International Journal of Reliability, Quality and Safety Engineering

MONITORING OF BATCH PROCESSES WITH VARYING DURATIONS BASED ON THE HAUSDORFF DISTANCE

PHILIPPE CASTAGLIOLA
IRCCyN / IUT de Nantes, La Chantrerie, Rue Christian Pauc, BP 50609
44306 Nantes, France

and

ARIANE FERREIRA PORTO ROSA
IRCCyN / Ecole des Mines de Nantes, La Chantrerie, 4 rue Alfred Kastler, BP 20722
44307 Nantes, France

Received (received date)
Revised (revised date)

In some industrial situations, the classical assumption used in the batch process monitoring that all batches have equal durations and are synchronized does not hold. A batch process is carried out in sequential phases and a significant variability generally occurs in the *duration of the phases* such that events signifying the beginning or the end of a phase are generally misaligned in time within the various batches. The consequence is that the variable trajectories, in the different runs of the same batch process, are unsynchronized. In this case, data analysis from process for performing the multivariate statistical process control can be difficult. In this paper, we propose several innovative methods for the *off-line* and *on-line* monitoring of batch processes with varying durations, all based on the Hausdorff distance. These methods have been successfully tested on a simulated example and on an industrial case example. The conclusion is that these methods are able to efficiently discriminate between nominal and non-nominal batches.

1. Introduction

A batch process is a discontinuous system of production particularly used in the chemical & pharmaceutical industries (see Vanbergen[1]). The use of batch processes in the production and the treatment of raw materials and products presented a significant growth during the last decades and represents today an alternative to the classical continuous mode of production. One of the main reason that makes the use of batch processes more and more attractive is the possibility to produce small or medium series depending on the changes of the economic situation of the market. For example, in the field of the fine chemistry, the trend is to put on the market specialized products (with very specific and high added values) for which the required quantities are small. The batch process approach corresponds to the increasing request in *customized manufacturing* (i.e. the production of products with characteristics defined for each customer) Davis[2] and in *Mass Customization* (i.e. the manufacturing adapted to the customer requirements) Da Silveira[3]. A batch process can be characterized in 3 sequential phases:

1. raw materials are loaded in a container,

2. these raw materials are processed (physical treatment, mixture, chemical reactions, ...) during a certain period corresponding to the duration of the batch,
3. at the end of this period the final product is unloaded.

A significant variability generally occurs in the duration of these phases, so that events signifying the beginning or the end of a phase are generally misaligned in time within the various batches. During the time of the treatment, the trajectory of several variables (such as temperature, pressure, acidity, ...) can be measured. After the end of each batch, the product is analyzed in order to check if it corresponds to the desired standards of quality. Generally, there is a significant variability within the durations of the phases in a batch and within the various batches as well. These variations are considered as acceptable small deviations of the variable trajectories around their nominal/reference trajectories. If one (or several) special causes affect the process during a given batch run, this will generate significant deviations on the nominal trajectories of one or several process variables.

The traditional process control consists in building fundamental models that are based on the theoretical equations corresponding to the mass, energy and momentum balances and in optimizing over these models or using them for the process monitoring and fault detection. However, in many batch processes there are measurements on other variables that often represent the electrical and mechanical parts of the process, such as pump speed, agitator power, for example. These variables can rarely be included in fundamental models. For this reason, the use of empirical models based on plant data has been recommended in complement to the traditional process control.

The methods developed during the last decade by Nomikos & MacGregor[4] for the monitoring of batch processes are based on the measurement of a deviation from a "nominal" behavior of the process, summarized by a large number of process variable trajectories collected throughout of the batch. These methods consist in the application of multivariate control charts, in the use of multivariate statistical projection methods for the on-line analysis, fault detection and process diagnosis. The use of multivariate statistical projection methods like the Multiway Principal Component Analysis (MPCA) and the Multiway Partial Least Squares (MPLS) have been thoroughly investigated by Kourti et al.[5], Martin et al.[6] and Wold et al.[7]. Other multivariate statistical projection methods such as PARAllel FACtor analysis (PARAFAC) and Tucker models are suggested for the treatment of batch data by Bro[8] and Smilde[9].

However, in some industrial situations, the classical assumption used in the batch process monitoring that all batches have equal durations and are synchronized does not hold. In many industries, different runs of the same batch process may have a different duration. Another common situation arise when the duration of various stages within the batches (or process transitions) is not the same. As examples, we can cite the case of batch processes where an exothermic reaction occurs that varies in duration between summer and winter, and the case of polymerization reactors where there can be batch-to-batch variations in impurities and in the initial charges of the raw materials.

Generally, in these processes, the determination of the total operational time of a single batch depends of the value required for one or more variables on the process. For example, a batch polymerization can reach the same level of monomer conversion at different time lengths and impurities or fluctuations in the rate of heat removal may affect this length. Another case is where variables are not present during the entire duration of the batch. In all these cases, it is necessary to do both alignment and synchronization of the variable trajectories before applying the multivariate statistical process control. A review of the methods for the statistical process control of batch processes with varying durations has been presented in Kourti[10,11]. More recent works on this subject have been presented by Kaistha et al.[12,13].

The work presented in this paper addresses the case of batch processes with varying durations and the aim of this paper is to present three innovative methods for both *off-line* and *on-line* monitoring without the need of alignment and/or synchronization strategies. All the proposed methods are based on a specific distance called the *Hausdorff Distance*.

2. Batch data

We assume that we processed $i = 1,\ldots,I$ "reference" batches for which the quality of the final product was proven to be high enough. For each of these batches, we monitored $j = 1,\ldots,J$ process variables from time $k = 1$ up to time $k = K_i$ (completing time) that depends on the batch index i (varying duration). These $i = 1,\ldots,I$ "reference" batches reflect the natural variability and correlation of the $j = 1,\ldots,J$ process variables under nominal operation conditions. Let $x_{i,j,k}$ be the value of the variable $j = 1,\ldots,J$ at time $k = 1,\ldots,K_i$ for the batch $i = 1,\ldots,I$ and let $K = \min(K_i)$, $i = 1,\ldots,I$, be the total duration of shortest batch. All the data $x_{i,j,k}$ could be represented in a (non cubic) three-way array $\underline{\mathbf{X}}$ but, instead, we found more convenient to define the following matrices

$$\mathbf{X}_{i,j}^{(k)} = \begin{pmatrix} 1 & x_{i,j,1} \\ 2 & x_{i,j,2} \\ \vdots & \vdots \\ k & x_{i,j,k} \end{pmatrix} \qquad \mathbf{X}_{i}^{(k)} = \begin{pmatrix} x_{i,1,1} & x_{i,2,1} & \cdots & x_{i,J,1} \\ x_{i,1,2} & x_{i,2,2} & \cdots & x_{i,J,2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i,1,k} & x_{i,2,k} & \cdots & x_{i,J,k} \end{pmatrix}$$

where $\mathbf{X}_{i,j}^{(k)}$ is the $(k,2)$ matrix describing the trajectory of variable j, for batch i, up to time k, and where $\mathbf{X}_{i}^{(k)}$ is the (k,J) matrix describing the trajectory of all variables $j = 1,\ldots,J$, for batch i, up to time k. For simplicity, we also defined matrices $\mathbf{X}_{i,j} = \mathbf{X}_{i,j}^{(K_i)}$ and $\mathbf{X}_{i} = \mathbf{X}_{i}^{(K_i)}$. Any new batches that will occur in the future will be indexed $i^* = I + 1,\ldots,I + n$.

3. The Hausdorff distance

The Hausdorff distance is the core of the methods that will be presented in the next section. Let $\mathbf{X}$ and $\mathbf{Y}$ be two polylines of $\mathbb{R}^p$ with respectively m and n vertices

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_m^T \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_n^T \end{pmatrix}$$

where $\mathbf{x}_i^T = (x_{i,1},\ldots,x_{i,p})$ and $\mathbf{y}_j^T = (y_{j,1},\ldots,y_{j,p})$. Let $d_2(\mathbf{x}_i,\mathbf{y}_j)$ be the euclidian distance between $\mathbf{x}_i$ and $\mathbf{y}_j$, i.e.

$$d_2(\mathbf{x}_i,\mathbf{y}_j) = \left(\sum_{k=1}^{p} (x_{i,k} - y_{j,k})^2 \right)^{1/2}$$

Let $h_1,h_2,\ldots,h_{m+n}$ be $m + n$ distances defined by

$$\begin{aligned} h_i &= \min_{j=1,\ldots,n} (d_2(\mathbf{x}_i,\mathbf{y}_j)) \qquad (i = 1,\ldots,m) \\ h_{m+j} &= \min_{i=1,\ldots,m} (d_2(\mathbf{x}_i,\mathbf{y}_j)) \qquad (j = 1,\ldots,n) \end{aligned}$$

By definition, the Hausdorff distance $d_H(\mathbf{X},\mathbf{Y})$ between the polylines $\mathbf{X}$ and $\mathbf{Y}$ is equal to

$$d_H(\mathbf{X},\mathbf{Y}) = \tilde{h}$$

where $\tilde{h}$ is the sample median of the $m+n$ distances $h_1,h_2,\ldots,h_{m+n}$. By construction, the Hausdorff distance is symmetrical, i.e. $d_H(\mathbf{X},\mathbf{Y}) = d_H(\mathbf{Y},\mathbf{X})$. We can also notice that for the calculation of the $h_1,h_2,\ldots,h_{m+n}$, alternate distances can be chosen, like

$$d_1(\mathbf{x}_i,\mathbf{y}_j) = \sum_{k=1}^{p} |x_{i,k} - y_{j,k}|$$

or

$$d_\infty(\mathbf{x}_i,\mathbf{y}_j) = \max_{k=1,\ldots,p} |x_{i,k} - y_{j,k}|$$

A possible generalization of the Hausdorff distance is to replace the definition $d_H(\mathbf{X},\mathbf{Y}) = \tilde{h}$ by $d_H(\mathbf{X},\mathbf{Y}) = \hat{h}_\alpha$, where $\hat{h}_\alpha$ is the α-percentile, and in particular $\tilde{h} = \hat{h}_{0.5}$. In our paper, we will use the Hausdorff distance based on the sample median.

4. Off-line methods

In this section, we will propose three new methods (denoted #1, #2 and #3) for the *off-line* monitoring of batch processes with varying durations. Each *off-line* method #1, #2 and #3 consists in two sequential phases:

- **Phase 1** (data analysis based on the "reference" batches) consists in computing a (I,J) matrix $\mathbf{Z}$ for method #1 or a $(I-1,J)$ matrix $\mathbf{Z}$ for methods #2 and #3

$$\mathbf{Z} = \begin{pmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,J} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,J} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

- **Phase 2** (monitoring of future batches) consists in computing a $(1,J)$ row vector $\mathbf{z}_{i^*} = (z_{i^*,1},z_{i^*,2},\ldots,z_{i^*,J})$ for each new batch $i^* = I+1,\ldots,I+n$.

The way of computing both matrix $\mathbf{Z}$ and vector $\mathbf{z}_{i^*}$ will depend on the selected method. These methods are described below.

4.1. *Off-line method #1*

- **Phase 1** consists in computing I row vectors $\mathbf{z}_1,\ldots,\mathbf{z}_I$ in $\mathbb{R}^J$, where $\mathbf{z}_i = (z_{i,1},\ldots,z_{i,J})$ and where $z_{i,j}$ is defined by:

$$z_{i,j} = \min_{\substack{\ell=1,\ldots,I \\ \ell\neq i}} (d_H(\mathbf{X}_{i,j},\mathbf{X}_{\ell,j}))$$

- **Phase 2** consists in computing the row vector $\mathbf{z}_{i^*} = (z_{i^*,1},\ldots,z_{i^*,J})$ for each new batch $i^* = I+1,\ldots,I+n$, where $z_{i^*,j}$ is defined by:

$$z_{i^*,j} = \min_{\ell=1,\ldots,I} (d_H(\mathbf{X}_{i^*,j},\mathbf{X}_{\ell,j}))$$

4.2. *Off-line method #2*

– **Phase 1** consists in computing $I-1$ row vectors $\mathbf{z}_1,\ldots,\mathbf{z}_{I-1}$ in $\mathbb{R}^J$, where $\mathbf{z}_i=(z_{i,1},\ldots,z_{i,J})$. In order to compute the components $z_{i,j}$, we firstly have to find the batch index R which is supposed to be "the most representative" among the I reference batches using the following equation:

$$\min_{i=1,\ldots,I}\left(\max_{\substack{\ell=1,\ldots,I\\ \ell\neq i}} d_H(\mathbf{X}_i,\mathbf{X}_\ell)\right)=\max_{\substack{\ell=1,\ldots,I\\ \ell\neq R}} d_H(\mathbf{X}_R,\mathbf{X}_\ell) \tag{1}$$

and then to compute

$$z_{i,j}=d_H(\mathbf{X}_{i,j},\mathbf{X}_{R,j})$$

– **Phase 2** consists in computing the row vector $\mathbf{z}_{i^*}=(z_{i^*,1},\ldots,z_{i^*,J})$, for each new batch $i^*=I+1,\ldots,I+n$, where $z_{i^*,j}$ is defined by:

$$z_{i^*,j}=d_H(\mathbf{X}_{i^*,j},\mathbf{X}_{R,j})$$

4.3. *Off-line method #3*

– **Phase 1** consists in computing $I-1$ row vectors $\mathbf{z}_1,\ldots,\mathbf{z}_{I-1}$ in $\mathbb{R}^J$, where $\mathbf{z}_i=(z_{i,1},\ldots,z_{i,J})$. In order to compute the components $z_{i,j}$, we firstly have to find, for each variable $j=1,\ldots,J$, the batch index R_j which is supposed to be "the most representative" using the following equation

$$\min_{i=1,\ldots,I}\left(\max_{\substack{\ell=1,\ldots,I\\ \ell\neq i}} d_H(\mathbf{X}_{i,j},\mathbf{X}_{\ell,j})\right)=\max_{\substack{\ell=1,\ldots,I\\ \ell\neq R_j}} d_H(\mathbf{X}_{R_j,j},\mathbf{X}_{\ell,j}) \tag{2}$$

and then to compute

$$z_{i,j}=d_H(\mathbf{X}_{i,j},\mathbf{X}_{R_j,j})$$

– **Phase 2** consists in computing the row vector $\mathbf{z}_{i^*}=(z_{i^*,1},\ldots,z_{i^*,J})$, for each new batch $i^*=I+1,\ldots,I+n$, where $z_{i^*,j}$ is defined by:

$$z_{i^*,j}=d_H(\mathbf{X}_{i^*,j},\mathbf{X}_{R_j,j})$$

Methods #2 and #3 are very similar indeed. The main difference is that for method #2 there is a single "most representative" batch index R, while for method #3, there are J "most representative" batch indices R_j, one per variable.

5. On-line methods

In this section, we will propose three new methods (denoted #1, #2 and #3) for the *on-line* monitoring of batch processes with varying durations. Like the off-line methods presented in the previous section, each *on-line* method #1, #2 and #3 consists in two sequential phases:

– **Phase 1** (data analysis based on the "reference" batches) consists in computing K matrices $\mathbf{Z}^{(2)},\ldots,\mathbf{Z}^{(K)},\mathbf{Z}^{(K+1)}$ where

$$\mathbf{Z}^{(k)}=\begin{pmatrix} z_{1,1}^{(k)} & z_{1,2}^{(k)} & \cdots & z_{1,J}^{(k)}\\ z_{2,1}^{(k)} & z_{2,2}^{(k)} & \cdots & z_{2,J}^{(k)}\\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Each matrix $\mathbf{Z}^{(2)},\ldots,\mathbf{Z}^{(K+1)}$ will be used for the *on-line* monitoring of the process up to time k, while the matrix $\mathbf{Z}^{(K+1)}$ will be used for the *on-line* monitoring of the process later than time K. The dimension of each matrix $\mathbf{Z}^2,\ldots,\mathbf{Z}^{(K)},\mathbf{Z}^{(K+1)}$ is (I,J) for method #1 and $(I-1,J)$ for methods #2 and #3.

– **Phase 2** (monitoring of future batches) consists in computing the $(1,J)$ row vector $\mathbf{z}_{i^*}^{(k)} = (z_{i^*,1}^{(k)},z_{i^*,2}^{(k)},\ldots,z_{i^*,J}^{(k)})$, for each new batch $i^* = I+1,\ldots,I+n$, at each instant k.

The way of computing matrices $\mathbf{Z}^{(k)}$ and vector $\mathbf{z}_{i^*}^{(k)}$ will depend on the selected method. These methods are described below.

5.1. *On-line method #1*

– **Phase 1** consists in computing the elements $z_{i,j}^{(k)}$ of each matrix $\mathbf{Z}^{(2)},\ldots,\mathbf{Z}^{(K+1)}$:
  – if $k = 2,\ldots,K$ then

$$z_{i,j}^{(k)} = \min_{\substack{\ell=1,\ldots,I \\ \ell\neq i}} (d_H(\mathbf{X}_{i,j}^{(k)},\mathbf{X}_{\ell,j}))$$

  – if $k = K+1$ then

$$z_{i,j}^{(K+1)} = \min_{\substack{\ell=1,\ldots,I \\ \ell\neq i}} (d_H(\mathbf{X}_{i,j},\mathbf{X}_{\ell,j}))$$

– **Phase 2** consists in computing the elements $z_{i^*,j}^{(k)}$ of the $(1,J)$ row vector $\mathbf{z}_{i^*}^{(k)}$, for each new batch $i^* = I+1,\ldots,I+n$, at each instant $k \geq 2$

$$z_{i^*,j}^{(k)} = \min_{\substack{\ell=1,\ldots,I \\ \ell\neq i^*}} (d_H(\mathbf{X}_{i^*,j}^{(k)},\mathbf{X}_{\ell,j}))$$

5.2. *On-line method #2*

– **Phase 1** consists in computing the elements $z_{i,j}^{(k)}$ of each matrix $\mathbf{Z}^{(2)},\ldots,\mathbf{Z}^{(K+1)}$. In order to compute these elements, we firstly have to find the batch index R which is supposed to be "the most representative" among the I reference batches and satisfies Eq. (1). The elements $z_{i,j}^{(k)}$ can then be computed:
  – if $k = 2,\ldots,K$ then

$$z_{i,j} = d_H(\mathbf{X}_{i,j}^{(k)},\mathbf{X}_{R,j})$$

  – if $k = K+1$ then

$$z_{i,j}^{(K+1)} = d_H(\mathbf{X}_{i,j},\mathbf{X}_{R,j})$$

– **Phase 2** consists in computing the elements $z_{i^*,j}^{(k)}$ of the $(1,J)$ row vector $\mathbf{z}_{i^*}^{(k)}$, for each new batch $i^* = I+1,\ldots,I+n$, at each instant $k \geq 2$

$$z_{i^*,j}^{(k)} = d_H(\mathbf{X}_{i^*,j}^{(k)},\mathbf{X}_{R,j})$$

5.3. *On-line method #3*

- **Phase 1** consists in computing the elements $z_{i,j}^{(k)}$ of each matrix $\mathbf{Z}^{(2)},\ldots,\mathbf{Z}^{(K+1)}$. In order to compute these elements, we firstly have to find the batch index R_j, $j=1,\ldots,J$, which is supposed to be "the most representative" for the variable j among the I reference batches and satisfies Eq. (2). The elements $z_{i,j}^{(k)}$ can then be computed:
  - if $k=2,\ldots,K$ then
$$z_{i,j}^{(k)} = d_H(\mathbf{X}_{i,j}^{(k)},\mathbf{X}_{R_j,j})$$
  - if $k=K+1$ then
$$z_{i,j}^{(K+1)} = d_H(\mathbf{X}_{i,j},\mathbf{X}_{R_j,j})$$
- **Phase 2** consists in computing the elements $z_{i^*,j}^{(k)}$ of the $(1,J)$ row vector $\mathbf{z}_{i^*}^{(k)}$, for each new batch $i^*=I+1,\ldots,I+n$, at each instant $k\geq 2$

$$z_{i^*,j}^{(k)} = d_H(\mathbf{X}_{i^*,j}^{(k)},\mathbf{X}_{R_j,j})$$

6. A modified Hotteling T^2 control chart

For all the methods proposed in the two previous sections, a natural approach for monitoring vectors $\mathbf{z}_{i^*}$ (*off-line* case) and $\mathbf{z}_{i^*}^{(k)}$ (*on-line* case) could be to use a Hotteling T^2 control chart based on the matrix $\mathbf{Z}$ (*off-line* case) or matrices $\mathbf{Z}^{(2)},\ldots,\mathbf{Z}^{(K+1)}$ (*on-line* case). Unfortunately, the components $z_{i^*,j}$ and $z_{i^*,j}^{(k)}$ of vectors $\mathbf{z}_{i^*}$ and $\mathbf{z}_{i^*}^{(k)}$ are, by definition, positive distances. Consequently $\mathbf{z}_{i^*}$ and $\mathbf{z}_{i^*}^{(k)}$ are not multinormal random vectors and the use of the traditional Hotteling T^2 control chart, in the case of our methods, must be considered with care. An alternative is to use a normalizing transformation, on each variable, making the transformed random vector distribution close to a $(\mathbf{0},\mathbf{I})$ multinormal distribution and then apply the Hotteling T^2 control chart on the transformed random vector. Possible candidates for such a transformation are the three parameters lognormal distribution or the Johnson distributions. For simplicity, in this paper, we chosen to use the lognormal distribution/transformation. The strategy we suggest is described below for both *off-line* and *on-line* cases.

- For the *off-line* case, the matrix $\mathbf{Z}$ has to be transformed into a matrix $\mathbf{U}$ using J different lognormal transformations on each column (variable) $j=1,\ldots,J$ of matrix $\mathbf{Z}$. Then the row vector $\mathbf{z}_{i^*}$ corresponding to a new batch, has to be transformed into a row vector $\mathbf{u}_{i^*}$ using the same J different lognormal transformations on each component $z_{i^*,1},\ldots,z_{i^*,J}$ of vectors $\mathbf{z}_{i^*}$. Finally, the Hotteling statistics that must be computed is
$$T^2(\mathbf{u}_{i^*},\mathbf{U})$$
- For the *on-line* case, each matrix $\mathbf{Z}^{(2)},\ldots,\mathbf{Z}^{(K+1)}$ have to be transformed into a matrix $\mathbf{U}^{(2)},\ldots,\mathbf{U}^{(K+1)}$ using J different lognormal transformations on each column (variable) $j=1,\ldots,J$ of matrix $\mathbf{Z}^{(k)}$. Then each row vector $\mathbf{z}_{i^*}^{(k)}$ corresponding to a new batch, has to be transformed into a row vector $\mathbf{u}_{i^*}^{(k)}$ using the J different lognormal transformations (applied for matrix $\mathbf{Z}^{(k)}$)

on each component $z_{i^*,1}^{(k)},\ldots,z_{i^*,J}^{(k)}$ of vectors $\mathbf{z}_{i^*}^{(k)}$. In this case, the Hotteling statistics that must be computed are

$$T^2(\mathbf{u}_{i^*}^{(k)},\mathbf{U}^{(k)}) \qquad 2 \leq k \leq K$$
$$T^2(\mathbf{u}_{i^*}^{(k)},\mathbf{U}^{(K+1)}) \qquad k \geq K+1$$

7. Illustrative examples

The goal of this section is to demonstrate the use of our methods for the detection of out-of-control batches with varying duration. Both a simulated example and an industrial example are presented.

7.1. *A simulated example*

Our simulated example consists in $I = 50$ reference batches with $J = 3$ process variables that are supposed to have a "nominal" behavior and generating conforming products. The trajectories of these $I = 50$ reference batches are plotted in Fig. 1 (left side). These reference batches have been simulated using the following approach:

- For each process variable, we defined a nominal profile (function of time) starting at time $k = 0$ and ending at time $k = 100$.
- For each batch i, a Poisson ($\lambda = 5$) random variable Y_i is generated and the profile of each variable is scaled such that batch i terminates at time $K_i = 100-Y_i$. Thus, the average time of duration for any batch is $100-5 = 95$.
- For each variable j, a normal $N(0,\sigma_j)$ random variable $Z_j^{(k)}$ is added to the scaled profile at time k. The values used are $\sigma_1 = 0.6$, $\sigma_2 = 0.4$ and $\sigma_3 = 0.2$.

In our example, the total time durations K_i of these $I = 50$ nominal batches vary from 89 to 100. Additionally, we have four extra batches. Two of them are supposed to be nominal with total time durations 97 and 98. The J variables of both nominal extra batches have the same variability of the references batches (i.e. $\sigma_1 = 0.6$, $\sigma_2 = 0.4$ and $\sigma_3 = 0.2$). Another one is supposed to be non-nominal with total a time duration 83 (i.e. a "too short" batch). Finally, the last one is supposed to be non-nominal with a total time duration 114 (i.e. a "too long" batch). The J variables of both non-nominal extra batches have $\sigma_1 = 0.8$, $\sigma_2 = 0.6$ and $\sigma_3 = 0.4$. In Fig. 1 (right side) we plotted the trajectories of these 4 extra batches.

- Use of *off-line* methods: in Fig. 2, we plotted (using a logarithmic scale) the Hotelling T_i^2 statistics corresponding to the transformed statistics $\mathbf{u}_i$ for the 3 *off-line* methods and for the 50 reference batches + 4 extra batches. As we can see, all the batches have $T_i^2 < UCL$ except for the two last ones that have a very large value for T_i^2 confirming the non nominal behavior of these batches.
- Use of *on-line* methods: in Fig. 3 and Fig. 4 we plotted the Hotelling control charts corresponding to the transformed statistics $\mathbf{u}_{i^*}^{(k)}$ for the 3 *on-line* methods and for time $k \geq 2$. In Fig. 3 we plotted the Hotelling control charts for the two extra batches that have a nominal behavior. In Fig. 4 (left side), we plotted the Hotelling control charts for the extra batch which has a non-nominal behavior and is "too short" and, in Fig. 4 (right side), we plotted the Hotelling control chart for the extra batch which has a non-nominal behavior and is "too long". As we can see in these figures, $T_k^2 < UCL$ for the nominal extra batch while $T_k^2 > UCL$ for both non-nominal extra batches.

7.2. *An industrial example*

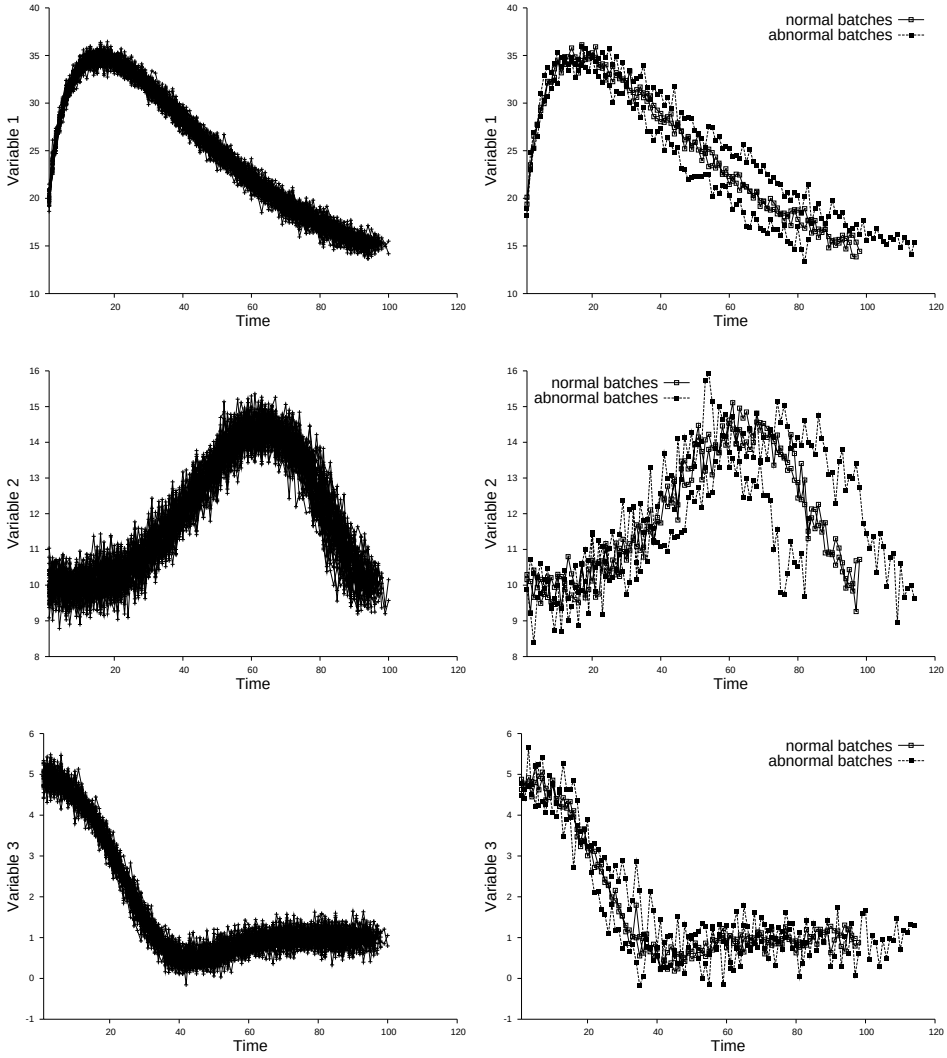

Fig. 1. Trajectories of the $J = 3$ variables for the $I = 50$ reference batches (left side) and for the 4 extra batches (right side)

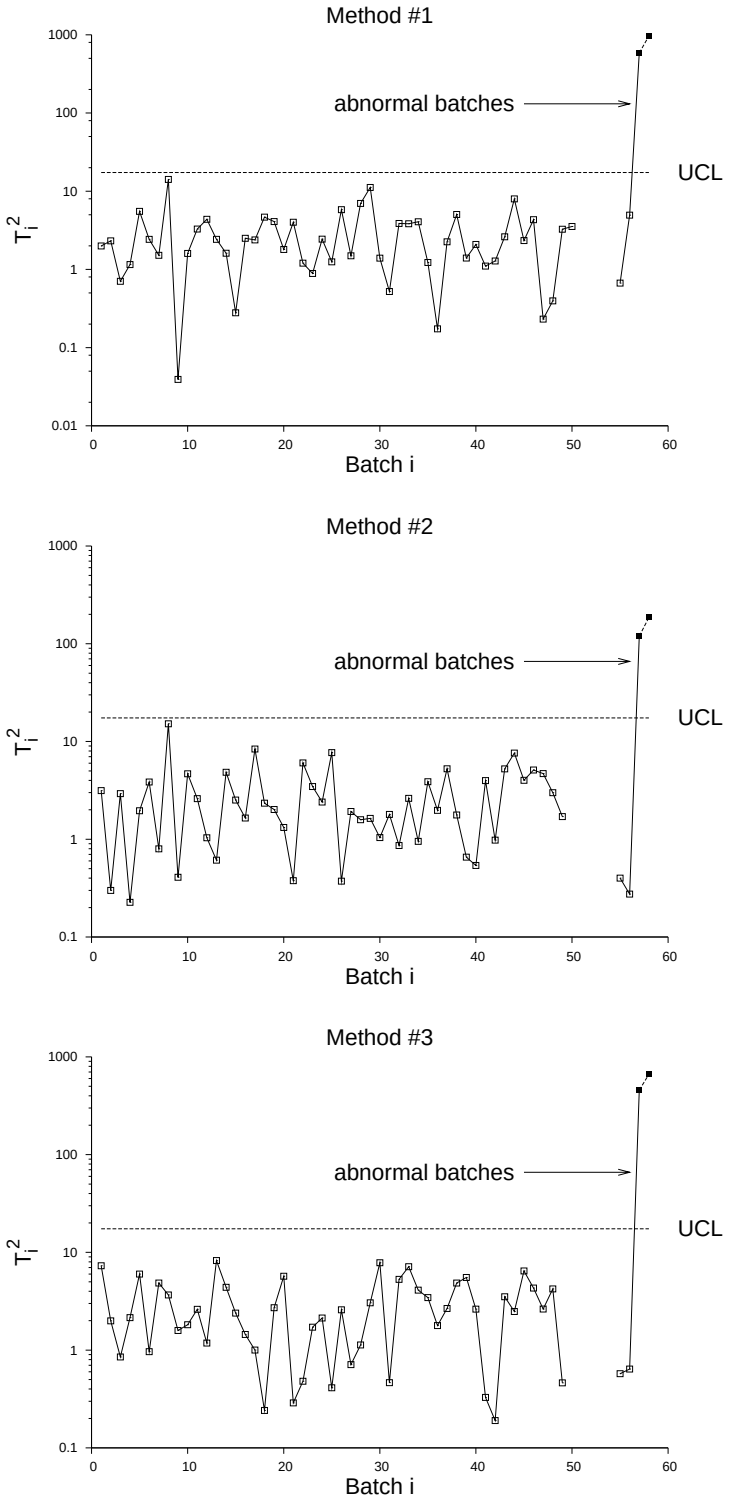

Fig. 2. Hotelling T_i^2 statistics corresponding to the transformed statistics $\mathbf{u}_i$ for the 3 *off-line* methods and for the 50 reference batches + 4 extra batches

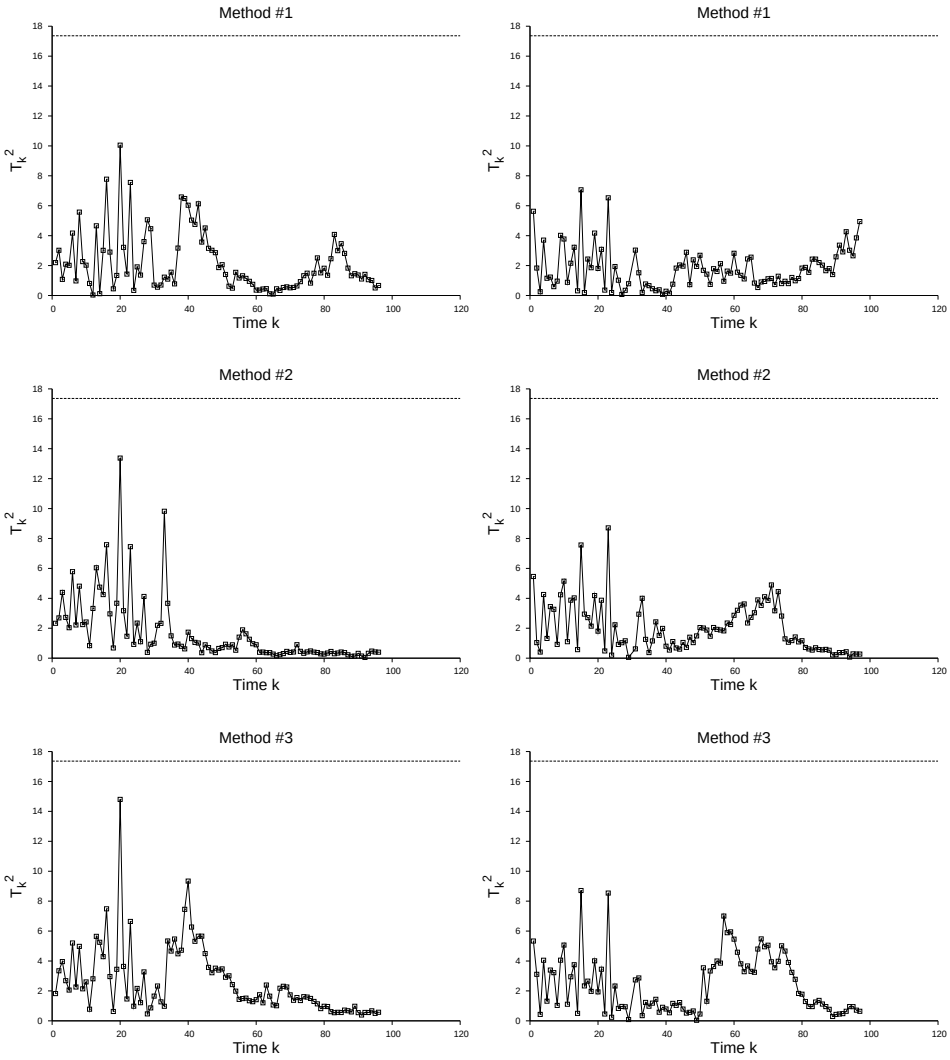

Fig. 3. Hotelling control charts corresponding to the three *on-line* methods for the two extra batches that have a nominal behavior

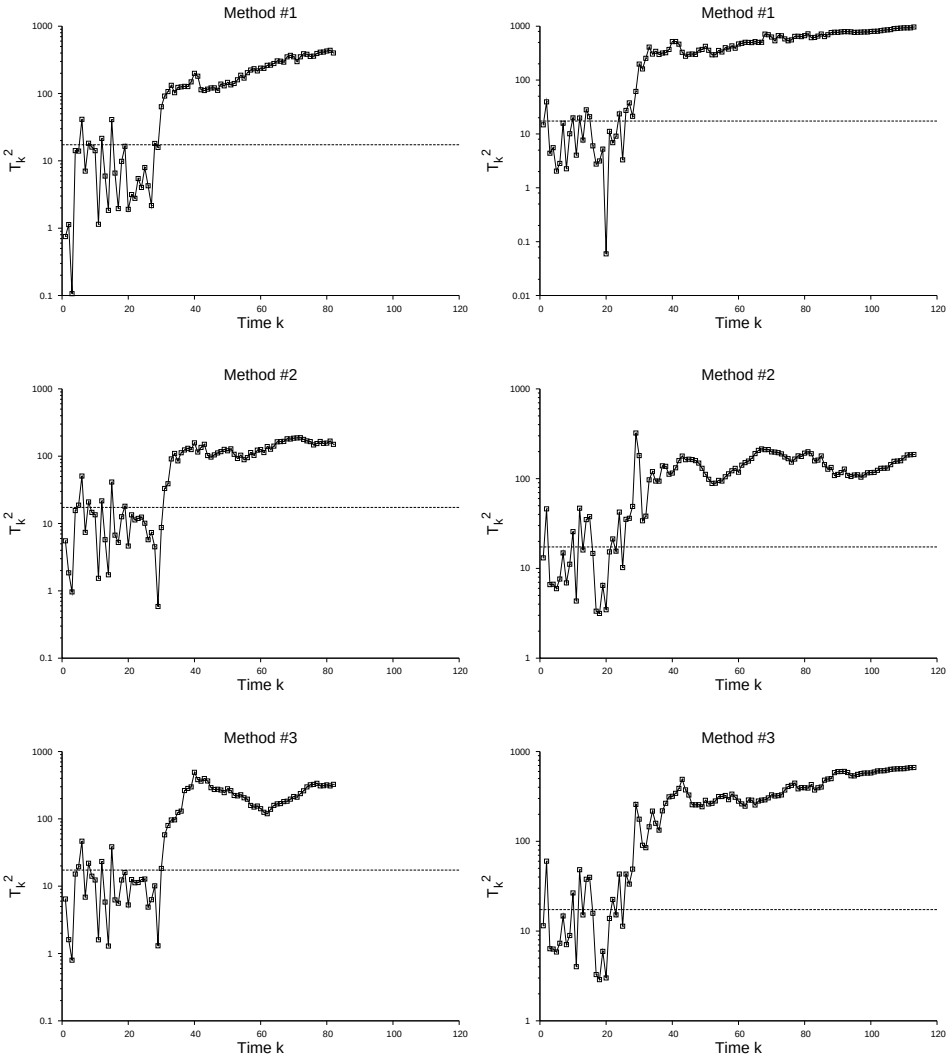

Fig. 4. Hotelling control charts corresponding to the three *on-line* methods for the two extra batches which have a non-nominal behavior

In this section, an industrial batch process of rubber for tires is presented. The number of monitored variables in this process is $J = 4$ and the historical data are composed of $I = 88$ "reference" batches that have a "nominal" behavior and generate conforming products. The total time durations K_i of these $I = 88$ nominal batches vary from 40 to 50. In Fig. 5 (left side) we plotted the trajectories of these $J = 4$ variables for these $I = 88$ batches. Additionally, we have four extra batches. Two of them are supposed to be nominal with total time durations 43 and 50. Another one is supposed to be non-nominal with a total time duration 38 (i.e. a "too short" batch). Finally, the last one is supposed to be non-nominal with a total time duration 64 (i.e. a "too long" batch). In Fig. 5 (right side) we plotted the trajectories of these 4 extra batches.

- Use of *off-line* methods: in Fig. 6(a), Fig. 6(b) and Fig. 6(c) we plotted the Hotelling T_i^2 statistics corresponding to the transformed statistics $\mathbf{u}_i$ (after the lognormal transformation) for the 3 *off-line* methods. As we can see, all the batches have $T_i^2 < UCL$ except for the two last ones that have a very large value for T_i^2 confirming the non nominal behavior of these batches.
- Use of *on-line* methods: in Fig. 7 and Fig. 8 we plotted (using a logarithmic scale) the Hotelling control charts corresponding to the transformed statistics $\mathbf{u}_i^{(k)}$ (after the lognormal transformation) for the 3 *on-line* methods and for time $k \geq 2$. In Fig. 7 we have the Hotelling control charts for the two extra batches that have a nominal behavior. In Fig. 8 (right side) we have the Hotelling control charts for the extra batch which has a non-nominal behavior and is "too short" and in Fig. 8 (left side), we have the Hotelling control charts for the extra batch which has a non-nominal behavior and is "too long". As we can see in these figures, $T_k^2 < UCL$ for the nominal extra batch and $T_k^2 > UCL$ for both non-nominal extra batches.

8. Comparison of methods #1, #2 and #3

In order to compare the three proposed methods, we suggest to reuse the simulated model presented in subsection "7.1. A simulated example" and to investigate *separately* the impact of the total duration time and the variability.

- **Impact of the total duration time**. In order to investigate the impact of the total duration time on the efficiency of the three methods, we suggest to vary the value of parameter $\lambda \in \{20, \ldots, -1, +1, \ldots, +20\}$ and for each value of λ to simulate $n = 3000$ batches for which each method will be tested. Negative values for λ means that the total duration time of batch i will be $K_i = 100 - Y_i$ (shorter batch) where Y_i is a Poisson $-\lambda$ random variable, while positive values for λ means that the total duration time of batch i will be $K_i = 100 + Y_i$ (longer batch) where Y_i is a Poisson λ random variable. The natural variability of the process variables, parametrized by $\sigma_1 = 0.6$, $\sigma_2 = 0.4$ and $\sigma_3 = 0.2$, is left unchanged.
- **Impact of the variability**. In order to investigate the impact of the variability on the efficiency of the three methods, we suggest to vary the value of a parameter $\tau \in [1,3]$ such that the new standard-deviation associated with the process variables are $\sigma_1' = \tau \times 0.6$, $\sigma_2' = \tau \times 0.4$ and $\sigma_3' = \tau \times 0.2$, and for each value of τ to simulate $n = 3000$ batches for which each method will be tested. If $\tau = 1$ then the variability of the variables is the reference variability. If $\tau = 3$ then the variability of the variables is three times the reference variability. The natural total duration time is left unchanged.

8.1. *Comparison of off-line methods*

- **Impact of the total duration time**. In Fig. 9 we plotted the percentage of

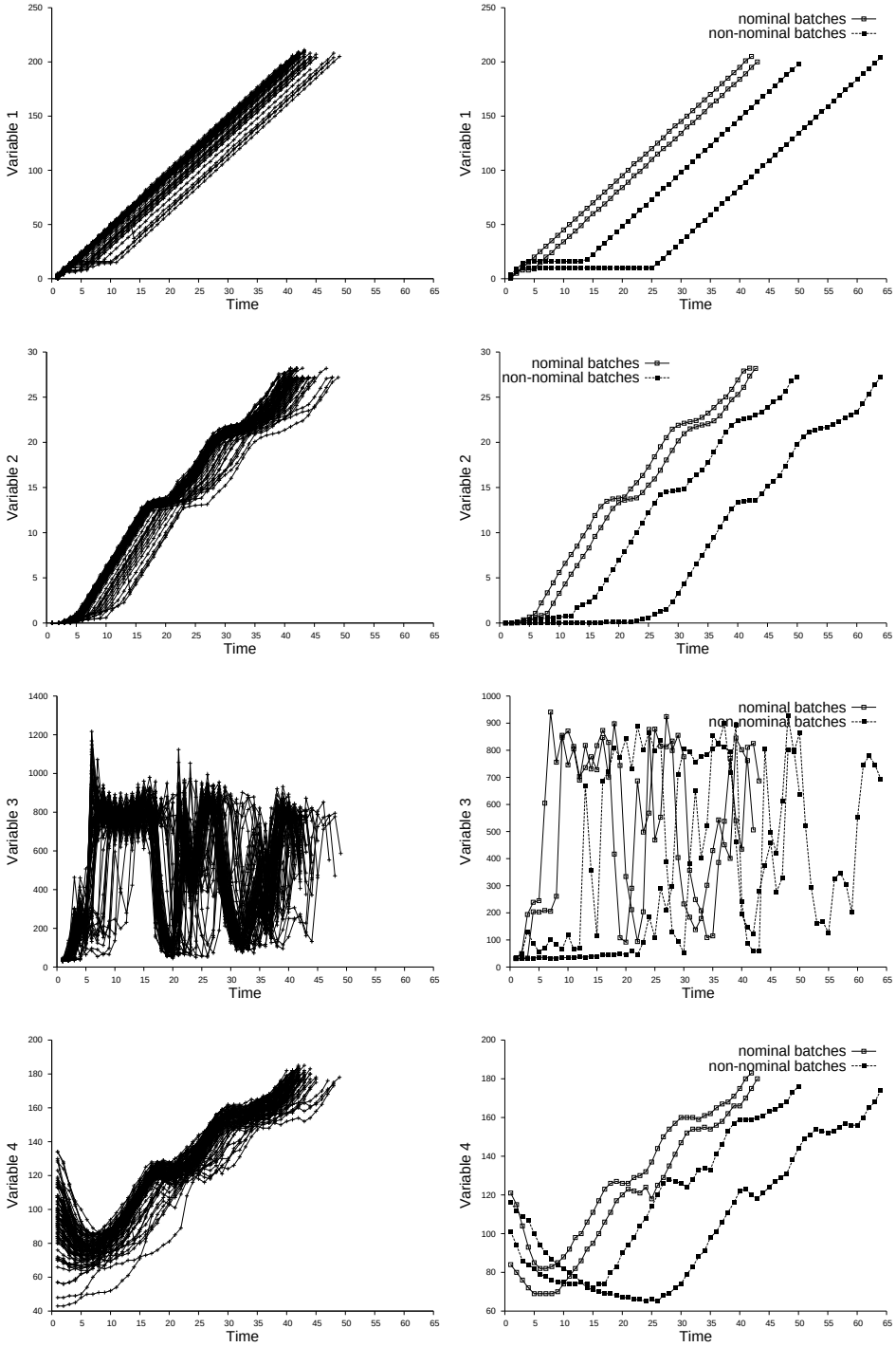

Fig. 5. Trajectories of the $J = 4$ variables for the $I = 88$ reference batches (left side) and for the 4 extra batches (right side)

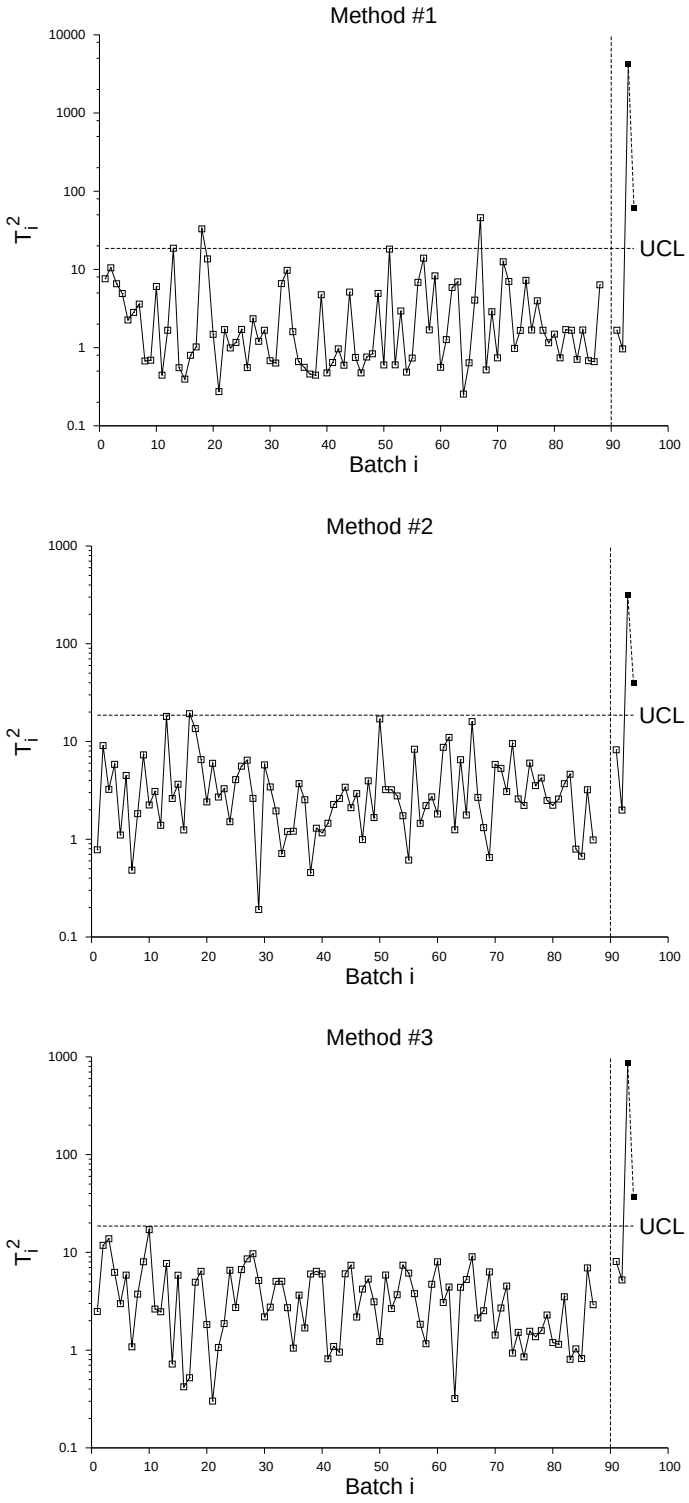

Fig. 6. Hotelling T_i^2 statistics corresponding to the transformed statistics $\mathbf{u}_i$ for the 3 *off-line* methods and for the 88 reference batches + 4 extra batches

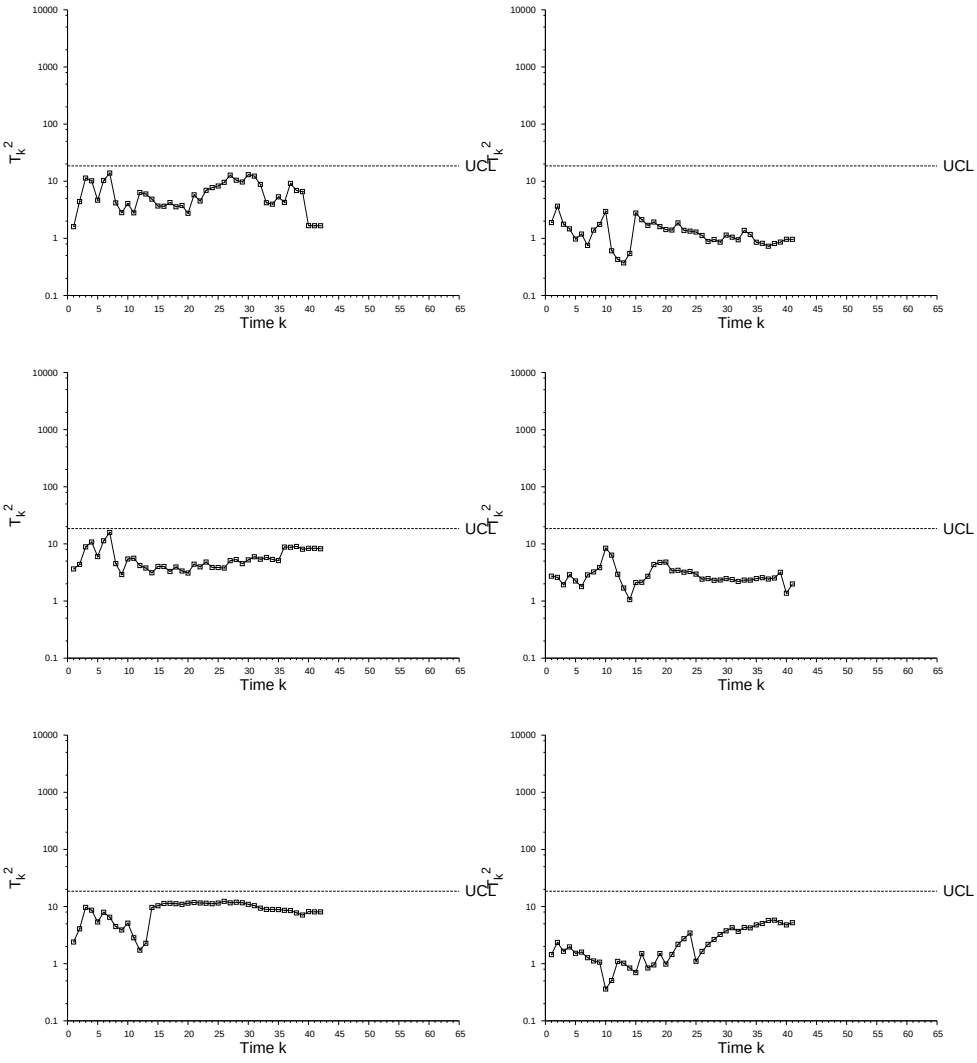

Fig. 7. Hotelling control charts corresponding to the three *on-line* methods for the two extra batches that have a nominal behavior

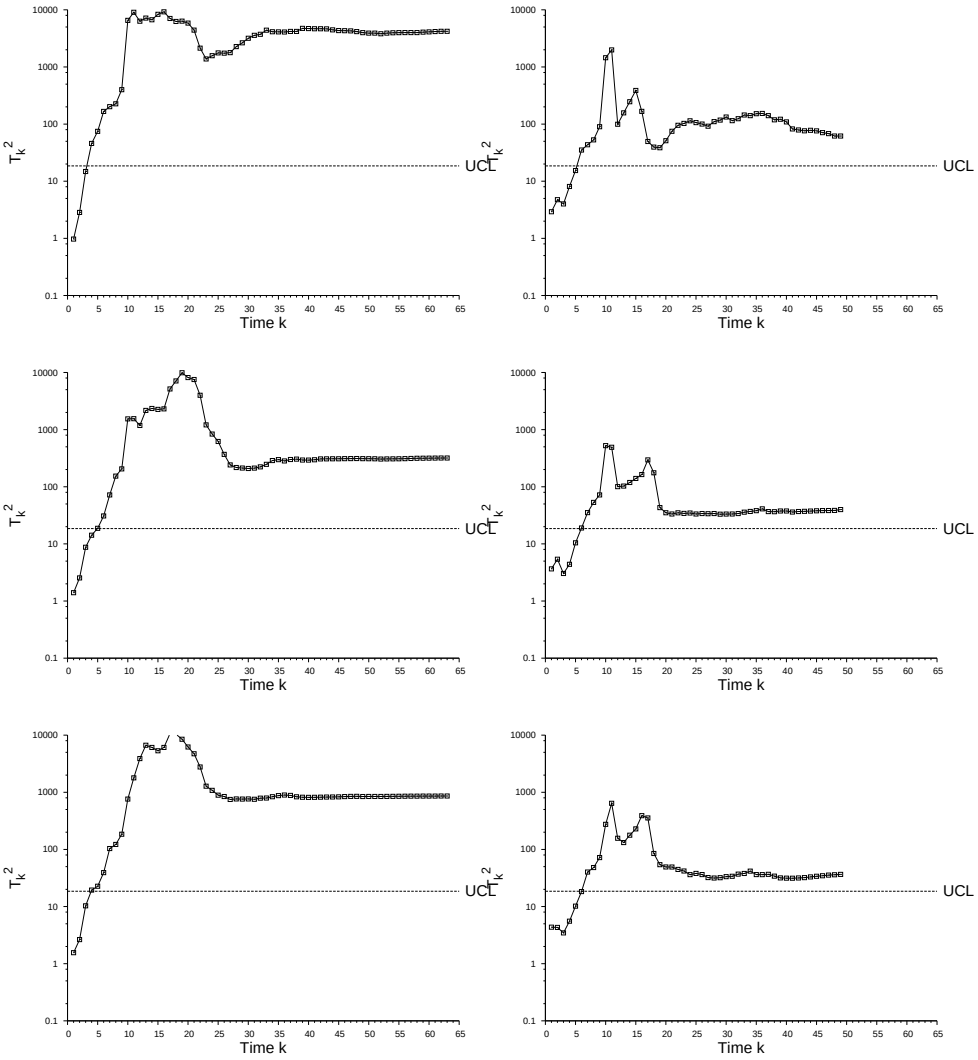

Fig. 8. Hotelling control charts corresponding to the three *on-line* methods for the two extra batches which have a non-nominal behavior

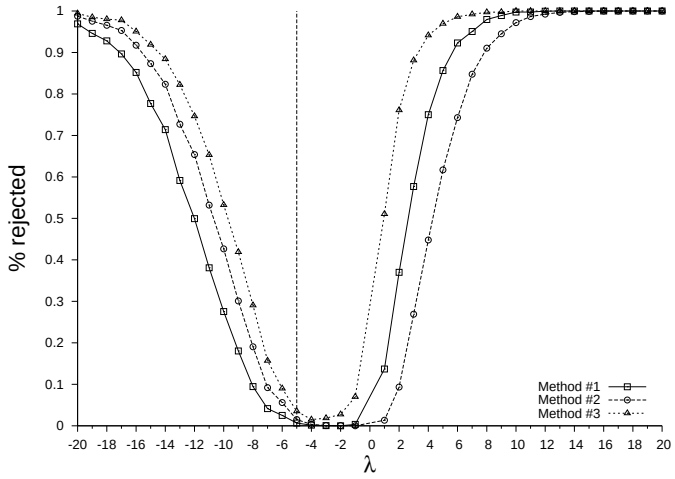

Fig. 9. Percentage of batch rejected using *off-line* methods #1, #2 and #3 versus parameter $\lambda \in \{20, \ldots, -1, +1, \ldots, +20\}$

batch rejected using off-line methods #1, #2 and #3 versus parameter λ. As we can see, whatever the value of λ, the method #3 is the most efficient, i.e. it rejects too short and too long batches more often than the other methods. For the too short batches ($\lambda < 0$), the method #1 is the less efficient, while for the too long batches ($\lambda > 0$) the method #2 becomes the less efficient. Concerning the nominal batches (the dotted line) the three methods behave the same.

– **Impact of the variability**. In Fig. 10 we plotted the percentage of batch rejected using off-line methods #1, #2 and #3 versus parameter τ. For a small increase of the variability $\tau \in$ [1,1.2] the method #3 is the most efficient, but for a larger increase of the variability $\tau \in$ [1.2,3] the method #1 becomes more efficient than the others.

8.2. *Comparison of on-line methods*

– **Impact of the total duration time**. In Fig. 11 we plotted the percentage of batch rejected using on-line methods #1, #2 and #3 versus parameter λ. For too short and too long batches, the method #3 is the most efficient, but for batches close to nominal (say, $\lambda \in \{-8, \ldots, -2\}$) the method #3 rejects a high percentage (more than 33%) of batches and, in that case, the use of method #1 seems more reasonable (even if this one still rejects 10% of the batches). In Fig. 12 we plotted the time for which the batches have been detected as non conforming. We can notice that, whatever the value of λ, the method that detects the fastest is method #2.

– **Impact of the variability**. In Fig. 13 we plotted the percentage of batch rejected using on-line methods #1, #2 and #3 versus parameter τ. The method which seems the most interesting is the method #1 since for a small increase of variability (close to nominal), the percentage of rejected batches is the lowest, while for a large increase of variability, the percentage of rejected batches is the highest. In Fig. 14 we plotted the time for which the batches have been detected as non conforming. We can notice that, whatever the value of τ, the method that detects the fastest is again method #2.

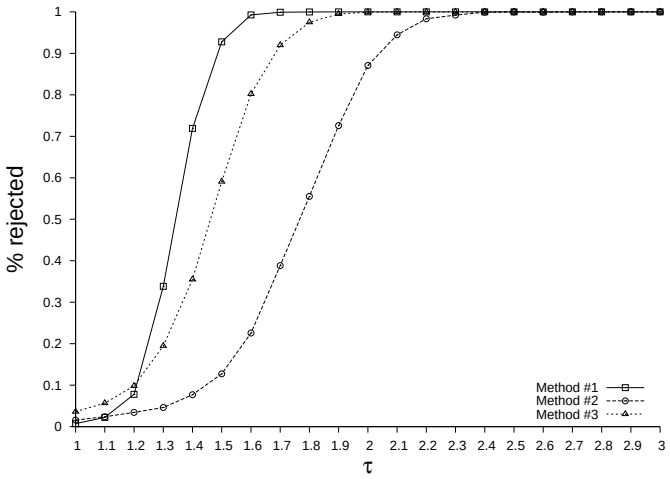

Fig. 10. Percentage of batch rejected using *off-line* methods #1, #2 and #3 versus parameter $\tau \in [1,3]$

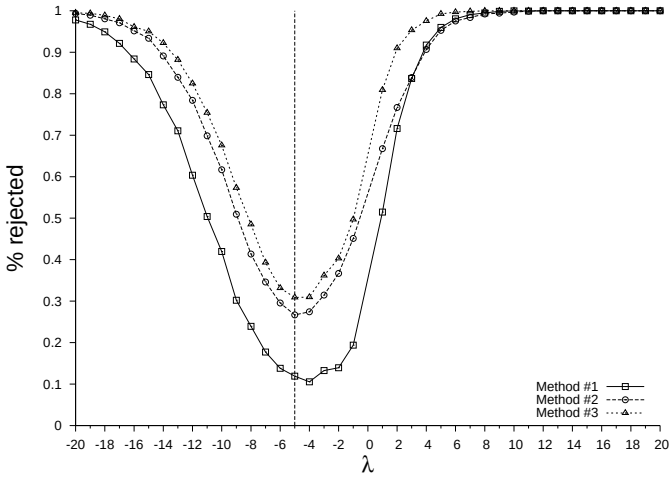

Fig. 11. Percentage of batch rejected using *on-line* methods #1, #2 and #3 versus parameter $\lambda \in \{20, \ldots, -1, +1, \ldots, +20\}$

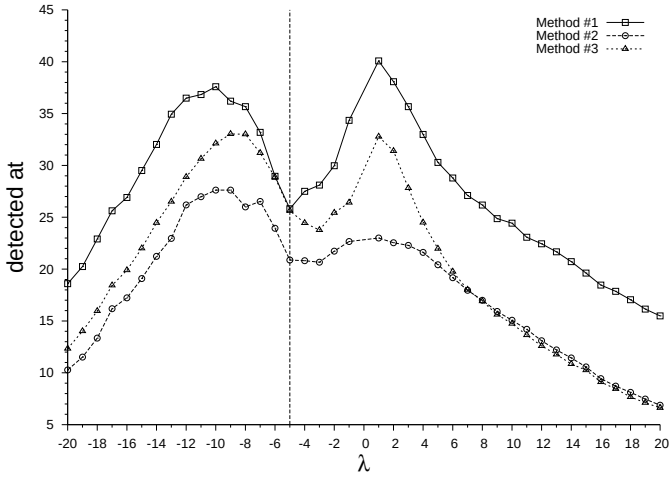

Fig. 12. Time for which the batches have been detected as non conforming using *on-line* methods #1, #2 and #3 versus parameter $\lambda \in \{20, \ldots, -1, +1, \ldots, +20\}$

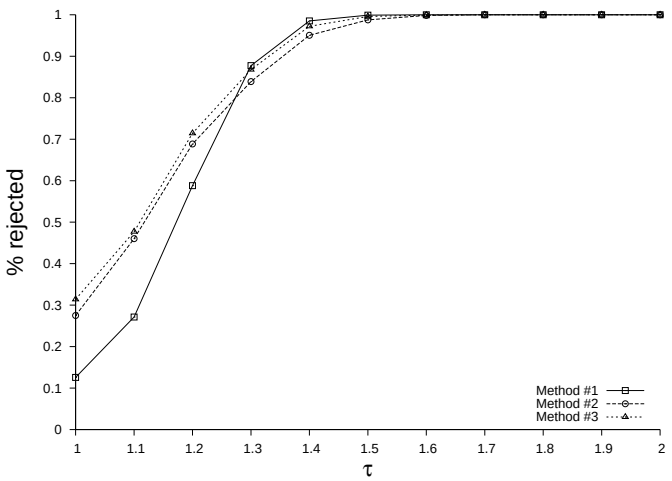

Fig. 13. Percentage of batch rejected using *on-line* methods #1, #2 and #3 versus parameter $\tau \in$ [1,3]

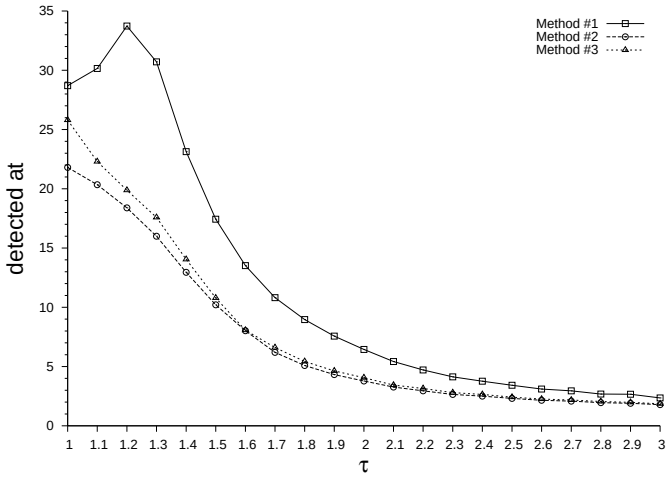

Fig. 14. Time for which the batches have been detected as non conforming using *on-line* methods #1, #2 and #3 versus parameter $\tau \in [1,3]$

9. Conclusions

This paper proposes different innovative methods for the *off-line* and *on-line* monitoring of batch processes with varying duration. The approach behind these methods is purely a geometrical one, i.e. the Hausdorff Distance. Concerning this distance, we choose the Euclidean distance as basic distance, but it is obvious that we are able to choose another basic distance as for example the Mahalanobis distance. This allows a kind of flexibility in the choice of the basic distance. However, depending on the basic distance chosen the time of execution of the algorithms could be increased.

Another important point is that our matrix $\mathbf{X}_i^{(k)}$ is equivalent to the variable-wise unfolding used in the methods developed for batch process monitoring with equal durations. In this way of unfolding, each point of the trajectory of every batch is considered as an object. This kind of unfolding is adjusted for application of our methods and with our geometric approach.

We would like to emphasize some key aspects in relation to the section "8. Comparison of methods #1, #2 and #3". For each method we investigated separately the impact of the total duration time and variability. Firstly, we want make some conclusions about the impact of the total duration time. In *off-line* batch control the method #3 is the most efficient because it rejects too short and too long batches more often than other methods. However, concerning the nominal batches the three methods behave the same. In *on-line* batch control for too short and too long batches, the method #3 is the most efficient, but for batches close to nominal the use of method #1 seems more reasonable (the method #1 rejects 10% of the batches against more than 33% rejected batches for the method #3). Secondly, we want make some conclusions about the impact of the variability. In *off-line* batch control the method #3 is the most efficient for a small increase of the variability. However, for a larger increase of the variability the method #1 becomes more efficient than the others. In *on-line* batch control the most interesting is the method #1 since for a small increase of variability (close to nominal), the percentage of rejected batches is the lowest, while for a large increase of variability, the percentage of rejected batches is the highest. Finally, for *on-line* batch control we can notice that the method #2 is the fastest method for detection of non-nominal batches, whatever increase of variability, but also whatever total duration time. In conclusion, for both

off-line and *on-line* batch control it seems the most interesting to make a control strategy with to use a combination of the different methods presented in according to characteristics of the process and objectives of the statistical process control. If the process presents variable profiles or total time duration with high variability it can be interesting to use in combination the methods #1 and #3. Nevertheless, if we want the fastest method for *on-line* detection of non-nominal batches it seems the most interesting to use the method #2.
It is also important to notice that for summarizing a large number of process variable trajectories collected throughout of the batch, the use of multivariate statistical projection methods is recommended. We did not apply any multivariate statistical projection methods in this paper because our applications consisted respectively of $J = 3$ and $J = 4$ variables. However, if the application of the multivariate statistical projection methods is necessary after application of our methods, we suggest the application of INDSCAL (Individual differences scaling) introduced by [14]. Further developments of the methods proposed here may mainly include non-parametric control charts based in our methods and search for diagnosis implementations.

Acknowledgments

Brazilian CAPES Grant BEX1101/01-3

References

1. J.P. Vanbergen Vue générale des problémes de l'automatisation des batches *In Ibra,editeur,Automatisation des processus mixtes: les systèmes dynamiques hybrides, ADPM'94, Bruxelles, Belgique* 265–270, 1994.
2. S.M. Davis "From Future Perfect":Mass Customizing *Planning Review*, 16–21, 1989.
3. G. Da Silveira, D. Borenstein and F.S. Fogliatto Mass Customization: Literature review and research directions *International Journal Prosuction Economics*, 72:1–13, 2001.
4. P. Nomikos and J.F. MacGregor. Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics*, 37:41–59, 1995.
5. T. Kourti, J. Lee, and J.F. MacGregor. Experiences with Industrial Applications of Projections Methods for Multivariate Statistical Process Control. *Computers Chemical Engineering*, 20:S745–S750, 1996.
6. E.B. Martin, A.J. Morris, and C. Kiparissides. Batch Process Monitoring for Consistent Production. *Computers Chemical Engineering*, 20:S599–S604, 1996.
7. S. Wold, N. Kettaneh, H. Friden, and A. Holmberg. Modelling and Diagnostics of Batch Processes and Analogous Kinetic Experiments. *Chemometrics and Intelligent Laboratory Systems*, 44:331–340, 1998.
8. R. Bro. Parafac. Tutorial and Applications. *Chemometrics and Intelligent Laboratory Systems*, 38:149–171, 1997.
9. A.K. Smilde. Three-Way Analysis. Problems and Prospects. *Chemometrics and Intelligent Laboratory Systems*, 15:143–157, 1992.
10. T. Kourti. Abnormal Situation Detection, Three-Way Data and Projection Methods; Robust Data Archiving and Modeling for Industrial Applications. *Annual Reviews in Control*, 27:131–139, 2003.
11. T. Kourti. Multivariate Dynamic Modeling for Analysis and Statistical Process Control of Batch Processes, Star-ups and Grade Transitions. *Journal of Chemometrics*, 17:93–109, 2003.
12. N. Kaistha, M.S. Johnson, C.F. Moore and M.G. Leitnaker. Online Batch Recipe Adjustments for Product Quality Control Using Empirical Models: Application to a

Nylon-6,6 Process. *ISA Transactions*, 42:305–315, 2003.

13. N. Kaistha, C.F. Moore and M.G. Leitnaker. A Statistical Process Control Framework for the Characterization of Variation in Batch Profiles. *Technometrics*, 46(1):53–68, 2004.
14. J.D. Carroll and J.J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of " Eckart-Young" decomposition. *Psychometrika*, 35:283–319, 1970.