



Universitat de Girona

**MULTIVARIATE STATISTICAL PROCESS
CONTROL AND CASE-BASED REASONING
FOR SITUATION ASSESSMENT OF
SEQUENCING BARCH REACTORS**

Magda Liliana RUIZ ORDÓÑEZ

ISBN: 978-84-691-6833-2

Dipòsit legal: GI-I299-2008

UNIVERSITAT DE GIRONA
DEPARTAMENT D'ENGINYERIA ELÈCTRICA, ELECTRÒNICA I AUTOMÀTICA

MULTIVARIATE STATISTICAL PROCESS CONTROL
AND CASE-BASED REASONING
FOR SITUATION ASSESSMENT
OF SEQUENCING BATCH REACTORS

by

Magda Liliana Ruiz Ordóñez

Advisor

Dr. Joan Colomer Llinàs

DOCTORAL THESIS
Girona, Spain
March, 2008

UNIVERSITAT DE GIRONA
DEPARTAMENT D'ENGINYERIA ELÈCTRICA, ELECTRÒNICA I AUTOMÀTICA

MULTIVARIATE STATISTICAL PROCESS CONTROL
AND CASE-BASED REASONING
FOR SITUATION ASSESSMENT
OF SEQUENCING BATCH REACTORS

A dissertation presented in partial
fulfillment of the requirements of the degree
of Doctor per la Universitat de Girona
en Tecnologies de la Informació

By

Magda Liliana Ruiz Ordóñez

Advisor

Dr. Joan Colomer Llinàs

Girona, Spain
March, 2008

ABSTRACT

MULTIVARIATE STATISTICAL PROCESS CONTROL AND CASE-BASED REASONING FOR SITUATION ASSESSMENT OF SEQUENCING BATCH REACTORS

by Magda Liliana Ruiz Ordóñez

ADVISOR: Dr. Joan Colomer Llinàs

March, 2008
Girona, Spain

This thesis focuses on the monitoring, fault detection and diagnosis of Wastewater Treatment Plants (WWTP), which are important fields of research for a wide range of engineering disciplines.

The main objective is to evaluate and apply a novel artificial intelligent methodology based on situation assessment for monitoring and diagnosis of Sequencing Batch Reactor(SBR) operation. To this end, Multivariate Statistical Process Control (MSPC) in combination with Case-Based Reasoning (CBR) methodology was developed, which was evaluated on three different SBR (pilot and lab-scales) plants and validated on BSM1 plant layout.

Results showed that, MPCA is a robust technique for monitoring and fault detection of SBR operation. The MPCA was successfully tested for on-line (real-time) monitoring of pilot scale-SBR performing nitrogen removal - the first time this is achieved (to our best knowledge). The MPCA methodology is now ready to be used as part of daily operation of the SBRs.

For the diagnosis part, a comprehensive evaluation of the CBR methodology for automatic diagnosis of SBR process operation (BIOMATH and LEQUIA) was performed - the first time an artificial intelligent method applied within WWTPs. The methodology was then tested on the BSM1 plant layout which were used to construct abnormal events, e.g. faults, sensor failures, etc. The CBR method used input from the MPCA (rather than raw process data) and the best descriptors for the assessment of the situation (cases) were found to be principal components and errors (Q) of the statistical model. The main results showed that the CBR successfully diagnosed a wide range of operational problems

such as sludge bulking, influent inhibition/toxicity, high influent flow and sensor faults. The diagnosis performance of CBR method using several statistical extensions such as MPCA, Dynamic PCA and PCA were also studied. This comparison showed that the MPCA + CBR combination has a good diagnosis performance. However, a more theoretical and in-depth study of which inputs and descriptors to use for the situation assessment step in the CBR are needed to further improve the diagnosis.

In addition, the ability of CBR to maintain and update the knowledge was also studied and tested successfully using DROP and IB family of algorithms. This showed that repeating the cycle of learning helps maintaining and updating the case-base of the CBR.

Overall, this adaptive and intelligent aspects of the method makes it a good candidate for helping the management in the daily plant operation as an automatic diagnosis and real-time warning tool. Such artificial intelligent methods are promising tools which has the potential to contribute to good management and operation of plants. Further research is, however, needed to improve and consolidate the application of CBR to WWTP operations, including input descriptors, retrieve and update algorithms and decision making rules. All in all, this is expected to save operational costs as well as improve plant performance to comply with the goals of urban water management.

*El futuro empieza hoy
y lo que actualmente se está investigando
condicionará nuestra vida en un mañana muy próximo*

Josep M. Orta

To
Lucho
and Esteban

Acknowledgments

This doctoral thesis is the result of not only my own efforts, but also those of many people who directly or indirectly have collaborated with me. However, with limited time and after four years it is very difficult to remember all of them. Therefore, I apologize in advance for not including in these lines some additional people who really deserve recognition.

First of all, I would like to express my gratitude to Dr. Joan Colomer Llinàs, who since the beginning trusted my responsibility, knowledge and abilities to participate in the project DPI2002-04579-C02-01. His patience helped me to understand how to express ideas when writing reports, papers, etc. He also knew when I needed support to be able to continue in my research.

I would like to thank Professor Dr. ir. Peter A. Vanrolleghem from Ghent University, who hosted me in his research group, giving me the opportunity to use his facilities and laboratories. Furthermore, his group's guidance, comments and suggestions helped me to turn some ideas into reality.

I thank Doctor Christian Rosen and Doctor Ulf Jeppsson from Lund University for providing information, experience and data, as well as for all their attention when I was working within the Lund group.

I thank Dr. Jesus Colprim and Dr. Ma. Dolors Balaguer who provided information, experience, optimism and guidance to this doctoral thesis.

I thank Drs. Joaquim Comas and Ignasi Rodriguez, researches from the LEQUIA Group who contributed with knowledge, experience and suggestions.

I thank Dr. Joaquim Meléndez who contributed with suggestions.

I thank Dr. Gurkan Sin from the Technical University of Denmark who among laughs and meetings gave me important suggestions and ideas, and of course his friendship.

I thank the Spanish government through the coordinated research project *Development of a system of control and supervision applied to a Sequencing Batch Reactor (SBR) for the elimination of organic matter, nitrogen and phosphorus DPI2002-04579-C02-01* which has given me economical support during the period of the research scholarship "BES-2003-1931".

I thank the Spanish government for economical support during my research visits to the BIOMATH group at Ghent University and the Department of Industrial Electrical Engineering and Automation at Lund University.

I thank Drs Gabriel Ordóñez, Gilberto Carrillo, Roberto Martínez, Jaime Barrero, Gabriel Plata and Oscar Gualdrón, professors and guides during my engineering studies at the Industrial University of Santander (UIS) in Colombia, who motivated me to start my PhD studies.

I thank the members of the eXiT research group (those who are always here, those who are finishing, as well as those who are starting) for their friendship and support during this time.

I thank the BIOMATH and LEQUIA groups and the IEA Department who cooperated with information and comments.

I thank my family in Girona which is growing more and more: My brothers Ronald, Alvarito and Sebastian, my sister-in-law Sabik, my nephew Alejandro and niece Violeta, and my cousins Andrea, Dayan and Camilo, who have made me feel close to Colombia.

I thank my family in Colombia: My father Alvaro and my mother Amanda, who from a distance always encouraged me to keep jumping over obstacles in life.

I thank my cousin Jennifer from the USA, whom I have recently known as a friend.

I thank my friends Claudia, Cesar, Maria, Juan, Daniel, Fabiana, Guillermo, Maira, Martha, Rodolfo, David, Rosa, Javier, Vicky and Sonia who have been my family in Girona.

I thank Xavi, who has helped me in my work when I thought I would give up.

I thank life for giving me the opportunity to know lovely countries, wonderful people, amazing cultures, exciting history and enriching experiences.

Last but not least, thanks to my husband, Lucho, for being there when I have needed a friend and partner in my life. As a fellow PhD he gave me invaluable suggestions. Thanks also to my son, Esteban, who is my inspiration to improve day by day.

Notation and Abbreviations

Notation

A	Instances of the same class Nearest neighbor
b	Number of successful classifications in number of attempts
C	Carbon
c_α	Standar deviation to a given α
E	Residual matrix
h	Number of instances stored in the data base
I	Number of batches
J	Number of variables
K	Number of samples
λ	Eigenvalue
m	Number of variables in a data series
N	Number of principal components
n	Number of variables in a data series
P	Loading matrix
p_j	Loading vectors
Q	Loading Y matrix
S	Covariance matrix
SI	Set of instances
si	one instance
Q	Q-statistics SPE-squared prediction error
T	Score matrix
t_j	Score vectors
TI	Training instances
T^2	Hotelling T^2 statistics D-statistic
σ	Standard deviation
σ^2	Variance
θ	Sum of eigenvalues
μ	Mean
V	Variance captured
X	Historical data matrix of process variables

x_k^T	m-dimensional observation vector
\underline{X}	Three-dimensional array
Y	Predicted matrix
z	Confidence limit

Abbreviations

$2D$	Two-dimensional data array
$3D$	Three-dimensional data array
AOC	Abnormal Operation Condition
AS	Auto scaling
ASM1	Activated Sludge Model No1
BIOMATH	Department of Applied Mathematics, Biometrics and Process Control
BSM1	Benchmark Simulation Model No1
BSM1_LT	Benchmark Simulation Model No1 long-term
CA	Cluster Analysis
CB	Case Base
CBR	Case-Based Reasoning
CS	Continuous Scaling
CUSUM	Cumulative Sum
DA	Discriminant Analysis
DPCA	Dynamic Principal Component Analysis
DO	Dissolved Oxygen
DROP	Decremental Reduction Optimization Procedure
ICA	Independent Component Analysis
ED	Equipment Defects
EF	Electrical Fault
EU	European Union
EWMA	Exponentially Weighted Moving-Average Chart
eXiT	Ingeniería de Control y Sistemas Inteligentes
GS	Group Scaling
IB	Instance-Based Learning
ILC	Influence Load Change
IAWQ	International Association on Water Quality
IWA	International Water Association
KLA	Mass transfer coefficient
KPCA	Kernel Principal Component Analysis
LEQUIA	Laboratorio de Ingeniería Química y Ambiental
MATLAB	Matrix Laboratory
MBPCA	Multi-Block Principal Component Analysis
MPCA	Multiway Principal Component Analysis
MPPCA	Multi-Phase Principal Component Analysis
MSPCA	Multi-Scale Principal Component Analysis

MSPC	Multivariate Statistical Process Control
MOPs	Memory Organization Packets
N	Nitrogen
NH_4^+	Ammonium
NIPALS	Non-linear Iterative Partial Least Squares
NN	Neural Network
NOC	Normal Operation Condition
NOx	Nitrogen dioxide
ORP	Oxidation Reduction Potential
P	Phosphorus
PC	Principal Component
PCA	Principal Component Analysis
pH	pondus Hydrogenium
PLS	Partial Least Square
	Projection to Latent Structures
SBR	Sequencing Batch Reactor
SNO	Nitrate and nitrite nitrogen
SNH	$NH_4^+ + NH_3$ nitrogen
SPC	Statistical Process Control
SSR	Solid State Relays
SVD	Singular Values Decomposition
TSS	Total Suspended Solids
VC	Variation in the Composition
WWTP	Wastewater Treatment Plant

Contents

Contents	xvii
List of Figures	xxi
List of Tables	xxv
1 Introduction	1
1.1 Legal framework	1
1.2 Project framework	1
1.3 Objectives	2
1.4 Contributions	3
1.5 Outline	4
1.6 Publications	5
2 Wastewater Treatment Plants	9
2.1 The continuous process	10
2.1.1 The COST/IWA simulation Benchmark	11
2.2 Sequencing Batch Reactor	13
2.2.1 Semi-Industrial SBR Pilot Plant at University of Girona (LEQUIA)	14
2.2.2 Lab-Scale Plant SBR at University of Girona (LEQUIA)	17
2.2.3 Lab-Scale Plant SBR at Ghent University (BIOMATH)	22
2.3 Conclusions	24
3 Multivariate Statistical Process Control	25
3.1 Preview	25
3.2 Univariate Statistical Process Control	28
3.3 Multivariate Statistical Process Control	31
3.3.1 Principal Component Analysis	31
3.3.2 Dynamic Principal Component Analysis	36
3.3.3 Partial Least Squares	37
3.4 MSPC for Batch Processes	39
3.4.1 Multiway PCA	40
3.5 Conclusions	44
4 Case-Based Reasoning (CBR)	45
4.1 Preview	45
4.2 The CBR Cycle	46

4.3	Incremental Reduction Optimization Procedure Algorithms	49
4.3.1	DROP1	50
4.3.2	DROP2	50
4.3.3	DROP3	52
4.3.4	DROP4	52
4.4	Instance-Based learning algorithms	53
4.4.1	IB1	54
4.4.2	IB2	55
4.4.3	IB3	56
4.5	Conclusions	58
5	Application of MPCA Methodology to SBR pilot plants	59
5.1	Semi-Industrial SBR Pilot Plant from the LEQUIA group	59
5.1.1	Types of batch processes	59
5.1.2	Application of MPCA	63
5.2	SBR Pilot Plant from BIOMATH group	66
5.2.1	Systematic comparison of PCA models	66
5.2.2	Results	69
5.2.3	Discussion	76
5.3	On-line MPCA application	77
5.3.1	Module for ON-LINE Monitoring	78
5.3.2	Module to build the models	80
5.3.3	Module to validate new batches	82
5.3.4	Conclusions	84
5.4	Analysis and Conclusions	85
6	Automatic Detection of Abnormal Situation in Process Operation	87
6.1	Methodology	87
6.1.1	Data	89
6.1.2	MPCA	89
6.1.3	CBR	89
6.2	Descriptors, case base and distance refining	90
6.2.1	Step 1: Definition of descriptors	90
6.2.2	Step 2: Building the MPCA model and the validation data set	91
6.2.3	Step 3: Building the Case-Base	92
6.2.4	Step 4: Retrieval	93
6.2.5	Step 5: Testing	94
6.2.6	Results	94
6.3	Application of descriptors and distance refining to the COST/IWA BENCH- MARK	97
6.3.1	Methodology	97
6.3.2	Results	98
6.4	Case base maintenance and updating	102
6.4.1	Building the statistical model	103
6.4.2	Rearrangement of data	104
6.4.3	CBR application	104

6.4.4	Results	105
6.5	Analysis and Conclusions	112
7	Conclusions and future work	115
7.1	Conclusions	115
7.2	Future work	118
7.2.1	MSPC	118
7.2.2	CBR	118
A	LAMDA application	121
A.1	A short introduction to the LAMDA algorithm	121
A.2	Semi-Industrial SBR Pilot Plant application	123
A.3	Lab-Scale Plant SBR application	126
A.4	Data mining	128
	Bibliography	131

List of Figures

2.1	Wastewater system	9
2.2	a)Simulation benchmark system b)Representation in the Simulink-MATLAB configuration: mixed tank 1, tank 2 and tanks 3, 4 and 5 aerated	12
2.3	a) Semi-industrial Pilot Plant b) Operational Schema of the semi-industrial pilot plant SBR	14
2.4	Operational schema of the semi-industrial pilot plant SBR	15
2.5	Storage tank Filling	15
2.6	Cycle applied to semi-industrial SBR pilot plant	16
2.7	Comparison of 5760 samples and 392 samples for variables	18
2.8	Lab-scale plant from LEQUIA	19
2.9	Period 1 cycle configuration	19
2.10	Period 2 cycle configuration	20
2.11	Lab-scale plant from BIOMATH	22
2.12	Operational scheme of the SBR	23
2.13	Cycle applied to lab-scale plant SBR	23
3.1	Classification of monitoring, fault detection and diagnostic algorithms	25
3.2	An illustration of the Shewhart chart. The rhombuses are observations. The process is said to be 'in control'	29
3.3	Multivariate statistical analysis vs. univariate statistical analysis and a comparison of the in-control status regions using T^2	30
3.4	Projection of the process variables in a new space using PCA	32
3.5	NIPALS algorithm	34
3.6	Q -statistic and D -statistic with 95.27% confidence limits	39
3.7	Arrangement of a three-way array	40
3.8	Decomposition of a three-way data array, $\underline{\mathbf{X}}$, by MPCA	42
3.9	Other decomposition of a three-way data array, $\underline{\mathbf{X}}$, by MPCA	42
4.1	CBR cycle	47
4.2	The distance between the new case or new problem and cases A and B . $X1$ and $X2$ are the characteristics that define the cases.	48
4.3	a)Central cluster instance b)Non-noisy border point c)Collection of border instances	49
4.4	DROP1 algorithm	51
4.5	DROP2 algorithm	52
4.6	DROP3 algorithm	53

4.7	DROP4 algorithm	54
4.8	IB1 algorithm	55
4.9	IB2 algorithm	56
4.10	IB3 algorithm	57
5.1	Score plot for batches. Dashed line is the model	60
5.2	DO (green line) and ORP (blue line) profiles when an EF occurs	60
5.3	ORP and DO profiles when and VC fault condition is presented a) NOC b) AOC	61
5.4	ORP and DO profiles when and ED fault occurs	61
5.5	ORP and DO profiles in presence of rainwater	62
5.6	ORP and DO profiles a)Good final quality b)Normal final quality	62
5.7	Types of events	63
5.8	Q -statistics and T^2 -statistics with 92.79% confidence limits for the Semi- Industrial Pilot Plant	65
5.9	MPCA Methodology applied to pilot-scale SBR	67
5.10	Scale process for variable wise models	69
5.11	The Q-Q distribution of the first principal component for models that are unfolded variable wise (left) and batch wise (right) and scaled with a) CS b) GS and c) AS approaches	71
5.12	a) Batches 1010 and 1011 b) New Gaussian distribution for Model 3	72
5.13	Loads graphics from components 1 to 5	73
5.14	Schematic representation of the interface	78
5.15	Interface to on-line monitoring in variable wise mode	79
5.16	Interface to on-line monitoring in batch wise mode	80
5.17	Contribution interface of each component	80
5.18	Interface to determine the number of principal components	81
5.19	Window of complementary information	82
5.20	Contribution analysis graphics	83
5.21	Contributions analysis graphics	83
5.22	Contributions analysis graphics	84
6.1	Methodology applied to SBR pilot plant	88
6.2	Test strategies to select descriptors and distances	91
6.3	Case Base 1	92
6.4	Case Base 2	93
6.5	Dispossession and unfolding of three-way data array	95
6.6	Projection of the process variables in a new space using PCA	97
6.7	Diagnosis using methodologies 1 and 2 for the evaluation data set	101
6.8	Diagnosis using methodology 3 for the evaluation data set	102
6.9	Case Base maintenance applied to pilot-scale SBR	103
6.10	Off-line variables a) Table used for the biological experts b) Table used for the monitoring experts	106
6.11	Loading plots for model 1 which corresponds to two reaction stages	108
6.12	Loading plots for model 2, which corresponds to three reaction stages	109
6.13	Score plots for models 1 and 2	110

6.14	Three dimensional representation of model 1 for one standard deviation . .	111
6.15	Three dimension representation of model 1 for two standard deviations . .	111
6.16	Learning evolution for tests 1, 2, and 3	112
6.17	Learning evolution for tests 4, 5, and 6	113
A.1	Basic LAMDA recognition methodology	123
A.2	LAMDA classification	124
A.3	Batch class composition according to type of batch process	124
A.4	Three dimensional representation for normal behavior (Class 2)	127
A.5	Three dimensional representation for abnormal behavior (Class 3)	127
A.6	Color levels for class 2	127
A.7	Color levels for class 3	127
A.8	Example of class with normal behavior (Class 2)	128
A.9	Example of abnormal behavior (Class 3)	128

List of Tables

2.1	Work schedule configuration from LEQUIA Lab-Scale Plant SBR	20
2.2	Three different lengths for anaerobic phase configuration	21
3.1	Principal component extraction of PLS example	38
3.2	Events exceeding limits a) Q -statistic b) D -statistic	38
3.3	Types of unfolding a three way data array	41
5.1	Types of events with AOC	62
5.2	Types of events with NOC	63
5.3	Principal component extraction	64
5.4	Batches detected using Q -statistic and T^2 -statistic	65
5.5	Names for each developed model	69
5.6	Variances for models 4, 5 and 6	72
5.7	Variances for models 1, 2 and 3	74
5.8	Criteria for performance assessment of the monitoring models in variable wise mode	74
5.9	Criteria for performance assessment of the batch wise monitoring	74
5.10	Performance assessment of Variable Wise (VW) considered models	75
5.11	Performance assessment of Batch Wise (BW) considered models	76
6.1	Names for each developed CBR	94
6.2	Specificity and sensitivity of each control charts	95
6.3	Specificity and sensitivity for Case Base 1 (CB1)	96
6.4	Sensitivity for Case Base 2 (CB2)	96
6.5	Division of the first data set	99
6.6	Assignment of class numbers for each event	100
6.7	Names for each methodology developed	100
6.8	Names for each test developed	105
A.1	LAMDA-descriptors used to define batches	122
A.2	Classes obtained by SALSA-LAMDA for semi-industrial pilot plant	125
A.3	Batch class composition according to principal component	125
A.4	Classes from SALSA-LAMDA for BIOMATH SBR pilot plant	126
A.5	Names for classes of the first classification	129
A.6	Names for classes of the second classification	129

Chapter 1

Introduction

1.1 Legal framework

The treatment of wastewater has become one of the important environmental topics. Wastewater treatment is an important part of maintaining the highest possible quality of natural water resources (rivers, lakes and seas). With new regulations for quality monitoring of WasteWater Treatment Plants (WWTP) under directive 98/15/CE (*Directiva 98/15/CE de la Comision, de 27 de febrero de 1998, por la que se modifica la Directiva 91/271/CEE del Consejo en relacion con determinados requisitos establecidos en su anexo I n.d.*), it is necessary to introduce new technology for control and supervision. The objective is to harmonize urban wastewater treatment legislation throughout the European Union (EU), in an attempt to protect the environment from any adverse effects. If the treatment of wastewater is insufficient in one member state of the EU, it often influences other members, affecting human integrity (de los Diputados de Espana 1978). The treatment of urban water must vary according to the receiving waters, which can be more sensitive or less sensitive, so the requirements for discharges from urban wastewater treatment plants are different (CEE 1991). In this way, legislations minimize the adverse effects on the environment of this discharges.

1.2 Project framework

Title:*Development of an Intelligent Control System applied to a Sequencing Batch Reactor (SBR) for the removal of Organic Matter, Nitrogen and Phosphorous. SICOTIN-SBR2*

This project is a continuation of a previous project DPI2002-04579 whose promising results prompted the consideration of a more ambitious control system. Goals are to improve the overall process performance and adapt it according to the influent wastewater characteristics in a wastewater treatment plant.

In the previous project, a system of Case-Based Reasoning (CBR) was elaborated. This system is qualified to identify the situation of the process when finalizing a cycle,

as well as to recover historical cases of operation in order to propose modifications to the current operation conditions of the process. First, qualitative trends were used to depict tendencies of the process in order to obtain variables profiles. Second, the cases were stored in a case-base. Finally, a comparison between the recovered historical cases and the diagnosis was developed (Rubio et al. 2004). In addition, due to the amount of collected data, a brief application of Multivariate Statistical Control technique was made with promising results.

The results using Multivariate Statistical Control suggest the continuation of this line of research in pursuit of other objectives, such as estimation of the characteristics of influent water and the quality of effluent water. The nature of the process (by batches, nonlinear, highly variable in time) and a complete system of data acquisition and storage indicates suitable tools for CBR, (with an initial version already implemented) Multivariate Statistical Process Control (MSPC), and a combination of both.

In this manner in this project, the use of CBR and MSPC is proposed to diagnose the state of the process and to consider the characteristics of influent and effluent water.

1.3 Objectives

In this thesis, the monitoring and diagnosis of WWTP are investigated. The main objective is to develop a methodology to assess Sequencing Batch Reactors (SBR) WWTP, focusing on Statistical Models and diagnosis using MSPC and CBR. The evaluation includes determining whether abnormal operation is present and defining the fault class and its features. More specifically, the objectives of this work are the following:

To develop a methodology to detect and diagnose Normal and Abnormal Operation Condition (AOC) using historical data from several Wastewater Treatment Plants. The information will be processed in order to obtain parameters that determine the real situation into the processes. This methodology could be used in the future for situation assessment in full-scale plant.

To introduce a Multivariate Statistical Process Control (MSPC) approach in a Sequencing Batch Reactor plant for on-line monitoring. Since the process has many sensors measuring variables for a long time, and since these data are highly correlated, Principal Component Analysis (PCA) and its extensions are proposed for reducing the dimensionality of the problem. Combined with other techniques, this will help improve and accelerate the monitoring and diagnostic processes.

To include Case Based Reasoning (CBR) to improve the results obtained from the MSPC methodology. CBR is an expert system which applies the experience and knowledge from experts about past situations. This experience can often provide a solution to new problems to help operators in their daily management and operation of the plant.

To validate the developed methodology in several processes with different kind of operating condition Since the main objective is situation assessment of WWTPs, several plants with different types of problems and several operation condition are considered in order to refine and improve the best methodology with the ability for determining the situations in whatever kind of SBR process.

1.4 Contributions

The main contributions presented in this work are the following:

- Mainly, this doctoral thesis makes a rigorous evaluation and testing of Multiway PCA (MPCA) + CBR approaches on several systems with different scales allowing a realistic assessment of the methods and their feasibility to practice.
- A new approach to situation assessment to detect the abnormal behavior in Wastewater Treatment Plants is proposed. This approach uses MSPC and CBR. MSPC is used to reduce the dimensionality and to remove non-linearity in the data. CBR is used to build the Case-base to diagnose future events. Maintenance and updating are made through the learning capacity of this tool.
- Several combinations of the above approaches are performed using two pilot plants, one semiindustrial pilot plant and one Benchmark simulation model. These processes have differences between them, for instance influent, size of reactors, problems and operating conditions. The influent for the first two plants are prepared in the laboratory with several ingredients (syntectic influent). In semi-industrial pilot plant the influent is taken directly from the real wastewater treatment plant which mainly comes from residential area. In this aspect, a rigorous evaluation and testing of the methodology using various systems with different scales have been performed. This means that it should be possible to generalise the obtained results.
- Several options of data scale before building the model using MSPC are studied in order to find the best option for this kind of process.
- A full implementation of the CBR approach for a WWTP is performed including case base building, maintenance and updating and on line application.
- The MPCA methodology in a WWTP is implemented on-line. A new module for the application of the proposed methodology has been added to the existent monitoring systems.
- A combination of the MSPC with the LAMDA algorithm (monitoring + clustering) to situation assessment in WWTPs is performed to identify normal data and to classify situations.

1.5 Outline

The structure of the thesis consists of six chapters and two appendices where supplementary materials are provided.

In the present Chapter (Chapter 1), the background, the general context and the problem statement of the research carried out in this thesis is provided. Next the research objectives are outlined, which guide the research carried out throughout the thesis. Last, the structure of the book is given.

Chapter 2 provides a general description of the wastewater treatment systems used throughout the thesis to test and develop the MPCA and CBR methodology. In total, four different systems were used. The first one is the semi-industrial SBR pilot which performs nitrogen and organic matter removal. The second one is a lab-scale SBR performing nitrogen and organic matter removal. Both of these plants were hosted and operated in the laboratory of LEQUIA research group (University of Girona, Catalonia, Spain). The third one is a lab-scale SBR that performs biological nitrogen, organic matter and phosphorus removal hosted and operated in the laboratory of BIOMATH (Ghent University, Belgium). The last one, is the BSM1 benchmark plant layout which is developed by Task Group on Benchmarking of Control Strategies for WWTPs.

Chapter 3 presents a review of multivariate statistical methods used mainly for process monitoring purposes. To this end, the classification of monitoring, fault detection and diagnoses algorithms and a state-of-the-art in the PCA applications are provided. Basic concepts and various methodologies developed within the context of univariate and multivariate statistical process control are introduced. Special attention is given to multi-way principal component analysis with the possible unfolding and control charts for batch monitoring process are explained.

Chapter 4 provides the theoretical background of the case-base reasoning (CBR) cycle. The four fundamental R's of the CBR that is retrieve, reuse, revise and retain, are explained. The artificial intelligence capacity of the approach to adapt and learn using decremental reduction optimization procedure (DROP) and instance-based learning (IB) algorithms are explained in detail. The detailed introduction of the DROP family of algorithms was felt necessary since it is the first time (to our best knowledge) a full implementation of case-based reasoning (CBR) methodology to wastewater treatment plant is done. In the same way, IB is also explained indepth, which ensures a continuous update of the case-base.

Chapter 5 provides results from the evaluation of the MPCA methodology at two SBR systems (semi-industrial pilot plant at LEQUIA group and lab-scale BIOMATH SBR). First, a preliminary work was performed using the data from the semi-industrial plant. Due to the correlation performed with the variables by means of the statistical model, several types of situations could be determined. Second, an indepth analysis of the application of MPCA to lab-scale SBR was done. The research in this section was done in a didactic sense to help find out how to build good MPCA models for process monitoring

purposes. In this regard, issues such as type of scaling and unfolding, number of principal components were investigated in view of their impact on process monitoring performance. Finally, implementation of the MPCA for on-line monitoring of the semi-industrial pilot plant (at LEQUIA group) were performed and the results were given.

Chapter 6 describes results from the evaluation of the methodology combining CBR with the MPCA approach. Three historical data were used for the evaluation: lab-scale BIOMATH SBR, COST/IWA simulation benchmark and lab-scale LEQUIA SBR. The development and evaluation of the methodology was carried in two parts. In Part one, the descriptors, the building of case-base and the retrieve algorithms of the CBR methodology were addressed. This evaluation was performed using historical data from lab-scale BIOMATH SBR plant. The objective was to find the best combination of descriptors, the retrieve procedure and the case-base structure. Having found that, the CBR methodology was tested/validated using COST/IWA simulation benchmark generated set of operational data. In part two, the maintenance and updating algorithms of the CBR method were investigated using the data from lab-scale LEQUIA SBR. In this way, the CBR ability to delete redundant information and learn automatically were added.

Finally, the main conclusions obtained from the research results as well as recommendations for future work are described in Chapter 7. Additionally, results from the combination of the LAMDA algorithm with the MPCA methodology are given in the Appendix A.

1.6 Publications

The following articles were published from the research results generated in this thesis study. The contribution of the author has been mainly to develop the MPCA and CBR algorithms and the analysis and interpretation of results.

Book chapters

Ruiz M., Colomer J. and Meléndez J. (2006) "Monitoring a sequencing batch reactor for the treatment of wastewater by a combination of multivariate statistical process control and classification technique". *Frontiers in Statistical Quality Control* ISBN 10 3-7908-1686-8 Physica-Verlag Heidelberg New York.

Contribution: MPCA modeling, analysis, interpretation and writing.

International Journal Publications

Mujica L, Vehí J., Ruiz M., Verleysen M., Staszewski W. and Worden K. (2008) "Multivariate Statistics Process Control for Dimensionality Reduction in Structural Assessment" *Mechanical Systems and Signal Processing*, 22:155-171.

Contribution: MPCA analysis and interpretation of preliminaries results.

Ruiz M., Villez K., Sin G., Colomer J., Rosen C. and Vanrrolleghem P.A. (2008) "Different PCA approaches for monitoring nutrient removing batch process:Pros and Cons" in preparation for publication in Water Science and Technology.

Contribution: MPCA modeling, analysis, interpretation and writing.

Ruiz M., Sin G., Colprim J. and Colomer J., "MPCA and CBR methodology for monitoring, fault detection and diagnosis in wastewater treatment plant" (2008) in preparation for publication in Water Science and Technology.

Contribution: MPCA modeling, CBR algorithms, analysis, interpretation and writing.

National Journal Publications

Ruiz M., Colomer J. and Melendez Q. (2006) "Combination of statistical process control (SPC) methods and classification strategies for situation assessment of batch process" Revista Iberoamericana de Inteligencia Artificial. 29:99-107.

Contribution: MPCA modeling, analysis, interpretation and writing.

International Conferences

Villez K., Ruiz M., Sin G., Rosen C., Colomer J. and Vanrrolleghem P.A. (2007) "Combining Multiway Principal Component Analysis (MPCA) and clustering for efficient data mining of historical data sets of SBR processes" Proceedings of the 3rd International IWA Conference on Automation in Water Quality Monitoring (AutMoNet2007), Ghent, Belgium, September 5-7, 2007, appeared on CD-ROM.

Contribution: Preliminary LAMDA methodology.

Ruiz M., Rosen C. and Colomer J. (2007) "Diagnosis of a continuous treatment plant using Statistical Models and Case-Based Reasoning", Proceedings of the 3rd International IWA Conference on Automation in Water Quality Monitoring (AutMoNet2007), Ghent, Belgium, September 5-7, 2007, appeared on CD-ROM.

Contribution: Modeling and diagnosis analysis using PCA, DPCA, MPCA and CBR, interpretation and writing.

Jaramillo M., Ruiz M., Colomer J. and Melendez J. (2007) "Multiway Principal Component Analysis and Case Base Reasoning methodology for abnormal situation detection in a Nutrient Removing SBR" Proceedings of the European Control Conference, Kos, Greece, July 2-5, 2007, appeared on CD-ROM.

Contribution: Modeling and diagnosis analysis using MPCA and CBR, interpretation and writing.

Ruiz M., Villez K., Sin G., Colomer J. and Vanrolleghem P.A. (2006) "Influence of scaling and unfolding in PCA based monitoring of nutrient removing batch process" 6th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes. September, 2006, Beijing P.R. China.

Contribution: MPCA modeling, analysis, interpretation and writing.

Ruiz M., Colomer J., Rubio M., Melendez J. and Colprim J. (2004) "Situation assessment of a sequencing batch reactor using multiblock MPCA and fuzzy classification" BESAI Workshop in Binding Environmental sciences and Artificial Intelligence, ECAI 2004 European Conference on Artificial Intelligence, ISSN.0922-6389, August, 2004, Valencia (Spain).

Contribution: MPCA modeling, LAMDA classification, analysis, interpretation and writing.

Ruiz M., Colomer J., Colprim J. and Melendez J. (2004) "Multivariable statistical process control to situation assessment of a sequencing batch reactor", CONTROL 2004, pp.11, ISBN.0 86197 130 2, September, 2004, Bath (UK).

Contribution: MPCA modeling, analysis, interpretation and writing.

Ruiz M., Colomer J., Rubio M. and Melendez J. (2004) "Combination of multivariate statistical process control and classification tool for situation assessment applied to a sequencing batch reactor wastewater treatment" ISQC Intelligent Statistical Quality Control pp.257-267. ISBN.83-88311-69-7. June, 2004, Warszawa (Poland).

Contribution: MPCA modeling, LAMDA classification, analysis, interpretation and writing.

Ruiz M., Melendez J., Colomer J., Sanchez J. and Castro M. (2004) "Fault location in electrical distribution systems using PLS and NN" Proceedings of the International Conference on Renewable Energies and Power Quality (ICREPQ'04), Barcelona, Spain, 2004, appeared on CD-ROM.

Contribution: PLS modeling, NN classification, analysis, interpretation and writing.

Chapter 2

Wastewater Treatment Plants

Every community produces solid and liquid wastes. Liquid waste refers to water after of residential, industrial and commercial sectors usage (wastewater). If it is accumulated and stagnated bad-smelling gases are generated, including a big number of human harmful microorganisms. Also, it includes nutrients favoring the growth of aquatic plants which contain toxic compounds. To prevent this situation, the European Union has regulated the final quality of urban wastewater with the new directive 98/15/CE (*Directiva 98/15/CE de la Comision, de 27 de febrero de 1998, por la que se modifica la Directiva 91/271/CEE del Consejo en relacion con determinados requisitos establecidos en su anexo I n.d.*). The main objective is to protect the environment from the negatives effects of this wastewater. As a consequence, the biological nutrient removal technology has been increased around the world in WWTPs (Figure 2.1).

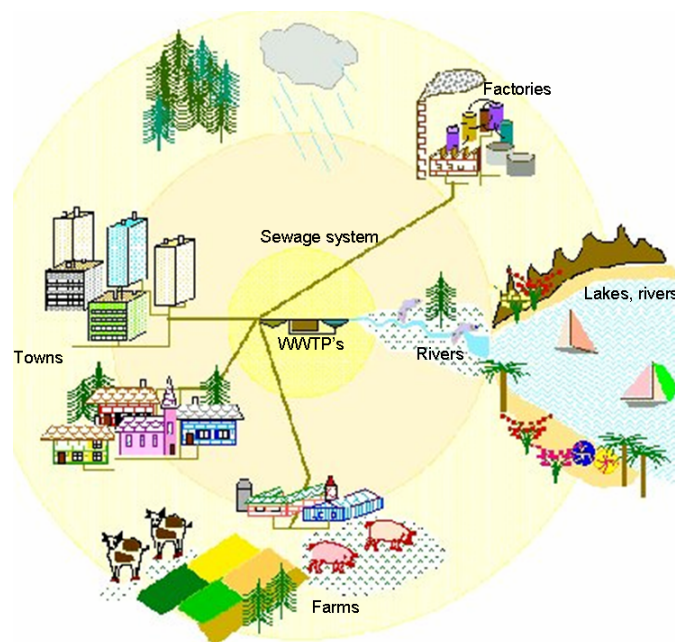


Figure 2.1: Wastewater system

Extracted from Benchmark-Web 2007

In an Activated Sludge (AS) process, the most commonly used technology for municipal wastewater treatment is the biomass. This is composed of a wide mixed culture of microorganisms is blended with the wastewater, which is composed of organic matter, suspended solids and nutrients. The mixture is then discharged to another reactor which is typically a settling tank to separate the biomass from the treated water.

2.1 The continuous process

Treatment plants perform primary treatment (physical removal of floatable and settleable solids) and secondary treatment (biological removal of dissolved solids).

Primary treatment involves (Federation 2003):

1. Screening - to remove large objects such as stones or sticks, which could plug lines or block tank inlets.
2. Grit chamber - slows down the flow to allow grit to fall out.
3. Sedimentation tank (settling tank or clarifier) - settleable solids settle out and are pumped away, while oils float to the top and are skimmed off.

Secondary treatment consists of a biological conversion of dissolved and colloidal organic compounds into stabilized, low-energy compounds and new biomass cells, caused by a very diversified group of microorganisms that respire in the presence of oxygen. Three options are explained below (Comas 2000):

1. Activated Sludge - The most common option uses microorganisms in the treatment process to break down organic material with aeration and agitation. The mixture is continually recirculated back to the aeration basin to increase the rate of organic decomposition.
2. Trickling Filters - The wastewater is sprayed on stone or plastic beds, allowing it to trickle. Microorganisms growing on the beds break down organic material in the wastewater. Trickling filters drain at the bottom; the wastewater is collected and then undergoes sedimentation.
3. Lagoons - These are slow, cheap, and relatively inefficient, but can be used for various types of wastewater. They rely on the interaction of sunlight, algae, microorganisms, and oxygen (sometimes aerated).

In this thesis, the goal is to diagnose normal and abnormal operation condition in WWTPs. Two kinds of plants are considered, a COST/IWA simulation benchmark (Copp 2002) and a Sequencing Batch Reactor (SBR) process. Each of these processes is explained in the next sections.

2.1.1 The COST/IWA simulation Benchmark

The International Association on Water Quality (IAWQ) held a meeting in 1983 in which a group was formed to promote and develop practical applications of models for the design and operation of biological wastewater treatment systems (Jeppsson 2007). To date, several objectives have been developed. One is the COST/IWA simulation benchmark which compares and evaluates different control strategies for a biological nitrogen removal process. In “benchmark simulation the goal is to obtain good performance and cost-effectiveness in wastewater management systems, given detailed descriptions of plant layout, model parameters and simulation models. The benchmark simulation compares past, present and future control strategies without reference to a particular facility collecting large amounts of data (Copp 2002).

The benchmark simulation system includes a plant layout, simulation models and parameters, a detailed description of influent disturbances (dry weather, storm and rain events), as well as performance evaluation criteria to determine the relative effectiveness of proposed control strategies (Copp 2002). The plant has five completely mixed reactors with a total volume of $5999 m^3$ of which tanks 1 and 2 are each $1000 m^3$ and tanks 3,4 and 5 are each $1333 m^3$ (see Figure 2.2a)). The biological process is modeled using the Activated Sludge Model No1 (ASM1) (Henze et al. 1987), and the settling processes are described using the Tákacs ten-layer model (Takacs et al. 1991). Several platforms have been used to develop for the Benchmark simulation using C/C++, Fortran and Simulink-MATLAB among others. In this thesis, the Simulink-MATLAB platform is used, see Figure 2.2b).

The Benchmark Simulation Model No1 (BSM1) has seen continuous improvements to the control system, procedure and evaluation criteria, however, it does not allow for Long-Term (LT) evaluation. To overcome this inconvenience, Rosen et al. (2004) and Jeppsson et al. (2006) have proposed long-term monitoring strategies (*BSM1_LT*) and another extension to allow control strategy development and performance evaluation at a plant-wide level (*BSM2*). Among other changes in *BSM1_LT*, the toxic components have been characterized by their concentration and not as a percentage of toxicity.

The final version of BSM1 is still in evolution; in this work, the most recent prototype has been used in order to acquire data. This version is closest to reality and a well known benchmark plant for evaluating the methodology developed in this doctoral thesis. In total, 9 sensors were simulated to monitor the process; they are: *flow_rate*, Nitrate and nitrite nitrogen (*SNO*), units $gN m^{-3}$: *SNO_reactor2*, *SNO_reactor5*; $NH_4^+ + NH_3$ nitrogen (*SNH*), units $gN m^{-3}$: *SNH_reactor5*; Total Suspended Solids (*TSS*), units mg/l : *TSS_reactor5*, *SNIT_plantininput*; Mass transfer coefficient(*KLA*), units m/s : *KLA_reactor3*, *KLA_reactor4*, *KLA_reactor5*. 96 samples per variable were collected. 609 days were simulated so that the dynamic influent data become steady state. The first 63 days are disregarded would become steady state. 364 days were used to identify and train the statistical models and the CBR approach. Immediately afterwards, 182 days were used to evaluate the monitoring models and diagnosis. The AS process is a complex system with operational problems. One of these problems has been simulated:

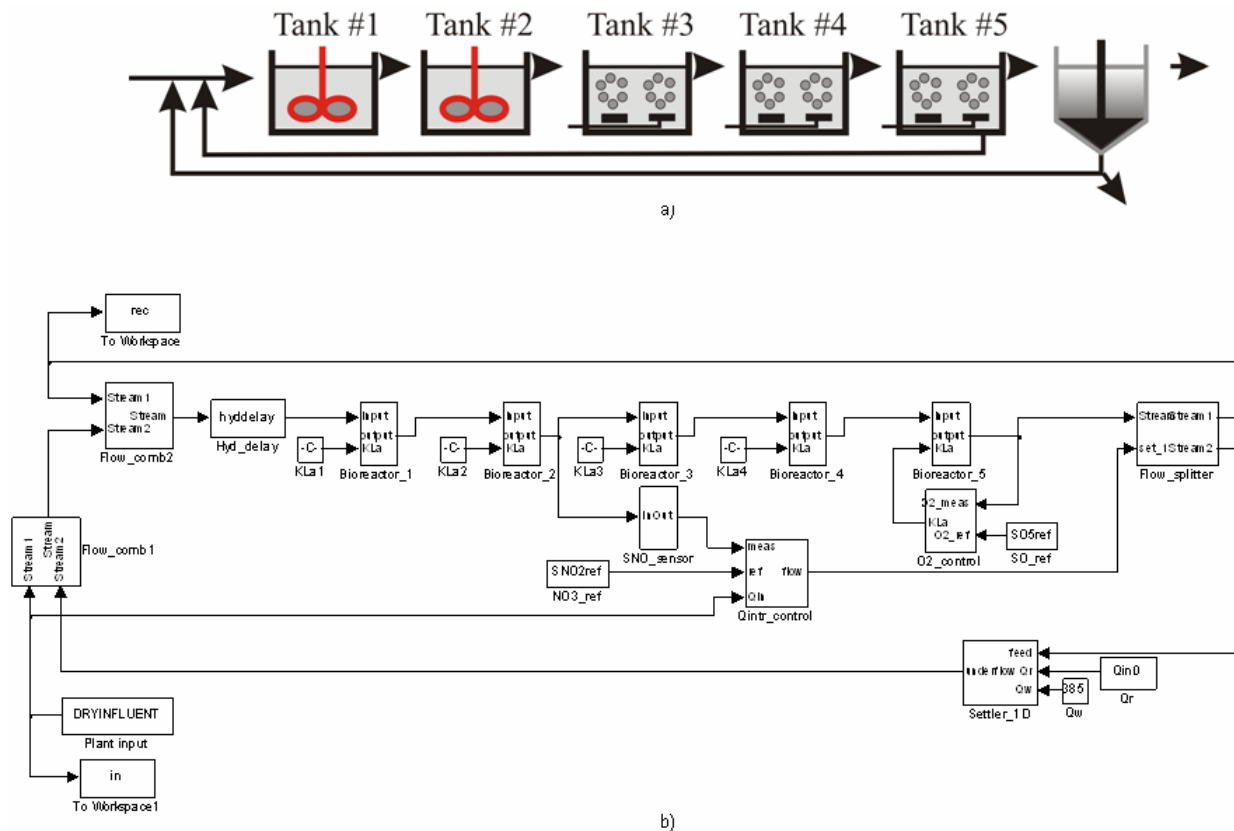


Figure 2.2: a)Simulation benchmark system b)Representation in the Simulink-MATLAB configuration: mixed tank 1, tank 2 and tanks 3, 4 and 5 aerated

Filamentous bulking (Bulking event). A bulking event is mainly caused by low DO in the aeration tank, causing growth of filamentous bacteria. This makes the separation of the biomass from the treated water difficult (bad settling). The events used in benchmark simulation were the following:

- Training set (364 days): Two bulking events starting from day 30 to 46 and 329 to 355. Five incidences of low level of inhibition however enough to affect the bacterial population due to toxicity starting from day 72 to 76, day 92 to 94, day 154 to 155, day 261 to 263 and 285 to 295. 41 aleatory days with high flow rate event. Finally, Day 180 with nitrate sensor fault.
- Evaluating set (182 days): Bulking event starting from day 122 to 147. Two incidences of low level of inhibition, with a soluble carrier start in day 60 to 62, day 110 to 112. Another inhibition/toxicity event with a particulate carrier is imposed day 72. Finally, 21 aleatory days with high flow rate event

2.2 Sequencing Batch Reactor

The main characteristic of the SBR is that whole process occurs in the same reactor following a sequence of phases, while in a continuous wastewater process plant such as the *BSM1_LT* plant shown above each phase occurs in different reactors. The SBR process has been shown to be an effective alternative to treat wastewater from domestic and industrial waste.

The advantages of the SBR process can be attributed to:

1. Design:
 - the clarification occurs in the same reactor.
 - a portion of the treated water is replaced by untreated wastewater for each cycle, distinguishing the SBR process from other continuous flowtype activated sludge systems.
 - influent and effluent flows are uncoupled by time sequencing.
2. Microbiology: Biological processing is cyclic.
3. Operation: The process operation can be easily adapted for different requirements by changing the duration of each phase.

The operation in a SBR process is performed by means of repeating a defined cycle. This cycle has four basic phases:

1. Fill: The influent wastewater is pumped into the reactor to be treated. The reactor can be filled under different conditions depending on operating conditions.
2. Reaction: Aerobic and anoxic conditions are combined in order for the biomass to consume the substrate from the influent wastewater.
3. Settle: This phase occurs when the aerobic and anoxic conditions finish. Normally, this phase is quicker than in a continuous process. The excess sludge is drained.
4. Draw: When the process finishes, the treated water is drawn from the reactor. In this way, it is ready to start the process with new influent wastewater.

Filling and reaction phases can be combined and configured in different ways and several times. This combination depends on the main objective of the treatment, the organic matter and nitrogen removal. The settle and draw phases are the last ones in the cycle structure. The most common structure is based on a combination between anoxic and aerobic conditions ending with the settling and draw phases (Corominas. 2006).

The SBR plant carries out advanced treatment, in which the nitrogen is removed in two steps as follows (Vives et al. 2001):

Nitrification : Ammonia is converted to nitrate by aerobic nitrifying (autotrophic) microorganisms.

Denitrification : Nitrate is converted to N_2O or reduced to nitrogen gas under anoxic (without oxygen) conditions by anoxic heterotrophic microorganisms.

In this thesis, three SBR pilot plants have been used: two from the Laboratorio de Ingeniería Química y Ambiental (LEQUIA) and one from the Department of Applied Mathematics, Biometrics and Process Control (BIOMATH). The expert knowledge of the process is the interpretation of profiles of some state variables which will be used to decide special events into the processes. In SBR pilot plants from the LEQUIA, these variables are Oxidation Reduction Potential (ORP) (mV), pondus Hydrogenium (pH), Dissolved Oxygen (DO) (mg/L) and Temperature (C). In the BIOMATH pilot plant, the variables are ORP, pH, DO, weight, conductivity and Temperature. These variables provide information about the biological reactions and the process state. The on-line measurements in the settling and draw phases of the SBR are usually not consistent due to changing properties/dynamics of settling in each batch (Sin et al. 2004) (Lee and Vanrolleghem 2003b). As a consequence, these phases are not considered in the development of the methodologies.

2.2.1 Semi-Industrial SBR Pilot Plant at University of Girona (LEQUIA)

Generalities

The semi-industrial pilot plant is located at a real wastewater treatment plant in Spain (Catalonia), seen in Figure 2.3a. It is composed of a metal square reactor of $1m^3$. The minimum and maximum volumes of the reactor are 483 liters and 683 liters respectively, and a volume of 200 liters of water to process (see Figures 2.3b and 2.4).



Figure 2.3: a) Semi-industrial Pilot Plant b) Operational Schema of the semi-industrial pilot plant SBR

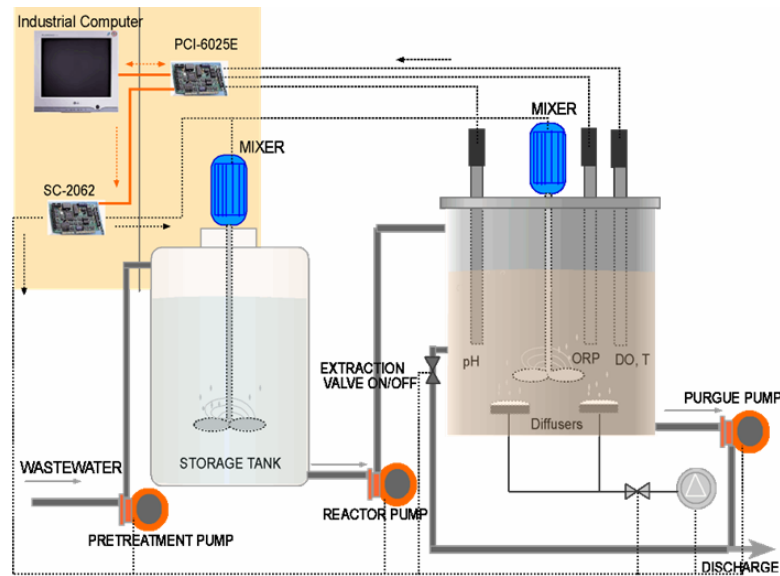


Figure 2.4: Operational schema of the semi-industrial pilot plant SBR

Wastewater is taken directly from the real WWTP by means of a peristaltic pump (Watson Marlow 621 F/R 77 RPM, flow=0 – 50L.h⁻¹) after passing through a grit chamber, sand and grease removal units (see Figure 2.5) in order to be stored in a storage tank under mixing conditions without refrigeration.



Figure 2.5: Storage tank Filling

Next, the wastewater is pumped to the reactor by means of another peristaltic pump according to the operating conditions. During the reaction phase, the mixed liquor is maintained under suspension and homogeneous conditions using a marine helix. The energy provided for mixing is used to regulate the distribution of mixed liquor solids in the reactor. The aerobic condition is achieved by four air filters (SKS-80 EW) through porous diffusers located at the bottom of the reactor. The air supply is controlled by an ON-OFF

valve in order to achieve complete nitrification and avoid high DO concentration when the anoxic conditions start (Corominas. 2006).

The monitoring and control systems consist of three parts: acquisition, monitoring and control system. The SBR process is equipped with DO-Temperature (OXIMAX-W COS 41), pH (CPF 81) and ORP (CPF 82) Endress-Hauser sensors. These signals are captured by a data acquisition card (PCI-6025E from National Instruments). The whole process is controlled by software in LabWindows (from National Instruments). The control is performed by a power relay output board (SC-2062 from National Instruments) (Puig et al. 2004).

Operating conditions and cycle description

The semi-industrial SBR pilot plant is run with a fixed cycle found by Vives (2004), which optimizes the cycle to achieve complete nitrification and denitrification. The duration of operation stages are fixed. Each cycle takes 8 hours and has 5760 samples (obtained every 5 seconds) per variable. There are six anoxic-aerobic phases of reaction, with filling only occurring during the anoxic condition. The applied operation stages are shown in Figure 2.6. The cycle is divided into 395 minutes of reaction phase, with 46% of aerobic conditions and 54% of anoxic condition, 60 minutes of settling, and finally 25 minutes of draw. The reaction phase is divided into 212 minutes of anoxic conditions and 183 minutes of aerobic conditions (Corominas. 2006).

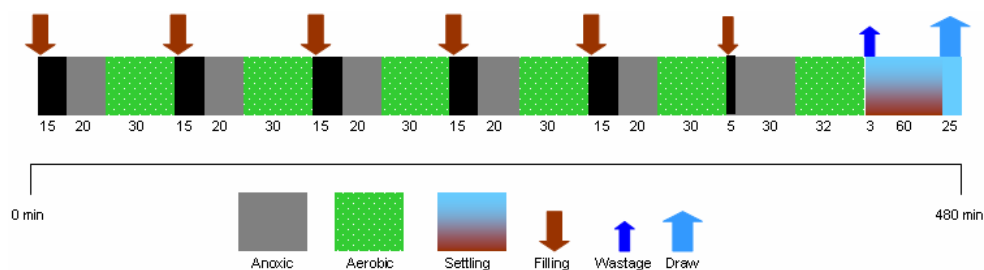


Figure 2.6: Cycle applied to semi-industrial SBR pilot plant

Re-sampling and time warping

The plant ran continuously for 60 days. Each batch took 8 hours with 5760 samples for each variable (one sample every 5 seconds). Due to computer limitations only 392 samples per batch are used. To test whether the samples per variable could correctly determine the operation of the process, the profile of each variable was studied. The profiles are important because they contain important points that provide valuable information about the beginning and ending of the biological reactions. In this way, one sample for each minute is considered. The 5760 and 392 time instants are contrasted in Figure 2.7, to

verify that the variables profiles do not change. Settling and drawing have not been taken into account because they are usually not consistent due to changing properties/dynamics of settling in each batch (Sin et al. 2004) (Lee and Vanrolleghem 2003b).

In Figure 2.7, the first variable is ORP for both lengths, the normal range of values is around $-300mV$ in anoxic stages and 0 to $50mV$ in aerobic stages. In anoxic stages, there is a bending-point called the nitrate knee. It occurs when the denitrification reaction has finished; this is perceived in both profiles. The third variable is pH, which has two important points that provide information about the end of nitrification and denitrification. Comparing both profiles, it can be seen that the profiles are equal. This implies that the SBR biological process changes slowly.

2.2.2 Lab-Scale Plant SBR at University of Girona (LEQUIA)

Generalities

The lab-scale plant SBR is located at the University of Girona (Catalonia-Spain). The maximum capacity of this SBR pilot plant is 30 liters, and the minimum operating capacity is 20 liters (see Figure 2.8). This minimum capacity is the residual volume at the end of each SBR cycle. The influent wastewater is synthetic. It is a blend of carbon source, an ammonium solution, a phosphate buffer, alkalinity control and microelements solution. The influent wastewater is stored in a tank with a capacity of 150 liters. The temperature in the storage tank is $4^{\circ}C$ to minimize microbial activity. The reactor operates in a predefined cycle of fill, reaction, settle and draw modes. This reactor is located in a thermoregulated room at $20^{\circ}C$.

The influent wastewater is transferred from the storage tank to the reactor by means of a peristaltic pump (Watson Marlow). Similar peristaltic pumps are used to fill, purge and draw. The sludge and wastewater are mixed under homogeneous conditions. For this purpose, a marine helix is used with a nominal value of 400 rpm. The reactor is operated under anoxic and aerobic conditions. Injecting compressed air creates aerobic conditions, without dissolved oxygen control. The compressed air is injected at the bottom of the reactor. The dissolved oxygen is controlled inside the reactor by means of an electrovalve. When the reaction has finished, the settling phase starts to separate the sludge from the treated water, decanting at the bottom of the reactor. Finally, the treated water is discharged. To monitor essential variables the SBR process is equipped with DO (WTW OXI 340), Temperature (PT 100), pH (EPH-M10) and ORP (ORP M10) Endress-Hauser probes. These sensors are connected directly to the control panel. The signal is processed by a data acquisition and control card PCI-821PG, afterwards sending a digital signal in order to drive the power relay output board controlling the orders to fill, draw, mix and air supply for the process (PCLD-885). The whole process is controlled by software in LabWindows (from National Instruments). This program has a user-friendly user interface which makes it easy to create and change operating cycles.

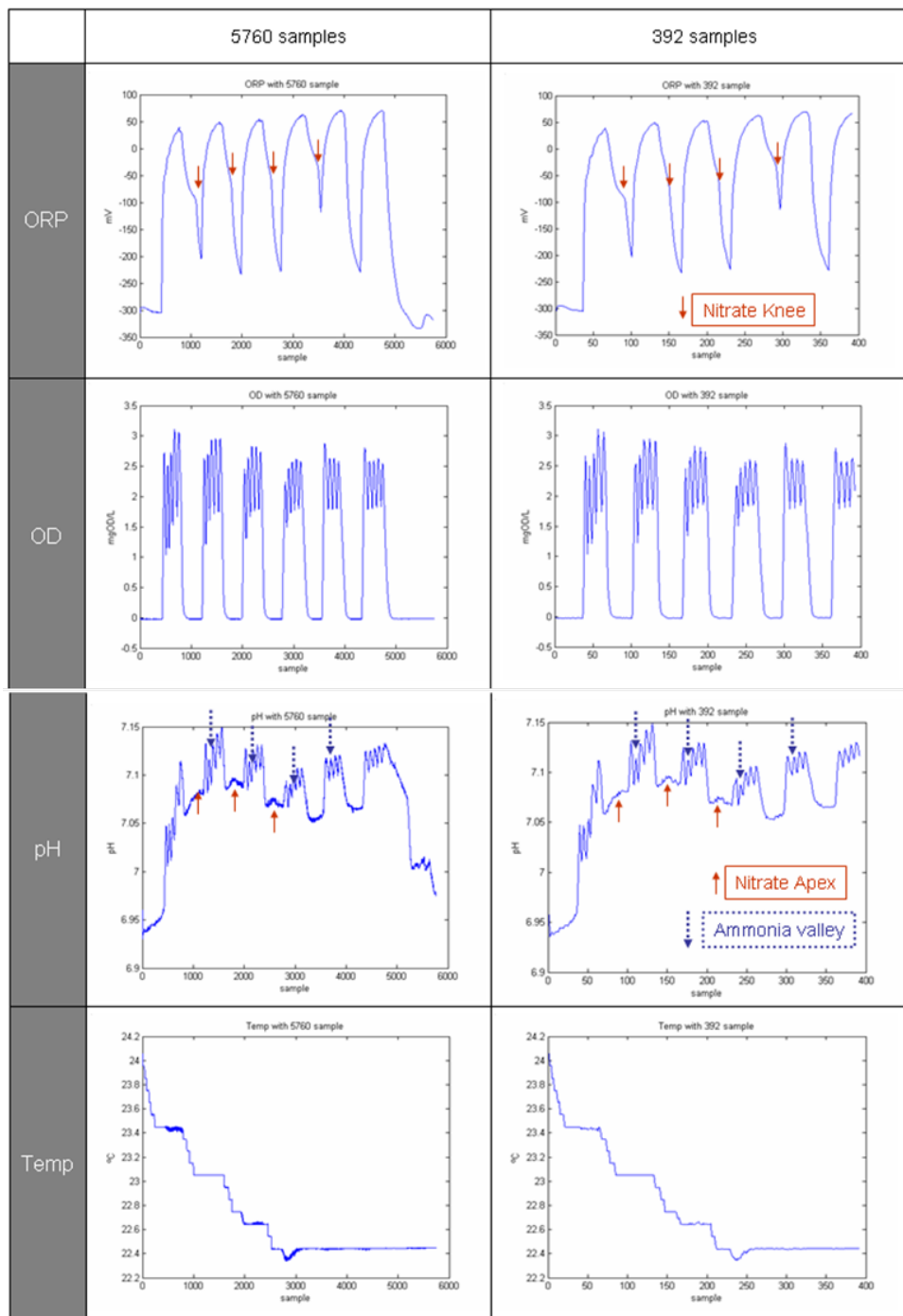


Figure 2.7: Comparison of 5760 samples and 392 samples for variables

Operating conditions and cycle description

The duration of operation stages is fixed. Each cycle takes 8 hours, divided into reaction, settling and discharge. Two combinations of the anoxic and aerobic conditions are implemented in this Lab-Scale SBR Pilot Plant, in which the number of filling events, anoxic and aerobic conditions are alternated.

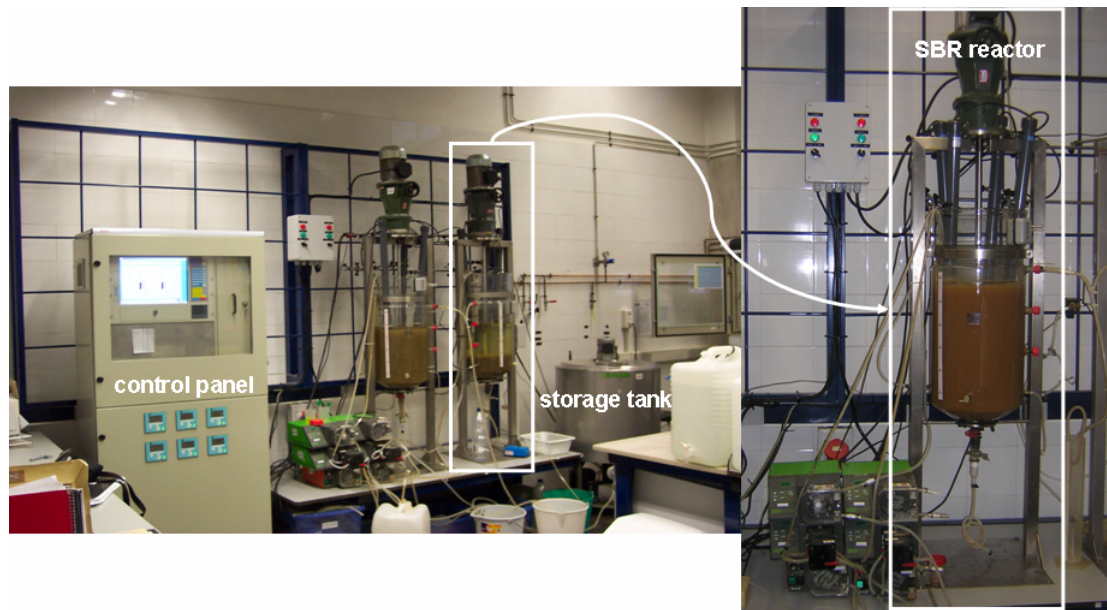


Figure 2.8: Lab-scale plant from LEQUIA

- Period 1: Three reactions are configured. The first reaction phase is a combination of anaerobic and aerobic conditions. The second reaction is a combination of anoxic and aerobic conditions and the third reaction is a different combination of anoxic and aerobic conditions. Filling only occurs at the beginning of each reaction phases (see Figure 2.9).

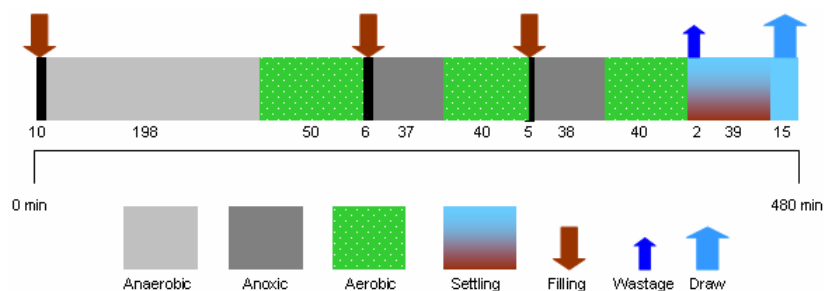


Figure 2.9: Period 1 cycle configuration

- Period 2: This period has only two reactions configured. The final combination of anoxic and aerobic conditions is eliminated from period 1 configuration (see Figure 2.10).

Re-sampling and time warping

The data sets from both periods are contained in text files, representing data retrieved from the wastewater treatment plant during 8 hours working time (duration of a complete cycle of treatment). At the beginning of these files are found several header lines

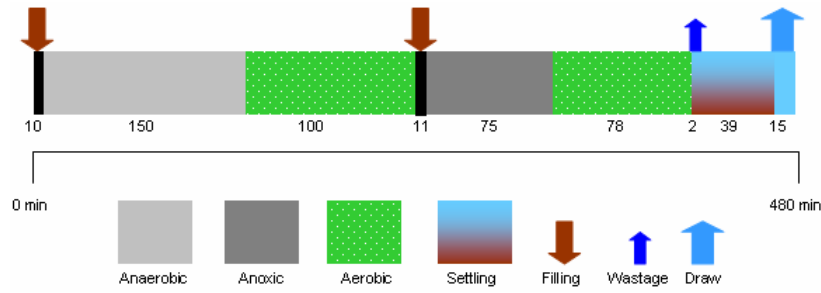


Figure 2.10: Period 2 cycle configuration

including information related to the measured variables, and other information. Next to these header lines, and lasting until the end of the file, are the measured values of the variables with a sample time of 5 seconds. The number of data contained in each file is 5760.

Taking into account the large number of samples in each file and that the variation of the treatment process does not occur suddenly, and in order to reduce the computational load, it is necessary to reduce the number of samples.

Work schedule	3			2								
	A		B	A			B			C		
Phases	SBR2802	SBR1-241105	SBR1912005	SBR12012005	SBRs15022006	04042006 SBRs	190406SBRs	27042006SBRs	30052006SBRs	130606SBRs	030806SBRs	21082006SBRs
Fill 1	10	10	10	10	10	10	15	15	15	15	15	15
Anaerobic	198	198	150	150	150	150	150	150	150	188	188	188
Aerobic 1	50	50	100	100	100	100	100	100	100	100	100	100
Fill 2	6	6	6	11	11	11	6	6	6	6	6	6
Anoxic 1	37	37	37	75	75	75	75	75	75	75	75	75
Aerobic 2	40	40	40	78	78	78	78	78	78	40	40	40
Fill 3	5	5	5	0	0	0	0	0	0	0	0	0
Anoxic 2	38	38	38	0	0	0	0	0	0	0	0	0
Aerobic 3	40	40	38	0	0	0	0	0	0	0	0	0
Wastage	2	2	2	2	2	2	2	2	2	2	2	2
Settling	39	39	39	39	39	39	39	39	39	39	39	39
Drawing	15	15	15	15	15	15	15	15	15	15	15	15

Table 2.1: Work schedule configuration from LEQUIA Lab-Scale Plant SBR

At the same time, the obtained data present several operating plans that are shown in Figure 2.1, notated as 3A, 3B, 2A, 2B and 2C, where the numeric value represents the number of cycles in the process, and the character distinguishes between the different time phases configurations in which the processes packed in the same group are divided. Of those divisions, the last 3 (Wastage, Settling and Drawing), will not be taken into account since the information added is not very important in this study. The strategies studied to reduce the data are:

- Independent data treatment: Each working plan will be treated independent the others, reducing the number of data samples to 1 sample per minute in every phase used. The expression for this new sample value is noted below:

$$X = \frac{\sum_{i=1}^N x_i}{N} \quad (2.1)$$

where:

x : new sample obtained for the actual time period.

X : sample value placed at position i for the actual time period.

N : number of samples needed to form a minute of real time.

One minute of sampling is sufficient since the time constant of biological reactions are in the order of hours hence one sample per minute is selected because this is the maximum value recommended by experts when dealing with biological processes.

- Grouped data treatment: The strategy consist of packing all working plans that share the same number of cycles into one single pack, so the number of samples taken must be unified in order to analyze the data. Two options are analyzed.
 1. Reduction to the minimum value: As in the independent data treatment, the maximum sample frequency allowed when working with biological processes is 1 sample per minute. This criterion will be taken as a reference for the minimum lasting time value, and the values that are greater will be undersampled to this value. As an example, if the different length of the anaerobic phase are taken (Figure 2.2), the number of samples that should be used is 150, so it is the minimum value.

Work schedule	3		
	A		B
Phases	SBR2802	SBR1-241105	SBR1912005
Fill 1	10	10	10
Anaerobic	198	198	150

Table 2.2: Three different lengths for anaerobic phase configuration

When reducing to the minimum value, it is assumed that processes with a shorter length are more critical than others that have a longer length. Also it is deduced that processes with a longer duration have a slower reaction; in order to acquire the same amount of information as in the shorter processes, the sample time must be greater. The formula to compute the new samples of the data is the same as that shown in Equation 2.1, with the number of periods needed to have 1 minute of real time changed as follows:

$$N = \frac{\text{Data from phase}}{\min(d_1, \dots, d_k)} \quad (2.2)$$

where d is the length of the analyzed phase among the working plans that share the same number of cycles.

2. Reduction to the maximum value: This time the more critical processes are those that have a longer length. Those that have a shorter length value must be artificially enlarged. Keeping in mind that all processes have the same sample time (1 sample per minute), it has been decided to use the following strategy:

- Compute the mean values of each time instant of those working plans that have the maximum length.
- For each working plan take 1 sample per minute, using Equation 2.2.
- For processes that have a smaller length than the maximal value, add the mean values of the time instants needed to reach the maximal value at the end of the new samples.

If the mean value is added when computing the data, these new values represent 0, so the new mean will not be affected.

2.2.3 Lab-Scale Plant SBR at Ghent University (BIOMATH)

Generalities

This pilot plant is located at Ghent University in Belgium. The maximum capacity of this SBR pilot plant is 64 liters in which synthetic sewage is used as influent wastewater which mimics real pre-settled domestic wastewater (see Figures 2.11, 2.12). Detailed information about the synthetic influent wastewater characterization can be found in (Insel et al. 2006).



Figure 2.11: Lab-scale plant from BIOMATH

The system consists of a PC, an analog/digital interface card, sensors, transmitters and Solid State Relays (SSR). This system controls the on-off status of the parasitical pumps, air supply and mixer and the duration of each phase; it also has a friendly interface (Lee et al. 2005). The data acquisition, pump and valve control loops are programmed in the

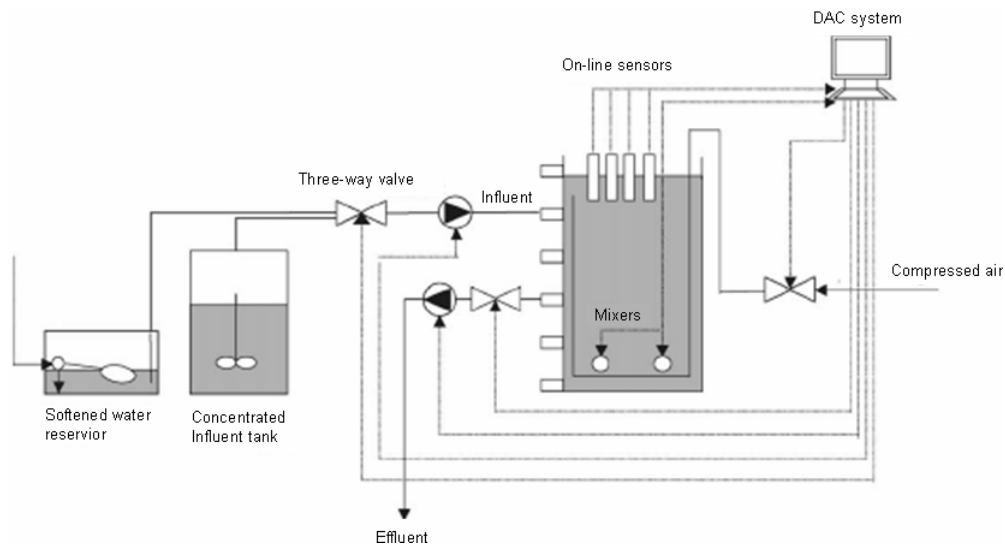


Figure 2.12: Operational scheme of the SBR

LabView platform. The sensors for pH, ORP, DO, temperature, weight and conductivity are connected to the individual sensors. These measurements are recorded every 1 min (360 time instants per cycle). The operating conditions are displayed on the computer, and the collected data is stored in a data log-file. The aeration is controlled by means of an on-off valve (Insel et al. 2006).

Operating conditions and cycle description

The pilot-scale SBR operation consists of 6-hour cycles (i.e., 4 cycles per day). The scheduling of phases, optimized in (Sin et al. 2004), is shown in Figure 2.13. The fill phase comprises minutes 1 to 60 of each cycle. In the reaction phase (minutes 61 to 270) the operation is switched 4 times between aerobic (20 minutes) and anoxic conditions (32.5 minutes). The last aerobic phase from 271 to 300 minutes is followed by the settling phase (45 minutes) and a draw phase (15 minutes). The excess sludge is wasted at the end of the second aerobic phase for each cycle.

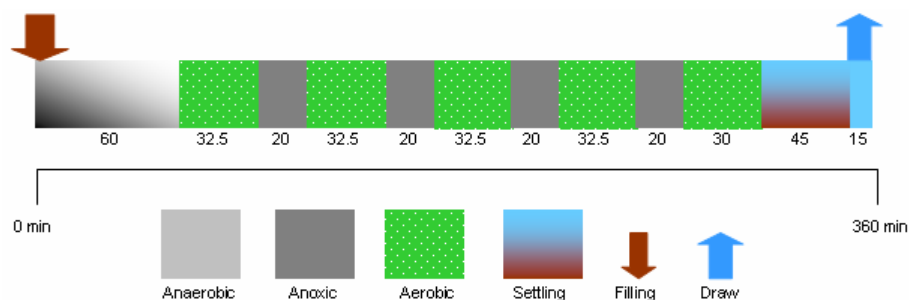


Figure 2.13: Cycle applied to lab-scale plant SBR

Re-sampling and time warping

The measurements of pH, ORP, DO, temperature, weight and conductivity are recorded every minute, resulting in 360 measurements per variable per cycle in one batch run. However, only the first 300 time instants of each batch run are used (Sin et al. 2004) (Lee and Vanrolleghem 2003b).

2.3 Conclusions

Wastewater treatment consists of the elimination of contaminants in the water. The treatment used depends on the type of process. The process can be structured into three main blocks based on their nature: physical, chemical and biological. All three are based on the separation of wastewater in two phases, one containing clean treated water and another containing solids. In this work two kinds of plants were used: a benchmark simulation plant with a continuous activated sludge process configuration (BSM1) and the other is sequential batch reactor (SBR) with three different scales and configurations (1 semi-industrial pilot plant, 1 lab-scale for COD and N removal and 1 lab-scale for COD, N and P removal). While in benchmark simulation the treatment occurs in several reactors, in SBR the whole process occurs in the same reactor following a sequence of phases. The main goal in benchmark simulation is determine the best control strategies including a plant layout, simulation models and parameters, as well as a detailed description of the influent disturbances. The main goal in the SBR process is to combine the filling and reaction phases in different ways and several times. Depending on these combinations, the organic matter and nitrogen are removed.

In next chapters the characteristics, operating conditions and requirements associated with each one of the these plants will be described. Then, explanations about proposed tools to evaluate the performance of a methodology to detect and diagnose normal and abnormal operation condition are described. Several tests are developed in order to obtain the best methodology.

Chapter 3

Multivariate Statistical Process Control

3.1 Preview

Many strategies for monitoring, fault detection and diagnosis are referenced in the bibliography. According to Venkatasubramanian et al. (2003), fault diagnosis methods can be classified in three general categories: quantitative model based methods, qualitative model based methods and process history based methods, as illustrated by Figure 3.1.

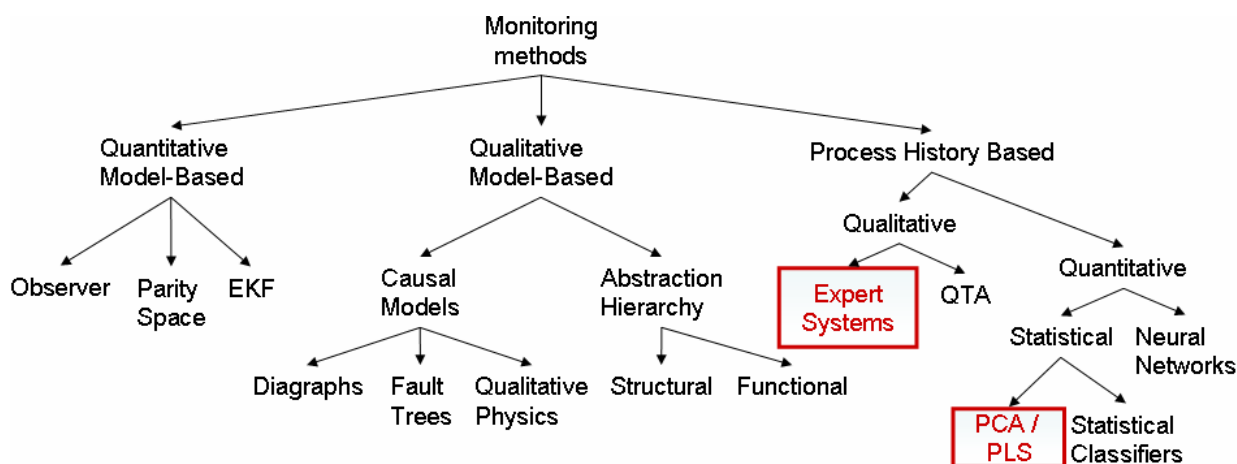


Figure 3.1: Classification of monitoring, fault detection and diagnostic algorithms

The solution proposed in this thesis falls into the category of process history based methods and specifically into the subgroups of statistical methods and expert systems (see Chapter 4). The history of statistics dates back to the Egyptians, where the pharaohs gathered information about the population and wealth. Later on, the Roman Empire improved these techniques. They carried out a census of the population every five years and recorded births and deaths. In the Middle Ages, this practice was forgotten until it was revisited by men such as Leonardo de Vinci, Nicholas Copernicus, Galileo, Neper, William Harvey, Sir Francis Bacon and Rene Descartes (Marte 2003). Between 1800

and 1820, two fundamental concepts were generated for Statistical Theory: Laplace and Gauss developed the theory of probability and the least squares approximation method (Schuldt 1998). In the late nineteenth century, Sir Francis Gaston developed the Correlation Method which measures the relative influence of factors on variables, and it led to the development of the coefficient of correlation by Karl Pearson. Other important cultivators of Biometry Science, such as J. Pease Norton, R. H. Hooker and G. Udny Yule, carried out studies of the Measure of Relations.

Statistical process control (SPC) began with Walter Shewhart in the 1920s. He emphasized the importance of adapting management processes to create profitable situations for both businesses and consumers, promoting the utilization of *the SPC control chart* (Hare 2003). SPC eventually became more than the application of control charts and it began to be used in manufacturing processes. Harold Dodge, Harry Romig, W. Edwards Deming and Eugene Grant have been other important developers. Eugene Grant is the author of the classic text *Statistical Quality Control*, first published in 1946. During this time, the formation of control chart limits had been transformed from Shewhart's original concept of economic limits to probability limits usually based on group variation. The term SPC has become much more than the application of control charts alone. Topics such as acceptance sampling, data analysis and interpretation, and managing for quality have been gathered into the discipline (Hare 2003).

The problems of modern processes are highly complex and operate using a large number of samples and variables, which will increase even more with further developments in sensor technology. Therefore, the control model must consider the amount and the correlation structure between variables (Ferrer 2003), characterized by the covariance matrix, that arise due to the existing relationship between variables and processes. When Statistical Process Control is used within batch processes, false alarms are often generated (Lee and Vanrolleghem 2003b). Fortunately, this problem can be solved using Multivariate Statistical Process Control (MSPC). MSPC compresses the multidimensional information into a few latent variables which explain the variability of the measured variables, including their relationships. This chapter contains a description of SPC, with an explanation of MSPC techniques, particularly Principal Component Analysis (PCA). MSPC has been widely used in different fields of science, mathematics, medicine, chemistry and biological processes, among others. With regard to the last field, a review of applications in biological processes is given.

PCA is a tool for data compression and information extraction which finds combinations of variables or factors that describe major trends in a data set (Wise et al. 1999). The history of PCA goes back to 1933, when Harold Hotelling linked Hotelling T^2 statistics with principal components. He precisely formulated the idea of a component based in the mathematical knowledge, pointing out the implications, setting forth computational procedures, and discussing statistical inference. Six decade after, Nomikos and MacGregor suggest the use of statistical models for monitoring batch process within the framework of MSPC (Nomikos and MacGregor 1994a)(Nomikos and MacGregor 1994b). The normal process behavior is captured in the statistical model, which is trained on a historical data set reflecting the normal operation conditions (NOC). Future observations

are projected onto that model and the resulting statistics is checked against its in-control limits. One of the most important benefits of such an approach for process monitoring is that no detailed mechanistic knowledge is required to assess whether the process is operating in its normal condition. In addition to monitoring of full batches, the progress of a new batch can be monitored as well while running. PCA has been increasingly used in diverse fields, from medical research (Ondusi et al. 2005), (Palmer et al. 2003), (Das et al. 2004) to eco-hydrological studies (Gonzalez-Silvera et al. 2004), in hydraulics (Zhan et al. 2004), structural health monitoring (Mujica et al. 2008) and spectroscopy (Stadlthanner et al. 2004).

PCA applications in wastewater treatment are relatively recent. Rosen and Olsson (1998) demonstrated the applicability of statistical models for the detection of process disturbances, making a comparison between PCA and PLS modeling. In the literature, integration of PCA techniques with other data-driven modeling techniques is common. For instance, Yoo et al. (2003) integrated PCA with adaptive credibilistic fuzzy-C-mean (CFCM) adaptive discriminant monitoring index and a Takagi-Sugeno-Kang (TSK) fuzzy model to predict the important output variables in a full-scale WWTP which treats cokes wastewater from an iron and steel making factory. Recently, Grieu et al. (2005) integrated multi-layer neural networks, K-means clustering and PCA to estimate the process quality and efficiency in the Saint Cyprien WWTP (France). First, the data were treated by K-means clustering and in turn PCA was used to improve the results of the next step, the neural network training. The main advantage of the PCA application is elimination of redundancies and correlation from the data set, which results in a better convergence in the neural network training step. Another hybrid approach is shown in Singh et al. (2005) where Cluster Analysis (CA), Discriminant Analysis (DA), PCA and PLS were used to analyze the composition of wastewater. The CA generated six groups of drains on the basis of similar characteristics. PCA was then used to extract information on seasonal variations and differences between domestic and industrial wastewaters. PLS-DA was applied to determine the most important (i.e., most discriminating) characteristics of the studied wastewater. Several extensions of PCA-based process monitoring have been reported in the literature. Among others, these extensions are denoted as 'Adaptive', 'Dynamic' (DPCA), 'Kernel' (KPCA), 'Multi-block' (MBPCA), 'Multi-phase' (MPPCA), 'Multi-scale' (MSPCA), 'Multi-way' (MPCA) or various combinations of these. As an alternative to PCA, Independent Component Analysis (ICA) and Kernel PCA have been applied to process monitoring by Lee and Dorsey (2003), Lee, Yoo and Lee (2004a) and Lee, Yoo and Lee (2004b). Several applications of the aforementioned extensions and alternatives are reported in the wastewater treatment field. Lee, Yoo and Lee (2004a) compared the application of PCA and ICA fault detection to a benchmark simulation of a WWTP. Lee, Yoo and Lee (2004b) compared PCA, Dynamic PCA, ICA and Dynamic ICA for process monitoring of a simulated multivariate dynamic process. Lee and Dorsey (2003) evaluated the integrated application of adaptive, multiblock, multiway PCA to identify the major sources of process disturbances in a pilot-scale Sequencing Batch Reactor (SBR) for biological nutrient removal. It is claimed that the adaptive structure allows accounting for non-linear process variation, while the multi-block approach allows for systematically identifying the phase(s) which the eventual disturbances occur. Ruiz, Colomer, Rubio, Melendez and Colprim (2004) Multi-block was used as monitoring tool

for a SBR wastewater system. Rosen and Lennox (2001) proposed during the application of wavelet transformations to account for process dynamics at different time-scales which resulted in a Multi-Scale PCA (MSPCA) model. Lee et al. (2005) applied multi-scale, adaptive MPCA to detect and to analyze a wide range of faults and disturbances in a pilot-scale WWTP. All variable trajectories were subjected to wavelet decomposition before PCA modeling and for each resulting scale an adaptive MPCA model was developed. Adaptive modeling refers to the automated updating of the covariance structure to deal with acceptable process changes. Yoo et al. (2004) applied MPCA and Multiway ICA (MICA) to monitoring of a WWTP and explains the calculation of the statistical confidence limits for the IC scores based on kernel density estimation. Aguado et al. (2006) compared different predictive models for a SBR WWTP: Principal Component Regression (PCR), Partial least Squares (PLS) and Artificial Neural Networks (ANNs) as well in (Aguado, Ferrer, Ferrer and Seco 2007) and (Aguado, Ferrer, Seco and Ferrer 2007) applied PCA to find the best way for modeling SBR process. In consequence, MSPC has been recently started to use as a tool for monitoring with successful results. In this way, this chapter contributes to development of this potent approach in order to detect abnormal situations in WWTP.

3.2 Univariate Statistical Process Control

The objective of SPC is to monitor a process over time in order to detect statistically significant events or abnormalities (Lennox 2003). A univariate statistical method can be used to determine the thresholds for each observation variable (a process variable observed through a sensor reading), where these thresholds define the boundary, and any violation of these limits indicates fault (Keats and Hubele 1989).

This demarcation typically employs the *Shewhart chart* (Russell et al. 2000) (see Figure 3.2) which has a baseline or central line L_0 , two lines L_1, L_2 (UCL) above L_0 and two lines L'_1, L'_2 (LCL) below L_0 . Some of the suggested rules for taking action are one or a combination of the following depending on the configuration of the successive plotted points (Rao 1973):

1. If a point falls above L_1 or below L'_1
2. Two successive points between L_1, L_2 or between L'_1, L'_2
3. A configuration of three points such that the first and third are between L_1, L_2 and the second between L_0, L_2 , and equivalent situation with respect to L_0, L'_2, L'_1

Measurements are plotted on the chart against time. The baseline for the control chart is the accepted value, an average of the historical standard values. A minimum of 100 standard values is required to establish an accepted value. The upper *UCL* and lower *LCL* control limits are:

$$UCL = \text{Accepted value} + k^* \text{ process standard deviation}$$

$$LCL = \text{Accepted value} - k^* \text{ process standard deviation}$$

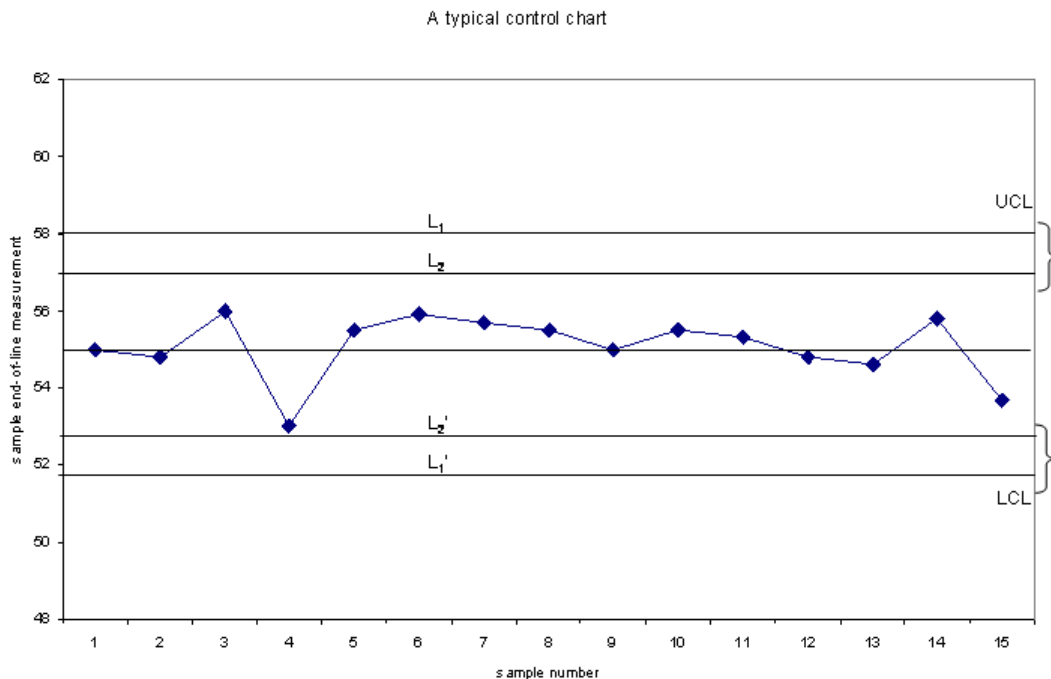


Figure 3.2: An illustration of the Shewhart chart. The rhombuses are observations. The process is said to be 'in control'

where the *process standard deviation* is the standard deviation computed from the standard database. The interest is in assessing individual measurements (or averages of short-term repetitions). Thus, the standard deviation over time is the appropriate measure of variability (NIST 2003). Generally, the control limits are chosen to be $\pm 3\sigma$ (Colomer et al. 2000). Montgomery (2000) showed another control chart which uses the same range.

Univariate control chart monitoring does not take into account that variables are not independent of each other and their correlation information can be important for understanding process behavior. In contrast, multivariate analysis takes advantage of the correlation information and analysis the data jointly (Chen 2001). The difficulty of using independent univariate control charts is illustrated in Figure 3.3. In this figure, the ellipse represents a contour for the in-control process with high confidence limits; circles and triangles represent observations from the process. Individual Shewhart charts are plotted for each quality variable, and it is observed that each individual Shewhart chart appears in a state of statistical control, and none of the individual observations gives any indication of a problem (Chen 2001) because the univariate statistical charts do not consider the information contained in the other variables and in the dynamic dependencies of the quality variables (Barcel and Capilla 2002).

Cumulative sum (CUSUM) and Exponentially Weighted Moving-Average Chart (EWMA) are other procedures for a single variable (Cinar and Undey 1999). CUSUM charts incorporate all the information a data sequence to highlight changes in the process average level, and are effective with samples of variable. EWMA is a weighted average of several

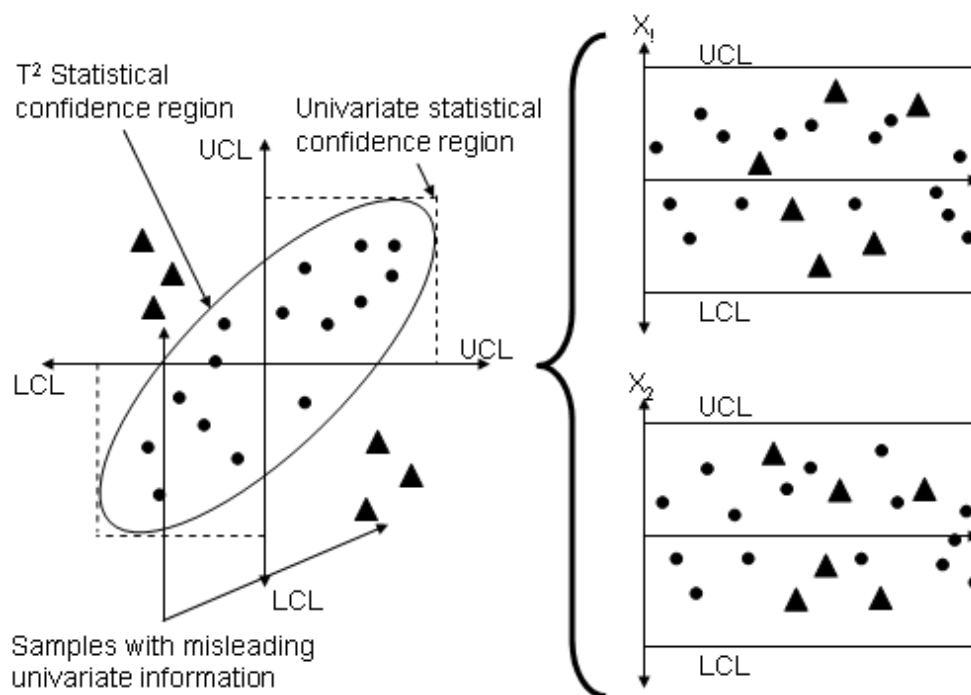


Figure 3.3: Multivariate statistical analysis vs. univariate statistical analysis and a comparison of the in-control status regions using T^2

consecutive observations, which is insensitive to non-normality in the distribution of the data (Cinar and Undey 1999). EWMA is also known as geometric moving average, exponential smoothing or first order pole filter. However, these charts (Shewhart, CUSUM and EWMA) do not consider the relation contained in others variables.

T^2 Statistics or so-called Hotelling's T^2 takes into account the correlations between the variables. T^2 is used as a tool for fault detection (Norvilas et al. 1998). T^2 is based on the level of significance (α), where α specifies the degree of trade-off between the false alarm rate and the missed detection rate, so T^2 can be determined by assuming that the observations are randomly sampled from a multivariable distribution (Russell et al. 2000). This is represented in Figure 3.3 as an elliptical confidence region. The number of samples and variables has been further increased and the processes are highly complex (Kourti 2003b) (Castell et al. 2002). Because of this, projecting the data onto a lower dimensional space that accurately characterizes the state of the process has been developed. These techniques of dimensionality reduction can greatly simplify and improve process monitoring procedures. PCA and Partial Least Square (PLS) are dimensional reduction techniques. These methods address all of the above problems and provide analysis results that are easy to present and interpret. In the same way, CUSUM and EWMA have versions for multivariable analysis.

3.3 Multivariate Statistical Process Control

Businesses have different goals: utilities (outsources, consultations and commercial), industrial or production (changes of raw material), financial (banks, securities) and virtual. All these businesses manage a great quantity of information and have large volumes of historical data stored in databases. Exploitation of these data is a critical component in the successful operation of any industrial process over the long term, however, until a decade ago, nothing has been done with them, due to the nature of these data. This amount of data is enormous and often highly correlated. To utilize this data, a database must be able to deal effectively with all these difficulties. Research has been focused on developing models by using latent variable methods such as PCA and PLS (Kourti 2002). Another method is autoregressive moving average ARMAX. This model can accurately represent a high order ARX model containing a large number of parameters, where the ARX model is the mathematical relation between the output at time t and the past h inputs and outputs. To avoid the problems of the classical approach, a class of system identification methods for generating state space models, called subspace algorithms, has been developed in the past few years. The most common subspace algorithms are: numerical algorithms for subspace state space system identification (N4SID), multivariable output-error state space (MOESP) and Canonical Variate Analysis (CVA). The CVA algorithm is actually a dimensionality reduction technique in multivariate statistical analysis involving the selection of pairs of variables from the inputs and outputs that maximizes a correlation measure (Russell et al. 2000). MSPC has been applied in different areas, including diversification of the financial system (Skonieczny and Torrisi 2003), applications in medicine (Ambroisine et al. 2003), semiconductor processes (Wise et al. 1999) (Li et al. 2000), desulphurization process (Dudzic and Quinn 2002), and monitoring of a bioprocess (Cimander and Mandenius 2002).

3.3.1 Principal Component Analysis

PCA is the favorite tool of chemometricians for data compression and information extraction which finds combinations of variables or factors that describe major trends in a data set (Wise et al. 1999). The aim of PCA is to describe a given data-set in a space whose dimension is smaller than the number of variables, in order to easily visualize similarities and differences. In Figure 3.4 three process variables are represented in which two principal components have been calculated.

That is, PCA is concerned with explaining the variance-covariance structure through a few linear combinations of the original variables. Its general objective is a reduction of dimensionality, which means to produce a lower dimensionality in which the correlation structure between the process variables is preserved (Russell et al. 2000).

The multivariate data can be organized in m variables and n samples per variable as is defined in Equation 3.1:

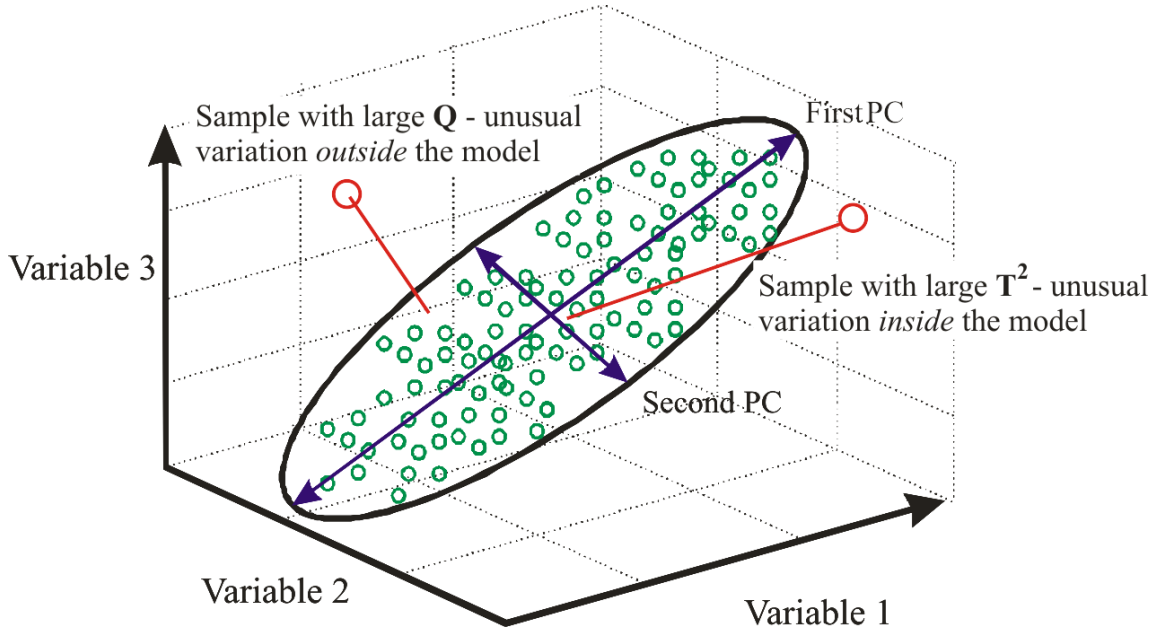


Figure 3.4: Projection of the process variables in a new space using PCA

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1m} \\ x_{21} & x_{22} & \cdot & \cdot & \cdot & x_{2m} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot \\ x_{n1} & x_{n2} & \cdot & \cdot & \cdot & x_{nm} \end{bmatrix} \quad (3.1)$$

X can be decomposed into the noise or residual part (\underline{E}) and the systematic part which is conformed by scores T and loadings P according to (Nomikos and MacGregor 1994a) (Kourti 2002):

$$X = TP^T + E = t_1p_1^T + t_2p_2^T + \dots + t_Np_N^T + E = \sum_{n=1}^N t_n p_n^T + E \quad (3.2)$$

where N is the number of principal components selected for the model t_n and p_n . Before applying PCA, it is necessary preprocess the data matrix \mathbf{X} (Martin et al. 2002). Several studies of this have been presented in the literature (Law et al. 1984), (Westerhuis et al. 1999) and (Gurden et al. 2001). Preprocessing refers to any transformation of the original data set which is performed before developing the main analysis model. Law et al. (1984) describes several reasons for preprocessing as follows:

1. To adjust the data set for the model by removing unwanted conditions. This condition means that the variables of the data set must be strictly proportional between themselves.
2. To emphasize relationships among patterns of change of the data set.

3. To weight or ignore particular data during the analysis or inversely to make equal the influence that different variables have on the form of the final solution.
4. To make equal the size of presumed error variance across the data set.
5. To standardize the data that later can be compared in an uncomplicated manner.
6. To standardize the data so that useful interpretations are possible.

Two classes of preprocessing are distinguished: data conversion and data adjustment (Law et al. 1984).

1. Data conversion that transforms one kind of data into a different kind, implicating a change in the form of the model. Two types of data conversion principles are identified:
 - Profile conversion of the data into covariances.
 - Multidimensional scaling as scalar products.
2. Data adjustment that does not require any change of the model that will be after performed. Two types of data adjustment principles are identified:
 - Centering the data, which is an additive adjustment. The mean of the data set is placed to the origin or zero-point.
 - Rescaling or normalizing the data, which is a multiplicative adjustment. The data is adjusted in such a way that the variance is unity.

Sometimes a third type of adjustment can be used: linearizing the data, which is a nonlinear adjustment.

Once the variables have been standardized, two algorithms for performing PCA can be performed: the Singular Values Decomposition (SVD) and Non-linear Iterative Partial Least Squares (NIPALS).

Singular Values Decomposition (SVD) algorithm

First, the X covariance matrix is calculated (equation 3.3.)

$$S = \frac{1}{n-1} X^T X \quad (3.3)$$

The matrix P, in the columns, are the eigenvectors of S, and the diagonal matrix λ , with eigenvalues of S on the main diagonal are found in equation 3.4:

$$S\hat{P} = \hat{P}\lambda \quad (3.4)$$

Each eigenvalue is associated with an eigenvector. The eigenvector with the highest eigenvalue represents the most important pattern in the data, i.e., it contains the largest quantity of information. Thus this vector is called the principal component of the data set.

Ordering the eigenvectors by their eigenvalues, highest to lowest, gives the components in their order of significance. In order to reduce the dimensionality, the less important components can be eliminated (information is lost, but if the eigenvalues are small, this information loss is minimal), then only the first n eigenvectors are chosen (loading vectors and denoted by P) and the final data set will be n -dimensional. The projected matrix T (or score vectors) in the new space is defined in equation 3.5:

$$T = XP \quad (3.5)$$

Projecting T over the new-dimensional space results in:

$$\hat{X} = TP^T \quad (3.6)$$

where the difference between X and \hat{X} is the residual matrix E :

$$X = \hat{X} + EX = TP^T + E \quad (3.7)$$

Non-linear Iterative Partial Least Squares (NIPALS)

NIPALS is the most common algorithm used to calculate the principal components. This algorithm was developed by Herman Wold first to solve PCA problems and later to generate models using PLS. Fundamentally, the method starts when one column is selected from matrix X as the first principal component t_1 . In this manner, the iterations begin using the "alternating least squares" method and the corresponding loading value p_1 is calculated using the "regression" method. Next step removes the variance explained by this component and the process is repeated for obtaining the next component.

```

X1=X, T0=[ ], P0=[ ], tol = 0.001
σx2=∑i=1m(σ2[X(:,i)])
is k=1 then lv
u=Xk(:,1), conv=1
while tol < conv
pk=Xkt u (utu)-1
pk= $\frac{p_k}{\|p_k\|}$ 
tk=Xk pk
conv = (tk - u)t (tk-u)
u=tk
end
Tk=[Tk-1 | tk]
Pk=[Pk-1 | pk]
Vk= $\frac{100\sigma(t_k)}{\sigma_x^2}$ 
Xk+1=Xk - tkpkt
end

```

Figure 3.5: NIPALS algorithm

The NIPALS algorithm calculates the matrix T (scores or principals components), matrix P (loadings) and V (variance captured per each principal component)

Control Charts

The score matrix and the residual matrix can be used in order to detect abnormal operation in a process. With this aim, the Q-or SPE-statistics and the D-statistics (Hotelling T^2 statistics) are used. These methods are based on the assumption (generally stemming from the central limit theorem) that the underlying process follows approximately a multivariate normal distribution where the first explain vector is zero (see Figure 3.4).

Multivariate control charts based on T^2 can be plotted based on the first N principal components as follows (MacGregor 2003):

$$T^2 = \sum_{j=1}^N \frac{t_j^2}{\sigma_{t_j}^2} = \sum_{j=1}^N \frac{t_j^2}{\lambda_{t_j}} \quad (3.8)$$

where

$$\sum_{i=1}^N \sigma_i^2 = \sum_{i=1}^N \lambda_i \quad (3.9)$$

This control chart will only detect variation in the plane of the first N principal components which are greater than what can be explained by the common-cause variations.

The statistical confidence limit for T^2 can be calculated by means of the equation

$$T_{k,m,\alpha}^2 = \frac{k(m-2)}{m-k} F_{k,m-k,\alpha} \quad (3.10)$$

where m is the number of samples used in the PCA model, k the number of principal components included in the model and $F_{k,m-k,\alpha}$, the value for the F -distribution for k and $m-k$ degrees of freedom and a given α .

When a new type of special event occurs which was not present in the in-control data used to build the PCA model, the new observations will move off the plane. This type of event can be detected by computing the Q -statistic or Squared Prediction Error (SPE) of the residual for new observations. It is defined as (MacGregor 2003) (Yoon and MacGregor 2000):

$$Q_X = \sum_{j=1}^N (x_j - \hat{x}_{j,new}) = (x_{new} - \hat{x}_{new})^T (x_{new} - \hat{x}_{new}) \quad (3.11)$$

where \hat{X}_{new} is computed from the reference PLS or PCA model using Equation 3.6 (Kourti 2002). Normally, Q -statistics is much more sensitive than T^2 . This is because Q is very small and therefore any minor change in the system characteristics will be observable. T^2 has great variance and therefore requires a great change in the system characteristics for it to be detectable.

The Q confidence limits can be calculated according to the equation

$$Q_\alpha = \Theta_1 \left[\frac{c_\alpha \sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0 (h_0 - 1)}{\Theta_1^2} \right]^{\frac{1}{h_0}} \quad (3.12)$$

where

$$\Theta_i = \sum_{j=k+1}^n \lambda_j^i \quad \text{for } i = 1, 2, 3 \quad (3.13)$$

and

$$h_0 = \frac{2\Theta_1\Theta_3}{3\Theta_2^2} \quad (3.14)$$

where k is the number of principal components retained in the model, λ_j is the eigenvalue associated with the j th principal component, n is the total number of principal components and c_α is the standard normal deviation corresponding to a given α . It is assumed for the validity of the Q statistics and its confidence limits that the measurement errors are independent and normally distributed and that the rank of the PCA model is correct.

Information about the control sample can also be obtained from the plot of scores for the relevant principal components. When there is a change in the system, the scores of the new spectrum for the control sample will be very different from the previous scores, and the change will be detected. However, this information is also included in the Hotelling T^2 statistics since it is calculated using the scores. Moreover, the Q statistics gives us additional information which is not included in the scores plot, because it is related to the variation which is not accounted for by the model, and the plots of Q and T^2 are a hypothesis test which clearly signals any out of control sample whereas the inspection of the scores plot is a qualitative tool (Rius et al. 1997).

3.3.2 Dynamic Principal Component Analysis

When the data contains dynamic information, applying PCA on the data will not reveal the exact relations between the variables. Alternative PCA extension can be used augmenting each observation vector with the previous V observations and stacking the data matrix in the following manner:

$$X = \begin{bmatrix} x_k^T & x_{k-1}^T & \cdot & \cdot & \cdot & x_v^T \\ x_{k-1}^T & x_{k-2}^T & \cdot & \cdot & \cdot & x_{k-v}^T \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & & & \cdot \\ x_{k+v-n}^T & x_{k+v-n-1}^T & \cdot & \cdot & \cdot & x_{k-n}^T \end{bmatrix} \quad (3.15)$$

Where x_k^T is the m -dimensional observation vector in the training set at time k interval. Afterwards, the monitoring can be performed using normal PCA on the data matrix in equation 3.15.

3.3.3 Partial Least Squares

PLS, also known as **Projection to Latent Structures**, is a dimensionality reduction technique that maximizes the covariance between the predictor matrix X (identical to PCA, see equation 3.1) and the predicted matrix Y for each component of the space (Russell et al. 2000).

$$Y^T = \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & 1 \end{bmatrix} \quad (3.16)$$

The predicted matrix $Y \in R^{n \times m}$ (m is the number of variables and n is the number of samples) contains the fault location or quality index. There are two possibilities for creating Y (Wold et al. 1987):

1. PLS1: each of the p predicted variables is modeled separately, resulting one model for each class
2. PLS2: all predicted variables are modeled simultaneously.

PLS requires calibration and prediction steps. The goal of PLS is to determine the loading and score vectors which are correlated with Y while describing a large amount of the variation in X (Cimander and Mandenius 2002). The most popular algorithm used in PLS to compute the parameters in the calibration step is **NIPALS** (Appendix A) (Wise et al. 2003) (Kourti 2002)

The PLS is achieved by decomposing X and Y into a combination of loadings P and Q (these are determined by orthogonal vectors), scores T (the projections of the loading vectors associated with the first singular values), weights W and residual matrices E and F (Wise et al. 2003) (Martin et al. 2002).

$$X = TP^T + E \quad (3.17)$$

$$Y = TQ^T + F \quad (3.18)$$

The matrix product TP^T can be expressed as the sum of the product of the score vectors t_j (the j^{th} column of T) and loading vectors p_j (the j^{th} column of P). Similarly, Y is decomposed as the sum of the product of the score vectors t_j (the j^{th} column of T) and loading vectors q_j (the j^{th} column of Q) (Russell et al. 2000).

$$X = \sum_{j=1}^N t_j p_j^T + E \quad (3.19)$$

$$Y = \sum_{j=1}^N t_j q^{T_j} + F \quad (3.20)$$

where N is the number of principal components deemed to be significant. Control charts can also be applied as in PCA.

For example, fault location in the number of electrical distribution systems where the matrix X has a size of 100×7 where 100 are voltage sag cases (voltage sags are a short duration reduction in *rms* voltage), which are registered faults in a $25kV$ Spain Electrical Facility, and 7 corresponds to descriptors of voltage sags. The descriptors are: three phase sag magnitude, three phase sag duration, starting time, ending time and the minimum, the maximum and the average PN-factor (PN-factor is the difference between positive-sequence and negative-sequence of voltage (Zhang and Bollen 1998)). The goal is to locate voltage sags based on descriptors, so that the matrix Y contains the voltage sags location, $Y \in R^{100 \times 1}$, where *one* denotes voltage sags in distribution and *zero* represents the transmission voltage sags (equation 3.16).

Percent Variance Captured by Regression Model

LV #	---X-Block---		---Y-Block---	
	This LV	Total	This LV	Total
1	84.42	84.42	61.09	61.09
2	10.85	95.27	4.42	65.51

Table 3.1: Principal component extraction of PLS example

The PLS model obtained with this data is given in Table 3.1. The charts Q -statistic and T^2 contain all the information required to identify voltage sags. Figure 3.6 shows that some events are outside the limits. These events have been identified as interruption, overvoltage, and not fault recovery (see Table 3.2). They are not sags, and therefore extracted and a new model is developed. The principal components have incorporated descriptors and location information.

Finally, the two principal component resultants of PLS were introduced into a Neural Network (NN) for voltage sag location (Ruiz, Melendez, Colomer, Sanchez and Castro 2004).

Sample	File	Remark
14	Salt 17-07-2002_10-26-41	interruption
16	Salt 22-10-2002_13-37-50	It is not fault recovery
17	Salt 22-10-2002_13-41-11	It is not fault recovery
18	Salt 22-10-2002_13-53-13	It is not fault recovery
77	Acobsa 10-12-2002_18-55-28	It is not fault recovery
83	Salt_TRI 31-1-2003_15-36-59	overvoltage
93	Salt_tr35 14-02-2003_21-23-56	interruption
95	Salt_tr35 19-02-2003_13-26-52	interruption

a)

Sample	File	Remark
74	Salt 16-12-2002_09-54-00	interruption
94	Salt_tr35 19-02-2003 13-21-24	interruption
97	Salt_tr35 21-02-2003_23-23-11	interruption

b)

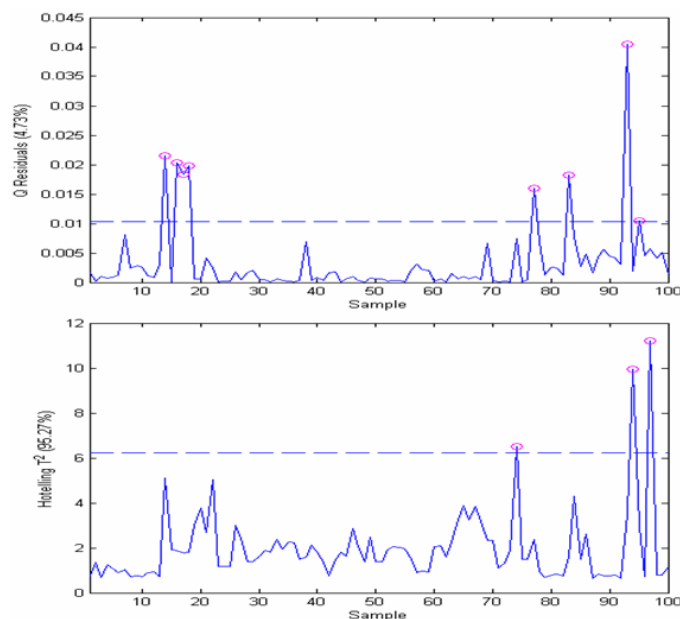


Figure 3.6: Q -statistic and D -statistic with 95.27% confidence limits

Table 3.2: Events exceeding limits a) Q -statistic b) D -statistic

The use of PCA and PLS to build low dimensional models for the analysis and monitoring of process operations is now well established and many large industrial applications exist (MacGregor 2003), but this information corresponds only to one process. At the present time, the processes are becoming more heavily instrumented and data (measurements of process variables) are being recorded more frequently, e.g. in pharmaceutical, chemical, biochemical, microelectronic processes.

3.4 MSPC for Batch Processes

A batch process is an industrial process, such as manufacturing, that goes through a pre-defined cycle and has a definite beginning and end. The processes designed to be time varying: several inputs are added and conditions are altered throughout its run to convert raw materials into a final product (Singh 2003). The trajectories of process variables contain significant information about product quality since the properties of the final product are affected by the operating conditions during the batch (Undey and Cinar 2002) (Kourti 2003a).

Some batch processes include a single step, whereas many others are carried out in a sequence of stages. Events taking place in each step have an impact on the final product yield and quality (Undey and Cinar 2002). The heart of the matter is to achieve high-quality production. The main goal of batch process monitoring is to detect and identify process errors. The bases of MSPC for batch processes are the extensions of PCA and PLS (Nomikos and MacGregor 1994b)(Lee and Vanrolleghem 2003b)(Nomikos

and MacGregor 1994a)(Lee and Vanrolleghem 2003a). These methods have proven particularly suitable for handling noisy and collinear or correlated data (Smilde 2001). Some applications of MSPC for batch processes are: Polymerization (Norvilas et al. 1998), pharmaceutical processes (Lopes et al. 2002), SBR processes (Yoo et al. 2004) (Ruiz, Colomer, Rubio and Melendez 2004), (Ruiz, Colomer and Melendez 2006), (Aguado, Ferrer, Ferrer and Seco 2007), (Aguado, Zarzo, Seco and Ferrer 2007) and industrial batch processes (Kosanovich et al. 1994) (Kourti 2003a) (Flores and MacGregor 2004).

To understand the nature of the data available with which to monitor batch processes, consider a typical batch run in which $j = 1, 2, \dots, J$ variables are measured at $k = 1, 2, \dots, K$ time intervals throughout the batch. Similar data will exist on a number of such batch runs $i = 1, 2, \dots, I$. All the data can be summarized in the $\underline{\mathbf{X}}$ ($I \times J \times K$) array illustrated in Figure 3.7, where different batch runs are organized along the vertical axis, the measurement variables along the horizontal axis, and their time evolution occupies the third dimension. Each horizontal slice through this array is a ($J \times K$) data matrix representing the time histories or trajectories for all variables of a single batch (i). Each vertical slice is a ($I \times J$) matrix representing the values of all the variables for all batches at a common time interval (k)(Wold et al. 1987) (Nomikos and MacGregor 1994a) (Westerhuis and Coenegracht 1997).

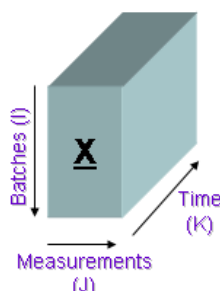


Figure 3.7: Arrangement of a three-way array

3.4.1 Multiway PCA

The objective of MPCA is to decompose the three-way $\underline{\mathbf{X}}$ into a large two-dimensional matrix \mathbf{X} . The method accomplishes this decomposition in accordance with the principle of PCA and separates the data in an optimal way into two parts. MPCA is equivalent to performing ordinary PCA on a large two-dimensional matrix constructed by unfolding the three-dimensional array (Kosanovich et al. 1994) (Duchesne et al. 2003). Six possible ways of unfolding the three-way data array $\underline{\mathbf{X}}$ are indicated in Table 3.3 as suggested by Westerhuis et al. (1999).

When aiming at PCA-based monitoring, unfolding types B and D will lead to models that are equivalent to the models constructed using the C , respectively E unfolded matrices. Matrix F is the transpose of A , and a PCA would simply switch the scores and

Type	Structure	Direction
A	IK x J	variable
B	JI x K	time
C	IJ x K	time
D	I x KJ	batch
E	I x JK	batch
F	J x IK	variable

Table 3.3: Types of unfolding a three way data array

loadings of the two matrices if no centering or scaling is applied. The unfolding used by Nomikos and MacGregor is of type *D*. This is straightforward for analysis of historical data and monitoring of batch processes because subtracting the mean of each column of the matrix \mathbf{X} removes the main nonlinear and dynamic components in the data. Nevertheless, batch-wise unfolding (type *D* and *E*) presents a problem for monitoring in real time (ON-line monitoring) since the new batch is incomplete during the progress of the batch (Nomikos and MacGregor 1994b). Nomikos and MacGregor (1995) suggest *three* ways to overcome the problem of incomplete batches, while not changing the unfolding type. Alternatively, Wold et al. (1987) suggest a variable-wise unfolded PLS approach, which does not require complete batches. Applications of batch-wise unfolding (type *D* or *E*) in biological batch processes can be found in Lee et al. (2005), Lee and Vanrolleghem (2003b), and Wold et al. (1998). In this thesis, methods *A* in variable wise (Figure 3.8) and *E* in batch wise (Figure 3.9) are used. Type *E* was chosen instead of the mathematically equivalent type *D*, for simplicity of interpretation

Batch wise unfolding

It is important to determine differences between batches and to project new batches on the model. $\underline{\mathbf{X}}$ contains vertical slices(*I*) side by side to the right, starting with the one corresponding to the first time interval. The resulting two-dimensional matrix has size ($I \times JK$), see Figure 3.8 (Lee and Vanrolleghem 2003b) (Kourti 2003a). This unfolding method allows us to analyze the variability among the batches in $\underline{\mathbf{X}}$ by summarizing the information in the data with respect both to the variables and their time variation (Wold et al. 1987), (Westerhuis and Coenegracht 1997) and (Nomikos and MacGregor 1994a).

Variable wise unfolding

Another suggestion is to unfold the three-way array into a two-way matrix of size ($KI \times J$) by preserving the variable direction, as shown in Figure 3.9 ((Undey and Cinar 2002), (Undey et al. 2003) and (Kourti 2003a)).

The MPCA algorithm derives directly from the NIPALS algorithm and has the following formulation (Wold et al. 1987) (Nomikos and MacGregor 1994a):

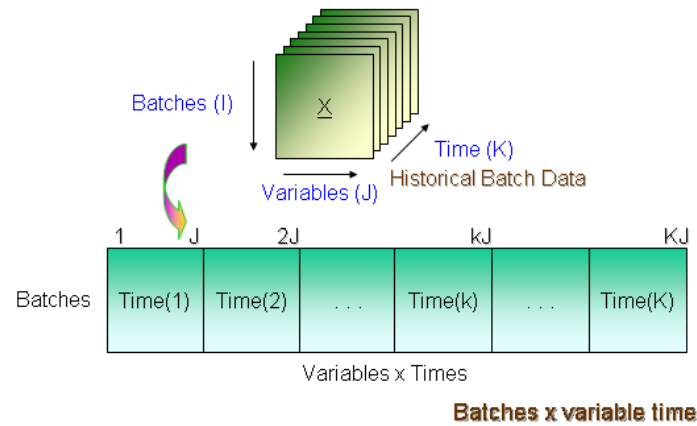


Figure 3.8: Decomposition of a three-way data array, $\underline{\mathbf{X}}$, by MPCA

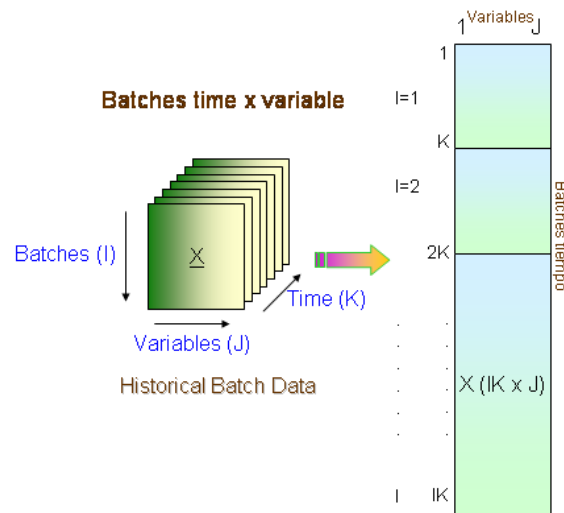


Figure 3.9: Other decomposition of a three-way data array, $\underline{\mathbf{X}}$, by MPCA

- i. Scale $\underline{\mathbf{X}}$ by subtracting from each column its mean and dividing by its standard deviation.
- ii. Arbitrarily choose a column of $\underline{\mathbf{X}}$ as t
- iii. $\underline{\mathbf{E}} = \underline{\mathbf{X}}$
 1. $P = \underline{\mathbf{E}}' . t$
 2. $P = \frac{P}{\|P\|}$
 3. $t = \underline{\mathbf{E}} \circ P$
 4. If t has converged then go to step 5, otherwise go to step 1.
 5. $\underline{\mathbf{E}} = \underline{\mathbf{E}} - t \otimes P$
 6. where \otimes denotes the Kronecker product

iv. Go to step 1 to calculate the next principal component.

The matrix operations in the above algorithm are:

$$\underline{\mathbf{E}}'(j, i, k) = \underline{\mathbf{E}}(i, j, k) \quad (3.21)$$

$$P = \underline{\mathbf{E}}' \cdot t \quad (3.22)$$

$$P(k, j) = \sum_{i=1}^I \underline{\mathbf{E}}'(i, j, k) t(i) \quad (3.23)$$

$$\| P \| = \sqrt[n]{\sum_{k=1}^K \sum_{j=1}^J P(k, j)^2} \quad (3.24)$$

$$t = \underline{\mathbf{E}} \circ P \quad (3.25)$$

$$t(i) = \sum_{j=1}^J \sum_{k=1}^K \underline{\mathbf{E}}(i, j, k) P(k, j) \quad (3.26)$$

$$t(i) = \sum_{j=1}^J \sum_{k=1}^K \underline{\mathbf{E}}(i, j, k) P(k, j) \quad (3.27)$$

$$\underline{\mathbf{X}} = t \otimes P \quad (3.28)$$

$$\underline{\mathbf{X}}(i, j, k) = t(i) P(j, k) \quad (3.29)$$

The array $\underline{\mathbf{X}}$ is a summation of the product of score vector t_r and loading matrixes P_r , plus a residual matrix \underline{E} , that is minimized in a least-squares sense as

$$\underline{\mathbf{X}} = \sum_{n=1}^N t_n \otimes P_n \quad \text{or} \quad X = \sum_{n=1}^N t_n P_n^T + E = \hat{X} + E \quad (3.30)$$

MPCA decomposes the three-way $\underline{\mathbf{X}}$ array ($\underline{\mathbf{X}} = t \otimes P$ is $\underline{\mathbf{X}}(i, j, k) = t(i) P(j, k)^T$) and N denotes the number of principal components retained. The first equation in (3.30) is the 3-D decomposition while the second equation shows the more common 2-D decomposition (Undey and Cinar 2002).

Control Charts

Control charts can be applied to both unfolding methods. Abnormal operation is identified by projecting the new batches onto the model (Qin 2003). Control charts that are used for monitoring batch processes are based on the *Q-statistic* and T^2 , in which control limits are used to determine whether the process is in control or not.

The *Q-statistic* is a measure of the lack of fit with the established model. For end-of-batch i , Q_i is calculated as:

$$Q_i = \sum_{j=1}^J \sum_{k=1}^K (e_{jk})^2 \sim gx_{(h)^2} \sim gx_{(h)}^2 \quad (3.31)$$

where e_{jk} are the elements of E and Q_i indicates the distance between the actual values of the batch and the projected values onto the reduced space. The distribution of the calculated Q_i values can be approximated by a chi-square distribution, gx_h^2 , where g is a constant and h is the effective degrees of freedom of the chi-square distribution.

The Hotelling T^2 or *D-statistic*, measures the degree to which data fit the calibration model:

$$T_i^2 = t_i^T S^{-1} t_i \frac{I(I-N)}{N(I^2-1)} \sim F_{R,I-R} \quad (3.32)$$

where N denotes the number of principal components, I is the number of batches in the reference set, and S is the estimated matrix of scores. The T^2 statistic gives a measure of the Mahalanobis distance in the reduced space between the position of a batch and the origin that designates the point with average batch process behavior (Lee and Vanrolleghem 2003b) (Undey and Cinar 2002).

3.5 Conclusions

MSPC is a tool has been satisfactorily used in complex manufacturing processes, including process industries and batch-oriented manufacturing types such as pharmaceutical, chemical, petrochemical, and pulp & paper manufacturing. This tool provides early information about abnormal operation conditions reducing the complexity of the process and thereupon simplifying the monitoring and situation assessment of the processes. The particularity of a batch process is due to the nonstationary behavior of the process. Solving this characteristic problem, MSPC is an efficient tool for process understanding, monitoring and diagnosing assignable causes for special events.

MSPC should be applied in complex processes when large amounts of historical data are obtained from process sensors. This is the case for many Wastewater Treatment Plants. For this reason MSPC is proposed as an adequate tool for situation assessment of the plants described in Chapter 2.

Chapter 4

Case-Based Reasoning (CBR)

4.1 Preview

In this thesis, Case-Based Reasoning (CBR) is proposed as an Artificial Intelligence approach which can be applied to improve expert supervision by exploiting data acquired from the MSPC approach. The main CBR advantage is that the **Case Base** is built just once, in addition, maintaining and updating is accomplished through the learning capacity of this tool. CBR is a relatively recent problem solving technique that is attracting increasing attention. However, the numbers of people with first hand theoretical or practical experience of CBR is still limited. The main objective of this Chapter is to provide the theoretical concepts related with CBR. In CBR systems, a library of past cases expertise is built. Each case typically contains a description of the problem and a solution. The knowledge and reasoning process used by an expert to solve the problem is not recorded, but is implicit in the solution.

The first ideas about CBR started in 1977 with Schank and Abelson (1977). They proposed that the general knowledge about situations should be recorded as scripts which allow the set up of expectations and performance of inferences. These situations should describe information about past events, such as going to a restaurant or visiting a doctor. However, the experiments were not satisfactory due to incomplete theory of memory representation, people often confused events with similar situations. In 1982 Schank (1982) continued exploring the role of the memory with respect to the roles previous situations, situation patterns, or memory organization packets (MOPs) play in both problem solving and learning. At the same time, Gentner (1983) developed a theoretical framework for analogy which also has relevance to CBR. This work has been a philosophical basis for CBR (Aamodt and Plaza 1994). The work of Roger Schanks group at Yale University in the early eighties produced a cognitive model for CBR, and the first CBR applications were based upon this model. Derek Sleemans group from Aberdeen in Scotland studied the use of cases for knowledge acquisition, developing the REFINER system (Sharma and Sleeman 1988). Likewise, Mike Keane, from Trinity College Dublin, undertook cognitive science research into analogical reasoning that has subsequently influenced CBR (Keane 1988). Agnar Aamodt at the University of Trondheim investigated the learning facet of CBR and the combination of cases and general domain knowledge resulting in CREEK (Aamodt et al. 1989), (Aamodt 1991). Wang and Wang (2005) described the

development of web based expert systems for a steel company which involved Rule Based Reasoning (RBR) and CBR to detect equipment failure.

CBR has been applied in different areas, such as recommendation systems, help desks, decision support systems, medicine and others. Mujica et al. (2005) and Chang and Lai (2005) proposed a hybrid system that combines the self-organizing map (SOM) Neural Network with CBR, the former detecting damage of structures and the latter predicting the sales of new books. CBR has also been used for supervision of WWTP. Comas (2000) explained the development of a module for supervision of the WWTP in Granollers (Spain). In Rodriguez-Roda et al. (2002), cases represent a day's knowledge. Each case is described through the most relevant measurements and observations (quantitative and qualitative) and describes a specific situation to be diagnosed. The set of specific cases is stored in a structured way in a hierarchical case library. More applications to WWTP are found in Nunez et al. (2004) and Martinez et al. (2006).

4.2 The CBR Cycle

CBR is a tool used in expert systems. CBR is an approach to problem solving that is able to use specific knowledge of previous experiences (de Mantaras and Plaza 1997). A new problem is solved by matching it with a similar past situation. In the case of diagnosis, solving the problem means that the CBR-system proposes a solution satisfactory enough to identify the new fault. When building a CBR system, it is necessary to select an appropriate case base. It can be either an empty base that will grow by learning from new situations or a base containing some previous events. Reduction of the data base is applied to get a data base without information redundance.

The processes involved in CBR can be represented by a schematic cycle (see Figure 4.1). (Aamodt and Plaza 1994) describe CBR typically as a cyclical process comprising the four R's:

Retrieve If a past situation is similar to the new one. It is necessary to define a metric function and the number of cases to retrieve from the case base.

Several algorithms have been implemented to retrieve appropriate cases, including serial search (Navichandra 1991); (Acorn and Walden 1992), hierarchical search (Maher and Zhang 1991) and simulated parallel search (Domeshek 1993). Among well known methods for case retrieval are nearest neighbor, induction, knowledge guided induction and template retrieval. These methods can be used alone or combined into hybrid retrieval strategies. Wilson and Martinez (2000) list extensive information about possible methods. Before applying the retrieval step, it is necessary to normalize the attributes because they have different orders of magnitudes. The number of neighbors or cases k to retrieve from the case base will be the value that produces the best diagnosis results. Chapter 6 explains several empirical proofs developed as a part of this work in order to find the best way to retrieve the neighbor cases.

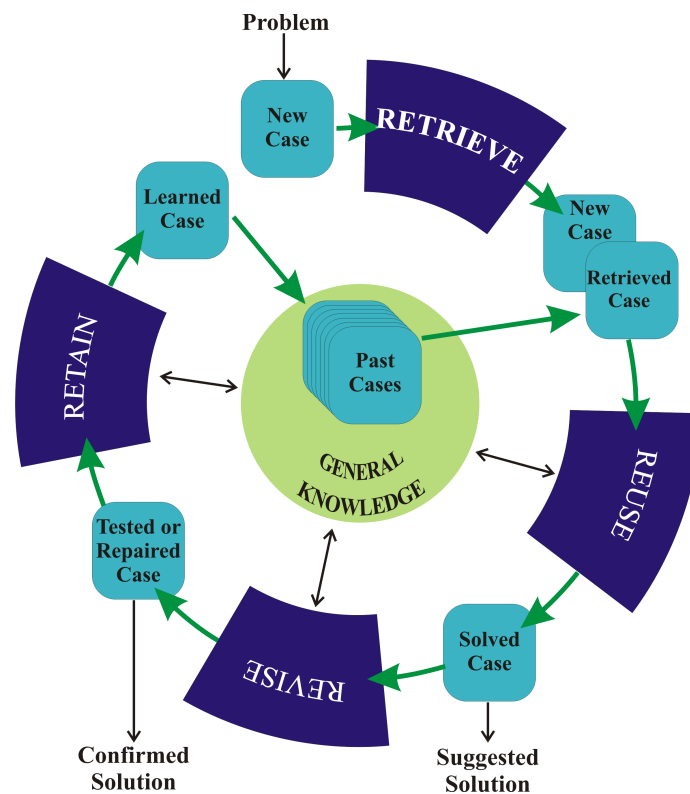


Figure 4.1: CBR cycle

Reuse Once the k nearest cases are extracted, they are used to propose a possible diagnosis.

Once the solution to the new presented case is proposed, it has to be revised. If the solution is considered to be correct and is accurate enough, it is not necessary to retain the new case. On the other hand, if it is considered wrong or has poor accuracy, the new case will be kept in the case data base. The revision analyzes how the cases that constitute the adapted solution are performing the diagnosis.

Revise According to a similarity index, a voting technique, or adapting the cases, the system provides a possible diagnosis. The adaptation can be transformational or derivational (Aamodt and Plaza 1994). The former uses the past case solution instead of the past method that constructed the solution to evaluate the suggested solution.

Retain The revision can be evaluated in a model or in the real world, and the CBR knowledge base can be updated with the newly learned case. After the revision process, according to the proposed solution, it must be decided if it is useful to retain the knowledge obtained from the new problem. Control and supervision of the process is a domain where CBR can rapidly extend its benefits because data is systematically collected and registered for further analysis.

To revise and retain it is necessary that the CBR adapts and learns. The main idea is to select potential cases to be stored into the case dictionary. For this purpose, Instance-Based learning algorithms (IBs) have been used.

Another important aspect is related to the building of the case base. The case base is an array in memory organizing all the cases to facility the search for the case most similar to the current problem. When building a CBR methodology, the case base is the most important item as it is important to select the correct initial case base (Leake 1996). The case base can be initiated empty and grow by learning with new problems, or a case base that start with some initial problems. In accordance with Sheppard and Simpson (1998) the selection can be $\pm 20\%$ and $\pm 50\%$ from the original or first problems. These problems are distributed well enough in order to cover a possible set of typical faults. For this, algorithms are used to reduce the number of problems inside the database. When the dictionary of cases is generated, it generally has a large size and wastes memory space while keeping noisy and redundant cases. The reduction algorithms are based on the improvements of the nearest neighbor and they are called Decremental Reduction Optimization Procedure.

Finally, the sum of the similarity of all attributes multiplied by a weighting factor can be used to calculate the distance to find the neighbor (Sanchez-Marre et al. 1997) (Watson 1998). In Nuez et al. (2002), seven different distances were checked for two environmental systems: Euclidean, Manhattan, Clark, Canberra and L'Eixample. Wiese et al. (2004) developed another Euclidean distance implementation for a WWTP. Mujica et al. (2005) applied neural networks in order to find the best neighbor. The technique of "neighbor" determines the gap between the characteristics of the new case or problem and the cases that already exist, finding the case that is closest.

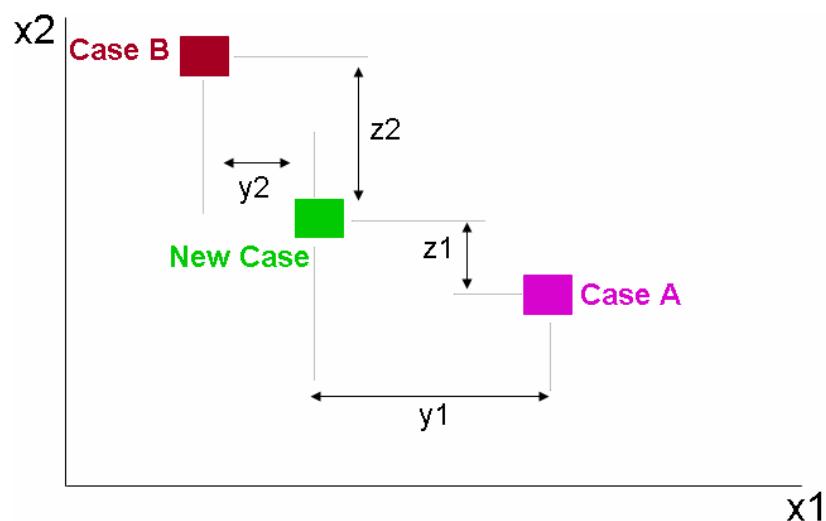


Figure 4.2: The distance between the new case or new problem and cases A and B. X_1 and X_2 are the characteristics that define the cases.

Normally to highlight a feature, the technique considers the distance, with a weight w_i , lets that the new case will be near or more particularly, as shown in the following equation:

$$distance(C, C_0) = \sum_{j=1}^n f(C_i, C_{0j})w_j \quad (4.1)$$

where f is a function of similarity, C is the new cases, C_0 is a prior case and n is the number of attributes of the case.

The technique of "Inductive Recovery" requires prior preparation of a decision tree. The drawbacks with this technique is that if you do not have all the answers, it is impossible find one similar case.

4.3 Decremental Reduction Optimization Procedure Algorithms

The Decremental Reduction Optimization Procedure Algorithms (DROP) family of algorithms are used to remove noisy instances. A noisy instance si has a different class than the class of the instances that have been associated (Wilson and Martinez 2000).

DROP removes a set of instances si in the center of a cluster surrounded by instances of the same class $A = A_1, A_2, A_3, \dots, A_K$ as is shown in Figure 4.3a. On the other hand, if si is surrounded by enemies (instances that different class), then si can not be removed because it leads to misclassifying A_1 as is shown in Figure 4.3. In this manner, the DROP family of algorithms tends to keep non-noisy border collection points (Figure 4.3c).

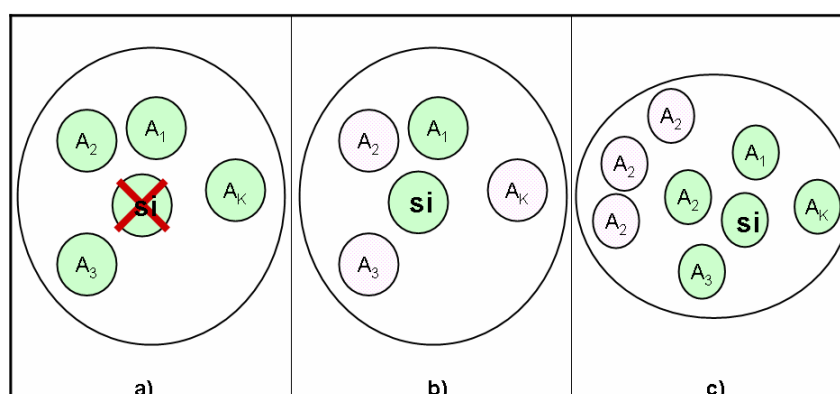


Figure 4.3: a)Central cluster instance b)Non-noisy border point c)Collection of border instances

4.3.1 DROP1

The main idea of this algorithm is to check whether an instance si has to be removed from the instance set SI ($SI = TI$ where TI is the initial data base and SI is the new data base without redundancy cases or instances for the first iteration) according to the following rule (Wilson and Martinez 2000):

If the number of associates of si classified correctly without si is greater than or equals to the number of associates of si classified correctly with si , then si is removed from SI . Associates are those instances from SI that have si as one of its nearest neighbors. The goal of this rule is to avoid the elimination of instances that alter the behavior of the training set, preserving its correct classification ratio. When si is removed, each of its associates uses its $k+1$ st nearest neighbor to determine its class. If this new instance class is different than the one of the associate, its classification is weakened, reaching a misclassification in the worst case. On the other hand, if the class of si is different from the instance class, the classification is strengthened, leading to a better classification ratio and increasing the number of correct instances classified.

In essence, this rule improves the efficiency of determining the class of each instance, since with fewer instances in the training set the computational cost is reduced to a linear relation between computational time and size of the training set. In order to better understand this algorithm, the procedure is explained as follows:

The algorithm builds a list of nearest neighbors for each instance in SI and the associate list. One instance in SI is removed if and only if its removal does not alter the classification of the remaining instances in SI . When an instance si is removed, each associate must replace si with a new nearest neighbor in order to maintain $k+1$ nearest neighbors in their list. Whenever a new neighbor N is found, then each associate of si is added to N 's associate list. Therefore, at all the times every instance in SI has a current list of neighbors and associates.

4.3.2 DROP2

As an evolution of DROP1, DROP2 improves the selection of instances to remove by using more information and sorting the order of removal. DROP1 removes the noisy instances without taking into account any specific order (Wilson and Martinez 2000). This can cause the elimination of several neighbors that are considered to be noisy instances, but as a consequence of this action, the instance hoped to be removed can be kept in the training set. Then, the search for the new neighbors begins. This can provoke the inclusion of associate instances of the same class, located on the other side of the decision boundary. If this is the case, this instance can not be removed, due to it being surrounded by enemies.

DROP2 solves this problem by considering the effect on the original training set TI , instead of SI . This provokes a modification of the basic rule explained in DROP1 (section 4.3.1), restated as follows:

```

function instance set  $SI = \text{DROP1}(\text{Training set } TI)$ 
   $SI = TI$ 
  for  $si = 1$  to number of instances in  $SI$ 
    find( $k+1$  nearest neighbors)
    add( $si$  at each neighbors' associates list)
  endfor
  for  $si = 1$  to number of instances in  $SI$ 
    with = number of associates correctly classified
      with  $si$ 
    without = number of associates correctly classified
      without  $si$ 
    if without  $\geq$  with
      remove( $si$  from  $SI$ )
      for  $a = 1$  to number of associates of  $si$ 
        remove( $si$  from nearest neighbors of  $a$ )
        find(new nearest neighbor of  $a$ )
        add(new neighbor to associates list of  $a$ )
      endfor
      for  $n = 1$  to number of neighbors of  $si$ 
        remove( $si$  from associates list of  $n$ )
      endfor
    endif
  endfor
endfunction

```

Figure 4.4: DROP1 algorithm

If the number of associates of si classified correctly without si is greater than or equals to the number of associates of si classified correctly with si , then si is removed from SI and the associates are extracted from TI . With this variation, instances in SI can have as an associate an instance that does not belong to SI , and the instances removed from SI do not have a list of associates because they are no longer a neighbor of any instance. Moreover with the use of additional information to estimate the generalization accuracy, it avoids the removal of entire clusters.

The order of removal of instances used by DROP2 is based on sorting by the distance to the nearest enemy starting with the furthest value of distance. This removes instances located far from the decision boundary, which tends to keep non-noisy border points.

After these modifications, the DROP2 algorithm states as follows:

```

function instance set  $SI = \text{DROP2}(\text{Training set } TI)$ 
   $SI = TI$ 
   $SI = \text{sortBy}(SI, \text{furthest enemy distance})$ 
  for f = 1 to number of instances in  $SI$ 
    find( $k+1$  nearest neighbors)
    add( $si$  at each neighbors' associates list)
  endfor
  for si = 1 to number of instances in  $SI$ 
    with = number of associates correctly classified
      with  $si$ 
    without = number of associates correctly classified
      without  $si$ 
    if without  $\geq$  with
      remove( $si$  from  $SI$ )
      for a = 1 to number of associates of  $si$ 
        remove( $si$  from nearest neighbors of  $a$ )
        find(new nearest neighbor of  $a$ )
        add(new neighbor to associates list of  $a$ )
      endfor
    endif
  endfor
endfunction

```

Figure 4.5: DROP2 algorithm

4.3.3 DROP3

As distinguished from DROP2, DROP3 adds a noise filter before starting to analysis the instances in SI (Wilson and Martinez 2000). This is because starting by the borders tends to remove central points. If a noisy instance is in the center, surrounding instances are considered border points, avoiding its removal. Any instance misclassified by its k nearest neighbors is removed due to being considered a noisy instance. After this, the order of removal is the same as DROP2. The modifications are shown in Figure 4.6.

4.3.4 DROP4

DROP4 applies the same idea as in DROP3, but uses more caution in the filter noise step (Wilson and Martinez 2000). This removal depends on two conditions:

1. Any instance misclassified by its k nearest neighbors.
2. Its removal does not alter other instances classification.

Sometimes DROP3 eliminates too many instances, causing poor generalization accuracy. DROP4 avoids this situation at the expense of a bit more storage requirements. The

```

function instance set  $SI = \text{DROP3}(\text{Training set } TI)$ 
   $SI = TI$ 
   $SI = \text{noise filter}(SI)$ 
   $SI = \text{sortBy}(SI, \text{furthest enemy distance})$ 
  for  $si = 1$  to number of instances in  $SI$ 
    find( $k+1$  nearest neighbors)
    add( $si$  at each neighbors' associates list)
  endfor
  for  $si = 1$  to number of instances in  $SI$ 
    with = number of associates correctly classified
      with  $si$ 
    without = number of associates correctly classified
      without  $si$ 
    if without  $\geq$  with
      remove( $si$  from  $SI$ )
      for  $a = 1$  to number of associates of  $si$ 
        remove( $si$  from nearest neighbors of  $a$ )
        find(new nearest neighbor of  $a$ )
        add(new neighbor to associates list of  $a$ )
      endfor
    endif
  endfor
endfunction

```

Figure 4.6: DROP3 algorithm

algorithm is presented in Figure 4.7.

4.4 Instance-Based learning algorithms

The main objective of the Instance-Based learning algorithms (IB) algorithm is to give a classification for each new instance drawn from the problem domain, "*Similar instances have similar classifications*". To make this classification, the following components are necessary (Aha et al. 1991):

1. Similarity function: This function looks for the similarity between a training instance TI and the new problem given. This similarity is a numeric value.
2. Classification function: From the results of the similarity function and according to the classification performance record, a classification is given to TI .
3. Concept Description Updater: This function is used to decide which instances are kept in the training set, according to its classification performance, the similarity function and the actual training set.


```

function instance set  $SI = \text{DROP4}(\text{Training set } TI)$ 
   $SI = TI$ 
   $SI = \text{carefully noise filtering}(SI)$ 
   $SI = \text{sortBy}(SI, \text{furthest enemy distance})$ 
  for  $si = 1$  to number of instances in  $SI$ 
    find( $k+1$  nearest neighbors)
    add( $si$  at each neighbors' associates list)
  endfor
  for  $si = 1$  to number of instances in  $SI$ 
    with = number of associates correctly classified
      with  $si$ 
    without = number of associates correctly classified
      without  $si$ 
    if without  $\geq$  with
      remove( $si$  from  $SI$ )
      for  $a = 1$  to number of associates of  $si$ 
        remove( $si$  from nearest neighbors of  $a$ )
        find(new nearest neighbor of  $a$ )
        add(new neighbor to associates list of  $a$ )
      endfor
    endif
  endfor
endfunction

```

Figure 4.7: DROP4 algorithm

IB algorithms differ from most of the other supervised learning methods in the following way:

1. IB algorithms do not construct explicit abstractions, like decision trees or rules.
2. IB algorithms do not store explicit generalizations from instances because they do not use matching criteria.
3. The work load is higher because the computation of similarities between saved instances and those that are new is done every time.

4.4.1 IB1

The IB1 algorithm is a variation of the nearest neighbor algorithm in terms of values range normalization, the incremental processing of instances, and a simple policy for tolerating missing values (Aha et al. 1991).

For each instance x in the training set TI , the most similar instance y in SI has to be found according to the similarity function described in 4.4. If the class of x is different than the class of y , then the classification is incorrect; therefore, the classification record

```
function instance set  $SI = \text{IB1}(\text{Training set } TI)$ 
   $SI = \emptyset$ 
  for  $x = 1$  to number of instances in  $TI$ 
    for  $y = 1$  to number of instances in  $SI$ 
       $\text{sim}[y] = \text{similarity}(x,y)$ 
    endfor
     $y_{\text{max}} = \text{max}(\text{sim}[y])$ 
    if ( $\text{class}(x) \neq \text{class}(y)$ )
      classification = incorrect
    else
      classification = incorrect
    endif
     $SI = SI \cup x$ 
  endfor
endfunction
```

Figure 4.8: IB1 algorithm

of y has to be updated with a misclassification. Otherwise, the classification record of y is updated with a correct classification.

The main goal of the IB1 algorithm is to maintain the classification record in order to classify each instance in SI according to its class. This is achieved because only instances near the decision boundary are needed to produce an accurate classification of instances. Unfortunately, this set of instances is not known without a complete knowledge of the problem, but can be approximated by the set of instances misclassified by the IB1 algorithm. This is the basis for the IB2 algorithm, which is presented in the next section.

4.4.2 IB2

The IB2 algorithm is similar to the IB1 algorithm, but it only saves instances in SI if and only if the instance is misclassified. This modification is shown in Figure 4.9.

For each instance x in the training set TI , the most similar instance y in SI has to be found according to the similarity function described in Section 4.4. If the class of x is different than the class of y , then the classification is incorrect. Therefore, the classification record of y has to be updated with a misclassification and x will be saved in the training set. Otherwise, only the classification record of y is updated with a correct classification.

By only saving the misclassified instances, the storage space used is reduced dramatically. This reduction slightly reduces the classification ratio that would be reached with the IB1 algorithm. A study of the relation between the degree of sacrifice and accuracy lost was presented in (Aha et al. 1991). The main conclusion of this study was that IB2 is too sensitive to noisy instances, leading to the development of the IB3 algorithm.

```

function instance set  $SI = \text{IB2}(\text{Training set } TI)$ 
   $SI = \emptyset$ 
  for x = 1 to number of instances in  $TI$ 
    for y = 1 to number of instances in  $SI$ 
       $\text{sim}[y] = \text{similarity}(x,y)$ 
    endfor
     $y_{\text{max}} = \text{max}(\text{sim}[y])$ 
    if ( $\text{class}(x) == \text{class}(y)$ )
      classification = correct
    else
      classification = incorrect
       $SI = SI \cup x$ 
    endif
  endfor
endfunction

```

Figure 4.9: IB2 algorithm

4.4.3 IB3

The IB3 algorithm is a noise tolerant version of the IB2 algorithm that employs a simple selective utilization filter to decide which of the saved instances have to be used in the classification decision.

The main differences between IB3 and its predecessor are:

1. IB3 maintains a classification record, the number of correct classifications divided by the number of classification attempts, with each of the saved instances. This record can predict the future performance of this instance according to the performance with the training instances.
2. IB3 employs a significance test to decide which are the good classifier instances and those that can be considered as noisy instances and removed from the classification set.

These variations modify the algorithm as shown in Figure 4.10

IB3 is based on discriminating acceptable and unacceptable instances. An instance is acceptable if its classification accuracy is greater than its class classification. This comparison is made using a confidence interval, and if this value is extremely poor, it is removed from the training set. The confidence interval is based on the actual instance performance (number of successful classifications) and its class classification (number of instances correctly used of the class during the classification process). An instance is accepted if its lower endpoint of the confidence interval is greater than the higher endpoint of the class record. On the other hand, if the higher endpoint is less than the class lower

```

function instance set  $SI = \text{IB3}(\text{Training set } TI)$ 
   $SI = \emptyset$ 
  for  $x = 1$  to number of instances in  $TI$ 
    for  $y = 1$  to number of instances in  $SI$ 
       $\text{sim}[y] = \text{similarity}(x,y)$ 
      if  $\exists y \in SI \mid \text{acceptable}(y)$ 
         $y_{\max} = \max(\text{sim}[y])$ 
      else
         $y_{\max} = \text{randomInstance}(SI)$ 
      endif
      if  $\text{class}(x) == \text{class}(y)$ 
        classification = correct
      else
        classification = incorrect
         $SI = SI \cup x$ 
      endif
      for  $y = 1$  to number of instances in  $SI$ 
        if ( $\text{sim}[y] \geq \text{sim}[y_{\max}]$ )
          update classification record of  $y$ 
          if record classification of  $y$  is
            significantly poor
             $SI = \text{remove}(y \text{ from } SI)$ 
          endif
        endif
      endfor
    endfor
  endfor
endfunction

```

Figure 4.10: IB3 algorithm

endpoint, the instance is dropped. The equation to compute the confidence interval is shown in equation 4.2:

$$S = \frac{b + \frac{z^2}{2h} \pm z\sqrt{b\frac{(1-b)}{h} + \frac{z^2}{4h^2}}}{1 + \frac{z^2}{h}} \quad (4.2)$$

where:

- $b = \frac{\text{number of successful classifications}}{\text{number of attempts}}$,
- $z =$ is the confidence limit (0.9 for acceptance; 0.7 for dropping) and
- $h =$ number of instances stored in the data base.

For further comparison of the performance of IB algorithms more details can be found in Aha et al. (1991).

4.5 Conclusions

Case Based Reasoning is the process of solving new problems based on the solutions of past situations. The CBR is one of the fastest growing areas in the field of knowledge based systems. This paradigm is often used to solve a myriad of situations, this being one of the main reasons for its acceptance in the Artificial Intelligence research community. The quality of such systems depends on the experiences that are stored and the ability to understand, adapt, assess and repair new cases. Several steps are necessary in order to develop the CBR cycle. The aim is to reuse these cases for solving new problems by analogy. A problem is solved by retrieving a similar problematic situation (case or cases) from the past and reusing its solution in the new situation. Reusing implies a procedure of adapting the retrieved solution, which is then completed with the revision. Once a diagnosis is proposed, it must produce a review of the cases to decide whether the new case will be entered in the database, or if the content of the base is already sufficient to diagnose the new problem. Immediately afterwards, it considers whether the new problem is retained or eliminated from the case base. In this way, *DROP4* and *IB3* are used to reduce the dimension of the case base.

As MSPC, CBR should be applied to complex processes with large amounts of historical data. In the next chapters, CBR will be proposed to complement the MSPC tool (MPCA) within a situation assessment methodology. In this way, MPCA can be used as a dimensionality reduction tool while CBR can use the results obtained from the application of the MPCA approach as descriptors for the case.

Chapter 5

Application of MPCA Methodology to SBR pilot plants

The main objective of this chapter is to evaluate the behavior of MPCA for monitoring of biological nutrient removing Sequencing Batch Reactors (SBRs) in wastewater treatment in order to improve the operation of this kind of process. The chapter is divided into two sections. The first section describes work using MPCA for the Semi-Industrial Pilot Plant SBR from the LEQUIA group. This pilot plant is pertinent to the project "Development of an Intelligent Control System applied to a Sequencing Batch Reactor (SBR) for the removal of Organic Matter, Nitrogen and Phosphorous. SICOTIN-SBR2". The second section describes the methodology used to find the optimal method for building a statistical model for a SBR process; for this, the Pilot Plant SBR from the BIOMATH group was used. This pilot plant was used during the research period in Ghent-Belgium. Both pilot plants have different combinations of filling and reaction phases. In this way, the chapter will aim to prove the efficiency of the methodology for this kind of process.

5.1 Semi-Industrial SBR Pilot Plant from the LEQUIA group

5.1.1 Types of batch processes

The SBR process is very complex. For this reason, chemical engineers (LEQUIA) and the control engineers (eXiT) together created a classification for the types of batches. Two types of analysis were necessary. The first one was based on *the analytical methods* proposed in Puig et al. (2004) in which the sludge reaction is explained. The second analysis is a *preliminary MSPC application*, in which some batches are found outside the control limit. Figure 5.1 shows the preliminary results of this technique. Batches with an Abnormal Operation Condition (AOC) are outside the control limit, whereas those groups with Normal Operation Condition (NOC) are inside the limit (solid line).

As a result of the combined knowledge of the engineers, five types of events in the Semi-Industrial SBR process were proposed.

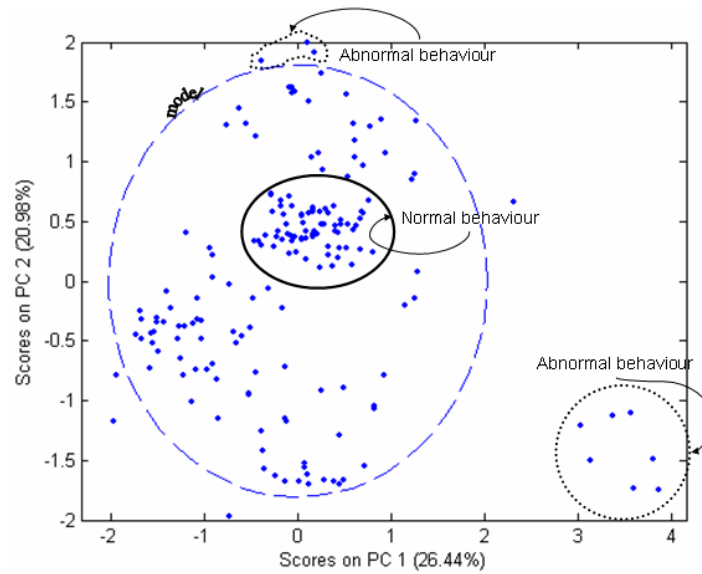


Figure 5.1: Score plot for batches. Dashed line is the model

1. Electrical Fault (EF): Corresponds to voltage sags. Voltage sags are short duration reductions in *rms* voltage that are caused by short circuits, overloads and starting of motors. The interest in voltage sags stems mainly from the problems that they cause for several types of equipment (Bollen 2000). Figure 5.2 shows the response when a voltage sag is presented. At the moment that a voltage sag is present, there are sensor problems; when the voltage sag disappears, the sensors resume their normal operation.

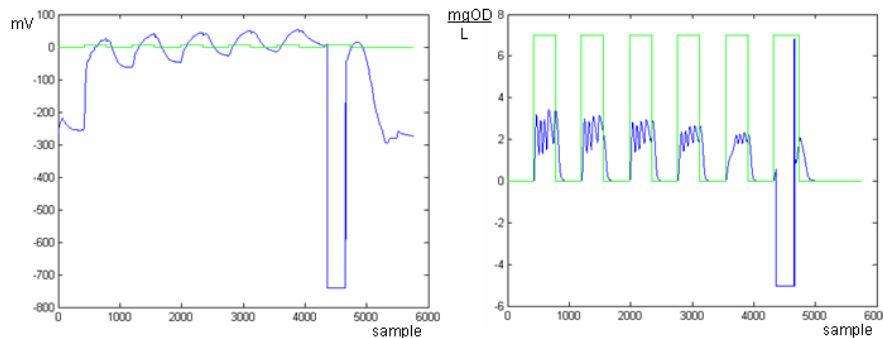


Figure 5.2: DO (green line) and ORP (blue line) profiles when an EF occurs

2. Variation in the Composition (VC) can be due to problems in either microorganisms or the influent composition that cause disturbances in the process variables (Puig et al. 2004). The univariate plot of two of these process variables (Oxidation Reduction Potential (ORP) and dissolved oxygen (DO)) is shown in Figure 5.3 for NOC and AOC.

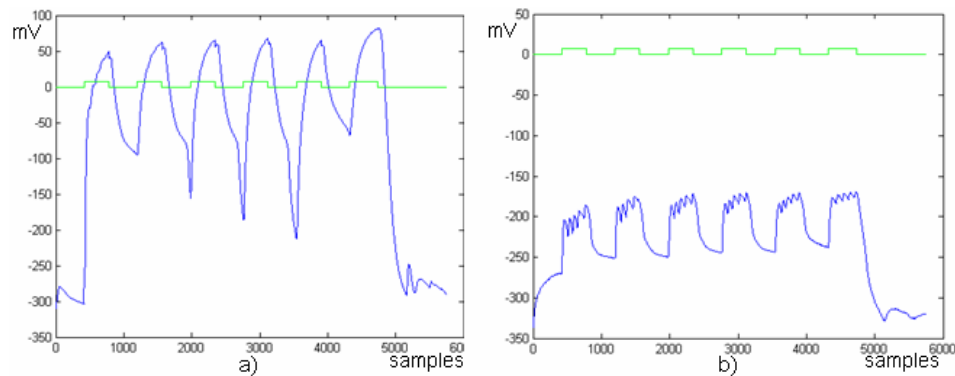


Figure 5.3: ORP and DO profiles when and VC fault condition is presented a) NOC b) AOC

3. Equipment Defects (ED) are present when the computer does not register data due to permanent faults in sensors or its acquisition card or due to missing data, (Figure 5.4).

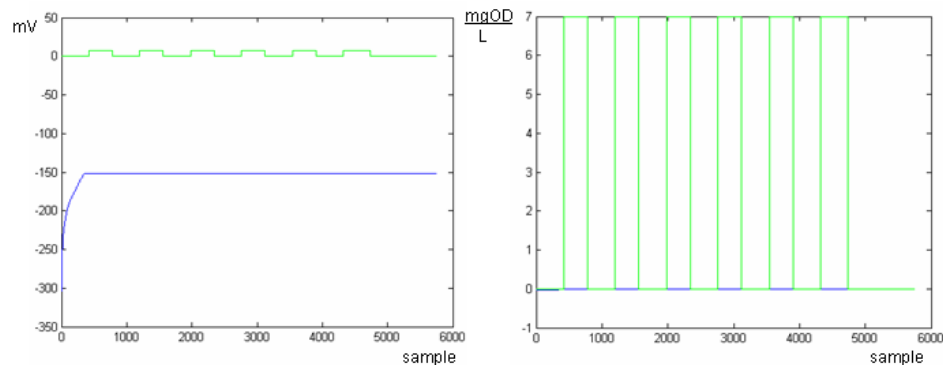


Figure 5.4: ORP and DO profiles when and ED fault occurs

4. Influent Load Change (ILC) correspond to a rain period. Wastewater contains a lot of rainwater; therefore, the concentrations of organic matter and nitrogen is low (see Figure 5.5).
5. Normal Operation Condition (NOC) corresponds to all batches where nitrogen and organic matter removal are satisfactorily obtained. Based on the knowledge of the chemical engineers and checking the off-line variables (effluent measurements) and compare with the discharge limits, it was possible to categorize the final quality of the water as excellent, good or normal. Figure 5.3 shows a profile of the ORP when an excellent final quality of the water was obtained. Figure 5.6 shows the same variable for good and normal qualities of the water. The difference between these is the variable magnitude.

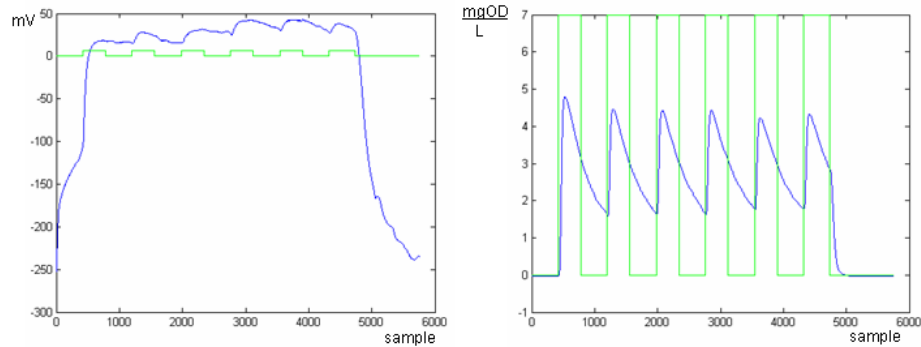


Figure 5.5: ORP and DO profiles in presence of rainwater

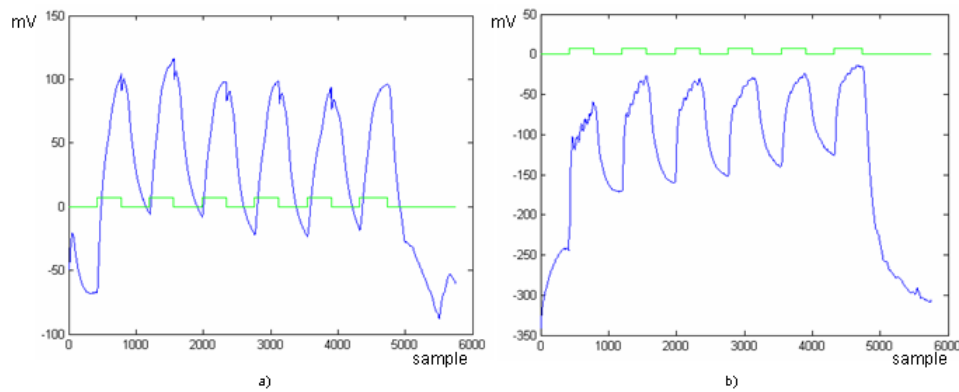


Figure 5.6: ORP and DO profiles a) Good final quality b) Normal final quality

From this classification, it is possible to determine the number of the batch for each group. Tables 5.1 and 5.2 summarize all batches of the process. There are 60 (equivalent to 33.5%) batches with AOC, and these have been divided into several events: Electrical Faults (EF), Variation in the Composition (VC), Equipment Defects (ED) and Influent Load Change (ILC) (AC). The NOC is the most common type of event, and 66.5% of these have a nitrogen efficiency higher than the legally required effluent standards. These are classified according to the final quality of the water.

AOC	Amount	%
ILC	17	9,50
ED	8	4,47
VC	33	18,44
EF	2	1,12
TOTAL	60	33,52

Table 5.1: Types of events with AOC

NOC	Amount	%
Excellent	98	54,75
Good	14	7,82
Normal	7	3,91
TOTAL	116	66,48

Table 5.2: Types of events with NOC

Finally, Figure 5.7 summarizes the types of events present in the Semi-Industrial Pilot Plant.

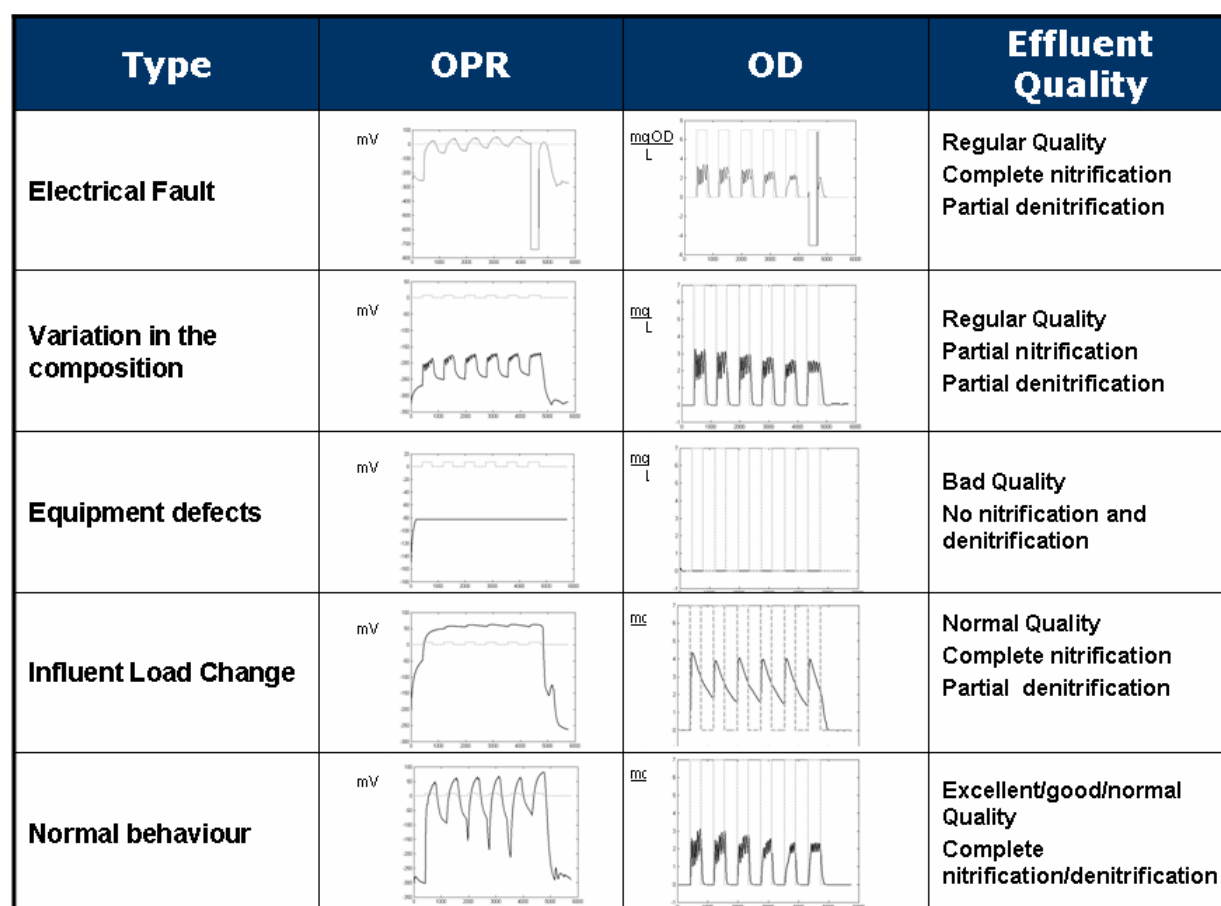


Figure 5.7: Types of events

5.1.2 Application of MPCA

The MPCA algorithm was applied to the three-way data array, $\underline{\mathbf{X}}$, which has dimensions $179 \times 4 \times 392$. In this array, $K = 392$ is the number of time instants throughout the batch (samples), $J = 4$ is the number of process variables (ORP, DO, pH and temperature),

and $I = 179$ is the historical data set. The three-way array \mathbf{X} has been unfolded in the batch wise ($I \times KJ$) array, which is the same (179×1568). Afterwards the model was built using only 8 principal components. Therefore, the array is a matrix with a size of (179×8). The model explains 92.79% of the total variability (see Table 5.3)

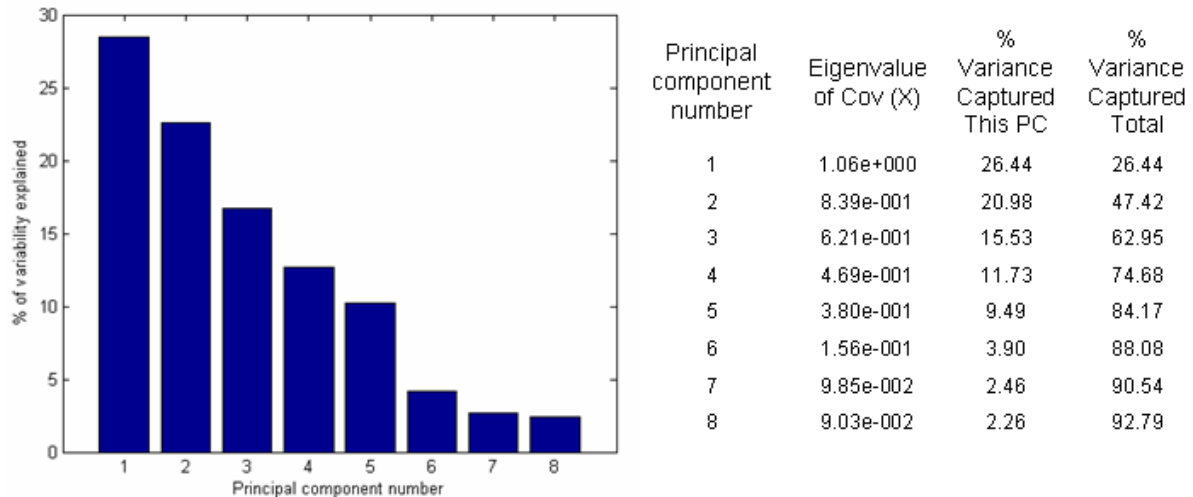


Table 5.3: Principal component extraction

To examine the process data in reduced space, defined by a small number of latent variables, the first and the second principal components were plotted (see Figure 5.1). In this figure, several groups of batches are outside the model space. These batches correspond to AOC batches, more precisely, they represent the variation in the composition. Figure 5.8 shows the Q -statistic and T^2 -statistic distances for all batches.

Using this knowledge, each batch is analyzed. In this study, it was found that 60 batches contain AOC. It is possible to determine whether the batches were correctly classified as NOC or AOC. Table 5.4 summarizes the batches appearing outside the control limit. The Q -statistic detects only onethird of the total of AOC occurrences, and there are 8 false alarms. The T^2 statistic detects 20 batches with AOC (without false alarms). These are distributed as follows: 4 are influent load change, 6 are equipment defects, 8 are variation of the composition and 2 are electrical faults. Using the Q -statistic and T^2 -statistic, 31 batches with AOC were found; of these, 9 batches were found in both control charts.

Combining MPCA and the knowledge from the experts made it possible to classify the batches into five types of events. Individually, the results have been satisfactory because they made it possible to build a model. The model can detect faults in the processes. The results of this section were published as a book chapter in *Frontiers in Statistical Quality Control* (Ruiz, Colomer and Melndez 2006).

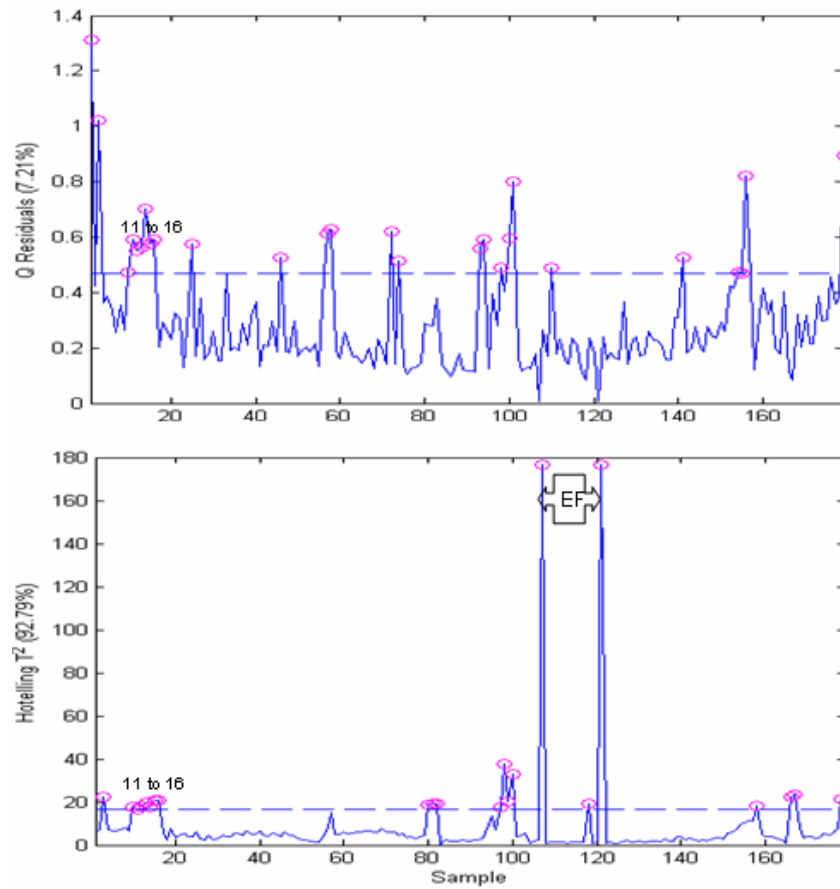


Figure 5.8: Q -statistics and T^2 -statistics with 92.79% confidence limits for the Semi-Industrial Pilot Plant

Q -statistic			T^2 -statistic		
AOC	Amount	%	AOC	Amount	%
ILC	9	5.03	ILC	4	2.23
ED	0	0.00	ED	6	3.35
VC	11	6.15	VC	8	4.47
EF	7	0.00	EF	2	1.12
TOTAL	20	11.17	TOTAL	20	11.17
NOC			NOC		
	Amount	%		Amount	%
Excellent	5	2.79	Excellent	0	0.00
Good	3	1.68	Good	0	0.00
Normal	0	0.00	Normal	0	0.00
TOTAL	8	4.47	TOTAL	0	0.00

Table 5.4: Batches detected using Q -statistic and T^2 -statistic

5.2 SBR Pilot Plant from BIOMATH group

In Camacho and Pic (2006) and Aguado, Ferrer, Ferrer and Seco (2007) several studies have been developed in order to test the best way for the modeling of batch processes and SBR process. However, normal questions about what items are necessary to have in mind when the data is collected from the process?. In this section, it is didacticism and empirically explained several steps in order to perform a MPCA model. Two important degrees of freedom in building MPCA models are investigate: unfolding and scaling by means of several proof in order to demonstrate which is the better methodology for applying MPCA in a SBR WWTP. During the time of this study, a large amount of knowledge was collected. The large amount of data collected in this plant permits facilities to divide the data into different sets, such as a set to build the model and sets to validate the model. In this study, the problem of multiple comparisons does not occur because the variables are entirely dependent. The main objective is to show the effects of the unfolding and preprocess effects on fault detection and diagnosis in a SBR process. Initially, knowledge from experts was used to classify the types of batch processes. This knowledge, which contained both qualitative and quantitative information on process variable, was necessary in order to apply MSPC.

5.2.1 Systematic comparison of PCA models

In total, 1711 complete batches were available. Each of these batches contained 6 different trajectories of 300 samples each. These included the weight of the reactor, temperature, pH, DO, ORP and the conductivity in the reactor. The approach taken to compare the discussed options for PCA-based monitoring is as follows (Figure 5.9).

Step 1

The three-way data array is first unfolded in a batch wise manner (type E, see Section 3.4.1). The data $X(I \times JK)$ were normalized using auto-scaling as suggested by Nomikos and MacGregor (1994b) and Westerhuis et al. (1999) to construct an MPCA model for data screening. The resulting MPCA-model identified 248 batches as showing abnormal operation, and these batches were thus excluded from the data set for future use.

Step 2

The 1711 remaining batches were used to compare different unfolding and scaling approaches to PCA-based monitoring of the pilot-scale SBR. The data base (1711x6x300) is divided in order to first develop models with 80% of the total data in a three way (1369x6x300) data matrix and then validate the models with 20% of the data in a (342x6x300) data matrix. Two options with regard to unfolding are available when considering batch process monitoring:

1. The first one involves unfolding the data in a batch wise manner (type E). This unfolding is used for OFF - LINE monitoring.

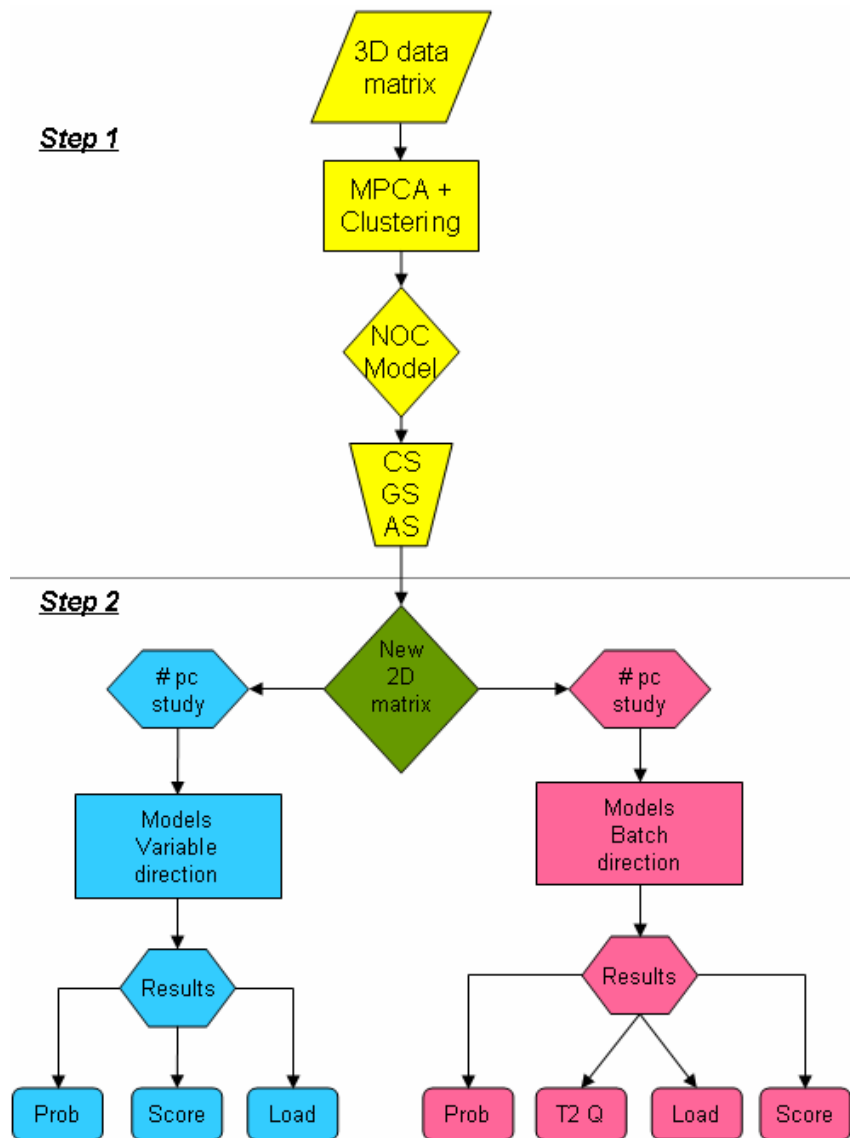


Figure 5.9: MPCA Methodology applied to pilot-scale SBR

2. The second option involves unfolding the data in a variable wise manner (type A, see Section 3.4.1). This method is typically used for ON - LINE monitoring.

For both options; the data in X ($I \times JK$) are again normalized using three different options: continuous scaling (CS), group scaling (GS), and auto scaling (AS).

1. continuous scaling (CS) was applied by Wold et al. (1987), and this method treats data as if all samples were drawn from the same distribution. For each variable, j , one mean and one standard deviation are calculated on the basis of all the data (all batches and all time instants). By doing so, J mean values (μ_j) and J standard deviations (σ_j) are obtained. They are calculated as follows:

$$\mu_j = \frac{\sum_i^I \sum_k^K x_{ijk}}{IK} \quad (5.1)$$

$$\sigma_j = \frac{\sum_i^I \sum_k^K (x_{ijk} - \mu_j)}{IK} \quad (5.2)$$

All of the measurements (throughout the entire set of batches and time instants) are then scaled with the same mean and standard deviation. In general, however, it is not expected that samples during a batch run are drawn from the same distribution; this may lead to poor performance of the monitoring model.

2. Group scaling (GS) avoids the poor performance of CS by removing the trajectory from the variables throughout the batch run. This is done by defining a separate mean for each variable at each time instant in the batch runs and thereby obtaining a mean trajectory for each variable. After scaling of the mean trajectories, one standard deviation is calculated per variable in a manner similar to that of CS. Then, one obtains $J.K$ mean values (μ_{jk}) as is shown in Equation 5.3, and J standard deviations (σ_j) as is presented in Equation 5.2.

$$\mu_{jk} = \frac{\sum_i^I x_{ijk}}{I} \quad (5.3)$$

Even if the trajectory of the data is removed from the data set by GS, the standard deviation of the variables is assumed not to change during the batch runs.

3. Auto scaling (AS) calculates the mean and standard deviation of each variable calculated at each time in the batch over all batches. Then, $J.K$ mean values and standard deviations are calculated as is presented in Equations 5.3 and 5.4.

$$\sigma_{jk} = \frac{\sum_i^I (x_{ijk} - \mu_{jk})}{IK} \quad (5.4)$$

When prior knowledge is available, the scaling approach may be defined by an operator or expert. For instance, Westerhuis et al. (1999) argued that auto scaling is appropriate in systems where the variables have different units (e.g. temperature, pressure and concentration).

In total, six models for monitoring SBRs were thus generated (3 using the variable way and 3 using the batch way). In variable wise the scale process was performed in two steps: firstly, in batch wise eliminating the major nonlinear and non-stationary behavior of the process and after re-scaled in v . Secondly, in variable wise (see Figure 5.10).

Table 5.6 summarizes how the respective models were labeled in the framework of this research. Before making further inferences, the normal probability plot for the first score is made. This allows a visual inspection of validity of one of the assumptions in PCA modeling (i.e. that the scores exhibit a Gaussian distribution). Immediately after this

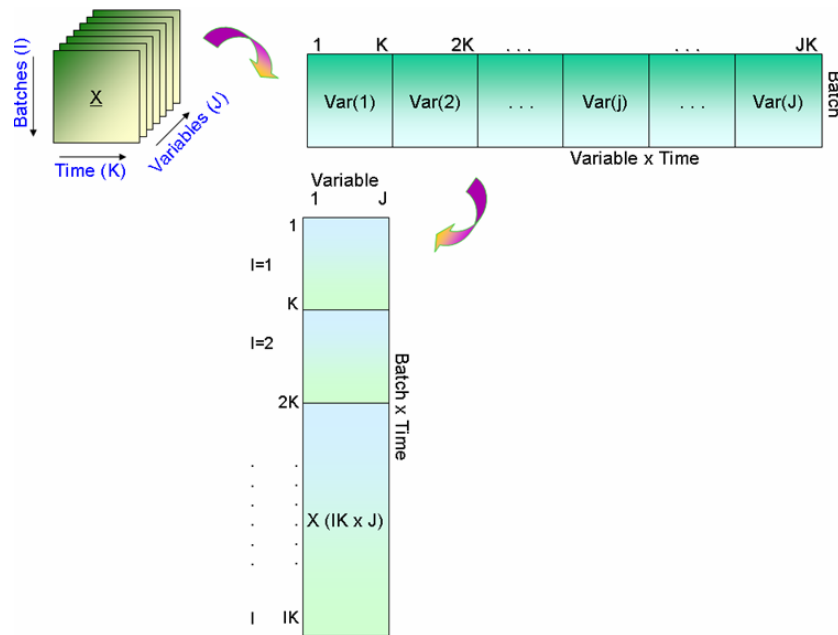


Figure 5.10: Scale process for variable wise models

inspection, the number of principal components is determined and analysis of the models are performed.

	Type A	Type E
CS	Model1	Model4
GS	Model2	Model5
AS	Model3	Model6

Table 5.5: Names for each developed model

5.2.2 Results

Normal probability plot

Before doing a detailed model comparison, it is useful to evaluate whether; first, the extent of linearization of the original data, which are typically non-linear and dynamic, by different combinations of unfolding and scaling approaches, satisfies the assumption for the normal linear model and second to have an approximately normal distribution (Yoo et al. 2004). Two hypotheses have been performed in accordance with Giudici (2003). The first one states that the value of the response variable is a linear combination of the explanatory variables. This is plainly stated for the linear combinations of PCA. The second hypothesis concerns the data set.

If the data comes from populations with a normal distribution, these could be tested using the measures of Kurtosis and the Quantile-Quantile plot (q-q plot). Figure 5.11 organized as in Table 5.5 with models 1 to 6 arranged from top to bottom and left to right, shows the q-q plots for the first PC of all models. If the values of the first PC come from a normal distribution, then the plots should appear as a linear curve.

The plots show that the data from models 1, 3, 4, 5 and 6 can be considered as a normal distribution. The plots are approximately linear for these models. However, in model 2 a distortion of the operation is found due to the presence of two batches (1010 and 1011) reflecting a specific fault in the temperature sensor, which generates (see Figure 5.12). These two batches were removed.

A gap is clearly visible at the left hand side of the q-q plot for variable wise unfolded models, whereas batch wise unfolded models show a bump in the same region of the plot. These non-linearities were not considered to be extreme violations of the assumptions of linearity, but, some errors or omissions may be expected when the process monitoring is developed. It should be noted that none of the six unfolding and scaling combinations removed these non-linearities in the data set in an acceptable manner. When faced with extreme non-linearities, the monitoring process may be improved by applying non-linear methods like Kernel PCA (Yoo et al. 2006).

Determination of the number of principal components using contribution plots

A critical step in PCA modeling is the determination of the number of principal components to be retained in the model. Qin and Dunia (2000) and Al-Kandari and Jolliffe (2005) elaborate on this subject within the framework of PCA-based sensor validation and reconstruction. In this section, an empirical method to use the loading plots of the principal components will be shown and evaluated. The selected principal components are limited to a number of components that capture all the present variables. The principal components are ordered along their captured variance (equivalent to ordering by their eigenvalues) from high to low and are evaluated in that order. When the dominant variables of the principal component under consideration are already dominant in the retained components, this and all following components are omitted. To illustrate the method, the selection for model 6 is explained in detail below.

Figure 5.13 shows the contribution plots of the first five components (with the 5 highest captured variances) for model 6. It can be observed that temperature and ORP are the dominating variables in PC1. The same holds for conductivity in PC2 and pH in PC3. PC4 is dominated by the weight. If the inclining method had been used, the PC4 will be the last one. But, in PC5, the DO is the dominating variable. DO is a very important variable in this process, but it is detected until PC5, because it is a variable very controlled. Thus, all the variables are well represented using the first principal components. Based on these observations, five principal components are thus selected, which capture 85.16% of the total variance. The rest of PC can be considered as noise. The same approach was used to select the principal components for models 4 and 5. It was observed

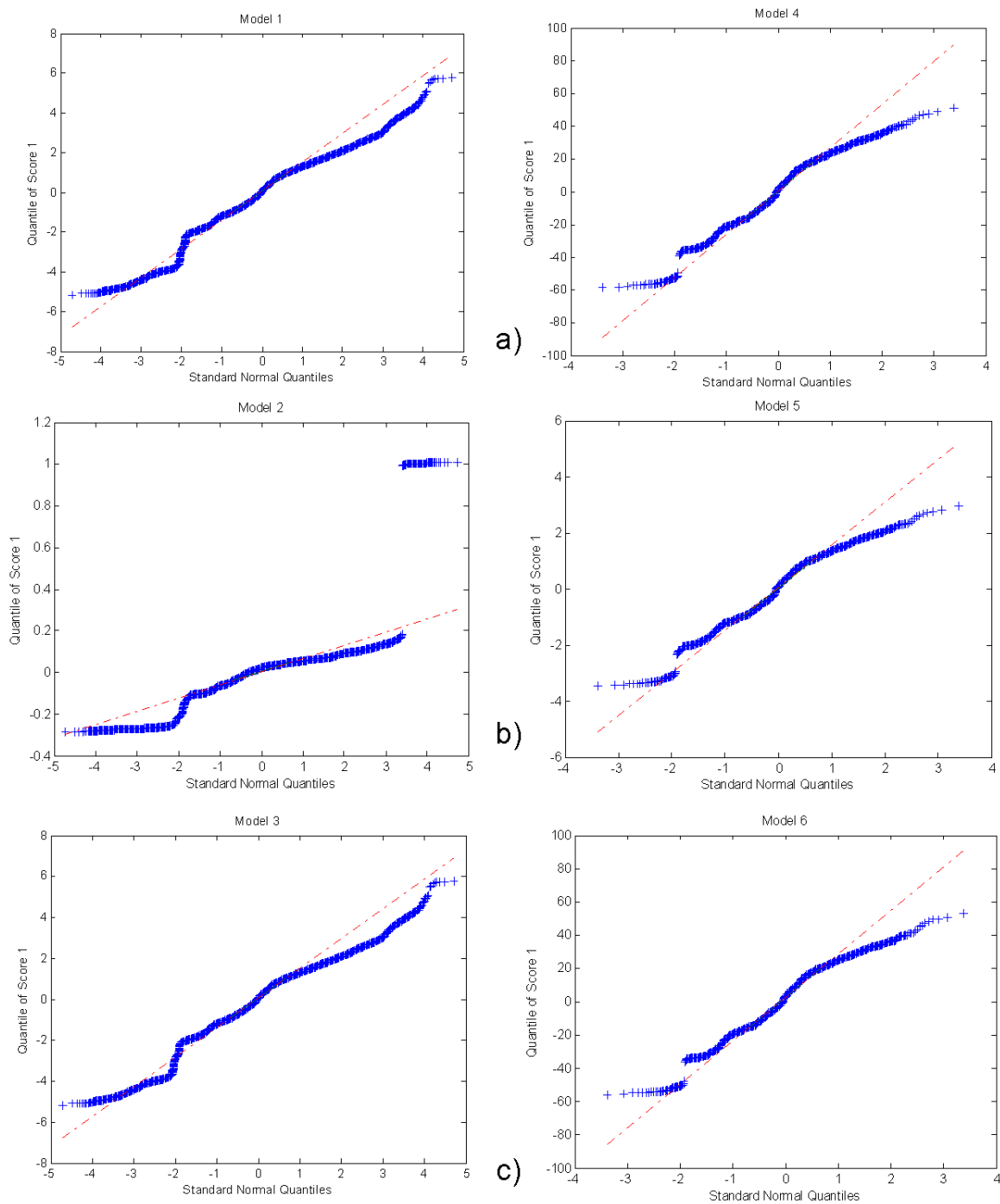


Figure 5.11: The Q-Q distribution of the first principal component for models that are unfolded variable wise (left) and batch wise (right) and scaled with a) CS b) GS and c) AS approaches

(data not shown) that the variables could all be found as dominant variables in the first five components. Table 5.6 summarizes the number of principal components and the total percentage of captured variance for each batch wise model.

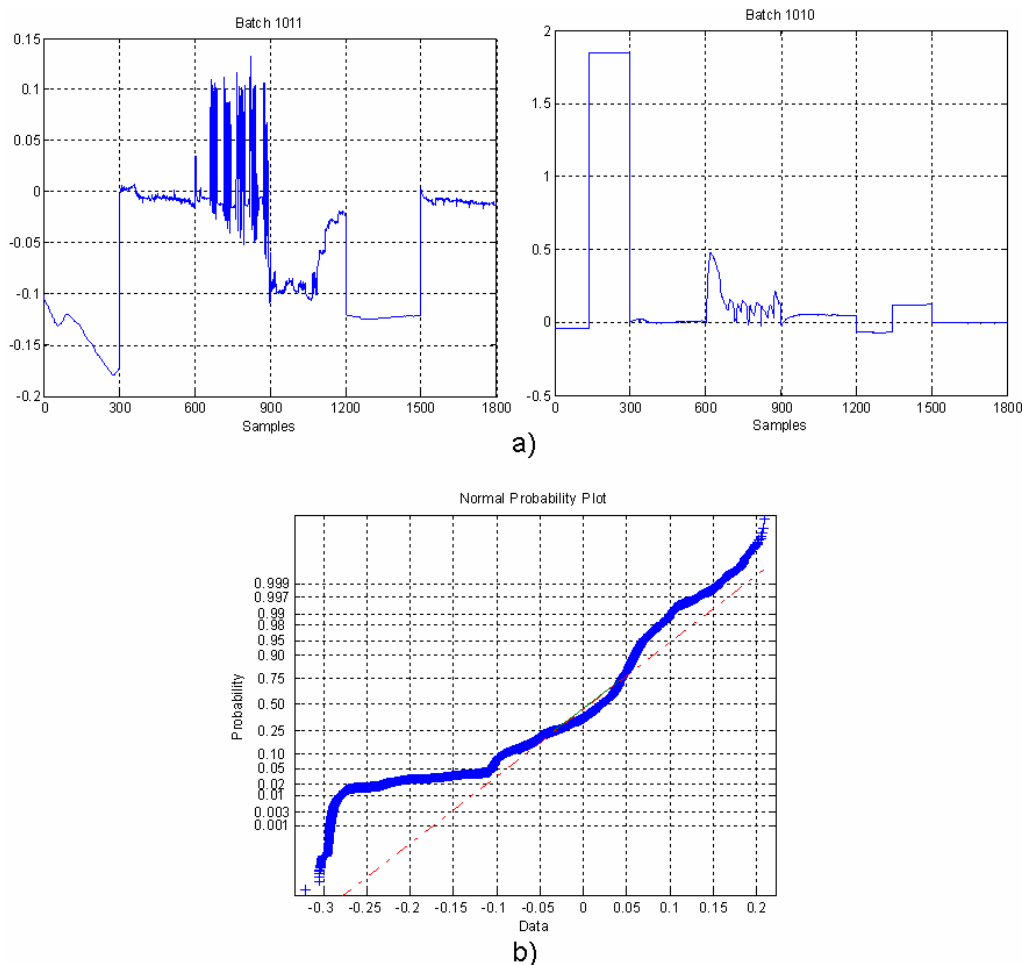


Figure 5.12: a) Batches 1010 and 1011 b) New Gaussian distribution for Model 3

Scaling process	Name of model	# of Principal Components	% Variance Captured Total
CS	Model4	5	91.50%
GS	Model5	5	85.27%
AS	Model6	5	85.16%

Table 5.6: Variances for models 4, 5 and 6

It was possible to observe that the three types of scaling led to similar captured variances until component number four. The fifth component magnified the difference between continuous scaling and other types of scalings. It should be kept in mind that auto-scaling and group scaling cause a larger decrease in the total variance than continuous scaling.

To find an adequate number of principal components using a variable wise method, Yoo et al. (2003) used the cross validation of the prediction residual sum of squares method.

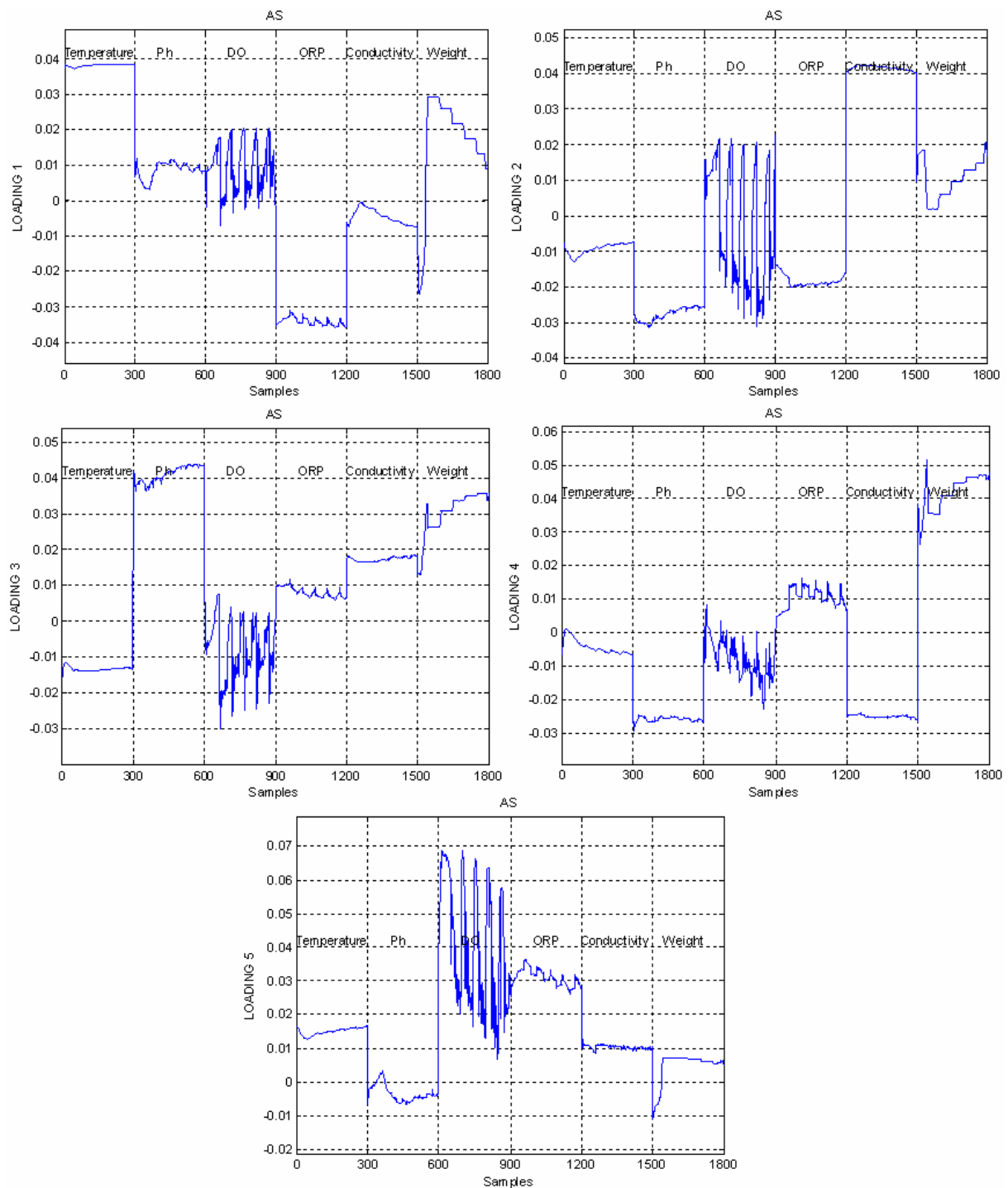


Figure 5.13: Loads graphics from components 1 to 5

In this work, six principal components were selected for models 1, 2 and 3 (see Table 5.7). Each PC represents different variables with 100% of the total variance for each model.

Scaling process	Name of model	# of Principal Components	% Variance Captured Total
CS	Model1	6	100%
GS	Model2	6	100%
AS	Model3	6	100%

Table 5.7: Variances for models 1, 2 and 3

Process monitoring by means of the resulting MPCA models

To evaluate the performance of the resulting MPCA models, the validation data set was projected onto the models. Each batch in the validation set was projected onto the 6 models, and the corresponding statistics were calculated and checked against their in-control limits. This means that each batch was classified 6 times. The analysis was divided in two parts:

1. First, each model was evaluated separately.
2. Second, the results were interpreted in terms of the effect of the scaling method by comparing the models within the group of batch-wise models and variable-wise models separately. Variable wise models are normally used for ON-LINE monitoring and batch wise models are used for OFF-LINE monitoring. In Tables 5.8 and 5.9, the criteria for fault detection are shown.

T^2	Alarm	NOC	AOC
0	0	True acceptance	False acceptance
1	1	False alarm	True alarm

Table 5.8: Criteria for performance assessment of the monitoring models in variable wise mode

Variable wise models: The main application of this kind of analysis is for ON - LINE monitoring. This monitoring allows the detection of abnormal operation in the process before the end of the batch. This kind of monitoring provides an opportunity for performing reconfiguration of the process when an abnormal operation is detected.

Q	T^2	Alarm	NOC	AOC
0	0	0	True acceptance	False acceptance
0	1	1	False alarm	True alarm
1	0	1	False alarm	True alarm
1	1	1	False alarm	True alarm

Table 5.9: Criteria for performance assessment of the batch wise monitoring

Batch wise models: The main application of this kind of analysis is for OFF - LINE monitoring. It takes place when the trajectories of the process variables are finished. The operator reconfigures the process to correct faults when the batch has finished.

The Hotelling T^2 statistic and Q -statistic charts with 95% confidence limits were used for the batch-wise unfolded models. For variable-wise unfolded models, only the Hotelling T^2 -statistic was available since the number of PCs was equal to the number of variables (i.e. the residual matrix is zero).

For proper evaluation of the performance of the models, three numbers were calculated from the batch runs in each of the aforementioned categories:

1. The sensitivity is the ratio of true alarms to the number of batch runs in the AOC data set. A higher sensitivity thus signifies a higher chance of detecting an abnormal batch run.
2. The specificity is the ratio of true acceptances to the total number of normal batch runs. A higher specificity thus signifies a lower number of false alarms. Generally speaking, higher sensitivity correlates with lower specificity and vice versa. Sensitivity and specificity have thus to be bargained against one another.
3. An overall performance index was calculated as the mean of the sensitivity and specificity; in this study, it was assumed that a false alarm has the same weight as a false acceptance in the decision process.

By doing so, the effect of unequal size of the AOC and NOC validation data sets is countered. In Tables 5.10 and 5.11 the number of true alarms, false alarms, false acceptances, and true acceptances are given together with the calculated sensitivity, specificity, and overall performance. The false alarms and false acceptances were determined by the 95% confidence limit.

Scaling	CS	GS	AS
Model #	1	2	3
True alarms (-)	43	430	426
False acceptances (-)	517	130	134
False alarms (-)	12	37	40
True acceptances (-)	268	243	240
Sensitivity (%)	7.7	76.8	76.1
Specificity (%)	95.7	86.8	85.7
Overall performance (%)	51.7	81.8	80.9

Table 5.10: Performance assessment of Variable Wise (VW) considered models

As can be seen, models 1 and 4 (CS models) exhibit poor performance (Individual analysis). Model 1 (CS, VW) shows an overall performance of just 51.7% and model 4

Scaling	CS	GS	AS
Model #	4	5	6
True alarms (-)	228	500	488
False acceptances (-)	332	60	72
False alarms (-)	38	43	47
True acceptances (-)	242	237	233
Sensitivity (%)	40.7	89.3	87.1
Specificity (%)	86.4	84.6	83.2
Overall performance (%)	63.6	87.0	85.2

Table 5.11: Performance assessment of Batch Wise (BW) considered models

(CS, BW) exhibits 63.6% for the same index. The latter index differs by more than 15% compared to the second worst model. Clearly, scaling the data with one overall mean for all variables produces poor performance. This remains true for both types of unfolding that were tested. These models (1 and 4) were therefore excluded from further comparisons. The remaining models exhibit sensitivities higher than 75%, specificities higher than 80%, and overall performances higher than 80%.

The results regarding the comparison between AS and GS models for each unfolding are similar. GS model 2 shows a slightly higher sensitivity than AS model 3 of +0.6%. The sensitivity increases +2.1% when model 5 is compared to model 6. These increases in sensitivity are accompanied by an increased specificity (+1.1 and +1.4%, respectively). A bargain is thus not at hand in this case. Preference is logically given to the GS scaled models even though the difference between these and the AS models is minor.

5.2.3 Discussion

In this chapter, a study was conducted to compare the effect of different types of unfolding and scaling of PCA models in terms of their capabilities for ON - LINE and OFF - LINE monitoring in a SBR Pilot Plant.

The primary result shows that scaling the batch run data with mean trajectories (as in AS and GS) is essential. Scaling with a single overall mean (CS) for each variable was shown to lead to poor performance.

Second, the need for monitoring during the batch run may be an important point of interest. In such a context, it is most logical to use both types of unfolding in parallel (i.e. to use a variable-wise unfolded model as the batch progresses and using a batch-wise unfolded model once the batch is finished). This will permit detection of a major fraction of the abnormal disturbances during the batch run, while ensuring an overall optimal performance.

With respect to the scaling to be used, a slight preference is noted for group scal-

ing (GS) over autoscaling (AS). Still, the difference in performances does not suggest a strong recommendation for the use of one particular type of scaling. It is possible that the slightly poorer effect of autoscaling can be explained by a magnification of the noise in non-informative variables, and this idea was suggested by Gurden et al. (2001) in the context of multivariate regression models. Additionally, it may be true that in our case, it is more meaningful to apply a single standard deviation for each sensor removed so that existing autocorrelations are not broken in the data preprocessing step.

From the ON - LINE monitoring, the best model found resulted in a sensitivity of 76.8%, specificity of 86.8% and overall performance of 81.8%. From the OFF - LINE monitoring, the best model found resulted in a sensitivity of 89.3%, specificity of 84.6%, and overall performance of 87.0%. In conclusion, the best scaling for monitoring is group scaling (GS). It is important to explain here that these results may be improved by either adjusting the in-control limits on the basis of kernel density estimation or applying an inherently non-linear model (Lee, Yoo, Choi and Vanrolleghem 2004). However, this was not considered within the scope of this study. In addition to the notes above, it is important to note that all results presented here stem from a cyclic system with equal cycle and phase lengths. Auto-scaling and group-scaling models require equal phase and equal cycle lengths. Because of this requirement, the corresponding conclusions cannot be extrapolated easily to systems with unequal phase lengths and/or cycle lengths. More importantly, a continuous scaled (CS) model with variable-wise unfolding is the only model that is readily applicable in such a case when no prior knowledge is available even though this model delivers the worst performance of all the considered models. Results from this section were presented as oral presentation in 6th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, September 2006, Beijing P.R. China (Ruiz, Villez, Sin, Colomer and Vanrolleghem 2006).

5.3 On-line MPCA application

This section is the result of a final project for industrial engineering with the purpose of applying knowledge compiled until now. This application is installed in the semi-industrial pilot plant described in Subsection 2.2.1. The tool allows statistical analysis for batches establishing a connection between the data base where the values of the variables are stored and the monitoring module.

The application implements the LABVIEW platform in which the monitoring module is implemented. It allows the selection of different MSPC models, including the models already created as well as models created from the new batches and/or updated by them. For the creation and importation of these models, it is necessary to create connections between MATLAB (*PLS_Toolbox* 3.5) and LABVIEW (see Figure 5.14). The models are created internally within the same application. The application presents a set of tables and graphs. It allows visualization of the results obtained from the MSPC analysis.

The tool allows to make analysis in real time of the process, knowing the evolution of

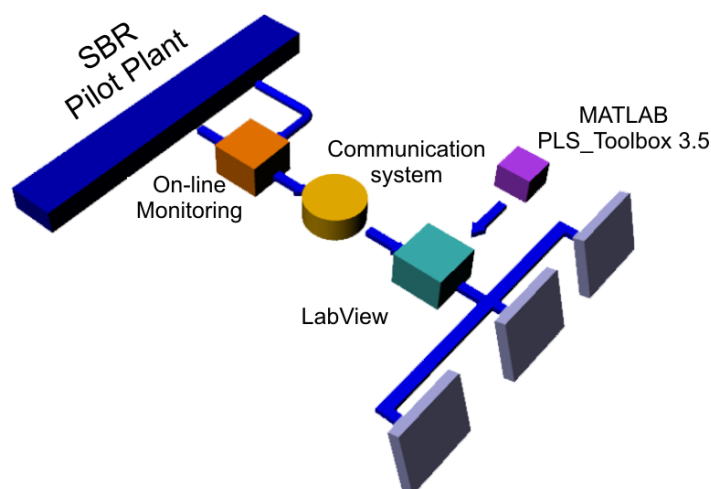


Figure 5.14: Schematic representation of the interface

a new batch and the operation of the previous batch. Also, it is possible to make off-line analysis. The on-line application is formed by three modules, each one with determined functions.

1. Module for on-line Monitoring (variable and batch wise)
2. Module to build the models (variable and batch wise)
3. Module to validate the new batches (batch wise)

5.3.1 Module for ON-LINE Monitoring

The main goal of this module is to automatically make the statistical analysis in real time. Two statistical analysis methods can be used:

1. Variable wise: made over the new batch in real time for each sample.
2. Batch wise: made once the last aerobic stage is finished allowing to know the batch operation before it is concluded.

In conclusion, this module presents the information quickly. The information is shown in a simple manner and is updated periodically. The module works automatically, includes the calculation algorithms and displays the statistical analysis of batches. It calculates and displays the statistical analysis per batch connecting periodically with the sensor files once per minute (Section 2.2.1). The values obtained from the sensors are stored. This connection allows to obtain the parameters like the number of batches validated, the stage of the present batch, and the information tables.

Once the process data have been imported and when the new batch starts, the statistical analysis is made. At any instant during the processing of the batch, the operator can decide if the reference model should be changed. In this case, the data of the new model selected is made by means of the connection with MATLAB.

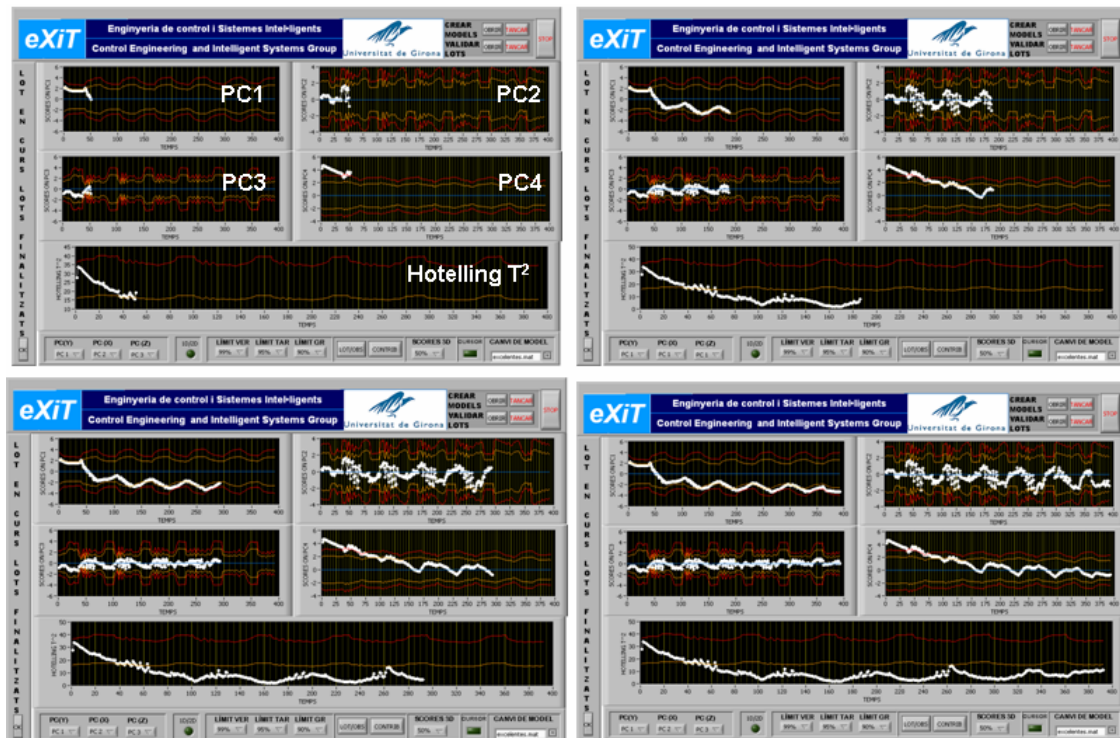


Figure 5.15: Interface to on-line monitoring in variable wise mode

Figure 5.15 shows the interface for ON - LINE monitoring (variable wise). The scores vector corresponding to the main components and the value of the T^2 statistic from the statistical analysis in variable wise mode are calculated. These values are shown within the corresponding principal components (PC1, PC2, PC3 and PC4). In this way, the operator knows the present state of the batch for any moment.

When the connection between the process and the data base detects that one new batch is starting the purge phase, the statistical analysis in batch wise mode is made immediately (see Figure 5.16). The fact that the purge and sedimentation data do not contribute to the biological information (in accordance with Section 2.2) allows that the analysis of the batch to be made half an hours early. Therefore, it is possible to reconfigure the system before a future batch starts. As in the variable wise analysis, the values of the score vector, Q statistic, and T^2 statistic are calculated and stored. In addition, the percentage contribution of each component is obtained (see Figure 5.17). Each new result is stored in chronological order (from oldest to most recent batch).

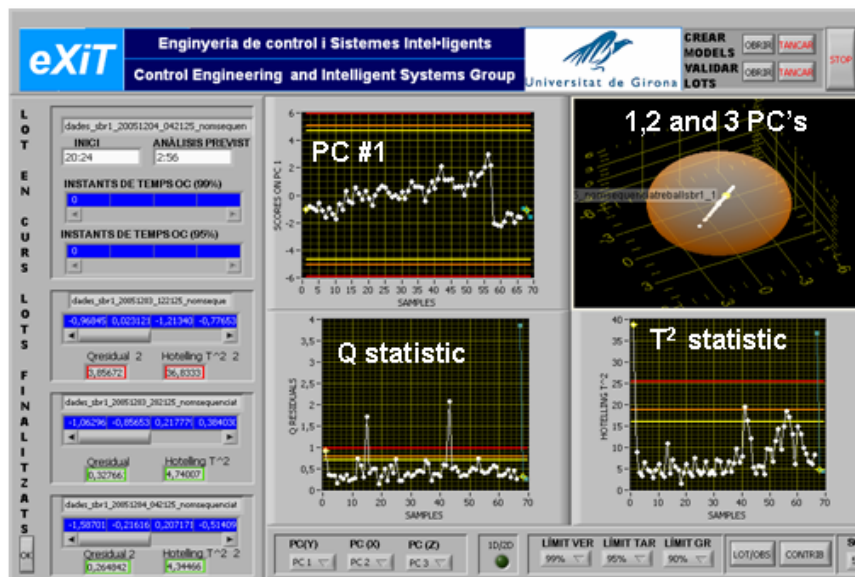


Figure 5.16: Interface to on-line monitoring in batch wise mode

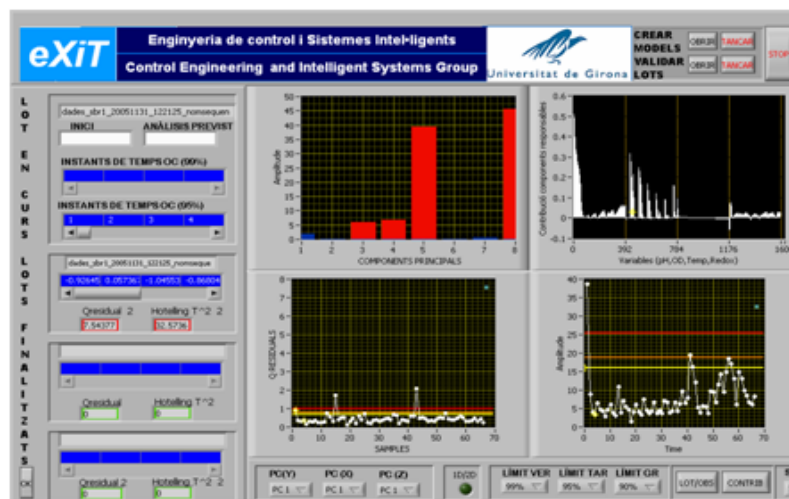


Figure 5.17: Contribution interface of each component

5.3.2 Module to build the models

This module allows the operator to create and/or update models. These models can be created in two different ways:

- Using a new batch.
- From batches stored in the results files.

In both cases, once the model has been created; it can be edited to eliminate any batch and thus build a new model. This module allows the display of models details.

Before creating the model, it is necessary to determine the number of principal components. The module allows up to twenty principal components. Figure 5.18 shows the graphical interface in order to apply the methodology developed in Section 5.2.

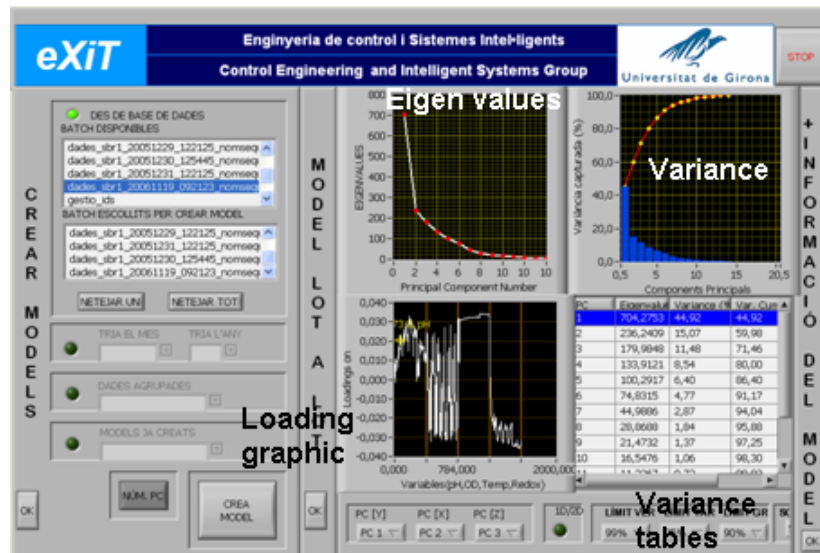


Figure 5.18: Interface to determine the number of principal components

When the number of principal components has been determined, the model can be built. The operator can decide which kind of batches to use in order to build the model. The module for validation has a periodic connection with the results files. This allows the building of models using a specific operation or batches of a particular period of time.

Once the necessary information is loaded, the models are built using MATLAB. The communication between LABVIEW and MATLAB is done using the block of LABVIEW Matlab Script. When a new model is built, all the information is stored inside the results files. There are two types of information:

1. The threedimensional matrix which is used to build the model
2. Information about what kind of scaling was used, the name of the new batch, start time, and others. In Figure 5.19, the window of complementary information is shown.

The operator can modify or update the models in variable and batch wise modes. When the data set is ready, the model is built. In this module, the contribution plots of the variables are calculated. When one or more batches are detected as Abnormal Operation Condition (AOC) the contributions plots are used to find the variable with the fault. Figure 5.20 shows the contribution of the principal components and the contribution of a specific batch.

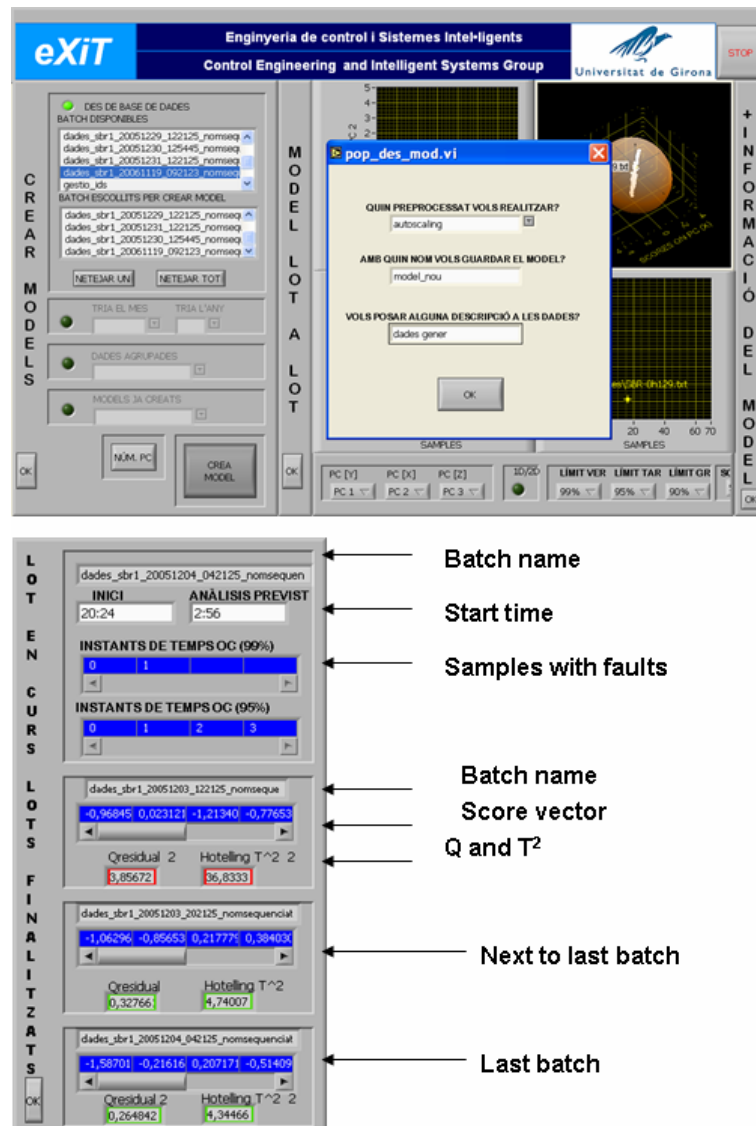


Figure 5.19: Window of complementary information

In general, the module allows the operator to check all the information that is generated when the models are built. The operator can select the graphic. Figure 5.21 shows other windows generated by the module.

5.3.3 Module to validate new batches

The application has a special module in order to validate new batches. In this module, the operator selects the data set stored in the results files in order to project it onto the model. The operator can change the model as is necessary. Using this module, it is possible to know the operation of one or more new batches. The validation is done using the batch wise method.

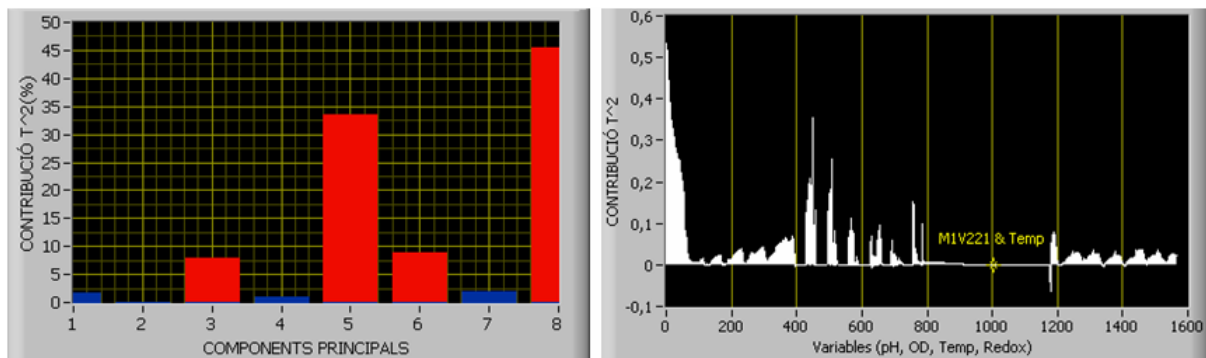


Figure 5.20: Contribution analysis graphics

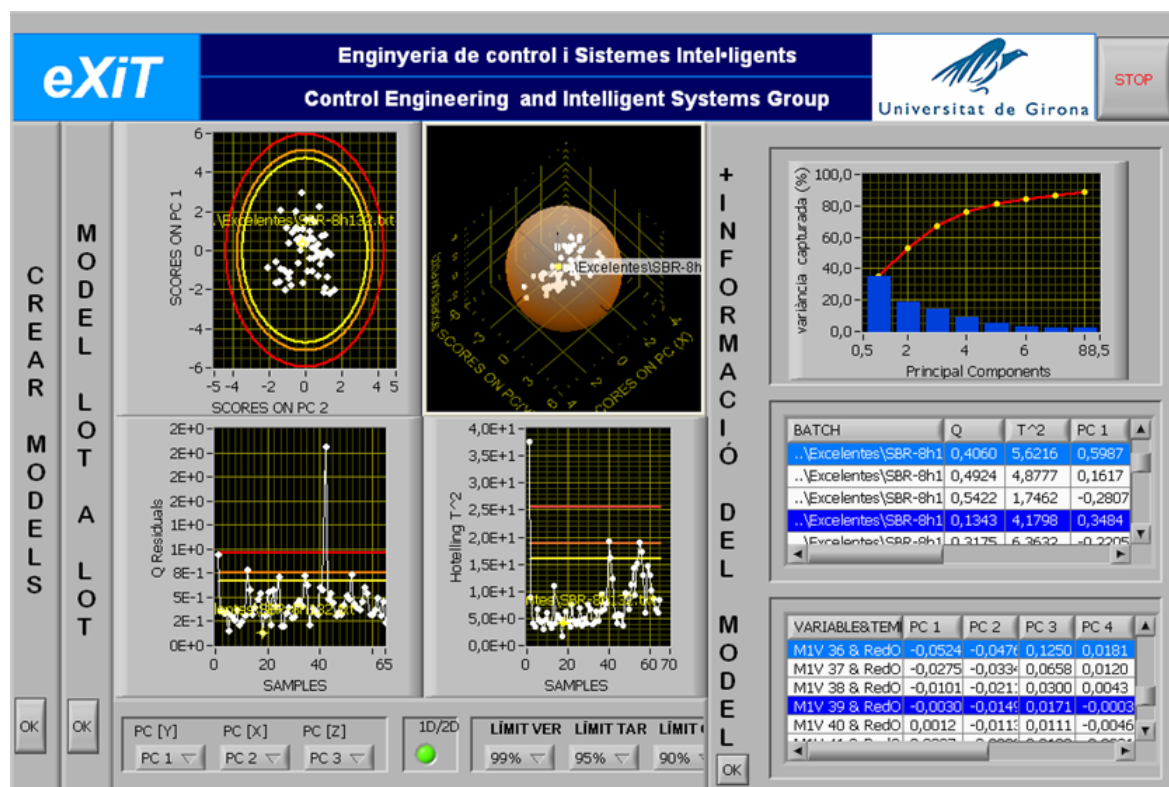


Figure 5.21: Contributions analysis graphics

The module is connected periodically with the results files. In this manner, the information is updated constantly. The batches can be projected over any model. The calculation is done using the platform LABVIEW. The preprocessing of the validation data set uses the information stored from the models. The results from the previous analysis can be pictured in their corresponding graphics and tables. This information represents the score vector, Q statistic and T^2 statistic, as well as the contribution analysis (see Figure 5.22).

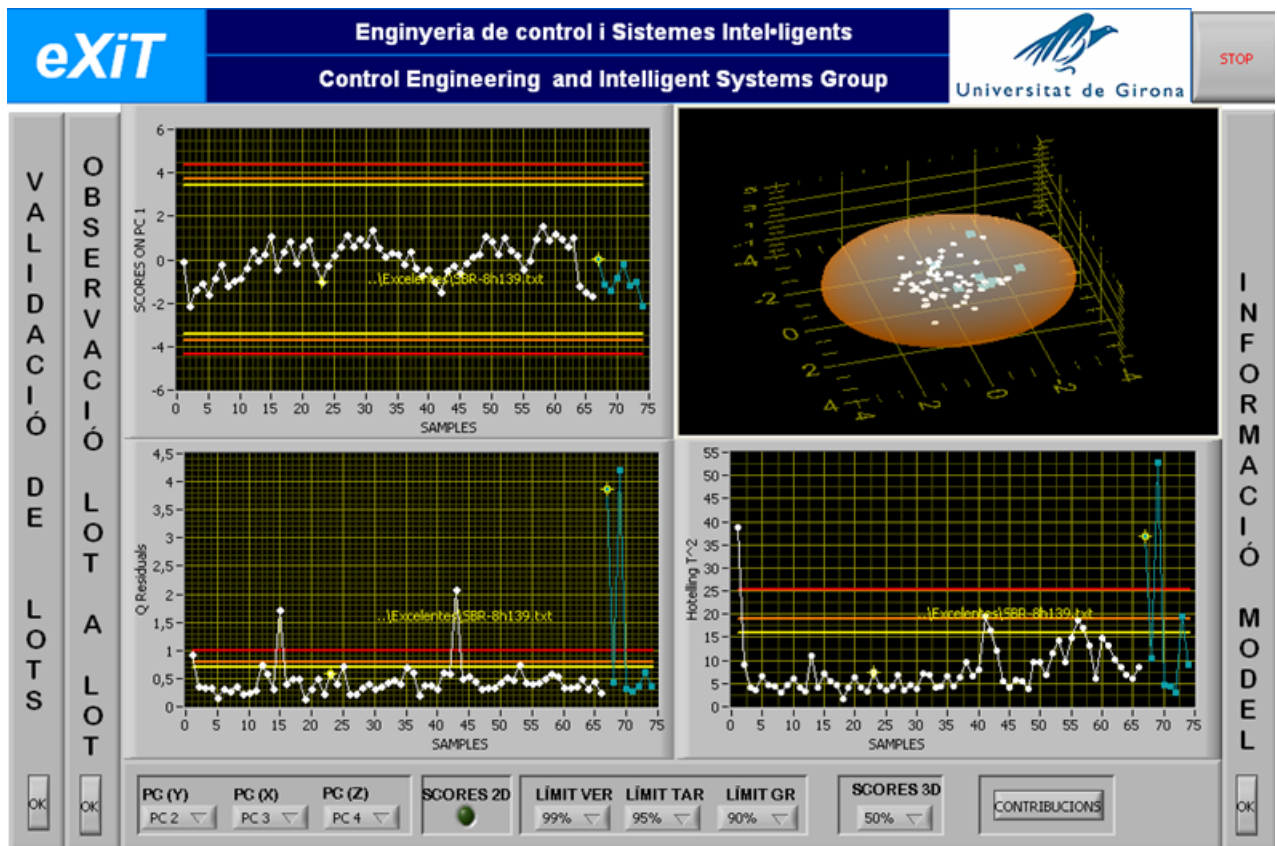


Figure 5.22: Contributions analysis graphics

5.3.4 Conclusions

The on-line application has been developed using a friendly interface making easy the daily operation of a semi-industrial SBR pilot plant. Three modules were developed which a basic knowledge about MPCA is needed. (i) The module for on-line monitoring was based over the variable wise mode. The model represents the variables trajectories through the real time. One, two or more samples are detected if they are outside of control limit. (ii) The module to build models was based over the batch wise mode. The models (batch wise and variables wise) are updated wherever it is consider needed. (iii) The module to validate new batches is performed using batch wise mode. In this case, the knowledge about monitoring and fault detection are performed until last aerobic condition. As settling and draw phases are not included in the model, the batch situation is found with one and half hour allowing the process reconfiguration before beginning a new batch. Results from the last two section are in preparation for publication in Water Science and Technology (Ruiz, Villez, sin, Colomer, Rosen and Vanrolleghem 2008).

5.4 Analysis and Conclusions

In the methodology applied for the SBR pilot plant from LEQUIA, a model was built using the whole data set without any distinction of NOC and AOC batches. In general, MPCA and knowledge from the experts are tools that complement one another very well. Both tools contribute to a general knowledge of the state of a process. However, this technique did not perceive all of the batches with AOC, and the methodology was complemented by a LAMDA classification tool (see Appendix A). The results were promising, and therefore it was decided to conduct an in depth study of which methodologies should be applied to this kind of process. Therefore, the work on the SBR pilot plant from BIOMATH was proposed.

In the methodology applied to the SBR pilot plant from BIOMATH, several approaches for PCA-based biological process monitoring that have been discussed in the literature were compared. The constructed models exhibited different methods for monitoring (ON -LINE and OFF -LINE) and three types of scaling (continuous scaling, group scaling and autoscaling).

ON -LINE and OFF - LINE monitoring differs, because variable-wise models exhibit the advantage of being readily applicable for monitoring of running batches, whereas batch-wise models deliver a better detection performance. As a consequence, it is suggested that both types of models be used in parallel to both detect faults when possible during a running batch and ensure an optimal overall detection rate.

In general, MPCA is an effective dimension reduction technique in data mining. To apply MSPC methodology in SBR processes requires normalization of the original data by appropriate scaling. Utilization of the PCA method requires proper unfolding of the original $3D$ data array into a $2D$ matrix in order to permit proper monitoring of the process. Another important contribution of this chapter is the way by which the best number of principal components is selected.

Chapter 6

Automatic Detection of Abnormal Situation in Process Operation

The main objective of this chapter is to present a methodology for diagnosis of the process. Villez et al. (2006) obtain good results by combining PCA with LAMDA clustering. Case-Based Reasoning (CBR) is proposed as an Artificial Intelligence approach that can be applied to improve expert supervision by exploiting data obtained from the MPCA results. The advantage of CBR is that the Case Base is built just once; maintenance and updating are accomplished through the learning capacity of this tool. In Sanchez-Marre et al. (1997), Nuez et al. (2002), Wiese et al. (2004) and Martinez et al. (2006), some applications of CBR can be appreciated directly from the sensor of WWTP's. Here, an implementation of the MPCA approach with CBR is proposed. For this purpose, and in accordance with Ruiz, Villez, Sin, Colomer and Vanrolleghem (2006), MPCA has been used as a dimensionality reduction tool which is able to obtain good representation of the process in few variables. Next, the results from MPCA are used as descriptors by CBR.

This chapter is organized as follows: First, the work methodology is explained, special attention is on descriptors and distance definition (retrieve) and case base maintenance and updating. Immediately afterwards, the results are shown. Finally, the chapter finishes with conclusions and discussion of future work.

6.1 Methodology

The main goal of this doctoral thesis is to develop a methodology to monitoring, fault detection and diagnosis using historical data from several WWTPs. According to the results and analysis from the previous chapter, it is necessary to build a MPCA model following the below steps:

- Scaling process data is performed in order to obtain samples with similar range.
- A specific number of PC's are selected reducing the dimension of the data.
- T^2 and Q statistics are used to determine faults.

In this way, ON-LINE and OFF-LINE monitoring are performed. Results are good, however some omissions and false alarms are presented. As complement to MPCA, the use of CBR is proposed in this chapter. An initial Case Base (CB) is built using indices from MPCA. These are: PC's, T^2 and Q -statistic. The selection of these indices to describe one case and the way to retrieve the neighbor cases need a specific study. Afterwards, this CB needs maintenance and updating in order to further improve the results. The methodology developed for combining MPCA and CBR for situation assessment is illustrated in Figure 6.1.

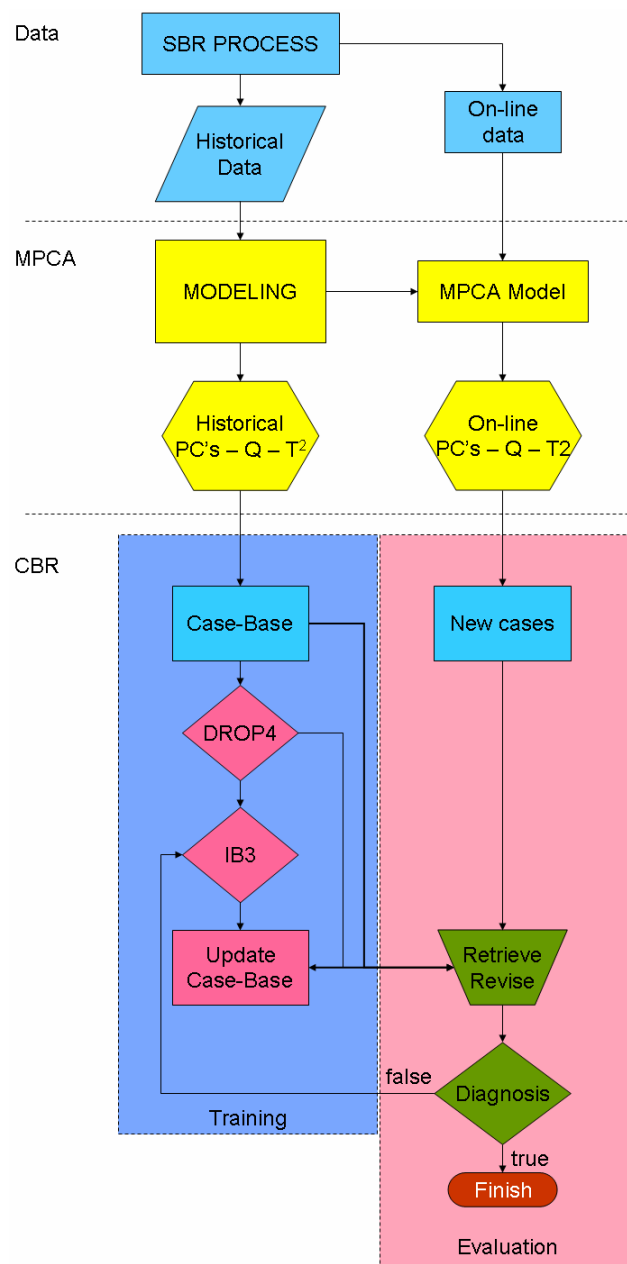


Figure 6.1: Methodology applied to SBR pilot plant

When a new case is presented, the MPCA model has to be applied to this new data in order to obtain the descriptors which will characterize the case. In this way, the new case can be compared with the CB. When a similar case is retrieved, it is selected and reused to diagnose the current situation. If this new case does not have solution, it is revised and retained into the CB for future diagnosis. In the next subsections, some aspects the proposed methodology application is explained.

6.1.1 Data

Three processes have been used to test the proposed methodology in order to solve the crucial points of this chapter: select the appropriate combination of descriptors, distance definition (retrieve) and case base maintenance and updating. The data sets were explained in Chapter 2:

1. The first data set comes from the SBR pilot plant from the BIOMATH group (see Section 2.2.3).
2. The second data set comes from the COST/IWA Benchmark (see Section 2.1.1).
3. The third data set comes from the SBR pilot plant from the LEQUIA group (see Section 2.2.2).

These data sets are presented in different formats which should be treated properly to be used within a working method.

6.1.2 MPCA

The data are organized to be scaled, and also organized in a three-dimensional matrix to be unfolded in two dimensions depending on the desired direction. At this point, it is possible to obtain the respective statistical models. MPCA obtains indices that characterize each one of the batches. These indices are the principal components PCs, Q -statistic, and T^2 -statistic with 95% confidence limit. These indices are called Descriptors. Depending on the model, the number of principal components can change according to the selection criteria. It is possible to detect some faults and false alarms, if MPCA is used as the only tool for fault detection (Ruiz, Villez, Sin, Colomer and Vanrolleghem 2006). However, it could be possible to finalize the methodology by this step; though, some omissions can be produced. For this reason, the CBR has been used as a complement to MPCA, because the best aspect of CBR is its capacity for learning from experience.

6.1.3 CBR

Once the descriptors are selected, it is fundamental to select a system to retrieve the nearest neighbors and create a CB which handles maintenance and updating. Two blocks are necessary to enhance the CBR application and both are developed at the same time.

1. Training block: Each batch stored in the CB is called CASES. Each case is defined by descriptors and the operation condition. When the DB is built, it is necessary to remove repetitive cases which could induce wrong situation assessment (*DROP4* application). Additionally, to cover all possible situations, it is necessary that the CB learns about new situations; for this, *IB3* is applied. The CB is under examination constantly because if one new case is not correctly diagnosed, as a consequence, this case will be stored automatically in the CB for future situation assessments.
2. Evaluation block: The remaining data is used in order to apply the CBR algorithm. This means that each case of the remainder sets are compared with each case stored in the DB by means of distance measurements from searching their neighbors.

To develop this methodology, it has been necessary to solve two main problems.

1. Descriptors and distance refining: The goal in this phase is to find the best combination of descriptors as well as the best method for calculation of the distance between new problems and the cases stored in the case base.
2. Case base maintenance and updating: The goal in this phase is to apply the *DROP4* and *IB3* algorithms in order to achieve two main properties of CBR: maintenance (clearing of redundant cases) and updating (learning of the case base).

6.2 Descriptors, case base and distance refining

In this implementation, the data used comes from the SBR Pilot Plant from the BIOMATH group (see Section 2.2.3). In this data set, 1588 batches were used, with a collection of 6 variables per batch, where each variable contains 300 samples. In this manner, 1800 samples per batch are used to apply the MPCA and CBR methodology.

In order to find the best combination of descriptors and the best way to define the selection system of the nearest neighbors, several steps were necessary to develop the approach taken to compare the discussed options in PCA-based + CBR monitoring. The diagnosis is explained in Figure 6.2 and it is further explained below.

6.2.1 Step 1: Definition of descriptors

Each new problem and each case stored in the *CB* are described by means of the results from the MPCA approach. These results are called descriptors. In consequence, seven descriptors were used as follows:

- Descriptor 1 corresponds to the Q value;
- Descriptor 2 corresponds to the T^2 value;
- Descriptors 3 until 7 are the five Principal Components (PCs)

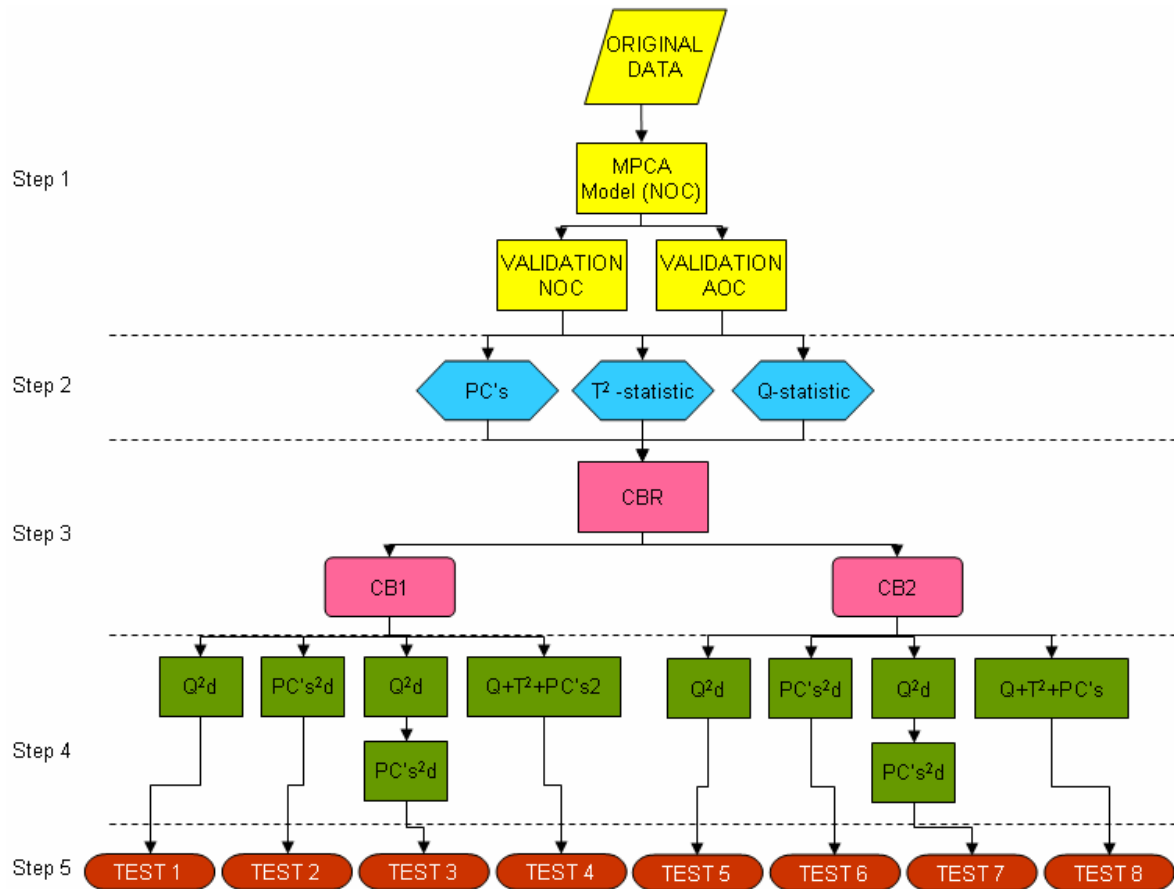


Figure 6.2: Test strategies to select descriptors and distances

6.2.2 Step 2: Building the MPCA model and the validation data set

The data are divided into three sets in accordance with Villez et al. (2006). The first is used to build the model. The second and third sets are then used to validate the model.

- **Set 1** Building the model: 1119 batches with normal operation conditions are used to build the MPCA model. To guarantee that the seed will be homogeneous, the Q value larger than 1 due to were removed. In consequence, 1020 batches were used to develop the CB .
- **Set 2** Validation data set of AOC: In this set, there were 560 batches with abnormal operation conditions.
- **Set 3** Validation data set of NOC: 280 batches with normal operation conditions.

Using MPCA, the model number 5 built and the knowledge obtained from Chapter 3 was used. In this model only batches with Normal Operation Condition (NOC) are used.

1119 batches were selected and stored in a three-dimensional data array and it was unfolded as suggested by Nomikos and MacGregor (1994a). Next, the data were scaled, calculating one standard deviation per variable and one mean for each sample throughout the batch. This data scaling is commonly called "group scaling".

The model with the T^2 -statistic and Q -statistic was developed using a 95% confidence limit. Five principal components were selected with 85.27% of the variance captured. The validation data sets (NOC and AOC) were stored, with unfolding and scaling using the same methodology. Using the T^2 -statistic and Q -statistic, it was possible to determine the number of batches outside the limit for each validation data set.

6.2.3 Step 3: Building the Case-Base

In some implementations, simulation situations of normal and abnormal operation are necessary in order to build the Case Base (CB). But, in this work, the data process contained a sufficient number of batches. In this way, some batches are separated in order to build and check the better option. Two CB are constructed. Both CB contain NOC and AOC data sets. The difference between them is the number of batches.

- $CB1$ is composed of the entire data set of batches used to build the model in MPCA. It contains 1020 batches with NOC as well as 20% of batches from the validation data set of AOC (112 batches). As a consequence, this CB has a total of 1132 batches, as shown in Figure 6.3. The entire validation data set of NOC (280) is used in order to test the CBR. The remaining 80% of the batches from the validation set of AOC were tested as well; thus, 448 batches with AOC were used to check the CBR approach.

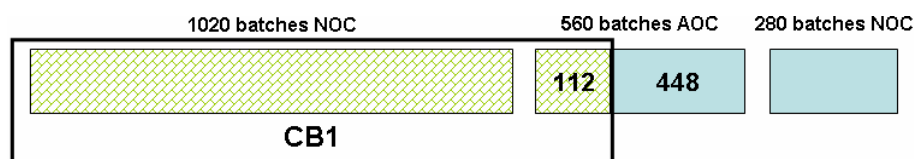


Figure 6.3: Case Base 1

- $CB2$ is composed of just 448 batches from the data set used to build the model. The same number of batches were selected from the validation data set of AOC or 80% of the total set. Thus, the CB has a total of 896 batches, as seen in Figure 6.4. In the same way as above, the entire data set of NOC was used to test the CBR and the remaining 20% of the batches from the AOC data set. Because $CB2$ only uses a small portion of the model data set, the remaining portion is used to validate the CBR approach; the remaining portion contains 572 batches, the rest of the model data set.

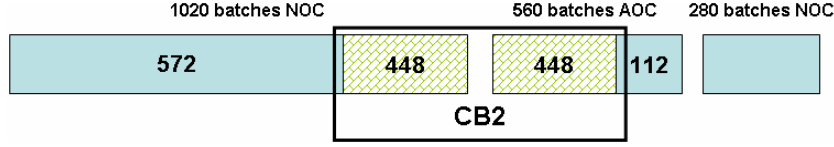


Figure 6.4: Case Base 2

6.2.4 Step 4: Retrieval

For calculation of the distances to the nearest neighbors, four different options were considered:

- *Q*-distance: The distance is calculated by comparing the new problem and batch-bybatch i from the case base using only the Q descriptor for the problem and the CB.

$$Q_d^2 = (Q_{new} - Q_{iCB})^2 \quad (6.1)$$

- *PC*'s-distance: the distance is calculated using only the *PC*'s as a unique descriptor. It is accomplished by comparing each *PC* $p=1,..5$ from the new problem with each batch i from the CB. Of particular relevance is the weight (W_p) assigned to each *PC*'s. In this work, the weights are the eigenvalues calculated by equation 3.4. These eigenvalues represent the percentage of variance for each of the principal components. This means that the highest value represents the most important pattern in the data, the second highest value represents the second most important pattern in the data and so forth.

$$PC' s_d^2 = \sum_{p=1}^P W_p * (PC' s_{pnew} - PC' s_{ipCB})^2 \quad (6.2)$$

- Two-steps retrieval: Q and *PC*'s distances are implemented in this method, and two steps are take into account. First, the Q descriptor is used to retrieve the 30 nearest neighbors in accordance with equation 6.1. Second, the *PC*'s distance (equation 6.2) uses the 30 pre-selected cases and only the first five nearest neighbors. Finally, the solution for the new case is determined by voting. For instance, if the nearest neighbors are 3 NOC and 2 AOC, the case will be categorized as NOC because NOC is the set with more population.

- $(Q + T^2 + PC's)$ distance: This distance is calculated by employing all descriptors including T^2 , which was not taken into account by the other distances.

$$\begin{aligned}
 Q_d &= (Q_{new} - Q_{iCB})^2 \\
 T_d^2 &= (T_{new}^2 - T_{iCB}^2)^2 \\
 PC's_d^2 &= \sum_{p=1}^P W_p * (PC's_{pnew} - PC's_{ipCB})^2 \\
 distance &= \sqrt{Q_d + T_d^2 + PC's_d} \tag{6.3}
 \end{aligned}$$

6.2.5 Step 5: Testing

To improve the performance of the MPCA and CBR methodology, CB1 and CB2 were evaluated using the four distances explained above. In other words, for both created *CBs*, four different distances are checked. Table 6.1 summarizes how the respective options were labeled.

	<i>CB1</i>	<i>CB2</i>
<i>Q</i> -distance	TEST 1	TEST 5
<i>PC's</i> -distance	TEST 2	TEST 6
Combining distances	TEST 3	TEST 7
$(Q + T^2 + PC's)$ distance	TEST 4	TEST 8

Table 6.1: Names for each developed CBR

6.2.6 Results

In this section, the MPCA results are given and the performance of the CBR methodologies is presented.

MPCA

The data are stored in a three-dimensional data model array that has a size of (1119 x 6 x 300), where 1119 corresponds to the number of batch runs, 6 represents the number of process variables, and 300 represents the sample number. In figure 6.5, the first three PC's are illustrated, where the circle is the in-control limit of the model. In order to compare the results obtained using the classical control charts (Q and T^2), a batch is considered to be abnormal if Q or T^2 are outside their control limits.

Results of validation set of AOC: In this set, 500 batches are detected with abnormal operation. This is a sensitivity of 89% where sensitivity is defined as the true alarms. To obtain these results, it is necessary to join the results from the Q -statistic and T^2 -statistic charts. Table 6.2 shows the percentage of batches detected by each control

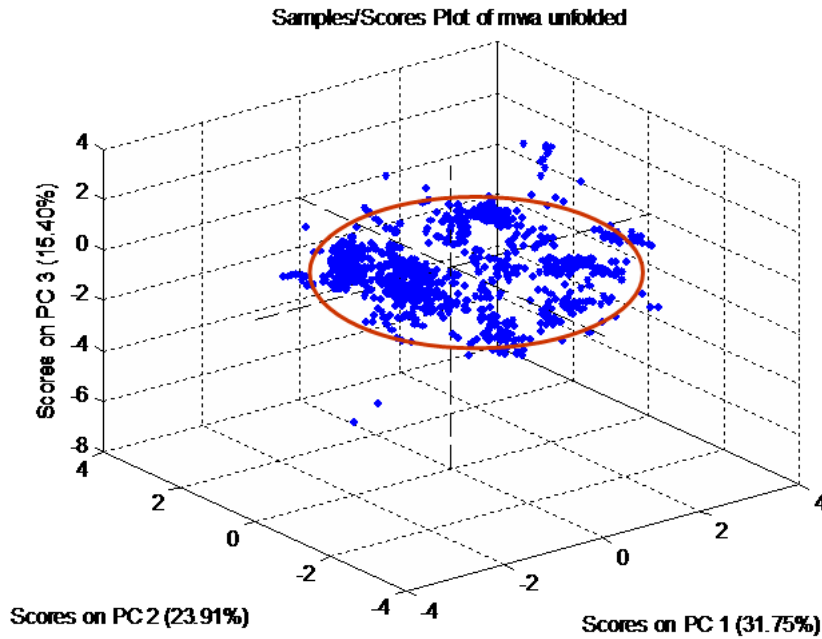


Figure 6.5: Dispossession and unfolding of three-way data array

chart.

Results of validation set of NOC: In this set, 43 batches are detected with abnormal operation. This is an error or false alarm rate of 15.36%. To obtain this result, it is necessary to join the results from the Q -statistic and T^2 -statistic charts (see Table 6.2).

Parameter	Specificity	Sensitivity
Q-statistic	91.79	82.14
T-statistic	91.43	61.25

Table 6.2: Specificity and sensitivity of each control charts

In agreement with the results obtained from the previous study of this process, a robust discrimination between different anomalies and operational changes will be used to diagnose the SBR process (Villez et al. 2006). CBR is only used for detection in order to identify the better methodology.

CBR

To improve the performance of the MPCA and CBR methodologies, CB1 and CB2 were evaluated based on several methods of how to calculate distances as explained previously (Table 6.1). Two groups of simulations were performed in order to select the best combination between the CB and the distances calculation. The analysis and discussion of the results are divided into two parts. In the first analysis, the simulations from each CB are

evaluated separately. For proper performance evaluation, the sensitivity and specificity percentages are calculated. In the second analysis, all simulations are analyzed together in order to compare CBs.

First analysis: Table 6.3 exhibits the first group of simulations applied to CB1. The performance level was low when the validation set of AOC was checked. The reason for these results may be the amount of NOC batches that are loaded into the CB1 which is higher than the number of AOC. However, the results improve when the validation set of NOC are inspected. In spite of TEST 3 was the lowest detecting AOC, this has a specificity of 100 due to the number of cases stored as NOC . In general when TEST 1, 2, and 4 are proved, the specificity is good.

	Specificity NOC set	Sensitivity AOC set
TEST 1	92.5	85.49
TEST 2	97.1	43.75
TEST 3	100	7.37
TEST 4	95.7	70.1

Table 6.3: Specificity and sensitivity for Case Base 1 (CB1)

With respect to CB 2, Table 6.4 shows the results of these test. The validation sets are the same as those used in CB 1. Nevertheless, the batches that were not taken into account for CB 2 are inspected as well. This set of batches is given the name NOC rest. When the validation data set of NOC and NOC Rest are inspected, the sensitivity was found to be good for TEST 5, TEST 6, and TEST 8. When the AOC set is checked for TEST 5, TEST 6, and TEST 8, the sensitivity performance was poor. However, these results improve and show the sensitivity to 100% when the data sets are checked using TEST 7.

	% Specificity NOC set	Sensitivity AOC set	Specificity NOC Rest set
TEST 5	90.36	62.5	97.2
TEST 6	88.57	47.3	91.43
TEST 7	100	100	100
TEST 8	92.86	53.57	97.73

Table 6.4: Sensitivity for Case Base 2 (CB2)

Second analysis: In relation to the selections of which Case Base (CB) is better, the results are clear; using CB2, the performance of the methodology for validation of AOC is improved. In conclusion, CB2 is the best option for this work. The results of this section were presented as oral presentation in the European Control Conference 2007, July 2-5, 2007, Kos, Greece (Garcia et al. 2007). A study using the results from this section was applied to the COST/IWA BENCHMARK. The goal of the application is to check the response of the methodology in a continuous process.

6.3 Application of descriptors and distance refining to the COST/IWA BENCHMARK

The goal of this section is to assess normal and abnormal operation conditions from a WWTP using the COST/IWA simulation benchmark (Copp 2002). The "benchmark simulation" goal is the performance and cost-effectiveness of wastewater control systems providing detailed descriptions about the plant layout, model parameters and simulation models. The benchmark simulation provides a base for comparing past, present and future control strategies without reference to a particular facility collecting a large amount of data. In this way, Multivariate Statistical Process Control (MSPC) has proven to be a powerful tool for monitoring, compressing and extracting data and identifying linear combinations of variables that describe major data trends (Russell et al. 2000). Due to the process nature, different MSPC techniques, including Principal Component Analysis (PCA), Multiway Principal Component Analysis (MPCA) and Dynamic Principal Component Analysis (DPCA) are examined. Where one day corresponds to one complete process.

6.3.1 Methodology

The methodology is tested using the BSM1 benchmark plant layout. One common inconvenience in CBR is the weight assigned to each descriptor, here the percentage of variance per each principal component is used. The necessary steps as shown in Figure 6.6.

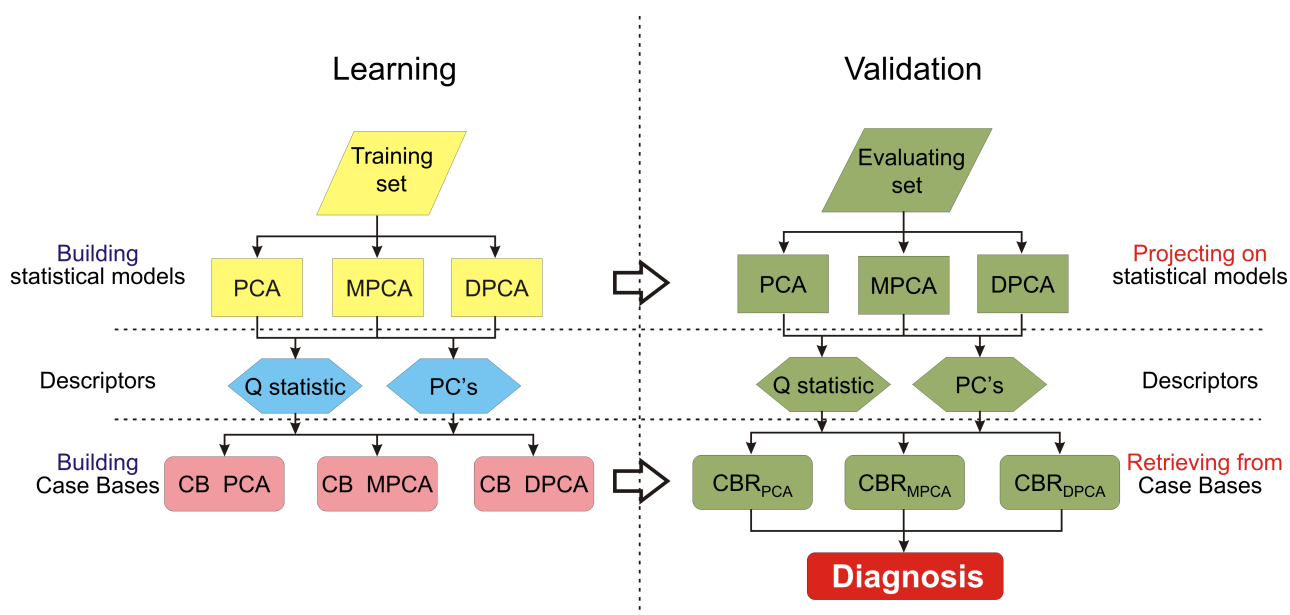


Figure 6.6: Projection of the process variables in a new space using PCA

Building the statistical models

Days with Normal Operation Conditions (NOC) were simulated in the training set. They are used to build several statistical models:

- One model using the PCA approach is developed. This model represents the relation between the variables.
- Another model using MPCA is developed. Using unfolding, as is explained in Section 5.2.1, the model represents the day-to-day relations of the process.
- DPCA is used to develop another statistical model. In this case, several simulations are used to identify the best stacking (size of window), then one option is selected to develop the model.

In conclusion, three statistical models are developed, one for each approach. In addition, the Q -statistic is calculated with 95% confidence limits.

Descriptors: Each new problem and each case stored in the CB are described by means of the results from the statistical models. These results are called descriptors. As a result, three sets are created per model each with four descriptors as follows:

- Descriptor 1 is the Q -statistic value;
- Descriptors 2 to 4 are the three first principal components from each statistical model.

Retrieval: As was shown in Section 6.2, the best way to retrieve neighbors is combining distances. For this, two steps are taken into account. First, the Q distance is calculated comparing the Q_{new} with each Q_i stored in the CB (equation 6.1). Using this distance, the first 30 nearest neighbors are selected. Next, the Principal Components distances (PCs) are calculated (equation 6.2). These descriptors are scaled using the variance captured by each PC by comparing each PC ($p=1, 2$ and 3) from the new problem with each data stored in the CB. Here, only the first five nearest neighbors are inspected. According to the largest number of repetitions, the event will be marked.

6.3.2 Results

The result of the methodology is applied to the COST/IWA simulation Benchmark. First, the data is organized in order to apply the methodology proposed. Second, the statistical model results are given. Finally, the results of the CBR methodologies are presented.

Organizing data

The data from the benchmark simulation is separated into two sets, as was explained in Section 2.1.1:

1. First data set: This set included 364 days for identifying and training the statistical models and CBR approach. This historical data set is divided in accordance with the knowledge about the events of the process per day. In this manner, NOC days are exclusively selected in order to build the statistical models. Days with events considered as Abnormal Operation Conditions (AOC) are stored in another set. Table 6.5 shows the distribution of days in this data set.

	Days	%
NOC (to build statistical models)	261	71.7
AOC (to validate models)	103	28.3
Total days first data set	364	100

Table 6.5: Division of the first data set

2. Second data set: This data set contains 182 days and is labeled as evaluation data set. The methodology proposed in this work will be validate using this set. Different types of events with Abnormal Operation Conditions (AOC) are imposed: bulking events, inhibition-toxicity events, high flow rate events and finally, nitrate sensor faults.

Statistical models

PCA model: The NOC data are organized in a matrix $X1 \in R^{261 \times 9}$ corresponding to 261 days, and scaled. Three principal components are selected in order to represent the model with 87% of the variance captured. The remainder of the data was organized and scaled using the same methodology.

MPCA model: The NOC data are stored in a $3D$ array X with a size (261 x 9 x 96) where 261 are the number of days, 9 process variables with 96 samples. The data was unfolded in a $2D$ matrix $X2 \in R^{261 \times 864}$, as is shown in Figure 3.8. Immediately afterwards, the data were scaled calculating one standard deviation and one mean for each sample per day. Commonly, this data scaling is called "auto scaling" (see more Section 5.2 Step 2). Three principal components are selected in order to represent the model with 71.26% of the variance captured. The rest of the data were stored, unfolded and scaled using the same methodology.

DPCA model: The NOC data are organized in a matrix $X1 \in R^{261 \times 9}$. To find the best stacking, several iterations are performed. In this manner, three stacking methods were selected: 10, 20 and 30 observations. Therefore, three models with three principal components are calculated. The variances captured are 83.88%, 79.87% and 76.39% respectively. The rest of the data were organized using the same stacking principle and scaled using the same methodology.

Building Case Base (CB)

From the first data set the Case Base (CB) is built. Days with AOC (103 days) are projected onto the statistical models. Immediately afterwards, the descriptors are composed in accordance with Section 6.2. The same amount of NOC days are selected to complete the CB. These NOC have been used to build the statistical models. In this manner, the CB includes 103 NOC days and 103 AOC days. Each case stored in the CB has a class number, which allows diagnosing new cases. The class numbers assigned to each event are shown in Table 6.6. In addition, seven AOC days stored in the CB are present in more than one event; their names are a combination of their class numbers.

Event	Class	Amount of days	%
NOC	1	103	50
Bulking	2	39	18.92
Inhibition/toxicity	3	22	10.6
High flow rate	4	34	16.5
Nitrate sensor fault	5	1	0.49
Bulking/flow	2.4	5	2.43
Inhi-toxi/flow	3.4	2	0.98
Total		206	100

Table 6.6: Assignment of class numbers for each event

CBR

Each case stored in the CB is compared with the new cases. In this work, two sets are tested: the remaining data set from the statistical models not used in the CB (only NOC), and 182 days evaluation data set to validate the methodology. The events present in the evaluation data set are event classes 1,2, 3 and 4; therefore, nitrate sensor faults did not occur.

In this work, the goal is to find the best methodology for diagnosing the events presented. Three methodologies are evaluated (Table 6.7).

Methodology 1	Methodology 2	Methodology 3
PCA	MPCA	DPCA
+	+	+
CBR	CBR	CBR

Table 6.7: Names for each methodology developed

Remaining data set: Sensitivity is the number of true diagnoses done using a specific methodology. The sensitivity for methodology 1 is 89%, and 100% for methodology 2. In

methodology 3, three sensitivity levels are used (stacking 10, 20 and in 30), resulting on 67%, 57% and 63%, respectively.

Evaluation data set: Figure 6.7 shows the final diagnosis per day for methodologies 1 and 2. In methodology 1, the sensitivity level is 60.44%. Using this methodology, only 2 events are diagnosed. Methodology 2, with a sensitivity level of 70%, detects three events. The inhibition/toxicity event is not detected.

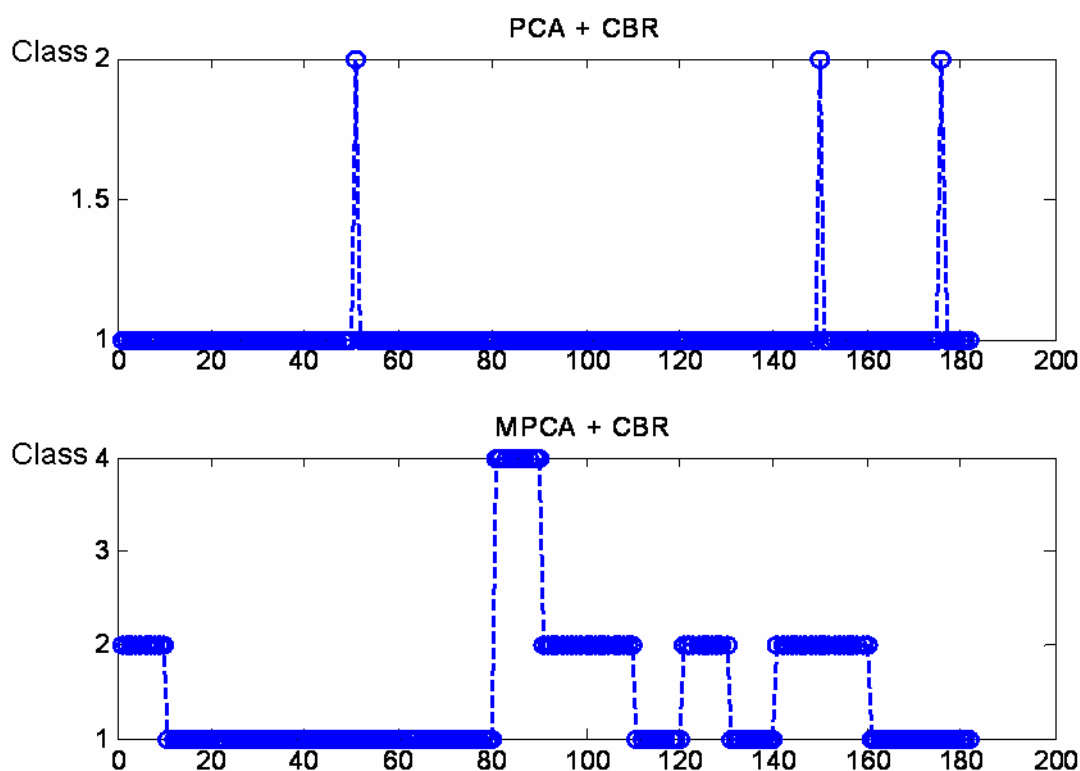


Figure 6.7: Diagnosis using methodologies 1 and 2 for the evaluation data set

Figure 6.8 shows the final diagnosis per sample for three different stackings (10, 20 and 30). Few days with 10 observations are diagnosed in events 2, 3 and 4. Using 20 or 30 observations result in an increment in diagnosis of AOC events compared with methodology 1. The sensitivity levels for each stacking are 57.86%, 62% and 67% respectively. This methodology successfully detects the four events present in this evaluation set.

The results of this section were presented as poster in 3rd International IWA Conference on Automation in Water Quality Monitoring (AutMoNet2007), September 5-7, 2007, Ghent, Belgium (Ruiz, Rosen and Colomer 2006).

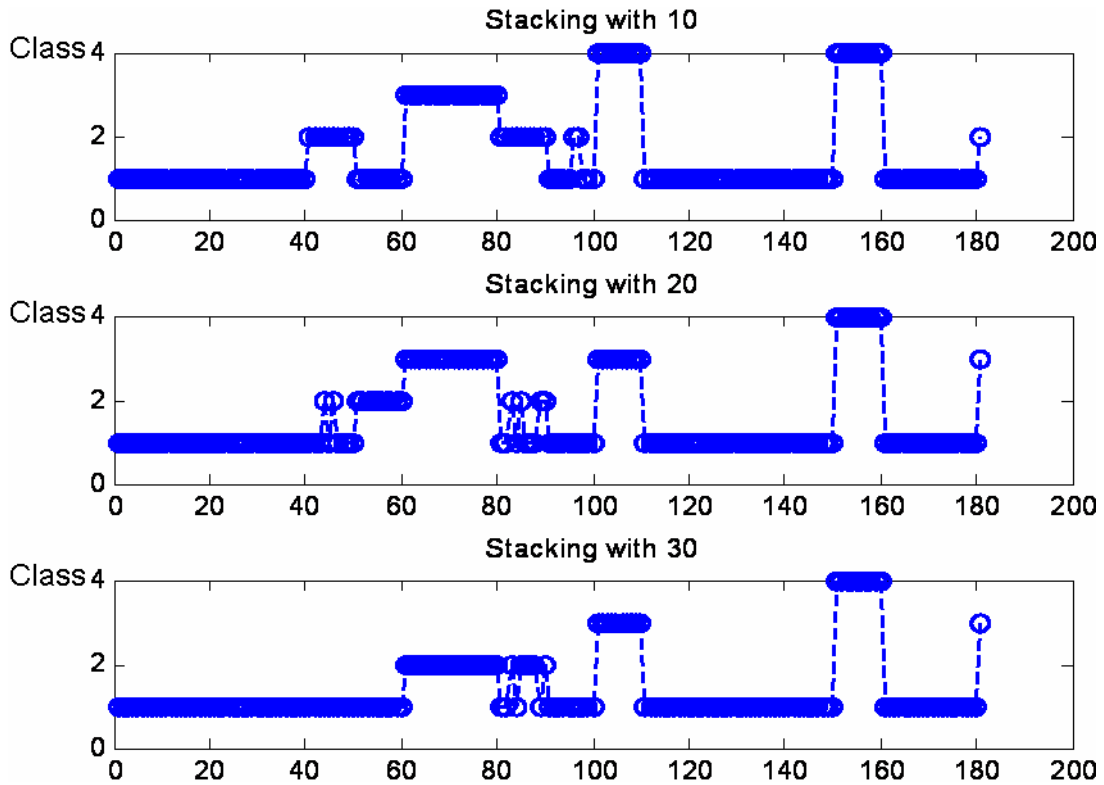


Figure 6.8: Diagnosis using methodology 3 for the evaluation data set

6.4 Case base maintenance and updating

For this implementation, based on the methodology described in Section 6.1, the data used comes from the SBR Pilot Plant from the LEQUIA group (see Section 2.2.2). The particularity of this data set is the configuration variation of process operating conditions as a result of this group's research. Each batch of the data set has 4 process variables with 424 samples per variable. In this manner, two data sets are used:

1. Set 1: this set corresponds to 98 batches with two reaction stages.
2. Set 2: this set corresponds to 227 batches with three reaction stages.

The goal of this implementation is to apply the *DROP4* and *IB3* algorithms achieving the best abilities of CBR: maintenance and updating. The case base removes similar cases which can produce errors at the moment of retrieving a solution and learns from new situations. The implementation is divided in three parts, clearly distinguishable as follows:

1. Building the statistical model
2. Rearranging of the data
3. Applying the CBR

The step of building the statistical models is repeated because some graphical improvements were applied in order to simplify the monitoring process for the operators. Figure 6.9 shows the implementation developed.

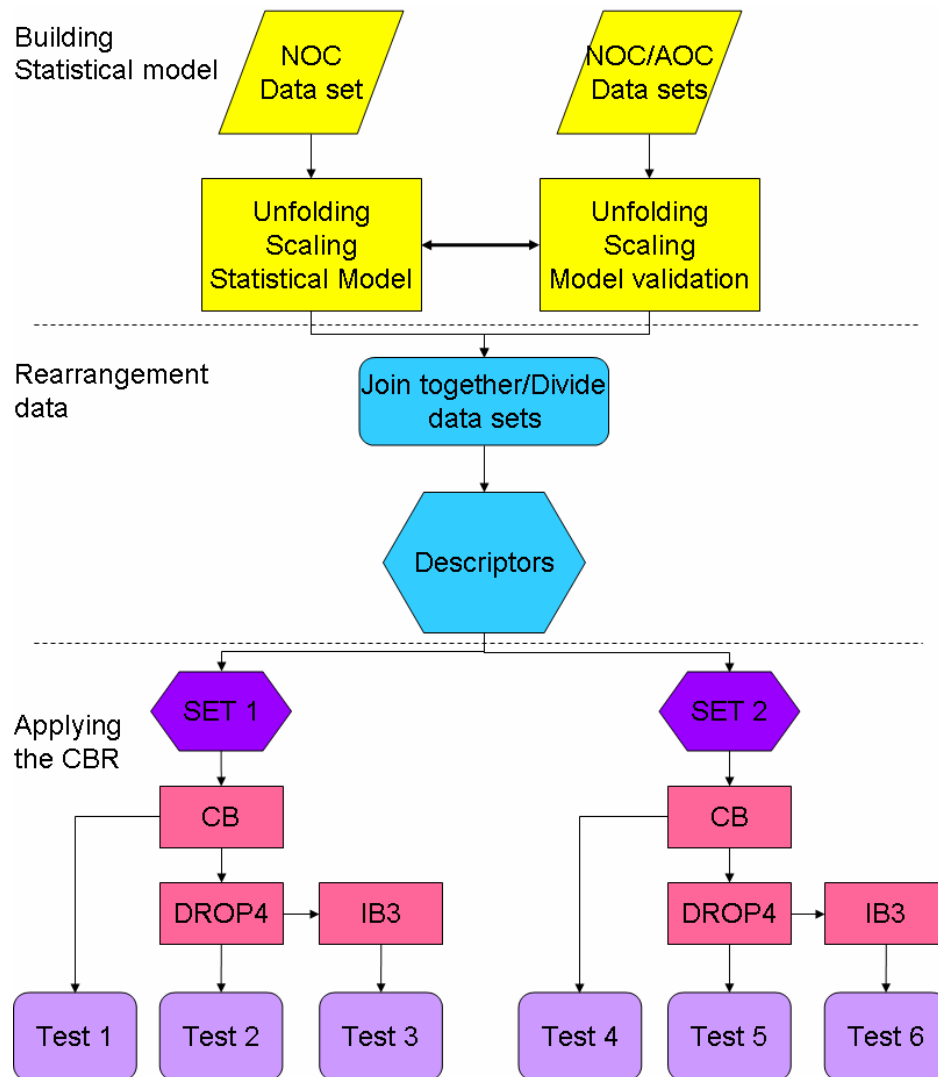


Figure 6.9: Case Base maintenance applied to pilot-scale SBR

6.4.1 Building the statistical model

In order to understand the operation of the process, it is necessary to build a statistical model using batches with normal operation (Normal Operation Condition - NOC) according to previous knowledge from the expert. Thus, when a new batch is projected on the model, its principal components and distances Q and T^2 determine the operation of that new batch.

The original data is divided in two three-dimensional (3D) arrays; one of them contains part of the NOC batches, the other one contains the rest of the NOC batches and the whole set of Abnormal Operation Condition (AOC) batches.

For building and validating the model, the 3D arrays must be unfolded, and thus the data becomes collected in a 2D matrix. According to the study performed in Section 5.1, the data must be scaled as well. The model is built using just NOC batches. The validation of the model is performed using NOC and AOC batches.

In this way, a new representation of every batch is obtained: A batch previously defined by process variables (i.e four variables: ORP, DO, pH, Temperature) directly acquired from the process (300 samples per variable) becomes defined by 5 principal components (PC's) and the distance Q .

6.4.2 Rearrangement of data

Once all batches are projected on the new space, the indices that represent each batch are obtained. These indices are called descriptors in CBR. In accordance with testing previously performed, each case can be described by PC's and Q -statistic (see section 6.2.6). In turn, these descriptors must be weighed using the eigenvalue for to each PC, as was presented in step 4 in Section 6.2.4. Now, all descriptors have been balanced in accordance with their percentage variance participation.

When the descriptors have been weighed, a group are selected in order to build the case base which contains batches or cases with NOC and AOC. The remaining descriptors are used for testing the specificity and sensitivity of the methodology.

6.4.3 CBR application

In the same way as the MPCA was tested and validated (Section 5.2.1), this CBR application is performed in two stages: creation of the case base and validation. The creation of the case base is based on the criteria that the best case base is that one which has an equal number of cases of NOC and AOC (see Section 6.2.6). These groups of descriptors are part of the training block (Figure 6.1). As distinguished in Section 6.2.6, the case base is initiated by randomly choosing cases from both groups (NOC and AOC). Once the first case base is built, *DROP4* and *IB3* algorithms are applied. The maintenance of the case base is done by means of the *DROP4* algorithm. The updating of the case base is made by means of the *IB3* algorithm. In consequence, three case bases are built and used to compare the results:

1. Full Case base: Contains all batches selected randomly with an equal number of NOC and AOC cases and any treatment is applied to this case base.

2. *DROP4* Case base: A clean-up algorithm is applied to the initial case base (full case base). *DROP4* (Decremental Reduction Optimization Procedure algorithm) eliminates the redundant information and the noise generated by neighbor classes.
3. *DROP4 + IB3* Case base: A new algorithm is applied to the *DROP4* case base. *IB3* (Instance-Based learning algorithm) makes of a classification each new case by looking for the similarities between the stored cases and the new case.

Descriptors for the evaluation block (Figure 6.1) are used for testing the effectiveness and sensibility of every case base (Full, *DROP4* and *DROP4+IB3*). When a new batch is presented for diagnosis, the CBR cycle proposes retrieving batches from the case base with similar descriptors. The retrieval step is performed in two stages; firstly, the descriptor Q is used for selecting the 30 batches closest to the new batch. Finally, from these 30 batches, five are selected using the Principal Component values. From these five batches a diagnosis is performed using voting (Chapter 6 step 4). If the diagnosis is correct, the procedure ends, but if it is incorrect, the classification is revised and retained into the case base by means of the *IB3* algorithm (see Figure 6.9).

As the configuration variation of process operating conditions change, two data set are used, and as a consequence in total six tests for this implementation are generated. Table 6.8 summarizes how the respective tests were labeled.

Case base	Set 1	Set 2
Full	test1	test4
<i>DROP4</i>	test2	test5
<i>DROP4 + IB3</i>	test3	test6

Table 6.8: Names for each test developed

6.4.4 Results

Determining parameters

Before applying the methodology, some parameters must be defined. In the case of the statistical model, the methodology developed in Section 5.2.2 is applied in order to determine the number of principal components. The scaling process is selected according to expert knowledge. In the case of specification for CBR parameters, the size of neighbors, number of friends and enemies in the *DROP4* algorithm, and the size of neighbors of cases with the same class in the *IB3* algorithm are refined by means of several iterations.

Other important parameters necessary in this implementation are the different anomalies and operation changes registered in the process. This provides extra information in

relation to an in depth investigation from the experts and the quality variables or off-line variables (Carbon (C), Ammonium (NH_4^+), Nitrogen dioxide (NOx) and Phosphorus (P)). However, for the case of the data set used in this implementation (LEQUIA data set), the extra information comes as shown in Figure 6.10.

Data		INFLOW				OUTFLOW				REACTOR				Global valoration		
Sample	Days	ppm O	ppm N	ppm N	ppm P	ppm O	ppm N	ppm N	ppm P	ppm	work plan	Removal				
		DQO _E	[N-N _{TOTAL}] _E	[N-NH ₄ ⁺] _E	[P-PO ₄ ³⁻] _E	DQO _S	[N-NH ₄ ⁺] _S	[N-NO ₂] _S	[P-PO ₄ ³⁻] _S	SST _R		C	NH ₄ ⁺	NO ₂ ⁻	P	
03-03-05	1	468	77,09	63,2	6,71	114	15,6	1,35	2,41	2640	SBR28027	⊕	×	✓	×	✖
04-03-05	2					48	39,1	3,62	1,34	2210	SBR280210	✓	×	✓	⊕	✖
05-03-05	3	464	87,00	76,9	4,53	73	36,9	1,5	1,62		SBR280213	✓	×	✓	⊕	✖
07-03-05	5	355	82,40	67,7	4,17	74	36,6	3,62	0,41	2455	SBR280119	✓	×	✓	⊕	✖
09-03-05	7		80,49	69,4	6,82		26	3,67	1,7		SBR280225	✓	×	✓	⊕	⊕
11-03-05	9	476	84,70	72	6,41						SBR280231	✓	×	✓	⊕	⊕
14-03-05	12	350	83,76	68,9	4,63	53	0,9	9,17	1,06		SBR280240	✓	✓	⊕	⊕	✓
16-03-05	14	513	80,09	60,9	4,44	61	1	6,59	1,296		SBR280246	✓	✓	⊕	⊕	✓

a)

Data	work plan	Removal				Global valoration	
Sample		C	NH ₄ ⁺	NO ₂ ⁻	P		
03-03-05	SBR28027	1	3	2	3	3	18
04-03-05	SBR280210	2	3	2	1	3	12
05-03-05	SBR280213	2	3	2	1	3	12
07-03-05	SBR280119	2	3	2	1	3	12
09-03-05	SBR280225	2	3	2	1	1	12
11-03-05	SBR280231	2	3	2	1	1	12
14-03-05	SBR280240	2	2	1	1	2	4
16-03-05	SBR280246	2	2	1	1	2	4

b)

Figure 6.10: Off-line variables a) Table used for the biological experts b) Table used for the monitoring experts

In Figure 6.10, three options of quality per variable are taken into consideration: good, regular and bad. Every variable has been associated with a number (Figure 6.10b)).

Good = 1

Regular = 2

Bad = 3

In this way, nine possible combinations are presented. These nine combinations are used as subclasses in order to check the specific situation of one operation.

Application of MPCA

Two statistical models are built, and each one corresponds to a data set from the LEQUIA SBR pilot plant.

- Model 1: The data is stored in a 3D array (78 x 4 x 424) where 78 batches are the 80% from data set number 1, 4 is the number of process variables, and 424 are the number of samples per variable. The contribution of the three principal components represents a variance of 65.49%. The loading plots are shown in Figure 6.11. The stages of the process in every variable are also shown in the figure along with their phases (fill, anaerobic, aerobic1, fill2, anoxic 1 and aerobic 2).
- Model 2: The data is stored in a 3D array (181 x 4 x 424) where 181 batches are the 80% from data set number 2, 4 is the number of process variables, and 424 is the number of samples per variable. The contribution of the three principal components represents a variance of 69.42%. The loading plots are shown in Figure 6.12. The stages of the process in every variable are also shown in the figure along with their phases (fill, anaerobic, aerobic1, fill2, anoxic 1, aerobic 2, fill3, anoxic 2 and aerobic 3).

Figure 6.13 chronologically shows the displacement of the batches in models 1 and 2.

In Figures 6.14 and 6.15, the 3D representation for model 1 can be checked. These figures contain lines which show standard deviations 1 and 2. Clicking over any point, the name of the batch is displayed (not shown). The three principal components of every batch are projected there.

Rearrangement data and Descriptors generation

When all the information has been projected in the new space, the principal components are obtained together with Q -statistic. These descriptors represent all batches. As was indicated in Section 6.4.2, the principal components are multiplied by their eigenvalues. In the same manner, the value of the Q -statistic is divided by the value of the Q limit, according to Equation 6.4. The uniformity of the batches into the database is assured by scaling the distance of the Q statistic.

$$Q_{descriptor} = Q_{distance}/Q_{lim} \quad (6.4)$$

In this way, the Q -statistic farthest batches are eliminated from the case base, maximizing margin between sensitivity and specificity.

Case base building

Once the entire information from the statistical model step has been scaled, it is joined together in the same group. Several options are possible to build the case base. It is possible to start with an empty base which will grow by learning from new situations or a base containing some cases. In relation to this last option, 80% of the data per set is selected. The selection of these sets are random, that is to say, any criteria can be used

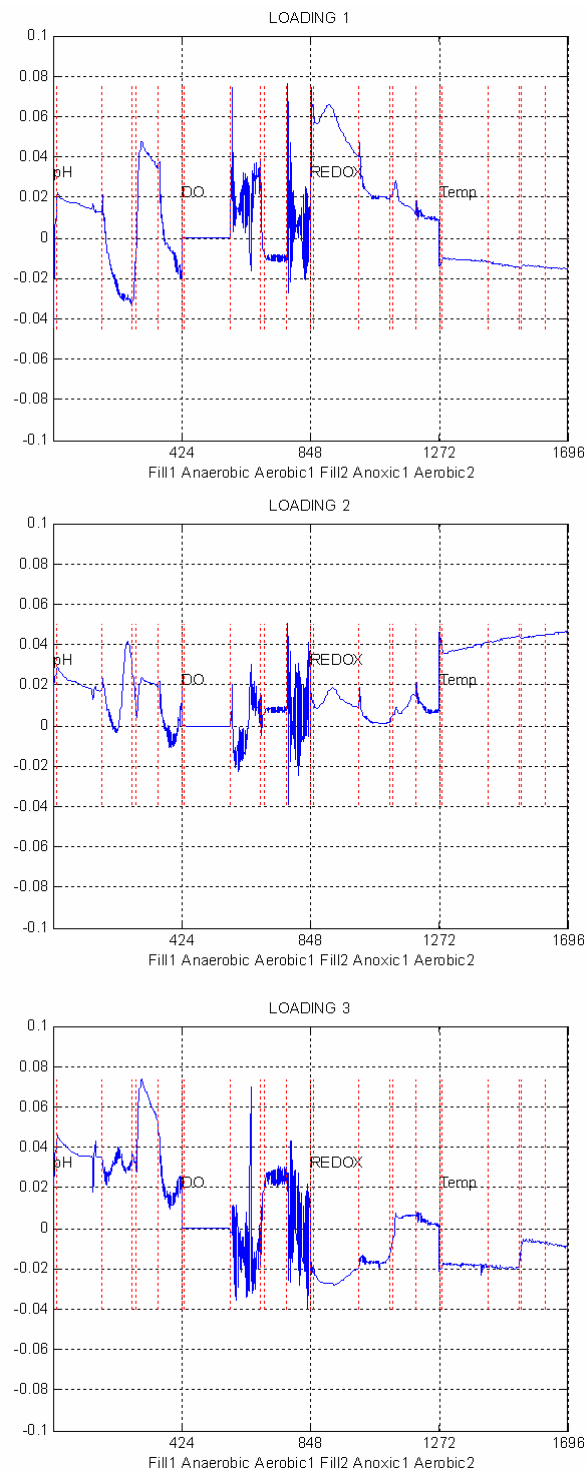


Figure 6.11: Loading plots for model 1 which corresponds to two reaction stages

for selecting the batches. This procedure starts with the TRAINING block in Figure 6.9. This block makes reference to the DB calibration. The remaining 20% of the data per set is used to evaluate the CBR methodology; this procedure starts with the EVALUATION

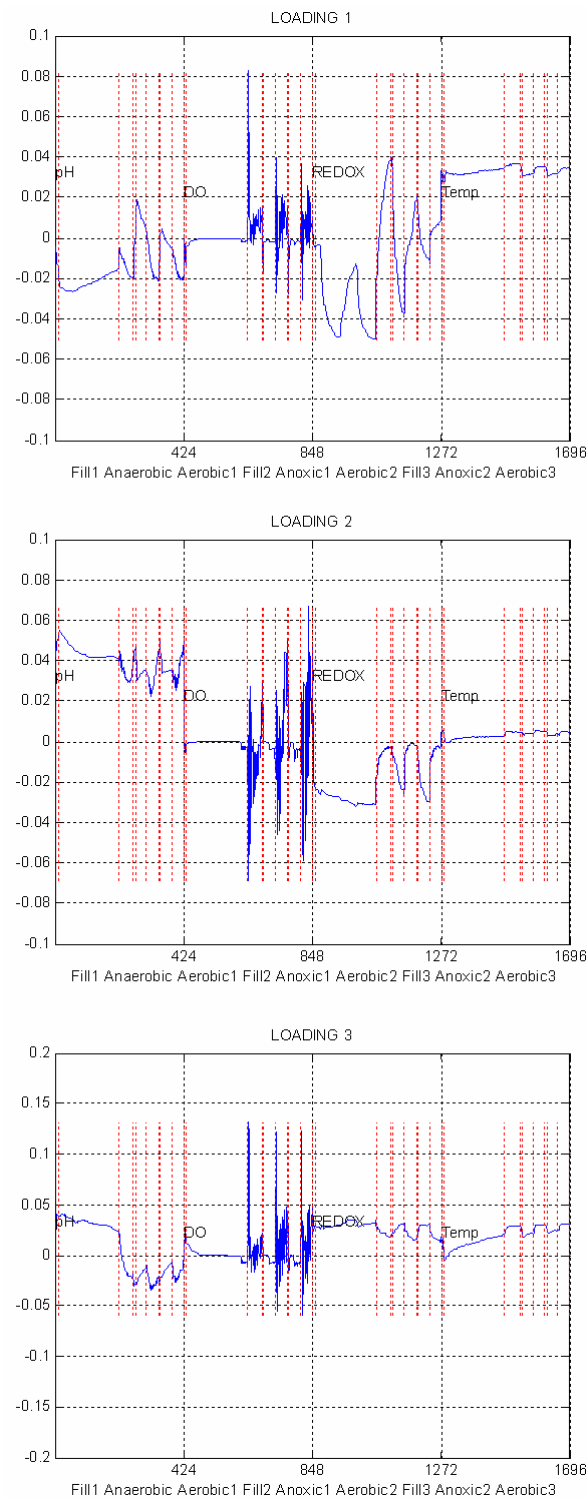


Figure 6.12: Loading plots for model 2, which corresponds to three reaction stages

block in Figure 6.9. Each of these blocks are explained in next section.

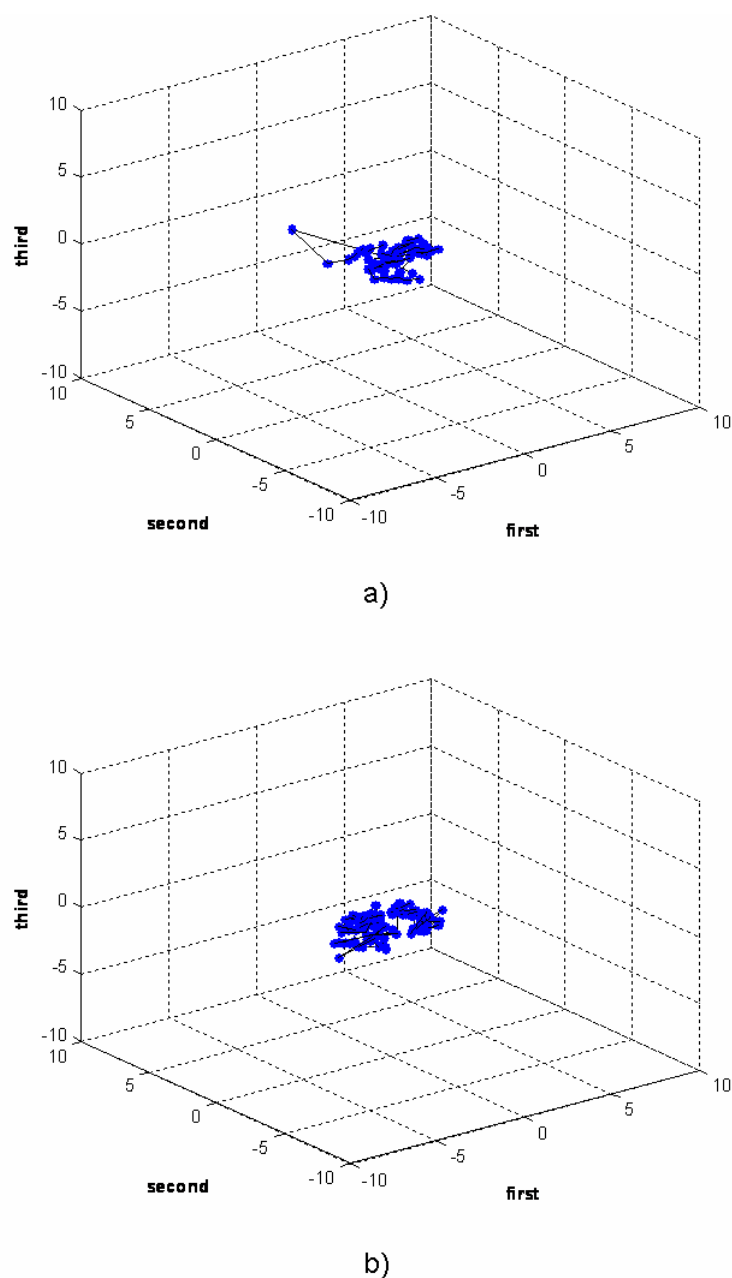


Figure 6.13: Score plots for models 1 and 2

Applying *DROP4* and *IB3*

To understand the results, it is necessary to bear in mind that two data sets are used. The first one (set 1) has two reaction stages. The second one (set 2) has three reaction stages. Additionally, each data set is tested using three different case bases (Full, *DROP4*, and finally *DROP4 + IB3*). According to Table 6.8, the data sets are compared with the three DBs in order to check which case base is the best for diagnosing the normal and abnormal operation of the process. Also, these tests will provide extra information indispensable for a proper correction of the process when a fault or event is presented. In this

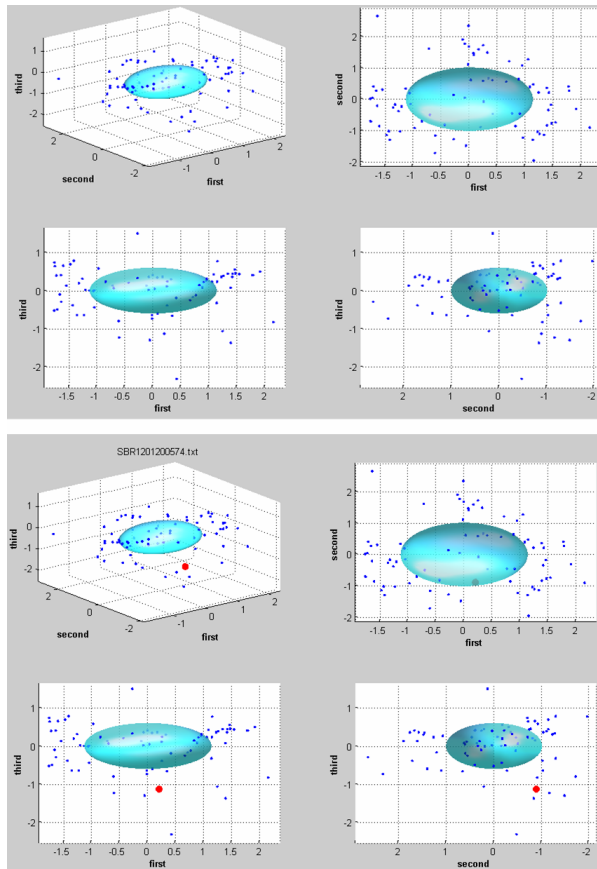


Figure 6.14: Three dimensional representation of model 1 for one standard deviation

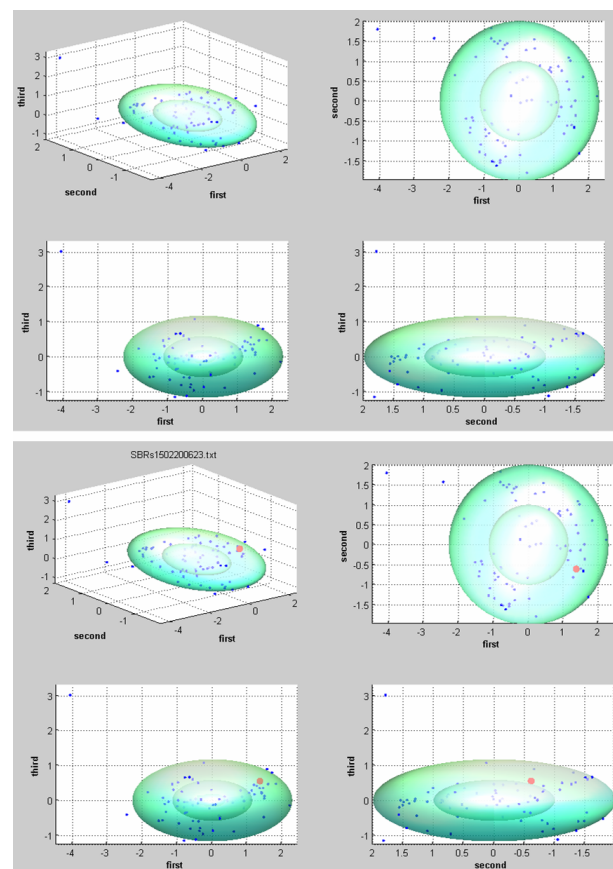


Figure 6.15: Three dimension representation of model 1 for two standard deviations

way, several simulations are performed in order to test what happens when the *DROP4* and *IB3* algorithms are used.

In Figure 6.16, the diagnosis of the operation condition improves when new cases are added to the case base because the correct diagnosis increases as new cases are presented to the *IB3*. This means that when a new case is stored in the case base, the entire case base is reevaluated in order to determine if this new case base is better for diagnosis. However, in Example 4, it is shown that the increase has a limit. As was explained in Section 4.3 referring to the *DROP* algorithm, this algorithm is applied only once, to the full case base and for this reason it is possible to see that the line is constant during the entire simulation. Example 4 improves because group 4 may contain repetitive cases of the previous groups.

In Figure 6.17, the overall percentage seems to be worse than the previous ones shown in Figure 6.16 due to the greater area of intersection between the different operation conditions; however, its overall tendency is satisfactory. In the same way as for the previous results, the *DROP4* case base shows similar behavior. Results from this section are in preparation for publication in *Water Science and Technology* (Ruiz, Sin, Colprim and Colomer 2008).

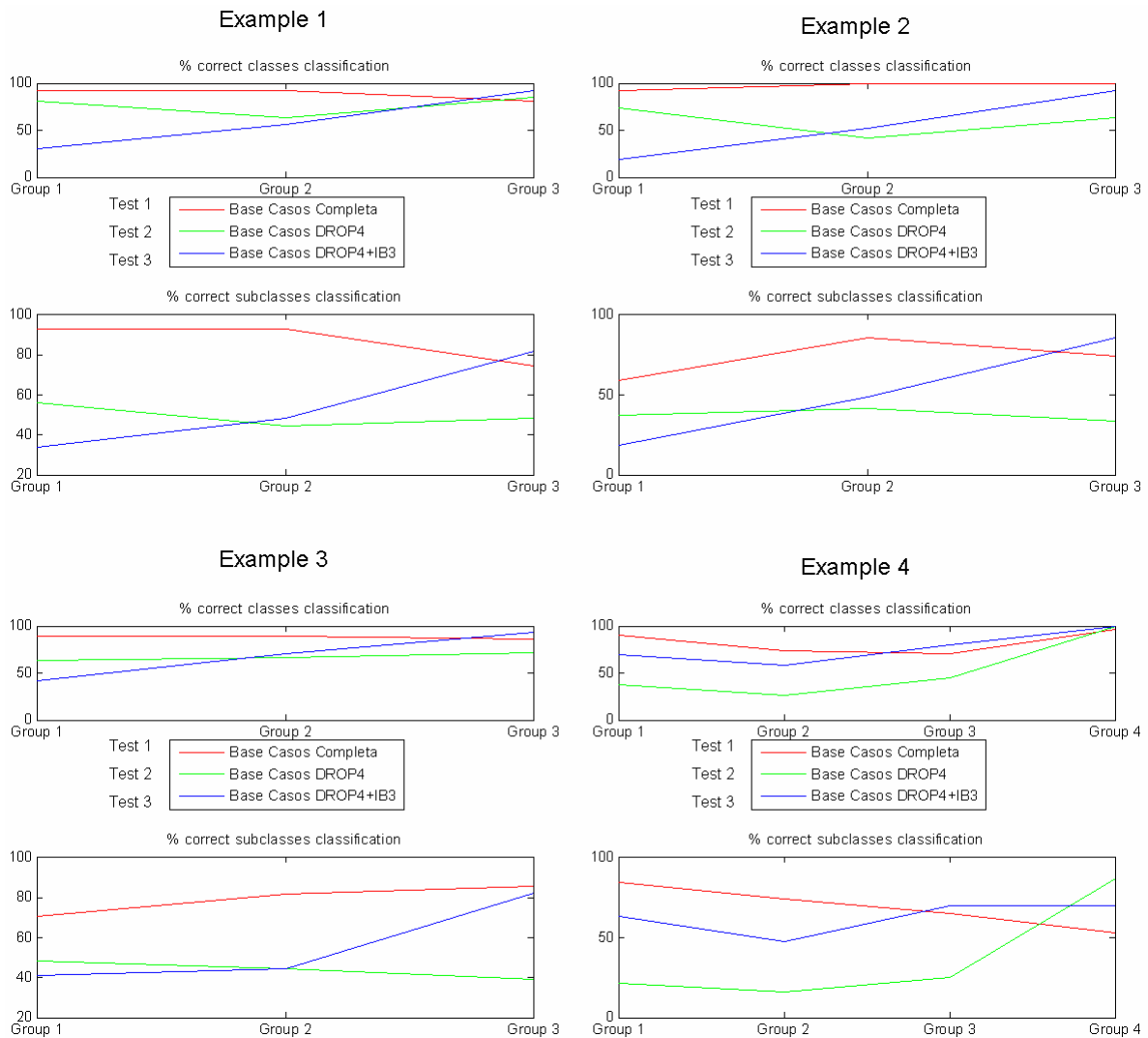


Figure 6.16: Learning evolution for tests 1, 2, and 3

6.5 Analysis and Conclusions

In this chapter, a new methodology for the diagnosis of a WWTP using statistical models and CBR is proposed and several strategies based on this methodology are studied using several data sets: data from the BIOMATH pilot plant were used to refine descriptors and distances, and data sets from the LEQUIA pilot plant were used for implementation of the *DROP4* and *IB3* algorithms.

To refine descriptors and distances, several simulations have been tested, combining different possibilities. In the first one, indices from the MPCA application (PC's, Q -statistic and T^2 -statistic) have been chosen as descriptors for case representation. The

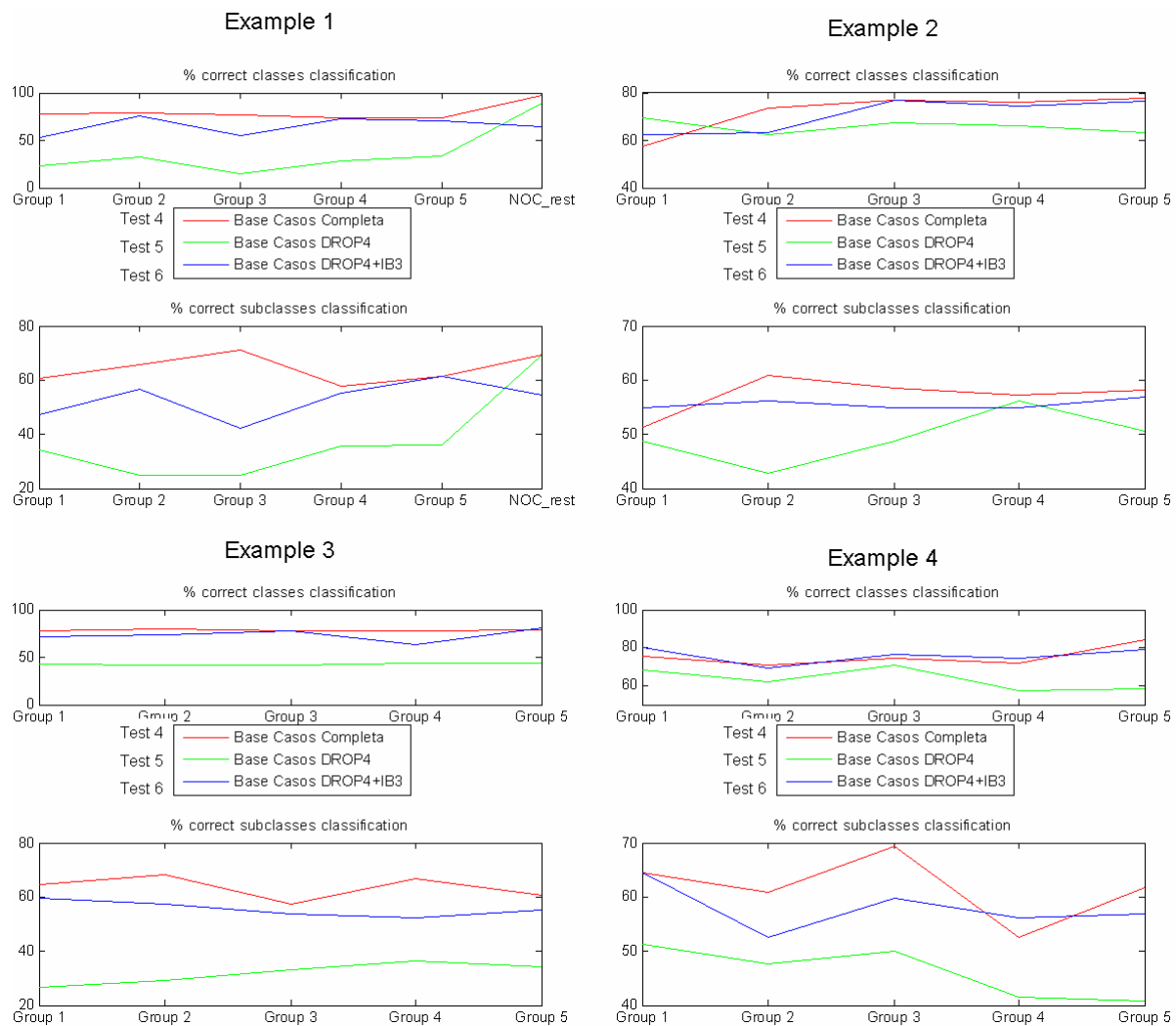


Figure 6.17: Learning evolution for tests 4, 5, and 6

combination of the Q statistic and the Principal Components in two steps offered the best results among the other possible combinations. This is because they contain complementary information. T^2 is not needed because it has the same information as the principal components. Another important selection is the initial case base. The case base with the same number of Normal and Abnormal Operation Conditions offered the best results. In addition, it is proven (in this specific case) that the combination of MPCA and CBR offer better results than using only MPCA for fault detection.

From application of the methodology to the COST/IWA simulation benchmark data set, the best solution is using MPCA + CBR. However, DPCA + CBR method gives good results, in this way, future work will be focused in stacking decision. In addition, redundance situations will be deleted, and the learning techniques will be used to obtain a better diagnosis and reconfiguration.

For case base maintenance and updating, it was necessary to develop two blocks at the same time: training and evaluation. In each block, it is necessary to take several actions. First, an initial case base has been selected. Then, the redundancy cases are eliminated from the case base by means of *DROP4* and supervised for learning by means of *IB3*. The processes of cleaning and learning were performed by means of the knowledge obtained from application of the algorithms in Section 4.3.4 and Section 4.4.3. When a new case is incorrectly diagnosed, this is added automatically into the case base. This is the best CBR characteristic, the ability to learn about errors and the ability to use the knowledge from past situations in order to bring solutions to new problems.

Chapter 7

Conclusions and future work

7.1 Conclusions

The methodology developed through this doctoral thesis based on situation assessment for monitoring, fault detection and diagnosis of WWTP's, specially to SBR processes; using MPCA and CBR allows to determine the operation condition of each case. In general the methodology uses the MPCA models as a preprocessing tool for the SBR system. The dynamic information of process variables measured is compacted in a suitable form to be used for the definition of cases; in others words the descriptors are obtained. The MPCA models are implemented in an interface for monitoring (on-line and off-line) and detecting operational conditions of a SBR process. CBR can be seen as a complement to statistical models that improves the fault detection task towards a more useful event diagnosis system when an initial CB is built. Afterwards, *DROP4* and *IB3* are applied in order to achieve the maintenance and updating of the initial CB. When similar cases are retrieved, the diagnosis of the case is determined by means of a voting system reusing the information stored in the selected situation. New cases are revised and retained to update the CB.

In Chapter 2, description of the data sets were explained: the SBR Semi-Industrial pilot plant data set from the LEQUIA group was used for the first study using MPCA and the SBR pilot plant data set from the BIOMATH group was used for in depth analysis of MPCA and the effects of the scaling process of the data. This data set was also used for the first test of selection of descriptors and distance in order to apply CBR. Finally, SBR pilot plant data sets from the LEQUIA group were used for the application of maintaining and updating algorithms in CBR. In addition data from the COST/IWA BENCHMARK simulation were used in order to assess situations by means of CBR in addition to PCA, DPCA and MPCA. In this point, it is important to emphasize the difference between SBR and COST/IWA BENCHMARK simulation which the methodology gets a global application in WWTPs. SBR is a batch process, while COST/IWA is the simulation of a continuous process. The characteristics of the SBR process can be attributed to a) the clarification that occurs in the same reactor, b) biological process takes place in a cyclic way, c) a portion of the treated water is replaced by untreated wastewater for each cycle, so the SBR process is distinguished from other continuous flow type activated sludge systems, and d) influent and effluent flows are uncoupled by time sequencing. The SBR process has a high correlation structure between variables as characterized by the

covariance matrix. Furthermore, it is highly nonlinear, time-varying and subject to disturbances with a large amount of data collected.

Main concepts about the MSPC and the CBR approaches were provided in two chapters. In Chapter 3, special emphasis in monitoring for batch process was given. In Chapter 4, detailed information about DROP and IB were presented making easy that anyone develops these algorithms in any platform.

In Chapter 5, the work started with a first MPCA implementation in a semi-industrial SBR pilot plant (LEQUIA) which without previous knowledge different groups of batches were found. Using the MPCA and the knowledge from the experts, five types of events are detected, four abnormal and one normal operation condition: Electrical Faults (EF), Variations in the Composition (VC), Equipment Defects (ED), Influent Load Change (ILC) and Normal Operation Condition (NOC). These tools complement each other very well. Both tools contribute to a general knowledge from the state of process. In this way, Multiway Principal Component Analysis (MPCA) was demonstrated to be a powerful data tool for compression and information extraction, permitting detection of linear combinations of variables that describe major trends in a data set. However, this technique did not detect all the batches with AOC. As a consequence, the work using the SBR pilot plant from BIOMATH was proposed. In this research, several models were developed in order to find the best methodology to apply MPCA to SBR WasteWater Treatment Plants. From this work, on-line and off-line monitoring can be used in parallel to detect faults. Likewise, scale is essential for a correct and precise methodology where group scaling presents a small difference when compared with auto scaling. However, this small difference does not allow a strong recommendation for one or the other. Based on the results, the scaling decision should be related to the objectives to the process operation. In this thesis, the appropriate scale approach was GS. Finally, a module for a real application at semi-industrial SBR pilot plant was developed.

In Chapter 6, CBR was added to the MPCA methodology. In this union several doubts emerged, for example: The process variables are available, but could be used the results from MPCA methodology, in this way, which descriptors could be used to describe a case? Which could be the best way to build a case-base? Which should be the best procedure to retrieve cases from the data-base? How many cases will be necessary to retrieve? How could be possible to take the decision about a new case diagnosis? How could be cleaned (maintenance) the data base from redundancy information? How update the case base? To solve all of these questions, the work was divided in two parts. In the first part, the descriptors, case base and retrieve questions were solved. In the second part, maintenance and updating questions were developed.

Descriptors, case base and distance refining was performed using historical data from lab-scale BIOMATH SBR plant because it had big amount of data in which several operations conditions were determined by means of MPCA and LAMDA algorithm (see Appendix A). In spite of the fact that the process variables could be used directly as descriptors of the cases, the values of principal components, Q-statistic and T2-statistic were used as descriptors because they were correlated by means of lineal combination increas-

ing the relation and eliminating the possible sensor noise. Once, the descriptors had been selected, several combinations were performed (not all showed). Next, the case bases were built. The cases used for this purpose should perfectly be well-known. Quality and reliability of the data was an important aspect. If these aspects were insufficient, probably the solution would be incorrect. Three results were obtained: First, Case base should started with same amount of normal and abnormal condition guarantying all possible situations. Second, the best way to describe cases was using Q-statistic and PC. Third, performing two steps retrieval was the best way to get the most similar cases where Q-statistic behaved as filter characterizing the similitude with the nature of the new case and PC looked for the accurate space location of new cases. These results were used in a short study of the methodology response over continuous processes. The historical data set was the COST/IWA simulation benchmark. The dynamic of the benchmark process is different than SBR, then other statistical extensions like dynamic PCA and PCA were checked together with MPCA. The results showed the methodology can satisfactorily work in this kind of process. The methodology could be used as a final application in wastewater systems. However new questions were presented, why redundancies cases from the data base should be removed? why updating should be implemented? The answers were: because, the case base could increase enormously making complex the diagnosis about the situation due to when a new case was presented, it was compared with each case stored into the case base. Similarly, when a new situation was presented, it was incorrectly diagnosed.

Case base maintenance and updating were performed using historical data from lab-scale LEQUIA SBR plant. This historical data set had two operational conditions. The first operational condition (set 1) had two reaction stages and the second (set 2) had three operational stages. Visual load windows were improved to simplify the monitoring process. Maintenance and updating CBR abilities were obtained by means of decremental reduction optimization procedure (DROP) and instance-based learning (IB) algorithms.

Important decisions were resolved when the next questions were analyzed: how much redundancy information should be removed? And how many enemies and neighbors could be selected? In addition, extra information about anomalies and operational changes were introduced. Several proofs were developed to test the sensitivity and specificity of the best CBR. In this way, three tests were developed: (i) using a full data base generated randomly. (ii) employing DROP to delete redundant cases. (iii) used a full implementation of the CBR abilities by means of DROP + IB algorithms. The results successfully showed that, after two iterations, a full implementation of CBR surpassed the other two, guaranteeing the learning procedure saving computational cost.

In conclusion, the methodology developed in this doctoral thesis can be used to monitoring, fault detection and diagnosis for WWTPs, with some specific changes mentioned in future work. For instance, the case base can be improve storing more information about the process, causes, solutions and comments.

7.2 Future work

This thesis is intended to contribute to the development of new techniques for diagnosis in Wastewater Treatment Plants. Nevertheless, this work is only a particular approach to the wide field of situation assessment. Future research is certainly needed to obtain a robust, automatic and general tool. The immediate future work should be focused on the following subjects:

7.2.1 MSPC

Implementing multi-phase algorithm

Recently, multiples multi-stage models have been proposed to performance monitoring in batch process increasing the fault knowledge and adjustment to the process nature. The new methodology is called Multi-Phase principal component analysis (MPPCA). In this way, MPPCA could be implemented for detection of phases during the batch process.

Missing data

The historical data lengths were worked without take into account the missing data problem. However, in real systems missing measurements are a common problem of the processes. It can be due to sensor failures, sensor routine maintenance, samples not collected at the required times, data discarded by gross measurement errors, or sensors with different sampling periods. This problem could be solved applying methods to estimate the possible scores values or using MPPCA algorithm.

7.2.2 CBR

Optimal neighbors number

It is necessary to find the optimal number of neighbors k . In this thesis, this number was found by means of several iterations, ranging from 5 to 100, and choosing the number of iterations that procures the best performance. However, it is necessary to perform in depth research to extract the optimum number for k .

Initial Case Base

The initial case base was made by randomly selecting and conserving an equal number of NOC and AOC. Other modes to begin the case base can be explored in order to see if some improvements are obtained in the final results.

Refining of distance

Several combinations have been used in order to find the best option to calculate the distance between the new case and its neighbors. Options such as trigonometrical functions together with the indexes should be investigated.

Decision making rules

Actually, at the end of retrieval step, there are five nearest neighbors. The final decision to determine the operation of a new case is solved using a voting rule. The results are acceptable when the situation assessment is only distinguished in two classes of operation (normal and abnormal). However, this rule has problems when there are more classes, so it is important to elaborate a deeper research in this direction.

Giving more extra information

The case base can be programmed using more extra information. This will allow, for example, the gathering of important information about the process, possible causes for abnormal operation, possible solutions and complementarily comments using the off line variables N-NH and P-PO, among others.

To extrapolate this methodology to other process

This methodology has been applied only to SBR processes. Implementation in others batch processes should be proposed. If a modification in the Benchmark model *BSM1_LT* is made, the final methodology developed in this doctoral thesis could be applied also here.

To improve the on-line application

The thesis developed a previously available monitoring application. However, the application has not implemented the CBR approach. Future work could focus on the finalization of this approach applying it to a real case.

Improving the Benchmark application

The COST/IWA simulation benchmark evaluates different control strategies. At the same time, the monitoring, fault detection and diagnosis can be studied. The application in this work was superficial since sensitivity was not evaluated. Full implementation of a CBR methodology (retain, reuse and revise) should be focused as a future work.

Reconfiguration of actuators

Until this moment, the diagnosis can show the possible solutions when a fault occurs, but the operator has to reconfigure manually the process. The union of diagnosis and automatic control of actuators will provide a quick solution saving time and cost in WWTPs.

As a final recommendation, when the engineers know the process better methodologies, analysis and interpretation can be developed. As a consequence, more implication between the studied process and monitoring and control engineers must be considered.

Appendix A

LAMDA application

A.1 A short introduction to the LAMDA algorithm

LAMDA methodology (Learning Algorithm for Multivariate Data Analysis) is a classification strategy developed by Piera (1987). It was based on an original idea by Aguliar-Martin and Lopez (1982). LAMDA represents a system of classes by means of logical connection of the marginal information. In this manner, the global configuration of one object to a class is calculated starting from the marginal configuration of each attribute following a heuristical rule. The object will be a member of one class when it has the larger adequation scale.

When one object is not classified, it is confronted with one prototype or concept for each of the existent classes. This diffuse nature of prototypes creates imprecision when the concepts are formed. In this way, LAMDA is a conceptual methodology of classification.

The main feature of LAMDA is the diversity of problems that can be solved. This flexibility is due to properties of the supervised and nonsupervised learning, the employment of quantitative and qualitative attributes and sequential learning. The learning capacity depends on specific functions.

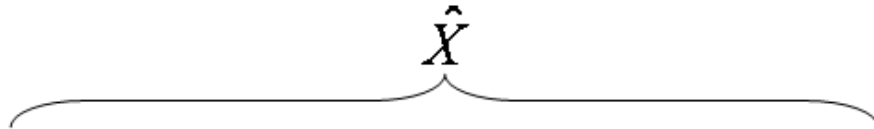
An application of LAMDA is SALSA. SALSA is a userfriendly interface developed by the Diagnosis Supervision and Control group (DISCO) of the Laboratory for Analysis, Architecture of Systems (LAAS-CNRS) (Toulouse-France) which was used in eXit and Disco (2004). SALSA provides an easy interface which has the capability to work in two stages:

- off-line stage: Design and construction of a classification system.
- on-line stage: Classification system used to determine the function and state of the process plant.

In this manner, it is possible to determinate the current status of the process providing more information about the operation in accordance with the methodology applied

for diagnosis. In conclusion, SALSA is an interface to LAMDA, which is a method that combines algorithms of numeric and symbolic classification using fuzzy logic and hybrid connectives (Aguliar-Martin and Lopez 1982). In this work, the classification process is described as follows:

Matrix \hat{X} is conformed by the LAMDA-descriptors, where each row characterizes one case. Each row is conformed to the principal components extracted in the MPCA model. The descriptors report the normal or abnormal operation of each batch (see Figure A.1).



Batch	X1	X2	X3	X4	X5	X6	X7	X8
1	-1,19040	0,40769	-0,16347	0,68946	-0,37174	0,23784	-0,53304	-0,00425
2	-0,60437	-0,46155	-0,06614	0,76171	-0,24992	0,01876	-0,54621	-0,45446
3	-0,37465	-1,41540	-0,03215	0,70066	-0,18338	0,03463	-0,30230	-0,39654
4	-0,16948	-0,22701	-0,00589	-0,07695	0,00882	-0,15824	-0,02901	-0,08701
5	0,43591	0,19642	-0,02061	-0,19565	0,07733	-0,30996	0,07951	0,35192
.
.
.
177	-0,52598	-0,95484	0,73638	-0,70182	1,06240	0,76620	0,09412	0,88625
178	-0,52763	-0,94670	0,73593	-0,70479	1,06320	0,77200	0,08568	0,90762
179	-0,52763	-0,94670	0,73593	-0,70479	1,06320	0,77200	0,08568	0,90762

Table A.1: LAMDA-descriptors used to define batches

In the classification process, each descriptor is assigned to one "class" (Piera 1987). The class (k_i) is defined as the universe of descriptors, which characterizes one set of descriptors as pictured in Figure A.1.

In accordance with Aguado (1998) and Aguliar-Martin and Lopez (1982), in LAMDA two steps are necessary in order to obtain the final classification:

1. The MAD (Marginal Adequacy Degree): a term related of how one descriptor is related to one class. Each descriptor is compared with each existing class. This step has a possibility function to estimate the distribution of the descriptors based on a "fuzzification" of the binomial probability function computed as equation A.1:

$$MAD(d_i x_j / \rho_{i/k}) = \rho_{i/k}^{d_i x_j} (1 - \rho_{i/k})^{1 - d_i x_j} \quad (\text{A.1})$$

where

$d_i x_j$ = descriptor i of the object

$j \rho_{i/k}$ = descriptor i and class k

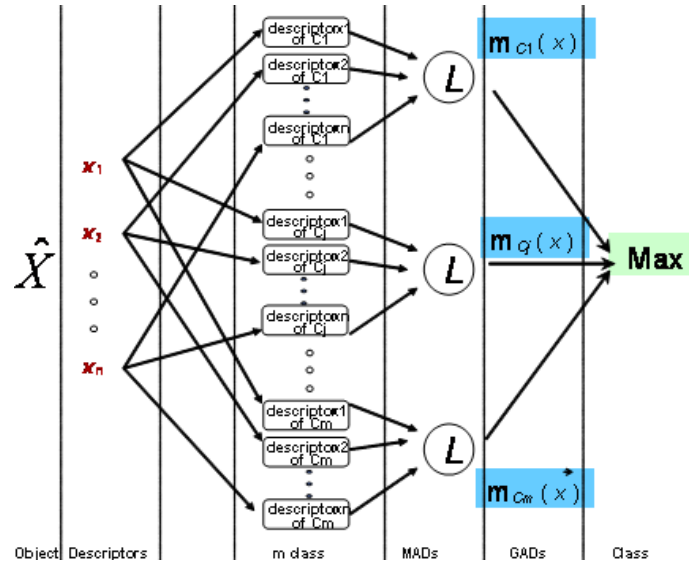


Figure A.1: Basic LAMDA recognition methodology

- GAD (Global Adequacy Degree) is defined as the pertinence degree of one object to a given class, as in fuzzy membership functions ($m_{c_i}(x)$) where the partial results from the MAD step are aggregated to form an individual class. The GAD step is performed as an interpolation between a t-norm and a t-conorm by means of the β parameter such that $\beta = 1$ represents the intersection and $\beta = 0$ means the union (in this application $\beta = 1$):

$$GAD = \beta T(MAD) + (1 - \beta)S(MAD) \quad (A.2)$$

where t-norm can be used with minimum / maximum values.

A.2 Semi-Industrial SBR Pilot Plant application

In this work, only the off-line stage was used for the batch classification. \hat{X} is the principal components of each batch, with dimensions 8 x 179 (Table A.1). The training method was not supervised but the exigency level was the maximum. Figure A.2 shows a part of the frame tool when the SALSA-LAMDA algorithm is used in order to classify the batches of operation from the semi-industrial pilot plant (LEQUIA group). In this Figure, the batches are on the x axis and the classes are on the y axis.

In this way, the tool automatically classified the data into eleven classes. According to these results and the analysis made by the chemical and the control engineers (Table 5.7), it is possible to identify classes that only contain batches with equipment defects, electrical faults, atmospheric changes or variation in the composition. Classes 1,9 and 10 correspond to Normal Operation Condition (NOC). Class 6 is due to atmospheric changes. Classes 3 and 11 are variations in composition. Classes 7 and 8 are electrical

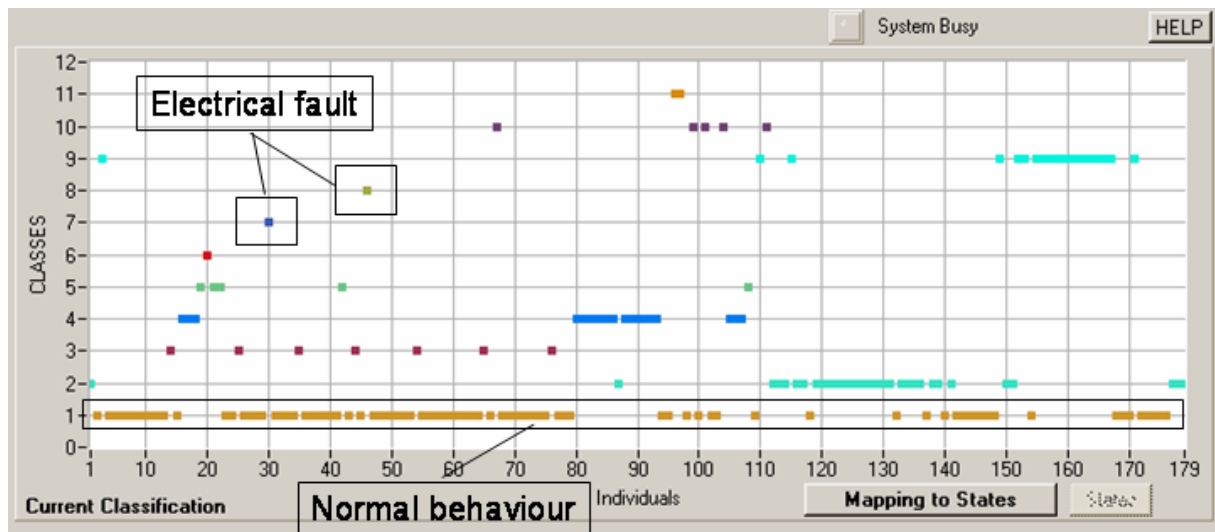


Figure A.2: LAMDA classification

faults. Finally, classes 2, 4 and 5 are composed of different types of batch processes with Abnormal Operation Condition (AOC) (Figure A.3).

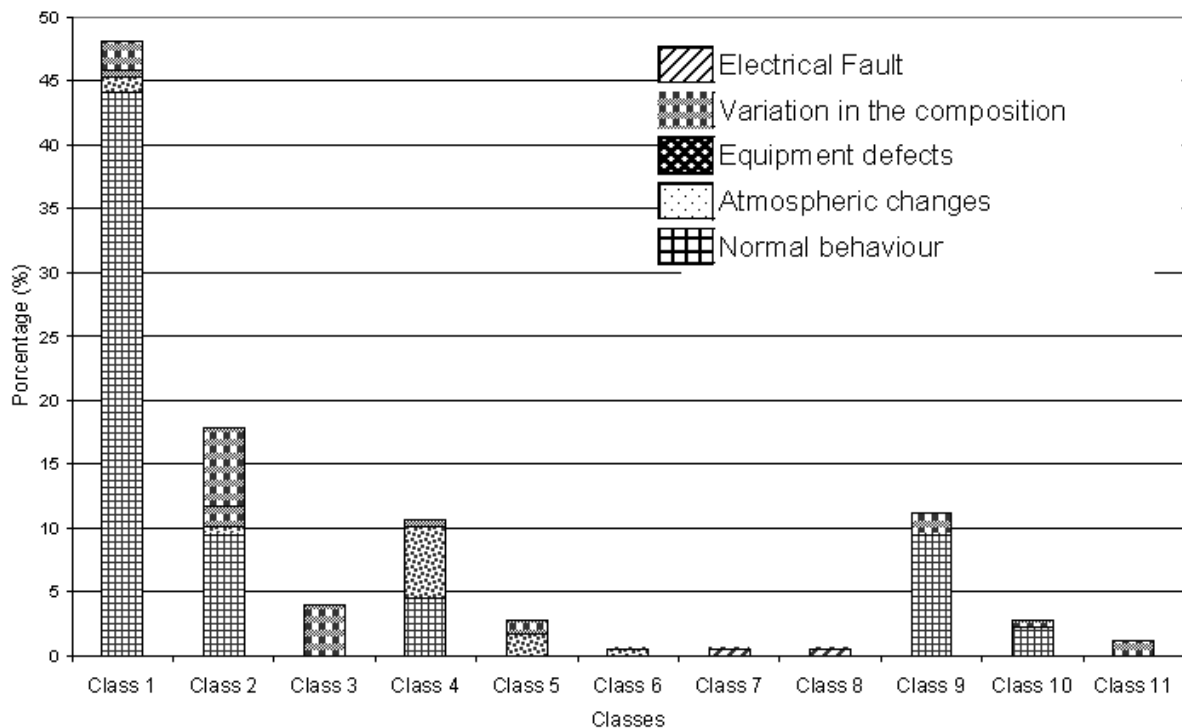


Figure A.3: Batch class composition according to type of batch process

Table A.2 shows a summary of the number and percentages of batches in each class, as well as the composition and the name assigned to each class. The predominant is class

1 which contains 48.04% of the total historical data; this class is called NOC (Normal Operation Condition). Some classes do not have one specific class of operation; for instance, class 5 is considered an Abnormal Operation Condition (AOC) because there are influent load change and equipment defects events.

CLASSES

	1	2	3	4	5	6	7	8	9	10	11
NOC	79	17	0	8	0	0	0	0	17	4	0
AT	2	1	0	10	3	1	0	0	0	0	0
EQ	1	3	0	0	2	0	0	0	0	0	2
VC	4	11	7	1	0	0	0	0	3	1	0
EF	0	0	0	0	0	0	1	1	0	0	0
Total	86	32	7	19	5	1	1	1	20	5	2
name	NOC	not classified	VC	not classified	AOC	AC	EF	EF	NOC	NOC	ED

Table A.2: Classes obtained by SALSA-LAMDA for semi-industrial pilot plant

The relationship between the class and principal components was also observed in this study. Table A.3 shows the result of this analysis. The 8th principal component has the lowest change of all the classes formed automatically. This indicates that the number of principal components of the MPCA model, and consequently, the LAMDA-descriptors (\hat{X} matrix), is seven. If only seven components are selected, the total variability will be 90.54%. In conclusion, the MPCA model will be with only 7 principal components.

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10	Class 11
1 st Component	0,250	0,064	0,977	0,267	0,313	0,313	0,313	0,313	0,230	0,563	0,300
2 nd Component	0,500	0,495	0,023	0,991	0,781	0,875	0,125	0,125	0,008	0,500	0,100
3 rd Component	0,817	0,815	0,804	0,813	0,798	0,739	0,125	0,875	0,813	0,792	0,755
4 th Component	0,832	0,840	0,818	0,845	0,813	0,875	0,375	0,125	0,828	0,806	0,767
5 th Component	0,832	0,840	0,818	0,810	0,844	0,625	0,750	0,125	0,833	0,806	0,900
6 th Component	0,500	0,500	0,500	0,500	0,969	0,875	0,500	0,500	0,516	0,375	0,900
7 th Component	0,002	0,005	0,023	0,009	0,031	0,875	0,125	0,125	0,008	0,042	0,100
8 th Component	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Table A.3: Batch class composition according to principal component

Results from this study were published in the Revista Iberoamericana de Inteligencia Artificial (2006) 29:99-107 Ruiz, Colomer and Melendez (2006).

CLASSES

	Amount	%	Name
Class 1	223	14.18	pH + DO
Class 2	259	16.47	NOC
Class 3	43	2.73	W + DO + Cond
Class 4	261	16.59	NOC
Class 5	139	8.84	DO
Class 6	74	4.70	pH + DO
Class 7	15	0.95	DO
Class 8	4	0.25	DO
Class 9	1	0.006	Temp + DO
Class 10	153	9.73	NOC
Class 11	32	2.03	ORP + W
Class 12	146	9.28	DO
Class 13	7	0.45	ORP + NOC
Class 14	8	0.51	pH + DO
Class 15	135	8.58	Cond + NOC
Class 16	54	3.43	DO
Class 17	19	1.21	Temp or DO

Table A.4: Classes from SALSA-LAMDA for BIOMATH SBR pilot plant

A.3 Lab-Scale Plant SBR application

The LAMDA tool was used in order to classify the batches from the BIOMATH lab-scale plant SBR (Chapter 5, section 5.2). There were six descriptors, and the MPCA model was developed using auto-scaling for the scale data step. In this way, 17 classes were formed automatically.

An analysis by the chemical and control engineers was made and as a result, the classes were assigned names: Class 1 comprises batches with Abnormal Operation due to pH and DO problems; class 2 represents batches with Normal Operation Condition (NOC); class 3 contains normal batches with problems in Weight (W), DO and Conductivity (Cond); class 4 is formed of batches with Normal Operation Condition; class 5 is made up of batches with AOC in the DO profile; class 6 contains two different types of events: pH and DO; classes 7 and 8 are made up of batches with perturbations in the normal evolution of the DO variable; class 9 is formed by just one batch with problems in Temp and DO; class 10 is made up by batches with Normal Operation Condition (NOC); class 11 is formed by some batches with AOC in ORP and W profiles; class 12 is formed by batches with problems in DO; class 13 formed by batches with NOC and small problems in the ORP profile; class 14 is made up of batches with DO and pH abnormal operation; class 15 is formed by batches NOC; class 16 is formed by batches with NOC but abnormal operation in DO; and class 17 is composed of batches with abnormal operation in Temp or DO.

Threedimensional figures are shown in order to make comparisons between classes with NOC and AOC. Figures A.4, A.6 and A.8 represent a class with normal behavior. Figures A.5, A.7 and A.9 are classes with problems in weight, DO and high values of conductivity. In Figures A.4 (class 2) and A.5 (class 3), a plane that crosses the origin is shown in order to help distinguish between classes with NOC and AOC. Using this plane, it is possible to check the values larger and less than zero (pink plane). Figures A.6 and A.7 highlight the classification using contrast of colors. Finally, Figures A.8 and A.9 show a cross-section for classes 2 and 3.

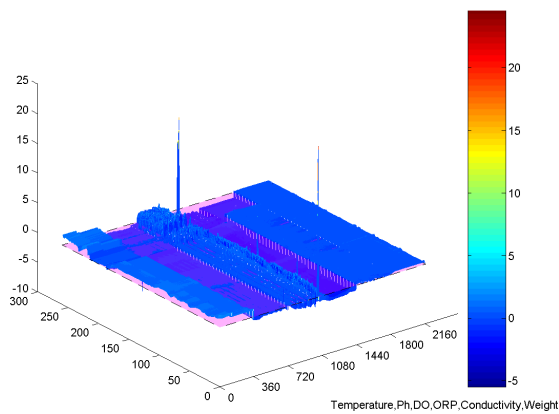


Figure A.4: Three dimensional representation for normal behavior (Class 2)

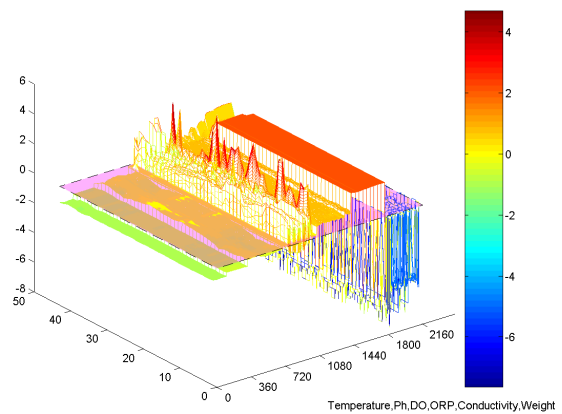


Figure A.5: Three dimensional representation for abnormal behavior (Class 3)

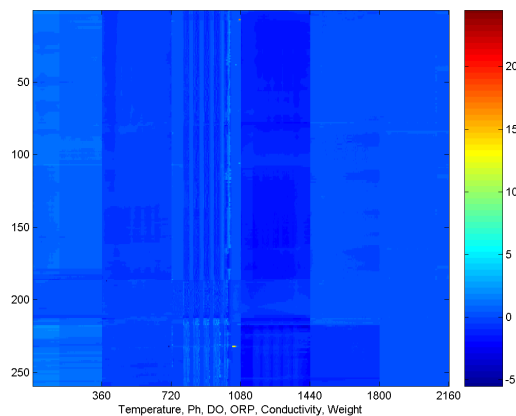


Figure A.6: Color levels for class 2

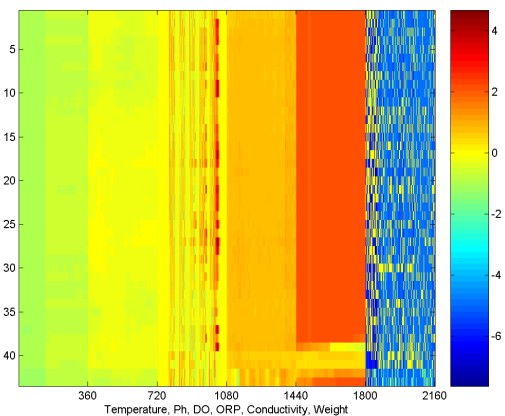


Figure A.7: Color levels for class 3

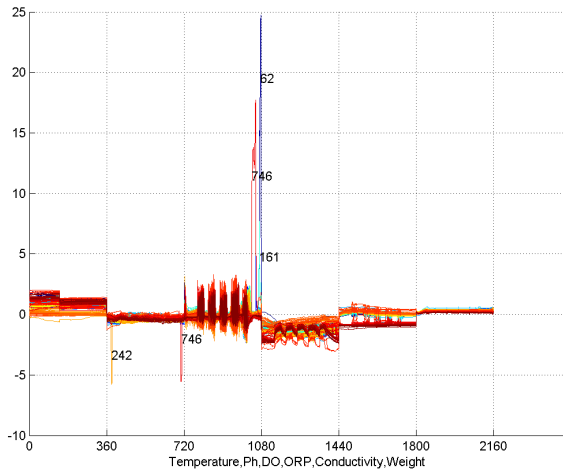


Figure A.8: Example of class with normal behavior (Class 2)

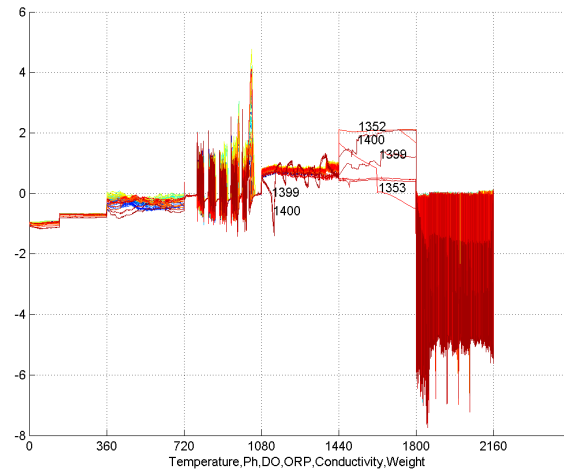


Figure A.9: Example of abnormal behavior (Class 3)

A.4 Data mining

Another study using the same Lab-Scale Plant SBR data from BIOMATH was developed, using MPCA to build the statistical model and LAMDA is used as a classification tool. However, in this particular work the procedure is performed twice. The first iteration discriminates between batches with normal and abnormal operation condition (NOC and AOC). The second iteration uses the same methodology, but using only batches with NOC obtained from the first iteration. This procedure helps to distinguish between batches with the same kind of operation.

In accordance with the conclusions from the previous section about the number of principal components and together with the study developed in Section 5.2 (Determination of the number of principal components), in this study six principal components and the Q -statistic are selected in order to constitute the descriptors.

LAMDA automatically classified 16 different classes. Each of these classes is subjected to an in depth investigation using the knowledge from the experts and the stored information of the data. One name is associated with each of the classes. Table A.5 shows the names together with the number of batches per class. Class 16 could not be marked, though this cluster exhibited only abnormal batches. Classes 1, 3, 5, 7 and 13 are identified as normal, corresponding to 73% of the data set. The 10 remaining classes could be marked with a specific fault or set of faults.

The descriptors from classes 1, 3, 5, 7 and 13 are used for the second iteration. Similarly to the first iteration, the obtained classes are investigated in detail in order to mark them. Table A.6 shows the obtained classes given with their number of batches and names.

Batches with the same or similar behavior are grouped together by means of the LAMDA algorithm. Diagnosis is then easily performed by investigation of a limited number of batches of those clusters. Combining MPCA and clustering therefore provides

an efficient and effective diagnosis tool for SBR processes. In the first iteration, where MPCA and LAMDA are combined, it is possible to discriminate several classes that uniquely correspond to a certain fault or set of faults. These classes represent 93% of the all abnormal batches found and the remainder of batches with AOC are identified by means of the second iteration. The results show that the combination of MPCA modelling and the LAMDA algorithm allows a discrimination of the operation and the alterations of the process. However, LAMDA can not learn when a new fault or problem is presented. More information about this work can be found in Villez et al. (2006).

Class	Amount	%	Label
1	219	11.18	normal 1 (low DO operation)
2	31	1.58	communication problem with balance
3	241	12.3	normal 2
4	187	9.55	cooler failure
5	607	30.99	normal 3
6	23	1.17	high pH
7	126	6.43	normal 4, recovery from cooler failure (cluster 4)
8	5	0.26	extreme DO
9	1	0.05	low ORP and extreme DO
10	72	3.68	conductivity probe failure
11	144	7.35	conductivity probe in repair
12	39	1.99	conductivity probe in repair and communication problem with balance
13	245	12.51	normal 5 (optimised operation)
14	7	0.36	low ORP measurement
15	1	0.05	multiple sensor failure (ORP, temperature and weight)
16	11	0.55	abnormal (no unique fault)

Table A.5: Names for classes of the first classification

Class	Amount	%	Operation	Label
1	210	14.59	normal 1	low DO setpoint, transient operation
2	95	6.61	normal 2	low DO setpoint, steady operation
3	98	6.82	normal 3	high SV ₃₀ (bad settling), low NO ₃ -N
4	49	3.41	normal 4	decreasing SV ₃₀ (improving settling), increasing NO ₃ -N
5	334	23.22	normal 5	filamentous bulking, decreasing/low NH ₄ -N, high NO ₃ -N
6	4	0.28	abnormal 1	high DO in anoxic phases
7	71	4.94	normal 6	increasing SV ₃₀ (worsening settling), high NO ₃ -N (10-20), increasing COD
8	94	6.54	normal 7	high SV ₃₀ (bad settling), high NO ₃ -N (>20), increasing COD
9	3	0.21	abnormal 2	high DO in aerobic phases
10	5	0.35	abnormal 3	high DO in aerobic phases (mixing too intense)
11	289	20.1	normal 8	filamentous bulking, high NH ₄ -N, decreasing/low NO ₃ -N
12	5	0.35	abnormal 4	cooling system failure
13	2	0.14	abnormal 5	pump control error: feeding too high in anaerobic phase
14	19	1.32	abnormal 6	pumping failure
15	116	8.06	normal 9	filamentous bulking, decreasing/low NH ₄ -N, high NO ₃ -N
16	1	0.07	abnormal 7	high DO in aerobic phases
17	43	2.99	normal 10	filamentous bulking, high NH ₄ -N, low NO ₃ -N

Table A.6: Names for classes of the second classification

Bibliography

- Aamodt, A.: 1991, *A Knowledge intensive approach to problem solving and sustained learning*, PhD thesis, University of Trondheim, Norwegian Institute of Technology. University Microfilms PUB 92-08460.
- Aamodt, A., Boose, J., Gaines, B. and Ganascia, J.: 1989, *Towards robust expert systems that learn from experience - an architectural framework*, number 311-326, EKAW-89: Third European Knowledge Acquisition for Knowledge-Based Systems Workshop.
- Aamodt, A. and Plaza, E.: 1994, Case-based reasoning: foundational issues, methodological variations, and systems approach, *AI communications* **7(1)**, 39–59.
- Acorn, T. and Walden, S.: 1992, *SMART: Support management cultivated reasoning technology Compaq customer service*, AAAI-92, MA: AAAI Press/MIT Press. Cambridge. USA.
- Aguado, D., Ferrer, A., Ferrer, J. and Seco, A.: 2007, Multivariate spc of a sequencing batch reactor for wastewater treatment, *Chemometrics and Intelligent Laboratory Systems* **85**, 82–93.
- Aguado, D., Ferrer, A., Seco, A. and Ferrer, J.: 2006, Comparison of different predictive models for nutrient estimation in sequencing batch reactor for wastewater treatment, *Chemometrics and Intelligent Laboratory System* **84**, 75–81.
- Aguado, D., Ferrer, A., Seco, A. and Ferrer, J.: 2007, Using unfolding-pca for batch-to-batch start-up process understanding and steady-state identification in a sequencing batch reactor, *Journal of Chemometrics* **22**, 81–90.
- Aguado, D., Zarzo, M., Seco, A. and Ferrer, A.: 2007, Process understanding of a wastewater batch reactor with block-wise pls, *Environmetrics* **18**, 551–560.
- Aguado, J.: 1998, *A Mixed Qualitative-Quantitative Self-Learning Classification Technique Applied to Situation Assessment in Industrial Process Control*, PhD thesis, Universitat Politecnica de Catalunya. Spain.
- Aguliar-Martin, J. and Lopez, R.: 1982, The process of classification and learning the meaning of linguistic descriptors of concepts, *Approximate Reasoning in Decision Analysis*, 165–175.
- Aha, D., Kibler, D. and Albert, M.: 1991, Instance-based learning algorithms, *Machine Learning* **6**, 37–66.

- Al-Kandari, N. and Jolliffe, I.: 2005, Variable selection and interpretation in correlation principal components, *Envirometrics* **16**, 659–672.
- Ambroisine, L., Guinot, C., Latreille, J., Mauger, E., Tenenhaus, M. and Guehenneux, S. (eds): 2003, *Relationship between clinical characteristics and biophysical parameters of the skin using PLS regression*, Third International Symposium on PLS and Related Methods (PLS'03), ISBN 2-906711-49-7.
- Barcelo, S. and Capilla, C.: 2002, *A multivariate statistical process control system to monitor a wastewater treatment process*, Second Annual Conference on Business and Industrial Statistics, Rimini, Italy.
- Bollen, M. H. J.: 2000, *Understanding Power Quality Problems*, 0-7803-1713-7, IEEE PRESS, Power Engineering.
- Camacho, J. and Pic, J.: 2006, Online monitoring of batch processes using multi-phase principal component analysis, *Journal of Process Control* **16**, 1021–1035.
- Castell, M. Z., Riquelme, A. F. and Villafranca, R. R. (eds): 2002, *Multivariate process control in improve the quality of PPOX production*, Second Annual Conference on business and industrial statistics, Rimini, Italy.
- CEE: 1991, *Diario Oficial n L 135 de 30/05/1991 P. 0040 - 0052 Directiva 91/271/CEE del Consejo, de 21 de mayo de 1991, sobre el tratamiento de las aguas residuales urbanas*, <http://www.gestion-ambiental.com/norma/ley/391L0271.htm>.
- Chang, P.-C. and Lai, C.-Y.: 2005, A hybrid system combining self-organizing maps with case-based reasoning in wholesaler's new-release book forecasting, *Expert Systems with Applications* **29(1)**, 183–192.
- Chen, K. H.: 2001, *Data-Rich Multivariate Detection and Diagnosis Using Eigenspace Analysis*, Doctor of philosophy in aeronautics and astronauti, Massachusetts Institute of Technology. Department of Aeronautics and Astronautics. USA.
- Cimander, C. and Mandenius, C.: 2002, Online monitoring of a bioprocess based on a multianalyser system and multivariate statistical process modelling, *Journal of Chemical Technology and Biotheology* (77), 1157–1168.
- Cinar, A. and Undey, C.: 1999, *Statistical process and controller performance monitoring: a tutorial on current methods and future directions*, Vol. 4(2625-2639), American Control Conference, Digital Object Identifier 10.1109/ACC.1999.786544.
- Colomer, J., Melendez, J. and Ayza, J.: 2000, *Sistemas de Supervisin*, 1 edn, Romany Valls, S.A.
- Comas, J.: 2000, *Development, Implementation and Evaluation of an Activated sludge Supervisory System for the Granoller WWTP*, PhD thesis, University of Girona. Spain.

- Congreso de los Diputados de Espana: 1978, *Spanish Magna Carta*, <http://www.congreso.es/funciones/constitucion/indice.htm>.
- Copp, J.: 2002, *The COST simulation benchmark. Description and simulator manual*, Office for Official Publications of the European Communities, ISBN 92-894-1658-0.
- Corominas., L.: 2006, *Control and optimization of an SBR for nitrogen removal: from model calibration to plant operation*, PhD thesis, University of Girona. Spain.
- Das, S., Lazarewicz, M. and Finkel, L. (eds): 2004, *Principal component analysis of temporal and spatial information for human gait recognition*, Vol. 2 of *Digital object identifier 10.1109/IEMBS.2004.1404267*, Proceedings of the 26th annual International Conference of the IEEE EMBS 2004.
- de Mantaras, R. L. and Plaza, E.: 1997, Case-based reasoning: An overview, *AI Communications* **10(1)**, 21–29.
- Directiva 98/15/CE de la Comision, de 27 de febrero de 1998, por la que se modifica la Directiva 91/271/CEE del Consejo en relacion con determinados requisitos establecidos en su anexo I.*
- Domeshek, E.: 1993, *A case study of case indexing: Designing index feature sets to suit task demands and support parallelism*. In, *Advances in connectionist and neural computation theory, Vol.2: Analogical connections*.
- Duchesne, C., Kourti, T. and MacGregor, J. F.: 2003, *Multivariate Monitoring of startups, restarts and grade transitions using projection methods*, number 0-7803-789, American Control Conference.
- Dudzic, M. and Quinn, S.: 2002, *Predictive modeling using adaptive PLS desulphurization reagent control system*, number 0-7803-729, American Control Conference.
- eXit and Disco: 2004, *SALSA Situation Assessment using LAMDA classification Algorithm*, CHEM Advanced Decision Support Systems for Chemical and Petrochemical Manufacturing Process, European Community G1RD-CT-2001-00466.
- Federation, W. E.: 2003, Wef.org web page (<http://www.wef.org>).
- Ferrer, A.: 2003, *Control Estadístico Mega Variante para los Procesos del Siglo XXI*, 27 Congreso Nacional de Estadística e Investigación Operativa (España).
- Flores, J. and MacGregor, J. F.: 2004, Multivariate monitoring of batch processes using batch to batch information, *AIChE Journal* **50(6)**, 1219–1228.
- Garcia, M., Ruiz, M., Colomer, J. and Melendez, J.: 2007, *Multivariate Principal Component Analysis and Case Base Reasoning methodology for abnormal situation detection in a Nutrient Removing SBR*, European Control Conference.
- Gentner, D.: 1983, Structure mapping - a theoretical framework for analogy, *Cognitive Science* **7**, 155–170.

- Giudici, P.: 2003, *Applied data mining Statistical methods for business and industry*, 047084678x edn, John Wiley & Sons Ltd, England.
- Gonzalez-Silvera, A., Santamaria, E., M., V., Garcia, T., Garcia, C., Milln, R. and Muller-Karger, F.: 2004, Biogeographical regions of the tropical and subtropical atlantic ocean off south america: Classification based on pigment (czcs) and chlorophyll-a (sea wifs) variability, *Continental Shelf Research* **24**, 983–1000.
- Grieu, S., Traor, A., Polit, M. and Colprim, J.: 2005, Prediction of parameters characterizing the state of a pollution removal biologic process, *Engineering Applications of Artificial Intelligence* **18**, 559–573.
- Gurden, S., Werterhuis, J., Rasmus, B. and Smilde, A.: 2001, A comparison of multiway regression and scaling methods, *Chemometrics and Intelligent Laboratory Systems* **59**, 121–136.
- Hare, L.: 2003: From chaos to wiping the floor web (<http://www.asq.org/pub/qualityprogress/past/0703/58spc0703.html>).
- Henze, M., Grady, C. J., Gujer, W., Marais, G. and Matsuo, T.: 1987, Activated sludge model no1, *Technical report*, IAWQ Scientific and Technical Report No1 - London, England.
- Insel, G., Sin, G., Lee, L., Nopens, I. and Vanrolleghem, P.: 2006, A calibration methodology and model-based systems analysis for SBRs removing nutrients under limited aeration conditions, *Journal of Chemical Technology and Biotechnology* **81**, 679–687.
- Jeppsson, U.: 2007, A general description of the iawq activated sludge model no.1, *I Jornada Tecnica de Modelatge d'EDAR. Benchmarking d'estrategies de control d'EDAR. Una eina til pel disseny, operaci i control d'EDAR*, University of Girona - Spain.
- Jeppsson, U., Rosen, C., Alex, J., Copp, J., Gernaey, K., Pons, M.-N. and Vanrolleghem, P.: 2006, Towards a benchmark simulation model for plant-wide control strategy performance evaluation of WWTPs, *Water Science and Technology* **53(1)**, 287–295.
- Keane, M. (ed.): 1988, *Where's the Beef? The absence of pragmatic factors in theories of analogy*, number 327-332, ECAI.
- Keats, J. B. and Hubele, N. F.: 1989, *Statistical Process Control in Automated Manufacturing*, 0-8247-7889-8, Marcel Dekker, Inc.
- Kosanovich, K., Piovoso, M., Dahl, K., MacGregor, J. and Nomikos, P. (eds): 1994, *Multiway PCA applied to an industrial batch process*, American Control Conference.
- Kourti, T.: 2002, Process analysis and abnormal situation detection: From theory to practice, *IEEE Control Systems Magazine* **22(5)**, 10–25.
- Kourti, T.: 2003a, *Abnormal Situation Detection, Three Way Data and Projection Methods; Robust Modeling for Industrial Applications*, 5th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes.

- Kourti, T.: 2003b, *Empirical Modeling with Latent Variables - From Theory to State of the Art Industrial Applications*, Third International Symposium on PLS and Related Methods (PLS'03), ISBN 2-906711-79-7, Lisbon (Portugal).
- Law, H., Snyder, C., Hattie, J. and McDonald, R.: 1984, *Research methods for multimode data analysis*, Praeger Publishers.
- Leake, D.: 1996, Case-based reasoning: experiences, lessons and future directions.
- Lee, D., Park, J. M. and Vanrolleghem, P.: 2005, Adaptive multiscale principal component analysis for on-line monitoring of a sequencing batch reactor, *Journal of Biotechnology* **116**, 195–210.
- Lee, D. S. and Vanrolleghem, P. A.: 2003a, Adaptive consensus principal component analysis for on-line batch process monitoring, *Technical report*, Fund for Scientific Research - Flanders (F.W.O.) and the Ghent University Research Fund, Coupure Links 653, B-9000 Gent, Belgium.
- Lee, D. S. and Vanrolleghem, P. A.: 2003b, Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis, *Biotechnology and Bioengineering* **82**(4), 489–497.
- Lee, J. H. and Dorsey, A. W.: 2003, Monitoring of batch process through state space models, *Revised for AIChE Journal MS* (7315RA).
- Lee, J., Yoo, C., Choi, S. and Vanrolleghem, P.: 2004, Nonlinear process monitoring using kernel principal component analysis, *Chemical Engineering Science* **59**, 223–234.
- Lee, J., Yoo, C. and Lee, I.: 2004a, Statistical monitoring of dynamic based on dynamic independent component analysis, *Chemical Engineering Science* **59**, 2995–3006.
- Lee, J., Yoo, C. and Lee, I.: 2004b, Statistical process monitoring with independent component analysis, *Journal of Process Control* **14**, 467–485.
- Lennox, B.: 2003, Multivariate statical process control, *Technical report*, Control Technology Centre Ltd School of Engineering University of Manchester, Dept. of Chemical and Process Engineering, University of Newcastle-upon-Tyne, UK.
- Li, W., Yue, H. H., Valle-Cervantes, S. and Qin, S. J.: 2000, Recursive PCA for adaptive process monitoring, *Journal of Process Control* (10), 471–486.
- Lopes, J., Menezes, J., Westerhuis, J. and Smilde, A.: 2002, Multiblock PLS analysis of an industrial pharmaceutical process, *Biotechnol Bioeng* (80), 419–427.
- MacGregor, J. F. (ed.): 2003, *Multivariate Statistical Approaches to Fault Detection and Isolation*, 5th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes.
- Maher, M. and Zhang, D.: 1991, *CADSYN: using case and decomposition knowledge for design synthesis*, *Artificial Intelligence in Design*.

- Marte, Y. A.: 2003, History of statistic web (<http://www.geocities.com/ymarte/trab/esthistor.html>).
- Martin, E., Morris, J. and Lane, S.: 2002, Monitoring process manufacturing performance, *IEEE Control Systems Magazine* **22**(5), 26–39.
- Martinez, M., Sanchez-Marre, M., Comas, J. and Rodriguez-Roda, I.: 2006, Case-based reasoning, a promising tool to face solids separation problems in the activated sludge process, *Water Science & Technology* **53**(1), 209–216.
- Montgomery, D. C.: 2000, *Introduction to Statistical Control Quality*, ISBN 0471316482.
- Mujica, L., Veh, J., Ruiz, M., Verleysen, M., Staszewski, W. and Worden, K.: 2008, Multivariate statistics process control for dimensionality reduction in structural assessment, *Mechanical Systems and Signal Processing* **22**, 155–171.
- Mujica, L., Vehi, J., Rodellar, J. and Kolakowski, P.: 2005, A hybrid approach of knowledge-based reasoning for structural assessment, *Smart Material and Structures* **14**(6), 1554–1562.
- Navichandra, D.: 1991, *Exploration and innovation in design:towards a computational model.*, Springer Verlag, New York. NY, US.
- NIST: 2003, National institute of standar and technology engineering statistics handbook, <http://www.itl.nist.gov/div898/handbook/index.htm>.
- Nomikos, P. and MacGregor, J.: 1995, Multivariate SPC charts for monitoring batch processes, *Technometrics* **37**(1), 41–59.
- Nomikos, P. and MacGregor, J. F.: 1994a, Monitoring batch processes using multiway principal component analysis, *AIChE* **40**(8), 1361–1375.
- Nomikos, P. and MacGregor, J. F.: 1994b, Multi-way partial least squares in monitoring batch processes, *First International Chemometrics InterNet Conference*.
- Norvilas, A., Tatara, E., Negiz, A., DeCicco, J. and Cinar, A.: 1998, *Monitoring and fault diagnosis of a polymerization reactor by interfacing knowledge based and multivariate SPM tools*, number 0-7803-453, American Control Conference.
- Nunez, H., Sanchez-Marre, M., Cortes, U., Martinez, M. and Poch, M.: 2002, *Classify environmental system situation by jeans of case-based reasoing: a comparative study*, iEMSs the internations Enviromental Modelling and software society.
- Nunez, H., Sanchez-Marre, M., Cortes, U., Comas, J., Martinez, M., Rodriguez-Roda, I. and Poch, M.: 2004, A comparative study on the use of similarity measures in case based reasoning to improve the classification of environmental system situations, *Environmetal Modelling & software* **19**, 809–819.

- Ondusi, K., Wollman, R., Ambrosone, C., Hutson, A., McCann, S., Tammela, J., Geisler, J., Miller, G., Sellers, T., Cliby, W., Qian, F., Keitz, B., Intengan, M., Lele, S. and Alderfer, J.: 2005, Detection of epithelial ovarian cancer using h-nmr-bases metabonomics, *Int. J. Cancer 2005* **113**, 782–788.
- Palmer, G., Zhu, C., Breslin, T., Xu, F., Gilchrist, K. and Ramanujam, N.: 2003, Comparison of multiexcitation fluorescence and diffuse reflectance spectroscopy for the diagnosis of breast cancer, *IEEE Transactions on Biomedical engineering* **50(11)**, 1233–1242.
- Piera, N.: 1987, *Connectius de lgiques no estandard com a operadors d'agregaci en classificaci multivariable i reconeixement de formes*, Doctoral dissertation, Universitat Politecnica de Catalunya, Spain.
- Puig, S., Vives, M., Corominas, L., Balaguer, M. and Colprim, J.: 2004, Wastewater nitrogen removal in SBRs, applying a step-feed strategy: From lab-scale to pilot plant operation, *Water Science and Technology* **50(10)**, 89–96.
- Qin, S. and Dunia, R.: 2000, Determining the number of principal components for best reconstruction, *Journal of Process Control* **10**, 254–250.
- Qin, S. J.: 2003, Statistical process monitoring: basics and beyond, *Journal of chemometrics* **17**, 480–502.
- Rao, C. R.: 1973, *Linear Statistical Inference and Its Applications*, John Wiley & Sons, Second Edition, ISBN 0-471-70823-2.
- Rius, A., Callao, M. and Rius., F.: 1997, Multivariate statistical process control applied to sulfate determination by sequential injection analysis, *Journal Analyst* **122**, 737741.
- Rodriguez-Roda, I., Sanchez, M., Comas, J., Baeza, J., Colprim, J., Lafuente, J., Cortes, U. and Poch, M.: 2002, A hybrid supervisory system to support WWTP operation: Implementation and validation, *Water Science and Technology* **45(4-5)**, 289297.
- Rosen, C., Jeppsson, U. and Vanrolleghem, P.: 2004, Towards a common benchmark for long-term process control and monitoring performance evaluation, *Water Science and Technology* **50(11)**, 41–49.
- Rosen, C. and Lennox, J. A.: 2001, Multivariate and multiscale monitoring of wastewater treatment operation, *Water Research* **35(14)**, 3402–3410.
- Rosen, C. and Olsson, G.: 1998, Disturbance detection in wastewater treatment plants, *Water Science and Technology* **37(12)**, 197–205.
- Rubio, M., Colomer, J., Ruiz, M., Colprim, J. and Melendez, J.: 2004, *Qualitative Trends for Situation Assessment in SBR Wastewater Treatment Process*, BESAI Workshop in Binding Environmental sciences and Artificial Intelligence, ECAI 2004 European Conference on Artificial Intelligence, ISSN.0922-6389.

- Ruiz, M., Colomer, J. and Melendez, Q.: 2006, Combination of statistical process control (SPC) methods and classification strategies for situation assessment of batch process, *Revista Iberoamericana de Inteligencia Artificial* **29**, 99–107.
- Ruiz, M., Colomer, J. and Melendez, J.: 2006, *Frontiers in Statistical Quality Control: Monitoring a sequencing batch reactor for the treatment of wastewater by a combination of multivariate statistical process control and classification technique*, Physica-Verlag Heidelberg New York, ISBN 10 3-7908-1686-8.
- Ruiz, M., Colomer, J., Rubio, M., Melendez, J. and Colprim, J.: 2004, *Situation assessment of a sequencing batch reactor using Multiblock MPCA and fuzzy classification*, BESAI-ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence.
- Ruiz, M. L., Colomer, J., Rubio, M. and Melendez, J.: 2004, *Combination of multivariate statistical process control and classification tool for situation assessment applied to a sequencing batch reactor wastewater treatment*, VIII International Workshop on Intelligent Statistical Quality Control, Printing House: Zaklad Poligraficzny, Warszawa, Poland ISBN 83-88311-69-7.
- Ruiz, M., Melendez, J., Colomer, J., Sanchez, J. and Castro, M.: 2004, *Fault location in electrical distribution systems using PLS and NN*, International Conference on Renewable Energies and Power Quality ICREPQ2004.
- Ruiz, M., Rosen, C. and Colomer, J.: 2006, *Diagnosis of a continuous treatment plant using Statistical Models and Case-Based Reasoning*, Proceedings of the 3rd International IWA Conference on Automation in Water Quality Monitoring (AutoMoNet2007). appeared on CD-ROM.
- Ruiz, M., Sin, G., Colprim, J. and Colomer, J.: 2008, MPCA and CBR methodology for monitoring, fault detection and diagnosis in wastewater treatment plant, *in preparation for publication in Water Science and Technology* .
- Ruiz, M., Villez, K., sin, G., Colomer, J., Rosen, C. and Vanrolleghem, P.: 2008, Different PCA approaches for monitoring nutrient removing batch process: pros and cons, *in preparation for publication in Water Science and Technology* .
- Ruiz, M., Villez, K., Sin, G., Colomer, J. and Vanrolleghem, P.: 2006, *Influence of scaling and unfolding in PCA based monitoring of nutrient removing batch process*, 6th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes.
- Russell, E. L., Chiang, L. H. and Braatz, R. D.: 2000, *Data-driven techniques for fault detection and diagnosis in chemical processes "Advances in Industrial Control"*, ISBN 1-85233-258-1, London.
- Sanchez-Marre, M., Cortes, U., R-Roda, I. and Poch, M.: 1997, Learning and adaptation in wastewater treatment plants through case-based reasoning, *Microcomputers in Civil Engineering* **12**, 251266.

- Schank, R.: 1982, *Dynamic memory: a theory of reminding and learning in computers and people.*, Cambridge University Press, Cambridge, UK.
- Schank, R. and Abelson, R.: 1977, *Scripts, Plans, Goals and Understanding*, Erlbaum, Hillsdale, New Jersey, US.
- Schuldt, J. E.: 1998, Walter Shewhart web (<http://www.geocities.com/wallstreet/exchange/9158/shewhart.htm>).
- Sharma, S. and Sleeman, D. (eds): 1988, *REFINER: A Case-Based Differential Diagnosis Aide for Knowledge Acquisition and Knowledge Refinement*, Vol. 88, European Working Session on Learning (EWSL).
- Sheppard, J. and Simpson, W. R.: 1998, *Research Perspectives and Case Studies in Systems Test and Diagnosis*, Vol. 13, Frontiers in Electronic Testing. Kluwer. Chapter 5. Inducing Inference Models from Case Data.
- Sin, G., Insel, G., Lee, D. and Vanrolleghem, P.: 2004, Optimal but robust N and P removal in SBRs: a model-based systematic study of operation scenarios, *Water Science and Technology* **50(10)**, 97–105.
- Singh, K., Malik, A., Mohan, D., Sinha, S. and Singh, V.: 2005, Chemometric data analysis of pollutants in wastewater: a case study, *Analytica Chimica Acta* **532**, 15–25.
- Singh, R.: 2003, Visualization tools for process condition monitoring, *Mathworks matlab*, MATHWORKS MATLAB.
- Skonieczny, G. and Torrisi, B.: 2003, *PCA, PLS and ANN: diversification of the Italian financial system*, Third International Symposium on PLS and Related Methods (PLS'03), ISBN 2-906711-49-7.
- Smilde, A. K.: 2001, Comments on three way analyses used for batch process data, *Journal of Chemometrics* (15), 19–27.
- Stadlthanner, K., Tom, A., Teixeira, A. and Puntonet, C.: 2004, *Kernel PCA denoising of artefact-free protein NMR spectra*, Vol. 3 of *Proceeding IEEE International Joint Conference on Neural Networks*, K. Stadlthanner and A.M. Tom and A.R. Teixeira and C.G. Puntonet.
- Takacs, I., Patry, G. and Nolasco, D.: 1991, A dynamic model of the clarification thickening process, *Water Research* **25(10)**, 1263–1271.
- Undey, C. and Cinar, A.: 2002, Statistical monitoring of multistage, multiphase batch processes, *IEEE Control Systems Magazine* **22(5)**, 40–52.
- Undey, C., Ertunc, S. and Cinar, A.: 2003, Online batch/fed-batch process performance monitoring, quality prediction, and variable-contribution analysis for diagnosis, *American Chemical Society* .

- Venkatasubramanian, V., Rengaswamy, R., Yin, K. and Kavuri, S.: 2003, A review of process fault detection and diagnosis part i: qualitative model-based methods, *Computer and Chemical Engineering* **27**, 293–311.
- Villez, K., Ruiz, M., Sin, G., Colomer, J., Rosen, C. and Vanrolleghem, P. A.: 2006, *Combining Multiway Principal Component Analysis (MPCA) and clustering for efficient data mining of historical data sets of SBR processes*, Automation in Water Quality Monitoring AutMoNet2007. appeared on CD-ROM.
- Vives, M.: 2004, *SBR technology for wastewater treatment: suitable operational conditions for nutrient removal*, PhD thesis, University of Girona, Spain.
- Vives, M., Balaguer, M., Garca, R. and Colprim, J.: 2001, Study of the operational conditions for organic matter and nitrogen removal in a sequencing batch reactor, *Technical report*, University of Girona, Spain.
- Wang, H. and Wang, H.: 2005, A hybrid expert system for equipment failure analysis, *Expert Systems with Applications* **28**, 615–622.
- Watson, I.: 1998, CBR is a methodology not a technology, *Research and Development in Expert Systems* **15**, 213223.
- Westerhuis, J. A. and Coenegracht, P. M.: 1997, Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock PLS, *Journal of Chemometrics* **11**, 379–392.
- Westerhuis, J. A., Kourti, T. and MacGregor, J. F.: 1999, Comparing alternative approaches for multivariate statistical analysis of batch process data, *Journal of Chemometrics* **13**, 397–413.
- Wiese, J., Stahl, A. and Hansen, J.: 2004, *Possible Application for case-based reasoning in the field of wastewater treatment*, 16th European Conference on Artificial Intelligence BESA workshop on Biding Environmental Sciences and Artificial Intelligence.
- Wilson, R. and Martinez, T.: 2000, Reduction techniques for instance-based learning algorithms, *Machine Learning* **38**, 257–286.
- Wise, B., Gallagher, N., Watts, S., White, D. and Barna, G.: 1999, A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process, *Journal of Chemometrics* **13**, 379396.
- Wise, B. M., Gallagher, N. B., Bro, R. and Shaver, J. M.: 2003, *PLS Toolbox 3 0*, Eigen Vector Research Incorporated.
- Wold, S., Geladi, P., Esbensen, K. and Ohman, J.: 1987, Multiway principal component and PLS analysis, *Journal of Chemometrics* **1**, 41–56.
- Wold, S., Kettaneh, N., Friden, H. and Holmberg, A.: 1998, Modelling and diagnostics of batch processes and analogous kinetic experiments, *Chemometrics and Intelligent Laboratory Systems* **44**, 331–340.

- Yoo, C. K., Lee, D. S. and Vanrolleghem, P. A.: 2004, Application of multiway ICA for on-line process monitoring of a sequencing batch reactor, *Water Research* **38**, 1715–1732.
- Yoo, C.-K., Villez, K., Lee, I.-B. and Vanrolleghem, P. A.: 2006, Multivariate nonlinear statistical process control of a sequencing batch reactor, *Journal of Chemical Engineering of Japan* **39(1)**, 43–51.
- Yoo, C., Vanrolleghem, P. and Lee, I.: 2003, Nonlinear modelling and adaptive monitoring with fuzzy and multivariate statistical methods in biological wastewater treatment plants, *Journal of Biotechnology* **105**, 135–163.
- Yoon, S. and MacGregor, J. F.: 2000, Statistical and causal model-based approaches to fault detection and isolation, *AIChE Journal* **46**, 1813–1824.
- Zhan, Z., Qin, Q., Wang, X. and Ghulam, A.: 2004, *Study on ecological indices from NDVI using NOAA/AVHRR data in Western Loess Plateau of China*, Vol. 6 of *Digital object identifier 10.1109/IGARSS.2004.1369910*, Proceedings IEEE International Geoscience and Remote Sensing Symposium.
- Zhang, L. and Bollen, M.: 1998, A method for characterizing unbalanced voltage dips (sags) with symmetrical components, *IEEE Power Engineering Letters* .