Computers and Chemical Engineering 34 (2010) 223–231

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

A robust clustering method for detection of abnormal situations in a process with multiple steady-state operation modes

Mauricio Maestri[a], Andrés Farall[b], Pablo Groisman[b], Miryan Cassanello[a], Gabriel Horowitz[a,c,*]

[a] PINMATE, Dep. de Industrias, FCEyN, Universidad de Buenos Aires, C1428BGA Buenos Aires, Argentina
[b] Instituto de Cálculo, FCEyN, Universidad de Buenos Aires, Argentina
[c] Centro de Tecnología Argentina, YPF, Argentina

ARTICLE INFO

Article history:
Received 8 September 2008
Received in revised form 1 April 2009
Accepted 15 May 2009
Available online 27 May 2009

Keywords:
Fault detection
Multiple operating modes
Multivariate statistical process monitoring

ABSTRACT

Many classical multivariate statistical process monitoring (MSPM) techniques assume normal distribution of the data and independence of the samples. Very often, these assumptions do not hold for real industrial chemical processes, where multiple plant operating modes lead to multiple nominal operation regions. MSPM techniques that do not take account of this fact show increased false alarm and missing alarm rates. In this work, a simple fault detection tool based on a robust clustering technique is implemented to detect abnormal situations in an industrial installation with multiple operation modes. The tool is applied to three case studies: (i) a two-dimensional toy example, (ii) a realistic simulation usually used as a benchmark example, known as the Tennessee–Eastman Process, and (iii) real data from a methanol plant. The clustering technique on which the tool relies assumes that the observations come from multiple populations with a common covariance matrix (i.e., the same underlying physical relations). The clustering technique is also capable of coping with a certain percentage of outliers, thus avoiding the need of extensive preprocessing of the data. Moreover, improvements in detection capacity are found when comparing the results to those obtained with standard methodologies. Hence, the feasibility of implementing fault detection tools based on this technique in the field of chemical industrial processes is discussed.

1. Introduction

Arising from the ever growing possibility of collecting immense amounts of data with modern monitoring and control systems, there has been increasing interest in pursuing methods that are capable of grasping the essentials in the data. Multivariate statistical process control (MSPC) tools are data driven techniques that generally reduce the dimension of process data and extract key features and trends that are of interest to plant personnel (Venkatasubramanian, Rengaswamy, Kavuri, & Yin, 2003). MSPC tools used to reduce the explaining dimensions of the process data, like Principal Component Analysis (PCA) and subsequent refinements, have shown great success. PCA is a method particularly suited to data sets comprising correlated and collinear variables. The methodology projects the process data onto a low dimensional subspace in order to capture the major sources of variability associated with the process. The principal eigenvectors (associated with the principal component loadings) of the sample variance–covariance matrix of the data set conform a base of the subspace; i.e., a set of orthogonal latent variables formed by linear combinations of the original process variables. New data of the process are projected onto the subspace to detect abnormal situations by computing statistics that quantify if the new data are within the limits specified as a normal control region. Relevant information leading to identification or diagnosis of the problem can be found by interrogating the contribution of each process variable to the principal component score.

In spite of the success of applying PCA based MSPC tools to process data for detecting abnormal situations, when these tools are applied to a process with multiple operating modes, many missing and false alarms can appear even when the process itself is operating under other steady-state nominal operating conditions (Zhao, Zhang, & Xu, 2004). This is not fortuitous; it is because many of the current techniques are based on the assumption that the process has one nominal operating region while real processes have many. Process data generally define different groups based, for instance, on variations in the operating capacity, seasonal variations or changes in the feedstock characteristics, and also on modifications in the operation strategies introduced purposely by the plant personnel through changes in the set points (Ge & Song, 2008).

* Corresponding author at: PINMATE, Dep. de Industrias, FCEyN, Universidad de Buenos Aires, C1428BGA Buenos Aires, Argentina. Tel.: +54 11 45763383; fax: +54 11 45763366.
E-mail address: ghorowitzer@ypf.com (G. Horowitz).

0098-1354/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2009.05.012

From a geometric point of view, whenever such a change occurs, the process data tend to group into a new cluster in a different location in the high dimensional space containing the process normal (meaning not faulty) operation region. If all the data is considered as belonging to a unique normal operation region, the volume of this region becomes incorrectly large. A monitoring tool considering such a region will lead to an increased number of missing and false alarms (Zhao et al., 2004; Zhao, Zhang, & Xu, 2006).

Some approaches have been proposed to address the issues associated with multiple operating modes under different assumptions. Lane, Martin, Kooijmans, and Morris (2001) adopted a common subspace model to monitor a semibatch process to produce several different products. The method is based upon the assumption that a common eigenvector subspace exists for the variance–covariance matrices of the individual product grades or recipes, and through a pooled sample variance–covariance matrix the principal component loadings of the multi-group model can be calculated. Hwang et al. (1999) proposed a monitoring method using a super-PCA model which considers that the number of retained eigenvectors is the same in each of the clusters defined by hierarchical clustering of the data. Chen and Liu (1999) proposed a method called Mixture Principal Component Analysis (MixPCA). In their approach, PCA is used to compress and extract process features and a heuristic smoothing clustering (HSC) algorithm based on the Gaussian filter automatically determine the proper number of clusters. Choi, Yoo, and Lee (2003) proposed a method based on Partial Least Squares (PLS) and credibilistic fuzzy-c-means (CFCM) for modeling and monitoring processes that undergo operating condition changes. Yoo, Vanrolleghem, and Lee (2003) used PCA to reduce the dimensionality and to remove collinearity of the data. Then, they applied adaptive credibilistic fuzzy-c-means to model diverse kinds of operating conditions, and also proposed an adaptive discrimination monitoring (ADM) method to distinguish between a large process change and a simple fault. In the approach proposed by Srinivassan, Wang, Ho, and Lim (2004), process data are first segmented based on regions of steady-state operations into modes and transitions. Then, a dynamic PCA (DPCA) based similarity factor clusters the transitions.

Different clusters certainly have different means; however, since the physical rules governing the process are the same, the covariance structures share common characteristics (Hwang et al., 1999). To enhance the monitoring performance considering the unchanged physical grounds, a statistical model of multiple normal distributions sharing a common covariance matrix is considered. Therefore, the robust clustering method proposed by Gallegos and Ritter (2005) can be used. This method considers that all the clusters share a common covariance matrix, computed as a pooled covariance matrix. Moreover, while defining the clusters, the method is able to cope with potential contamination of the data. This constitutes an important advantage considering that contamination of the data is unavoidable when monitoring real processes.

Standard monitoring tools can be easily extended using this model. In this work, examples are given with the statistics associated with Principal Component Analysis (PCA). The advantage of this approach is illustrated through its application to a toy example in two dimensions. Then, its performance is demonstrated by applying the tool to the Tennessee–Eastman Process (TEP) simulation benchmark and to industrial data belonging to a methanol plant subsection.

2. Model development and implementation

Clusters can be described as continuous regions of space containing relatively high densities of points, separated from other high-density regions by regions containing relatively low densities of points (Choi et al., 2003). Statistical approaches to cluster analysis have a strong theoretical background, and offer the advantages of being able to compute the cluster criteria to be optimized and to yield algorithms that effectively and efficiently reduce them (Gallegos & Ritter, 2005).

We propose to use a statistical model consisting of several normal distributions sharing a common covariance matrix for processes with multiple operation modes (MOM) where due to the physical relations between variables, the covariance structures share common characteristics. This will let us use a very powerful method to find the different clusters even in the presence of outliers.

The method proposed by Gallegos and Ritter (2005) considers precisely a contaminated set of n observations on d variables coming from g different, normally distributed, populations with a common covariance matrix. In their work, they first introduce a criterion for the clustering procedure (the trimmed determinant criterion—TDC) and demonstrate that it leads to maximum likelihood estimates of the parameters of the model (i.e., the means and the common covariance matrix of the g normal distributions). Moreover, they develop an algorithm and demonstrate that converges to the required minimum of the TDC in a finite number of steps. Finally, they compute the asymptotic breakdown values of the estimators and find results consistent with the robustness claim. Readers interested in the theoretical background are referred to the work by Gallegos and Ritter.

The algorithm partitions the r regular observations into g clusters and simultaneously detects $n–r$ outliers. It does so by choosing a subset of size r from the n observations and partition it into g clusters so that the pooled sum of squares and products (SSP) matrix has minimum determinant (TDC). The maximum likelihood estimate of the mean vectors of the different underlying normal distributions are the sample mean vectors of the various clusters, whereas that of the common covariance matrix is the pooled SSP matrix divided by r. The number $n–r$ of rejected data is a parameter of the model and the estimated means are fairly insensitive to the choice of this parameter provided it is not too large in comparison with the total available process data.

Starting from a configuration R (i.e., a subset of the data together with its partition into g clusters), the key step of the algorithm is to look for another configuration R_{new} such that the sum of square distances is smaller than the one in configuration R. This is done by assigning each observation to the cluster that minimizes the square distance $d_R(i,j)^2$. It has been demonstrated that the determinant of the pooled covariance matrix corresponding to the new configuration is smaller than the one corresponding to the previous one (Gallegos & Ritter, 2005).

The algorithm can be briefly described as follows: Given a starting configuration R, together with its mean vectors $\mathbf{m}_R$, and its SSP matrix $\mathbf{W}_R$,

$$\mathbf{W}_R = \sum_{j=1}^{g}\sum_{x \in R_j}(\mathbf{x}-\mathbf{m}_R(j))(\mathbf{x}-\mathbf{m}_R(j))^T \tag{1}$$

(i) Compute the Mahalanobis distance from each data point to the mean of each cluster:

$$d_R(i,j)^2 = (\mathbf{x}_i-\mathbf{m}_R(j))^T\mathbf{W}_R^{-1}(\mathbf{x}_i-\mathbf{m}_R(j)),$$
$$i \in 1,\ldots,n,\;\; j \in 1,\ldots,g \tag{2}$$

(ii) For each $i \in 1, \ldots, n$, find $j \in 1, \ldots, g$ that minimizes $d_R(i,j)^2$; that is, for each i determine the optimal cluster j.

(iii) Sort the square distances in ascending order.

(iv) Construct the new configuration R_{new}, considering the data subset corresponding to the first r sorted distances calculated in (iii) and assign to each point its optimal cluster j. Compute the new mean vectors $\mathbf{m}_{Rnew}$ and SSP matrix $\mathbf{W}_{Rnew}$ rejecting data whose distances sorted in (iii) are in the last n–r places (now considered outliers).

(v) If $\det(\mathbf{W}_{Rnew}) = \det(\mathbf{W}_R)$, stop. Else, $\mathbf{W}_R = \mathbf{W}_{Rnew}$ and $\mathbf{m}_R = \mathbf{m}_{Rnew}$, go to (i).

By iterating these steps, a sequence of configurations R_k that satisfies $\det(\mathbf{W}_{R_{k+1}}) \leq \det(\mathbf{W}_{R_k})$ is obtained. The process becomes stationary after a finite number of steps. The final configuration is one approximation to the minimum trimmed determinant. Multi-start optimization (i.e., starting by randomly assigning each data to any of the clusters) is applied to the foregoing iterative process; the limit configuration with the least value of the determinant of the corresponding SSP matrix is the final approximation to the minimum. Geometrically, the reduction in the determinant of the pooled covariance matrix represents a reduction in the volume of its associated ellipsoid (Bersimis, Psarakis, & Panaretos, 2007).

It is worth mentioning that the computed pooled covariance matrix could be biased. Since the algorithm willingly excludes the farthest n–r points, the variance will be underestimated if the number of actual outliers is less than n–r. This is not a problem for identifying the clusters, but the computed pooled covariance matrix will not be an appropriate estimator of the covariance matrix.

The partitioning procedure of assigning randomly each data point to a cluster has no physical meaning if successive data points correspond to successive times. The plant operation mode will never jump randomly from one cluster to another; points adjacent in time generally belong to the same cluster. Taking into account this consideration, the initial configuration is generated dividing the data set in g groups of consecutive (meaning sequential in time) data, imposing randomly the separation dates. In this way, fast convergence to minimum values of the trimmed determinants is attained.

3. Determination of the number of clusters

In the algorithm, the parameter g (number of clusters) is assumed to be given “a priori”. This is a limitation because, in real multivariate processes, the number of clusters is very often unknown. To overcome this limitation, a heuristic rule based on geometric considerations is proposed.

Taking into account that the determinant of the pooled covariance matrix is related to the volume of its associated ellipsoid (Bersimis et al., 2007), the algorithm is run for different number of clusters, g, and the volume associated with the underlying normal model, V, is computed as the square root of the pooled covariance matrix determinant of the final configuration.

$$V = \sqrt{\det(\mathbf{S}_{pool})} \tag{3}$$

Then, an objective function, Y, that relates the occupied volume of the space, V, with the number of clusters, g, and the space dimension, d, is defined as:

$$Y = V \cdot g, \quad g > d \tag{4}$$

$$Y = \left(\frac{g}{d}\right) V \cdot g + \left(1 - \frac{g}{d}\right) V \cdot 2^g, \quad g \leq d \tag{5}$$

The objective function considers that, when a new cluster is properly added, the space occupied by the ellipsoid associated with the pooled covariance matrix determinant should decrease with a factor related to the number of clusters. A cluster in excess will lead to a decrease of the associated volume which is less significant. The dimension of the space (i.e., the number of variables) will also affect the shrinking of the ellipsoid, which will be more important when the dimension is larger than the number of clusters. If a proper factor is used, the objective function will indicate when the addition of a new cluster does not reduce significantly the volume any more. Then, the value of g that minimizes Y will indicate the optimum number of clusters. Many factors have been tested with different simulations considering space dimensions between 2 and 100 and 2–20 clusters, and those indicated in Eqs. (4) and (5) gave the best results. Notwithstanding, research is still ongoing particularly comparing the developed heuristic method with others, more computing demanding, statistical procedures described in the literature, like the GAP statistic method (Tibshirani, Walther, & Hastie, 2000), to establish the best methodology for determining the number of clusters.

4. Standard and clustered statistics

A common procedure for reducing the dimensionality of the variable space is the use of projection methods like Principal Components Analysis (PCA) (Bersimis et al., 2007). These methods are based on reducing the sample variance–covariance matrix $\mathbf{S}$, to a diagonal matrix $\mathbf{L}$ by premultiplying and postmultiplying it by a particular orthonormal matrix $\mathbf{U}$ such that $\mathbf{U}^T\mathbf{S}\mathbf{U} = \mathbf{L}$. The diagonal elements of $\mathbf{L}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are the eigenvalues of $\mathbf{S}$, and the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{S}$, also called the loading vectors ($\mathbf{u}_i$). The covariance matrix $\mathbf{S}$ is calculated from a given a set of n vectors, corresponding to measurements of d variables under normal plant operation. When variables are measured in different units, the vectors must be normalized to standard units. In those conditions, the covariance matrix $\mathbf{S}$ calculated with the normalized vectors is the correlation matrix $\mathbf{R}$ of the original vectors. New measurements properly normalized and arranged in a $m \times d$ matrix $\mathbf{Y}$, are then projected by $\mathbf{T} = \mathbf{YU}$, or $\mathbf{t}_i = \mathbf{Yu}_i$ $i = 1, 2, \ldots, d$. The $\mathbf{t}_i$, called the score vectors, are the columns of $\mathbf{T}$ and different statistics can be calculated to decide whether the measurements remain under control or not. For instance, charts based on the Hotelling's T^2 can be plotted based on the first a principal components (Eq. (6)). Another useful statistics is the Q statistics (Eq. (7)). Each statistics defined by Eqs. (6) and (7) provide complementary information.

$$T^2 \text{ on the } a \text{ retained eigenvectors}: T_a^2 = \sum_{i=1}^{a} \frac{t_i^2}{\lambda_i} \tag{6}$$

$$Q \text{ statistics}: Q = \sum_{i=a+1}^{d} t_i^2 \tag{7}$$

This procedure can be easily extended to the case where multiple clusters are present, considering the different cluster means and the pooled covariance matrix instead of the global mean and the standard covariance matrix.

5. Results and discussion

5.1. Case 1: toy example

To exemplify the performance of the clustering technique, we have applied it to an artificially generated data set. The controlled data set consists of four uncontaminated clusters with the parameters detailed in Table 1.

The result of applying the algorithm with multistart optimization to the artificially generated data set is shown in Fig. 1. As observed in the figure, the proposed method succeeds in determining the clusters in spite of their different density.

To analyze the robustness of the method for handling contaminated data, the same data set is contaminated with 200 outliers, following a Gaussian distribution with a standard deviation of 10.

Table 1
Parameters of the artificially generated data set for illustrating the clustering method.

Cluster	Covariance matrix	Mean	Number of points
1	$\begin{bmatrix} 8.5 & 7.5 \\ 7.5 & 8.5 \end{bmatrix}$	[8 8]	400
2		[4 −4]	800
3		[−4 4]	1600
4		[−8 −8]	2000

Note that the number of outliers represents 4% of the data set and 50% of the less dense cluster.

The algorithm shows its robustness by successfully partitioning the data despite the presence of outliers. Fig. 2a and b shows the results of running the algorithm changing the parameter to account for 200 and 500 outliers, respectively. The former case corresponds to the “true” number of outliers, while the latter is more conservative, overestimating the number of outliers. Both cases lead to the same results (i.e., the same means and clusters sets) and can be used as a good starting point for a robust covariance matrix calculation.

5.2. Case 2: Tennessee–Eastman Process (TEP)

The Tennessee–Eastman Process (TEP) simulation benchmark, presented by Downs and Vogel (1993), is the simulation of a complex industrial chemical process. It has been used as a benchmark, especially for studying advanced control strategies. In the last decade, it has increasingly been used to test the performance of proposed MSPC tools. Fig. 3 gives the well known flow sheet of the TEP (Ricker, 1996). The process has five major units: a reactor, a condenser, a vapor–liquid separator, a recycle compressor, and a product stripper. It involves two simultaneous gas–liquid exothermic reactions that produce two desired products (G and H) and a byproduct F which is produced from two additional reactions, from four reactants A, C, D and E. Within the process there is also an inert B. The process has 12 manipulated variables and 41 measured variables for monitoring and control. About half of the measured variables are component compositions, available at discrete sampling intervals of 0.1 or 0.25 h. The remaining 22 measured variables are available at significantly higher sampling frequency. The original process is open-loop unstable and, in the absence of feedback, small perturbations eventually lead to a shutdown; then, a control strategy must be introduced. Here, the advanced decentralized control strategy presented by Ricker (1996) is employed for its capability of less variability in the product rate and quality, and of operating on-spec for long periods without feedback from composition measurements. The involved control loops for the considered strategy are indicated in Fig. 3. For the sake of practical consideration, the 22 continuous outputs among the 41 measurements are used for monitoring and the sampling interval is 0.01 h. The simulation programs are available at Ricker’s home page (Ricker, 2008).

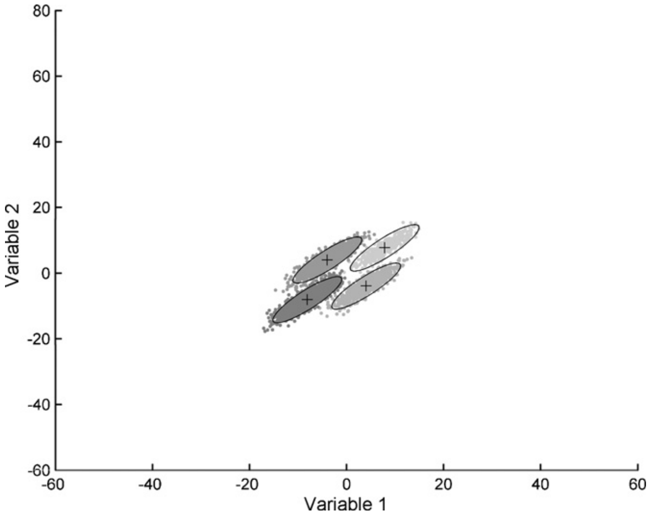

Fig. 1. Normal operation regions identified by the clustering methods for the toy example in two dimensions with uncontaminated data.

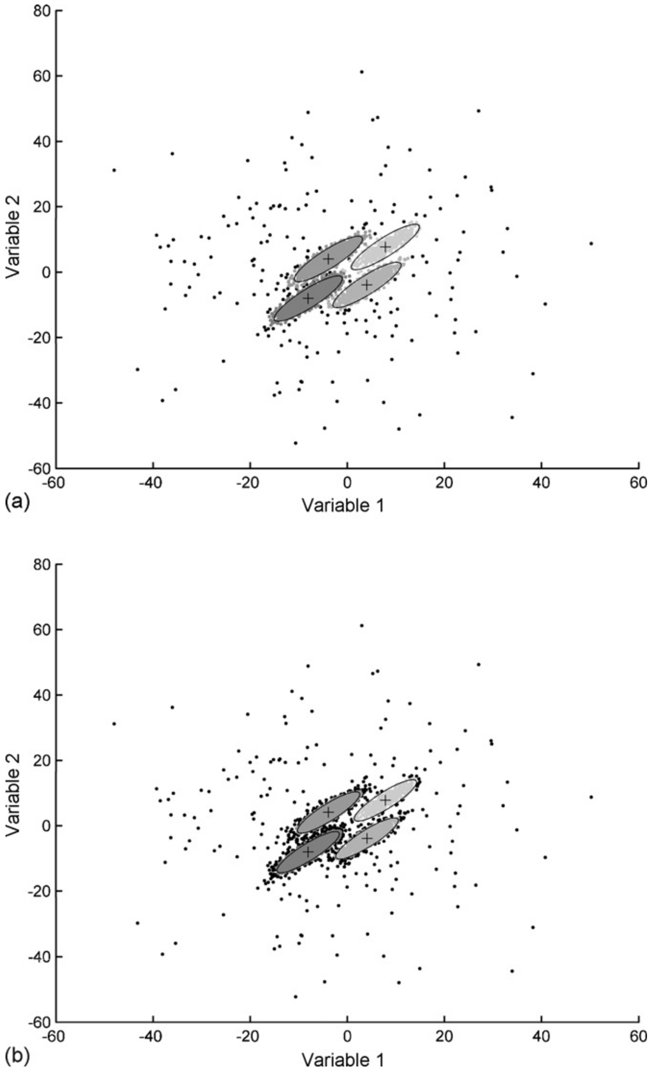

Fig. 2. Normal operation regions identified by the clustering methods for the toy example in two dimensions imposing different values of the parameter r to establish the number of outliers: (a) $n–r = 200$, outliers representing 4% of the total data; (b) $n–r = 500$, outliers representing 10% of the total data. Black dots indicate the outliers identified by the method.

The simulation programs have been implemented and data corresponding to a set of two “normal process operation” in the sense that they lead to a product with the same specification are obtained by imposing modifications in the operation strategy. The imposed modifications are indicated in Table 2 and arise from a controlled change in reactor pressure for which drifts in the production output

Table 2
Different steady-state modes of operation that define the clusters considered for the TEP example.

	Reactor pressure (kPa)	Stripper underflow drifts (m^3/h)
Cluster 1	2800	22.4–18.6 and 17.9–17.2
Cluster 2	2680	22.9–21 and 19.1–17.2

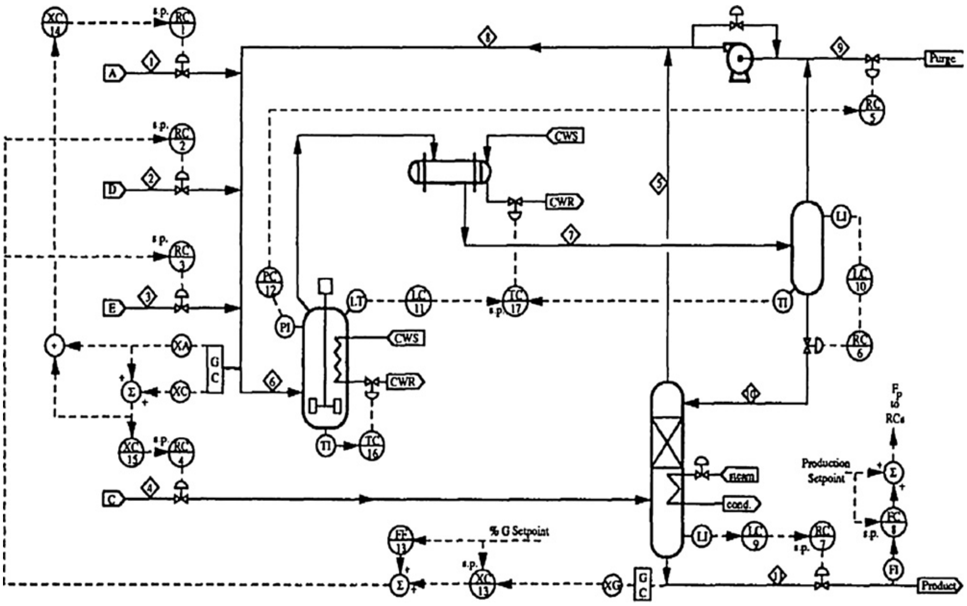

Fig. 3. Flow sheet of the Tennessee–Eastman Process (TEP) indicating the control loops used by Ricker (1996).

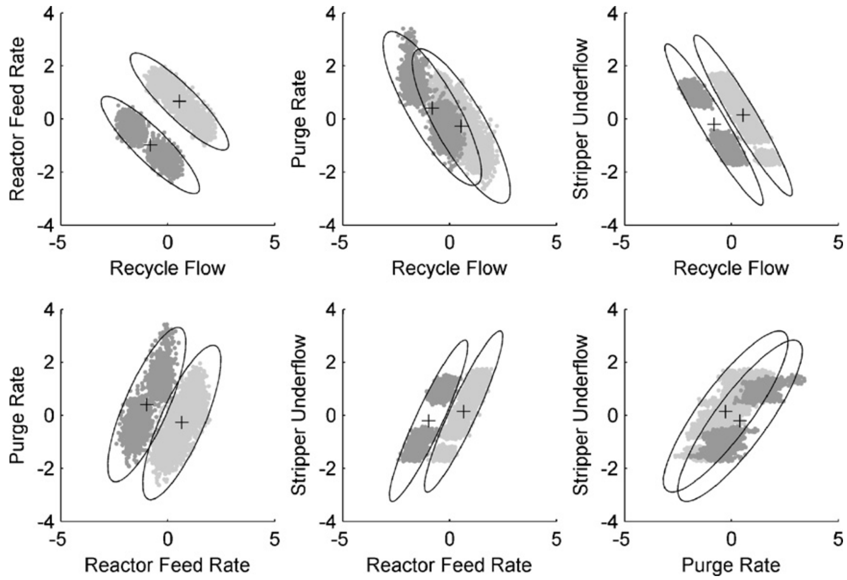

Fig. 4. Normal operation regions identified by the proposed clustering method for the TEP example considering no outliers.

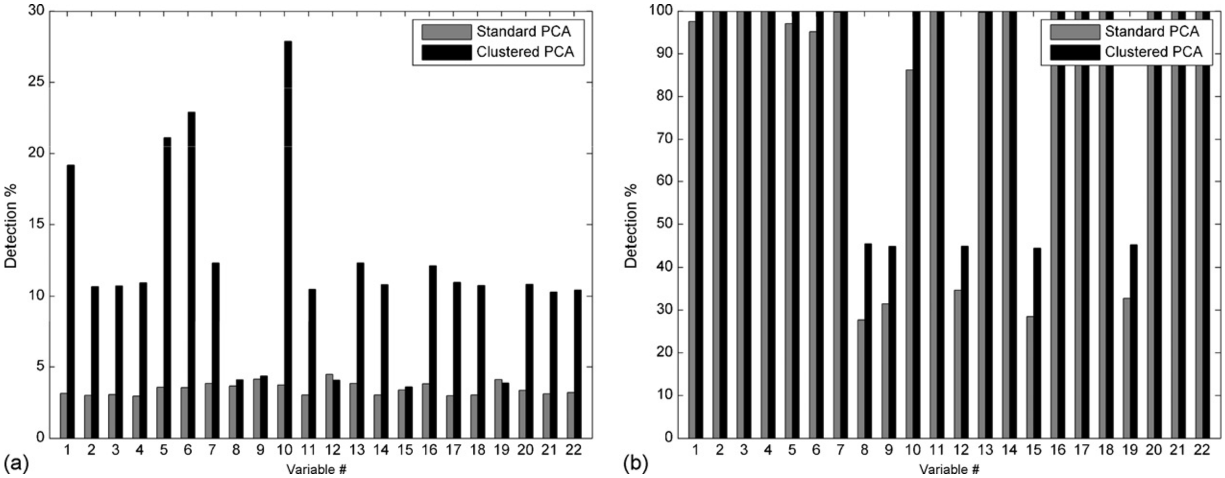

Fig. 5. Detection capability of the joint output of the complementary statistics, T_a^2 (Eq. (6)) and Q (Eq. (7)), using the standard and the proposed clustering method when forcing an out-of-range error of each of the supervised variables in the TEP for each instant. Levels of univariate drift: (a) 1 standard deviation; (b) 3 standard deviations.

Table 3
List of the standardized faults in the TEP simulation benchmark.

Fault	Description	Type
1	A/C feed ratio, B composition constant (Stream 4)	Step
2	B composition, A/C ratio constant (Stream 4)	Step
3	D feed temperature (Stream 2)	Step
4	Reactor cooling water inlet temperature	Step
5	Condenser cooling water inlet temperature	Step
6	A feed loss (Stream 1)	Step
7	C header pressure loss – reduced availability (Stream 4)	Step
8	A, B, C feed composition (Stream 4)	Random variation
9	D Feed Temperature (Stream 2)	Random variation
10	C feed temperature (Stream 4)	Random variation
11	Reactor cooling water inlet temperature	Random variation
12	Condenser cooling water inlet temperature	Random variation
13	Reaction kinetics	Slow drift
14	Reactor cooling water valve	Sticking
15	Condenser cooling water valve	Sticking
16	Unknown	–
17	Unknown	–
18	Unknown	–
19	Unknown	–
20	Unknown	–

Table 4
Comparison between missing and false alarm rates arising from monitoring with the standard statistics and the one computed by present method for the faults in the TEP example (similar cluster sizes).

	Standard			Clustered		
	T_a^2	Q	Joint	T_a^2	Q	Joint
False alarms (%)	1.00	1.00	2.00	1.00	1.00	1.99
Missing alarms (%)						
Fault 1	58.50	24.13	23.50	63.50	17.63	17.13
Fault 2	91.88	36.38	35.63	94.00	27.88	27.25
Fault 3	98.88	98.50	97.50	99.00	100.00	99.00
Fault 4	86.50	86.88	84.63	85.38	100.00	85.38
Fault 5	98.88	98.63	97.50	98.63	100.00	98.63
Fault 6	98.38	64.25	63.50	97.88	49.88	49.25
Fault 7	71.25	10.38	7.88	50.38	4.75	2.38
Fault 8	98.00	70.38	69.38	94.25	62.88	61.63
Fault 9	98.88	98.50	97.38	98.63	100.00	98.63
Fault 10	98.75	98.50	97.38	98.88	100.00	98.88
Fault 11	12.13	16.13	10.50	10.63	90.50	10.63
Fault 12	98.13	98.75	96.88	98.25	100.00	98.25
Fault 13	98.88	98.50	97.50	98.88	100.00	98.88
Fault 14	10.25	14.25	9.25	8.63	100.00	8.63
Fault 15	99.00	98.63	97.63	99.00	100.00	99.00
Fault 16	98.88	98.50	97.50	98.88	100.00	98.88
Fault 17	93.50	89.63	85.50	90.50	83.50	80.00
Fault 18	98.88	98.50	97.50	98.88	100.00	98.88
Fault 19	99.13	98.63	97.75	98.88	100.00	98.88
Fault 20	97.50	96.13	94.13	98.13	85.75	84.50

are imposed. It should be mentioned that previous works analyzing multiple operation modes for the TEP example have considered modes that lead to products with different compositions (Chen & Liu, 1999; Ge & Song, 2008; Zhao et al., 2004), which is not the goal of the present work. We have particularly analyzed the case of different operation modes that lead to a product with the same specification. It is also worthwhile to mention that the method performance highly increase its capabilities if the obtained products have modified composition, since the differences considered in this case are extremely more subtle.

Fig. 4 illustrates how the implemented algorithm clusterizes the data considering the 22 variables measured with relatively high frequency (0.01 h). The minimum in the heuristically defined objective function (Eqs. (4) and (5)) to estimate the number of clusters is obtained with two clusters. The figure presents typical relations found among several of the considered variables.

It can be observed that the method does separate operation performed under different reactor pressure. However, the drift in production output remains within the same cluster, even if a discontinuity in the drift was purposely imposed. This result arises from no changes in the majority of the physical relations which govern the correlation between variables under a given operation strategy, within a reasonable production output.

To illustrate the advantages of clustering the data in such a way, two examples based on the TEP are presented. In both cases, the PCA statistics are calculated for each test point with respect to the mean of each cluster, using the pooled covariance matrix resulting from the application of the proposed clustering technique. Each data is assigned to a particular cluster according to the minimum distance. The number of principal components is selected to explain 95% of the variability of the data, resulting in 6 for all cases. The control limit is set at the 99 percentile of the statistics calculated from data corresponding to normal operation. Then, missing and false alarms are compared with those obtained when the standard statistics (i.e., without clustering) are used. Given that PCA involves two different statistics, the results are summarized considering the joint information (i.e., if any statistic is over the control limit, a fault is detected, and both statistics under the control limit indicate normal operation).

In the first example, a drift is artificially introduced in each variable of each data point to simulate malfunctions in instruments. It is important to start from each data point to make sure that the results do not depend on the starting point. Then, the number of detections (i.e., occasions when the statistics is greater than its control limit) is computed using either the standard or the clustered statistics. Results are shown in Fig. 5, proving improvement in the sensibility to different drifts in measured variables at two levels of deviations from the mean.

The second example consists in the simulation of all the 20 faults defined for the TEP, starting from the same initial condition. The faults are listed in Table 3. Again, the statistics are calculated in the same way, but in this case the control limits are set to achieve, in all cases, 1% of false alarms as recommended by Russell, Chiang, and Braatz (2000). Two different situations are also compared. In

Table 5
Comparison between missing and false alarm rates arising from monitoring with the standard statistics and the one computed by present method for the faults in the TEP example (different cluster sizes).

	Standard			Clustered		
	T_a^2	Q	Joint	T_a^2	Q	Joint
False alarms (%)	1.00	1.00	2.00	1.00	1.00	1.99
Missing alarms (%)						
Fault 1	44.75	23.63	22.75	65.75	16.75	16.25
Fault 2	89.75	35.50	35.00	94.88	28.00	27.38
Fault 3	98.88	98.75	97.63	98.88	100.00	98.88
Fault 4	87.13	86.38	84.38	85.63	100.00	85.63
Fault 5	98.88	98.75	97.63	98.63	100.00	98.63
Fault 6	98.00	58.38	57.50	97.88	49.50	48.88
Fault 7	47.25	14.88	8.50	48.00	5.50	3.00
Fault 8	97.75	70.88	69.88	93.63	61.88	60.50
Fault 9	99.00	99.00	98.00	98.75	100.00	98.75
Fault 10	98.75	99.00	97.75	98.88	100.00	98.88
Fault 11	12.25	14.50	10.50	10.63	89.75	10.63
Fault 12	98.50	98.38	96.88	98.38	100.00	98.38
Fault 13	98.88	99.00	97.88	98.88	100.00	98.88
Fault 14	10.75	12.75	8.75	8.88	100.00	8.88
Fault 15	98.88	99.00	97.88	99.00	100.00	99.00
Fault 16	98.88	99.00	97.88	98.88	100.00	98.88
Fault 17	93.88	89.38	85.00	90.50	85.00	81.25
Fault 18	98.88	99.00	97.88	98.88	100.00	98.88
Fault 19	99.00	99.13	98.13	99.00	100.00	99.00
Fault 20	96.00	93.75	91.13	98.50	85.63	84.50

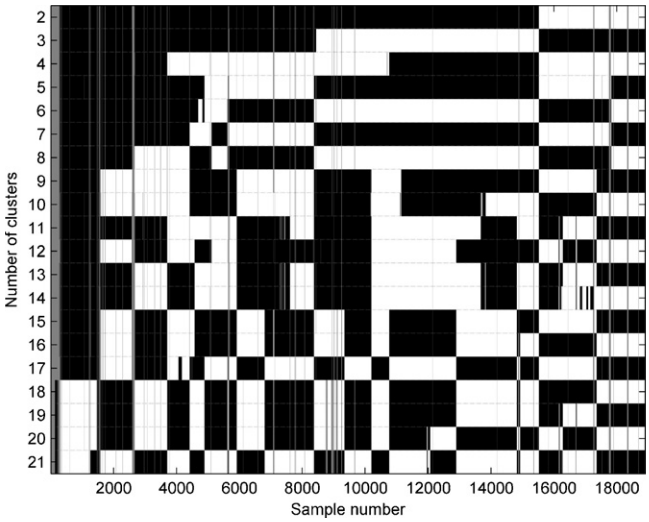

Fig. 6. Separation among clusters predicted by the algorithm, as different numbers of clusters, *g*, are imposed.

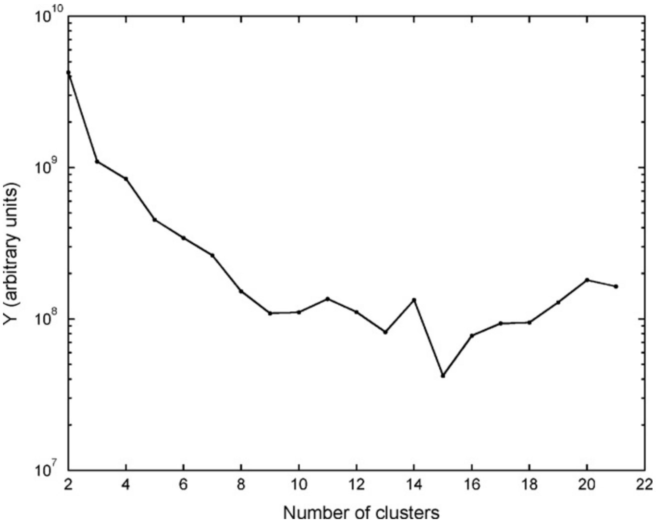

Fig. 7. Evolution of the objective function defined by Eqs. (4) and (5) as the number of clusters is progressively increased. Minimum found for $g = 15$ clusters.

the first, the two clusters corresponding to normal operation have a similar number of points, whereas, in the second case, the number of points in cluster 1 is ten times larger than the number of points in cluster 2. Results of the statistics defined by Eqs. (6) and (7) are detailed in Tables 4 and 5, respectively, for similar and different cluster sizes. Missing and false alarm rates evaluated through the joint information of the complementary statistics T_a^2 and Q are also detailed in the tables. From the results, it comes out that, even if many faults are not detected by either method, whenever there is a significant change in the missing alarm rate, the clustered statistics have improved detection capability. Note that the non-faulty testing data is different from the training data and that the control strategy is very strong. Actually, if a less advanced control strategy is used, the percentage of missing alarms noticeably diminishes for both methods and the performance of the clustered method is highly superior to the standard PCA.

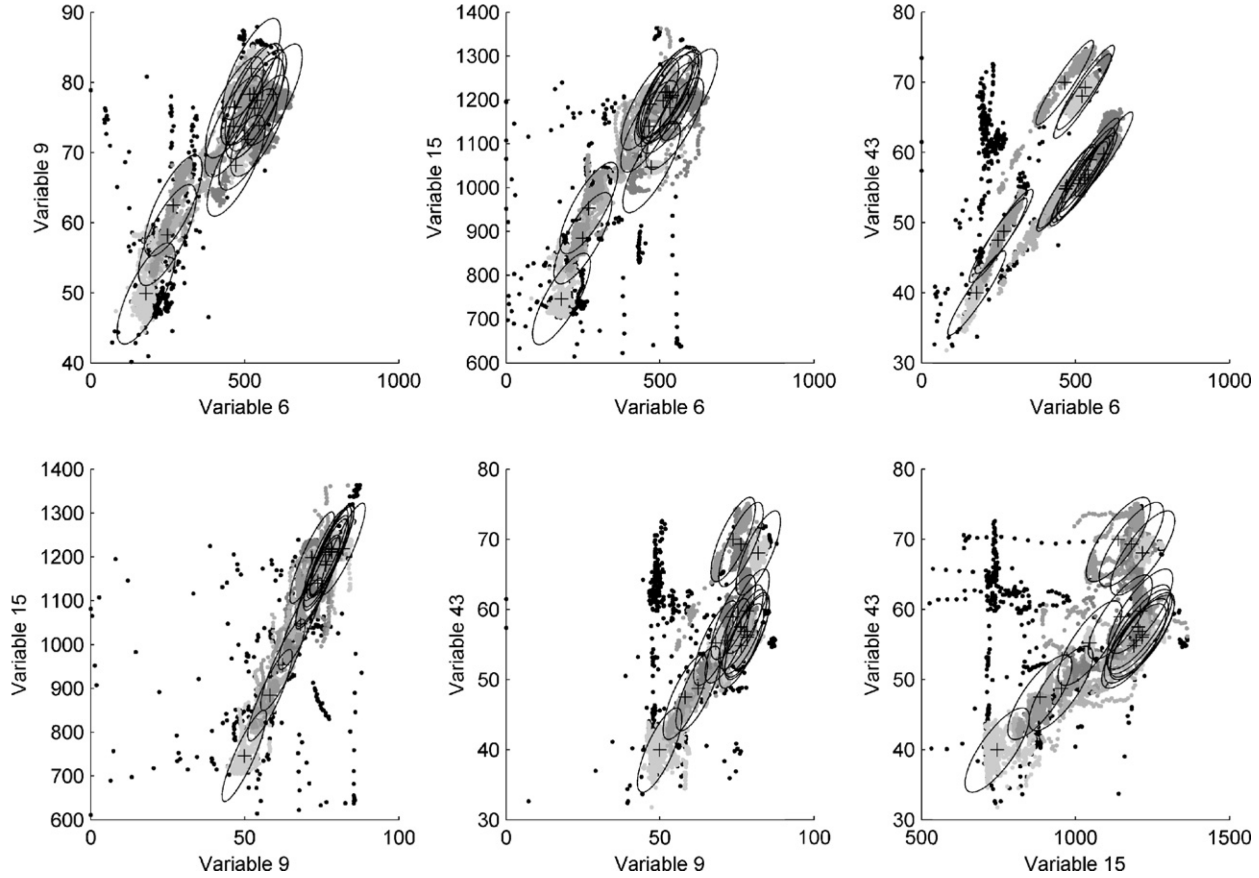

Fig. 8. Normal operation regions identified by the proposed clustering method for the methanol plant example considering that 5% of the data are outliers (identified by black dots).

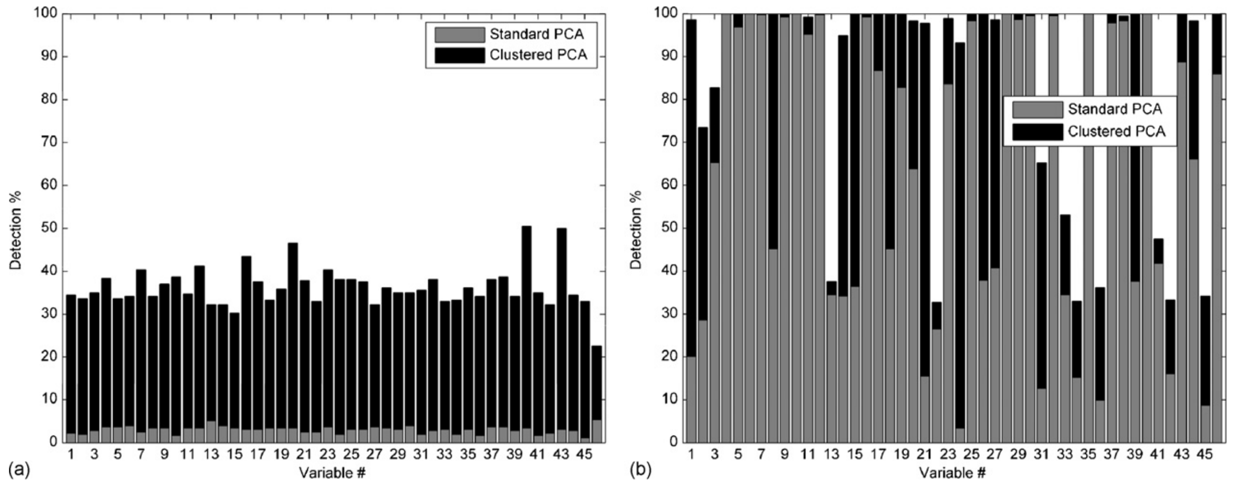

Fig. 9. Detection capability of the joint output of the complementary statistics T_a^2 (Eq. (6)) and Q (Eq. (7)), using the standard and the proposed clustering method when forcing an out-of-range error of each of the supervised variables in the methanol plant for each instant. Levels of univariate drift: (a) 1 standard deviation; (b) 3 standard deviations.

5.3. Case 3: methanol purification plant

Process data collected for monitoring the operation of the purification section of a methanol plant owned by YPF are employed for assessing the strategy in an actual industrial environment. The purification section consists of a standard arrangement of two distillation columns. The first column separates the light products from the methanol–water mixture. The bottoms of this column are treated in the second column, where the methanol is distilled in the top and process water is extracted from the bottom.

For monitoring and analysis, 46 variables were continuously followed for more than 3 years. The process data collected during periods of normal operation were used for identifying the clusters corresponding to different operation modes and for training the tools.

The proposed clustering algorithm is applied to the process data, considering different number of clusters (i.e., different values of g), and assuming that 5% of the data are outliers. The number of outliers is generally established based on the experience of the plant personnel since they depend mostly on the recollection and transmission processes and also on the periods when the plant should stop or operate with low capacity due to fuel restrictions and following starts-up.

The data clusters arising from applying the proposed algorithm are shown in Fig. 6 represented as a function of time units. Clusters are marked alternatively in black and white to remark the dates corresponding to a cluster change. It is remarkable that, whatever is the initial condition imposed, many cluster separation dates are coincident for different values of g. For instance, for $g > 2$, the algorithm always groups in different clusters data measured before and after the date corresponding to 15,400 time units, which is coincident with a change in plant operation strategy. The same observation is valid for the outliers (marked in grey); that is, there are points recursively classified as outliers by the algorithm independently of the number of clusters considered.

To establish the optimum number of clusters, the proposed heuristic rule based on minimising the objective function defined by Eqs. (4) and (5) was applied (Fig. 7). Once more, the number of clusters corresponding to the minimum of this figure is considered as a good choice by the YPF personnel, and cluster divisions can be successfully associated to modifications in the plant operation.

Some representative projections of the data, grouped in 15 clusters according to the result of the minimization, are illustrated in Fig. 8.

Following the same procedure as in the TEP example, a drift in each variable of each data point was artificially introduced to test the improvement in the sensibility of the proposed monitoring tool with respect to the standard methodology. For an explained

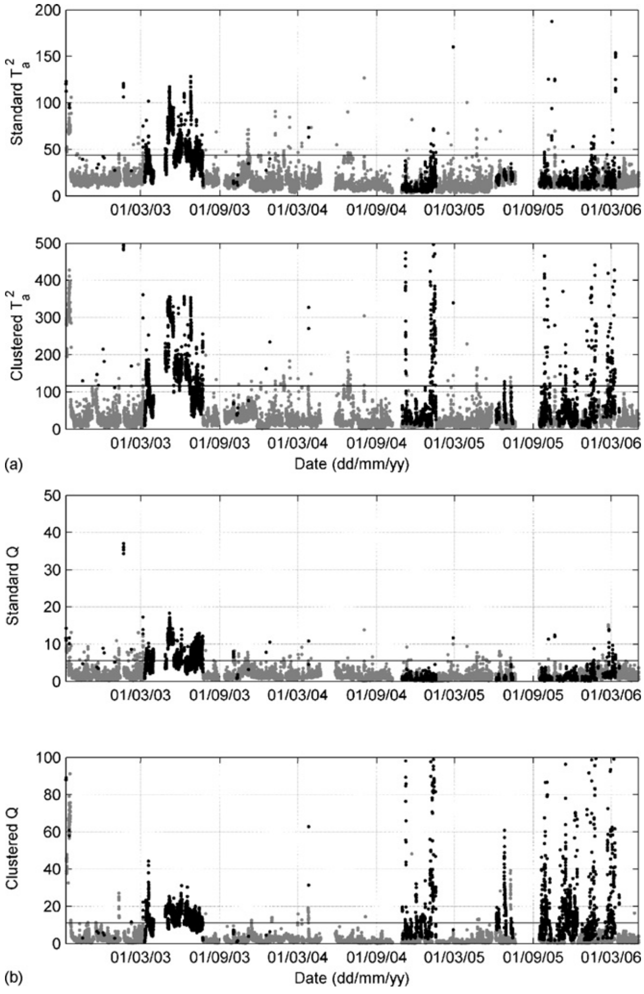

Fig. 10. Control charts for standard and clustered statistics. Grey points correspond to normal operation, black points correspond to faults and horizontal lines are the estimated control limits for each statistics. (a) T_a^2 (Eq. (6)); (b) Q (Eq. (7)).

Table 6
Comparison between missing and false alarm rates arising from monitoring with the standard statistics and the ones computed by present method for two different faults in the real methanol plant. Joint refers to using the joint information of the T_a^2 and the Q.

	Standard			Clustered		
	T_a^2	Q	Joint	T_a^2	Q	Joint
False alarms (%)	1.0	1.0	1.7	1.0	1.0	1.4
Missing alarms (%)						
Fault 1	51.0	46.3	33.3	30.2	39.0	22.9
Fault 2	98.5	99.3	98.0	46.1	87.9	46.1
Total	77.9	76.0	69.8	39.7	66.5	36.1

variance of 95%, the number of retained components was 22 for the standard methodology and 21 for the clustered one. There is a remarkable increase in detection sensibility at the two levels of drift examined (Fig. 9). When considering the joint information of the Q and T_a^2, the improvement is remarkable for all the variables particularly at the drift level of 1 standard deviation.

Fig. 10 shows the charts for standard and clustered statistics (Eqs. (6) and (7)) for a period spanning 40 months of plant operation. In both cases the limits for the control charts were set as the 99 percentile of each statistics calculated from the databank corresponding to normal operation. During this period, two faults occurred. The first one involved the malfunction of a manometer at the top of the second column (roughly between 01/03/2003 and 01/09/2003). The second one was an intermittent failure of a temperature sensor located mid-height of the second distillation column (starting about 2 months after 01/09/2004 and repeated intermittently until 01/03/2006).

Both methods succeeded in detecting appropriately the first fault. However, there is a clear improvement in the detection capability of the clustered method for the second fault, which is observed in the computed statistics. Missing and false alarms arising from these results are detailed quantitatively in Table 6.

While both methods lead to similar percentages of false alarms, significant differences are found in detecting the two documented malfunctions. These results highlight the benefits of using the suggested clustering methodology for properly taking into account the different operating modes that existed in the methanol plant.

6. Conclusions

A new MSPC technique is proposed in this paper to address the problem of monitoring processes with multiple operations modes. This approach relies on a robust clustering method, assuming that the different clusters share a common covariance matrix, preserving the physical relations between variables. Moreover, the method is capable to cope with the presence of outliers. A procedure to determine the optimum number of clusters is also proposed. The performance of this technique is tested on the TEP benchmark and with real data from a methanol plant, thus establishing the feasibility of its implementation in industrial environments.

Acknowledgments

Financial support from Fundación YPF, CONICET, ANPCyT and Universidad de Buenos Aires, Argentina, are gratefully acknowledged. Pablo Groisman and Miryan Cassanello are members of CONICET. The authors also wish to acknowledge the operations staff of the methanol plant of YPF (Plaza Huincul).

References

Bersimis, S., Psarakis, S., & Panaretos, J. (2007). Multivariate statistical process control charts: An overview. *Quality and Reliability Engineering International*, *23*(5), 517–543.

Chen, J., & Liu, J. (1999). Mixture principal components analysis models for process monitoring. *Industrial & Engineering Chemistry Research*, *38*, 1478–1488.

Choi, S. W., Yoo, C. K., & Lee, I. B. (2003). Overall statistical monitoring of static and dynamic patterns. *Industrial & Engineering Chemistry Research*, *42*(1), 108–117.

Downs, J. J., & Vogel, E. F. (1993). A plant-wide industrial process control problem. *Comparative Chemical Engineering*, *17*, 245–255.

Gallegos, M. T., & Ritter, G. (2005). A robust method for cluster analysis. *The Annals of Statistics*, *33*, 347–380.

Ge, Z., & Song, Z. (2008). Online monitoring of nonlinear multiple mode processes based on adaptive local model approach. *Control Engineering Practice*, *16*, 1427–1437.

Hwang, D.-H., & Han, Ch. (1999). Real-time monitoring for a process with multiple operating modes. *Control Engineering Practice*, *7*, 891–902.

Lane, S., Martin, E. B., Kooijmans, R., & Morris, A. J. (2001). Performance monitoring of a multi-product semi-batch process. *Journal of Process Control*, *11*, 1–11.

Ricker, N. L. (1996). Decentralized control of the Tennessee Eastman Challenge Process. *Journal of Process Control*, *6*(4), 205–221.

Ricker, N. L. (2008). *Tennessee Eastman Challenge Archive*. Available for download at http://depts.washington.edu/control/LARRY/TE/download.html.

Russell, E. L., Chiang, L. H., & Braatz, R. D. (2000). Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *51*, 81–93.

Srinivassan, R., Wang, C., Ho, W. K., & Lim, K. W. (2004). Dynamic principal components analysis based methodology for clustering process states in agile chemical plants. *Industrial & Engineering Chemistry Research*, *43*, 2123–2139.

Tibshirani, R., Walther, G., & Hastie, T. (2000). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *63*(2), 411–423 (2001)

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S. N., & Yin, K. (2003). A review of process fault detection and diagnosis. Part III. Process history based methods. *Computers and Chemical Engineering*, *27*, 327–346.

Yoo, C. K., Vanrolleghem, P. A., & Lee, I. B. (2003). Nonlinear modelling and adaptive monitoring with fuzzy and multivariate statistical methods in biological wastewater treatment plants. *Journal of Biotechnology*, *105*, 135–163.

Zhao, S. J., Zhang, J., & Xu, Y. M. (2004). Monitoring of processes with multiple operating modes through multiple principle component analysis models. *Industrial & Engineering Chemistry Research*, *43*, 7025–7035.

Zhao, S. J., Zhang, J., & Xu, Y. M. (2006). Performance monitoring of processes with multiple operating modes through multiple PLS models. *Journal of Process Control*, *16*, 763–772.