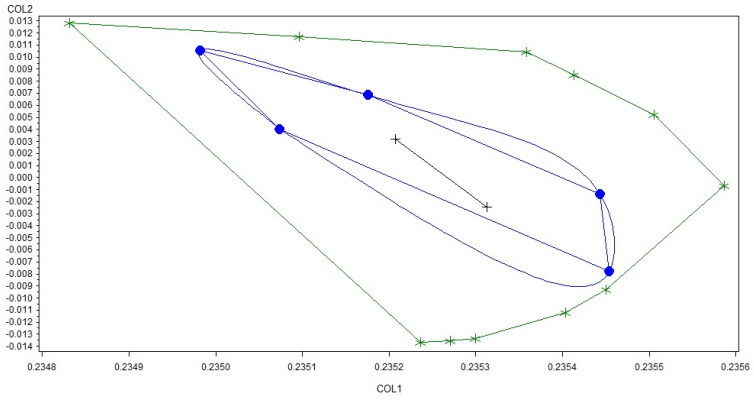

FIGURE 6. *Convex hull peeling, inner region and its corresponding B-spline curve*

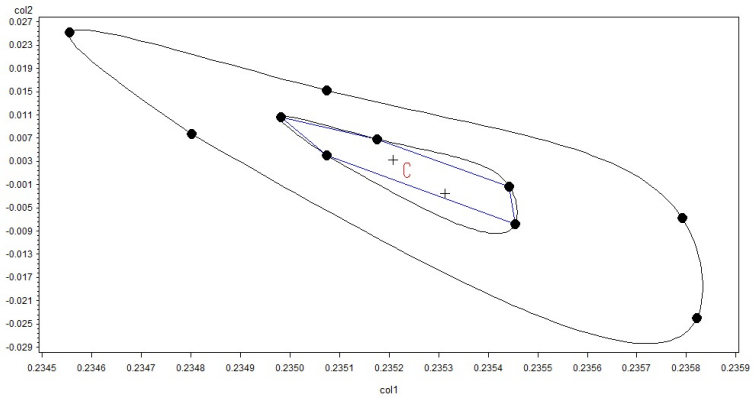

FIGURE 7. *Control region*

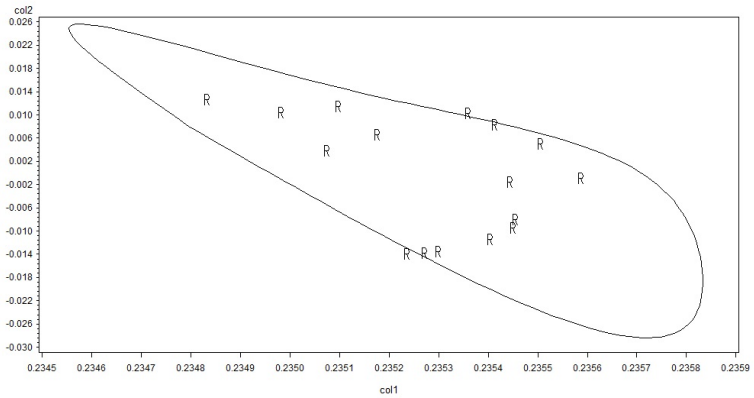

FIGURE 8. *Off line control chart- Plot of reference batches, labelled R*

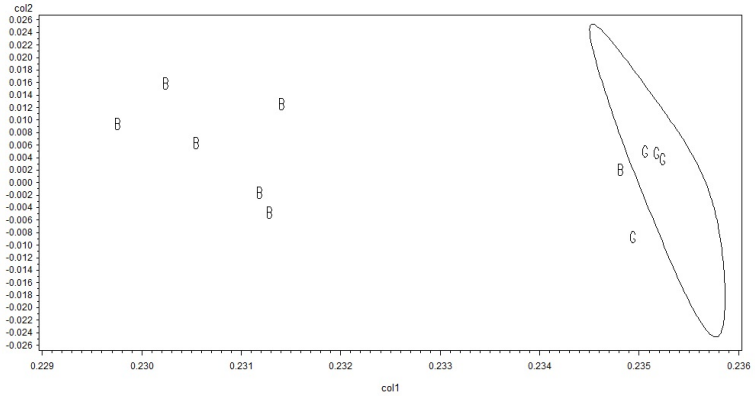

FIGURE 9. *Off line control chart- Plot of supplementary batches labelled G for good batches and B for bad batches*

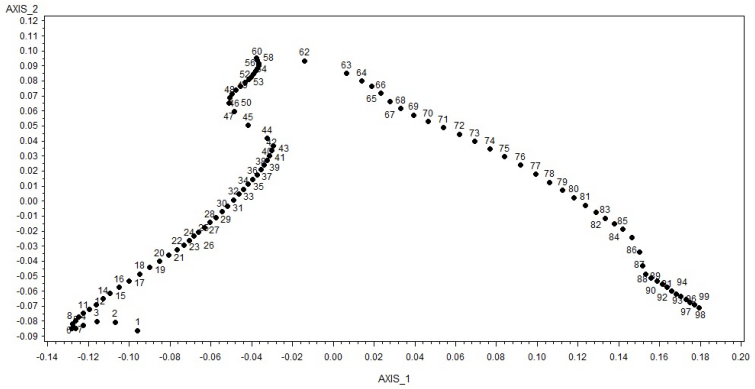

FIGURE 10. *Compromise plot of the 100 time instants*

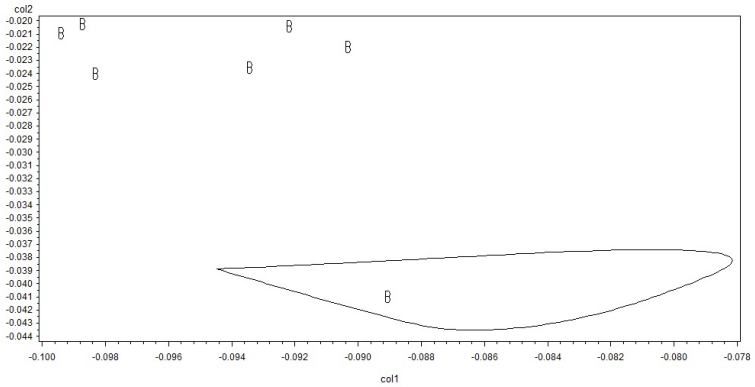

FIGURE 11. *CO control chart for time instant 20 - Plot of bad batches labelled B*

TABLE 3. *Values of upper bound for clusters: each reference table is partitioned into seven clusters*

Cluster	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
1	15	16	16	17	17	15	15	16	16
2	29	30	30	30	31	29	29	29	30
3	46	46	47	46	47	45	45	45	46
4	63	63	64	63	64	62	62	62	63
5	76	76	77	75	76	74	75	74	75
6	86	86	87	86	88	85	85	85	85
7	100	100	100	100	100	100	100	100	100
Cluster	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}	P_{17}	P_{18}
1	16	16	16	17	17	18	17	17	17
2	29	30	30	31	31	31	31	31	31
3	45	46	46	47	46	46	45	46	46
4	63	63	63	64	63	61	61	62	61
5	75	75	75	76	75	74	73	73	73
6	85	86	85	87	86	86	86	86	87
7	100	100	100	100	100	100	100	100	100

It is due to the fact that the reference sample is comprised exclusively of data from batches that yielded products within specifications.

The compromise matrix obtained using equation (4) is then diagonalized and the two first eigen pairs were retained representing 86.8% of the total variability among compromise points. Figure 10 displays the compromise position of time instants onto the first factorial plan.

Data matrices $\mathbf{W}^i$ were then projected on this plan allowing the graphical visualization of all time instants for all tables. Selecting points corresponding to a given time instant provides the CO graph presented in section 3.2 from which the CO control charts are derived. Figure 11 corresponds to time instant 20 and shows detection of departures from the reference model in the 6 batches identified as strongly out-of-control on the off-line control chart.

6.2. On-line control of batch process

We apply the method proposed for on-line control to the 18 batches of the reference sample. The multivariate extension of the W.D.Fisher's constrained clustering method was applied to each batch with K, the number of clusters in the partitions, varying from 3 to 20. We grouped the 100 time instants into 7 clusters which were the most similar regarding the interval bounds, giving the shortest critical period as explained in section 4. Table 3 shows the values of the clusters' upper bounds. For example, for cluster 1 the upper bounds vary from 15 to 18, and then the latest first instant was choosen equal to 18.

The final instants of the critical periods were 18, 31, 47, 64, 77, 88 and correspond to the time instants for on-line control. For the first on-line control chart at time instant t=18 , the first 18 rows of the complete reference tables were used to form the reference sample on which STATIS is applied. The two first eigen-pairs were retained representing 98.5% of the total variablitity in the partial data. The associated control chart is shown in Figure 12. Only five out of seven bad batches are detected to be out-of-control. The good batches fall inside the control region.

For the next on-line control chart, at time instant t=31, the first 31 rows of the complete

 ISSN: 2102-6238

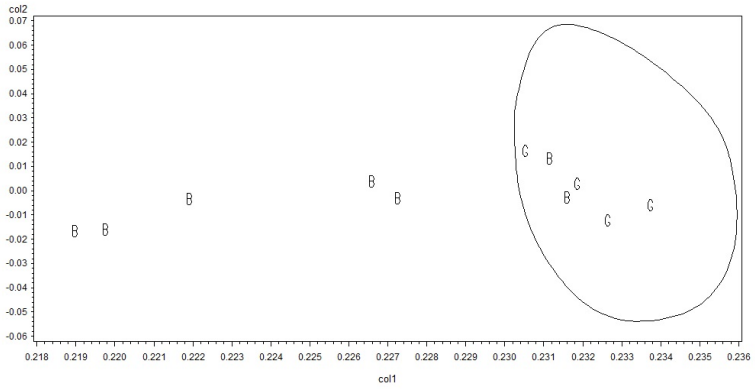

FIGURE 12. *On line control chart for time instant 18 - Plot of supplementary batches labelled G for good batches and B for bad batches*

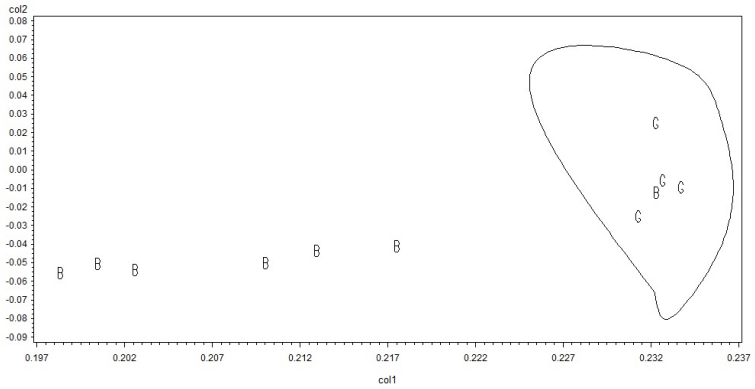

FIGURE 13. *On line control chart for time instant 31 - Plot of supplementary batches labelled G for good batches and B for bad batches*

reference tables were used to form the reference sample on which STATIS is applied. The two first eigen-pairs were retained representing 98.6% of the total variablitity in the partial data. The associated control chart is shown in Figure 13.The good batches are found to be in control. One more bad batch is signalized out-of-control.

The control chart for time instant 64 signalizes the boundary bad batch. That explains the weak departure from the reference distribution identified for that batch in section 6.1: the batch only starts presenting abnormal behavior in latter time instants in its trajectory. It is also noteworthy that the off-line CO charts signalize the other 6 bad batches at time instant 20, i.e. earlier than the on-line charts. That may be explained by the fact that in the off-line control scheme information from all time instants are already available, and the charts become more sensitive to abnormal situations in the batches.

 ISSN: 2102-6238

7. Conclusion

In this paper we have presented a method for quality control of batch processes with same duration. In the proposed method data matrices are directly analyzed using the STATIS method for off-line control. Process monitoring is implemented using two non-parametric control charts. In the IS CC the global behavior of batches may be verified combining all time instants and process variables, with respect to a reference sample containing batches that yielded products within specifications; such control chart is indicated for off-line control of batches. In the CO charts the trajectory of process variables at each time instant is analyzed, and significant detours from the reference trajectories may be detected. After applying a constrained clustering step on the reference dataset, the method quite directly allows the on-line control of batches.

The proposed method is applied on a real dataset where ten process variables are monitored in 100 time instants. The obtained results illustrate the good performance of the proposed control charts, both for off-line and on-line monitoring of the process.

Dimension reduction of process data using DUAL-STATIS would allow monitoring of batches with variable duration (Niang et al., 2009). That is left for future research.

Other natural extensions of the work presented here could include (i) a more formal evaluation of the method's performance; (ii) a comparative study of results obtained using the proposed method and other methods available in literature, with special emphasis on the approach based on dynamic time warping proposed by Kassidas et al. (1998); and (iii) the development of diagnosis methods for the out-of-control points signalized in the proposed control charts.

APPENDIX: Fisher algorithm main steps [3]

Consider T ordered objects labelled $(1,2,...,T)$

Clusters are constrained to be intervals of objects $(I,I+1,...,J)$.

step 1: compute $w(I,J)$ for the cluster $(I,I+1,...,J)$ for all I, J such that $1 \leq I < J \leq T$

step 2: optimal solution for K=2
For I, $2 \leq I \leq T$, compute $W(I,K) = min[w(1,J-1)+w(J,I)]$ over the range $2 \leq I < J \leq I$

step 3: optimal solutions for k= 3 to K
For L, $3 \leq I \leq k$, compute $W(I,L)$, $L \leq I \leq T$ by
$W(I,L) = min[W(J-1,L-1)+w(J,I)]$ over the range $L \leq J \leq I$

step 4: final step: $\underline{P}(T,K)$ optimal partition of T objects into K clusters is obtained from the values $w(I,J)$ and $W(I,L)$ which have to be stored, by

a) first finding J so that : $W(T,K) = W(J-1,K-1)+w(J,T)$. The last cluster is then $(J, J+1, ... , T)$

b) next finding J' so that $W(J-1,K) = W(J-1,K-1)+w(J',J-1)$. The second-to-last cluster is then $(J', J'+1, ... , J-1)$

[3] R software was used to write the corresponding program

 ISSN: 2102-6238

c) an so on...

Acknowledgements

We would like to thank the anonymous reviewers for their comments which helped improving the quality of our work. We would also like to acknowledge the support provided by doctoral student Mory Ouattara in R and Latex programming. Prof. Fogliatto's research is supported by CNPq (Grant no. 303059/2011-7).

References

Abdi, H., Williams, L. J., Valentin, D., and Bennani-Dosse, M. (2012). Statis and distatis: optimum multitable principal component analysis and three way metric multidimensional scaling. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):124–167.

Bellman, R. (1961). On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6):284–284.

Castagliola, P. and Ferreira, A. (2006). Monitoring of Batch Processes with Varying Durations Based on the Hausdorff Distance. *International Journal of Reliability, Quality and Safety Engineering*, 13(3):213–236.

des Plantes, H. L. (1976). *Structuration des Tableaux à Trois Indices de la Statistique*. PhD thesis, Université de Montpellier, Montpellier (France).

Doan, X. and Srinivasan, R. (2008). Online monitoring of multi-phase batch processes using phase-based multivariate statistical process control. *Computers and Chemical Engineering*, 32:230–243.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., and Wold, S. (2001). *Multi-and Megavariate Data Analysis*. Umetrics, 2nd edition.

Escoufier, Y. (1987). Three-mode data analysis: the statis method. In N.C., F. B. . L., editor, *Methods for multidimensional data analysis*, pages 259–272. ECAS.

Fisher, W. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53(284):789–798.

Flores-Cerrillo, J. and MacGregor, J. (2002). Control of particle size distribution in emulsion semibatch polymerization using mid-course correction policies. *Industrial & Engineering Chemistry Research*, 41:1805–1814.

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

Hastie, T., Tibshirani, R., and Friedman (2001). *The elements of statistical learning: data mining, inference and prediction*. Springer.

Hébrail, G., Hugueney, B., Lechevallier, Y., and Rossi, F. (2010). Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomputing*, 73(7):1125–1141.

Jackson, J. (1991). *A User's Guide to Principal Components*. Wiley.

Jackson, J. E. and Mudholkar, G. S. (1979). Control procedures for residuals associated with principal component analysis. *Technometrics*, 21(3):341–349.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

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.

Kassidas, A., MacGregor, J. F., and Taylor, P. A. (1998). Synchronization of batch trajectories using dynamic time warping. *AIChE Journal*, 44(4):864–875.

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(1):93–109.

Kourti, T. and MacGregor, J. F. (1996). Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology*, 28(0):409–428.

Lavit, C., Escoufier, Y., Sabatier, R., and Traissac, P. (1994). The ACT (STATIS method). *Computational Statistics & Data Analysis*, 18(1):97–119.

Lechevallier, Y. (1990). *Recherche d'une partition optimale sous contrainte d'ordre total*. Rapport de recherche INRIA RR-1247.

 ISSN: 2102-6238

Liu, R. Y. (1995). Control charts for multivariate processes. *Journal of the American Statistical Association*, 90(432):1380–1387.

Liu, R. Y. and Tang, J. (1996). Control charts for dependent and independent measurements based on bootstrap methods. *Journal of the American Statistical Association*, 91(436):1694–1700.

Lombardo, R., Vanacore, A., and Durand, J. (2008). Non parametric control chart by multivariate additive partial least squares via spline. In Preisach, C., Burkhardt, H., Schmidt-Thieme, L., and Decker, R., editors, *Data Analysis, Machine Learning and Applications*, pages 201–208. Springer.

Lowry, C. and Montgomery, D. (1995). A review of multivariate control charts. *Computational Statistics & Data Analysis*, 27(1):800–810.

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

Montgomery, D. (2001). *Introduction to Statistical Quality Control*. John Wiley & Sons.

Niang, N., Fogliatto, F., and Saporta, G. (2009). Batch process monitoring by three-way data analysis approach. In *The XIIIth International Conference Applied Stochastic Models and Data Analysis ASMDA-2009*, pages 463–468.

Nomikos, P. and MacGregor, J. (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37:41–59.

Rosa, A. (2005). *Maîtrise statistique de procédés par lots à temps variable*. PhD thesis, Université de Nantes, Nantes (France).

Scepi, G. (2002). Parametric and non parametric multivariate quality control charts. In C., L., J., A., V., E., and G., S., editors, *Multivariate Total Quality Control*, pages 163–189. Physica-Verlag.

Wierda, S. (1994). Multivariate statistical process control - recent results and directions for future research. *Statistica Neerlandica*, 48:147–168.

Zani, S., Riani, M., and Corbellini, A. (1998). Robust bivariate boxplots and multiple outlier detection. *Computational Statistics & Data Analysis*, 28(3):257–270.

 ISSN: 2102-6238