Full Length Research Paper

New monitoring method based principal component analysis and fuzzy clustering

Khaled Ouni^{1*}, Hedi Dhouibi¹, Lotfi Nabli² and Hassani Messaoud²

¹The High Institute of Applied Sciences and Technology of Kairouan, Kairouan, Avenue Beit ElHekma, 3100 Kairouan, Tunisia.

²The National Engineering School of Monastir, Avenue Ibn ElJazzar, 5019 Monastir, Tunisia.

Accepted 7 May, 2012

This work concerns the principal component analysis applied to the supervision of quality parameters of the flour production line. Our contribution lies in the combined use of the principal component analysis technique and the clustering algorithms in the field of production system diagnosis. This approach allows detecting and locating the system defects, based on the drifts of the product quality parameters. A comparative study between the classification performance by clustering algorithms and the principal component analysis has been proposed. Locating parameters in defect is based on the technique of fault direction in partial least square.

Key words: Fuzzy clustering, fault detection, fault location, principal component analysis (PCA).

INTRODUCTION

During the last years, health monitoring of complex manufacturing, which is markedly useful, durable and dependable for the modern industrial drive systems has become very significant in decreasing the unprogrammed downtimes of the appropriate production lines. The accurate and fast isolation of defects can assess the performance process, and the current defect flour processing industry improves the efficiency process and product quality. Early defect detection may help avoid many breakdown and incidents in a flour production line. To use historical data for process monitoring, we have to isolate abnormal process data from a mixture of normal data and abnormal historical data. Accordingly, detecting defects is essential to obtain a high flour quality. The validity of the delivered information is so important to compute an efficient control.

A flour production line is a multi variable complex system. For this kind of system, it is generally not appropriate to use the analytical redundancy methods through an input-output model. In fact, it is often difficult

to design a mathematical model which allows obtaining an efficient diagnosis system. To reach this target, we may consider the implicit modeling approaches (Yingwei and Yang, 2012), based on the data driving techniques such as the Principal component analysis (PCA). These methods are well adapted to emphasize the relationships between the plant variables, without the explicit expression of the system model (Kresta et al., 1991).

In the literature associated with the field of diagnosing defects in the industrial systems, we find many statistical techniques designed to make use of the historical data and relations between the variables. Among these techniques, we can mention the use of the correlation matrix in the data sample, using the PCA to capture the most important directions in the correlation amongst the variables.

These statistical techniques are pre-calculated, based on a reference data sample; and as soon as the new data become available, the statistical approach can be quickly calculated on the parameters. These statistical methods

*Corresponding author. E-mail: ouni_khaled@yahoo.fr. Tel: 0021697787676.

have also been spread out for the monitoring of flexible processes of a complex manufacturing production. During the treatment of the process containing a great number of variables, the PCA can be used to reduce them, by grouping them into new sub-spaces of a reduced dimension. In determining the structure and classifying data, the fuzzy clustering techniques offer an important insight, while producing the affiliation functions for each group or cluster. A considerable number of fuzzy clustering algorithms have been developed with widely known methods like the Fuzzy C-Means (FCM), the Gustafson Kessel (GK) and Gath-Geva (GG) algorithms (Pal et al., 2005; Oliveira and Pedrycz, 2007).

This literature evaluation indicates an increasing tendency concerning the contribution of the statistical method like the PCA with the artificial intelligence method, as well as neural networks and the fuzzy logic. Bouhouche et al. (2007) proposed a combined use of the PCA and SOM algorithms. In this combined method, new detection indexes, based on metric distances, were used; all these indexes were used in the diagnosis phase. The SOM algorithm showed a bad classification compared to the PCA and the PCA-SOM. Ahvenlampi et al. (2005) proposed in these works, an approach of diagnosis of a digester system. In this study, measurements and some statistical variables are combined with fuzzy logic to produce key factors for the diagnosis. The combination between the PCA and the fuzzy logic has shown interesting results for the detection of defects and the diagnosis.

Also, many works have combined the PCA with the clustering techniques for the detection of defects and the classification of regions. Sebzalli and Wang (2001) presented in this work, an industrial study that uses the PCA and clustering algorithms to identify the operational spaces. This is a case study applying the PCA and the Fuzzy C-Means technique to the data of a refinery liquid. Ahvenlampi and Kortela (2005) presented an application on a quality parameter called kappa. It is a quality measurement in the kitchen processes. The clustering technique and the defect diagnosis system are used to control the quality variables. Peter et al. (2005) proposed in their work, a combined approach between the PCA and the k-means clustering algorithm. The defects are detected by the PCA and the classifier, in the form of regions with and without defects, using the k-means clustering. This approach is efficient in detecting and isolating defects in a process of industrial films. Ouni et al. (2011) presented a combined approach between the non linear PCA and the partial least square (PLS) for the monitoring in the tobacco manufacturing industry. The SPE statistic and the PLS-2 are applied to the defect detection and the region classification, respectively.

This work puts forward a combined approach of detecting and locating parameters of quality, in defect based on real data in a flour production line.

In this paper, we present a new analysis of flour

quality variables using the statistical methods and fuzzy clustering techniques. The PCA is used to find the correlations between the quality parameters. The PCA and its statistical indexes SPE and T^2 guarantee the detection of defects in all the data space. Contribution calculations can facilitate the isolation of defects, but these contributions do not always provide variable results. Improving the isolation of parameters in defect can be realized by the method combining the PCA and the Fisher Discriminant Analysis (Peter et al., 2005). Likewise, our combined approach based on the PCA and fuzzy clustering technique aims at separating between the normal and abnormal data in the form of regions in a 2D space. The directions of these regions are used as a new approach of contribution calculations in PLS.

In this paper, we suggest to extend this approach by using a PCA proposed in Ouni et al. (2011) and Chaouch et al. (2011). Also, we present an improvement of the operating region definition by studying the dynamic of the squared predictor error and the statistic T^2 . These regions are visualized in a 2D space, using the clustering algorithms. Therefore, we suggest the technique of PLS defect direction between the regions two to two. The weights of its directions are used in isolating the defected quality parameters.

This paper is organized as follows: The lines coming afterwards are firstly devoted to the situation of our contribution. After a brief description of the PCA technique, our contribution is highlighted at the level of the different modules of functional decomposition of the proposed diagnosis system. The last part of this work is devoted for an industrial application on a flour production line, and to a comparative study between the classification performance by FCM, GK, GG clustering and the PCA.

Principal component analysis (PCA)

The PCA is a multidimensional statistical method, which allows us to synthesize a set of data, while identifying the existing redundancy in them (Jolliffe, 1986). If at the origin of its development, it has been known to be attractive, while showing how to graphically represent the data groups, we should highlight the correlations between observations and variables. It is actually with the recent developments that a method of quantitative appreciation has known an informative content of the observations, which allows us to comprehend many problems such as searching a model structure, identifying model parameters, detecting aberrant values, detecting changes of the functioning flow rates, and diagnosing the systems functioning (Harket et al., 2006).

The PCA principle

We consider a vector of centered measurements

Figure 1. Detecting and locating defected quality parameters.

(nil average), $x = [x_1, x_2, \dots, x_m]^T \in \mathfrak{R}^m$, and $\Sigma \in \mathfrak{R}^{m \times m}$ is a symmetrical covariance matrix which represents the estimated variance of the data matrix $X \in \mathfrak{R}^{N \times m}$, where X , is a matrix of N observations at m variables. The transformation matrix $P \in \mathfrak{R}^{m \times m}$ is called the loading matrix of the covariance or correlation matrix of the original data samples. T is called the score matrix or the principal component matrix, which is the projection of original data in the sub-space of principal components.

$$T_{N \times m} = X_{N \times m} P_{m \times m} \quad (1)$$

$P = [p_1, p_2, \dots, p_m]$ is a set of vectors, orthonormal in $\mathfrak{R}^{m \times m}$. A PCA model is often established from the collected data in the form of information that can be measured on sensors or indicators of quality. In fact, the data matrix X is decomposed into the two following parts:

$$X = \hat{X} + E \quad (2)$$

Where $\hat{X}$, the estimated part of X and E presents the variations caused by modeling errors. The two components are orthogonal, one to the other, because they are in the complementary sub-spaces of $\mathfrak{R}^{N \times m}$.

$$\hat{X} E^T = 0 \quad (3)$$

Determining the structure of PCA model

In the structure of the PCA model, we first choose an adequate number of principal components to represent the process optimally (Valle et al., 1999). Dunia and Qin

(1998) proposed to choose the number of principal components based on best reconstruction of the process variables. An important characteristic of this approach shows that the index has a minimum, corresponding to the best reconstruction. The variance of reconstruction error for the i^{th} sensor is:

$$v_i = \frac{\tilde{\xi}_i^T \Sigma \tilde{\xi}_i}{\left(\tilde{\xi}_i^T \tilde{\xi}_i \right)^2}, \quad i = 1, \dots, m \quad (4)$$

Where $\tilde{\xi}_i = (I - \hat{C}) \xi_i$ and $\tilde{\xi}_i^T \tilde{\xi}_i = (1 - c_{ii})$. The VRE is defined to represent all the sensors:

$$VRE = \sum_{i=1}^m \frac{v_i}{\text{var}(x_i)} \quad (5)$$

Where $\text{var}(x_i) = \xi_i^T \Sigma \xi_i$ signifies the variance of the i^{th} element of the observation vector. $VRE(\ell)$ can be calculated recursively, using only the values and loadings of the covariance matrix Σ , until $\ell \leq m$; until the method of the variance of the reconstruction error chooses the number of principal components that gives a minimum VRE.

Once the number of components ℓ is determined, the PCA model is then identified, and the data matrix $X \in \mathfrak{R}^{N \times m}$ can be approximated from the first principal components.

$$\hat{X} = \sum_{i=1}^{\ell} T_i p_i^T = \sum_{i=1}^{\ell} X p_i p_i^T \quad (6)$$

Knowing that $\sum_{i=1}^{\ell} p_i p_i^T = \hat{P} \hat{P}^T$ and that we will note it as $\hat{C}$, the estimation of X is formulated as follows:

$$\hat{X} = X \hat{C} \quad (7)$$

Proposed monitoring approach

In Figure 1, we show the steps of our monitoring approach on the measurements of quality parameters in a factory of flour manufacturing. These measurements contain at the same time, the normal measurements and illogical others, which show the non quality of the product. These abnormal or defected data can always lead to controlling the production quality, which not only causes a great energy loss in the form of wastes, but also affects the quality of flour. To manufacture flour, it can influence the ordinary quality of flour and dough.

Our proposed monitoring approach is composed of four steps: The pre-analysis, the detection of defects and the prediction of the number of classes, the classification and visualization of classes and the location of defected parameters.

The pre-analysis

During this phase, the measurements of quality parameters of flour are treated in a way to be significant and valid. This pre-analysis phase consists in centering and reducing the data.

The detection of defects and the prediction of the number of classes

After formulating the PCA model, the defect detection is realized by the statistics SPE (Square Prediction Error), T^2 Hotelling and the combined statistic ϕ . According to the evolution of detection statistics, their signals are divided up into classes with and without defects.

Classification and visualization of classes

According to the evolution of its detection statistics, we distinguish regions without defects and others with defects that go beyond classification of these regions and the visualization of these classes in a 2D space. The classification and visualization phase is introduced, using the clustering algorithms. A comparative study of classification is proposed using the FCM, GK and GG algorithms and the PCA. In our application, the classification is applied on the detection statistics SPE and T^2 .

The location of defected parameters

The location of defected quality parameters is realized by the principle of defect direction in PLS between classes.

In our application we have used the steps of the algorithm for the offline monitoring while respecting the following instructions:

1. Get the data that represent the process at its state of normal functioning.
2. Center and reduce the data.
3. Get the PCA model, by determining the number of principal components.
4. Determine the control limits for the statistics SPE, T^2 and ϕ .
5. Classify the normal and abnormal regions using the fuzzy clustering tools.
6. Use the defect directions two by two in PLS to isolate variables in defect.

7. Use the contributions of weight and their directions to isolate the parameters in defects.

Pre-analysis

Beforehand, an essential pre-treatment consists in centering and reducing the variables to obtain an independent result of the units used for each variable.

Detection of defects

The detection of defects using the PCA models is normally accomplished based on the statistics SPE, T^2 and ϕ . The SPE statistic, known as the Q statistic measures the projection of an x vector in the residual space.

$$SPE(k) = e^T(k)e(k) = \sum_{i=1}^N (x_i(k) - \hat{x}_i(k))^2 \quad (8)$$

Where $e = (I - P_t P_t^T)x$ is the residue of the x vector which represents the distance squared at each observation, perpendicular to the sub-space of principal components, and measures the residues that cannot be represented by the PCA model. The process is considered wrong at the instant k if;

$$SPE(k) < \delta_\alpha^2 \quad (9)$$

Where δ_α is the trust threshold of the SPE. The approximation of the detection limit for the SPE statistic with an α covariance level is then represented by:

$$\delta_\alpha = \theta_1 \left[\frac{1 - \theta_2 h_0 (1 - h_0)}{\theta_1^2} + z_\alpha \frac{(2\theta_2 h_0^2)^{\frac{1}{2}}}{\theta_1} \right]^{\frac{1}{h_0}} \quad (10)$$

$$\text{Where } \theta_1 = \sum_{i=1}^m \lambda_i, \theta_2 = \sum_{i=1}^m \lambda_i^2, \theta_3 = \sum_{i=1}^m \lambda_i^3, h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2}$$

The statistic T^2 , which measures the variation of the score vector in the space of principal components, is expressed by:

$$T^2 = t^T A^{-1} t = x^T P A^{-1} P^T x = x^T D x \quad (11)$$

Where A , is a diagonal matrix containing the principal components used in the PCA model, and $D = P A^{-1} P^T$ is a semi positive matrix. The detection limit is obtained,

using the distribution of Fisher (Vermasvuori, 2008) and will depend on the freedom degrees available for the estimation of Σ .

Case 1: Known covariance matrix

This statistics is used as a big-sized sample whose distribution follows the rule of chi-square centered with ℓ degrees of freedom. A limit threshold in a α trust interval is expressed by $\chi_{\alpha,n}^2$.

Case 2: Unknown covariance matrix

When the covariance matrix Σ is unknown, it should be estimated using the T^2 statistics given by:

$$T^2 = (x - \mu) S^{-1} (x - \mu)^T \quad (12)$$

Where S is the estimation of the covariance matrix Σ . The exact detection limits are given at the α trust level:

$$\begin{aligned} HCL &= \frac{m(N^2 - 1)}{N(N - P)} F_{1-\alpha}(m, N - m) \\ LCL &= \frac{m(N^2 - 1)}{N(N - P)} F_{\alpha}(m, N - m) \end{aligned} \quad (13)$$

Where HCL, high detection limit at the trust level $(1 - \alpha)$, and LCL, low detection limit at the trust level α .

The statistics SPE and T^2 are complementary one to the other and can measure the variation in the whole measurement space. The combined statistic proposed by Yue and Qin (2001) introduces the two metrics SPE and T^2 together. This combined index is defined as an addition of the statistic SPE and T^2 , balanced against their threshold limit.

$$\varphi = \frac{SPE(x)}{\delta^2} + \frac{T^2}{\tau^2} = x^T \Phi x \quad (14)$$

Where $\Phi = \frac{I - PP^T}{\delta^2} + \frac{PA^{-1}P^T}{\chi_{\alpha}^2(\ell)}$. The detection threshold limit

for the combined index is given in the following equation:

$$\varphi \geq g \chi_{\alpha}^2(\ell) = \zeta^2 \quad (15)$$

Classification and visualization with clustering algorithms

In today's flour production industry, the massive quantity of data is easily available; whereas, visualizing

highly-dimension data is difficult. However, the radial plots (Nottingham et al., 2001) and parallel coordinates (Albazzaz et al., 2005) methods are usually applicable to small sets of data; because of this limitation, these methods cannot be applied to the visualization of flour quality parameters due to the big number of samples and variables. Generally, if the highly-dimension data cannot be represented in a space of 2 or 3 dimensions, the use of clustering algorithms provides a general frame for visualizing and exploring big sets of multi-varied data.

The clustering is a technique of a non supervised study; its objective is to group the points of similar data. A clustering algorithm assigns a great number of data points to a smaller number of groups such as the data points in the same group sharing the same properties, whereas in the different groups, they are dissimilar. The clustering technique has many applications, including the image segmentation, the knowledge extraction, the form recognition and the classification (Liang et al., 2005; Hung et al., 2006; Luukka, 2009).

Fuzzy C-means classification (FCM)

The FCM calculates the distance measured between two vectors, where each component is a trajectory instead of a real number. Thus, this FCM algorithm is rather flexible than being able to eliminate the classes and their combination. The FCM is a classification technique, which introduces the fuzzy set notion in the presentation of classes. In the traditional FCM algorithm, an object is assigned only to one group (Rezaee et al., 1998; Dulyakran and Ransanseri, 2001). This is valid before that the groups are split or separated. But if the groups are nearing one another or are overlapping, then, an object can belong to more than one group. In this case, the FCM technique is the best to be used. Particularly, the version proposed by Bezdek is the best applied one. It is based on minimizing the following objective function (Bezdek, 1981; Höpper et al., 1999; Pedrycs, 1997).

$$J_m = \sum_{i=1}^N \sum_{j=1}^C (u_{ij})^m d^2 \|x_i - c_j\|^2, 1 \leq m \leq \infty \quad (16)$$

Where m , is a real number higher than 1; μ_{ij} is the membership degree of x_i in the j cluster; x_i is the i^{th} data of measurements; c_j is the cluster center, and $\|\cdot\|$ is the norm between a measured datum and the center.

The fuzzy division is affected by an iterative optimization of the J_m function, by updating the μ_{ij} membership and the cluster center c_j :

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{1}{m-1}}} \quad (17)$$

Where $c_j = \frac{\sum_{i=1}^N u_{ij}^m x_i}{\sum_{i=1}^N u_{ij}^m}$. The k iterations stop when

$$\max_{ij} |u_{ij}^{k+1} - u_{ij}^k| < \varepsilon$$

Where ε is a benchmark between 0 and 1.

Fuzzy C-means model steps

1. Randomly select the j number of the center (c_j).
2. Determine the fuzzy membership μ_{ij} starting from equation 17.
3. Compute the fuzzy center c_j .
4. Repeat steps 2 and 3 until the minimum of J_m is achieved.

The gustafson-kessel algorithm

The Gustafson-Kessel algorithm is an extension of the FCM. This technique uses an adaptative distance norm which detects the clusters of different geometric forms in a data set (Graves and Pedrycz, 2007; Krishnapuram and Kim, 1999). Each cluster has its own norm expressed as follows:

$$D_{ikA}^2 = (x_k - v_i)^T A_i (x_k - v_i), \quad 1 \leq i \leq c, \quad 1 \leq k \leq N \quad (18)$$

The objective function of the GK algorithm is the following:

$$J(X; U, V, A) = \sum_{i=1}^c \sum_{k=1}^N (\mu_{ik})^m D_{ikA_i}^2 \quad (19)$$

The Gath-Geva algorithm

The Gath-Geva algorithm is also known as the mixed gaussian decomposition. It is an algorithm similar to the FCM, where the Gaussian distance is used instead of the Euclidean distance. The clusters do not have any definite form and can have various sizes. The Gaussian distance is expressed as follows:

$$d_{ie} = \frac{1}{P_i} \sqrt{\det(A_i)} \exp\left(\frac{1}{2} (x_e - v_i)^T A_i^{-1} (x_e - v_i)\right) \quad (20)$$

The different steps of the Gath-Geva are detailed in many works (Gath et al., 1989; Park et al., 2004).

Locating defects

We apply fault direction in the PLS to normal data and to each class of defect data to find a defect direction, which optimally moves each defect of data apart from normal data. We use weights in defect directions to generate contribution plots for defect diagnosis. The PLS-DA is a PLS-based model, where the general model form can be as follows:

$$Xb = y \quad (21)$$

Where y is a column vector of observations of a dependent variable and $X \in \mathbb{R}^{N \times m}$ is a matrix which results from N observations of the m variables. The column vector b contains the m regression coefficients. To solve this equation, we have suggested (Vance, 1996):

$$X^T Xb = X^T y \quad (22)$$

We have the vector y whose components take the classes $c_1 \dots c_p$. In our study, the vector y consists of p classes.

$$y = [c_1 \dots c_1, c_2 \dots c_2, c_3 \dots c_3, \dots, c_p \dots c_p]^T \quad (23)$$

So at the first step, to bring to fruition the algorithm of least partial squares, the standardization of the vector y consists in having a nil average $\alpha_1 c_1 + \alpha_2 c_2 + \dots + \alpha_p c_p = 0$ where, α_p is the number of occurrence of c_p . y is an n -dimensional vector whose components take on the two considered classes c_1 and c_2 , which are the class of normal data and each class of defect data, respectively. Then, to satisfy the normalization of the first step of the PLS algorithm, we should rescale y to have a zero mean

$\alpha_1 c_1 + \alpha_2 c_2 = 0$, where α_i is the number of occurrence of c_i . We arbitrarily choose $c_1 = 1$, which implies

$c_2 = -\frac{\alpha_1}{\alpha_2}$. Accordingly, the right hand side of the equation (21) is reduced to:

Figure 2. Locating defects for p classes.

$$X^T y = \sum_{j=1}^{n_1} a_{kj} - \frac{\alpha_1}{\alpha_2} \sum_{j=n_1+1}^n a_{kj} \quad (24)$$

Where a_{kj} is an element of X and $1 \leq k \leq m$. This can be written as follows:

$$X^T y = \alpha_1 (\mu_1 - \mu_2) \quad (25)$$

Where μ_1 is the m -dimensional mean whose corresponding variable has a c_1 class. As a result, $X^T y$ is in the same direction as the line connecting the means of the two classes of y . Then, the final solution vector b obtained by the PLS is given by:

$$X^T X b = \alpha_1 (\mu_1 - \mu_2) \quad (26)$$

$$b = \Sigma^{-1} (\mu_1 - \mu_2) \quad (27)$$

Where the covariance matrix is $\Sigma = \frac{1}{n} X^T X$.

Therefore, we have a PLS direction, where $\phi_i = (\mu_1 - \mu_2)$, which is the needed one. So, we define this ϕ_i direction as the defect direction for class c_i . This weight in ϕ_i is used to generate the contribution plot for the defected class.

$$\phi_i = [\phi_1, \phi_2, \dots, \phi_j, \dots, \phi_m] \quad (28)$$

For the fault direction, the j^{th} element ϕ_j is the contribution from the j^{th} variable. For p classes, we determine two to two the class fault direction. All the abnormal data are used by the PLS model to determine the directions of defects (Figure 2). The weights of these defect directions are calculated within the regions of normal and abnormal data two by two.

Application

Process description

Grinding wheat into flour or semolina is done in a progressive industrial mill. Wheat flour is worked out from wheat grains; we partially eliminate the bran and the germ and we crush the rest into enough fine powder. In Figure 3 we present the different steps of a line of grinding wheat into standard flour.

Water level (H): The flour water level is an important parameter that must be between 10 and 16% in such a way that we can properly preserve the flour. The moisture meter measurement ensures a rapid determination of the level of water. The moisture meters are used on the sites of reception. The measurement of an electric characteristic of the grains, which are variable in function of their moisture state, is linked after calibration to the water level of the grains.

Ash content (C): The ash rate is an official means used

Figure 3. Line of grinding wheat.

Table 1. Signification of quality parameters of the chopin alveograph.

Symbols	Signification
P	It is called pressure, and it measures the tenacity and firmness of the dough and its resistance to deformation
G	The rising tallies with the air quantity instilled into the dough, till its bursting.
L	It is the width of the graphic design, and it shows the curve extensibility and indicates the dough elasticity and the processing extension
W	This value measures the necessary work to deform the dough roll till bursting; we use also the term "baking power" of the flour
$P/L = r$	It is the link that shows the balance and unsteadiness between the dough's tenacity and extensibility
le	Rate of elasticity: $le (\%) = (P_{200}/P) \times 100$

for characterizing the flour purity. The ash determination makes it easy to know the global mineral material content of wheat and its derivatives.

Protein content (Pr): Having a good idea about the content of protein, combined with that of wheat variety, provides significant information about the technological flour capacity. We have picked out relatively high protein contents, which vary between 11.45 and 17%. The determination of protein content by infrared spectrometry is also a known method.

In addition, we distinguish six other indicators of flour quality by a method of indirect measurement using a device called the alveograph of CHOPIN, with which we measure the dough's resistance and elasticity. That is why we will make the dough to fall down completely. These different measurements permit us to get the flour baking power. The principle measurement is based on raising a dough sample subjected to air pressure. The formed bubble volume is determined in relation to the extensibility of the dough. The bubble pressure evolution in function of time is measured and carried forward in the

form of a curve called alveogram. In Table 1, we present the signification of quality parameters of the CHOPIN alveograph.

Pre-analysis

We set out the historical measurements of nine variables, representing the quality parameters of flour. The data base is used in our application form of historical measurements of quality parameters in a flour manufacturing line. The measurement (528 × 9) is the data test composed of 528 observations of 9 flour quality parameters. First, we should start by creating normalized data. In Figure 4, we present the reduced centered variables.

Among the principle components, we should not keep only the one carrying significant data, allowing us to explain the different variables, to estimate those which are not original. Then, we determine the structure of the PCA model; that is, we determine the number of components to keep or to retain in the PCA model.

Figure 4. Evolution of reduced centered variables.

Table 2. Variance of reconstruction error with $m = 8$ and $\min_{\ell} \sum_{i=1}^m \rho_i(\ell) = 3$

i	ℓ							
	1	2	3	4	5	6	7	8
1	0.25	0.98	0.56	0.79	0.91	0.54	0.96	0.95
2	0.18	0.83	0.38	0.40	0.82	0.54	0.23	0.24
3	0.17	0.11	0.06	0.12	0.75	0.47	0.14	0.13
4	0.66	0.06	0.11	0.05	0.06	0.01	0.01	0.11
5	0.92	0.04	0.04	0.05	0.05	0.01	0.01	0.05
6	0.54	0.02	0.02	0.02	0.02	0.01	0.01	0.01
7	0.97	0.01	0.01	0.01	0.01	0.01	0.01	0.01
$\sum_{i=1}^m v_i(\ell)$	3.69	2.05	1.18	1.44	2.62	1.59	1.37	1.5

Determining the number of principal components

To determine the number of principal components, we use the method of the method of the variance of reconstruction error. The number of components which minimizes $VRE(\ell)$ is $\ell = 3$. We have three principle components in the PCA model (Table 2).

Once the component number ℓ is determined, the PCA model is then identified. In Figures 5 and 6, we present the measurement development of the variables P and L , as well as their estimation determined by the PCA model

Detection of defects and prediction of classes

In Figures 7, 8 and 9, we present the evolution of the statistics SPE, T^2 and ϕ . Thus, we notice that the derivatives are only detected by the statistics SPE and

T^2 . The detection by the statistic ϕ is not effective, so this statistic will not be used in the classification phase. From the development from the statistics SPE and T^2 , we distinguish 5 regions, where B and D are two regions in defect and A, C and E make up three regions without defects.

Classification and visualization of regions

We present, in this section, the classification and visualization of quality parameters of flour by the clustering algorithms. The distances determined by the FCM, GK and GG algorithms are projected into a 2D space.

The validity indexes of clusters used in our application are the partition coefficient PC and the partition entropy

Figure 5. Measurements and estimation of pressure P .

Figure 6. Measurements and estimation of width L .

Figure 7. The evolution of the statistic SPE.

Figure 8. The evolution of the statistic T^2 .

Figure 9. The evolution of the statistic φ .

Table 3. Validation Parameters of the FCM, GK and GG clustering.

	PC	EC	PI	SI	XB
FCM	0.6851	0.4137	0.2568	0.0004	3.1271
GK	0.7820	0.5811	0.1990	0.0007	6.2296
GG	0.9228	NaN	NaN	NaN	NaN

EC. They are sensitive to noise and to the variation of the exponent m . The optimal cluster number is pointed out by the minimum *PC* value. The *EC* index measures the cluster fuzziness, which is similar to a division coefficient. The optimal cluster number is pointed out by the maximum *EC* value.

The other indexes *PI* and *XB* are proposed, respectively by Fukayama, Sugeno and Xie-Beni (Xie et al., 1991; Fukuyama et al., 1989). *PI* is sensitive to high and low values of m , and *XB* gives good responses onto a great choice for $c = 2, \dots, 10$ and $1 < m \leq 7$. The *XB* index measures how many clusters are compact; and *PI* is

a further term that measures how many clusters are separated. The distance-based separation index *SI* is the most used function that reduces the total distance in the cluster variation to the minimum. To determine these validation parameters of these algorithms, we have used the Fuzzy Clustering toolbox (Balasko et al., 2005).

Classification of SPE and T^2 signals

The classification is applied to both SPE and T^2 signals, using the clustering algorithms. In our application of the FCM, GK and GG algorithms, we have fixed the number of clusters $c = 5$ and $1 \leq m \leq 7$. The validation parameters of the FCM, GK and GG clustering algorithms are gathered in the Table 3.

During the application of clustering algorithms, both FCM and GK algorithms ensure the classification, whereas the GG algorithm diverges.

When applying the FCM algorithm, the *PC* index is minimum and the *EC* index is maximum, compared with the GK algorithm. So, the cluster optimum number is

Figure 10. FCM and Gk clustering of the data $m = 2$ and $c = 5$, and the recalculated membership contours.

Figure 11. PLS defect direction contribution plot on the classes of SPE and T^2 , (a, b) PCA contribution plot.

indicated by the minimum PC value and the maximum EC value. In our application, for the FCM and GK algorithm, we have a minimum PI coefficient that indicates a better division of clusters. So, the SI separation index guarantees a minimum separation distance between clusters. The XB index indicates how many clusters are compact; we have almost 3 groups, showing the three regions without defects. On the other hand, for the GK algorithm, we have almost 6 compact groups, which show that the classification by the GK algorithm does not ensure a separation between the regions with and without defects.

In the Figure 10, the blue dots are the measurements, the 'o's in red indicate the center of the cluster or class.

The contour-map circles points out the limits of each cluster. The contour-map of Figure 10a of the FCM classification presents three compact classes, which are the classes with defects and two other separated classes. The classification Figure 10b by the GK algorithm shows an unclear classification; we cannot distinguish between the classes.

Based on expertise, we notice that the two classes D and B are the two regions in defect. The groups A, C and E are more compact clusters than the two other groups D and B, which make up the A, C and E regions without defects.

The classification of the detection signals SPE and T^2 , using the clustering algorithms, is efficient for the

Figure 12. Simulated measurements of flour quality parameters.

separation between the data with and without defects.

RESULTS AND DISCUSSION

After classifying the process data in classes with and without defects, we put forward in Figure 11 the contribution plot based on directing PLS defects and PCA for all the used measurement. Figure 11c and d show the contribution plot of the defects B and D. The PCA based contribution plot draw in Figure 11a and b.

Defect B

In Figure 11c, the PLS defect direction contribution of variables has shown that parameter H (water level: moisture) and C (ash content) have the highest contribution, compared with other variables. The development of variables in Figure 12 shows that the parameters G (inflating), H and C are the cause of this B defect. However, the contribution calculation is based on the classical PCA in Figure 11a; only the variable G is identified, but other C and H parameters are not identified.

Defect D

In Figure 11d, the PLS defect direction contribution of variables has shown that the variable L (elasticity), r (ratio

of tenacity and stretchability of the dough) G and C have the highest contribution, compared with other variables. Thus, these parameters are considered in defect.

The PCA-based contribution calculation, in Figure 11b, shows the variables G and le (elasticity rate) which have the highest contribution. Therefore, these two parameters are considered in defect. The development of parameters in Figure 12 shows that the variables L, r, G, H and C are the cause of the defect D. Then, the location of defects by the PLS defect direction technique is much more efficient than the PCA contribution calculation. So, the location of parameters in defect on the SPE and T^2 signals is validated, and the location results are efficient on the detection signals SPE and T^2 .

The results of this analysis show that combining the PCA and the fuzzy clustering techniques is useful in extracting the abnormal data out of the set of flour quality parameters starting from these multidimensional, and in simplifying the data interpretation to detect and isolate defects. A comparative study of our monitoring approach and the isolation by PCA contribution calculations is well presented (Figure 11). Extracting abnormal regions, thanks to the combination of the PCA and the fuzzy clustering technique, has proved its contribution in isolating defects in PLS.

Our approach of isolating abnormal data from the set of historical measurements is efficient in relation to the isolation by linear PCA contribution. The isolation is based on the classification of regions by fuzzy clustering, and the PLS defect direction has a positive impact in determining parameters in defect.

Conclusion

In this study, the PCA defect detection and the clustering technique, for the monitoring of quality parameters, are applied and validated to the process of producing flour. The proposed monitoring approach implemented in our process is based on combining the PCA and the clustering algorithms.

In our application, we suggest data-processing to extract the detection signals SPE and T^2 , in order to make a classification depending on their development. Then, we suggest a comparative study between the classification by the clustering algorithms FCM, GG and GK, and the classification by the PCA. The development of the results of locating defected parameters, using the classification on the SPE and T^2 signals, is validated. The location technique of defect direction in PLS has shown its performance, compared with the location, using the PCA for determining defected parameters.

The method has been assessed with historical measurements of flour quality and with the possibility of carrying out this combined approach of monitoring in an online automatization system.

REFERENCES

- Albazzaz H, Wang XZ, Marhoon F (2005). Multidimensional visualization for process historical data analysis: a comparative study with multivariate statistical process control. *J. Proc. Cont.* 15:285-294.
- Ahvenlampi T, Tervaskanto M, Kortela U (2005). Diagnosis System for Continuous Cooking Process. 16th IFAC World Congress. 16:1-6.
- Ahvenlampi T, Kortela U (2005). Clustering Algorithms in Process Monitoring and: Control Application to Continuous Digesters, *Informatica*. 29:101-109.
- Bezdek JC (1981). Pattern recognition with fuzzy objective function algorithms. Plenum, New York.
- Bouhouche S, Lahreche M, Moussaoui A, Bast J (2007). Quality Monitoring Using Principal Component Analysis and Fuzzy Logic Application in Continuous Casting Process. *Am. J. A Sci.* 4(9):637-644.
- Balasko B, Abonyi J, Feil B (2005). Fuzzy Clustering and Data Analysis: Matlab Toolbox. pp. 75-79.
- Chaouch H, Ouni K, Nabli L (2011). ECG signal monitoring using linear PCA. *Int. J. Comput. Appl.* 33(8):48-54.
- Dulyakran P, Ransanseri Y (2001). Fuzzy c-means clustering using spatial information with application to remote sensing. 22nd Asian Conference on Remote Sensing. <http://www.actapress.com/Abstract.aspx?paperId=31118>.
- Dunia R, Qin SJ (1998). A subspace approach to multidimensional fault identification and reconstruction. *AIChE J.* 44(13):1813-1831.
- Fukuyama Y, Sugeno M (1989). A new method of choosing the number of clusters for the fuzzy c-means method. *Proc. Fifth Fuzzy Syst. Symp.* pp. 247-250.
- Gath I, Geva AB (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 11(7):773-781.
- Graves D, Pedrycz W (2007). Fuzzy C-Means, Gustafson-Kessel FCM, and Kernel-Based FCM: A Comparative Study. *Adv. Soft Comput.* 41:140-149.
- Harket MF, Mourot G, Ragot J (2006). An improved PCA scheme for sensor FDI: Application to an air quality monitoring network. *J. Proc. Control* 16(6):625-634.
- Hung WL, Yang MS, Chen DH (2006). Parameter selection for suppressed fuzzy c-means with an application to MRI segmentation. *Pattern Recogn. Lett.* 27(5):424-438.
- Höpper F, Klawonn F, Kruse R, Renklu T (1999). Fuzzy cluster analysis for classification data analysis and image recognition. John Wiley and Son. pp. 35-37.
- Jolliffe I (1986). Principal Component Analysis. Springer Verlag.
- Krishnapuram R, Kim J (1999). A note on the Gustafson-Kessel and adaptive fuzzy clustering algorithms. *IEEE Trans. Fuzzy Syst.* 7(4):453-546.
- Kresta JF, MacGregor JF, Marlin TE (1991). Multivariate statistical monitoring of process operating performance. *Can. J. Chem. Eng.* 69(1):35-47.
- Liang GS, Chou TY, Han TC (2005). Cluster analysis based on fuzzy equivalence relation. *Eur. J. Oper. Res.* 166(1):160-171.
- Luukka P (2009). PCA for fuzzy data and similarity classifier in building recognition system for post-operative patient data. *Expert Syst. Appl.* 36(2):1222-1228.
- Nottingham QJ, Cook DF, Zobel CW (2001). Visualization of multivariate data with radial plots using SAS. *Comput. Ind. Eng.* 41:17-35.
- Ouni K, Nabli L, Messaoud H (2011). A Monitoring Method Based on Fuzzy Detection and PCA Diagnosis. *Int. J. Intell. Control Syst.* 16(1):19-27.
- Ouni K, Dhoubi H, Nabli L, Hassani M, Simeu-Abazi Z (2011). Monitoring Approach Using Nonlinear Principal Component Analysis, *CCCA*. 1:1-6.
- Oliveira JV, Pedrycz W (2007). Advances in Fuzzy Clustering and its Applications. John Wiley and Sons. DOI:10.1002/9780470061190.
- Pal NR, Pal K, Keller JM, Bezdek JC (2005). A Possibilistic C-Means Clustering algorithm. *IEEE Trans. Fuzzy Syst.* 13(4):517-530.
- Peter Q, He S, Qin J, Wang J (2005). A New Fault Diagnosis Method Using Fault Directions in Fisher Discriminant Analysis. *AIChE J.* 51(2):555-571.
- Park HS, Yoo SH, Cho SB (2004). A Fuzzy Clustering Algorithm for Analysis of Gene Expression Profiles. *Trends Artif. Intell.* 3157:967-968.
- Pedrycz W (1997). Fuzzy clustering with partial supervision. *IEEE Transactions on Systems, Man and cybernetics, Part B. Cybernetics*. 27(5):787-795.
- Rezaee MR, Lelieveldt BPF, Reiber JHC (1998). A new cluster validity index for the fuzzy c-mean. *Pattern Recogn. Lett.* 19(3-4):237-246.
- Sebzalli YM, Wang XZ (2001). Knowledge discovery from process operational data using PCA and fuzzy clustering. *Eng. Appl. Artif. Intell.* 14:607-616.
- Vance F (1996). Partial Least Squares, Conjugate Gradient And The Fisher Discriminant. Los Alamos National Laboratory, Los Alamos New Mexico. 3958:1-17.
- Valle S, Li W, Qin SJ (1999). Selection of the number of principal components: The variance of the reconstruction error criterion with a comparison to other methods. *Am. Chem. Soc.* 38(11):4389-4401.
- Vermasvuori M (2008). Methodology for utilising prior knowledge in constructing data-based process monitoring systems with an application to a dearomatisation process. Ph. D, Helsinki University of Technology.
- Xie L, Beni G (1991). A validity measure for fuzzy clustering, *IEEE Transaction. Pattern Anal. Mach. Intell.* 13(8):841-847.
- Yue H, Qin SJ (2001). Reconstruction-based fault identification using a combined index. *Ind. Engin. Chem. Res.* 40(20):4403-4414.
- Yingwei Z, Yang Z (2012). Process monitoring, fault diagnosis and quality prediction methods based on the multivariate statistical techniques. *IETE J. Res.* 27(5):406-420.