UNIVERSITAT DE GIRONA
GIRONA INTERNATIONAL GRADUATE SCHOOL
Department of Electrical, Electronics and Automatics Engineering

INDUSTRIAL COMPUTER SCIENCE AND AUTOMATICS
MASTER THESIS

Case-Based Diagnosis in the Principal Components Space. From definition to applications

Submitted by
Xavier Berjaga Moliné

Supervisor:
Dr. Joaquim Meléndez

Dedicated to

To my beloved parents. Because they are always there to put me in the good direction. Not to mention that with their tips and education have made of me the person I am now. I hope someday I can give back a little part of all they have given to me.

To my family, that is far in distance from me, but very close in presence. Because it is always a pleasure to visit them on holidays, and laugh an talk... You are always with me and will never be forgotten.

To my godson, the most precious thing in my life and the most important thing to take care of. You always makes me have a good time with you, and playing with you is always a new adventure.

To Paco, a friend from Lloret that was there in the bad moments and also to guide me and teach me invaluable things.

Acknowledgments

First of all, I would thank you Joaquim Meléndez, my supervisor in this Master Thesis for his ideas and support, and because sometimes it can be difficult to get along with me.

I also would greet the financial support of the Generalitat de Catalunya in the projects I have been involved (DPI2005-08922, DPI2006-09370 and COLL-CT-2006-030339).

To all friends from P-IV for those good times in the lab and to outside.

To Magda, for her help in this work, as well to listen my questions and be patient enough with me.

To Alberto Ferrer, from the Universidad Politécnica de Valencia for his help and guidance in the MPCA methodology.

To Sergio Herraiz and Victor Barrera for their ideas and collaboration in the electrical field, where their help was very valuable.

To Encarna Escudero and Francesco Puliga from ASCAMM Foundation for their help and collaboration to understand the injection moulding process.

I thank the LEQUIA research group for providing the data for the Wastewater Treatment scenario as also their knowledge and advise.

I thank TNO for providing the data and information for the moulding injection field.

I thank ENDESA Distribución for providing the information and data for the Power Quality Monitoring as well as their invaluable help in this scope.

Contents

List of Figures

Chapter 1

Introduction

In this introductory chapter the motivations of this Master Thesis are explained, followed by three application domains that have been used as experimental domains. Next a brief explanation of the proposed methodology that will be applied is done, and finally, the outline of this Master Thesis is presented.

1.1 Problem statement

With the increase of the dependency degree of modern society in systems (vehicles, planes, trains, etc.) and complex technological systems (distribution networks and energy production, water, etc.), their availability and correct performance have become strategical points. Their wrong operation can cause financial losses, danger situations for the operators, users inconveniences, among others [1]. Because all of that, the control of those process is one of the most important tasks nowadays.

A process is classified as out of control whenever a fault appears. A fault is whenever a non-allowed deviation of part of the system appears, what causes the system to not accomplish the original function it was originally designed for [1]. Fault detection and diagnosis (usually addressed as Fault Detection and Isolation (FDI)) is strongly dependant of the a priori knowledge available [2]. A priori knowledge needed for fault diagnosis consists of a description of Normal Operation Conditions (NOC) and additionally information related to abnormal operation conditions. Depending on available knowledge to describe NOC, the fault diagnosis techniques can be divided in two groups: model-based methodologies [3] and process history, also known as data driven methods or model free methods. At the same time, the model-based approach can be divided into qualitative and quantitative. Qualitative models are high level models that describe the influence among variables (casualty). For example, they can describe functional and / or structural properties of the systems. On the other hand, the quantitative models rely on mathematical relationships, typically mathemati-

cal equations describing the process from first principles. Ordinary Differential Equations (ODE) and Algebraic relationships are typically used to describe physical behaviours and mass or energy blains. Figure 1.1 represents this classification, and different approaches are included. For a brief introduction to each approach and further references refer to [2], [4] and [5].

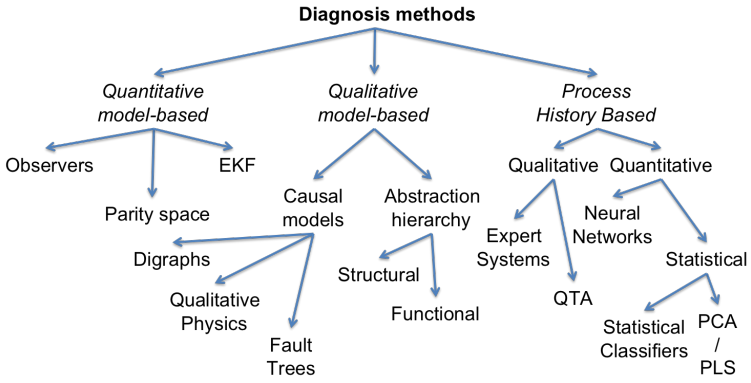


Figure 1.1: Classification of diagnosis methodologies according to a priori knowledge strategy

Statistical techniques have been used for process monitoring from the earlies 20's. A set of tools known as Statistical Process Control (SPC) has been developed for this purpose. This methodology was firstly introduced by Walter Shewhart in the 1920s and was based on the usage of *SPC control charts* to adapt the management processes [6]. This would allow the creation of profitable situations for both consumers and producers. As time passed, the usage of the SPC was more than only the application of control charts and eventually became used in the manufacturing process. This evolution was also reflected in the change of the original idea of basing the control limits in economic limits to use the process history to compute statistical control limits based on the probability of group variations.

This Master Thesis will be centred in the history data based approach for monitoring finite duration processes. More exactly, in the usage of a Principal Component Analysis (PCA) variation for model creation and fault detection to take into account the correlation among variables at different time instants [7]. After that, a Case-Based Reasoning approach will be used for fault diagnosis based on the new variables obtained from the previous model.

1.2 Motivation and application fields

The motivation of this Master Thesis comes from one of the research lines of the eXiT group: statistical process control of industrial processes using multivariate techniques. The methodology this Master Thesis will expose consists in basically two steps. Firstly a PCA model is created to capture and model the data structure constrained by the Normal Operation of the process. Then, the properties of this model are exploited to define a case based diagnosis strategy in the projection space. This methodology is explained in detail in Section 1.3.

There exist two antecedents of this methodology: a doctoral thesis ([8]) and a master thesis ([9]) directed in the same research group, eXiT of the UdG. In this work a new implementation of the method has been done and several improvements on the neighbourhood in the projection space is proposed as retrieval mechanism. Three application domains (wastewater treatment plants, power quality monitoring and monitoring of injection moulds) have been selected to test and validate the methodology and study the behaviour of PCA models and similarity criteria. Those application fields are included in three different projects: DPI2005-08922 includes the wastewater field, DPI2006-09370 includes the power quality monitoring field and COLL-CT-2006-030339 defines the moulding process field. Now, a brief introduction to each of the application fields is presented.

1.2.1 Wastewater Treatment Plants (WWTP)

All communities produce solid and liquid wastes daily. When liquid wastes come after of residential, industrial or commercial usage is called wastewater. The accumulation and stagnation of it can cause bad-smelling gases, as well as human harmful microorganisms. Another important point is that wastewater contains nutrients that accelerate the growth of plants with toxic compounds. In order to avoid these situations, wastewater treatment plants have become one of the most important environmental topics. Moreover, due to the sparsely distribution of rains nowadays all procedures that help the reduction and reuse of water are key topics. To harmonise urban wastewater treatment, the European Union (EU) has approved a more protective legislation with the environment. This legislation requires the introduction of new technologies to control and supervise the treatment phase to intervene before any problem occurs.

In the past, control of these processes was delivered to some human operators. But the increase in the signals to be controlled and the huge amount of information that the operators receive at every instant has led to a necessity for an automatic control procedure. The high complexity of biological processes and relations is one of the most important handicaps to overcome. Because of that, process historical data based techniques are increasingly being used in the model creation step of biological processes. For instance, the usage of PCA in WWTP was firstly introduced in [10], where the applicability of statistical procedures

for detection of process disturbances was demonstrated with a comparison between PCA and Partial Least Squares (PLS). Using PCA as a previous step or as a part with other techniques is not a new idea. In [11], PCA was combined with a Credibilistic Fuzzy-C-Mean (CFCM) and Takagi-Sugeno-Kang (TSK) fuzzy model to predict the important variables output in a full-scale WWTP. Also, in [12], a Multi-Layer Neural Network, k-Means clustering and PCA were integrated to estimate process quality and efficiency in Saint Cyprien WWTP (France). And as a last example, in [13] a Cluster Analysis (CA), Discriminant Analysis (DA), PCA and PLS are presented to study wastewater composition.

In this scenario, the methodology explained in Section 1.3 will be used in a first step for Normal Operation Conditions (NOC) determination according to historical data, and later, for fault detection and diagnosis of known Abnormal Operation Conditions (AOC).

1.2.2 Power Quality Monitoring

The aim of power quality monitoring is to automatically evaluate disturbances registered by power monitors installed in power distribution substations. Utility companies have increased the number of power quality monitors installed in the distribution substations and are very interested in developing reliable methods to efficiently exploit the information contained in these registers. In this example domain, the relative location of disturbances known as voltage sags in order to determine its origin up or downstream of the measuring point is studied.

According to the IEEE Standard [14], a voltage sag is the reduction of the nominal voltage of one phase between 10 % to 90 % and with a duration time from 200 ms to 1 minute (Figure 1.2). This kind of disturbances is the most common in the actual electrical sector and utility companies are investing a great amount of money in order to locate its origin rapidly and effectively.

Determining whether sags have occurred in the distribution or transmission networks precedes the localisation and mitigation stages [15]. Typical classification according to the origin consists in discriminating between transmission (or high voltage) and distribution (or medium voltage) origins. For this purpose, phase analysis and an unsupervised method were compared in [16] by extracting some temporal descriptors from the RMS representation of sags and using a Learning Algorithm for Multivariate Data Analysis (LAMDA). Recent research has also identified similarities among sags using the variability in the information contained in the waveform in statistical analyses based on Principal Component Analysis (PCA), which allows dimensionality reduction before similarity criteria are applied to sags, assigning them to different classes. In [16] sags are categorised into three classes using certain features run through a fuzzy system. A more recent method for locating the origin of a voltage sags in a power distribution system using the polarity of the real current component relative to the monitoring point has been introduced in [15].

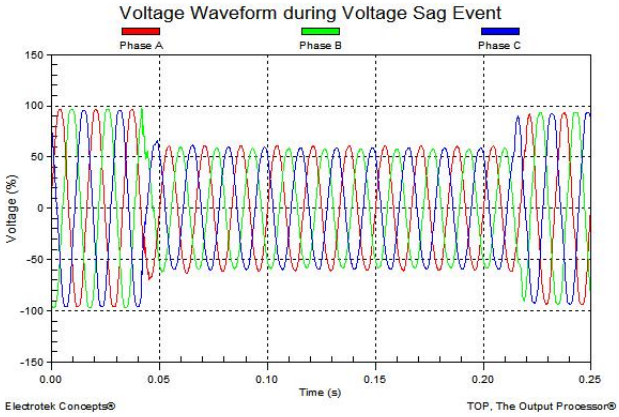


Figure 1.2: Example of voltage sag

The usage of statistical techniques in this field is a very rare approach, finding few references. For instance, a first proposal to work with the SPC methodology was presented in [17], where the PLS methodology was combined with a Neural Network for the relative allocation of voltage sags. And finally, in [18] and [19] the model obtained from the multivariate statistics is used directly for classification.

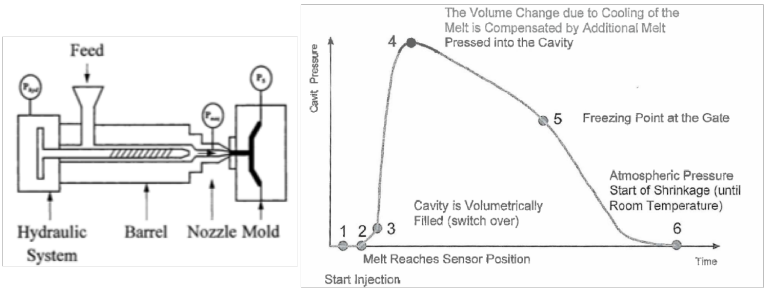
In this case, the methodology will be used as a classification tool to relatively locate the origin of voltage sag using directly the voltage and current values, avoiding the definition of data-derived descriptors and additional computations.

1.2.3 Plastic injection moulding processes

Injection moulding is one of the most important polymer processing operations in the plastic industry nowadays. Due to its ability to produce complex-shape plastic parts with good dimensional accuracy and very short cycle times, the injection moulding has become one of the processes that are greatly preferred in manufacturing industry [20].

The injection moulding process is a cyclic process that can be divided in four main parts: filling, packing, cooling and ejection. The filling stage consists in filling the mould with hot polymer met at injection temperature. In the packing stage, new polymer melt is packed into the mould at a higher pressure in order to compensate the shrinkage produced by the polymer solidification. In the cooling stage the mould is cooled until its content is rigid enough to be ejected. Finally, in the ejection stage the mould is open, the part is ejected and closed another time, waiting for the beginning of the next cycle. In Figure 1.3 an

example of an injection machine (a) as well as the characteristic pressure curve measured in this processes during the whole injection cycle (b) are presented.



a) Moulding injection machine example b) Characteristic pressure curve in moulding process

Figure 1.3: Example of an injection mould machine (a) and the characteristic pressure curve during all injection cycle

In the literature exist two studied fields related to the moulding process: the study of the moulding machine and the control of the moulding process. Related to the first one, there are several topics of interest, such as determining the best strategy for the monitoring process [21] or its benefits [22], which are the effects of the variations of the parameters during the process [23][24], the conceptual design of the process [25][26], automatic selection of the best parameters setting [27][28][29], or optimisation of the scheduling and performance [30][31].

On the other hand, the application of artificial intelligence techniques is commonly used in the process control step. For example Neural Networks are used for quality assurance [32], fault detection [33], process control [34][35]. Excluding Neural Networks, other techniques like Support Vector Machines [36] and pattern discovery [37] have been also tried in this context. At the same time, more common statistical procedure for process monitoring have been treated, such as Statistical Process Control (SPC) [23]. Finally, an Independent Component Analysis (ICA) approach as a previous step to be used with a Neural Network is presented in [38].

The utilisation of the multivariate statistics for process control in this scenario is a new contribution and will be used in first instance to create a model of the NOC region, and then use the CBR principles to identify fault sensors occurrence.

1.3 Proposed Methodology

The methodology proposed in this Master Thesis, and that has been tested in the above application fields, consists in:

1. Create the NOC region model of the process using PCA to define control charts based on T^2 and Q statistics.

2. Refine the model obtained in the previous step according to the statistical control limits and determine the out of control situations.

3. Use a CBR through historical data projected in the principal component space, and using multivariate statistics from the model as attributes.

In order to achieve these main steps the following tasks has been conducted:

1. Study of the previous work done in the research group for every application fields to deeply understand the methodology basis and characteristics, as well as find some solutions to the initial limitations or handicaps.

2. Analyse the original data organisation to select the treatment required to apply the methodology.

3. Implement a first version of the methodology over the application domains and compute and analyse its results.

4. From the conclusions obtained in the previous task and the process understanding, generate a more accurate scenario to test and compare results.

5. Expose the constraints to apply the methodology. Also determine the benefits and limitations of each of the composing methodologies (PCA and CBR), as well as their interaction.

6. Proposal of future actions to overcome the limitations and new investigation fields, like new techniques (Support Vector Machines (SVM) for classification, Dynamic Principal Component Analysis (DPCA) for time dependency determination) or paradigms (Multiphase Principal Component Analysis (MPPCA) to study each phase independently) to complete the PhD.

1.4 Outline

Now, the outline this Master Thesis will be exposed to clarify its organisation, as also a brief description of what can be found in each chapter.

In this first chapter (Chapter 1), a brief introduction of the main motivation of this Master Thesis, the application domains and previous work taken as reference have been explained.

In Chapter 2, the main basis of PCA and a variation to work with finite duration processes will be explained. Then, how these techniques can be used for process monitoring, and finally, its main benefits and limitations will be exposed.

A case-based approach to work in the PCA space is introduced in Chapter 3. First, the CBR fundamentals are presented. Then the adaptations needed to work in the PCA space are exposed. And at the end, the conclusions extracted from this combination are detailed.

Chapter 4 presents the methodology used for validation and the results obtained. The validation can be divided in 2 main parts: the first one will present the steps required before the application of the procedure to represent the inherent process behaviour. The other one will concern the steps taken during the methodology application for performance evaluation. After that, a brief introduction of the domain and its results will be carried.

The main conclusions and contributions of this Master Thesis are presented in Chapter 5, as also the future tasks to complement this work.

Finally, some bibliographic references that have been consulted during the writing of this document are presented.

Chapter 2

Process Monitoring based on latent structures. The Multiway Principal Component Analysis (MPCA) case

In this chapter it will be introduced a variation of the original PCA technique to deal with finite duration processes: Multiway Principal Components (MPCA). First, the basis of PCA will be explained. Then, the previous steps, also known as pre-process tasks, to apply MPCA will be described. Next, how this procedure reduces the dimensionality, this is, how to select the number of principal components to keep. In the next 2 sections how the model obtained can be used for fault detection and diagnosis is detailed. Finally, the main advantages and drawbacks of this methodology will be exposed.

2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a technique for data compression and information extraction. PCA is used to find combinations of variables or factors that describe major trends in a data set [39]. That is, PCA is concerned to explain the variance-covariance structure through a few linear combinations of the original variables. It is commonly used as a reduction technique and for interpretation of data structure [40].

Processes involving a large number of variables can be monitored using this technique. Observations during normal operation conditions are used to build a

data model. Further it is used to assess the behaviour of the process by checking new observations against this model in the principal component space (fault detection). In case of detecting an abnormal situation it is possible to identify (fault location, diagnosis) the variables, in the original space, responsible for that [40].

2.1.1 Mathematical Formulation

Multivariate data (observations during normal operation conditions) is expected to be organised in a matrix structure, X, with m variables and n observations. Variables are assumed to be centred (zero mean) and standardised (unit variance).

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & ... & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix} \tag{2.1}$$

The sample covariance matrix (S) can be computed with the following expression:

$$S = \frac{1}{n-1} X^T X \tag{2.2}$$

And solving an eigenvalue decomposition of the sample covariance matrix S, the loading vectors for this sample can be obtained:

$$S = \frac{1}{n-1} X^T X = V \Lambda V^T \tag{2.3}$$

The orthonormal column vectors in the matrix V are commonly known as *loading vectors*, and the variance of the training set projected along the direction described by the $i-th$ column of V , i.e. σ_i, corresponds to the root square of the $i-th$ element of the diagonal matrix Λ. That is, the diagonal matrix Λ contains the non-negative real eigenvalues of decreasing magnitude ($\lambda_1 > \lambda_2 > \cdots > \lambda_m \geq 0$) with ($\lambda_i = \sigma_i^2$).

This is equivalent to solve the stationary points of the following optimisation problem:

$$\underset{v \neq 0}{max} \left(\frac{v^T X^T X v}{v^T v} \right) \tag{2.4}$$

Where $v \in R^m$ are the loading vectors.

2.1.2 Projection

The most important characteristic when applying PCA is the dimensionality reduction in the number of variables. This reduction is attained by selecting the first r columns of the loading matrix to build the matrix $P \in R^{m \times r}$; i.e. the loading vectors eigenvectors) associated with the first r eigenvalues ($r < m$). The projections of the observations in X onto the lower dimensional space are contained in the *score matrix*, T, computed as follows:

$$T = XP \tag{2.5}$$

And the projection of scores, T, back onto the m-dimensional observation space can be computed with:

$$\hat{X} = TP^T \tag{2.6}$$

The difference between X and $\hat{X}$ is the residual matrix E, [41]. It contains a vector for each observation orthogonal to the principal components (scores) and resumes the variance not captured for the r components selected in the new space (see Equation 2.7). The principal components represent the selection of a new coordinate system obtained by rotating the variables after pre-processing (Subsection 2.2.1 and 2.2.2) and projecting them onto the reduced space defined by the first r few principal components, where the data are described adequately and in a simpler and more meaningful way. The principal components are ordered such that the first one describes the largest amount of variation in the data, the second one the second largest amount of variation, and so on [42]. By retaining only the first r principal components, the X matrix is approximated by Equation 2.7 [43]. Thus, the complete PCA model can be mathematically expressed as follows [44]:

$$X = \sum_{j=1}^{r} t_i p_j^T + E \tag{2.7}$$

Where r is the number of principal components selected following some criteria and grouped in the score vector $T = t_1, .., t_r$ presented in Equation 2.6. For example the analysis of cumulative variance captured for the considered principal components can be used:

$$\sum_{i=1}^{r} \sigma_i^2 = \sum_{i=1}^{r} \lambda_i \tag{2.8}$$

2.2 Multiway Principal Component Analysis

The PCA methodology presented before can be directly applied on two-dimensional matrices (*observations* $\times$ *variables*). Finite duration processes are usually represented by time series of variables representing the execution of the process. Consequently, a three dimensional matrix is needed to represent the data set

(*observations* × *variables* × *time*) as shown in Figure 2.1. This added complexity implies to perform a two steps pre-processing before applying the PCA methodology: unfolding and scaling.

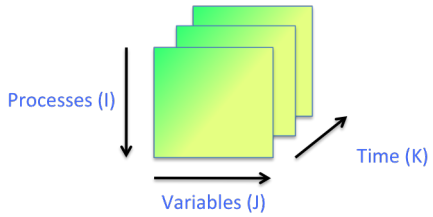


Figure 2.1: 3D data matrix associated with finite duration processes

2.2.1 Unfolding the data

From the six feasible unfolding directions, only 2 of them are meaningful for monitoring: unfold in the process direction (batch-wise or Nomikos-MacGregor approach [42]) and unfold in the variable direction (variable-wise or Wold approach [45]).

The batch-wise approach (Figure 2.2) fixes the processes axis (kept as rows) and collapse as columns the product of variables × time. So, a row will be representing each of the processes, and a column represents a time instant of a given variable for each process. This approach can only be applied whenever all data of the process is available, what is, when the process has finished, since a whole row has to be projected. However, in [46] are proposed 3 ways to solve this problem maintaining the unfolding:

- The remaining values are considered to be the mean trajectory of the processes.
- All future values have the same deviation than the last sample.
- All future values are predicted from the observed values.

On the other hand, when the normalisation step is carried with this approach, variations observed in the unfolded matrix represent variation with respect the mean trajectory [47].

The variable-wise approach (Figure 2.3) fixes the variables axis (kept as columns) and the product processes × time are the rows. In this case, one row will represent a time instant of a given process and the columns are the values of one variable for all processes at every time instant. This approach does not

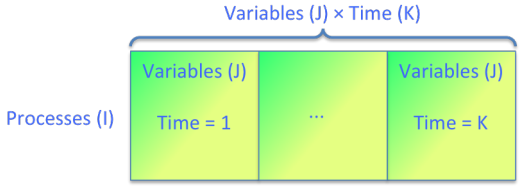


Figure 2.2: Batch-wise unfolding of the original 3D matrix

require any prediction to be applied at every time instant, but has the drawback of leaving the non-linear time variations in the normalised data matrix [47].

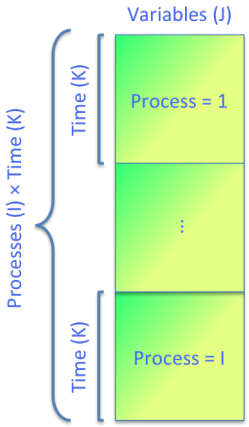


Figure 2.3: Variable-wise unfolding of the original 3D matrix

2.2.2 Scaling of the data

The PCA methodology requires the data to be mean-centred, that is, the axis origin of the new projection space will be the mean value. But there are cases where some variables present different value range or deviations. In this case, the data not only has to be centred, it also has to be scaled. In this subsection the 3 main normalisation procedures when dealing with finite duration processes will be presented.

Continuous scaling (CS)

The continuous scaling procedure assumes that variables in the data matrix share the same distribution. So, it will compute 1 mean and 1 standard deviation

for each of the original variables (J) the following way:

$$\mu_j = \frac{\sum_i^I \sum_k^K x_{ijk}}{I \times K} \quad \sigma_j = \frac{\sum_i^I \sum_k^K (x_{ijk} - \mu_j)}{I \times K} \tag{2.9}$$

This type of normalisation is not usually used because during the process is not mandatory that variables follow the same distribution, especially if the process can be divided in phases. Another point to note is that this procedure does not remove the mean trajectory of the variables along time (the same drawback that the variable-wise unfolding), what can end in a bad performance of the monitoring model.

Group scaling (GS)

Group scaling tries to overcome this problem by computing one mean for each variable for all time instants, so the mean trajectory is eliminated. However, it still maintains one standard deviation per variable, what results in J × K means and J standard deviations, computed as follows:

$$\mu_{jk} = \frac{\sum_i^I x_{ijk}}{I} \quad \sigma_j = \frac{\sum_i^I \sum_k^K (x_{ijk} - \mu_{jk})}{I \times K} \tag{2.10}$$

Computing only 1 standard deviation per variable, it is supposed that the variability is kept along time in the whole process.

Auto scaling (AS)

When variability changes during the process, it is necessary to compute one standard deviation at every time instant. This is the basis of auto scaling, so there will be J × K means and standard deviations, computed:

$$\mu_{jk} = \frac{\sum_i^I x_i jk}{I} \quad \sigma_{jk} = \frac{\sum_i^I (x_{ijk} - \mu_{jk})}{I} \tag{2.11}$$

Finally, Table 2.1 shows the number of means and standard deviations computed for each normalisation.

Procedure	Number of means (μ)	Number of standard deviations (σ)
Continuous scaling	J	J
Group scaling	J × K	J
Auto scaling	J × K	J × K

Table 2.1: Number of means and standard deviation for each normalisation technique

2.3 Dimensionality reduction and MPCA modelling

One of the most important points in the MPCA model generation is the number of principal components to retain, since the two control statistics available are sensitive to it, due to noise in sensor measurements, the no representation of some sensors or the redundancy of information related to one sensor. The combination of the previous situations can lead to miss detections (non detected fault occurrence). Moreover, fault isolation depends on the correct selection of the principal components to retain [48].

Several are the techniques in the literature to decide the number of principal components to keep in the model. Most of them rely on the analysis of the eigenvalues of the covariance matrix, and are based on analytical and graphical strategies [49].

2.3.1 Percent Variance Explained

This procedure is based in the percentage of variance explained by the eigenvalues obtained from the covariance matrix (Equation 2.12) [48]. The main assumption is that until a fixed value of percentage the information retained is from the variations within the process, and the rest is considered to be noise.

$$pvf_i = \frac{\sum_{j=1}^{i} \lambda_j}{\sum_{k=1}^{J} \lambda_k} \times 100 \tag{2.12}$$

Where pvf_i stands for the percentage of variability explained by factor i, λ refers to the eigenvalues of the covariance matrix and J is the number of original variables. Although it is a very simple methodology, it is not very often considered since each process presents a different variation and normally it is an unknown value.

2.3.2 Kaiser-Guttman Criterion

This method was presented by Kaiser in 1960 and Guttman in 1954, and is based on keeping only those principal components with an associated eigenvalue greater than one, because it has no interest to maintain a variable that brings less information that the original standardised variable [49]. An example of the application of the methodology is presented in Figure 2.4.

The main drawback of this method is that leads to a rather arbitrary decision like throw away a factor with an associated eigenvalue of 0.99 and keep another one with a 1.01 value. Another point to take into account is that this method tends to overestimate the number of principal components.

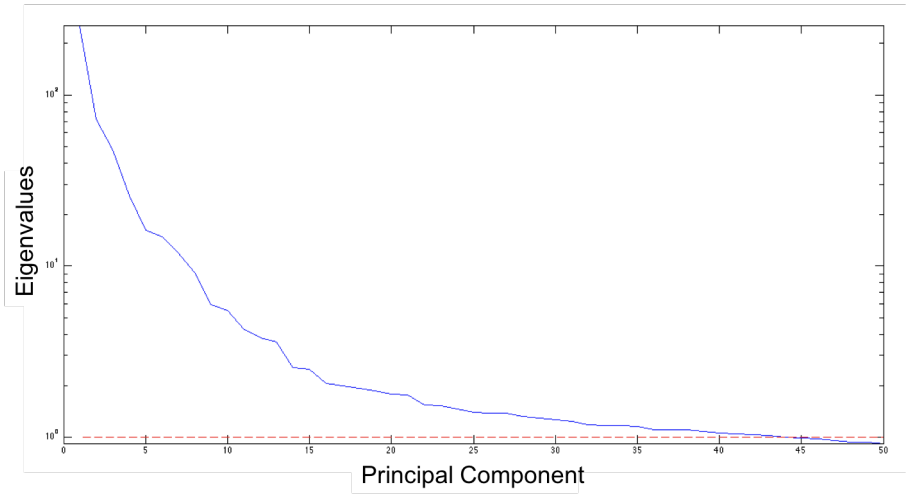


Figure 2.4: Example of the Kaiser-Guttman criterion

2.3.3 Cattell's Scree Test

This method was developed by Cattell after observing that plots of eigenvalues versus their respective principal component number presented a characteristic shape (Figure 2.5). Eigenvalues tend to drop off quickly at the beginning, with a continuous decreasing up to a curve. After that, the remaining eigenvalues present a quasi-linear fall. The curve change represents the separation between the process variation and noise and linear relationships between variables [48].

The main counterpart of this procedure is its subjectivity, since there is no formal definition for the cut-off point. Contrary to what happened with the Kaiser-Guttman methodology, this procedure tends to overestimate the number of principal components to keep.

2.3.4 Representation of original variables

This approach is based on keep principal components until all variables of the process, or al least the most important ones, are represented. A graphical example for representation of a variable is shown in Figure 2.6.

In this Master Thesis, it will be used a consensus among Kaiser-Guttman, Cattell's Scree Test and the representation of the original variables.

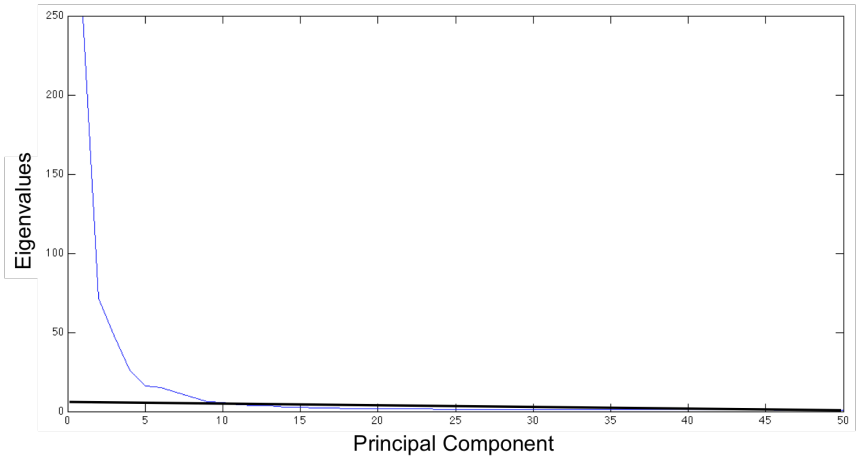


Figure 2.5: Example of the Cattell's criterion

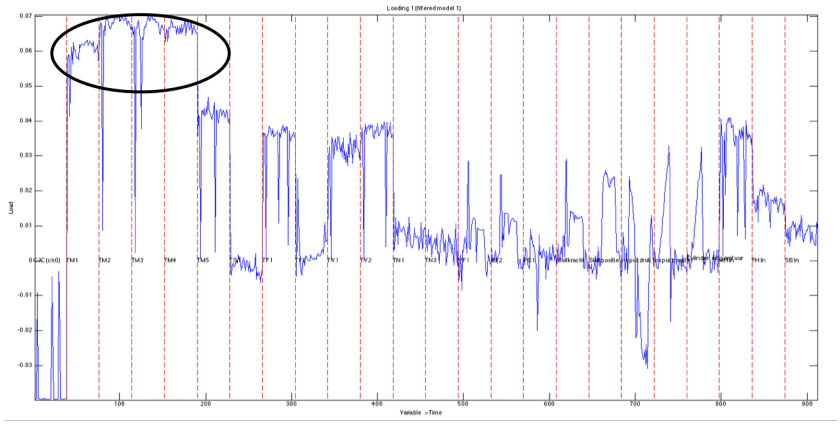


Figure 2.6: Example of the some variables considered represented

2.4 Fault detection : T^2 and Q

Two complementary control charts are usually used for multivariate process monitoring using PCA. The purpose is to assess new observations against the PCA model built during normal operation conditions. T^2 and Q statistics are used to build them. Control charts based on T^2 can be plotted based on the first r principal components as follows [43]:

$$T^2 = \sum_{j=1}^{r} \frac{{t_j}^2}{\sigma_j^2} = \sum_{j=1}^{r} \frac{{t_j}^2}{\lambda_j} \tag{2.13}$$

Where T^2 can be computed for each observation by adding the square of the r components ${t_j}^2$ weighted by their variances σ_j^2 (eigenvalues). This results in a measure of distance (Mahalanobis distance) of each observation to the centre of the model. A graph or control chart, built with this data is useful to detect variations in the plane of the principal components (r) greater than common-cause variations but preserving the structure gathered by the PCA model. Nevertheless, when a new event, x, in the process produces a large variation out of the hyperplane described by the r principal components this implies that the data structure has been broken. This type of event are detected by computing the Q-statistic or Squared Prediction Error (SPE) of the residual of each assessed observations defined as ([42], [50]):

$$Q = \sum_{j=1}^{m} (x_j - \hat{x}_j)^2 = (x - \hat{x})^T (x - \hat{x}) \tag{2.14}$$

Where $\hat{x} = (\hat{x}_1, \cdots, \hat{x}_j, \cdots, \hat{x}_r)$ is computed from the reference PCA model using Equation 2.6. Q-statistic is much more sensitive than T^2 to changes in the process structure. This is because Q during normal operation conditions is very small (typically associated to noise) and therefore any minor change in the process will affect the correlation structure of observed data. T^2 represents a greater variance and therefore it is less sensible to small variations in the process. Figure 2.7 represents the NOC region where T^2 and Q thresholds are defined.

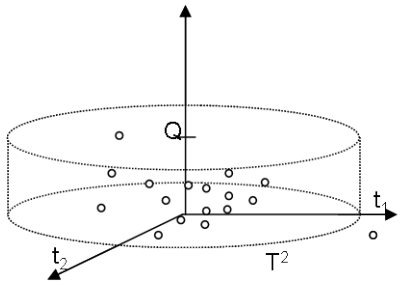


Figure 2.7: Graphical representation of Q and T^2 statistic

2.5 Fault diagnosis: contribution analysis

The role of the contribution plots to fault isolation is to indicate which of the variables are related to the fault rather than to reveal the actual cause of it [51].

Once a faulty situation has been detected, the contribution of each variable in the original space to the individual scores can be analysed and quantified following these steps if the fault was detected using the T^2 statistic:

1. For a faulty observation x, check the normalised scores (t_i/σ_i^2) and determine the $s \leq r$ scores responsible for the fault situation (or out-of control status in terms of statistical process control). For instance, those scores with:
$$\left(\frac{t_i}{\sigma_i}\right)^2 > \left(\theta_\alpha^2\right)^{\frac{1}{r}} \tag{2.15}$$
(t_i is the score of the observation projected onto the $i-th$ loading vector, σ_i^2 is the corresponding singular value, r is the number of principal components used in the representation, θ_α^2 is the T^2 fault detection threshold)

2. Calculate the contribution of each variable x_j (in the original space) to the s out of control scores t_i:
$$cont_{i,j} = \frac{t_i}{\sigma_i^2} p_{i,j}(x_j - \mu_j) \tag{2.16}$$
($p_{i,j}$ is the element $(i,j)-th$ of the loading matrix P (see subsection 2.1.2); μ_j is the mean value of x_j)

3. For each process variable x_j calculate the total contribution, taking into account only the positive contributions ($cont_{i,j}$ has to be set equal to zero):
$$CONT_j = \sum_{i=1}^{s} cont_{i,j} \tag{2.17}$$

4. Select the variables responsible to the faulty situations from a representation of all $CONT_j$.

If the statistic that detected the fault was Q, the contribution of the variable that caused the fault is directly extracted from:

$$cont_{j,new} = (x_{new} - \hat{x}_{new})^2 \tag{2.18}$$

Where x_{new} are the measured variables of a new case and $\hat{x}_{new}$ is the reconstruction of the values using Equation 2.6 with the retained components.

This contribution computation is centred on the analysis of the data of every new case. Another feasible approach could be using a comparison. For instance, all NOC region processes can be used to compute the average and control limits of the contributions, as in [9]. Another possibility is compare two consecutive batches (a and b) in order to detect trends in the scores ([52]). This can be done conducting the following steps:

1. Compute the contribution of each component by:

$$cont_{i,j} = \sum_{j=1}^{J} p_{i,j} \times (x_{b,j} - x_{a,j}) \tag{2.19}$$

2. Assign $cont_{t_i} = 0$ if the contribution was negative. With this, the detectably becomes higher.

3. Compute the total contribution as in Equation 2.17:

2.6 Conclusions: advantages and drawbacks

As has been shown up to the moment, MPCA is a powerful tool that compresses information such a way that the variability of the original set is not modified. Another interesting point of the methodology is that the dimensionality reduction does not imply to lose the data, actually, the data can be projected back to the original space by using Equation 2.6. Also, this methodology is based on a graphical representation of the data, what makes easier the process understanding and explanation to external people, as also provides some control statistics for process monitoring. Something to take into account is that the methodology return variables that are independent among them (orthogonal), invariant to scaling because the original data has been mean-centred, and with optimal dimension reduction using the first r principal components, what is, the minimum reconstruction error is granted.

Although all these great advantages, it has some drawbacks that have to be considered. First of all, this methodology requires a great amount of data in order to construct a reliable model. Another initially important requirement was that the data matrix had to present continuous variables. This is not a strict requirement nowadays since research in this area is focused on finding methods to compute the covariance matrix with categorical variables, like in [53].

Chapter 3

Case-Based Diagnosis in the principal component space

In this chapter it will be explained how Case-Based Reasoning (CBR) has been applied in order to solve some of the drawbacks of the MPCA procedure and how CBR takes advantage of some of the characteristics of the MPCA space. First, the basic methodology of the CBR will be explained. Next, cases and case base definition will be exposed, following with the definition of the similarity functions and neighbourhoods in the principal component space. Finally, the conclusions, advantages and drawbacks of the combination of both approaches will be commented.

3.1 Case-Based Reasoning: the 4R-Cycle

Case-Based Reasoning (CBR) is a reasoning approach to problem solving capable of using the knowledge acquired by previous experiences [54]. It has demonstrated to be a good option for solving problems in several domains (diagnosis, prediction, control, planning, etc.)[55]. Like in many machine learning algorithms, the independence of attributes involved in the retrieval of cases is usually assumed, i.e. when the Euclidean distance is used to defined the neighbourhood. As it is exposed in [56], attribute independence also lets a classifier to collect the evidence for a class from individual attributes separately. So, the contribution of an attribute to a class can be determined independently from the other attributes. This requirement, not only simplifies the learning algorithms but it also results in a robust performance and simpler models.

The basic functions that all CBR present are known as the 4-Rs [55], and can be organised in a cycle as depicted in Figure 3.1:

1. RETRIEVE the most similar cases of the new case.
2. REUSE the information in these cases to solve the new problem.

3. REVISE the proposed solution.

4. RETAIN the new information of the new experience in order to solve new similar problems.

To solve a new problem, the most similar cases are retrieved from the experiences previously stored. The information contained in these retrieved cases is then reused to propose a possible solution. Once the solution is evaluated, the case is retained, if necessary, for further classifications.

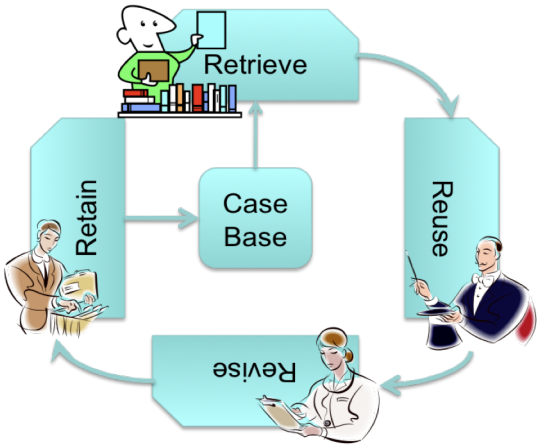


Figure 3.1: The CBR cycle

In this Master Thesis the retain function will not be implemented, and the revise procedure will be commented in Chapter 4 in Section 4.1, since it involves several steps.

3.2 Case and Case Base definition in the projection space

The basis of Case-Based Reasoning is the case definition. A case is the minimum representation of a past experience and its solution [57]. When several cases are available, they can be grouped in a Case Base.

A basic case structure composed of original observations, scores and the basic statistics is proposed:

$$\mathrm{c} = \{x_1, ..., x_m, t_1, ..., t_r, T^2, Q, l\}$$

Where x_1 to x_m stands for the m original variables, t_1 to t_r are the retained r first principal components (scores). They are obtained with the loadings (projection operator) obtained with previous observations of the process during normal operation conditions. T^2 and Q are the statistical indices used to measure the adequacy of each observation to the projection model (normal operation conditions) as it has been explained in a previous section. Finally, l refers to a diagnostic of the observation. In this formulation, for simplicity, l can be associated with a label reducing the diagnose problem to a classification one.

3.3 Case-Based reasoning in the principal component space

Using the principal components as the descriptors used in the CBR adds the following interesting features:

- Independence of the new variables, since each component is orthogonal to the previous one.
- One of the most important aspects in CBR is deciding the importance degree of each attribute or descriptor. However, when using the scores as attributes, this task becomes trivial, since the associated eigenvalues of the retained principal components relates the percentage of global variation explained, what is the relative weight of each principal component.
- The principal components are inherently ordered, since they are found in a decreasing order that assure the minimum reconstruction error of the original data set.

When combining both approaches some benefits should be noted. It can be seen that the CBR methodology relaxes the number of past experiences needed for the model generation using the MPCA methodology, because its inherent capacity to learn from past experiences, that can be used to improve the original model quality. Another important point is that MPCA provides two control statistics that separate cases within the NOC region (not important for monitoring) and cases in the AOC region, that are the ones to be detected. Thanks to that, the CBR will be used to specify the type of fault, and as a support to classify those cases near the statistical control limits. So, it can be seen that the utilisation of MPCA as a previous step of the CBR simplifies and complements it.

3.4 Similarity based methods for diagnosis in the principal component space

According to CBR methodology, case reuse will be based on the nearest neighbours criterion. Consequently, neighbourhood based on distance or similarity

criteria has to be defined. In the following section several similarity criteria in the principal component space are defined and interpreted for monitoring purposes. Then, in the following subsection they are combined to identify appropriate neighbourhoods.

3.4.1 Distance criteria in the projection space

Three basic similarity criteria are proposed. The first one is simply an Euclidean distance in the principal component space whereas the second and third are basically the comparison of Q and T^2 statistics between observations.

Euclidean Distance between observations in the principal component space

Taking advantage that the application of PCA results in new r independent components. The space defined by scores will be appropriate to compute an Euclidean distance between observations projected on it:

$$d_t(c_a, c_b) = \sqrt{\sum_{i=1}^{r}(t_{c_a,i} - t_{c_b,i})^2} \tag{3.1}$$

Nevertheless, remember that Principal Components are ordered according to the variance captured in each direction (eigenvalue). Consequently, it is better to weight each score according to the root square of its eigenvalue, or what is the same:

$$d_t(c_a, c_b) = \sqrt{\sum_{i=1}^{r}\frac{(t_{c_a,i} - t_{c_b,i})^2}{\lambda_i}} \tag{3.2}$$

Where r stands for the number of retained principal components, $t_{c_a,i}$ is the $i-th$ score of a case c_a, for example a new observation, and $t_{c_b,i}$ could represents the same for an observation in the case base. A geometrical interpretation of this distance is shown in Figure 3.2. This similarity criterion does not take into account the adequacy of projections to the model.

Q Similarity

As exposed in Section 2.4, the Q statistic index is related to the projection error. Consequently, observations with a low value of Q are consistent with the projection model (obtained with observations gathered during normal operation conditions) and they will close to the hyperplane defined by the r retained principal components. On the other hand, observations with a large Q are expected to be inconsistent with the model structure and consequently they are candidates to faulty situations.

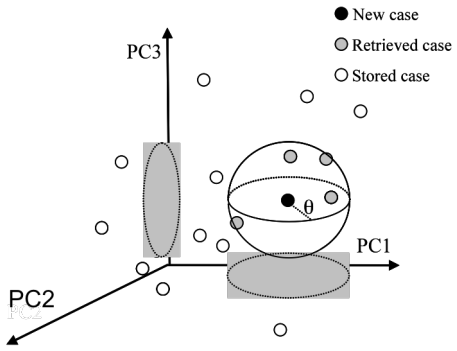


Figure 3.2: Similarity based on the Euclidean distance with two principal components

Therefore, observations with a similar Q can be used to identify similar operation conditions (normal or abnormal). A simple difference can be used to compute this similarity:

$$d_Q(c_a, c_b) = |(Q_{c_a} - Q_{c_b}| \tag{3.3}$$

A possible geometrical interpretation of this distance is showed in Figure 3.3.

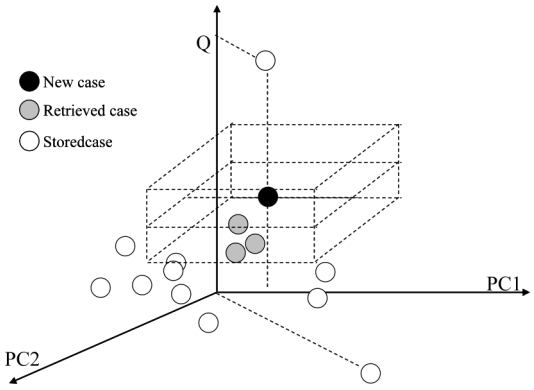


Figure 3.3: Similarity based on the Q statistic distance

T^2 Similarity

In Section 2.4 the statistic T^2 has been presented as a measure of the distance (Mahalanobis distance) of an observation to the centroid of the model. In fact,

it is a square distance and represents the dispersion from the mean of the model since the scores are normalised (unit variance) previous to compute the T^2 index.

Low values of T^2 represent observations close to mean whereas high values of T^2, over the control limits, are evidences of an abnormal behaviour; although it does not necessary implies that the correlation structure has been broken (this will depend on Q).

Similarity according to the statistic T^2 will be computed as follows:

$$d_{T^2}(c_a, c_b) = \left| T^2_{c_a} - T^2_{c_b} \right| \tag{3.4}$$

And a possible geometric interpretation is shown in Figure 3.4.

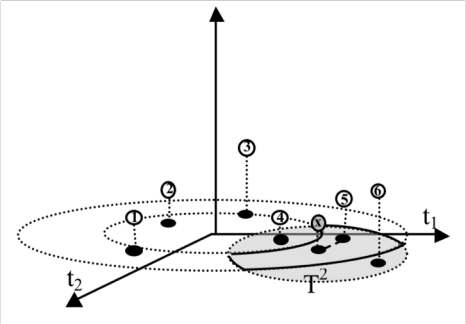


Figure 3.4: Retrieval space using the T^2 statistic

3.4.2 Neighbourhood in the monitoring space

The neighbourhood of an observation c_a computed with a distance d can be designed by the observations closer than a threshold θ as the following relation suggests:

$$N_d(c_a, \theta) = \{c_i / d(c_a, c_i) \leq \theta\} \tag{3.5}$$

Based on this definition several combinations can be defined to retrieve a set of observations useful for process monitoring. For example the neighbourhood of observations of normal operation conditions (NOC) are expected to be around the origin in the principal component space. Therefore, they would be retrieved as the neighbours of a representative theoretical case located in the origin of coordinates, c_0, with a confidence level α by selecting an appropriate value for the thresholds ($\theta_{T^2} = T^2_\alpha$ and $\theta_Q = Q_\alpha$).

$$N_{NOC} = N_{d_{T^2}}(c_0, T^2_\alpha) \cap N_{d_Q}(c_0, Q_\alpha) \tag{3.6}$$

Operating in a similar way is possible to select the nearest observations with a similar deviation with respect to the projection model, in terms of Q or T^2 using the following relations. Once an observation c_a has been projected and the resulting Q evidences that it is not consistent with the model structure, then a focalised search among neighbours with the same dissimilarity can be useful for diagnosis purposes.

$$N_{Q \wedge t}(c_a) = N_{d_t}(c_a, \theta_t) \cap N_{d_Q}(c_a, \theta_Q) \tag{3.7}$$

The intersection is proposed as a refinement of the neighbourhoods when specific search are required. In a similar way, neighbourhood can be restricted to the observations in the hyperplane defined by scores and with a similar value of the index T^2 using the following sentence (Figure 3.5):

$$N_{T^2 \wedge t}(c_a) = N_{d_t}(c_a, \theta_t) \cap N_{d_{T^2}}(c_a, \theta_{T^2}) \tag{3.8}$$

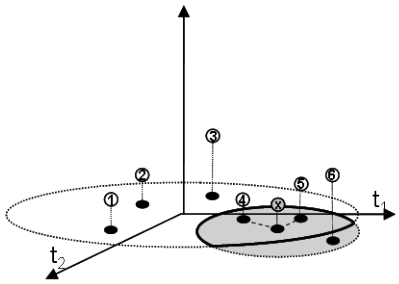


Figure 3.5: Neighbourhood of a new case based on similarity in the T^2 hyperplane

Other useful neighbourhoods can be those defied to retrieve the most similar cases of c_a inside the region defined as normal operation conditions (Figure 3.6).

$$N_{NOC \wedge t}(c_a) = N_{NOC} \cap N_{d_t}(c_a, \theta_t) \tag{3.9}$$

Or in case of focusing on the observations out of the NOC region:

$$N_{\neg NOC \wedge t}(c_a) = N_{d_t}(c_a, \theta_t) - N_{NOC \wedge t}(c_a) \tag{3.10}$$

3.5 Conclusions: advantages and drawbacks

In this chapter the usage of a CBR methodology has been presented. As was mentioned in the previous chapter, the application of the MPCA procedure produces new variables that are independent among them, a key point when applying any machine learning algorithm. With the combination of both techniques the necessity of a great number of processes to generate a model is relaxed, since

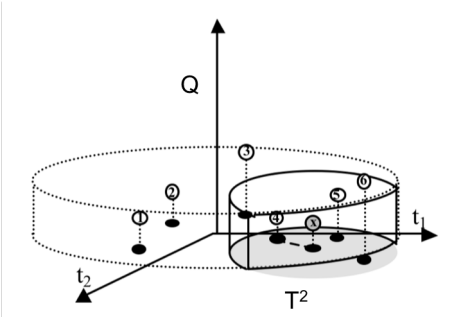


Figure 3.6: Neighbourhood of a new case included in the NOC region

the case base will complement the MPCA model. Another interesting feature obtained with the combination of both approaches is that the operator does not require a deep knowledge of the process to interpret the fault occurrence. For instance, when only using the MPCA contribution plots, the interpretations of the graphics were dependant to the knowledge of the operator in charge. With the utilisation of the CBR engine, the explanation of the errors can be in a more natural way, by specifying with comprehensive explanations the faults, although both approaches are complementary.

But the combination of both techniques also presents some drawbacks. The first one, and most significant is that past situations and its solutions (existence of an initial case base) correctly classified to apply the methodology. Related with the case base, the resulting combined model will be capable of classify the fault occurrence in the present typologies in the case base, as well as the information that can be extracted depends exclusively to the one contained in the cases. To solve this problem, novelty discovery techniques could be applied once a fault can be labelled with a unique class. And a common disadvantage to all supervised learning techniques is that when a new case arrives, it is not possible to know whether the classification was correct or not. Actually, one of the methodologies used to validate the classification can be used to compute the ratio of well classification as well as how behaves the model when wrongly classify cases, that is, computing the Confusion Matrix explained in the following chapter.

Chapter 4

Application domains, examples and validation

In this chapter, first the methodology to validate the classification performance of the methodology will be explained, followed for the results obtained in the tested fields.

4.1 Validation

In this section it will be presented the methodology to evaluate the performance of the classification in an objective way, or the so-called revise CBR functionality.

4.1.1 n-Fold Cross Validation

In n-Fold Cross Validation, the available data is divided into n folders containing approximately the same number of examples. The stratified version of this technique takes into account the several ratios among classes present in the original set. Once the data is divided, one of the n folds of samples is retained for validation of the model formed by the remaining $n-1$ data fold. This process is repeated n times (once for each fold) [58]. Figure 4.1 presents this methodology in a graphical way.

4.1.2 Confusion Matrix and performance indices

In order to evaluate the classification performance, the confusion matrix is used. A confusion matrix is a form of contingency table showing the differences between the true and predicted classes for a set of labelled examples, as is shown in Table 4.1 [59].

Where TP stands for True Positive (cases correctly predicted from the reference class), FP for False Positives (cases classified of the reference class with

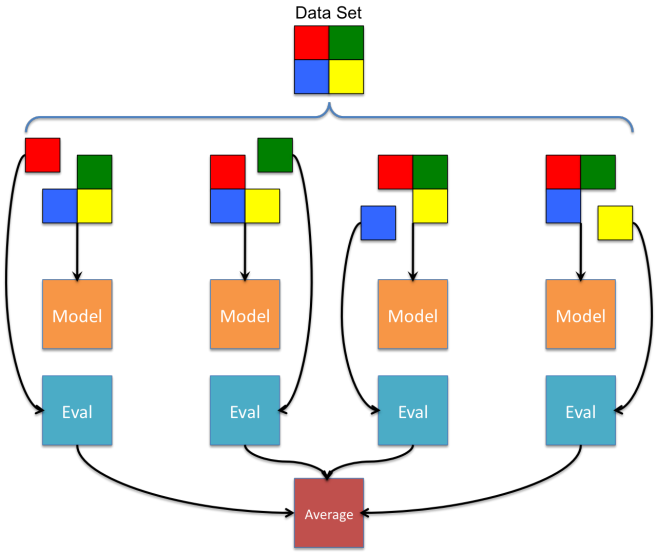


Figure 4.1: n-Fold Cross Validation graphical procedure

		Real Class	
		Ref	**No Ref**
Predicted Class	**Ref**	TP	FP
	No Ref	FN	TN

Table 4.1: Confusion Matrix elements

its real class being non reference), FN for False Negative (cases classified as non reference class and its real class being of the reference class) and TN for True Negative (cases correctly classified as non reference class). Using this information, some statistics can be computed, for example:

Accuracy

Accuracy measure the proportion of correctly classified cases among all the cases used for testing, and is computed as:

$$Accuracy(ACC) = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Precision

Precision measure the proportion of correctly classified cases from the reference class from all cases that were predicted as the reference one, and is computed by:

$$Precision(PRE) = \frac{TP}{TP + FP} \tag{4.2}$$

Sensitivity

Sensitivity measures the ratio between cases correctly classified as the reference class among all cases that have been predicted as the reference class, and is computed by:

$$Sensitivity(SEN) = \frac{TP}{TP + FN} \tag{4.3}$$

Specificity

Specificity measures the ratio between cases correctly classified as non reference class among all cases that have been predicted as non reference class, and is computed by:

$$Specificity(SPC) = \frac{TN}{TN + FP} \tag{4.4}$$

In this Master Thesis, the relation with FP and FN will be used to estimate the trust of a classifier when the real class is not known.

4.1.3 Receiver Operating Characteristic (ROC) curve

The ROC curve representation is a two-dimensional graph where the y-axis represents sensitivity and x-axis represents the False Positive Rate (FPR), or what is the same, 1- Specificity. Observe that the lower left point (0,0) represents a classifier that never classifies correctly the cases of the model. The upper left point (0,1) represents the perfect classifier (it never misses to classify the cases of the model, and also determine correctly the cases that are not represented by the model) and the upper right point (1,1) represents a classifiers that always classifies correctly cases fitting the model, but always classifies incorrectly cases different from the model [60]. Figure 4.2 presents a ROC curve constructed the way described in this section.

4.1.4 Area Under the ROC Curve (AUC)

Because in some operating points sensitivity can be increased with a minor loses in specificity and in others this is not possible, a non-ambiguous possible comparison of performance can be achieved by computing the Area Under the ROC Curve (AUC). A simple way of computing this value is using the trapezoidal integration method described in [59].

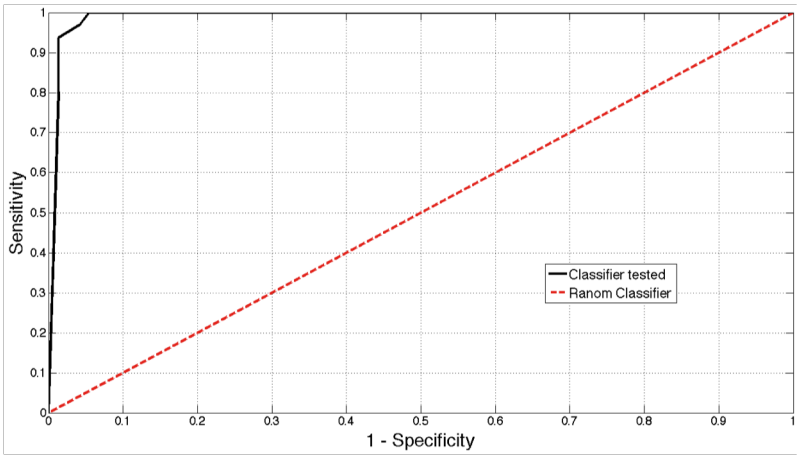


Figure 4.2: Example of ROC curve

$$AUC = \sum_i \{(1 - \beta_i \times \Delta\alpha) + \frac{1}{2}[\Delta(1-\beta) \times \Delta\alpha]\} \tag{4.5}$$

$$\Delta(1-\beta) = (1-\beta_i) - (1-\beta_{i-1}) \tag{4.6}$$

$$\Delta\alpha = \alpha - \alpha_{i-1} \tag{4.7}$$

Where β_i represents the specificity value of the actual point (i), β_{i-1} represents the specificity value of the previous point $(i-1)$, α represents the sensitivity value of the actual point (i) and α_{i-1} represents the sensitivity value of the previous point $(i-1)$. This value will be used to select the best classifier, since it measures the overall classification ability of the classifier.

4.1.5 Detection of outliers

Once the number of principal components has been fixed, the last step to conduct when constructing the model, is to remove the cases too far from the centre of the model to adjust the control limit. When the covariance matrix is used, the limit to consider one case as an outlier can be computed with the following equation:

$$T_\alpha^2 = \frac{(n-1)^2(m/(n-m-1))F_\alpha(m, n-m-1)}{n(1+(m/(n-m-1)))F_\alpha(m, n-m-1)} \tag{4.8}$$

However, in this Master Thesis this equation will not be used since the limits found in each field were too narrow resulting in a high FP ratio. In fact, and taking advantage of the CBR methodology to correct cases that could lead to wrong classifications, the following outlier detection limit is proposed:

$$OUTLIER(c_i) = \{T_i^2 \geq 3 \times T^2\_lim \vee Q_i \geq 3 \times Q\_lim\} \tag{4.9}$$

Where c_i is the case being studied, T_i^2 and Q_i are the respective control statistics values and T^2_lim and Q_lim are the control limits based on a confidence level computed as was explained in Chapter 2. In this case, the confidence level used was 95 %. Whenever outlier cases are found, a new model removing those cases has to be calculated and the methodology to decide the number of components has to be applied. This procedure is repeated until no outliers are found using this criterion. A graphical example of this criterion is presented in Figure 4.3.

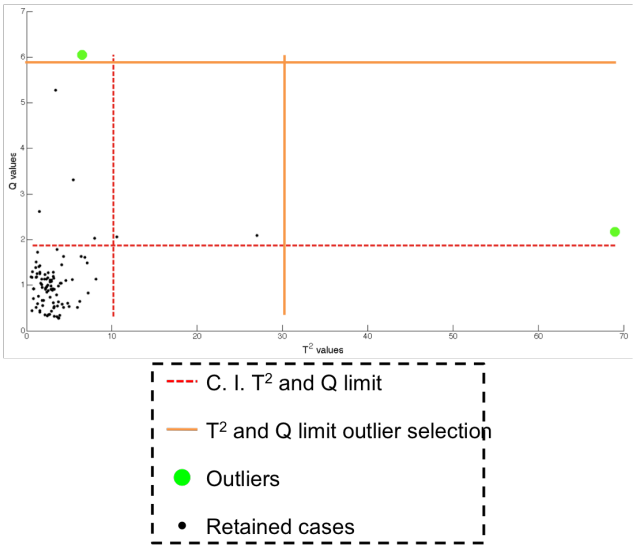


Figure 4.3: Example of two outliers found using the proposed outlier detection limits

4.2 Wastewater Treatment Plants

In this section the results obtained in a Sequencing Batch Reactor (SBR) for the wastewater treatment are presented. In order to do so, first a brief introduction to the process studied will be presented. Then the pre-treatment done to the original data will be exposed, followed by how the MPCA model has been created, as also the additional steps carried to best adequate it. Finally, some results obtained for faulty situation assessment will be presented.

4.2.1 Case study: Sequencing Batch Reactors (SBR) in LEQUIA research group

The main characteristic of a SBR is that the whole process occurs in the same reactor, following a sequence of phases, while in a continuous wastewater process plant, each phase occurs in different reactors. The SBR process is an effective alternative to treat wastewater from domestic and industrial waste [8]. This cycle consists of 4 main phases that are:

1. Fill: The influent wastewater is pumped into the reactor to be treated.
2. Reaction: Aerobic and anoxic conditions are combined to consume the substrate from the influent wastewater.
3. Settle: This phase occurs once all reaction phases have been done. In this phase, the excess sludge is drained.
4. Draw: This is the last phase and happens once the process finishes. The treated water is drawn from the reactor and waits until new water has to be treated.

The SBR plant of the LEQUIA group can be configured in several ways. The methodology this Master Thesis is based on will be applied only in a specific configuration that lasts 8 hours: 2 repetitions of the fill and reaction phases and 1 settle and draw phases combined as shown in Figure 4.4 . This decision is based in that this configuration presents the best characteristics for class separation, as explained in [9].

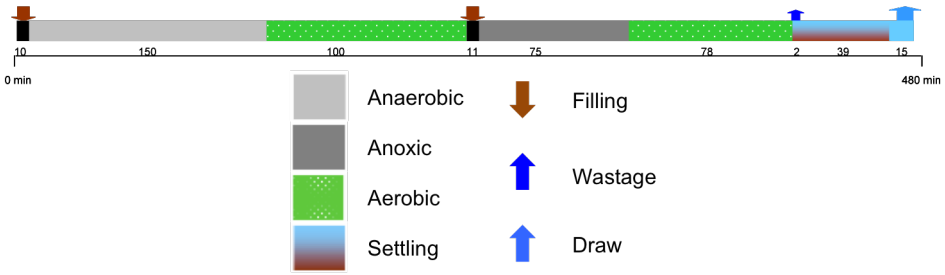


Figure 4.4: Configuration of the SBR of LEQUIA group used in this section

4.2.2 Pre-treatment of the data

The data given by LEQUIA consist of measurement of 4 variables (DO, pH, ORP and Temperature) sampled every 5 seconds. As mentioned before, the length of each process is of 8 hours, what makes a total number of time instants

measured 5760.

Since compute this huge amount of data would carry a very high computational cost, and taking into account that reactions in biological processes are very slow, the number of samples will be reduced to one for minute, considering the different phases that conform the wastewater cycle. This will be done according the following methodology:

1. Each batch is divided in its respective phases. Then, the data referred to the settling, drawing and wastage are discarded, since no biological information is contained in this phases [8].

2. For each remaining phase, at every 12 instances (1 real time minute) the minimum and maximum value are eliminated.

3. The mean ($\bar{x}$) of the remaining values (x_i) is computed, using the following equation:
$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} \tag{4.10}$$

4. $\bar{x}$ is the value stored and that will be used to apply the MPCA methodology.

The original (Figure 4.5 a)) and resulting (Figure 4.5 b)) of resampling the original data are nearly the same, so by this pre-processing step results won't be altered.

4.2.3 MPCA model creation

The unfolding procedure used in this scenario is the batch-wise, since the goal is to classify the whole process. 3 main classes were labelled by the LEQUIA experts and are presented in Table 4.2. Something to note in this scenario is that 52 processes couldn't be assigned a known class since didn't get an associated quality variable and were discarded. Finally, and as was explained in Chapter 2, the type of processes used in the model creation step are the NOC ones, that in this scenario is associated with the good quality batch.

Class	Shape of Quality Control Variables	Number of processes
Good	Correct	98
Regular	Correct but present a gain	74
Bad	Not correct	62
Unknown	–	52

Table 4.2: Subdivisions of the SBR plant data and related information

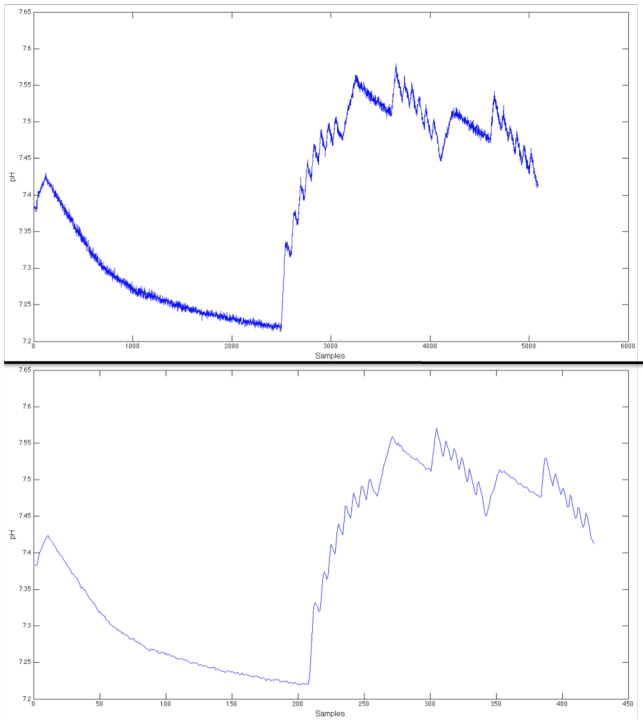


Figure 4.5: Comparison of the original data (up) with the resampled data (down)

After comparing the resulting models of both batch-wise and variable-wise. the scaling procedure chosen is the group scaling. Although both models presented a similar performance, the usage of the group scaling ended in better classification results. The remaining task to create the model is selecting the number of principal components to keep. Although only one principal component presents an associated eigenvalue greater than 1, adding the Cattell's criterion and representation of the original variables, the number of principal components chosen is 4, as shown in Figure 4.6. The loading vectors associated to each principal component are shown in Figure 4.7, with a red ellipsoid over the represented variables in the loadings. The resulting model explains a 74.6 % of the global variability.

Finally, the last task to perform is check for the presence of outliers. Using the proposed methodology (Equation 4.9). 5 cases have been labelled as outliers, presenting a completely different shape of those cases considered to be in the NOC region. So, in this case and according to Table 4.2, outliers will be associated to bad quality batches. An example of a case labelled as outlier in

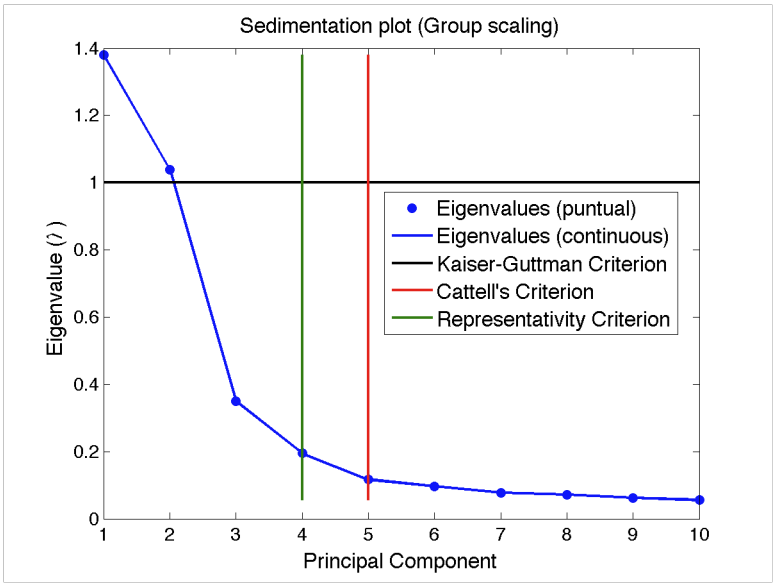


Figure 4.6: Graphical representation of all criteria used to decide the number of principal components

this scope and the mean shape of NOC region cases are presented in Figure 4.8.

4.2.4 Results

In this scenario, 234 cases are available. The number in which the case base will be divided using the Stratified n-Fold Cross Validation will be 4. According to the principle used in statistics to determine whether a set of numbers follow a random distribution, its number has to be at least 50. So, three of the fold will have 57 cases (with the same ratio as the whole case base) and the last one will have 59. The distribution of cases is shown in Figure 4.9, and in Figure 4.10 is presented how this cases are distributed in the Q and T^2 statistics in Fold 1.

As it can be seen, all folds present a similar distribution in the principal components space. So, it is expected that all folds present similar performance indices. The distance criteria used for testing will be all the ones exposed in Chapter 3: distance in the principal components space (Equation 3.2), Q-similarity (Equation 3.3) and T^2 similarity (Equation 3.4). On the other hand, three more distance criterion are proposed, and result from the combination of the previous ones. Those three distance criteria are also related to the neighbourhoods exposed in Chapter 3. From the ones exposed there, the following modifications have been done:

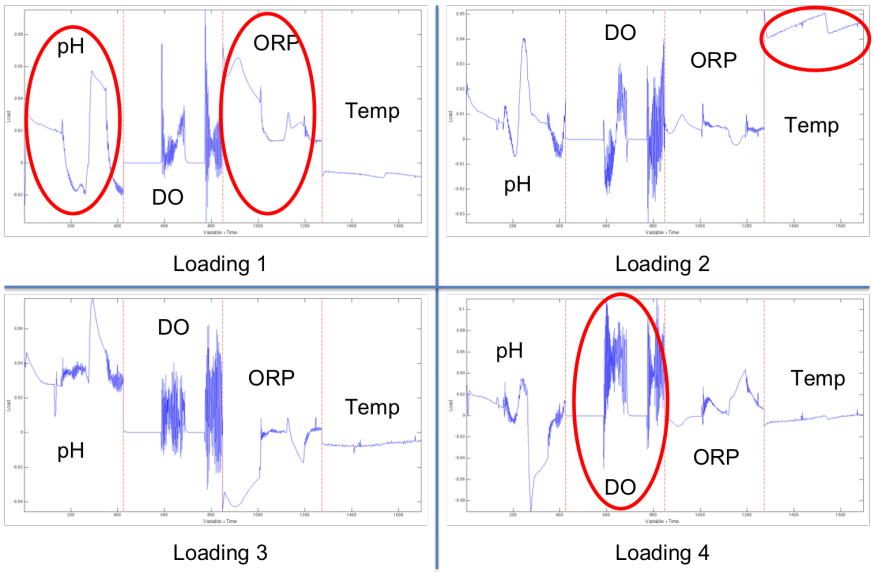


Figure 4.7: Loadings of the 4 principal components retained with the variables represented in each one

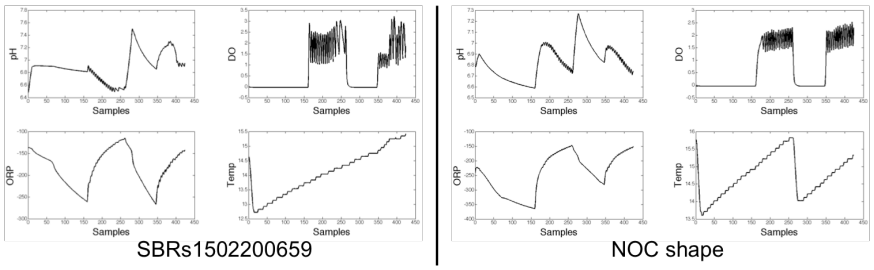


Figure 4.8: Outlier case (left) and mean shape of the NOC processes (right) of the SBR plant

- Distance in the NOC region. First, the Q distance for every new case using Equation 3.3. From all cases in the case base, only will be retained the k_1 cases with lesser distance. From the remaining cases, the T^2 similarity will be computed using Equation 3.4. Finally, only the k_2 cases with a lower value using this distance will be used to predict the class of the new case.

- Similarity in the Q region. This distance criterion consists in computing the Q distances using Equation 3.3 and then, retain those k_1 cases with a minimum distance of the new case. Next, the distance in the principal

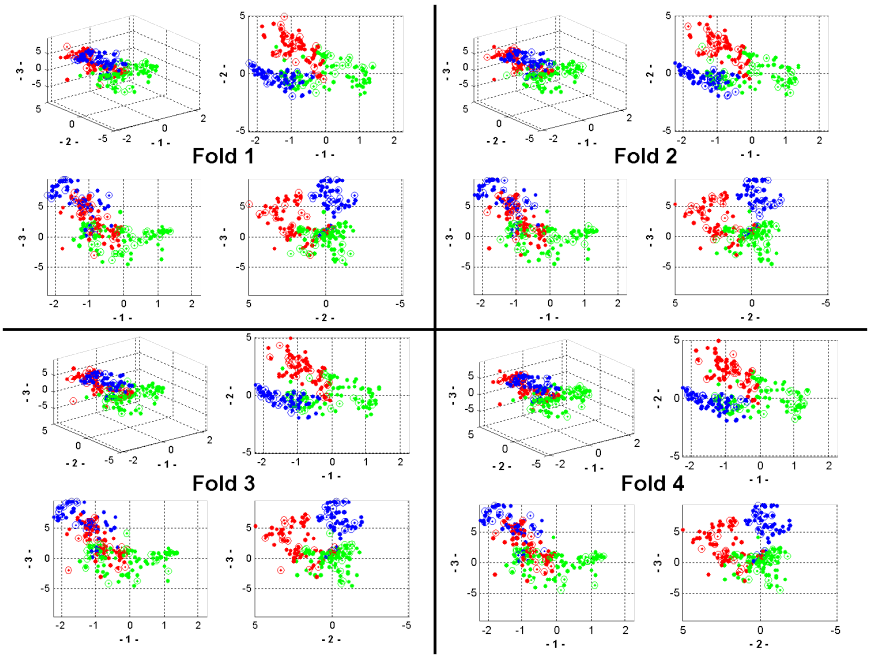


Figure 4.9: Distribution in the PC space of all the folds the case base has been divided the SBR data

components of the remaining cases is computed using Equation 3.2. Finally, the k_2 most similar cases will be used to predict the class of the new case.

- Similarity in the T^2 zone. First, the distance to the centre of the model is computed using Equation 3.4. Then the k_1 nearest cases are selected, and its similarity to the new case is computed using Equation 3.2. Finally, the k_2 nearest cases will be used to determine the class of the new case.

When comparing all the results obtained with each distance criterion, it was observed that all where the same. Those results are presented in Table 4.3, where k_1 stands for the number of neighbours kept in the first step of the combined distances, k_2 is the number of neighbours kept in the second level of the combined distances. TP are the True Positive classification, FN stands for the False Negative cases, FP are the False Positive ones and TN are the True Negative classifications that were presented in Subsection 4.1.2. ACC, SEN, PRE and SPC are respectively the accuracy, sensitivity, precision and specificity that were presented in Subsection 4.1.2. Finally, AUC is the Area Under the ROC curve computation that was presented in Subsection 4.1.4.

If the same results are obtained with several distance criteria, it means that the number of neighbours kept is more important than the distance itself. Two

(k_1, k_2)	TP	FN	FP	TN	ACC	SEN	PRE	SPC	AUC
(20,5)	85	5	8	131	0.943	0.913	0.946	0.964	0.9840
(20,3)	88	4	5	132	0.960	0.946	0.958	0.970	0.9870
(20,1)	92	2	1	134	0.987	0.989	0.980	0.985	0.9874
(15,5)	83	6	10	130	0.930	0.891	0.932	0.956	0.9647
(15,3)	84	6	9	130	0.934	0.903	0.935	0.956	0.9613
(15,1)	88	2	5	134	0.969	0.945	0.978	0.985	0.9656

Table 4.3: Classification results over the SBR plant

main options can be the cause this situation: the groups in the classification space are clearly separated or the method to determine the class of a new case has produced this effect. The distribution of the Q and T^2 values has studied to find if classes were clearly separated. The distribution of the first fold is presented in Figure 4.10. In here, it can be seen that exists intersections between the different classes, what means that the cause of the similarity of the results is due to the reuse function.

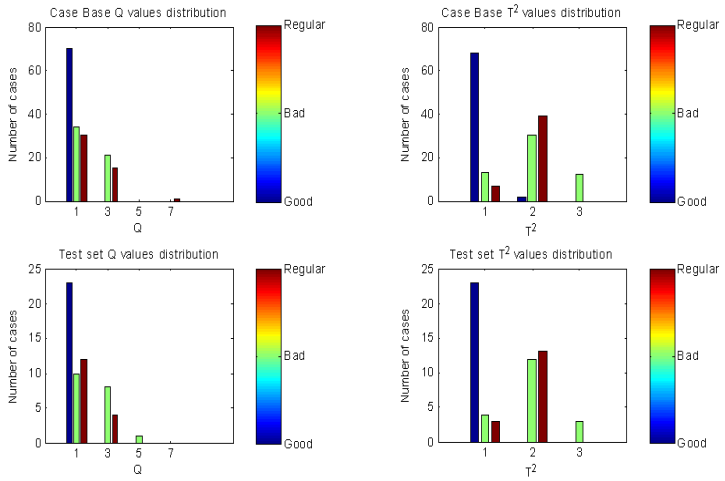


Figure 4.10: Distribution of cases grouping by Q and T^2 values in the training and test sets

Analysing the results obtained, the best classifier is the one with the greatest AUC (in this case the configuration $k_1 = 20$, $k_2 = 1$). It also presents the best classification indices, although this is not mandatory. This can be observed graphically when representing the ROC curves of all tested classifiers in Figure 4.11, because this classifier is the nearest classifier to the point (0,1) at all points.

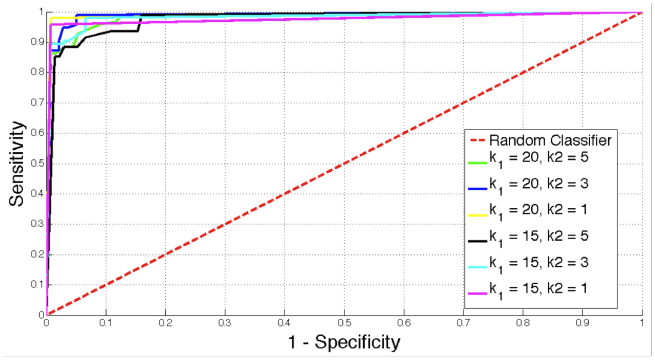


Figure 4.11: ROC curves of all tested classifiers in the WWTP field

Name	t_1	t_2	t_3	t_4	Q	T^2	l	d
SBRs1502200671 (MC)	-0.309	0.542	-0.210	-5.026	0.778	0.528	2	–
SBRs1502200663 (NN)	0.190	0.604	-0.448	-4.990	0.610	0.525	1	0.558
SBRs1201200561 (MC)	-0.957	-0.409	1.215	0.501	0.804	0.197	3	–
SBR1201200563 (NN)	-0.789	0.112	0.261	0.474	0.659	0.092	1	1.105
SBRs1502200626 (MC)	-0.087	0.148	4.168	-0.877	2.433	0.610	1	–
SBR1201200554 (NN)	-0.909	-0.577	3.339	0.729	2.259	0.536	3	2.114

Table 4.4: Information of the wrong classified cases and its nearest neighbour

Miss classified cases for this classifier are detailed in Table 4.4, where MC stands for miss classified case, NN for nearest neighbour, from t_1 to t_4 are the four retained principal components values of each, Q and T^2 are the statistical control limits of those cases, l is the associated class label of the case and d stands for the distance between the miss classified case and its nearest neighbour. Finally, when using the basic distances, the number of neighbour retained is the minimum value between k_1 and k_2.

As can be seen, the main cause of the erroneous classification of the cases is that the nearest neighbour is from another class. The main cause of this situation is that only one analytical result was available per test day, and since in one day there were 3 processes of 8 hours, it was decided to assume the same quality for all the processes of the same day. Figure 4.12 is presented the first miss classified case in this field. It can be seen that their shapes are practically the same, but if both shapes are compared with the mean shape of the NOC region cases presented in Figure 4.8, it can be seen that the nearest neighbour class should be consulted with an expert to decide whether is correctly labelled or not. The same criterion can be applied for the other two miss classified cases.

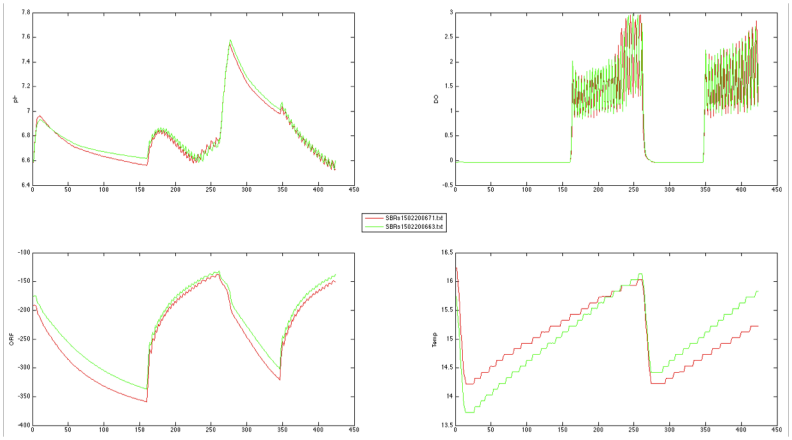


Figure 4.12: Comparison of the first case wrong classified in the SBR process with its nearest neighbour

4.3 Power Quality Monitoring

In this scenario, the methodology is used to determine the origin of voltage sags (upstream or downstream) registered in 25 kV substations of the catalan power network. A voltage sag is the reduction of voltage (10 % and 90 %) during a short time (between 0.5 periods and 1 second). MPCA has been used to model waveforms of voltage and currents of those sags registered by power quality monitors. This allows an enormous reduction of the dimensionality and at the same time the temporal dependency of data is avoided in the projection space.

4.3.1 Relative location of voltage sag in a real power distribution network

The goal in this scenario has been focused on the discrimination between sags originating in the transmission (HV) and distribution (MV) networks. With this aim, sags registered in three 25kV distribution substations have been used as case base. Additionally, the utility has provided information related to the relative origin, upstream (HV) or downstream (MV) from the transformer. More concretely, the classification method is based on the definition of similarity criterion in the projection space obtained when the PCA is applied to sags waveforms. The method proposes the exploitation of the whole information contained in the voltage and current waveforms instead of obtaining features from them. With this goal PCA is used to cope with the dimensionality problem at the same time that it provides statistical indices to assess the quality of projected data in terms of adequacy to the projection model. A graphical representation on this scenario is presented in Figure 4.13, as well as the characteristic shape of the HV and MV voltage sag.

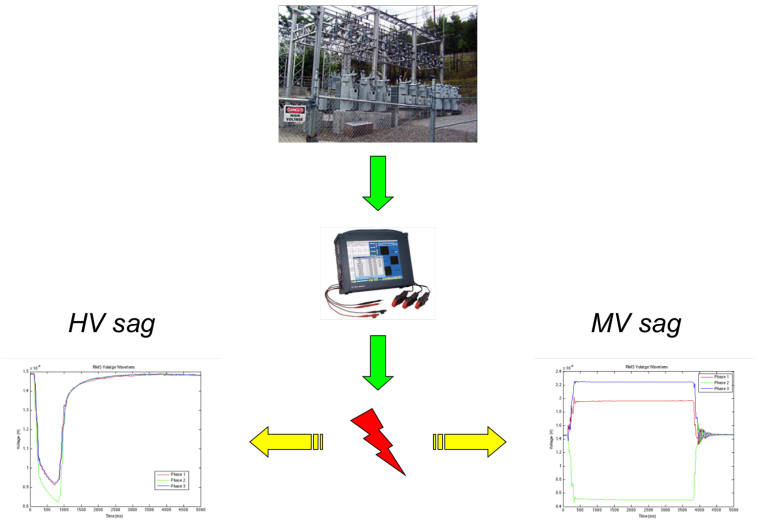


Figure 4.13: Graphical representation of relative location of voltage sags from a measure point, and characteristic shape of the HV and MV voltage sags

4.3.2 Data pre-treatment

The registers of the voltage sags used in this scenario were supplied by the Power Quality Department of ENDESA Distribución. They consist of 221 voltage sags captured during 2004 in a subset of 3 catalan substations, with 140 HV and 81 MV voltage sags. The information contained in those archives is the instantaneous simple and compound voltage and current measures. Due to voltage sags are defined over the RMS values, a Short Fourier Transform (SFT) has been used to calculate the RMS. A one cycle sliding window has been used for this purpose. Figure 4.14 depicts this computation of the RMS value.

After computing the RMS, the number of cases of each class was reduced to 100 HV and 73 MV voltage sags, since there were registers that not presented a pre-fault stage, what would cause misalignment among cases and resulting in a noisy model. An example of a register without pre-fault stage is presented in Figure 4.15.

4.3.3 MPCA model creation

The normalisation procedure used in this scenario is the group scaling, since it rendered the least control limits values and principal component range when

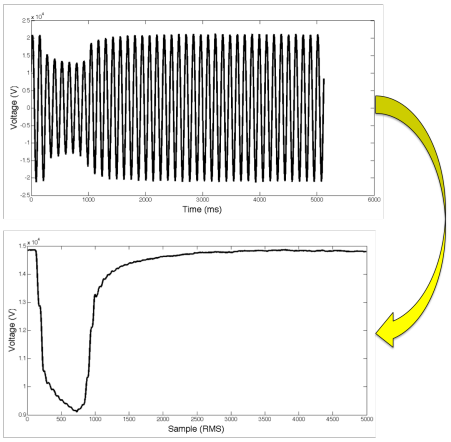


Figure 4.14: Graphical effect of computing the RMS value from the instantaneous ones

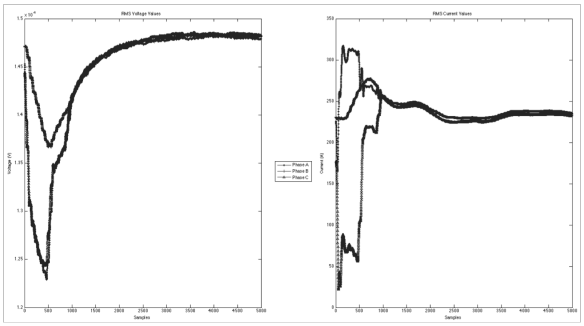


Figure 4.15: Voltage sag without a pre-fault stage

comparing with the auto scaling. In this scenario the HV class will be used to generate the MPCA model, since it presents the least variability of the voltage sags gathered from the 3 substations. Moreover, when comparing the models obtained with a single class with another one created with two classes, the first performs better because the bigger discriminant capability. A comparison between those two models is shown in Figure 4.16.

An example of an outlier case removed in this scenario is presented in Figure 4.17. After removing those cases, 2 principal components have been retained (Figure 4.18 a)). The associated loading vectors are shown in Figure 4.18 b) and c). Although the first principal component explains all the variables, it has been decided to keep an additional principal component by consensus with the other 2 criteria.

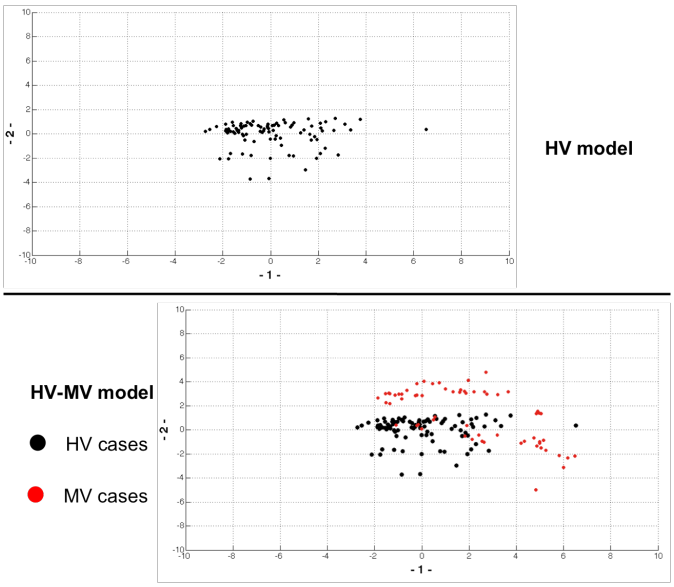


Figure 4.16: Comparison between a MPCA model created with a single class and a model with both classes

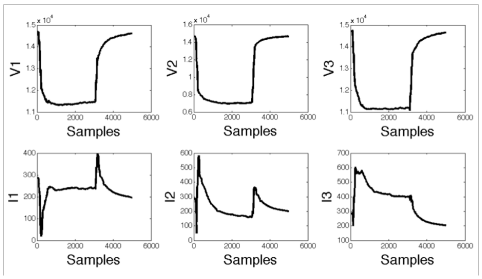


Figure 4.17: Example of an outlier case in the Power Quality Monitoring

4.3.4 Results

In this scenario, after removing all outlier cases and those voltage sags that not have a pre-fault phase, 169 cases are available. The criterion used to decide the number of folds in this field will be that all folds present at least 50 cases. So, the case base will be divided in 3 parts. The distribution in the principal component space is presented in Figure 4.19. As it can be seen, the distribution

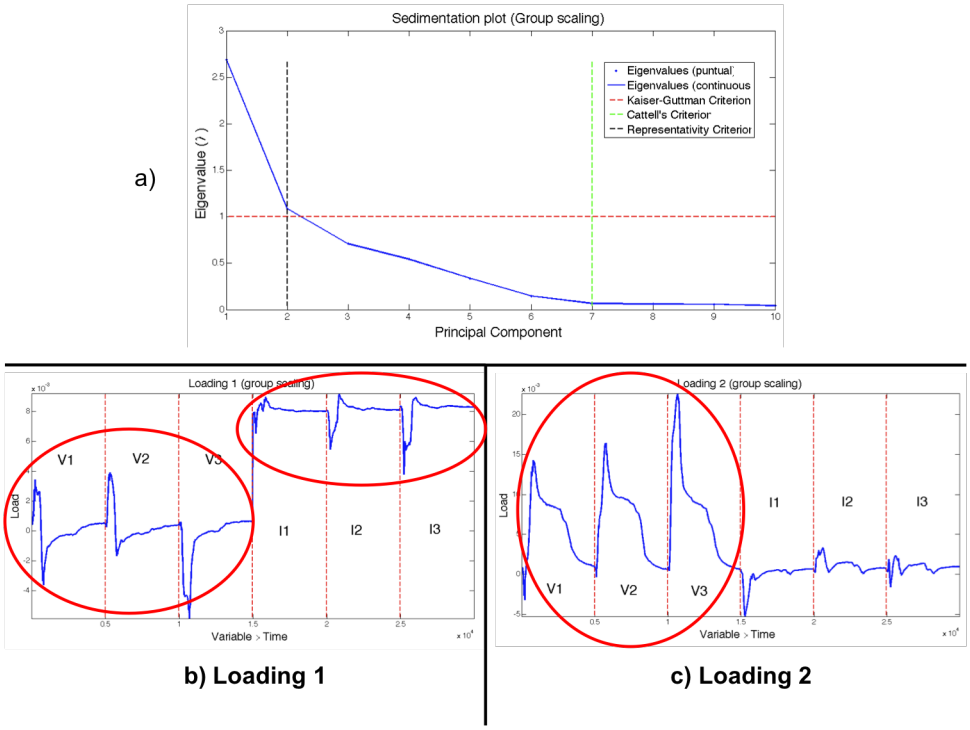


Figure 4.18: Sedimentation plot with the selection of principal components criteria and the associated loading vectors of the retained factors

of the cases and the classes is practically the same.

The distance criteria that have been used for testing the methodology are all distance computations explained in Chapter 3 (distance in the principal components space using Equation 3.2, Q similarity using Equation 3.3 and T^2 similarity using Equation 3.4). Also, the three following combinations of these basic distances will be used:

- Distance in the NOC region. First, the Q distance for every new case using Equation 3.3. From all cases in the case base, only will be retained the k_1 cases with lesser distance. From the remaining cases, the T^2 similarity will be computed using Equation 3.4. Finally, only the k_2 cases with a lower value using this distance will be used to predict the class of the new case.

- Similarity in the Q region. This distance criterion consists in computing the Q distances using Equation 3.3 and then, retain those k_1 cases with a minimum distance of the new case. Next, the distance in the principal

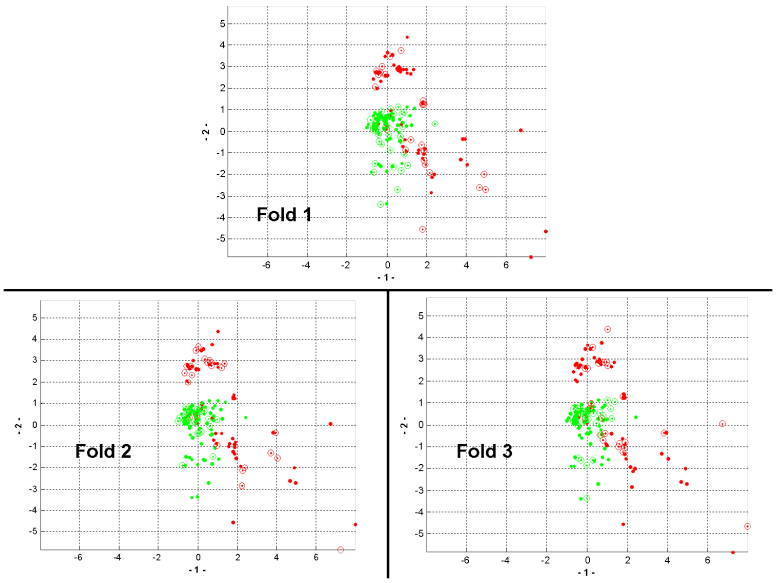


Figure 4.19: Distribution in the PC space of all the folds the case base has been divided

components of the remaining cases is computed using Equation 3.2. Finally, the k_2 most similar cases will be used to predict the class of the new case.

- Similarity in the T^2 zone. First, the distance to the centre of the model is computed using Equation 3.4. Then the k_1 nearest cases are selected, and its similarity to the new case is computed using Equation 3.2. Finally, the k_2 nearest cases will be used to determine the class of the new case.

The results obtained using all these distances were the same and are presented in Table 4.5, where k_1 is the number of neighbours kept in the first level, k_2 is the number of cases retained in the second level of the composed distances. TP, FN, FP and TN stands respectively for True Positive, False Negatives, False Positives and True Negative that were presented in Subsection 4.1.2. ACC, SEN, PRE and SPC are the related performance indices of a Confusion Matrix presented in Subsection 4.1.2 and that stand for accuracy, sensitivity, precision and specificity. Finally, AUC refers to the Area Under the ROC Curve that was explained in Subsection 4.1.4.

The best classifier is the one with the highest value of AUC. In this case, this classifier is ($k_1 = 15$, $k_2 = 5$), although it didn't present the best individual performance indices in all measures. However, in Figure 4.20 is analysed, it can be seen that this classifier presents a more regular classification performance.

(k_1, k_2)	TP	FN	FP	TN	ACC	SEN	PRE	SPC	AUC
(20,1)	93	5	3	68	0.952	0.968	0.950	0.932	0.950
(20,3)	93	4	3	69	0.958	0.968	0.960	0.946	0.974
(20,5)	93	5	3	68	0.952	0.968	0.949	0.932	0.981
(15,1)	92	4	4	69	0.952	0.958	0.959	0.945	0.951
(15,3)	91	2	5	71	0.958	0.947	0.979	0.972	0.978
(15,5)	93	3	3	70	0.964	0.968	0.968	0.958	0.990

Table 4.5: Classification results of the Power Quality Monitoring scenario

On the other hand, the classifier that presents the better individual classification is ($k_1 = 15$, $k_2 = 3$) because has the nearest point to (0,1) in the ROC space. As a general note, it can be also observed that all tested classifications present a good and similar performance because the majority of points are close to the perfect classification zone (0,1).

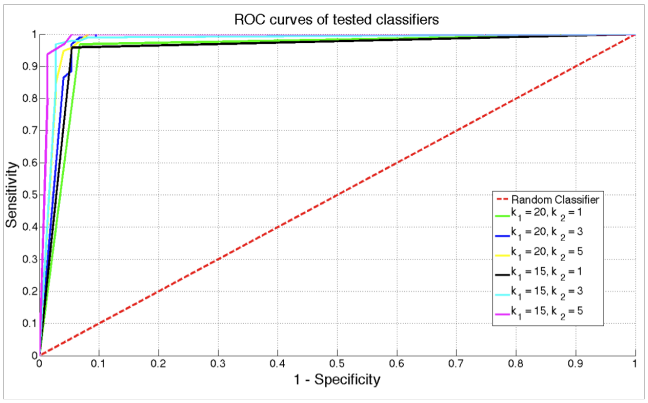


Figure 4.20: ROC curves of all tested configurations of the classifiers in the Power Quality Monitoring

Misclassified MV cases that have been labelled as False Positive are associated to a typology of electrical fault (transformer energising) that presents a similar shape than an HV cases in the voltage waveform, although they present an increase in the current waveform. An example of one of those misclassified cases associated is presented in Figure 4.21.

On the other hand, HV cases misclassified as MV (False Negative) are those cases that present a similar distance between HV neighbours and MV neighbours. As an example, the distances to one of these misclassified cases is presented in Table 4.6, where NN stands for the position among the nearest neigh-

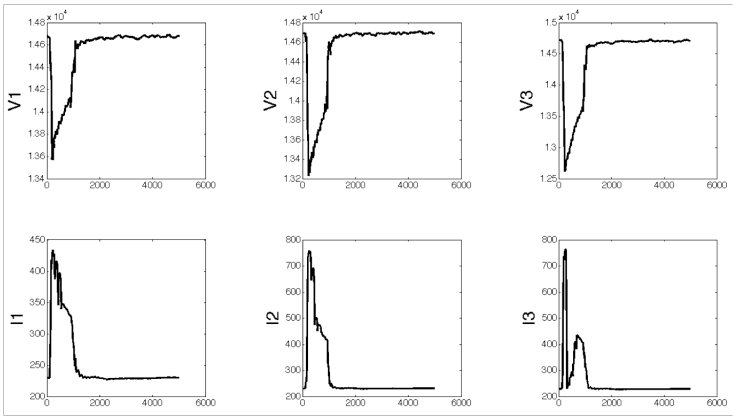


Figure 4.21: Example of a MV case that has been misclassified as an HV one

bours, Distance is the distance between the neighbour and the misclassified case and Class is the relative location of the voltage sag (HV or MV).

NN	Distance	Class
1	0,263	HV
2	0,436	HV
3	0,478	MV
4	0,552	HV
5	0,566	MV

Table 4.6: Nearest Neighbours of an HV misclassified case

A visual interpretation of this situation is presented in Figure 4.22, where the neighbourhood of the misclassified case in the Q-T^2 space is presented.

In this case, if a voting computation had been used, those cases would be classified correctly. Another possibility maintaining the actual method would be change the threshold value, but as can be seen in Figure 4.20 none of the classifiers presented a point with all cases classified correctly.

Finally, the last point to check is the reason because all distance criteria gave the same results. Two can be the main causes: classes are clearly separated or the class determination procedure. It can be seen that in Figure 4.23, actually there is intersection in the grouping of Q and T^2 values of both classes, so the responsible of obtaining the same results using all distances is the reuse function of the CBR.

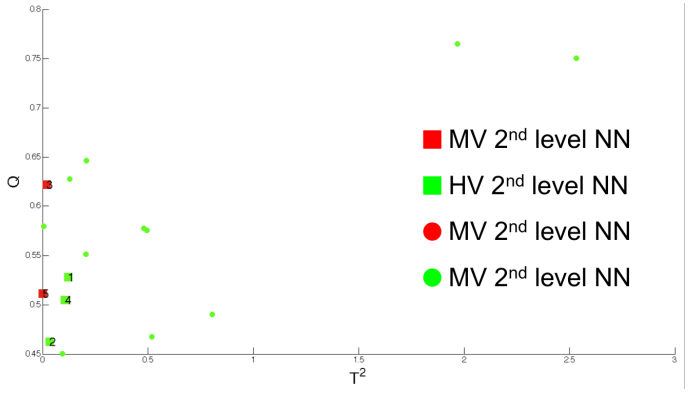


Figure 4.22: Distribution of a misclassified HV case in the Q-T^2 space and its nearest neighbours

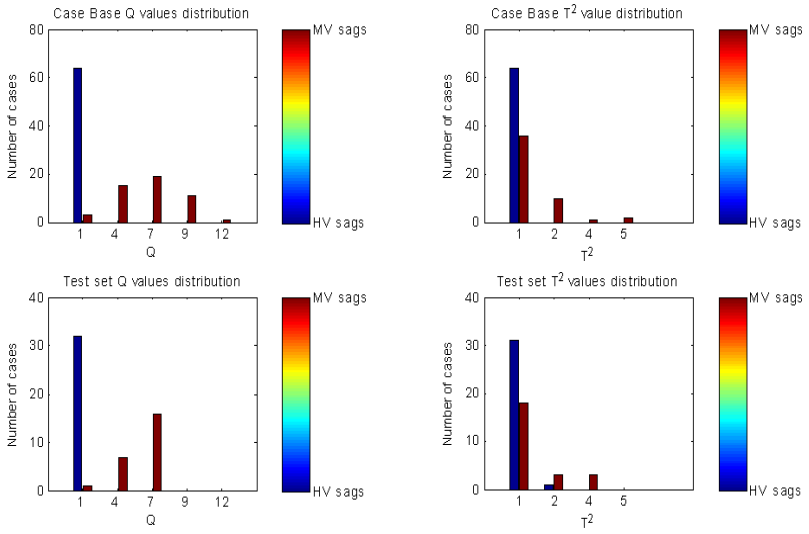


Figure 4.23: Distribution of cases grouping by Q and T^2 values in the training and test sets

4.4 Plastic injection moulding process

In this scenario, the methodology will be used in order to detect and diagnose fault occurrence of an injection moulding. First, a brief introduction about injection and moulding is done. Then the data pre-treatment applied to the original data set will be explained, followed by the steps carried to create the MPCA model used for fault diagnosis. Finally, some results in this scenario will

be explained.

4.4.1 Fault detection and diagnosis of a plastic injection moulding machine

This scenario is included in the EMOLD project (COLL-CT-2006-030339). The aim of this project is to propose a new concept for the plastic injection moulding industry: make moulds a networked element that can be accessed in real time to correct any deviations using embedded knowledge. This knowledge will be based on the information captured from the injection process sensors and experts. The main advantages of this project are mainly related with the improvement of the life cycle of the process.

More concretely, the injection moulding machine studied is presented in Figure 4.24. This process has 24 sensors that collect information of several temperatures, pressures and cylinder positions that will be used to determine the correct evolution of the injection.

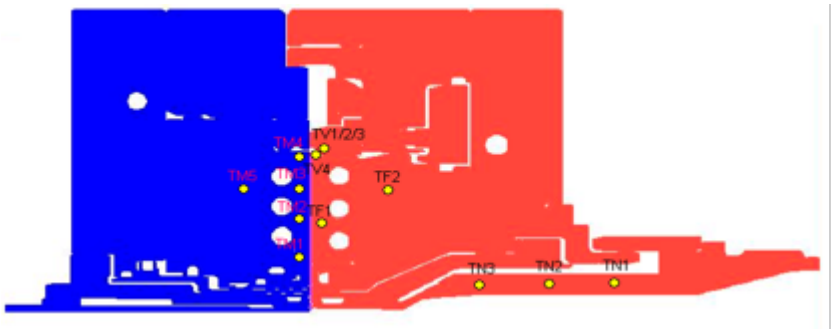


Figure 4.24: Plastic injection process studied

Four different faults have been defined in sensors:

- Gain error in sensors (T1).
- Saturations (T2).
- Loss of signal (T3).
- Blass error in sensor (T4).

Those misbehaviours have been created artificially by modifying some registers associated to real injections during normal operation.

The normal operation condition cases have been used to create the MPCA model. Whereas the modified ones are used in the test phase. The goal is to

observe regions in the projection space associated with specific misbehaviours that make feasible the use of neighbourhoods for fault detection and diagnosis.

After studying each combination of variable and fault type, it can be discriminated 4 regions of detectability:

- Non detectable faults (ND). Neither the principal components, the Q-residual nor the Hottelling's T^2 does not present a major trend for those faulty sensors.
- Faults detectable with Hottelling's T^2 (DT^2). As its name state, those faults present a higher value of T^2 for those injections with faulty sensors considered not in the NOC region.
- Q-statistic detectable faults (DQ). As the previous one, Q-statistic presents higher values for the faulty sensors.
- Full detectable faults (D). This typology not only presents variations on both control statistics, it also presents a major trend in a subset of principal components.

4.4.2 Data organisation

The original data set was formatted by a two dimensional matrix where rows represent all time instants registered in the file, multiple injections, and columns represent all measured variables. The dimensions of this data matrix are 32190 rows and 24 columns. Variable names will be labelled from V_1 for the first variable to V_{24} for the last variable. In order to obtain a model for the injections it is needed to identify the beginning and end of each injection and build the 3-D matrix. Special interest in this step is because the need of data perfectly aligned (correspondence among significant instants of different injections) when building the 3-D matrix. One of the most critical variables of the process, temperature sprew, represented by V_{16}, has been used to identify the beginning of each injection and the duration. The resulting division of the original data is presented in Figure 4.25. The resulting dimensions of the 3D matrix are 38 time instants per injection, 24 variables and 838 injections, or what is the same (38 $\times$ 24 $\times$ 838).

Since all variables presented an important variability after dividing the original data into injections, those with the lowest variation in the y-axis were erased after verifying their low information gain when building the statistical model. The remaining variables that will be used in the MPCA methodology can be seen in Figure 4.26. The total number of variables retained is 15, that will be labelled from X_1 to X_{15} and the new 3D matrix will be of the following dimensions: 38 time instants $\times$ 15 variables $\times$ 838 injections.

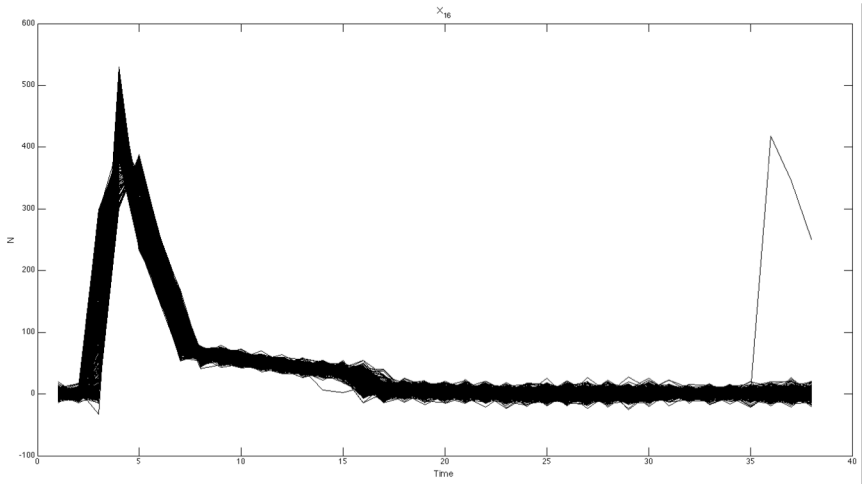


Figure 4.25: Division of the data in injections using V_{16}

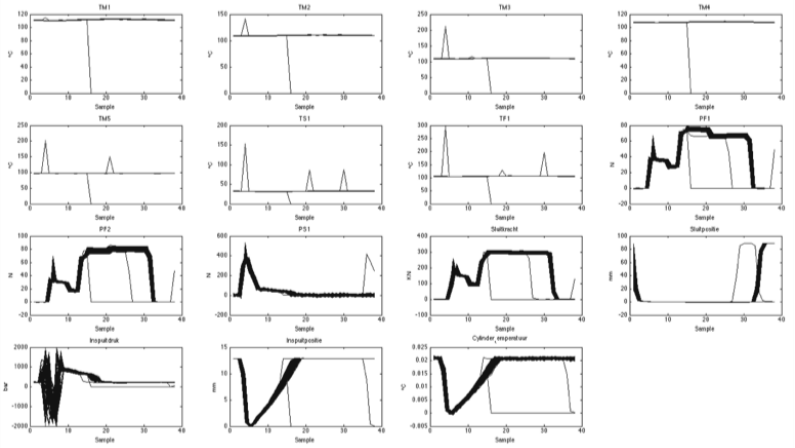


Figure 4.26: Variable retained for the MPCA model creation

4.4.3 MPCA model creation

The unfolding procedure that will be applied in this scenario because no qualitative data is available about the process will be the batch-wise.

The cases that will be used to create the MPCA model will be the whole set of available cases, and the statistical control limits will be used to detect undesired behaviours in the model. Because temperature and pressure present

different value ranges and variability, a normalisation step is needed. In this scenario, the auto scaling and the group scaling procedures gave different solutions as can be seen in Figure 4.27. In fact, the group scaling procedure couldn't eliminate the non-linear behaviour of the variables, reflected in a first principal component with a great amount of variability (Table 4.7) and at the same time, in the second principal component an inverted U distribution of cases can be observed [61]. So the auto scaling procedure will be the one used in this scenario.

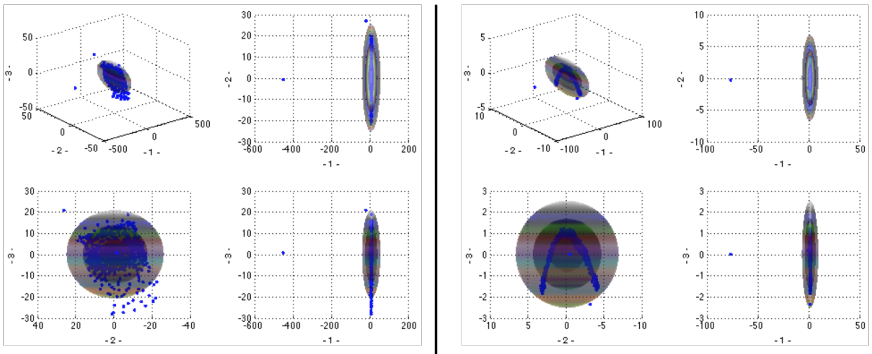


Figure 4.27: Comparison of the models obtained with the auto scaling and group scaling procedures

PC	λ	% Global Variance explained	% Global Variance accumulated
1	7.20	48.03	48.03
2	5.05	33.67	81.70
3	0.69	4.66	86.37
4	0.52	3.51	89.89
5	0.44	2.95	92.84

Table 4.7: Group scaling information related to the first 5 principal components in the plastic injection scenario

Once the normalisation step has been fixed, the next step is to find all outlier cases, that in this case will be associated to the AOC region. In this scenario, two of the injections labelled as outliers were too short injections where detected using the T^2 statistic as it is shown in Figure 4.28.

The values of these too short injections (Processes 128 and 838) are presented in Figure 4.29. The rest of outlier cases are the misaligned cases that could be observed in Figure 4.26.

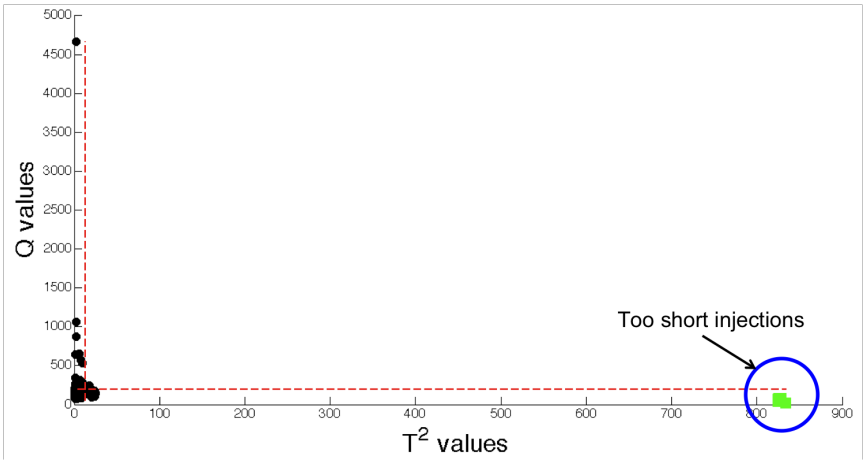


Figure 4.28: Detection of the too short injections in the Q-T^2 space

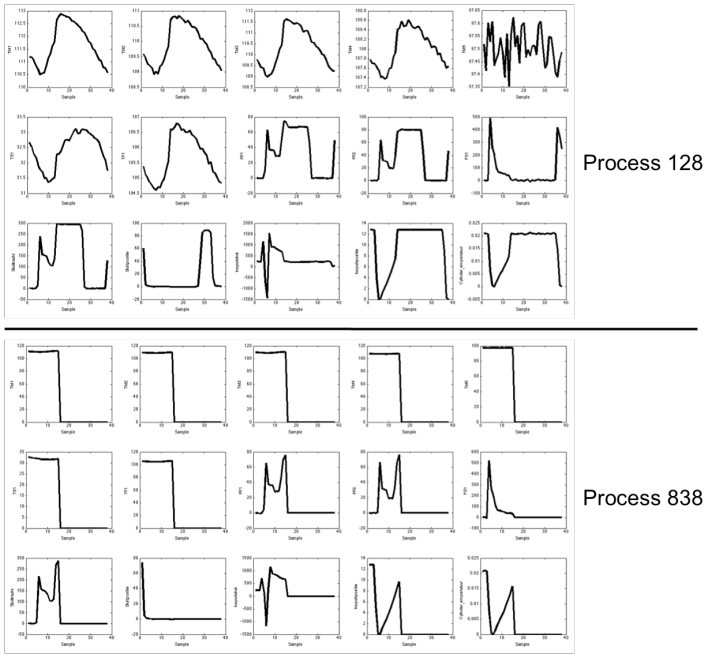


Figure 4.29: Two processes too short that were detected using the statistical control limits

When outliers have been removed from the training set, the last task is to decide the number of principal components to retain. In this case, the number

has been fixed to 3. According to Figure 4.30 b), most of the variables are explained in the first principal component. However, until loading vector 3, the remaining variables are not explained. The graphical representation to select the number of principal components is shown in Figure 4.30 and the loading vectors of the retained principal components are presented from 4.30 b) to d). The global variation explained by the model is 45.72 %. The residual percentage of this model can be divided in more information of the sensors (the procedure to select the number of components stops when all variables are represented) and noise in the processes.

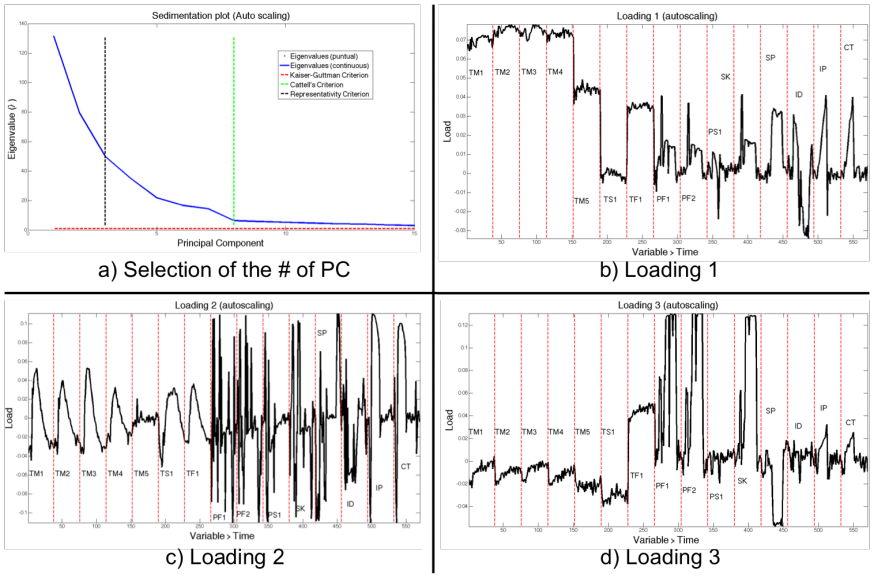


Figure 4.30: Graphical representation of the criterion used to select the number of principal components and the respective retained loading vectors

4.4.4 Results

After removing all outlier cases, 830 injections are available to evaluate the methodology in this scenario. As it was mentioned previously, 50 faulty cases for each combination of typology and type of sensor has been created and studied. In this section, and since the great number of combination that exists (15 variables $\times$ 4 fault typologies = 60 types of faults) only a subset will be studied using all types of faults presented previously:

- Fault in sensor 5 is full detectable (FD) whence the typology T1 and T3 are given, while for T2 it is non detectable (ND) and for T4 it is only

detectable using the Q statistic (DQ).

- Fault in sensor 6 is full detectable (FD) when T3 happens, it is detectable by Q (DQ) in T4 and T1 and it is non-detectable (ND) by T2.
- Fault in sensor 8 is detectable by the Q statistic (DQ) in all types of fault.
- Fault in sensor 12 is detectable using the Q statistic (DQ) in all types of faults, although in type T4 it also is detectable by T^2 (DT^2).
- Fault in sensor 15 is non-detectable (ND) for types T1 and T2, it is detectable using the Q statistic (DQ) when occurs type T3 and it is full detectable (FD) on type T4.

Table 4.8 presents the detection type (FD, ND, DQ, DT^2) for every combination of type of fault (T1, T2, T3 and T4) and sensor faults.

Faulty sensor	T1	T2	T3	T4
1	FD	ND	FD	DQ
2	FD	ND	FD	DQ
3	FD	ND	FD	DQ
4	FD	ND	FD	DQ
5	FD	ND	FD	DQ
6	DQ	ND	FD	DQ
7	FD	ND	FD	DQ
8	DQ	DQ	DQ	DQ
9	ND	DQ	DQ	ND
10	ND	ND	DQ	ND
11	ND	DQ	DQ	ND
12	ND	ND	DQ	DQ + DT2
13	ND	ND	DQ	ND
14	ND	ND	DQ	DQ
15	ND	ND	DQ	FD

Table 4.8: Type of detection of each faulty situation grouped by type of fault in the moulding injection process

In order to test the methodology proposed in this Master Thesis, 3 tests with three different case bases will be conducted. At every test, the full application of the validation methodology will be carried. The resulting cases bases will be projected to the same NOC region model explained before. The case bases used for testing are:

- Combination of all typologies (T1, T2, T3 and T4), sensors faults (fault in sensor 5, 6, 8, 12 and 15) and all NOC region cases (CB1).
- Combination of all typologies of faults (T1, T2, T3 and T4) for a given sensor fault. In this case, fault in sensor 15 has been selected because presents the most variability of detection types (FD, ND , DQ and DT^2) (CB2).

- Combination of all sensor faults (fault in sensor 5, 6, 8, 12 and 15) for a determined typology. The selected typology has been fixed to T4 because it presents all possible detection typologies (FD, ND, DQ, DT^2) (CB3).

In this scenario, since a large number of cases are available, each of the case bases will be divided in 10 folds (n = 10) to study the performance of the classifier. The distance criteria that will be used to test each case base will be all the basic distances exposed in Chapter 3: distance in the principal components space (Equation 3.2), Q similarity (Equation 3.3) and T^2 similarity (Equation 3.4). Also, three additional distances will be used, resulting of the combination of theses basic distances:

- Distance in the NOC region. First, the Q distance for every new case using Equation 3.3. From all cases in the case base, only will be retained the k_1 cases with lesser distance. From the remaining cases, the T^2 similarity will be computed using Equation 3.4. Finally, only the k_2 cases with a lower value using this distance will be used to predict the class of the new case.

- Similarity in the Q region. This distance criterion consists in computing the Q distances using Equation 3.3 and then, retain those k_1 cases with a minimum distance of the new case. Next, the distance in the principal components of the remaining cases is computed using Equation 3.2. Finally, the k_2 most similar cases will be used to predict the class of the new case.

- Similarity in the T^2 zone. First, the distance to the centre of the model is computed using Equation 3.4. Then the k_1 nearest cases are selected, and its similarity to the new case is computed using Equation 3.2. Finally, the k_2 nearest cases will be used to determine the class of the new case.

Table 4.9 presents the results obtained with CB1, where k_1 stands for the neighbours retained in the first level of the composed distances and k_2 the neighbours retained in the second level of the compound distances. TP, FN, FP and TN are respectively True Positives, False Negatives, False Positives and True Negative and were presented in Subsection 4.1.2. ACC, SEN, PRE and SPC are the confusion matrix presented in 4.1.2 and stand for accuracy, sensitivity, precision and specificity. Finally, AUC is the Area Under the ROC curve computed as it was presented in Subsection 4.1.4. The same abbreviations will be used for all three case bases, and when computing the basic distances, the minimum values between k_1 and k_2 is the number of neighbours to keep.

The first thing to point in CB1 is that the number of cases in the AOC region (all faulty situations) is greater than the number of cases within the NOC region. This has led to a classifier that is better predicting faulty cases than the ones in the NOC region. This can be observed since specificity is always greater than sensitivity. When observing the different values of AUC, it can be observed that the greater the value of k_2, the better the classification is. At

(k_1, k_2)	TP	FN	FP	TN	ACC	SEN	PRE	SPC	AUC
(20,1)	610	192	220	1008	0.797	0.734	0.760	0.84	0.787
(20,3)	625	192	205	1008	0.804	0.753	0.765	0.84	0.870
(20,5)	645	168	185	1032	0.826	0.771	0.794	0.86	0.891
(15,1)	605	187	225	1013	0.797	0.728	0.763	0.844	0.786
(15,3)	642	193	188	1007	0.812	0.773	0.770	0.839	0.873
(15,5)	643	168	187	1032	0.825	0.774	0.793	0.86	0.896

Table 4.9: Classification of CB1 in the injection moulding process

the same time, the lower k_1 is, the better results in classification are attained. This combination of events can only be observed in the compound distances. However, the same results where obtained with all distance criteria. The best classifier using this case base is (k_1 = 15, k_2 = 5), but it didn't had the best values in all performance indices. This can be observed visually in Figure 4.31, where it can be seen that this is the most regular classifier, but the classifier (k_1 = 20, k_2 = 5) has the point nearest to the point (0,1).

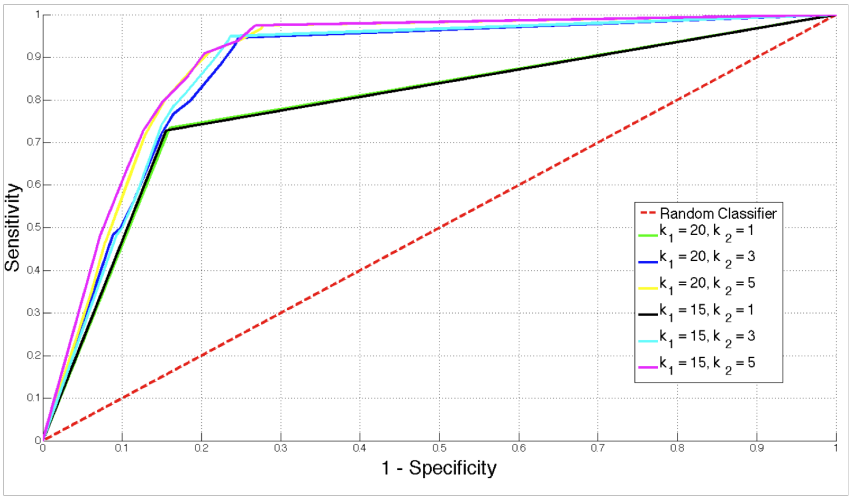


Figure 4.31: ROC curves of the CB1 case base tested classifiers

Table 4.10 presents the results obtained over the case base CB2. The abbreviations used in this table are the same that where presented for CB1. In this case, results obtained with all the distance criteria ended in the same values.

In this case base, the situation in CB1 has been reverted: there are more cases in the NOC region than in the AOC region. So, these classifiers will tend

(k_1, k_2)	TP	FN	FP	TN	ACC	SEN	PRE	SPC	AUC
(20,1)	769	61	61	139	0.881	0.926	0.928	0.695	0.8107
(20,3)	773	64	57	136	0.882	0.931	0.923	0.68	0.8527
(20,5)	789	57	41	143	0.904	0.950	0.932	0.715	0.8785
(15,1)	764	57	66	143	0.880	0.920	0.930	0.715	0.8177
(15,3)	785	59	45	141	0.899	0.945	0.930	0.705	0.8604
(15,5)	792	54	38	146	0.910	0.954	0.936	0.73	0.8781

Table 4.10: Classification of CB2 in the injection moulding process

to classify correctly NOC cases than AOC cases. This can be seen in that sensitivity is greater than the specificity. Also it can be observed that the overall classification ratio has increased, because accuracy and AUC values are greater than before. In this case base the best classifier is ($k_1 = 20$, $k_2 = 5$), although ($k_1 = 15$, $k_2 = 5$) presents almost the similar AUC values but higher individual performance indices. In Figure 4.32 it can be seen that both classifiers are very close, but at some points ($k_1 = 15$, $k_2 = 5$) it is slightly nearer to point (0,1).

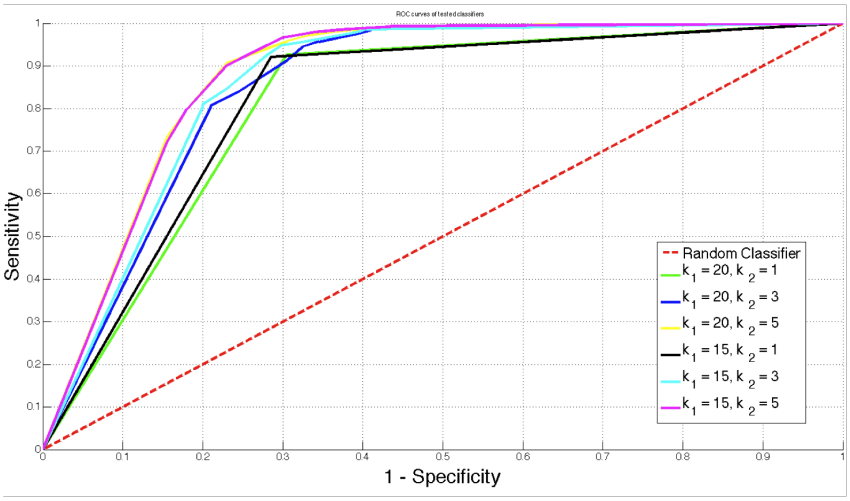


Figure 4.32: ROC curves of the CB2 case base tested classifiers

Finally, Table 4.11 presents the classification results of CB3. The abbreviations used in this table are the same than the tables of the other two cases bases (CB1 and CB2). The same results were obtained using all distance criteria (basic and composed).

(k_1, k_2)	TP	FN	FP	TN	ACC	SEN	PRE	SPC	AUC
(20,1)	811	11	19	289	0.973	0.977	0.986	0.963	0.970
(20,3)	801	13	29	287	0.962	0.965	0.984	0.958	0.986
(20,5)	801	7	29	293	0.968	0.965	0.991	0.976	0.987
(15,1)	811	12	29	288	0.972	0.977	0.985	0.96	0.968
(15,3)	802	10	28	290	0.966	0.966	0.987	0.966	0.986
(15,5)	799	8	31	292	0.965	0.962	0.990	0.973	0.989

Table 4.11: Classification of CB3 in the injection moulding process

In this last case base, the number of NOC is greater than the AOC cases, as happened in CB2. Also, AUC values of this case base are greater that the ones obtained in CB1 and CB2. Also it can be observed that in this cases, the classifier identifies correctly most of the NOC cases and AOC. This can be observed in that sensitivity and specificity are quite similar in all cases. The best classifier in this case base is ($k_1 = 15$, $k_2 = 5$) because it has the highest AUC value, although it does not have the best individual scores. In Figure 4.33 it can be observed that all classifiers are very close to point (0,1) with only little differences. As a main conclusion it can be said that its easier to distinguish between fault sensors given a fault typology (T4 in this case), that try to identify the type of fault that a sensor presents.

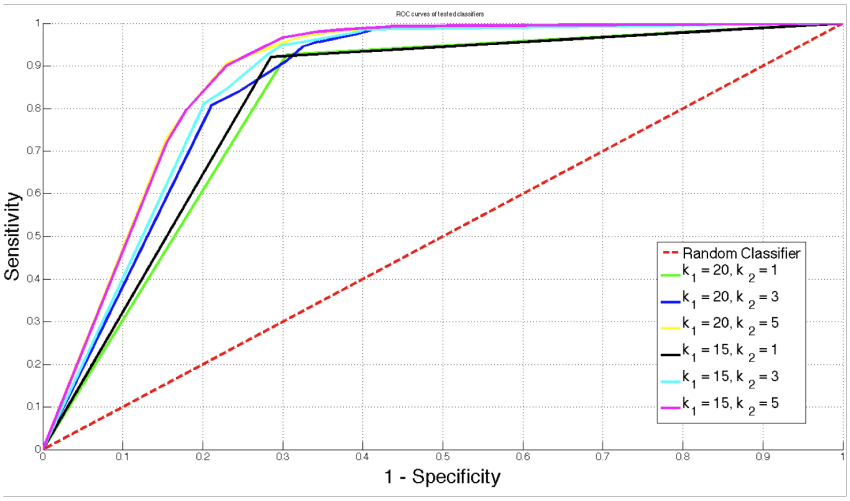


Figure 4.33: ROC curves of the CB3 case base tested classifiers

The last point to check is why all distance criteria gave the same results. Two are the main possibilities: classes are clearly separated, or it is because of

the way the new class is predicted. According to Figure 4.34, the distribution values of the Q statistic are not clearly separated, so the main cause of all distance criteria gave the same results is the reuse function of the CBR.

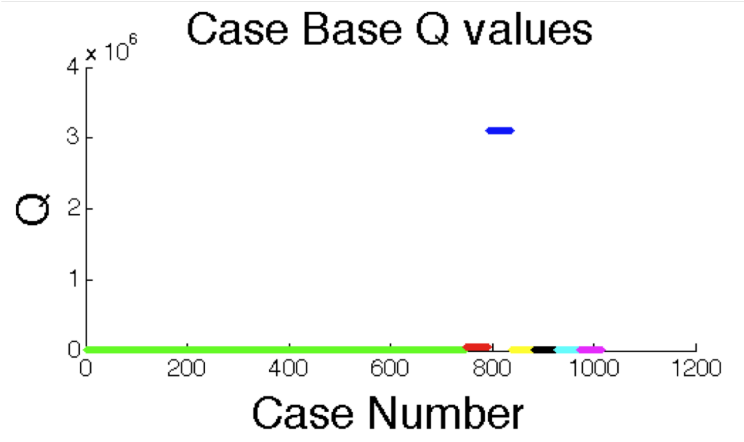


Figure 4.34: Distribution of the Case Base values of Q grouped by their class

Chapter 5

Conclusions and further work

In this final chapter the main conclusions, contributions and some related publications of this Master Thesis are exposed firstly, followed by the future tasks to improve the methodology exposed in this document will be stated.

5.1 Conclusions

The increasing overload of information actual processes throw away makes impossible to deal by means of well known but antiquate control techniques. The multivariate control of industrial processes is becoming one of the main attractions nowadays because its capacity to show the information of the monitored process in a graphical and simplified way. In fact, one of the most important handicaps of classical monitoring techniques (redundancy and dependency of information) is used to find and understand the latent variables that governs the process, that are not intuitively detectable.

Multivariate techniques are capable of detecting abnormal situations without the intervention of an expert in the process with the only necessity of historical data captured from the process, specially when this data comes from variables controlling the process in spite of the ones that control the final quality of the product. This information will be used to compute some control statistics that will detect processes that would generate an out of specifications product. Whenever a process is labelled as AOC, the procedure provides control charts that will show the variables related to the cause of this situation. Once the variable responsible of the problem are pointed, the help of an expert in the process will be needed in order to solve the problem, as well as to specify the root cause of the problem.

Taking into account this limitations and also that sometimes only few past

situations are available, the completion of the multivariate control of process with a CBR approach is proposed and taken to several implementation fields. The main advantage of applying a CBR in the projection space given for the MPCA methodology is that independence between of the new attributes is granted. This is the key point for all machine learning algorithms, since models obtained with this requirement present a better performance and are simpler. Another important advantage of this combination is that CBR can be dedicated exclusively to determine the type of fault that occurred, since NOC region cases can be detected using the statistical control limits of the MPCA model. Moreover, the CBR can complement the MPCA model in the borderline cases that can be easily misclassified because the proximity of a majority in the outskirts of faulty situations.

The results obtained in this work, as well as previous ones conducted in the research group eXiT, encourages the exploitation of this combined methodology to control finite duration processes, showing a constant behaviour of the methodology. Actually, despite the fact that in the Power Quality and WWTP the number of cases was not too large, the correct classification ratio was above 0.95 of AUC, which means the classifier tends to correct classify the majority of cases. The main modifications done to the original methodology presented in [8] are:

- The usage of the Stratified n-Fold Cross Validation over the complete set of cases to measure more accurately the behaviour of the classifier.
- The computation of the Confusion Matrix and its associated statistics to evaluate how the models misclassifies cases.
- The utilisation of the AUC as an objective evaluation method of tested classifiers.
- The modification of the reuse function to classify new cases arrived. However, this change has provoked that the number of nearest neighbour has became much more important than the distance criterion employed.

The published papers related to the methodology explained in this Master Thesis are:

- J. Meléndez, **X. Berjaga**, S. Herraiz, J. Sánchez and M. Castro. Classification of Voltage Sags based on k-NN in the Principal Component Space. International Conference in Renewal Energies and Power Quality (ICREPQ). Santander, Spain. 12-14 March, 2008.
- J. Meléndez, **X. Berjaga**, S. Herraiz, V. Barrera, J. Sánchez and M. Castro. Classification of sags according to their origin based on the waveform similarity. IEEE PES Transmission and Delivery Latin America, Transmission and Distribution Conference and Exposition. Bogotá, Colombia, 13-15 August, 2008.

- **X. Berjaga**, J. Meléndez and A. Pallarés. Statistical Monitoring of Injection Moulds. Congrés Català d'Intelligència Artificial (CCIA). St. Martí d'Empúries, Girona, Spain. 22-24 October, 2008.

A collaboration in a lateral publication can be found in:

- **X. Berjaga**, J. Meléndez and A. Pallarés. A Case-Based Centred Approach for Rapid Manufacturing: Definitions. Hybrid Intelligent Systems (HIS). Barcelona, Spain. 10-12 September, 2008.

And another paper related to the work done in the Power Quality Monitoring submitted and pending of acceptance is:

- V. Barrera, **X. Berjaga**, J. Melendez, S. Herraiz. Two New Methods for Voltage Sag Source Location. 13th International Conference on Harmonics and Quality of Power (ICHQP), Australia. 28th September 1st October, 2008.

5.2 Further work

The methodology explained in this Master Thesis has obtained good results in all the application fields it has been tested, specially in the Power Quality Monitoring (PQM) and plastic injection field. Therefore, this encourages the improvement of the combined approach of MPCA and CBR. Since this methodology is divided in two steps (first a statistical model is created and then the CBR is applied for fault diagnosis and as a complement for fault detection), the future tasks to conduct in both methodologies are presented separately.

5.2.1 Future improvements for the statistical model creation

In the WWTP and PQM scenarios the availability of the so-called quality variables, that is, the evaluation of the final state of the process that can only be computed once the process has ended, as can be seen in Figure 5.1. In the WWTP scenario, this quality variable are related to the analytical results of the batches used to test the scenario, and in the PQM scenario it is the relative location of the voltage sag. This information has been considered in the CBR step, more concretely, as the attributes used to dived the original space of all processes in different classes. PCA is a technique for dimensionality reduction that maintains the majors trends in the original data space (as commented in Chapter 2), but it is not oriented for classification purposes. In fact, there are some methodologies based on the principal components decomposition that take into account some discriminant variables (in this case the quality variables) to find the best composition of the original variables to separate between classes, like Partial Least Squares (PLS) and Principal Components Regression (PCR) among others. So, a future task would be study and compare the utilisation of

those techniques for fault detection and diagnosis, as well as its interaction with the CBR approach.

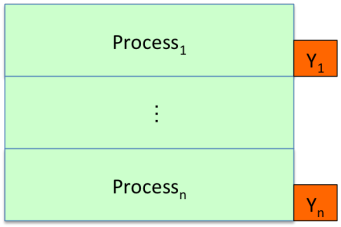


Figure 5.1: Representation of the difference of sample time for process variables and quality variables (Y)

Related with techniques oriented to classification, but this time without using the principal components decomposition procedure, Support Vector Machines (SVM) can be analysed. SVM only takes into account the marginal points, what is, the furthest points from its respective classes that are at minimum distance between them. A graphical example of those cases is shown in Figure 5.2.

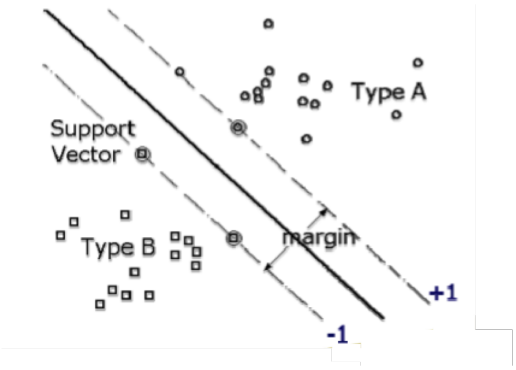


Figure 5.2: Example of marginal points used as a base for SVM

Another thing to point is that all statistical methods for model construction depends on the original data set was used to generate them. Since all processes evolve in time, two main points should be studied. The first one is evaluate how the classification performance is changing to decide if the original model is still

available for the data arriving at a certain moment.

As a last point to focus the future work in this part of the methodology, and taking into account the nature of the different fields (all application fields can be divided in phases and not necessarily all time instants must be used if its gain of information is insignificant), some other approaches can be revised. For example, to deal with processes that are conformed by some phases, the Multi-Phase Principal Component Analysis (MPPCA) can be used, and to determine time dependencies within the same process, Dynamic Principal Component Analysis (DPCA) can be applied.

5.2.2 CBR

As was firstly mentioned in Chapter 3, in this Master Thesis the Retain function has not been implemented, so this is a must in order to analyse the learning degree of the classifier in the applied fields and it is also related with the trust in the actual model. Another point is to study how results change using other approaches for the reuse step, since is the main bottleneck when analysing the results. Using the actual method has reduced the classification of cases to a binary classification. Related with this point, the analysis of the extended confusion matrix (when more than 2 cases have to be chosen from) should be conducted, as also the way to compute the related performance indices.

Bibliography

[1] M. Blanke. What is fault-tolerant control ? *Proceedings of IFAC SAFEPROCESS*, 35:123.126, 2000.

[2] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri. A review of process fault detection and diagnosis part i: Quantitative model-based methods. *Computers and Chemical Engineering*, 27:293–311, 2003.

[3] R. Milne. Strategies for diagnosis. *IEEE Transactions on Systems, Man, and Cyber*, 17(3):333–339, 1987.

[4] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri. A review of process fault detection and diagnosis part ii: Qualitative models and search strategies. *Computers and Chemical Engineering*, 27:313–326, 2003.

[5] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri. A review of process fault detection and diagnosis part iii: Process history based methods. *Computers and Chemical Engineering*, 27:327–346, 2003.

[6] L. Hare. From chaos to wiping the floor web. *http://www.asq.org/pub/qualityprogress/past/0703/58spc0703.html*.

[7] Alberto Ferrer. Control estadístico megavariante para los procesos del siglo xxi. *27 Congreso Nacional de Estadística e Investigación Operativa (ESPAA)*, 2003.

[8] M. Ruiz. *Multivariate Statistical Process Control and Case-Based Reasoning for Situation Assessment of Sequencing Batch Reactors*. PhD thesis, Universitat de Girona, 2008.

[9] A. Wong. Multivariate statistical process control (mspc) applied to a sequencing batch reactor for wastewater treatment. Master's thesis, Universitat de Girona, 2007.

[10] C. Rosen and G. Olson. Disturbance detection in wastewater treatment plants. *Water Science and Technology*, 37(12):3402–3410, 1998.

[11] C.-K. Yoo, K. Villez, I.-B. Lee, and P. A. Vanrolleghem. Multivariate nonlinear statistical process control of a sequencing batch reactor. *Journal of Chemical Engineering of Japan*, 39(1):43–51, 2003.

[12] S. Grieu, A. Traor, M. Polit, and J. Colprim. Prediction of parameters characterizing the state of a pollution removal biologic process. *Engineering Applications of Artificial Intelligence*, 18:559–573, 2005.

[13] K. Singh, A. Malik, D. Mohan, S. Sinha, and V. Singh. Chemotric data analysis of pollutants in wastewater: a case study. *Analytica Chimica Acta*, 532:15–25, 2005.

[14] M. H. J. Bollen. *Understanding Power Quality Problems: Voltage Sags and Interruptions*. IEEE Press series on Power Engineering, 2000.

[15] N. Hamzah, A. Mohamed, and A. Hussain. A new approach to locate the voltage sag source using real current component. *Journal of Electric Power Systems Research*, 72:113–123, 2004.

[16] J. Mora, D. Llanos, J. Meléndez, J. Colomer, J. Sanchez, and X. Corbella. Classification of voltage sags measured in a distribution substation based on qualitative and temporal descriptors. *17th Internation Conference on Electricity Distributions*, 2003.

[17] M. Ruiz, J. Meléndez, J. Colomer, J. Sanchez, and M. Castro. Fault location in electrical distribution systems using pls and nn. *International Conference in Renewal Energies and Power Quality (ICREPQ)*, 2004.

[18] K. Abbas, J. Meléndez, and J. Colomer. Classification of voltage sags based on mpca models. *Lecture Notes in Computer Science*, 4477/2007:362–369, 2007.

[19] K. Abbas, J. Meléndez, and J. Colomer. Multiway principal component analysis (mpca) for upstream/downstream classification of voltage sags gathered in distribution substations. *Studies in Computational Intelligence*, 116:297–312, 2008.

[20] A. T. Bozdana and O. Eyercioglu. Development of an expert system for the determination of injection moulding parameters of thermoplastic materials: Ex-pimm. *Journal of Materials Processing Technology*, 128:113–122, 2002.

[21] C. Collins. Monitoring cavity pressure perfects injection molding. *Assembly Automation*, 19(3):197–202, 1999.

[22] S. Orzechowski, A. Paris, and C. J. B. Dobbin. Process monitoring and control system for injection molding using nozzle-based pressure and temperature sensors. *Conference Proceedings Annual Technical Conference (ANTEC)*, 1(424-430), 1998.

[23] J. Cao, Y. S. Wong, and K. S. Lee. Applicationof statistical process control in injection mould manufacturing. *International Journal of Computer Integrated Manufacturing*, 20(5):436–451, 2007.

[24] S. Macfarlane and R. Dubay. The effect of varying injection molding conditions on cavity pressure. *58th SPE ANTEC Conference Proceedings*, pages 653–657, 2000.

[25] Z. Lou, H. Jiang, and X. Ruan. Development of an integrated knowledge-based system for mold-base design. *Journal of Materials Processing Technology*, 150(1-2):194–199, 2004.

[26] G. Yu. Application of genetic algorithms to conceptual design of injection mould. *Proceedings 3rd International Conference on Natural Computation (ICNC)*, 4:517–521, 2007.

[27] X. Jin and X. Zhu. Process parameters' setting using case-based and fuzzy reasoning for injection molding. proceedings. *3rd World Congress on Intelligent Control and Automation*, pages 335–340, 2000.

[28] S. L. Mok and C. K. Kwong. Application of artificial neural network and fuzzy logic in case-based system for initial process parameter setting of injection molding. *Journal of Intelligent Manufacturing*, 13:165–176, 2002.

[29] S. L. Mok, C. K. Kwong, and W. S. Lau. A hybrid neural network and genetic algorithm approach to the determination of initial process parameters for injection moulding. *International Journal of Advanced Manufacturing Technology*, 18(6):404–409, 2008.

[30] T. C. Bulgrin and T. H. Richards. The application of advanced control theory to enhance molding machine performance. *IEEE*, pages 94–102, 1994.

[31] C. M. Seaman and A. A. Desrochers. Multiobjective optimization of a plastic injection molding process. *IEEE Transactions on Control Systems Technology*, 2(3):157–168, 1994.

[32] J. Haeussler and J. Wortberg. Quality assurance in injection molding with neural networks. *51st SPE ANTEC Conference Proceedings*, 1:123–129, 1993.

[33] B. Ribeiro. Fault detection in a thermoplastic injection molding process using neural networks. *International Joint Conference on Neural Networks*, pages 3352–3355, 1999.

[34] S. J. Huang and T. H. Lee. Application of neural networks in injection moulding process control. *International Journal Advanced Manufacturing Technologies*, 21:956–964, 2003.

[35] C. H. Lu, C. C. Tsai, C. C. Liu, and Y. H. Charng. Predictive control based on recurrent neural network and application to plastic injection molding processes. *33rd Annual Conference of the IEEE Industrial Electronics Society (IECON)*, pages 792–797, 2007.

[36] B. Ribeiro. Support vector machines for quality monitoring in a plastic injection molding process. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, 35(3):401–410, 2005.

[37] S. L. B. Woll and D. J. Cooper. Onlinte pattern-based part quality monitoring of the injection molding process. *Polymer Engineering and Science*, 36(11):1477–1488, 1996.

[38] K. Salahsboor and A. Keshtgar. Statistical process monitoring via independent component analysis and learning vector quantization method. *Proceedings of the IEEE International Conference on Control Applications*, pages 2595–2600, 2007.

[39] B. M. Wise, N. B. Gallagher, S. Watts, D. D. White JR, and G. G. Barna. A comparison of pca, multiway pca, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process. *Journal of Chemometrics*, 13:379–396, 1999.

[40] R. A. Johnson and D. W. Wichern. *Applied Multiway Statistical Analysis.* Englewood Cliffs, Prentice-Hall International, 1992.

[41] E. L. Russell, L. H. Chiang, and P. D. Braatz. *Data-driven techniques for fault detection and diagnosis in chemichal process.* Advances in Industrial Control, 2000.

[42] P. Nomikos and J. F. MacGregor. Monitoring batch processes using multiway principal component analysis. *AIChE*, 40(3):1361–1375, 1994.

[43] J. F. MacGregor. Multivariate statistical approaches to fault detection and isolation. *5th IFAC SAFEPROCESS*, 2003.

[44] T. Kourti. Process analysis and abnormal situation detection: From theory to practice. *IEEE Control Systems Magazine*, 22(5):10–25, 2002.

[45] S. Wold, N. Kettaneh, H. Fridén, and A. Hamberg. Modelling and diagnosis of batch processes and analogous kinetic experiments. *Chemometrics and Intelligent Laboratory System*, 44:331–340, 1998.

[46] P. Nomikos and J. F. MacGregor. Multivariate spc charts for monitoring batch processes. *Technometriics*, 37(1):41–59, 1995.

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

[48] D. M. Himes, R. H. Storer, and C. Georgakis. Determination of the number of principal components for disturbance detection and isolation. *Proceedings of the American Control Conference*, 1994.

[49] G. Raîche, M. Riopel, and J.-G. Blais. Non graphical solutions for the cattell's scree test. *IMPS*, 2006.

[50] S. Yoon and J. F. MacGregor. Statistical and causal model-based approaches to fault detection and isolation. *AIChE Journal*, 46:1813–1824, 2000.

[51] T. Kourti. Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptative Control and Signal Processing*, 19:213–246, 2005.

[52] T. Kourti. Process analytical technology beyond real-time analyzers: The role of multivariate analysis. *Critical Reviews in Analytical Chemistry*, 36:257–278, 2006.

[53] H. Niitsuma and T. Okada. Covariance and pca for categorical variables. *9th Pacific-Asia Conference on Knowlegde Discovery and Data Mining (PAKDD)*, 2005.

[54] R. L. De Mantaras and E. Plaza. Case-based reasoning: An overview. *AI Communications*, 10(1):39–59, 1994.

[55] A. Aamodt and E. Plaza. Case-based reasoning: foundational issues,methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994.

[56] A. Jakulin and I. Bratko. Analyzing attribute dependencies. *7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 229–240, 2003.

[57] D. B. Leake. Case-based reasoning: experiences, lessons and future direction. *Press*, 1996.

[58] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of 14th International Joint Conference on Artificial Intelligence*, 2(12):1137–1143., 1995.

[59] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159., 1997.

[60] T. Fawcett. Pattern recognition. *Letters*, pages 861–874, 2006.

[61] J. Camacho and J. Picó. Multi-phase principal component analysis for batch processes modelling. *Chemometrics and Intelligent Laboratory System*, 81:127–136, 2006.