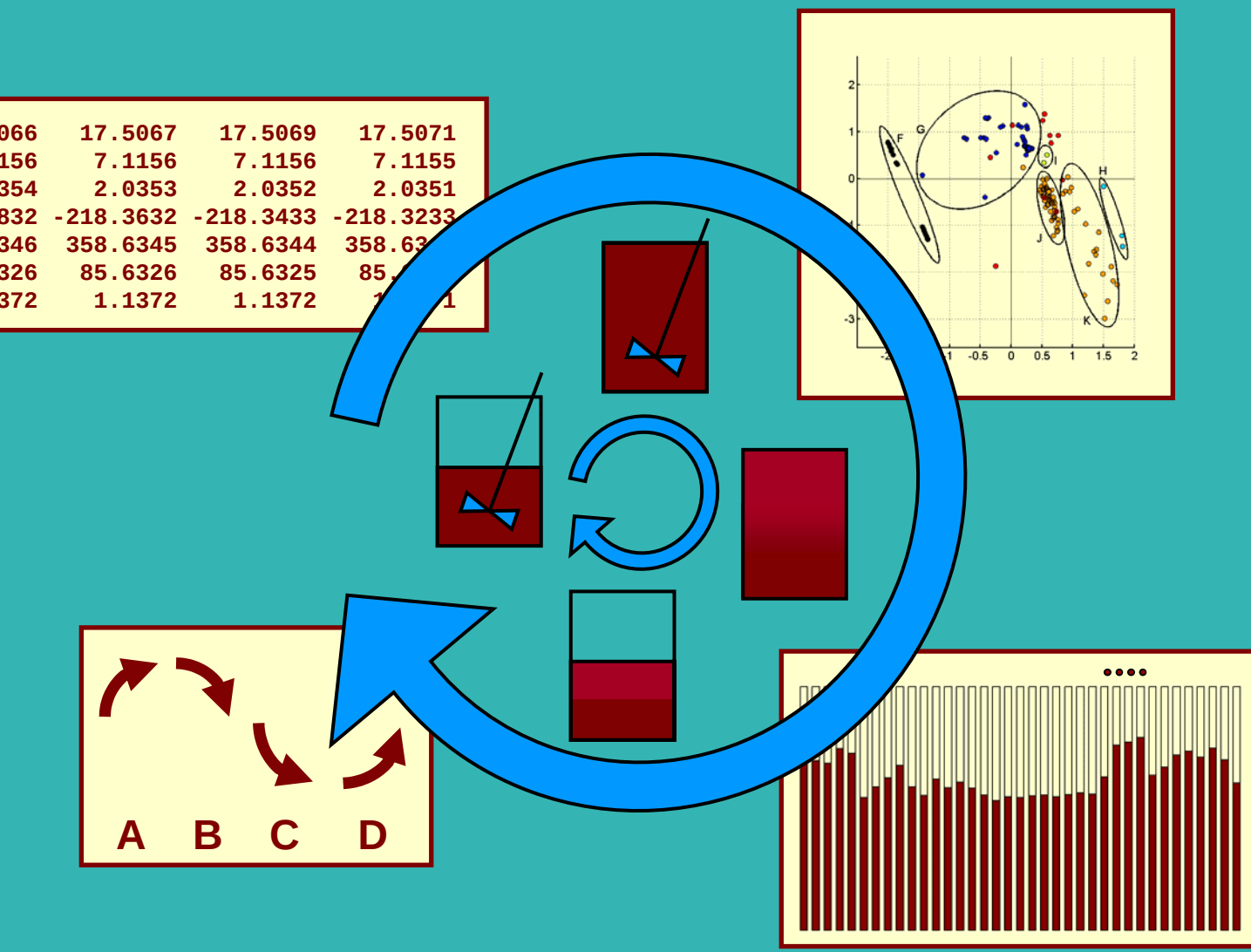
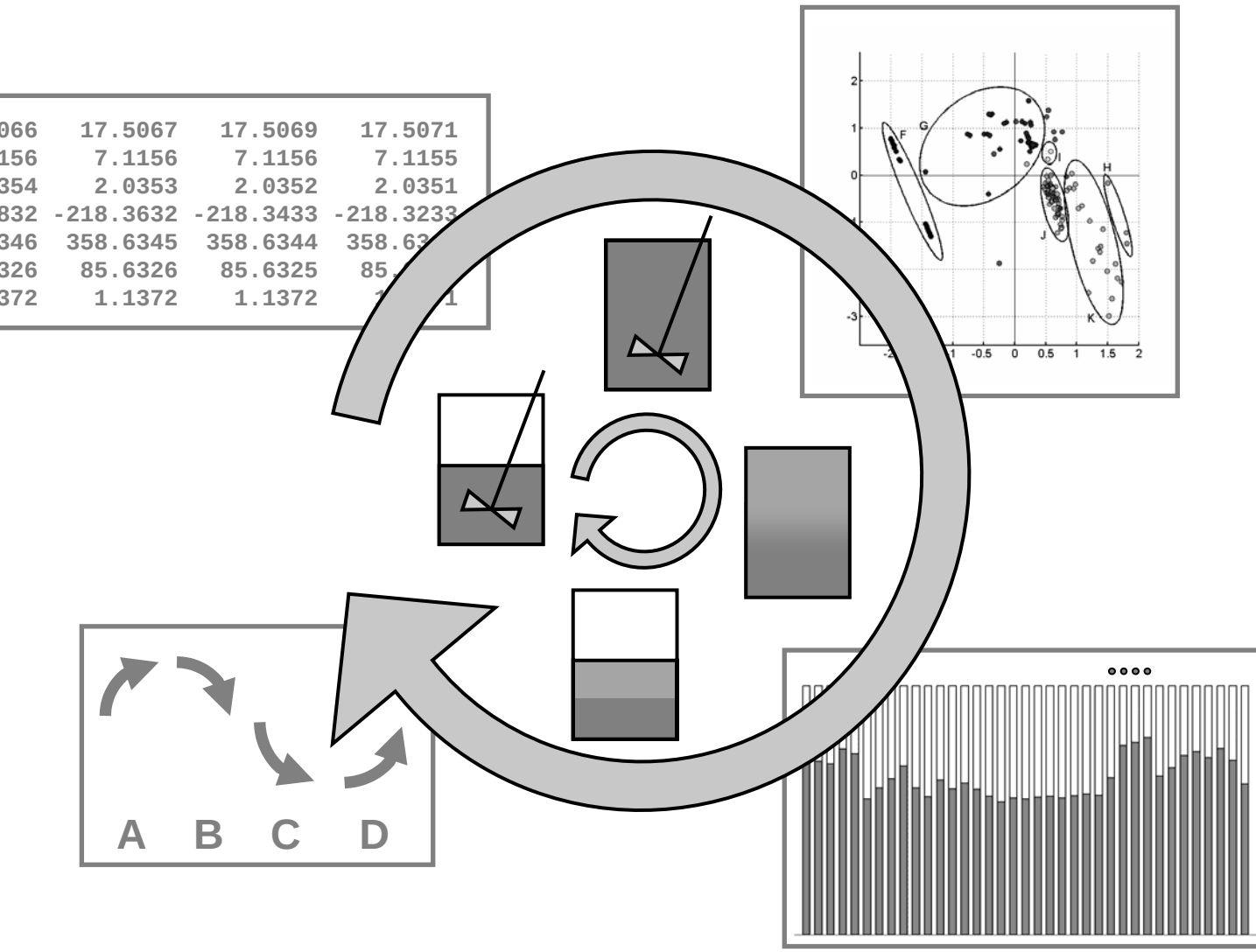


Multivariate and qualitative data-analysis for monitoring, diagnosis and control of sequencing batch reactors for wastewater treatment

ir. Kris Villez

UNIVERSITEIT GENT

FACULTEIT BIO-INGENIEURSWETENSCHAPPEN



Multivariate and qualitative data-analysis for monitoring, diagnosis and control of sequencing batch reactors for wastewater treatment

ir. Kris Villez

UNIVERSITEIT GENT

FACULTEIT BIO-INGENIEURSWETENSCHAPPEN

Quo clarior lux,
eo obscuriores umbrae

Examination Committee
Prof. Dr. ir. Marc Van Meirvenne (Ghent University)
Prof. Dr. ir. Olivier Thas (Ghent University)
Prof. Dr. ir. Willy Verstraete (Ghent University)
Prof. Dr. Jean-Pierre Ottoy (Ghent University)
Prof. Dr. ir. Shankar Narasimhan (Indian Institute of Technology Madras, India)
Prof. Dr. ir. Jean-Philippe Steyer (INRA-LBE, France)
Prof. Dr. ir. Krist Gernaey (Danish Technical University, Denmark)

Promotors
Prof. Dr. ir. Peter A. Vanrolleghem (Department of Applied Mathematics, Biometrics and Process Control, Ghent University)
Dr. ir. Christian Rosén (Veolia Water Solutions & Technologies, VA Ingenjörema), Malmö, Sweden

Dean
Prof. Dr. ir. Herman Van Langenhove

Rector
Prof. Dr. Paul Van Cauwenberge

ir. Kris Villez

Multivariate and qualitative data analysis for monitoring, diagnosis and control of sequencing batch reactors for wastewater treatment

Thesis submitted in fulfilment of the requirements for the degree of
Doctor (Ph.D.) in Applied Biological Sciences, Environmental Technology

Dutch translation of the title:
Multivariate en kwalitatieve data-analyse voor opvolging, diagnose en regeling van sequentiële batch reactoren voor afvalwaterzuivering

Please refer to this work as follows:
Kris Villez, 2007. Multivariate and qualitative data analysis for monitoring, diagnosis and control of sequencing batch reactors for wastewater treatment. Ph.D. thesis, Ghent University, Gent, Belgium.

ISBN 978-90-5989-211-8

Dankwoord

to choose is sometimes to lose they say
to lose what you haven't chosen but yes
I confess I want you all

Taken from *Nautilus & Zeppelin* by Samael

Met dit doctoraat wordt een einde gebracht aan een –(naar alle waarschijnlijkheid en hopelijk) niet-cyclisch batch– proces dat iets meer dan vier jaar geleden is begonnen. De nodige stappen vooruit, opzij en soms voorzichtig achteruit die voorafgingen aan dit sluitstuk zouden niet mogelijk zijn geweest zonder de steun en hulp van velen. Een dankwoord is hier en nu dan ook meer dan gepast.

Mijn dank gaat uit naar mijn promotoren, Prof. dr. Peter Vanrolleghem en dr. Christian Rosén, die mij de kans gaven om me onderzoeksmatig te ontwikkelen. Ik wil Peter in het bijzonder bedanken om mij te introduceren in een groep van zeer leuke mensen, voor zijn ongeëvenaarde gedrevenheid en voor de fantastische tijd in Québec. *Special thanks go out to Christian: for* no-problems, *discussions –scientific or shoe-matic but always fashionable–, for time spent together in Belgium, Sweden and other outskirts of our globe*. Dankewel, Peter! *Tuzen tack, Christian!*

Leuke bureau –achteraan en met zicht naar buiten–, laptop voor de neus en beginnen maar! Zo eenvoudig is het echter genoeg niet. Enkele uitzonderlijke mensen zijn ontegenzegbaar noodzakelijk gebleken om dit alles tot een goed einde te brengen. Zonder Griet en Tinne geen data, laat staan dit doctoraat. Zonder *sysadmin* geen uitgebreide data-analyse. Zonder Brecht, Gaspard en Lieven tekst noch leesbare formule. Dank daarvoor. Een aangename sfeer op de vakgroep was steeds gegarandeerd, mede dankzij fantastische cocktails en mega-fuiven. Dank aan alle collega's binnen en buiten de vakgroep; voor hulp en uitleg, voor begrip en om er eenvoudigweg te zijn.

Extra dank en groet gaat uit naar Jeroen, vriend en drinkebroer sinds jaren.

Een apart woordje plaats ik hier voor de studenten die ik mocht begeleiden. Zij beten zich koppig vast in een materie die hen niet altijd makkelijk lag. De hardnekkigheid en motivatie waarmee ze te werk gingen hebben absoluut bijgedragen tot dit doctoraat. Hou jullie goed!

Bureau al goed en wel, verrek, soms zat ik op een andere plek. Eerst was dat heel kort in Zweden om alles startklaar te maken. Later was een half-week-frenzy met Lieven in China voldoende om de begrippen geel, kip en vis voor de eeuwigheid te herdefiniëren. De buitenland-microbe had nu wel echt gebeten. Ik kon verder trippen in Zuid-Korea, Spanje, Polen, Vermont (VS), New York City (VS), Canada, Frankrijk en Indiana (VS). Met Jan en Katrijn (en Klaas-in-wording) werd een bijzondere vakantie in Canada ingelast. *For warm-hearted welcoming, steep learning curves and plain fun I want to thank all people abroad that have assisted me on my path of learning. Explicit thanks go out to to those that helped me out on the road –often unsolicited, at times most welcome and always appreciated.*

Met de eerste buitenlandse ervaringen drong iets vreemds zich op. Ik begon zowaar te sporten. Geïnspireerd door de Aziatische ervaringen, zette ik de eerste stappen in Aïkido. De Canadese ervaring zette bijkomend aan om te gaan duiken. Niet alleen een nieuwe balans maar ook een hoop nieuwe vrienden blijkt van al dat sporten het resultaat. Bedankt Aikido- en duikvrienden!

Uit de vorige eeuw stamt het begrip In Somnis (R.I.P.). Repeteren, podium-stress, road-trippen en –niet-te-vergeten– pintelieren hoorde er allemaal bij en wel 10 jaar lang. Bedankt voor creativiteit, karakter en plezier!

Bram, Els, ma & pa, in al mijn doen en laten durf ik al eens het noorden te verliezen. Jullie zijn een onontbeerlijk anker geweest om mee tot rust te komen en steeds weer koers vooruit te kunnen zetten. Bedankt ook aan de hele familie voor steun en toeverlaat. Merci!

So long, and thanks for all the coffee!

Kris,

Gent, december 2007

Contents

List of Abbreviations & Symbols

Abbreviations

AdMSPCA	Adaptive Multi-Scale PCA
AMBPCA	Adaptive Multi-Block PCA
ANN	Artificial Neural Network
ARMA	Auto-Regressive Moving Average
ATP	adenosinetriphosphate
BDPCA	Batch-Dynamic PCA
DPCA	Dynamic PCA
COD	Chemical Oxygen Demand
CSTR	Continuously Stirred Tank Reactor
CV	Cross Validation
CV-RMSR	Cross-Validated Root Mean Square Residual
DO	Dissolved Oxygen
FCM	Fuzzy C-Means clustering
FFT	Fast Fourier Transform
FSPCA	Function Space PCA
HPCA	Hierarchical PCA
IDPCA	Indirect Dynamic PCA
IT-PCA	Input Training PCA
IT-MPCA	Input Training Multiway PCA
IT-FSPCA	Input Training Function Space PCA
KPCA	Kernel PCA
KMPCA	Kernel Multi-way PCA
LS, LS_x, LS_y	Least-Squares, Least-Squares in x, Least-Squares in y
MixPCA	Mixture PCA
MixMPCA	Mixture Multi-way PCA

MPCA	Multi-way PCA
$MPCA_C$	MPCA model for Common-cause variation
$MPCA_R$	MPCA model for Residual variation
MBPCA	Multi-Block PCA
MPPCA	Multi-Phase PCA
MSPCA	Multi-Scale PCA
MVSPC	Multivariate Statistical Process Control
NH_4^+	ammonia
NH_4^+-N	ammonia nitrogen
NIPALS	Non-linear Iterative Partial Least Squares
NO_2^-	nitrite
NO_2^--N	nitrite nitrogen
NO_3^-	nitrate
NO_3^--N	nitrate nitrogen
NN	Neural Network
ORP	Oxidation Reduction Potential
PCA	Principal Component Analysis
PFR	Plug Flow Reactor
PID	Proportional-Integral-Derivative
PO_4^{3-}-P	inorganic phosphorus
QA	Qualitative Analysis
QR	Qualitative Reasoning
QRT	Qualitative Representation of Trends
RMSR	Root Mean Square Residual
SBR	Sequencing Batch Reactor
SRT	Sludge Retention Time
TAN	Total Ammonia Nitrogen
TN	Total Nitrogen
TP	Total phosphorus
TLS	Total Least Squares
UCL	Upper Control Limit

English symbols

a	approximation, low-pass signal
b, B	scalars
c	principal component index (scalar)
C	number of principal components (scalar)
g, G	scalars
cov ()	variance-covariance operator
d	detail, high-pass signal
e	measurement error
$\mathbf{e}$	vector of measurement errors
E ()	expected value operator
F_a	fuzzy covariance matrix
H_a	norm-inducing matrix
J	number of variables / objective function
k	time index (integer)
$\mathbf{K}$	Kernel matrix
$\mathbf{m}$	center of PCA model or fuzzy cluster
N	number of observations, (discrete) series length
p	scale index (integer)
$p_{j,c}$	j^{th} element of c^{th} principal component
$\mathbf{p}_{.,c}$, $\mathbf{p}$	(c^{th}) principal component, eigenvector
P	maximal scale index (wavelet analysis)
$\mathbf{P}$	matrix of principal components, eigenvectors
q	residual, deviation between estimated value and error-free value
$\mathbf{q}$	residual vector, vector of deviations between estimated values and error-free values
r	residual, deviation between estimated value and measured value
$\mathbf{r}$	residual vector, vector of deviations between estimated values and measured values
s	wavelet scale
$\mathbf{s}$	vector of scaling parameters for PCA model
s_x	standard deviation of variable x
t	time index (integer)
$t_{i,c}$	c^{th} principal score for i^{th} observation

tr $()$	trace operator
T^2	Hotelling's T^2 statistic
$\mathbf{T}$	matrix of principal scores
var $()$	variance operator
x, $\mathbf{x}$, $\mathbf{X}$, $\underline{\mathbf{X}}$	scalar, vector, 2-D matrix, 3-D matrix of measurements
X^2	Chi-square statistic
z, $\mathbf{z}$, $\mathbf{Z}$	scalar, vector, 2-D matrix of error-free variables
$\mathbf{x}_{.,j}$	j-th column in a matrix $\mathbf{X}$
$\mathbf{x}_{i,.}$	i-th row in a matrix $\mathbf{X}$
$\underline{x}_{i,j,k}$	element in three-dimensional measurement matrix with coordinates (i, j, k)
w	wavelet coefficient

Greek symbols

α	scalar parameter
χ^2	Chi-square distribution
δp	wavelet scale resolution (octaves per scale index increment)
$\delta(\cdot)$	Dirac delta function
ω	frequency
μ, $\boldsymbol{\mu}$	true process mean scalar/vector
ν	(fuzzy) cluster membership
ϕ	scaling function (time domain)
ϕ^*	scaling function (frequency domain)
ψ	wavelet function (time domain)
ψ^*	wavelet function (frequency domain)
ψ_o	mother wavelet function (time domain)
ψ_o^*	mother wavelet function (frequency domain)
ρ_j	relative magnitude of cluster region
σ_e	(white) noise standard deviation
τ	oscillation period

Other symbols

$\mathbf{a}^T$, $\mathbf{A}^T$	transpose of vector $\mathbf{a}$, matrix $\mathbf{A}$
$\Re(\cdot)$	real part
$\Im(\cdot)$	imaginary part
$*$	convolution operator

Part I

Introduction and background

Chapter 1

Introduction and problem statement

1.1 Introduction

Water plays a central role in the existence and activities of mankind. First of all, within the human body, water serves as the main vector for assimilation, excretion and transportation of chemical compounds and energy and is the matrix for virtually all physico-chemical conversion processes. As such, access to potable water is essential to the mere existence of human life. Secondly, mankind has increasingly cultivated additional activities over the centuries leading to intensified anthropogenic consumption of water, including food generation (e.g. irrigation in agriculture, fermentation processes), energy production, cleaning and leisure activities. Intensified consumption of water as well as the growing world population have led to ever increasing rates of water pollution and wastage.

Increased awareness of pollution processes and their intensification, concerns with health of human populations and the status of the earth's environment as well as the simple necessity of usable water have pushed forward the development of techniques for wastewater handling and treatment (Wiesmann et al., 2006). In

Mohenjo-Daro, one of the earliest systems for handling of wastewater, including toilets and sewer canals, has been traced back to 1500 BC. More recently, Roman culture engineered a sewer system for handling anthropogenic wastewater as well, having constructed a sewer line for the city of Rome, named the Cloaca Maxima. Effective treatment, rather than transportation from one environment to another, has been developed much later in human history (19^{th} Century), at first in view of guaranteeing human health. The activated sludge process, still the most popular wastewater treatment process, has come into existence by the works of Arden and Locket (beginning of the 20^{th} Century) and has been extended, optimized and increasingly applied since then.

Over the last century, increased understanding of the biochemical processes involved in wastewater treatment as well as technological developments in general, have resulted in improvements of wastewater treatment processes in terms of effectiveness, efficiency and costs. Design, control and optimization of wastewater treatment plants, dedicated measurement technologies and the mere understanding of their operation have been unrelinquished subjects for research and development.

1.2 Problem statement

In contrast to the promising theoretical and experimental results in the research field of wastewater treatment, the presented developments are not (easily) applied in practice. Low reliability of sensors and actuators are often indicated as significant barriers between results in research and development stages and industrial practice. In addition, biological (wastewater treatment) processes are characterized by changing characteristics which impedes a straightforward use of classic and non-adaptive control techniques. In view of the latter aspects of wastewater treatment practice, design and improvement of techniques for process monitoring, diagnostics and advanced control still represent a challenging field for research and development.

In relation to the aforementioned discrepancy between research and practice, the following aspects, typical for many industrial processes are to be considered:

- Several processes, i.e. (bio)chemical and physical processes, occur simultaneously and are characterized by complex causal relationships within and between them.

- Relationships (1) between measured variables and (2) between measured variables, process states and process performance are not always understood well.
- Increased use of sensors results in growing data bases that are data-rich and information-poor.

In addition, wastewater treatment systems share the following characteristics:

- Biological processes are typically slower than their chemical equivalents. Computational demands for process supervision are therefore not expected to represent a significant barrier. Important time scales may range from the minute scale (oxygen level control) to weeks or months (biomass growth, microbial population dynamics).
- Changing ambient and environmental conditions as well as past operation of wastewater treatment processes affect the process characteristics of biological systems. Mostly reported are process changes induced by seasonal changes in ambient temperature.
- Many important process variables are of a hidden nature. For example, the amount and activity of specific groups of bacteria (e.g. nitrifying bacteria, denitrifying bacteria, phosphorus accumulating organisms (PAO's)) and concentrations of biochemical component concentrations (e.g. amount of slowly biodegradable substrates, phosphorus content of microbial cells) cannot be measured on-line. Exact knowledge of relationships between directly controlled variables and hidden process states is inherently absent hereby troubling the formulation and acceptance of automated control laws in practice.
- In contrast to producing industries, where ingredients of low quality can be discarded and the availability of low quality products to the commercial market can be avoided or restricted, all wastewaters have to be dealt with and are therefore unalteredly accounted for during process performance evaluations, irrespective of their qualities. Not being able to properly handle certain wastewaters results in economic penalties and restriction of licenses in most industrialized countries.

State-of-the-art techniques in supervision and control for wastewater treatment systems lack necessary qualities (e.g. robustness, sufficient intelligence) for deployment in practice, often due to inadequate or insufficient knowledge of existing

relationships between monitored and controlled variables, lack of integration of several control loops and the inability to foresee disturbances to wastewater treatment plants (Olsson, 2006). Therefore, the main objectives of this dissertation are the evaluation, validation and improvement of techniques for process monitoring, diagnosis and control of wastewater treatment systems. More specifically, techniques will be searched for that allow:

- Efficient and effective detection of faults and key events
- The assessment and proper use of diagnostic information
- Advanced control of wastewater treatment systems

The detection and diagnosis of faults is a large research area in which several approaches co-exist. Developments can for example be categorized by their quantitative or qualitative nature or by their data-driven (inductive, black-box) or knowledge-driven (deductive, white-box) basis. A myriad of contributions exist in literature and it is hard to keep hold of them (Venkatasubramanian et al., 2003a,b,c). Due to the inherent complexity of biological processes and lack of direct measurements of key variables therein (e.g. metabolite concentrations, biomass activity), unique and universal descriptions of biological systems are non-existing (Alewell and Manderscheid, 1998). The latter characteristic turns the knowledge-driven or mechanistic approaches largely inappropriate for process supervision of biological systems such as wastewater treatment plants. With respect to (1) increased recording of data-rich and information-poor data sets of wastewater treatment plant operations and (2) the poor understanding of exact relationships between measured variables, a largely data-driven approach is believed to be more promising and is therefore selected. In order to enable the proposed application and evaluation steps, a pilot-scale SBR system for nutrient removal will be used as a study object throughout the larger part of this thesis.

1.3 Contributions

Results of the presented work are discussed in detail in Chapter 11, together with identified perspectives. Here, the major contributions of the work are:

- The identification of recent developments in PCA model identification which have not been adopted yet by the process monitoring research field, such as the introduction of Maximum Likelihood estimators for PCA models and the development of tools that allow bias-variance tradeoff in PCA modelling.
- The evaluation of an experimental biological process unit with respect to targets of research on process supervision of biological processes, including the identification of vulnerable subsystems and prospectives for better design.
- Validation of Multi-way Principal Component Analysis (PCA) as a tool for process monitoring and diagnosis of SBR's for wastewater treatment as well as the indication of limitations resulting from its conventional use.
- The conception, real-time implementation and successful validation of a data-driven control strategy for phase length optimization of cyclic systems.
- Improvement of an existing method for qualitative analysis of trends in time series in relation to inflection point assessment.
- Validation of the applicability of qualitative analysis of trends for process monitoring, diagnosis and control of wastewater treatment systems.

1.4 Outline of this thesis

This thesis is organized in 5 parts. Part 1 includes a general introduction, including the positioning of the work within the context of wastewater treatment and problem statement (Chapter 1) and an overview of state-of-the-art techniques in control of SBR's for wastewater treatment (Chapter 2). In Part 2, Materials and methods, an overview of PCA-based methods for process monitoring is given (Chapter 3), next to the description of the experimental pilot-scale SBR setup and resulting data sets used later on in this work (Chapter 4). Part 3 bundles the reported work based on PCA modelling of the SBR system, including results obtained in relation to process monitoring (Chapter 5), diagnosis (Chapter 6) and control (Chapter 7). Developments in the field of qualitative analysis, including method improvement (Chapter 8) and applications in a data mining context (Chapter 9) and process control (Chapter 10) are included in Part 4. Part 5 contains the conclusions and perspectives (Chapter 11).

Chapter 2

Control of Sequencing Batch Reactor systems

In this chapter, an overview of reported strategies in monitoring, diagnosis and control of SBR's is given. First, Sequencing Batch Reactors (SBR's) for wastewater treatment are introduced in general. Thereafter, separate sections will deal with typically deployed sensors, actuators and control strategies in SBR's for wastewater treatment.

2.1 Sequencing Batch Reactors for wastewater treatment

Sequencing Batch Reactors are systems for wastewater treatment characterized by their cyclic operation. A generic cycle for any SBR essentially consists of the following phases: fill, react, decant and idle. The following aspects are typical for any SBR configuration:

- Influent and effluent flows are uncoupled. While physical and safety limits for the volume typically exist, the volume is not constant by default in an SBR configuration, in contrast to Continuously Stirred Tank Reactor (CSTR) and Plug Flow reactor (PFR) configurations.
- The react phase can be split into several subphases, each with proper physical and (bio)chemical conditions, to promote different reactions in a single reactor system.
- The design, order and length of each phase and subphase is not constrained a priori nor necessarily fixed.

As a result of the above characteristics, major benefits of SBR configurations over CSTR (Continuously Stirred Tank Reactor) configurations include flexibility in design and operation (Irvine et al., 1997). SBR configurations for wastewater treatment necessarily include a settling phase before the decant phase, so that the (larger part of the) biomass, functioning as a catalyst for all biochemical reactions, is kept in the system. As such, clarification of the treated wastewater takes place in the same reactor, in contrast to other designs of wastewater treatment plants, which require a separate settling tank or settler. A generic scheme of the cyclic operation of SBR's for wastewater treatment is given in Figure 2.1.

To keep the biomass in the system, decantation can never be complete. A portion of the reactor volume is therefore removed at the end of a cycle after which new raw wastewater is added in the next cycle. Depending on the design, phase scheduling and applied control laws, SBR's may allow removal of organic compounds and nutrients (nitrogen and phosphorus) by proper selection of microbial communities (Wilderer et al., 2001). A commonly reported example is the cultivation of phosphorus accumulating organisms (PAO's) by means of the so-called *feast and famine* principle. In this, the PAO's are exposed to high concentrations of organic compounds (typically reported as Chemical Oxygen Demand (COD)), which they

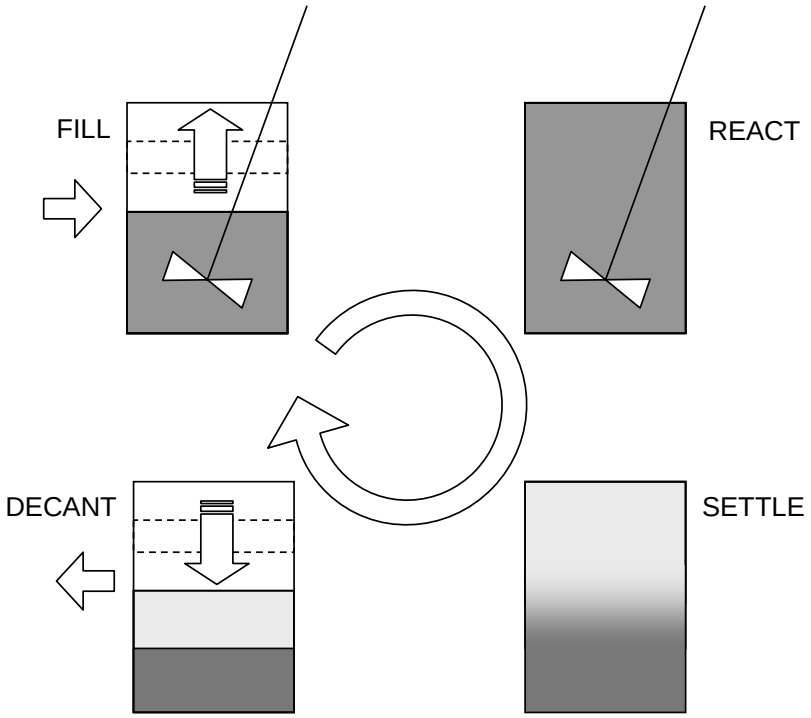


Figure 2.1: Schematic representation of the operation of SBR's for wastewater treatment.

accumulate inside their cells in the form of highly energetic PHA's (Poly-Hydroxy Alkanoates). In parallel, the PAO's release ortho-phosphate (PO_4^{3-}-P) into the bulk liquid during this *feast* state. This process is promoted under anaerobic conditions, i.e. oxygen and nitrate levels must be reduced to a minimum. In a following oxygenated or aerobic phase, (remaining) organic compounds are oxidized (to CO_2) and PAO's enter the so-called *famine* state. For maintenance and growth, PAO's oxidize the stored PHA (to CO_2) and take ortho-phosphate back in to produce the energy-carrying adenosinetriphosphate (ATP) molecule. The ortho-phosphate is stored as polyphosphate for later use in the next anaerobic *feast* period. By means of biomass (sludge) wastage after the latter process has occurred, effective phosphorus removal from the wastewater can be achieved.

Despite the reported benefits and the long-term existence of SBR configurations for wastewater treatment, conventional wastewater treatment plants are typically designed as CSTR tanks for biochemical conversion with additional settling tanks to clarify the treated water. SBR configurations have however gained increasing attention over the last decades (Cohen et al., 2003), possibly due to increased pressure on cost of land use, system liability and overall performance. Indeed, the capital costs of SBR's (i.e. construction) are lower compared to conventional designs (Cohen et al., 2003) and less land use is required (Andreottola et al., 2001).

Not unimportantly, wastewater treatment plants need to deal with changing flow rates and substrate concentrations irrespective of the process state, the qualities of the incoming water or needs for system maintenance. This situation is unlike production systems in other industries, where input flows can be adjusted in view of plant optimization, low quality inputs can be discarded when desired qualities are not met and processes can be put out of operation for maintenance without any but economic concerns. In view of changing flow rates and wastewater characteristics, SBR's offer a great deal of flexibility in dealing with those fluctuations. Processed volumes are allowed to change or can be controlled by buffer tanks, cycle and phase lengths can be adjusted to counter for changing susbtrate concentrations. In addition, operational costs (pumps, aeration) can be minimized as well. It is noted here that the question of how the flexibility of SBR's can be used to an optimal extent is still a matter of research. See for example in Demuynck et al. (1994); Chang and Hao (1996); Andreottola et al. (2001); Artan et al. (2001); Cohen et al. (2003); Traoré et al. (2005); Corominas et al. (2006); Marsili-Libelli (2006); Sin et al. (2006); Guo et al. (2007).

Advanced control of SBR's is challenged by the occurrence of abnormal events for which classic control systems are not designed. In order to enable automatic control of a system, a considerable amount of research has been aimed at detection and diagnosis of abnormal situations in SBR systems. The effective design of process monitoring and diagnosis tools allows to select among implemented control laws specific for normal and faulty situations. The design of supervisory control systems aims at the integration of techniques for process monitoring, diagnosis and control. Techniques which may eventually take part in supervisory control systems, such as tools for monitoring, diagnosis and advanced controllers for SBR's have been developed and tested in many studies, e.g. Stephanopoulos et al. (1997); Sarolta and Kinley (2001); Lennox and Rosén (2002); Ündey and Çinar (2002); Lee and Vanrolleghem (2003, 2004); Rubio et al. (2004); Ruiz et al. (2004); Lee et al. (2005); Ciappelloni et al. (2006); Yoo et al. (2006a,b).

2.2 On-line measurement technologies for wastewater treatment

A critical aspect of any wastewater treatment system is the selection of deployed sensors. Sensors which measure effluent quality variables such as ammonia, nitrate and phosphate concentrations are commercially available but are -as yet- often found too expensive for small wastewater treatment plants. Cost-effective sensors which provide information on the system's status or evolution are therefore needed (Serralta et al., 2004). Dissolved oxygen, pH and ORP measurements are the cheapest and most commonly reported on-line measurements. Not surprisingly, they often come as standard with any wastewater treatment plant. Less common is the use of conductivity measurements. More advanced technologies, such as ultraviolet-visible (UV-VIS) light spectrometry are less commonly reported in literature, but are getting rapidly increasing attention. In the following section, proposed on-line sensor measurements and their effectiveness in relation to monitoring of biological wastewater treatment are discussed.

2.2.1 Dissolved oxygen (DO)

Of all variables measured on-line, dissolved oxygen (DO) shares the longest history in use for control. On-line oxygen control has been proven effective in reducing the aeration costs of watewater treatment plants and is almost standard for any plant. Its practical use in SBR's is naturally limited to aerobic phases, i.e. when the oxygen concentration is intentionally non-zero. As the oxygen conecentration is the net result of oxygen supply (by aeration) and oxygen consumption by (biochemical) substrate oxidation, the oxygen supply and oxygen concentration jointly deliver information on the speed of oxidation reactions. The latter reactions are notably the (heterotrophic) oxidation of organic compounds, the (autotrophic) oxidation of ammonia to nitrate (via nitrite) and (heterotrophic) phosphorus uptake.

When a fixed gas flow rate is used, oxygen concentrations will typically show a large rise when all biodegradable substrates (COD, NH_4^+) are oxidized. The corresponding inflection point (see Figure 2.2) is referred to as the DO break point, DO flex point, α_{O_2}-point or α_{OUR}-point (Wareham et al., 1993; Plisson-Saune et al., 1996; Mauret et al., 2001; Puig et al., 2005). Importantly, when considerable amounts of organic compounds are present in the liquid at the start of an

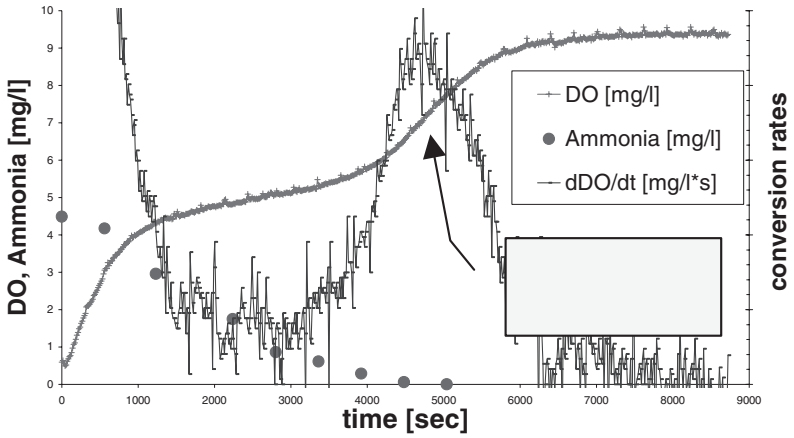


Figure 2.2: Dissolved oxygen and ammonia concentration measured in a laboratory SBR during the aerobic phase. Adopted from Cohen et al. (2003)

aerobic phase, this may lead to an additional inflection point occurring before the (targeted) DO break point as carbon compounds are oxidized much faster than ammonia. As a result, this earlier inflection point may erroneously be identified as the DO breakpoint (Cohen et al., 2003).

Irrespective of how the flow rate is controlled, an oxygen mass balance can be used to derive the rate of biological oxygen consumption, expressed as Oxygen Uptake Rate (OUR) (see Figure 2.3). In Demuynck et al. (1994); Surmacz-Gorska et al. (1996); Vives et al. (2003); Puig et al. (2005) the OUR is determined when aeration is put off (e.g. when an on-off aeration controller is applied). Note that the data used for OUR estimation needs to be sampled soon enough after switching the aeration off so that the DO concentration is high enough (i.e. information-rich). In the periods where aeration is off, biological oxygen uptake is (theoretically) the only ongoing process. As such, the OUR can be estimated by simple estimation of the derivative of the oxygen level during these off-phases. A drawback of this approach is that considerable time steps between consecutive estimates of the OUR are induced, thereby possibly rendering the timing of control actions based on OUR suboptimal. As noted already, the use of oxygen concentration measurements or related mass balances are limited to aerated phases of an SBR cycle. The oxygen measurement does not lend itself to track or control the process status under anaerobic or anoxic conditions.

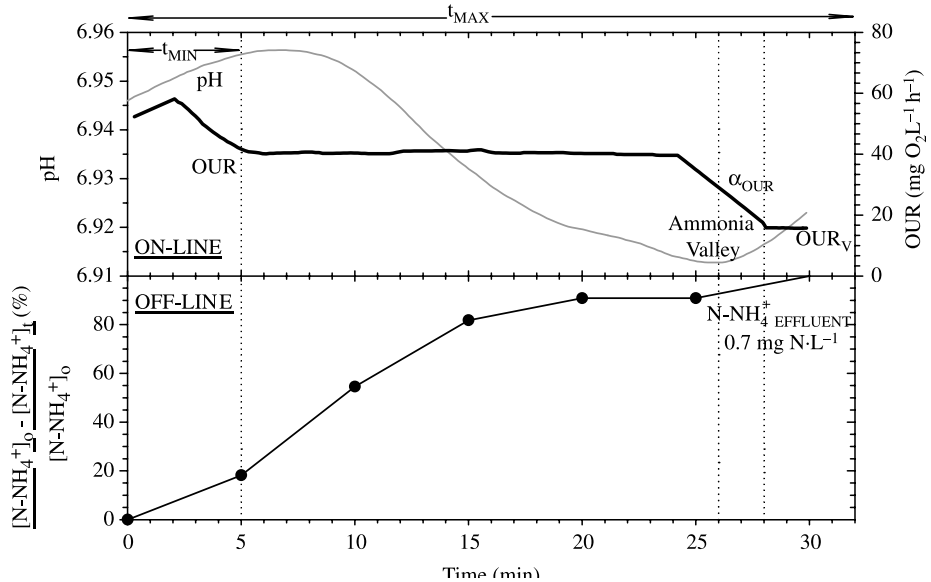


Figure 2.3: Aerobic phase of an SBR system. Profiles of OUR, pH and ammonia nitrogen. The α_{OUR}-point and ammonia valley are indicated. Adopted from Puig et al. (2005)

2.2.2 pH

The dynamics of pH data in SBR wastewater treatment systems are well understood and have effective meanings in both aerobic and anoxic or anaerobic conditions of wastewater treatment processes. The interpretation for these different conditions is given seperately.

2.2.2.1 Aerobic conditions

Under aerobic conditions, processes that reportedly influence the pH are CO_2-production and -stripping and the first step of nitrification, i.e. oxidation of ammonia to nitrite. Accumulation of phosphate into PAO's may affect the phosphate buffer system and thus the pH as well, yet, a sufficiently complete explanation in relation to phosphate uptake has not been given. In typical situations, a dominant

effect of CO_2-stripping, i.e. an upward trend in pH, is observed at the beginning of an aerobic phase (see Figure 2.3). As soon as the stripping rate lowers due to the exhaustion of CO_2 and the rate of ammonia oxidation increases, a net acidifying effect will result. When the oxidation of ammonia to nitrite is completed, acidification stops as well and an increase in pH is observed due to continued CO_2-stripping. The occurrence of the described minimum in pH, corresponding to the endpoint of ammonia oxidation, is denoted as the *ammonia valley*. The detection of the latter event has been the core of several control strategies, see e.g. Andreottola et al. (2001); Chen et al. (2004); Kishida et al. (2004); Li et al. (2004).

2.2.2.2 Anoxic/anaerobic conditions

Under anoxic conditions, i.e. with oxidized compounds such as nitrite and nitrate present but without presence of oxygen, biological conversion of nitrite and nitrate to nitrogen gas occurs. During the reported reduction processes, hydroxyl-ions (OH^-) are produced leading to an increase of the pH. When the latter processes are complete, the pH increase halts. The reported maximum in the pH profile is called the nitrate apex (see Figure 2.4). As continuation of anoxic or anaerobic conditions is uninteresting (assuming phosphate release is complete or does not need to continue), control of the anoxic phase length is then possible (Andreottola et al., 2001; Spagni et al., 2001; Wang et al., 2004). The nitrate apex is however less commonly used compared to the nitrate knee in oxidation reduction potential (ORP) time series (see below).

2.2.3 Oxidation Reduction Potential (ORP)

The Oxidation Reduction Potential (ORP) can be regarded as a measure for the amount of readily available electrons for oxidation in the bulk liquid of a system. More electrons lead to a higher ORP value or a more oxidized state, less electrons result in a lower ORP value or a more reduced state of the system. As such, ORP sensors relate to oxidation-reduction reactions similarly to pH sensors relating to acid-base reactions. As for the pH, the expected behaviour of ORP sensors is presented here separately for aerobic and anoxic/anaerobic conditions.

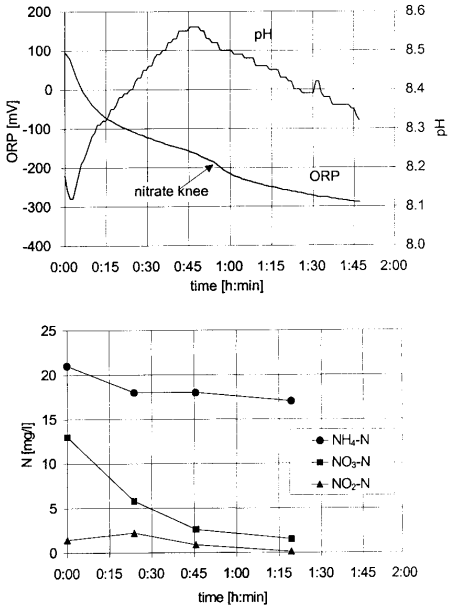


Figure 2.4: Anoxic phase of an SBR system. Profiles of ORP, pH and ammonia (NH_4^+), nitrite (NO_2^-) and nitrate (NO_3^-) nitrogen. The nitrate apex in the pH profile and the nitrate knee in the ORP profile are visible. Adopted from Andreottola et al. (2001)

2.2.3.1 Aerobic conditions

Under aerobic conditions, oxygen serves as an electron acceptor of several oxidation processes, including oxidation of organic carbon (to CO_2), ammonia (to nitrite) and nitrite (to nitrate). Also, for phosphorus uptake, PAO's consume oxygen. As long as oxidation processes continue, the ORP signal is expected to rise. A remarkable rise of the ORP value, observed as an inflection point, occurs when the nitrite-nitrate buffer is breached, i.e. when virtually all nitrite is converted to nitrate. This inflection point is called the nitrogen break point (Ra et al., 1999). If the second step of nitrification, i.e. nitrite oxidation, is faster than the first step, i.e. ammonia oxidation, the nitrogen break point in the ORP signal is observed shortly after the ammonia valley in the pH signal. It was observed for oxygen-based inference, that similar break points can occur for both the end of carbon oxidation and ammonia/nitrite oxidation. This effect plays a role in the interpretation

of the ORP signal as well. Indeed, complete oxidation of biodegradable organic compounds may result in a rise of the ORP signal, possibly confounded with the nitrogen break point (Peddie et al., 1990). For this reason, the ammonia valley (in pH) may be preferred in a univariate inference system, even if the ammonia valley indicates the end of the ammonia oxidation and not the end of the nitrite oxidation. Phosphate uptake by PAO organisms is reported to concur with increasing ORP values (Paul et al., 1998; Lee et al., 2001, 2004a) but the use of the ORP signal in relation to the phosphorus uptake process has not been reported in practice.

2.2.3.2 Anoxic/anaerobic conditions

Under anoxic conditions, reduction processes, including the reduction of nitrite and nitrate to nitrogen gas, result in the decrease of the ORP values. As soon as the nitrate reduction process is completed, the ORP value will decrease faster, due to the induced absence of the nitrate ORP-buffer (see Figure 2.4). This point is denoted as the nitrate knee (Peddie et al., 1990; Demuynck et al., 1994; Wareham et al., 1994). Lee et al. (2004a) link ORP signals with the phosphate release process but this has not been used in practice.

2.2.4 Conductivity

Measurements of conductivity, while cheap, are not reported often in practice. While the sensor often comes standard in the design of WWTP's, little use has been reported in practice. However, conductivity has been shown to be closely related to the biological phosphorus release and uptake processes (Maurer and Gujer, 1995; Serralta et al., 2004; Aguado et al., 2006). During phosphorus release, orthophosphate ions are released by the biomass into the bulk liquid. Simultaneously, the bacteria release ions (K^+, Mg^{2+}). As a result, a net increase of ions in the bulk liquid results, hereby reducing its electrical resistance, consequently leading to a higher conductivity measurement. Conversely, phosphorus uptake under aerobic conditions, results in the uptake of phosphate ions, balanced by incorporated ions (K^+, Mg^{2+}). As a result, the conductivity of the bulk liquid decreases. It is noted that studies concerning conductivity are limited in general and to processes with dominant phosphate-related processes.

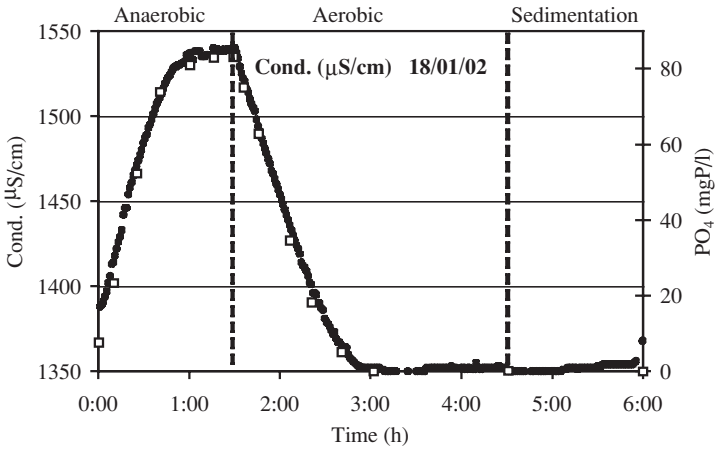


Figure 2.5: Complete cycle of an SBR system for phosphorus removal. Profiles of conductivity and phosphorus. Adopted from Serralta et al. (2004)

2.2.5 Process and effluent quality variables.

Two distinct types of sensors for quality variables, such as total suspended solids (TSS), ammonia, nitrite and nitrate are available on the commercial market, being on-line measurements and spectrophotometric measurements.

2.2.5.1 On-line sensors for quality variables

On-line measurement of ammonium, nitrate, orthophosphate and total suspended solids has proven to be technically feasible in practice. However, compared to DO, ORP, pH sensors and even optical nutrient & TSS sensors, the use of these sensors implies a larger capital cost as well as more intense maintenance. For example, filters, necessary for automated sample preparation, are widely known to clog and the sensors need be charged with sufficient and stable chemical solutions (e.g. for colometric reactions). Due to the high cost and high requirements for maintenance, on-line sensors for quality variables are not frequently used.

2.2.5.2 Ultraviolet-visible (UV-VIS) light spectrometry

Spectrometry on the basis of absorbance of light in the Ultraviolet-Visible (UV-VIS) range (wavelengths of 200 to 750 nm) is a relatively recent technology for on-line monitoring of wastewater treatment plants. In contrast to lab measurements, spectrometers can be used for in situ (i.e. in direct contact with the bulk liquid) measurement of key quality variables (e.g. Total Suspended Solids, nitrate nitrogen) (Langergraber et al., 2003, 2004; Rieger and Siegrist, 2006). In addition, much more frequent measurements are possible and multiple effluent quality variables can be measured by means of a single sensor. UV-VIS spectrometry has been applied for intensive measurement campaigns in sewer systems and wastewater treatment systems (Langergraber et al., 2004). Key to the use of UV-VIS spectrometry for monitoring is the establishment of the relationship between absorbance spectra and the targeted quality variables. The calibration and application of UV-VIS spectrophotometers has gained considerable attention lately (Gruber and Bretrand-Krajewski, 2005; van den Broeke et al., 2006, 2007; Maribas et al., 2007; Rieger et al., 2007; Torres and Bertrand-Krajewski, 2007; Winkler et al., 2007).

2.3 Control strategies for SBR's for wastewater treatment

Adjustment of water flow rates (influent and effluent), adjustment of the aeration intensity and addition of carbon source (for carbon-limited wastewaters) have been used as effective control handles in SBR's for wastewater treatment. In what follows, reported control strategies are reviewed. First, oxygen level controllers and carbon dosage are reported. Then, attention is given to phase scheduling of SBR's.

2.3.1 Oxygen control

Control during aerated phases of an SBR system is limited to aeration control in most cases. While oxygen control is not a necessity, considerable energy savings may be obtained by controlling the oxygen level to a fixed value. To this end, one strategy simply consists of putting compressors or aeration mixers on and off according to the measured values of DO (Traoré et al., 2005). Practically, the aeration system is put on when the DO level reaches a set minimum and is put off again as

soon as a set maximum is reached, resulting in a so called bang-bang controller. If the gas flow rate or mixing intensity can be adjusted in a continuous fashion, more advanced control can be opted for. Classic linear Proportional-Integral-Derivative control has however been shown to be suboptimal (Traoré et al., 2005) as the system properties (e.g gain and time constants) may change as oxidation reactions proceed or have been completed. Traoré et al. (2005) therefore propose the use of fuzzy control laws to allow adaptation of the control system to changing dynamic behaviour. Other approaches to circumvent this problem, including gain-scheduling, are given in Olsson and Newell (1999).

2.3.2 Control of carbon dosage

Wastewaters with a low ratio of organic compounds to nitrogen compounds, expressed as the C/N ratio, often present a challenge for wastewater treatment. Indeed, due to a low carbon content of the wastewater, complete reduction of oxidized nitrogen components (nitrite, nitrate) is difficult or impossible to achieve. In order to achieve complete reduction of the latter species, addition of an artificial carbon source can be considered. Products that are reported for this use are methanol, ethanol, acetate and wastewater of breweries (Peng et al., 2002; Chen et al., 2004; Kishida et al., 2004). The addition can be implemented at one time point in the batch run, at regular intervals or in a continuous fashion. Importantly, carbon dosage represents a considerable cost and should ideally not exceed the amount necessary to complete the reduction processes. Indeed, carbon source which is not consumed in the reduction process needs to be oxidized, thereby inducing additional aeration costs. It is in this context that real-time adjustment of carbon dosage is aimed for by Kim and Hao (2001) and Kishida et al. (2004). To this end, the detection of the nitrate knee (see Section 2.2.3.2) serves as stopping criterion for carbon dosage.

2.3.3 Phase scheduling

While SBR configurations are renown for their flexibility in operation, in particular for the flexibility in the scheduling of constituting phases, optimal adjustment of the applied schedule is still a matter of research. It is noted here that the continuation of phases beyond completing the desired biochemical conversion may not only represent excessive costs but may turn down the performance of plants as well. Wilderer

et al. (2001) report that excessive aeration negatively affects the settling properties of the biomass. This may lead to ineffective retention of biomass in the system and consequently result in worsened plant performance. On-line control aimed at optimizing the length of phases is typically conceived as a detection of the desired completion of biochemical reactions which is then linked to the shutdown of the ongoing phase and start of the next. Strategies based on oxygen measurements or OUR evaluation can be found in Demuynck et al. (1994); Surmacz-Gorska et al. (1996); Johansen et al. (1997); Klapwijk et al. (1998); Vives et al. (2003); Third et al. (2004); Balslev et al. (2005); Bisschops et al. (2006); Corominas et al. (2006); Guisasola et al. (2006); Shaw and Falrey (2007). Phase length optimization on the basis of endpoint detection on the basis of pH or ORP signals can be found in numerous articles (Demuynck et al., 1994; Wouters-Wasiak et al., 1994; Hao and Huang, 1996; Charpentier et al., 1998; Paul et al., 1998; Zipper et al., 1998; Ra et al., 1999; Cho et al., 2001; Kim and Hao, 2001; Yu et al., 2001; Li et al., 2004; Wang et al., 2004; Balslev et al., 2005; Cecil, 2007; Guo et al., 2007). Puig et al. (2005) combine extracted information on OUR and ORP for reaction endpoint detection while Peng et al. (2004) combine DO- and pH-based inferences. Andreottola et al. (2001); Ma et al. (2006); Marsili-Libelli (2006) provide algorithms based on joint analysis of DO, ORP and pH signals. In Puig et al. (2005); Corominas et al. (2006) a minimum absolute value for OUR is set to trigger the end of the aerobic phase, while reaching a set minimum ORP-value triggers the end of the anoxic phase. The latter thus chose not to identify reported inflection points explicitly. Motivations for this are that (1) ambiguity exists in the identified inflection points, as discussed above and (2) measurement noise impedes the direct and undelayed assessment of the inflection points. Model-based phase length optimization has been proposed and evaluated as well (Sin et al. (2006)), but has so far been limited to off-line calibration and optimization steps.

2.4 Concluding remarks

From the given review of state-of-the-art approaches in control for SBR's it can be concluded that most control strategies are based on the detection of certain endpoints in indirect measurements such as DO, OUR, pH or ORP. Conductivity measurement, yet simple and cheap, is covered minimally in literature and is neither applied in practice. Reported applications are typically based on rule-based evaluations, in turn based on available knowledge or experience. Detection of endpoints in indirect measurements always requires a minimum of intelligence in the detec-

tion system to enable effective discrimination between true and unreal key points in the analyzed trajectory. Upon future developments and increased availability on the commercial market, one may expect that the need for such intelligence in control may weaken as hidden variables (e.g. TSS, NH_4^+-N, NO_2^--N, NO_3^--N, PO_4^{3-}-P) are unveiled by spectroscopic technology. It is noted by (Steyer et al., 2006) that the best of opportunities may however be met when both advanced instrumentation as well as advanced control are put in use.

Part II

Materials and methods

Chapter 3

Methods

3.1 Introduction

In this chapter, methods used throughout this PhD are presented. Following this introduction, conventions with respect to notation are given. Then, Principal Component Analysis (PCA) is explained first and extensions of PCA for process monitoring are reviewed. The latter review is conceived as an extensive overview of available literature concerning PCA-based process monitoring techniques. Techniques used in this work are explained in detail while notes on other techniques are restricted. Section 3.3.1 (standard PCA), Section 3.3.3.1 (Multi-way PCA) and Section 3.3.2.3 (Mixture PCA) provide necessary details on the methods used in Chapter 5. In Section 3.4, Fuzzy C-means Clustering (FCM) is explained. This method is used in Chapter 6, in addition to Multi-way PCA (Section 3.3.3.1). In order to understand Chapter 7, the reader may limit himself to Section 3.3.1.5. In Section 3.5 an overview of methods for qualitative analysis of time series is given. Two methods are compared. and one of them is selected for further improvement (Chapter 8) and applications (Chapters 9 and 10). For a good understanding of the respective chapters, Sections 3.5.1 to 3.5.5 are considered essential.

3.2 Notations

Throughout this work, the following notations are used:

cov $()$: variance-covariance operator

E $()$: expected value operator

$f(x)$: function in x

tr $()$: trace operator

var $()$: variance operator

x: scalar

$\mathbf{x}$: column vector

$\mathbf{X}$: (2-dimensional) matrix

$\underline{\mathbf{X}}$: 3-dimensional matrix

x_i: scalar with coordinate i in vector $\mathbf{x}$

$x_{i,j}$: scalar with coordinates (i, j) in matrix $\mathbf{X}$

$\underline{x}_{i,j,k}$: scalar with coordinates (i, j, k) in matrix $\underline{\mathbf{X}}$

$\mathbf{x}_{i,.}$: i^{th} row vector in matrix $\mathbf{X}$

$\mathbf{x}_{.,j}$: j^{th} column vector in matrix $\mathbf{X}$

$\mathbf{X}_{i,.,.}$: i^{th} *horizontal* slab in matrix $\underline{\mathbf{X}}$

$\mathbf{X}_{.,j,.}$: j^{th} *lateral* slab in matrix $\underline{\mathbf{X}}$

$\mathbf{X}_{.,.,k}$: k^{th} *frontal* slab in matrix $\underline{\mathbf{X}}$

$\mathbf{a}^T$, $\mathbf{A}^T$:transpose of vector $\mathbf{a}$, matrix $\mathbf{A}$

3.3 Principal Component Analysis (PCA)

3.3.1 Standard Principal Component Analysis

Principal Component Analysis (PCA) is a data mining technique originally devised and still mostly used to express the information contained in a large number of variables by means of a smaller number of variables, called principal scores. The latter variables supposedly capture the larger part of the information contained in the original variables. As such, PCA can be used for data dimension reduction by converting so-called data-rich and information-poor data sets into new data for which the information to data ratio is higher.

Extensive introductions to PCA can be found in Jackson (1991) and Joliffe (2002). Next to dimension reduction, PCA models can also be used for process monitoring and regression, as will be shown.

In the sections that follow, some necessary definitions are given first whereafter an illustrative example is defined. This is followed by basic notes on PCA, including model identification procedures and illustrative results. Then, the use of PCA models for process monitoring and regression are reviewed.

3.3.1.1 Definitions

Variables and measurements. Consider a process for which a set of noise-free variables are defined. Define the two-dimensional matrix $\mathbf{Z}$ which contains the values of the M ($j = 1..M$) variables of this process, $z_{i,j}$, in N samples ($i = 1..N$):

$$\mathbf{Z} = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,j} & \cdots & z_{1,M} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,j} & \cdots & z_{2,M} \\ \vdots & \vdots & & \vdots & & \vdots \\ z_{i,1} & z_{i,2} & \cdots & z_{i,j} & \cdots & z_{i,M} \\ \vdots & \vdots & & \vdots & & \vdots \\ z_{N,1} & z_{N,2} & \cdots & z_{N,j} & \cdots & z_{N,M} \end{bmatrix} \tag{3.1}$$

Measurements are taken of these variables, hereby defining the matrix of the measured state variables as follows:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,j} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} & \cdots & x_{2,M} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i,1} & x_{i,2} & \cdots & x_{i,j} & \cdots & x_{i,M} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,j} & \cdots & x_{N,M} \end{bmatrix} \tag{3.2}$$

The relation between each measurement value and the corresponding value of the measured variable can generally be written as follows:

$$x_{i,j} = z_{i,j} + e_{i,j} \tag{3.3}$$

in which $e_{i,j}$ defines the deviation between the value of the state variable and the corresponding measurement value, further referred to as measurement errors. In the general case, only the values of $x_{i,j}$ are known to the experimenter. In what follows, it will be assumed that the means of the measurement errors are zero:

$$\mathrm{E}\left(e_{i,j}\right) = 0 \tag{3.4}$$

For statistical inferences, such as demonstrated in Section 3.3.1.5, satisfaction of additional requirements is necessary. More specifically, it is required that the process variables, $z_{i,j}$, follow a multivariate normal distribution and that the measurement errors, $e_{i,j}$, are independent and drawn from the same normal distribution:

$$\mathbf{z}_{i,\cdot} \sim MVN(\mu, \mathbf{\Sigma}) \tag{3.5}$$

$$e_{i,j} \sim N(0, \sigma_e) \tag{3.6}$$

These more restrictive requirements will only be assumed in Section 3.3.1.5.

Centering and scaling. A centered and scaled data matrix, $\tilde{\mathbf{X}}$, is obtained by application of the following transformation to each element of $\mathbf{X}$:

$$\tilde{x}_{i,j} = \frac{x_{i,j} - m_j}{s_j} \tag{3.7}$$

where m_j and s_j are predefined or estimated scalars for each of the variables (j=1..M). Typically, the values for m_j are set to the (estimated) means of the respective measurements:

$$m_j = \frac{1}{N} \cdot \sum_{i=1}^{N} x_{i,j}, \qquad j = 1..M \tag{3.8}$$

As a result the measurements of each variable (column) in the scaled data matrix have zero mean (*mean-centered* data). It is also common to set the value of s_j to 1 for all variables:

$$s_j = 1, \qquad j = 1..M \tag{3.9}$$

or, alternatively, to the (estimated) standard deviations obtained as follows:

$$s_j = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (x_{i,j} - m_j)^2}, \qquad j = 1..M \tag{3.10}$$

By means of the latter choice the total variance of the scaled measurements of each variable becomes one, i.e. the data are scaled to unit variance. If the values for m_j are estimated on the basis of the measurements themselves as by equation 3.8, equation 3.10 needs to be redefined as:

$$s_j = \sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_{i,j} - m_j)^2}, \qquad j = 1..M \tag{3.11}$$

The scaling parameters can be grouped into two vectors, $\mathbf{m}$ and $\mathbf{s}$, as follows:

$$\mathbf{m} = \begin{bmatrix} m_1 \; m_2 \; \cdots \; m_M \end{bmatrix} \tag{3.12}$$

$$\mathbf{s} = \begin{bmatrix} s_1 \; s_2 \; \cdots \; s_M \end{bmatrix}$$

Following the scaling, the covariance matrix of the scaled data matrix, $\mathbf{S}$, is defined as follows:

$$\mathbf{S} = \text{cov}\left(\tilde{\mathbf{X}}\right) = \frac{\tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}}}{N} \tag{3.13}$$

In case equations 3.8 and 3.9 are used then the scaling procedure is referred to as mean centering. By definition, the resulting covariance matrix of the scaled data set $\mathbf{S}$ is the covariance matrix of the original data set. The use of equations 3.8 and 3.11 is referred to as auto-scaling. $\mathbf{S}$ is then the correlation matrix of the original data set. As stated above, both options are common for typical PCA applications. Limited attention is generally given to the effect of scaling procedures. An exception is made by works in the context of errors-in-variables regression. More details on the latter are given in Section 3.3.1.6.

3.3.1.2 Examples

In order to demonstrate the use of PCA, two illustrative examples are introduced here. Both examples will be used throughout the following sections. The first example is simulated in particular to demonstrate the geometrical properties of PCA models. The second example is simulated to illustrate how PCA can effectively be used for dimension reduction.

Example 1. The following example was devised especially to demonstrate the geometrical aspects of PCA modelling. Consider a system containing three connected tubes without reservoir carrying water flows as depicted in Figure 3.1. Denote the respective flows at a given time instant i, $z_{i,j}$ ($j = 1..3$). If the signs of ingoing (outgoing) flows in the tubes are defined as positive (negative), then the mass conservation law, valid at all times i, can be expressed as follows:

$$z_{i,1} + z_{i,2} + z_{i,3} = 0 \tag{3.14}$$

Consider that measurements of the flow rates are taken simultaneously (equation 3.3). The measurement errors are independent and identically distributed (i.i.d.) and are drawn from a Gaussian distribution with zero mean and a respective standard deviation for each variable, ϵ_j:

$$e_{i,j} \sim N(0, \epsilon_j), \qquad i = 1..N, j = 1..3 \tag{3.15}$$

A data set was simulated with N samples. The values for $z_{i,1}$ and $z_{i,2}$ are drawn randomly from two independent Gaussian distributions with respective means μ_1 and μ_2 and standard deviations σ_1 and σ_2. The values for $z_{i,3}$ follow from equation 3.14.

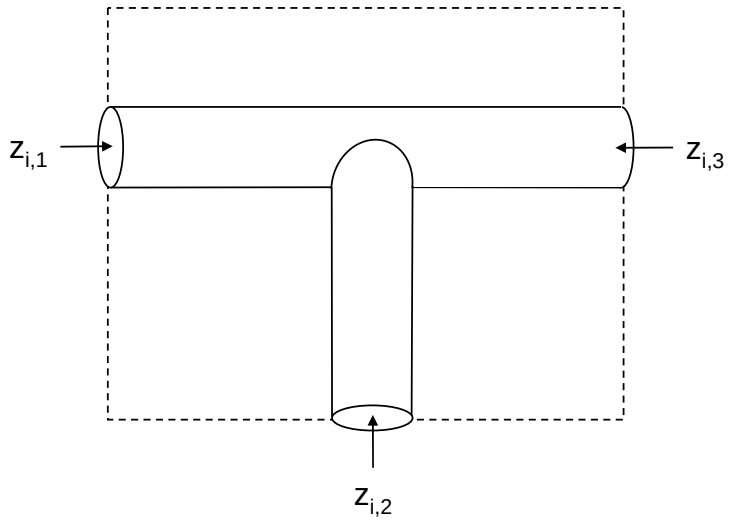


Figure 3.1: Scheme of the simulated system for Example 1.

The following parameter settings are used in all graphs for this example:

$$N = 300, \qquad \sigma_1 = 10,$$
$$\mu_1 = 50, \qquad \sigma_2 = 10,$$
$$\mu_2 = -20, \qquad \epsilon_j = 1, \quad j = 1..3$$

The simulated data are plotted in Figure 3.2 as well as the plane defined by equation 3.14. One can observe that the simulated data expectedly lie close to the latter plane. In fact, by definition, the error-free samples $(z_{i,1}, z_{i,2}, z_{i,3})$ lie on this plane. The deviations between the measurement samples $(x_{i,1}, x_{i,2}, x_{i,3})$ and their error-free value counterparts $(z_{i,1}, z_{i,2}, z_{i,3})$ are solely due to the introduced measurement error.

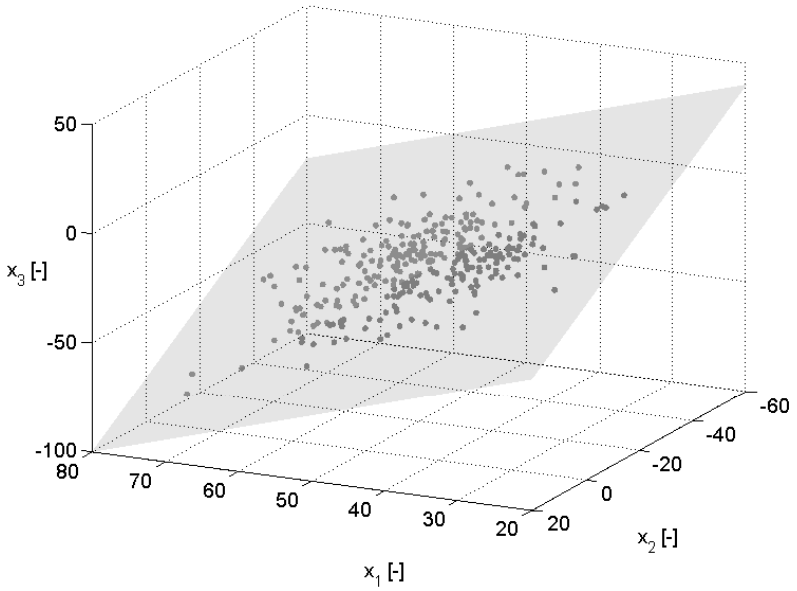


Figure 3.2: Example 1 – Simulated data. The shown plane represents the relationship between the error-free variables underlying to the measurements.

Example 2. A second example was simulated to demonstrate the use of PCA as a tool for dimension reduction and to illustrate PCA identification techniques. Consider the random generation of G Gaussian and independently distributed variables with N repetitions, $\eta_{i,g}, g = 1..G, i = 1..N$:

$$\eta_{i,g} \sim N(0, \zeta_g), \qquad g = 1..G, i = 1..N \tag{3.16}$$

The two variables are linearly transformed into M ($M > G$) new variables, $z_{i,j}$:

$$z_{i,j} = \sum_{g=1}^{G} \eta_{i,g} \cdot \gamma_{g,j}, \qquad i = 1..N, j = 1..M, g = 1..G \tag{3.17}$$

where:

$$\gamma_{g,j} = \frac{1}{\tau_g \cdot \sqrt{2\pi}} \cdot h^{\frac{-j-\nu_g}{2\tau_g^2}}$$

ν_g, τ_g: scalar parameters

It may be verified that the latter variables satisfy the following (set of) equation(s):

$$0 = \mathbf{z}_{i,\cdot} \cdot \left(I - \mathbf{\Gamma}^T \cdot \left(\mathbf{\Gamma}^T \right)^+ \right) \tag{3.18}$$

where:

$$\mathbf{\Gamma} = \left[\gamma_{1,\cdot}^T \ \gamma_{2,\cdot}^T \ \cdots \ \gamma_{G,\cdot}^T \right]^T \tag{3.19}$$

$$\mathbf{z}_{i,\cdot} = i^{\text{th}} \text{ row of } Z$$

$$\gamma_{g,\cdot} = i^{\text{th}} \text{ row of } \mathbf{\Gamma}$$

$$\left(\mathbf{\Gamma}^T \right)^+ \text{: the Moore-Penrose pseudoinverse of } \mathbf{\Gamma}^T \tag{3.20}$$

If the row vectors of $\mathbf{\Gamma}$ are linearly independent, then the Moore-Penrose pseudoinverse of $\mathbf{\Gamma}^T$, $\left(\mathbf{\Gamma}^T \right)^+$, in the formula above can be computed as follows:

$$\left(\mathbf{\Gamma}^T \right)^+ = \left(\mathbf{\Gamma} \cdot \mathbf{\Gamma}^T \right)^{-1} \cdot \mathbf{\Gamma} \tag{3.21}$$

Measurements of the variables, $z_{i,j}$, are taken simultaneously (equation 3.3). The measurement errors are assumed independent and identically distributed (i.i.d.) and are drawn from a Gaussian distribution with zero mean and a standard deviation for each variable, ϵ_j:

$$e_{i,j} \sim N(0, \epsilon_j), \qquad i = 1..N, j = 1..M \tag{3.22}$$

The following parameter settings are used for this example:

$$\begin{aligned} G &= 2, & \nu_1 &= 60, \\ M &= 256, & \nu_2 &= 170, \\ N &= 100, & \tau_1 &= 30, \\ \zeta_1 &= 5, & \tau_2 &= 40 \\ \zeta_2 &= 6, & \epsilon_j &= 0.005, \quad j = 1..M \end{aligned}$$

Simulated variables are shown in Figure 3.3. Additional *pure* samples are shown for the following values for the underlying variables, η_j:

- Pure sample 1: $\eta_1 = 3 \cdot \zeta_1$ and $\eta_2 = 0$
- Pure sample 2: $\eta_1 = 0$ and $\eta_2 = 3 \cdot \zeta_2$

These pure samples result in Gaussian bell shapes (see Figure 3.3) as may be derived from the provided equations as well. Each other simulated sample (mixture) reflects a linear combination of the pure Gaussian bell curves determined by the underlying variables, η_g. In this respect, the provided example reflects behaviour of spectral measurements. Note that this analogy is not complete as negative values are allowed. The simulated measurements are plotted in Figure 3.4. In both graphs, the same simulated mixture is highlighted. This sample will be used further on for interpretation of the identified PCA model.

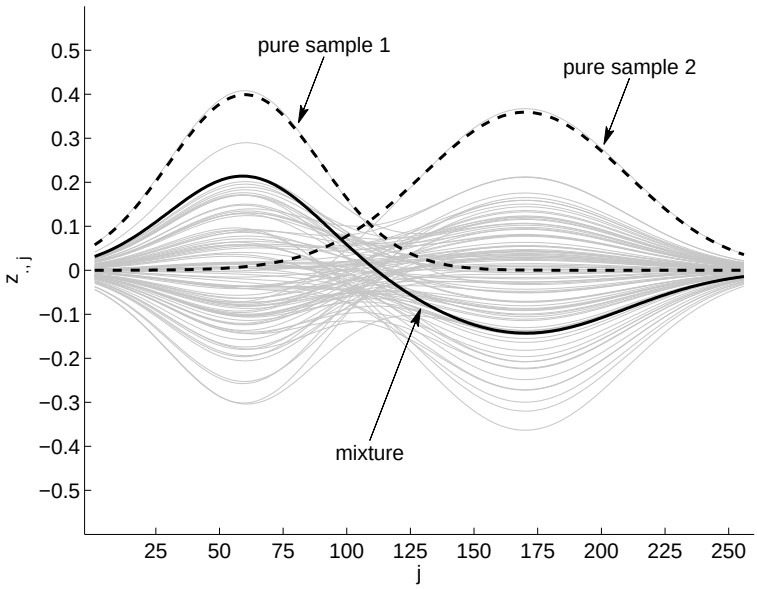


Figure 3.3: Example 2 – Simulated error-free variables, $z_{.,j}$. Each line represents a single sample. Pure samples and one mixture sample are highlighted.

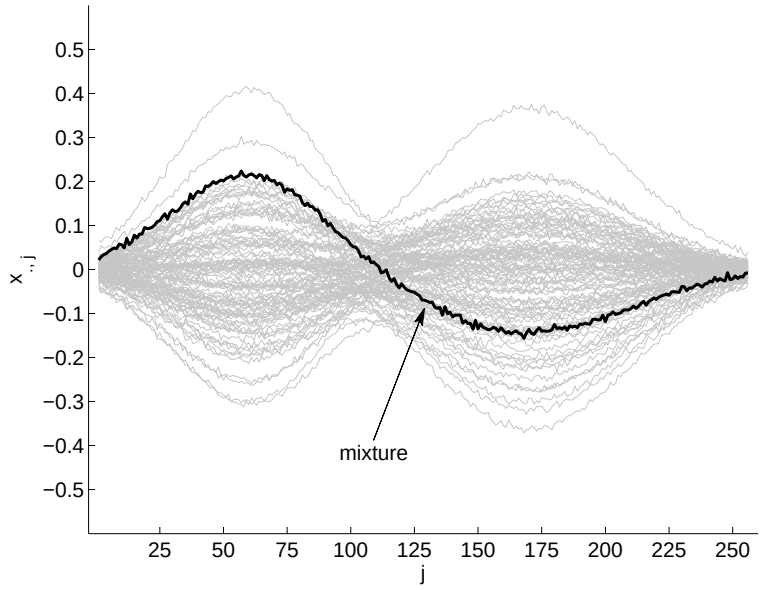


Figure 3.4: Example 2 – Simulated measurements, $x_{.,j}$. One mixture sample is highlighted.

3.3.1.3 Principal component identification

Consider a scaled data matrix, $\tilde{\mathbf{X}}$, as defined by equations 3.2 and 3.7 and write a set of C ($C \leq \max(M, N)$) new variables as linear combinations of the scaled variables as follows:

$$\begin{aligned} t_{i,1} &= \tilde{\mathbf{x}}_{i,.} \cdot \mathbf{p}_{.,1} = \tilde{x}_{i,1} \cdot p_{1,1} + \tilde{x}_{i,2} \cdot p_{2,1} + \cdots + \tilde{x}_{i,2} \cdot p_{M,1} \\ t_{i,2} &= \tilde{\mathbf{x}}_{i,.} \cdot \mathbf{p}_{.,2} = \tilde{x}_{i,1} \cdot p_{1,2} + \tilde{x}_{i,2} \cdot p_{2,2} + \cdots + \tilde{x}_{i,2} \cdot p_{M,2} \\ &\vdots \\ t_{i,C} &= \tilde{\mathbf{x}}_{i,.} \cdot \mathbf{p}_{.,C} = \tilde{x}_{i,1} \cdot p_{1,C} + \tilde{x}_{i,2} \cdot p_{2,C} + \cdots + \tilde{x}_{i,2} \cdot p_{M,C} \end{aligned} \tag{3.23}$$

Given the covariance matrix of the scaled data matrix as defined in equation 3.13, one can write the variances and covariances of these new variables as follows (Johnson and Wichern, 2002):

$$\text{var}\left(\mathbf{t}_{\cdot,\mathbf{c}}\right) = \mathbf{p}_{\cdot,c}^{T} \cdot \mathbf{S} \cdot \mathbf{p}_{\cdot,c} \qquad c = 1, ..., C \tag{3.24}$$

$$\text{cov}\left(\mathbf{t}_{\cdot,\mathbf{b}}, \mathbf{t}_{\cdot,\mathbf{c}}\right) = \mathbf{p}_{\cdot,b}^{T} \cdot \mathbf{S} \cdot \mathbf{p}_{\cdot,c} \qquad b, c = 1, ..., C \tag{3.25}$$

PCA modelling consists of determining the linear combinations as in equation 3.24 that are mutually uncorrelated and have maximal variance. Uncorrelated linear combinations satisfy the following equations:

$$\text{cov}\left(\mathbf{t}_{\cdot,b}, \mathbf{t}_{\cdot,c}\right) = 0 \qquad \forall b, c = 1, ..., C | b \neq c \tag{3.26}$$

To find the first principal component, one seeks the non-null vector $\mathbf{p}_{\cdot,\mathbf{1}}$ that maximizes $\text{var}\left(\mathbf{t}_{\cdot,1}\right)$ while being constrained to have unit norm:

$$\max_{\mathbf{p}_{\cdot,1}^{T} \cdot \mathbf{p}_{\cdot,1} = 1} \left(\text{var}\left(\mathbf{t}_{\cdot,1}\right)\right) = \max_{\mathbf{p}_{\cdot,1} \neq \mathbf{0}} \left(\text{var}\left(\mathbf{p}_{\cdot,1}^{T} \cdot \tilde{\mathbf{X}}\right)\right) \tag{3.27}$$

The second principal component (PC) is found by seeking the non-null vector $\mathbf{p}_{\cdot,\mathbf{2}}$ that maximizes $\text{var}\left(\mathbf{t}_{\cdot,2}\right)$ while being orthogonal to the first principal component, $\mathbf{p}_{\cdot,\mathbf{1}}$, and constrained to have unit norm:

$$\begin{aligned} &\max_{\mathbf{p}_{\cdot,2}^{T} \cdot \mathbf{p}_{\cdot,2} = 1,\, \text{cov}\left(\mathbf{t}_{\cdot,\mathbf{1}}, \mathbf{t}_{\cdot,\mathbf{2}}\right) = 0} \left(\text{var}\left(\mathbf{t}_{\cdot,\mathbf{2}}\right)\right) \\ = &\max_{\mathbf{p}_{\cdot,2}^{T} \cdot \mathbf{p}_{\cdot,2} = 1,\, \text{cov}\left(\tilde{\mathbf{X}} \cdot \mathbf{p}_{\cdot,1}, \tilde{\mathbf{X}} \cdot \mathbf{p}_{\cdot,2}\right) = 0} \left(\text{var}\left(\tilde{\mathbf{X}} \cdot \mathbf{p}_{\cdot,2}\right)\right) \end{aligned} \tag{3.28}$$

One continues to determine more components by maximizing the variance of additional linear combinations, constrained to be orthogonal to all previous principal components. For the c^{th} principal component one writes:

$$\begin{aligned} &\max_{\mathbf{p}_{\cdot,c}^{T} \cdot \mathbf{p}_{\cdot,c} = 1,\, \text{cov}\left(\mathbf{t}_{\cdot,b}, \mathbf{t}_{\cdot,c}\right) = 0} \left(\text{var}\left(\mathbf{t}_{\cdot,c}\right)\right) \\ = &\max_{\mathbf{p}_{\cdot,c}^{T} \cdot \mathbf{p}_{\cdot,c} = 1,\, \text{cov}\left(\tilde{\mathbf{X}} \cdot \mathbf{p}_{\cdot,\mathbf{b}}, \tilde{\mathbf{X}} \cdot \mathbf{p}_{\cdot,c}\right) = 0} \left(\text{var}\left(\tilde{\mathbf{X}} \cdot \mathbf{p}_{\cdot,c}\right)\right), \quad \forall b < c \end{aligned} \tag{3.29}$$

The maximum number of PC's that can be computed is the minimum of the number of variables and the number of samples available ($C \leq \max(M, N)$).

It is proven that the vectors resulting from the discussed procedure, $\mathbf{p}_{.,c}$, are eigenvectors of the covariance matrix of the scaled data matrix, $\mathbf{S}$ (Johnson and Wichern, 2002). Moreover, the corresponding eigenvalues, λ_c are equal to the variance of the corresponding linear combinations:

$$\lambda_c = \text{var}\left(\mathbf{t}_{.,c}\right) = \text{var}\left(\mathbf{p}_{.,c} \cdot \tilde{\mathbf{X}}\right) \tag{3.30}$$

Therefore, if one sorts the eigenvectors by descending values of their eigenvalues, λ_c and selects the first C of them, then one has exactly determined the principal components as in the former procedure. Importantly, one can thus identify the principal components by solving the following equation, which holds for all eigenvectors of $\mathbf{S}$:

$$\text{cov}\left(\tilde{\mathbf{X}}\right) \cdot \mathbf{p}_{.,c} = \lambda_c \cdot \mathbf{p}_{.,c} \tag{3.31}$$

and selecting the C eigenvectors with the C largest eigenvalues, λ_c.

An important property of the eigenvalues follows from the following equation 3.30 for the relative variance (RV) captured by the c^{th} principal component (Johnson and Wichern, 2002):

$$RV = \frac{\text{var}\left(\mathbf{t}_{.,c}\right)}{\text{tr}\left(\mathbf{S}\right)} = \frac{\lambda_c}{\sum_{b=1}^{\max(M,N)} \lambda_b} \tag{3.32}$$

The equation above can be coined in words as follows: *The proportional amount of variance captured by the c^{th} principal component is equal to the ratio of its corresponding eigenvalue to the sum of all eigenvalues*. Say one now selects the first C principal components, then the relative cumulative variance (RCV) captured by the components is the sum of their relative variances and is defined as:

$$RCV = \frac{\sum_{c=1}^{C} \text{var}\left(\mathbf{t}_{.,c}\right)}{\text{tr}\left(\mathbf{\Sigma}\right)} = \frac{\sum_{b=1}^{C} \lambda_b}{\sum_{b=1}^{\max(M,N)} \lambda_b} \tag{3.33}$$

Now (re-)define the following:

- The c^{th} *principal component* is the eigenvector of the covariance matrix $\mathbf{S}$ with the c^{th} largest eigenvalue, λ_c.
- The c^{th} *principal scores* are the linear combinations, $t_{.,c}$, defined by the c^{th} principal component as in equation 3.24.

and define the matrix, $\mathbf{P}$, containing the C first principal components, as follows:

$$\mathbf{P} = \left[\mathbf{p}_{.,1} \ \mathbf{p}_{.,2} \ \cdots \ \mathbf{p}_{.,C} \right] \tag{3.34}$$

Then, the matrix containing the C principal scores can be written as follows:

$$\begin{aligned} \mathbf{T} &= \left[\mathbf{t}_{.,1} \ \mathbf{t}_{.,2} \ \cdots \ \mathbf{t}_{.,C} \right] && (3.35) \\ &= \left[\tilde{\mathbf{x}}_{.,1} \ \tilde{\mathbf{x}}_{.,2} \ \cdots \ \tilde{\mathbf{x}}_{.,M} \right] \cdot \left[\mathbf{p}_{.,1} \ \mathbf{p}_{.,2} \ \cdots \ \mathbf{p}_{.,C} \right] && (3.36) \\ &= \tilde{\mathbf{X}} \cdot \mathbf{P} \end{aligned}$$

Given a defined PCA model and calculated scores, the original data of a sample can be reconstructed:

$$\hat{\tilde{x}}_{i,.} = \mathbf{t}_{i,.} \cdot \mathbf{P}^T \tag{3.37}$$

This reconstruction is not perfect in the general case due to the induced information loss by dimension reduction. The vector of residuals between the original sample and the reconstructed sample is written as follows:

$$\mathbf{r}_{i,.} = \hat{\tilde{\mathbf{x}}}_{i,.} - \tilde{\mathbf{x}}_{i,.} \tag{3.38}$$

In Figure 3.5(a), the data from Example 1 (see Section 3.3.1.2) are shown. All three PC's were computed for the mean centered data (equations 3.8 and 3.9 apply) and shown in the figure. Geometrically speaking, the principal scores can be obtained by orthogonal projection of the data samples onto these three principal components. In addition, the plane defined by the first two principal components (the plane in which both vectors lie) as well as the plane as defined by equation 3.14 are both plotted even though they cannot be discriminated visually. Figure 3.5(b) shows the same graph as 3.5(a) but the axes are rotated so that it becomes clear that the data lie close to the plane defined by the first two PC's (and thus orthogonal to the third PC). Note that also in this plot, the two planes are visually indiscriminable.

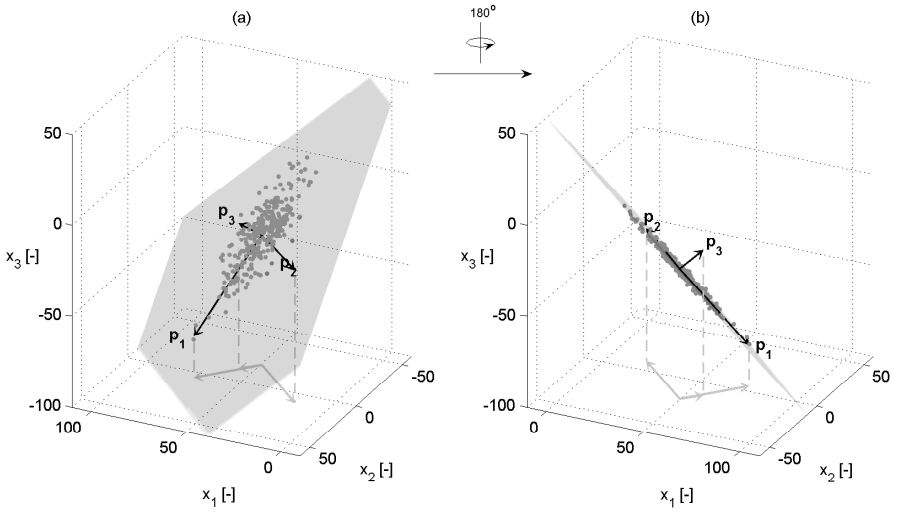


Figure 3.5: Example 1 – Simulated data, principal components (PC's) and (visually indiscriminable) planes defined (1) by the first two PC's and (2) by equation 3.14. (a) and (b) only differ by rotation of the axes. For reasons of visibility, the PC's are multiplied by 3 times the square root of the corresponding eigenvalue, resp. 59.8, 44.2 and 14.22

3.3.1.4 PCA for dimension reduction

By selection of a small number of principal components that capture the largest part of the total variance -that thus have largest eigenvalues-, one can thus replace the original set of M variables by a smaller number of variables, C, with minimal loss of captured variability. Application of PCA models for this purpose is illustrated here by means of the previously defined examples.

Consider Example 1 again (see Section 3.3.1.2) and the corresponding PC's as calculated in Section 3.3.1.3. Figure 3.6 shows the relative variance (RV) captured by the respective principal components and the relative cumulative variance (RCV) for each possible choice of number of PC's (3). As can be seen, the first PC captures the largest part of the variance (62.4%). The second PC captures 33.8%. The third PC captures the remainder (3.8%). If one chooses to keep the first PC only, then 62.4% of the variance of the original data set is captured in the new variable (the first principal score) while only using ⅓ of the original dimensions (1 variable

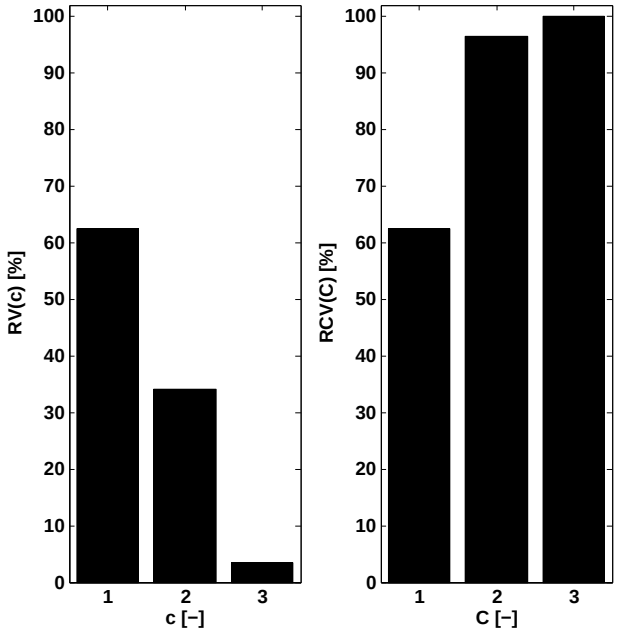


Figure 3.6: Example 1: Relative Variance (RV) and Relative Cumulative Variance (RCV).

instead of 3). By retaining 2 PC's, one captures a total of 96.2% (RCV=96.2%) while the dimension of the selected set of scores is 2/3 of the original number of dimensions (2 PC's vs. 3 original variables). Retaining all principal components does not result in a dimension reduction (3 PC's vs. 3 original variables) and no loss in captured variance (RCV=100%). Retaining 2 PC's may be a reasonable final choice, given that only 3.8% of the variance is lost while effectively reducing the dimension of the data set with 33.33%. This result is not surprising given that the true data samples lie on a two-dimensional plane defined by equation 3.14. Put otherwise, the dimensionality of the underlying error-free variables is 2 due to the constraint defined by equation 3.14. In practice, equation 3.14 may not be known a priori (in fact, knowledge of this equation turns the PCA exercise useless). Application of PCA for dimension reduction is then motivated by presumption that the amount of variance captured in a PCA model is proportional to the amount of information retained. This implies that the variance in the data set due to measurement error is said to be relatively small compared to the variance due to actual variation

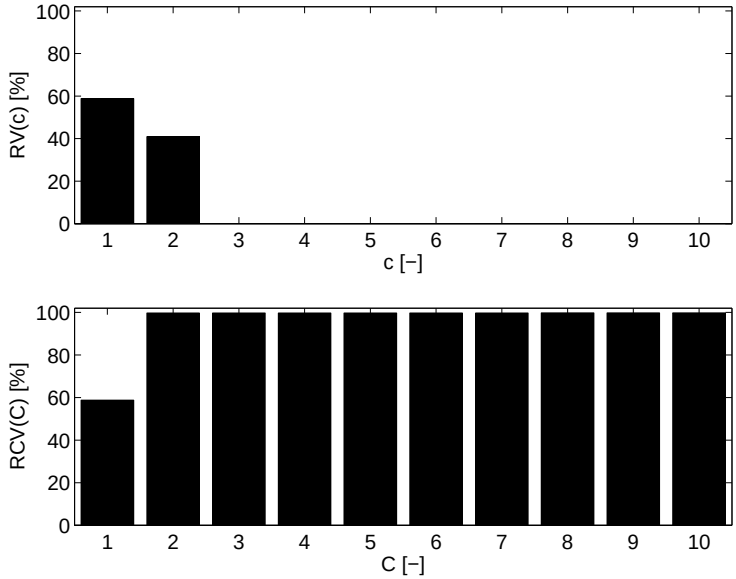


Figure 3.7: Example 2: Relative Variance (RV) and Cumulative Relative Variance (RCV).

in the underlying error-free variables. Other methods designed to determine the number of PC's are discussed in Section 3.3.1.7.

The first 10 PC's were identified for the mean centered data of Example 2 (see Section 3.3.1.2). To center the data, the true mean (i.e. the null vector) was used, assuming this is known a priori. Figure 3.7 shows the relative variance (RV) of the principal components and the relative cumulative variance (RCV) for each possible choice of number of PC's (3). As can be seen, the first 2 PC's capture the largest parts of the variance (58.8% and 40.9%). The third and all following PC's capture less than 0.01%. The PC's beyond the second PC jointly capture 0.3% of the variance. Selecting 2 PC's thus leads to an effective dimension reduction of 99.2% $((M - C)/M = (256 - 2)/256)$ with a (minimal) loss of 0.3% of the variance. Figure 3.8 shows the scores for the first 2 PC's for all simulated samples and the pure samples. Note that the pure samples (see Figure 3.3) were not part of the data set for PCA model identification.

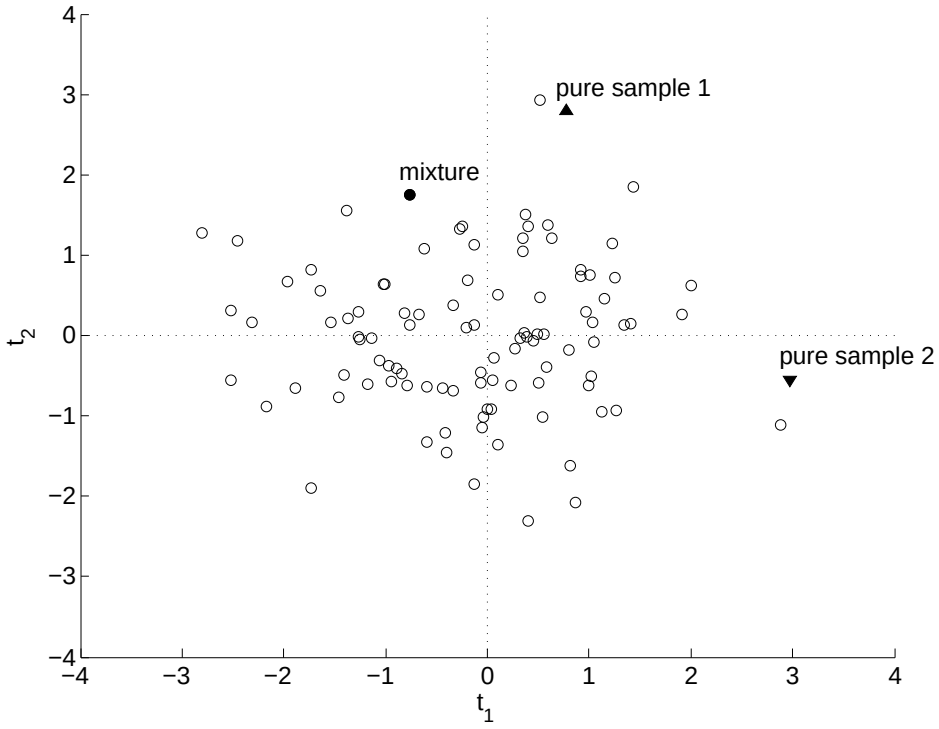


Figure 3.8: Example 2: Biplot of the first two principal scores. The pure samples and one mixture sample are indicated.

Interestingly, pure sample 1 lies close to the axis defined by the second PC and pure sample 2 lies close to the axis defined by the first PC. Given that the origin represents a mixture with zero values for both underlying variables, η_j, the given biplot can be used to interprete the other mixtures. Consider the highlighted mixture for example, also presented in Figure 3.3 and Figure 3.4. A positive value for the second score is observed as well as a negative value for the first score. By this observation it can be conceived that the sample reflects a linear combination of the curves of the pure samples in which the first pure sample is multiplied by a high (positive) number and the second by a low (negative) number. This can be verified in Figure 3.3. It is noted that such an interpretation of a PCA model is not generally available or may be difficult to establish (e.g. pure samples are not necessarily available). Also, it is noted that the pure samples are not expect to lie close

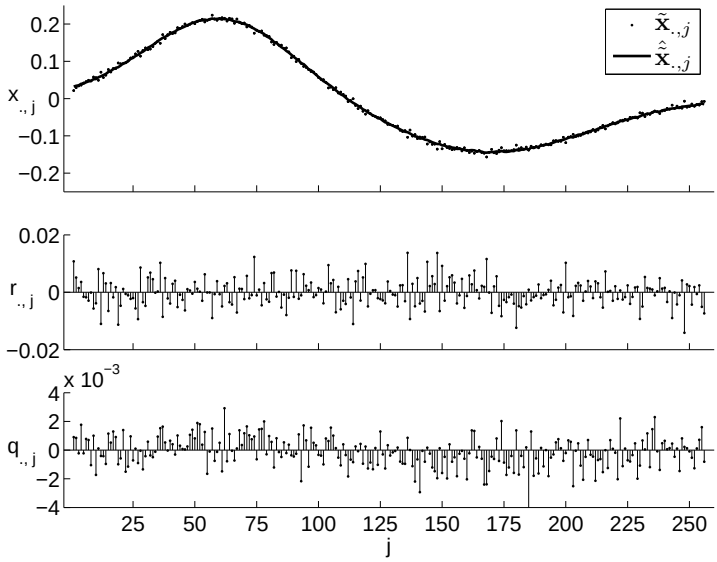


Figure 3.9: Reconstruction of a single sample by means of a 2-PC PCA model.

to the axes defined by the PC's by default nor do they lie exactly on the axes (due to rotational freedom of the PC's). When interpreting a PCA model in practice as indicated here, care should therefore be taken.

Reconstruction of the highlighted mixture samples was pursued for illustration. Given that for the simulation study the error-free variables, $\mathbf{z}_{i,.}$, are known as well, one can identify the deviations of the reconstructed samples from the error-free variables as follows:

$$\mathbf{q}_{i,.} = \hat{\tilde{\mathbf{x}}}_{i,.} - \tilde{\mathbf{z}}_{i,.} \tag{3.39}$$

In Figure 3.9, the respective data sample, $\tilde{\mathbf{x}}_{i,.}$, its reconstruction (by means of the 2-PC model), $\hat{\tilde{\mathbf{x}}}_{i,.}$, the reconstruction errors, $\mathbf{r}_{i,.}$, and the deviations from the true variables, $\mathbf{q}_{i,.}$, are shown. Visual inspection of $\mathbf{r}_{i,.}$ and $\mathbf{q}_{i,.}$ suggest that the 2-PC model permits a valid reconstruction of the original data as well as a valid estimation of the true variables. This underlines the former observation that minimal information is lost by reducing the 256 original variables to 2 principal scores.

3.3.1.5 Fault detection by means of PCA

In what follows, the advantages of multivariate statistical charts are explained. It is repeated here that the statistical charts demonstrated in this section require that process variables are distributed according to a multivariate distribution and that measurement errors are independent and drawn from a normal distribution as defined in equations 3.5 and 3.6.

Univariate charts Standard practices in industries for process monitoring include the construction of univariate charts for measurements of key variables in the considered process. If correlation between the monitored variables exist, then one will likely increase both the rate of false alarms (type I error) and the rate of false acceptance (type II error) for the given process. This is illustrated in Figure 3.10. A set of (normal) samples consisting of two measurements drawn from a from a bivariate normal distribution are plotted as well as an additional deviating sample. The latter sample does not correspond to the behaviour of the normal samples and its (automated) detection is therefore aimed for. Process monitoring techniques should exactly reveal this sample as a deviating one. We call the other samples the in-control samples. Figure 3.10(a) and 3.10(b) show the classic (univariate) Shewhart charts (Montgomery, 2005) for the theoretical 99% confidence limit. The two univariate charts define a rectangle in the biplot showing the two measurements (Figure 3.10(c)). Samples inside this rectangle are accepted as normal. It can be observed that some of the normal samples lie outside the rectangle defined by the univariate charts, thereby leading to undesired false alarms. Also, the deviating sample lies within the box defined by the univariate charts, thus indicating false acceptance of this sample.

Multivariate charts Given a process that produces data samples that are drawn from a M-variate distribution with mean vector μ and covariance matrix $\mathbf{\Sigma}$ then one can construct the X^2 statistic as follows:

$$X^2 = (\mathbf{x}_{i,\cdot} - \mu) \cdot \mathbf{\Sigma}^{-1} \cdot (\mathbf{x}_{i,\cdot} - \mu)^T \tag{3.40}$$

This statistic is a measure for the distance of a sample to the mean and follows a χ^2 distribution with M degrees of freedom. A chart for this statistic can be constructed by plotting the evaluation of this statistic and an upper control limit χ^2_α where α specifies the confidence limit ((1-α)·100%). The points in the M-dimensional space that deliver the same value for X^2 define an M-dimensional

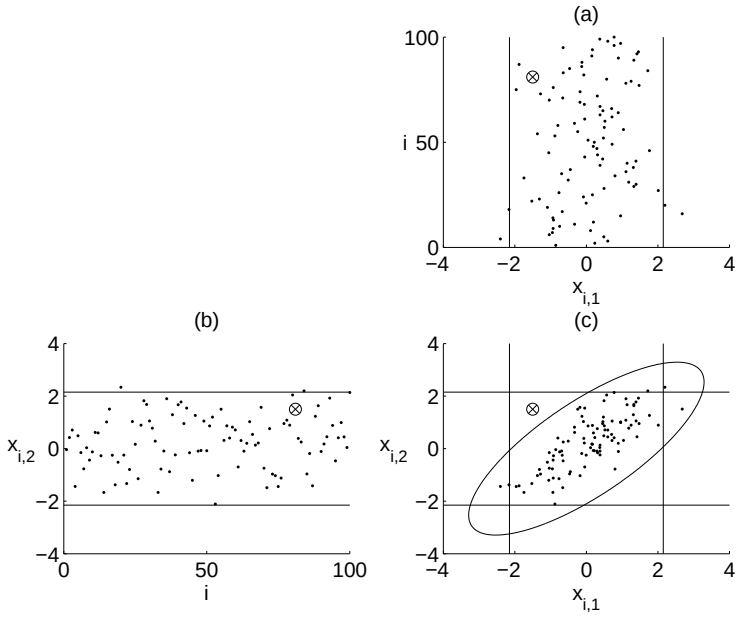


Figure 3.10: Benefits of multivariate statistics for process monitoring. • = normal samples, $\bigotimes$ = abnormal sample. (a),(b): The abnormal sample cannot be detected by univariate Shewhart charts (false acceptance) while a fair number of normal samples are outside limits in the univariate charts (false alarm). (c) By use of the elliptic boundary defined by the Hotelling's T^2 statistic. All shown boundaries are at the theoretical 99% limit.

ellipsoid. In Figure 3.10, the corresponding ellipse (points where $X^2 = \chi^2_{0.05}$) is shown for the bivariate case. Any point lying outside this ellipse will therefore be indicated as an anomaly. It can be observed that the deviating sample is indeed detected as such, while the normal samples lie inside the boundary. The bad performance of univariate boundaries is thus overcome by means of this new multivariate evaluation. The former statistic can however only be constructed when the true mean and covariance matrix are known. To overcome this problem, one can estimate the respective means and covariance matrix as in Section 3.3.1.1, followed by calculation of the Hotelling's T^2 statistic as follows:

$$T^2 = (\mathbf{x}_{i,\cdot} - \mathbf{m}) \cdot \mathbf{S}^{-1} \cdot (\mathbf{x}_{i,\cdot} - \mathbf{m})^T \tag{3.41}$$

The latter statistic follows an F-distribution with M and $N - M$ degrees of freedom if the distribution from which the data are sampled is a normal multivariate distribution:

$$T^2 \sim \frac{(N-1)\cdot(N+1)\cdot M}{N\cdot(N-M)} \cdot F(M, N-M) \tag{3.42}$$

Note that this is only valid for new observations, i.e. for samples that were not used for estimation of $\mathbf{m}$, $\mathbf{s}$ nor $\mathbf{S}$. The distribution of the Hotelling's T^2 statistic for samples included in the calibration data set follow a β distribution (Ramaker et al., 2004). Detailed information on the latter can be found in Tracey et al. (1992). The Hotelling's T^2 statistic as defined here is used in Chapter 7.

PCA-based charts Given an identified PCA model in which all PC's are kept ($C = M$), the Hotelling's T^2 statistic in equation 3.41 can be rewritten as follows, (Johnson and Wichern, 2002):

$$\begin{aligned} T^2 &= (\mathbf{t}_{i,.}) \cdot \mathbf{L}^{-1} \cdot (\mathbf{t}_{i,.})^T \\ &= \sum_{c=1}^{C} \frac{t_{i,c}^2}{\lambda_c} \end{aligned} \tag{3.43}$$

where:

$$\mathbf{L} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_C \end{bmatrix} \tag{3.44}$$

It can be proven that for lower number of selected PC's, C, this statistic remains following an F-distribution if the variables are drawn from a multivariate normal distribution as follows:

$$T^2 \sim \frac{(N-1)\cdot(N+1)\cdot C}{N\cdot(N-C)} \cdot F(C, N-C) \tag{3.45}$$

Therefore, one can construct an control chart for this statistic with the following upper control limit:

$$T_\alpha^2 = \frac{(N-1)\cdot(N+1)\cdot C}{N\cdot(N-C)} \cdot F_\alpha(C, N-C) \tag{3.46}$$

This property allows thus to construct a joint statistic for the selected principal components, which supposedly include the largest part of the variation of the dataset. As discussed before, the selected principal scores are expected to contain the largest part of the information of (future) samples for in-control situations. To evaluate to what extent a new sample belongs to the in-control situation, the Q statistic is constructed, also referred to as Squared Prediction Error (SPE):

$$Q_i = \mathbf{r}_{i,.} \cdot \mathbf{r}_{i,.}^T \tag{3.47}$$

It can be shown that, under the assumption of a normal multivariate distribution of the data, the upper control limit (UCL) of the Q statistic can be approximated as follows (Jackson and Mudholkar, 1979):

$$Q_\alpha = \theta_1 \cdot \left[t_\alpha \cdot \frac{\sqrt{2 \cdot \theta_2 \cdot h_0^2}}{\theta_1} + 1 + \frac{\theta_2 \cdot h_0 \cdot (h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{h_0}} \tag{3.48}$$

where:

α : significance level

t_α : upper $(1 - \alpha)$ percentile for standard normal distribution $(N(0, 1))$

$$h_0 = 1 - \frac{2 \cdot \theta_1 \cdot \theta_3}{3 \cdot \theta_2^2}$$

$$\theta_b = \sum_{c=C+1}^{M} \lambda_c^b$$

To demonstrate the use of the two discussed statistics, consider the following two events for the Example 1 (Section 3.3.1.2):

- A leak in the system, i.e. a constant but unmeasured negative flow is apparent. The mass balance as expressed in equation 3.14 is thus not valid anymore.
- An extreme-flow event, i.e. all flows are twice their normal means. The mass balance as expressed in equation 3.14 remains valid.

For each of the two cases, a single sample is simulated. Figure 3.11 is a repetition of Figure 3.5 additionaly showing the two added abnormal samples. Consider the leak event first. It can be seen in Figure 3.11(b) that this point lies remote from the planes defined by the first 2 PC's, further referred to as the PC-plane here. This is to be expected given that the PC-plane was identified to lie close to the plane defined by the mass balance equation 3.14, further referred to as the MB-plane. Indeed the mass balance equation is not valid for the considered event and therefore the sample is expected to lie remote from the MB-plane. Geometrically speaking, the squared orthogonal distance from the considered point to the PC-plane is exactly the Q statistic. In Figure 3.12, it can be seen that the Q statistic for the simulated leak event violates the 99% limit which confirms the deviating behaviour of the data. Consider now the simulated extreme-flow event. It can be observed that the simulated sample lies close to the PC- and MB-plane, thereby indicating that equation 3.14 remains valid. This is confirmed by inspecting the Q statistic for this sample as it remains below the set limit (Figure 3.12). Given

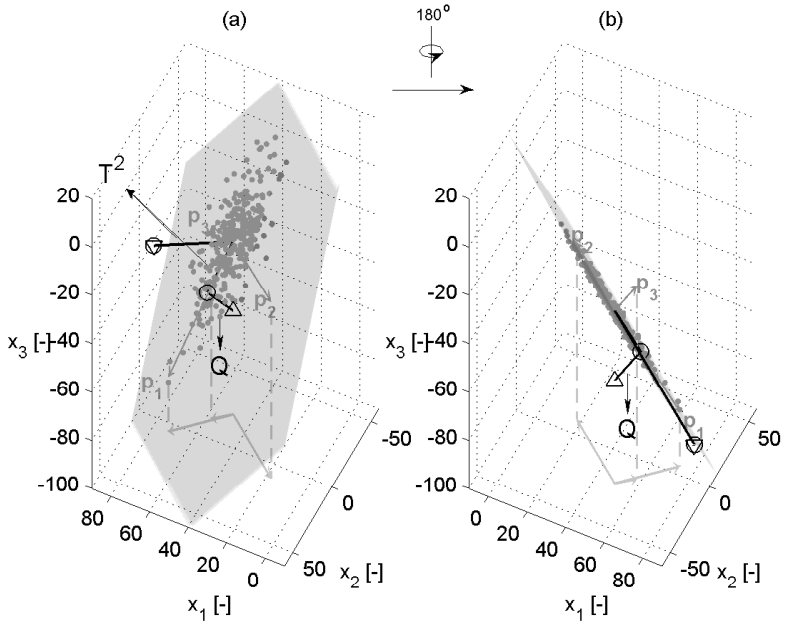


Figure 3.11: Example 1 – Geometrical interpretation of the Hotelling's T^2 and Q statistic. $\triangle$ = leak event, $\bigtriangledown$ = extreme-flow event, $\bigcirc$ = projection onto the identified PC-plane. (a) and (b) only differ by rotation of the axes.

the (proper) selection of the 2-PC model, the Q statistic thus indicates to what extent estimated relationship(s) (as defined by the PC-plane) are valid. The (set of) estimated relationship(s) is sometimes referred to as the correlation structure.

The geometrical interpretation of the Hotelling's T^2 statistic follows from the projection of the considered samples onto the identified PC-plane. The projected samples of both simulated events are indicated in Figure 3.11. As indicated, the Hotelling's T^2 is a measure for the distance from the resulting projected samples to the (estimated) mean. Given that (1) the scores are mutually uncorrelated and (2) the squared scores are divided by the respective eigenvalues (equal to the contained variance), the Hotelling's T^2 is by definition the squared Mahalonobis distance between the mean and projected sample, i.e. the Mahalonobis distance within the

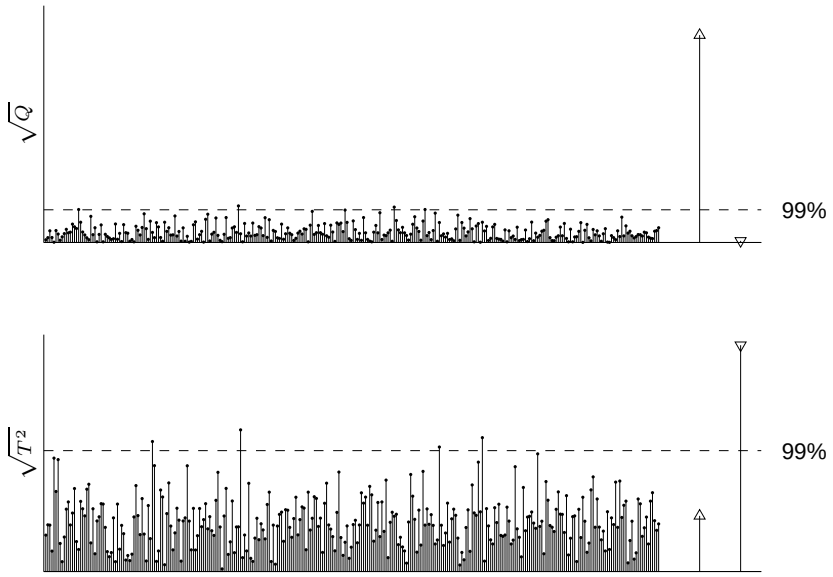


Figure 3.12: Demonstration of the Hotelling's T^2 and Q statistic for the leak and extreme-flow event. For illustrative purposes, square roots of the test statistics are plotted and the values for calibration data are included. • = calibration samples, △ = leak event, ▽ = extreme-flow event.

subspace defined by the PC plane. Consider again the extreme-flow event. In Figure 3.11, it can be observed that the simulated samples for this event lie within the PC-plane but remote from the mean. This is confirmed by a violation of the upper limit of the Hotelling's T^2 statistic (Figure 3.12). The violation of the Hotelling's T^2 (in absence of violation of the Q statistic) indicates the occurence of an abnormal event, though not violating the identified correlation structure. Note that the use of the Hotelling's T^2 presumes a valid PCA model, i.e. the Q statistic needs not to violate its limit for proper use.

Through an example with three variables, the geometrical interpretation of the constructed statistics and their intended use has been demonstrated in the paragraphs above. The given geometrical interpretations of the constructed statistics remain valid for higher numbers of variables and/or scores, even though visualization is impossible in the generic case. In an M-variate setting ($M > 3$), the subspaces defined by a number, C, of eigenvectors ($C < M$) are typically referred to as hyperplanes. Important for practice, the use of the 2 discussed separate statistics allows separating abnormal events in which the identified correlation structure is broken from those in which the identified correlation structure is not broken. In the absence such a feature, one would likely use a single Hotelling's T^2 chart.

3.3.1.6 PCA for regression

In order to demonstrate the application of PCA for regression, the following process is simulated, in which two state variables, $z_{i,1}$ and $z_{i,2}$, are generated from a single underlying variable, ν_i, for N ($i = 1..N$) repetitions, by means of a linear equation:

$$\begin{bmatrix} z_{i,1} & z_{i,2} \end{bmatrix} = \nu_i \cdot \begin{bmatrix} 1 \\ \beta \end{bmatrix} \tag{3.49}$$

The relation between the two variables, $z_{i,1}$ and $z_{i,2}$, can be written as:

$$z_{i,2} = \beta \cdot z_{i,1} \tag{3.50}$$

Now consider measurements taken from both variables, resp. $x_{i,1}$ and $x_{i,2}$, as considered before (equation 3.3):

$$x_{i,j} = z_{i,j} + e_{i,j}, \qquad i = 1..N, j = 1..2 \tag{3.51}$$

The measurements are taken simultaneously and measurement errors are independent and identically distributed (i.i.d.) and drawn from a Gaussian distribution with zero mean and a respective standard deviation for each variable, $e_{i,j}$:

$$e_{i,j} \sim N(0, \epsilon_j), \qquad i = 1..N, j = 1..2 \tag{3.52}$$

For simulation, ν is drawn from the following uniform distribution:

$$\nu \sim U(-\kappa, +\kappa) \tag{3.53}$$

and parameters are set as follows:

$$\begin{aligned} \beta &= 1 \\ \epsilon_j &= .1, \quad j = 1..2 \\ \kappa &= 1 \\ N &= 50 \end{aligned} \tag{3.54}$$

In Figure 3.13 the simulated equation 3.49 and simulated measurements are plotted.

PCA has been proposed for linear regression model identification for cases like the one depicted case above. More specifically, PCA can be used to identify the unbiased linear regression model when both input (predictive) and output (predicted) variables are known or assumed to be subjected to measurement error, i.e. Errors-In-Variables (EIV) regression. Such linear regression model fits the following linear relationship between the two underlying variables:

$$z_{i,2} = \hat{\beta} \cdot z_{i,1} \tag{3.55}$$

with $\hat{\beta}$ an estimator for β in equation 3.50. In the case that the measurements have mean zero and share the same measurement error standard deviation, PCA can be used to obtain an unbiased estimator of β. In the case presented, the expected second principal component is equal to the vector $\left[1 \ -\frac{1}{\beta}\right]^T$:

$$\mathrm{E}\left(\mathbf{p}_{.,2}\right) = \frac{1}{a} \cdot \left[1 \ -\tfrac{1}{\beta}\right]^T \tag{3.56}$$

where:

$$a\text{: norm of } \left[1 \ -\frac{1}{\beta}\right]^T \tag{3.57}$$

It can be shown that as a result, $\hat{\beta}$ in the following equation is an unbiased estimator of β (Wentzell et al., 1997):

$$\hat{\beta} = -\frac{p_{1,1}}{p_{2,1}} \tag{3.58}$$

Importantly, classic least-squares, assuming no measurement errors in one of the considered variables delivers a biased estimate of β. This is demonstrated in Figure 3.13. The linear fits corresponding to Least Squares in $x_{i,1}$ (assuming no error in $x_{i,2}$) and $x_{i,2}$ (assuming no error in $x_{i,1}$) as well as the TLS regression model (correctly assuming equal error standard deviations for both $x_{i,1}$ and $x_{i,2}$) are shown. It is visually clear that the TLS regression model gives a better approximation of equation 3.49.

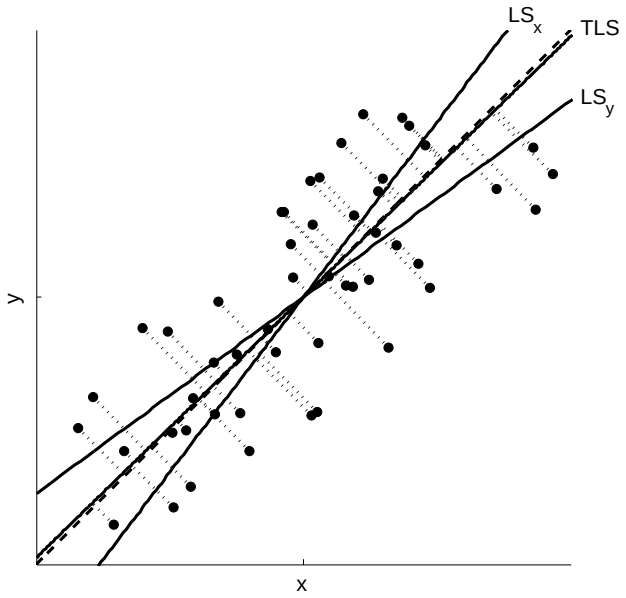


Figure 3.13: Least-squares and Total Least Squares regression. *dots*: simulated measurements; *dashed line*: simulated equation between process variables; *full line*, LS_1: Least Squares fit assuming errors in $x_{i,1}$; *full line*, LS_2: Least Squares fit assuming errors in $x_{i,2}$; *full line*, *TLS*: Total Least Squares fit assuming errors in both $x_{i,1}$ and $x_{i,1}$. Dashed lines indicate minimized distances by TLS.

The PCA-based estimation of the regression model can be extended for more variables and equations in the sense that for a set of M variables, a predefined number of linear and uncorrelated equations ($< M$) can be estimated (Wentzell et al., 1997). Such a model or set of equations is generally written as:

$$\mathbf{x} \cdot \mathbf{B} = 0 \tag{3.59}$$

Practically, $\hat{\mathbf{B}}$, the (unbiased) estimate of $\mathbf{B}$ for the latter model is then defined as the matrix containing the principal components with $M - C$ smallest eigenvalues. Using equation 3.38, the corresponding PCA model minimizes the following objective function for a given number of PC's:

$$J = \sum_{i=1}^{N} Q_i = \sum_{i=1}^{N} \sum_{j=1}^{M} r_{i,j}^2 \tag{3.60}$$

This objective function is equivalent to the objective function of (linear) Total Least Squares (TLS) regression. While a PCA model satisfies the objective in equation 3.60, this objective function is not sufficient to define a PCA model completely if more than 1 PC is retained ($C > 1$). Indeed, the given objective only defines the subspace spanned by the selected PC's, not their orientation within that subspace. It is noted here that the PCA method as presented can be used to assess how many linear equations are valid for a given data set. The number of (linear and orthogonal) estimated equations is exactly the number of PC's that are not included in the PCA model ($M - C$).

Importantly, the PCA method for TLS estimation assumes that the measurements share the same measurement error standard deviation. If this is not the case, standard PCA modelling will introduce a bias again into the estimated regression model. This can be overcome by scaling of the data with the respective measurement standard deviations, ϵ_j. The scaling parameters that result in equal standard deviations of the scaled measurements are called the optimal scaling parameters.

In Schuermans et al. (2005) the connections between PCA and TLS modelling are extensively studied and algorithms provided in the (so far) loosely connected research fields are compared. If the respective standard deviations of the measurements, ϵ_j, are not known a priori then the optimal scaling parameters cannot be identified directly. To overcome the latter problem, it is proposed by several authors (Wentzell et al., 1997; Tipping and Bishop, 1999b; Narasimhan and Shah, 2007) to estimate the measurement error covariance matrix and the PCA model. The so called probabilistic variants or Maximum Likelihood variants of PCA modelling

have been developed relatively recently compared to the original PCA method. To this end, the error covariance structure of the modelled variables is incorporated as a set of parameters which have to be estimated in addition to the principal components. Tipping and Bishop (1999b) use the Expectation-Maximization (EM) algorithm for this problem. However, the Probabilistic PCA (PPCA) model that they present requires that the error-covariance structure is a diagonal matrix with equal entries on the diagonal line. The PCA model itself is thereby still equivalent to the standard PCA model with mean centering. More precisely, the method assumes that (1) all measurements have the same measurement standard deviation and (2) no correlation between the measurement errors exists. Also, Tipping and Bishop (1999b) do not include the vector of means as a set of parameters, therefore assuming that (3) the mean is known exactly or its (prior) estimation does not affect the model. None of the latter hypotheses are generally valid. In this respect, the work by Wentzell et al. (1997) can be regarded as more generic as none of the hypotheses is required a priori in the framework provided. In Wentzell and Lohnes (1999), an effective procedure is provided for which the assumption on a diagonal measurement error covariance matrix has been dropped (i.e. so called oblique scaling or projection is allowed). The method does however assume prior knowledge or proper estimation of the scaling parameters (i.e. the mean vector and error covariance matrix). Narasimhan and Shah (2007) have recently provided an algorithm in which the error covariance matrix and the PCA model itself can be obtained jointly in the Maximum Likelihood (ML) sense. The provided ML-algorithm does not include mean vector estimation and lends itself therefore to cases where the origin (null vector) is (known to be) part of the hyperplane described by the retained PC's. In summary, the probabilistic modelling framework has led to reported algorithms for Maximum Likelihood estimation (ML-estimation) of PCA models. This has so far been reported without estimation of the mean vector in the Maximum-Likelihood procedure.

The probabilistic framework has not perceived large attention in the field of statistical process control (SPC) so far. This can in part be attributed to the required computational expenses (Wentzell and Lohnes, 1999). Indeed, standard PCA allows to calculate the principal components fast and in an incremental order (i.e. not all eigenvectors need to be calculated) without iteration. In contrast, the probabilistic approach requires that for a given number of PC's a complete iterative procedure for model identification is performed. In the absence of prior knowledge of the appropriate number of principal components, the iterative modelling procedure needs to be repeated for every viable choice for the number of PC's. Given that for many applications in chemometrics or process monitoring uncor-

related measurement errors and a unique measurement standard deviation for all variables is assumed, it may not be surprising that the probabilistic framework is seldomly used. Indeed, classic PCA can deliver the ML-model in case the former assumptions are true. Applications to process monitoring are limited to the one of Kim and Lee (2003). The latter find that essential improvements over classic PCA are the ability to handle missing data in an optimal fashion and the ability to extend the model to probabilistic mixture modelling (see Section 3.3.2). Improved performance in terms of fault detection or diagnostics of Probabilistic PCA models over classic PCA models has not been reported as yet. However, the availability of algorithm to optimally estimate the error covariance structure (and thereby the scaling parameters) in the Maximum-Likelihood sense may turn the -often arbitrary- choice of scaling procedures and the resulting ambiguity obsolete. A study of the work by Gurden et al. (2001) indeed suggests that applied scaling procedures are often chosen arbitrarily and without sound statistical support in the context of statistical process monitoring. Also, non-linear relationships -which render (linear) PCA models suboptimal- may affect the estimation of the error-covariance structure.

3.3.1.7 Identification of the numbers of principal components

It is repeated here that the underlying paradigm for dimension reduction by means of PCA is that captured variance relates to captured information. In other words, when using PCA, one underwrites the belief that an increase in captured variance by including an extra principal component corresponds to an increase in captured information. PCA modelling then boils down to a bargain between increased captured variance and decreased number of dimensions. In the examples above, the cumulative relative variance in the selected PC's has been illustrated to be a valid measure to base the PC selection on. The bargain made is often of a subjective nature and -as a result- may be ground for debate. Not suprisingly, a myriad of ways of selecting the number of components have been proposed. In the next paragraphs, the most common methods are reviewed. PC selection by means of captured variance, as already suggested in the previous section, will be repeated for completeness. Then, common approaches to PC selection will be discussed, being the eigenvalue scree plot, data scrambling and data reconstruction

Captured variance Determination of the number of PC's on the basis of (relative) captured variance is by far the simplest approach to PCA model identification. Before actually calculating the principal components, the user defines a minimal proportion of the variance that needs to be captured, e.g. 80%. Then, one proceeds to the calculation of the principal components and the relative captured variance. This is done for the whole data set available to the user. PCA model identification then finishes by selecting the minimal number of PC's necessary to meet the preset captured variance criterion. In both studied examples, 2 PC's would be retained with a preset 80% minimum for RCV. Given the simplicity of this method, it lends itself mostly to dimension reduction exercises where the efficiency of the dimension reduction is of minimal concern or when losses of small but significant proportions of information are not critical. Setting the preset captured variance criterion too high results in inefficient dimension reductions, setting the critical value too low results in loss of (potentially meaningful) information. The use of captured variance for PC selection was demonstrated before.

Eigenvalue scree plot Given that the selection by means of captured variance is often a too coarse way of selecting, a more refined approach towards PC selection has been developed on the basis of the eigenvalue scree plot. The eigenvalue scree plot itself is merely the plot of the eigenvalues corresponding to the principal components, ordered from large to small. Given that the eigenvalues are proportional to the variance captured by the respective PC's, one can equally choose to plot the RV-values. The eigenvalue plot for Example 2 (see Section 3.3.1.2) in Figure 3.14. Consider that one chooses the first PC only. Then, 58.8 % of the variance is captured. Assume now that one evaluates the addition of the second PC to the PCA model. One computes the marginal increase of captured variance by dividing the relative variance (RV) of the considered PC to the relative variance (RV) of the previous PC. This marginal increase is 69% (=RV(2)/RV(1)=40.9%/58.8). The relative captured variance in the second PC is thus of the same order as the first PC and one may therefore argue that PC 2 should be included in the model *if* PC 1 is retained as well (*if 58.8% is important enough than 40.9% should be important as well*). Now, consider to add the third PC as well. Then the marginal increase of RCV is now less than 0.02% ((RV(3)/RV(2)=0.0072%/40.90). This indicates that the increase in RCV by adding the third PC in the model is minimal compared to the increase observed by adding the second PC. This suggest that PC 3 should not necessarily be regarded as relevant if PC 2 is regarded as relevant. In other words, *if* 40.9% is important enough to retain it does not necessarily support the idea that 0.02% is an amount of variance important enough to retain. If one however would

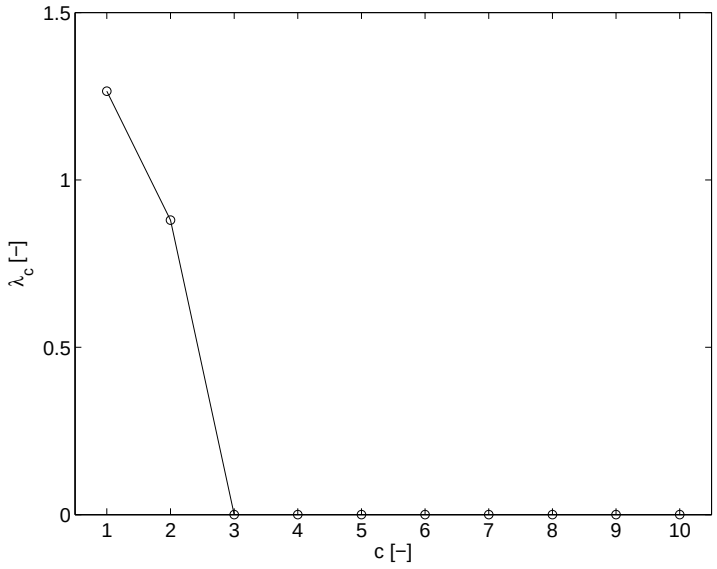


Figure 3.14: Example 2 – Eigenvalue scree plot. The drop in the curve at 3 PC's indicates that 2 PC is a valid choice for the number of PC's

still consider the inclusion of the third PC and so continue to evaluate the marginal increase in captured variance for following PC's, then one finds that the marginal increase ($RV(c)/RV(c-1) \cdot 100\%$) is higher than 89% for all following PC's. Following the argument given before *pro* inclusion of the second PC, this suggests that all PC's from the fourth until the last are important *if* one considers the third PC to be important. Indeed, all of these principal components capture a proportion of the variance that is of the same order as the proportion captured by their directly preceding PC's. As such, when one holds on to the arguments given, the following three options are available:

- Include none of the PC's. This means that all observations are replaced by the sample mean.
- Include 2 of the PC's.
- Include all PC's.

Without doubt, the second option is the most meaningful in the context of dimension reduction. Indeed, the first option leaves no dimensions at all, which means no information is retained (except for the mean vector), while the third option does not result in dimension reduction.

The rather extensive method described in the previous paragraph is performed fast and intuitively by identifying large drops in the eigenvalue scree plot or RV scree plot (going from left to right). Such a drop is easily seen at the third PC in Figure 3.14. This drop reflects the small marginal increase in captured variance by adding the 3rd PC in the model as discussed in the paragraph above. Identification of large drops in the eigenvalues leads therefore straightforwardly to PC selection. The selection method is common for explorative studies.

Scrambling A less popular method of PC selection is based on data scrambling. To do so, one generates a secondary data set with the same number of variables and samples by permuting the values in each column of the data matrix. This permutation process is called scrambling and supposedly removes any meaningful relationships between the measurements. Following the scrambling step, the principal components for the original and scrambled data (scaled in the same manner) are calculated as well as their eigenvalues. Since the scrambled data set is assumed to contain no meaningful information, the eigenvalue of a PC for the scrambled data relates to the expected proportion of variance captured by the considered PC when no meaningful information is contained in the data set. Now, if the corresponding eigenvalue for the original (non-scrambled) data is higher than this (scrambled) eigenvalue, one can conclude that more variance is captured in the considered PC than one would expect for a non-informative data set. The considered PC is then assumed to contain meaningful information. If a PC exhibits an eigenvalue lower than the expected eigenvalue for the scrambled data set, then one concludes that the considered PC contains less variance than would be expected for a non-informative data set. As a result, the considered PC is then judged not to contain (sufficient) meaningful variation. By means of the scrambling method, PC selection boils down to the selection of the first C PC's for the non-scrambled data that have a larger eigenvalue than the corresponding PC's for the scrambled data set. The scrambled and non-scrambled eigenvalues are shown for Example 2 in Figure 3.15. Two PC's are to be retained on the basis of this method, confirming former results.

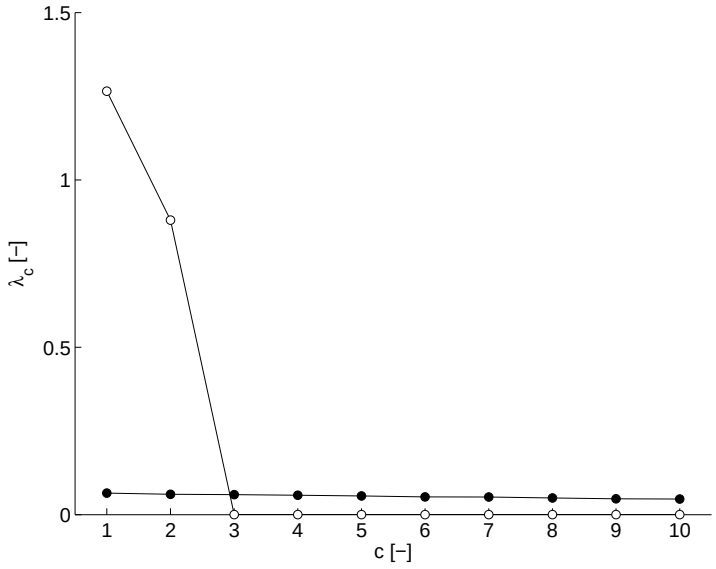


Figure 3.15: Example 2 – Eigenvalue scree plots with and without scrambling. Only the first 2 PC's for the non-scrambled data set are higher compared to those for the scrambled data set, indicating that 2 PC's are a valid choice for the PCA model.

Reconstruction-based selection Another method used for PC selection is based on the fact that PCA models deliver the Total Least Squares regression solution for a given number of PC's. For a given number of PC's, C, data reconstruction (equation 3.35) delivers the optimal estimate of the variables in the least-squares sense. A measure for the overall reconstruction error is expressed as the Root Mean Squared Residual (RMSR). The RMSR for a set of samples used for calibration is defined as follows:

$$RMSR = \sqrt{\frac{\sum_{i=1}^{N}\sum_{i=j}^{M} r_{i,j}^2}{N \cdot M}} \tag{3.61}$$

It can be proven that the PCA model delivers exactly the linear combinations that minimize this RMSR (Johnson and Wichern, 2002). Now, suppose one has a set of N_{val} samples that were not used for model calibration. Call this the validation data set. Then, using the same scaling parameters and the PCA model identified by means of the calibration data set, one calculates the reconstruction errors for the validation data set. Denote the values for the $RMSR$ calculated for the calibration and validation data set as $RMSR_{cal}$ and $RMSR_{val}$. Given that the model was

not optimized for the validation data set (the validation data were not part of the data set used to identify the PC's), the latter is a measure for how good the PCA model will approximate future samples, for which the model was not trained. As simulated examples enable the calculation of deviations between the reconstruction and the error-free variables (Equation 3.39), a corresponding $RMSR$, denoted $RMSR_{true}$, can be defined:

$$RMSR_{true} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{i=j}^{M} q_{i,j}^2}{N \cdot M}} \tag{3.62}$$

As will be shown, the trend in RMSR values may reach a minimal decrease as soon as the number of PC's exceeds the correct number of PC's. Detecting the point after which the RMSR value decrease only minimally thus allows effective dimension reduction.

The $RMSR_{val}$ measure can be made more robust by means of cross-validation. Instead of selecting a single calibration data set and a single validation set, one repeats the evaluation of $RMSR_{val}$ for multiple pair-wise calibration and validation sets. Suppose one wants to repeat the excercise B times. To do so, one splits the data set into B data sets by random selection with size N_b ($b = 1..B$). Preferably, these sets are of equal size ($N_b = N/B$). Each of these sets is then used once as a validation set while the corresponding calibration set is defined as the joint set of the $B-1$ other sets at each time. As such, B PCA models are constructed for all or a preselected set of choices for the number of PC's, each time delivering a measure of captured variance, $RMSR_{val,b}$, in the validation set. An average, $CV\text{-}RMSR$, can now be calculated for each number of PC's by weighted averaging over the single $RMSR_{val,b}$ measures for each number of PC's:

$$CV\text{-}RMSR = \sum_{b=1}^{B} \frac{N_b \cdot RMSR_{val,b}}{N} = \sum_{b=1}^{B} \frac{N_b \cdot RMSR_{val,b}}{\sum_{b=1}^{B} N_b} \tag{3.63}$$

As such, one obtains a value for $CV\text{-}RMSR$ for all choices of the number of PC's. The result is shown in Figure 3.16 for Example 5. The selection of PC's is no different than as for selection on the basis of a single validation set. Indeed, the number of PC's is identified by identifying the PC number after which the RMSR values decrease minimally. The added value of this approach lies in situations where the number of samples is limited. In such cases, the previous (simpler) approach with only one validation set may result in an unrepresentative sampling of the validation data set from the overall data set. As a result, the variance of the $RMSR_{val}$ statistic for a single validation set is higher than the variance of the $CV\text{-}RMSR$.

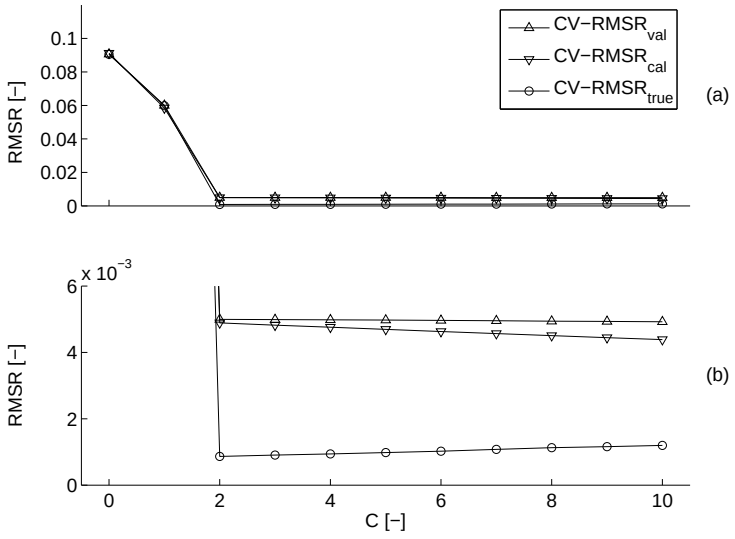


Figure 3.16: Example 2 – Cross-validated RMSR values. Large improvements are observed up to 2 PC's. Beyond 2 PC's, the decrease in $RMSR_{val}$ is minimal and 2 PC's are selected. This choice is confirmed by the observed minimum in $RMSR_{true}$ at 2 PC's.

The $CV\text{-}RMSR$ is thus a more certain measure for the goodness of reconstruction. Put otherwise, when assessing the $RMSE_{val}$ without cross-validation one may arrive in the unlucky situation that the samples in the validation data set are very different from the calibration data. This causes the $RMSR_{val}$ measure more likely to deviate from its expected value compared to the cross-validated (averaged) $CV\text{-}RMSR_{val}$ value. Cross-validation reduces the risk of such deviation and thus reduces the risk of a wrong choice for the number of selected PC's. How many blocks actually should be taken is not generally determined. As a general rule, 5 or 10 blocks are accepted as good compromises between bias and variance of the $CV\text{-}RMSR$ estimate. A special case of cross-validation is called Leave-One-Out (LOO) cross-validation or Jackknifing (Joliffe, 2002). In this case, one creates as many validation sets as samples, each time containing a single sample. This approach is the most computationally intensive of all choices for cross-validation as a maximal number of models has to be evaluated. It also leads to a minimal bias of the $CV\text{-}RMSR$ but may result in large variance.

In Figure 3.16, $CV\text{-}RMSR_{cal}$, $CV\text{-}RMSR_{val}$ and $CV\text{-}RMSR_{true}$ for up to 10 PC's are shown for Example 2. Ten blocks were used for the cross-validation procedure. The option to select 0 PC's (i.e. to reconstruct the data as their respective means) is added for completeness. As one can see, both $CV\text{-}RMSR_{cal}$ and $CV\text{-}RMSR_{val}$ decrease with increasing number of principal components. $CV\text{-}RMSR_{cal}$ is always lower than $CV-RMSR_{val}$ indicating that $CV\text{-}RMSR_{cal}$ is indeed an optimistic indicator for reconstruction performance. Beyond 2 PC's the decrease in $CV\text{-}RMSR_{val}$ is minimal, suggesting that retaining 2 PC's is a valid choice. This is confirmed by investigation of the $CV - RMSR_{true}$ values. At 2 PC's, a minimum is observed for $CV\text{-}RMSR_{true}$ indicating that selecting a too large number of PC's results in increasing deviation from the truly underlying variables. It is stressed that the latter check is impossible in practice as the error-free variables are not known then.

3.3.2 Extensions to Principal Component Analysis

In the following paragraphs, reported modifications and extensions to standard PCA modelling are reviewed. PCA-based methods specifically used for monitoring of batch processes are reviewed in Section 3.3.3.

3.3.2.1 Nonlinear variants of PCA

Standard linear PCA is limited by the underlying assumption that the relationships between measured variables are linear in nature or can be compressed into a limited number of linear principal scores with minimal information loss. This may not generally be true and for this reason, non-linear variants of PCA have been developed. In Figure 3.17 (a) and (c), the linear least-squares (LS) and total least-squares regression equation (TLS) are demonstrated for a bivariate case. The respective non-linear equivalents are found in Figure 3.17 (b) and (d). It was indicated before that PCA delivers linear approximations of data in the total least-squares sense. Equivalently, non-linear variants of PCA can be used to generalize this feature to non-linear regression. In the given example, the data are derived from a third order equation of one variable, y, in a second, x. The graphs show the biplots of the measurements of both variables, for which Gaussian errors with equal standard deviation were simulated.

A simple manner to account for non-linear relationships is to simply transform the data in such a fashion that the modelled relationships between the transformed measurements become linear. To this end, computing the log-transform, roots and/or powers of the original variables are typical choices. Two drawbacks of this approach are that (1) identification of the PCA model may be cumbersome if the structure of the non-linear relationships is not known a priori and (2) the measurements rather than the error-free variables are transformed. The latter possibly leads to unequal weighing of measurement errors which is likely to produce a biased model.

Kernel PCA (KPCA) is a nonlinear variant of PCA developed by Schölkopf et al. (1998b) and generalizes the former method. In this method, the original data are transformed into so-called features first. Standard PCA is then applied by analysis of these features. Figure 3.18 shows the two-step procedure schematically. A major benefit of KPCA is the computational efficiency by which such models can

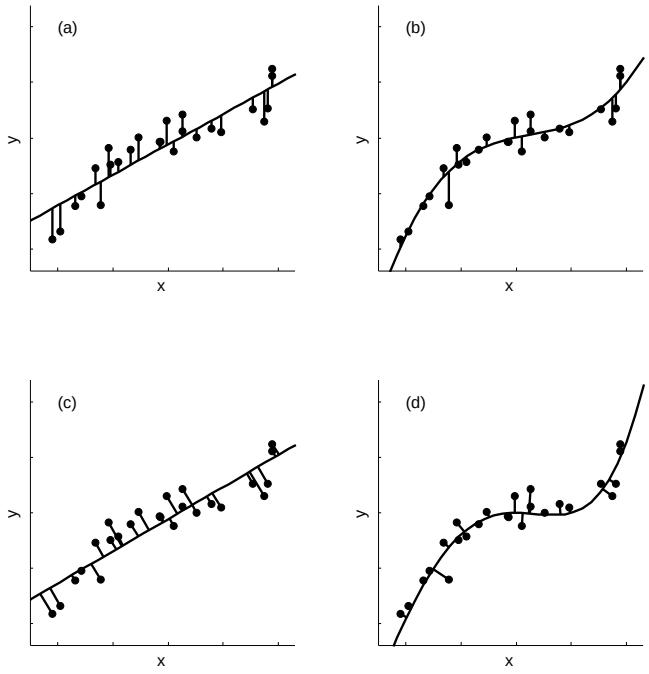


Figure 3.17: Comparing equation approximation by (a) linear Least-Squares regression, (b) non-linear Least-Squares regression, (c) (PCA-based) linear Total Least-Squares regression and (d) Non-linear Total Least Squares regression. After Dong and McAvoy (1996b).

be computed, thanks to the so called kernel trick. This trick is applicable when the transformation satisfies Mercer's theorem. A square kernel matrix, $\mathbf{K}$, with as many rows and columns as the number of samples results from this procedure. The kernel trick may help to resolve the first problem observed before, i.e a cumbersome search for appropriate transformations. Rosipal and Girolami (2001) provide a probabilistic version of Kernel PCA, in analogy to the probabilistic (linear) PCA model of Tipping and Bishop (1999b). Note that Kernel PCA does not solve the problem of non-linear transformation of measurement errors. Applications in the context of process monitoring can be found in (Lee et al., 2004b; Choi et al., 2005). While KPCA is reported to efficiently tackle non-linearity in view of process monitoring, diagnosis is more difficult to achieve as the KPCA models is based on

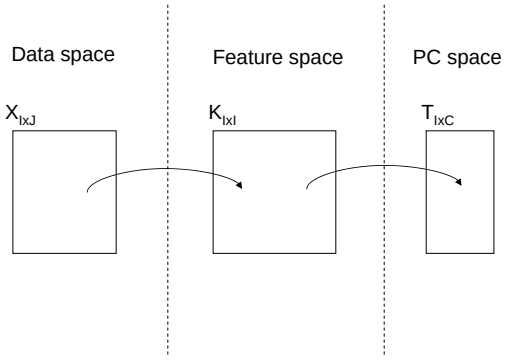


Figure 3.18: Schematic interpretation of Kernel PCA.

the transformed data, i.e. the features, for which no explicit meaning is directly available. A major drawback in this context is that the reverse transformation from the feature space to the original data space is at least diffult to identify (i.e from features to the original data). This problem is referred to as the *Pre-image* problem and troubles the accurate and precise interpretation of the resulting PCA model and its scores. For this reason, its applicability for diagnosis has been limited until today. Advances in solving the Pre-image problem may however relieve this drawback in the future; see for instance Kwok and Tsang (2004).

The principal curves method (Hastie and Stuetzle, 1989) generalizes principal components in the sense that the relationships between the measured data and the estimated (principal) score are not necessarily linear. Locally weighted regression is used to estimate the value of the (non-linear) scores. Importantly, this method does not imply non-linear transformation of the data and thereby does not suffer from inappropriate transformation of measurement errors. Jia et al. (1998) indicate the potential drawback that principal curves require the assumption that the non-linear function is a sum of non-linear functions of the original variables (i.e. an additive model is assumed). Also, principal curves cannot be used directly for projection of new samples as no explicit model is produced (i.e. one does not simply obtain a non-linear equivalent of equation 3.35). This is observed by Dong and McAvoy (1996b) and was consequently tackled by means of a neural network model. To obtain (estimates of) the scores of new samples, the samples are projected onto a neural network model, which approximates the (implicit) Principal Curves model. In the same work, applications for process monitoring are given. Process monitoring based on principal curves is also evaluated in Zhang et al. (1997).

Input-Training PCA (IT-PCA) is an alternative which allows the identification of non-linear PCA models and does not suffer from the assumption on additivity as for Principal Curves (Jia et al., 1998). Also, no non-linear transformation of the data is needed prior to modelling so that an unbiased model is obtained. In IT-PCA, a neural network model and scores are obtained by simultaneously optimizing both the input data (which are the scores) and the neural network model itself so as to predict the output data, which are then the observed data (Tan and Mavrovouniotis, 1995). By choosing a limited number of input variables, effective dimensionality reduction can be obtained. Jia et al. (1998) and Reddy and Mavrovouniotis (1998) use this type of model for process monitoring. To this end, the obtained neural network model is inversed so that from the observed data, the principal scores can be calculated. The original model is then used to give estimates of the original data, which are then used to construct the Q statistic.

3.3.2.2 Variants of PCA dealing with dynamics

Dynamic PCA is a variant of PCA which enables to account for autocorrelated measurements. Indeed, underlying to the application of PCA for process monitoring is the absence of autocorrelated behaviour of the measurements. In practice this may not be the case and for this reason Dynamic PCA (DPCA) has been proposed by Ku et al. (1995). To do so, the data matrix is augmented by defining a new augmented sample, $\mathbf{x}^a_{i,\cdot}$ as the vector of all measurements taken within the last W samples:

$$\mathbf{x}^a_i = \left[\mathbf{x}_{i,\cdot} \; \mathbf{x}_{i+1,\cdot} \; \cdots \; \mathbf{x}_{i+W,\cdot} \right] \tag{3.64}$$

The resulting matrix of all augmented samples, $\mathbf{X}^a$, has dimensions $(I - W) \times (J \cdot W)$. A PCA model is consequently constructed on this augmented matrix. By doing so, the relationships existing between contiguous samples as a result of autocorrelated behaviour can be accounted for. Once the model is established, monitoring charts can be constructed. Ku et al. (1995) provide a method to identify the window length, W. Note that in the provided method the PCA model itself is fixed once established, i.e. a time-invariant model is identified. In Li and Qin (2001), the original identification method is theoretically proven to deliver models that are inconsistent with the relationships underlying to the measured variables. An improved method, called indirect dynamic PCA (IDPCA) and inspired by the work of Chou and Verhaegen (1997), has been proposed and shown to deliver consistent results. Kruger et al. (2004) criticize that the original DPCA method

results in autocorrelated scores even if no autocorrelation in the process is present, hereby leading to invalid test statistics. In their work, this problem is consequently tackled by application of Auto-Regressive Moving Average (ARMA) filters to the data prior to PCA modelling.

In order to deal with (gradual) changes of the existing relationships between variables, i.e. changes in the covariance structure of the data, Wold (1994) proposes repeated identification of the PCA model as new data become available by means of an exponentially weighted moving window (exponential weights are given to each data sample). A more efficient algorithm, denoted Recursive PCA, is provided by Li et al. (2000) in which the covariance matrix (or the PCA model itself) is updated each time new data are available, conceptually similar to earlier work by Dayal and MacGregor (1997). Adaptive PCA is equally used to denote the use of PCA model updating procedures. Given the covariance structure at sample $i-1$, $\mathbf{S}_{i-1}$, updating the covariance structure for a new sample, x_i, is pursued as follows:

$$\mathbf{S}_i = \alpha \cdot \mathbf{S}_{i-1} + (1-\alpha) \cdot \tilde{\mathbf{x}}_i \cdot \tilde{\mathbf{x}}_i^T \tag{3.65}$$

where:

$$\alpha : \text{forgetting factor } (> 0)$$

The forgetting factor, α, controls the (relative) weight that is given to the past data. A high value will lead to a longer memory of past behaviours, while a low value leads to shorter memory of past behaviour. A more elaborate updating, including the updating of the mean vector, is given in (Lennox and Rosén, 2002). Lee and Vanrolleghem (2003) use an updating procedure based on a moving window. In this approach, the last w samples are used to construct a new covariance matrix at each time instant:

$$\mathbf{S}_i = \sum_{b=i-W+1}^{i} \tilde{\mathbf{x}}_b \cdot \tilde{\mathbf{x}}_b^T \tag{3.66}$$

where:

$$W : \text{window length } (> 0)$$

In the latter procedure, each sample (within the set window) is weighted equally.

Irrespective of the updating procedure, a new PCA model is constructed after each update. To do so, the selected number of PC's can be fixed (Lennox and Rosén, 2002) or adjusted (Li et al., 2000). Li et al. (2000) propose to pursue updating procedures only when the constructed statistical limits are not violated. Indeed, in the context of process monitoring one generally does not want to adapt to abnormal changes in process behaviour. It is noted here that the adaptive modelling implies that changes in process behaviour slow enough not to violate the constructed statistics will be adjusted for, irrespective of whether the changes are normal or abnormal. If faults occur which are characterized by changes of the monitored process (i.e. its covariance structure) which are slower than the speed of adaptation of the adaptive model, then undesired adaptation of the model will inevitably be pursued. Indeed, the model should only be adjusted for normal changes. Also, diagnosis on the basis of adaptive PCA models may be difficult due to the changing properties of the model.

In Kano et al. (2000), a modified version of Adaptive PCA is found, called Moving PCA and confoundingly abbreviated as MPCA (see Section 3.3.3). Rather than simply updating the PCA model and using classic test statistics, the changes in the model itself are tracked by monitoring the (angular) distances between the corresponding PC's. By doing so, the adaptation of the model itself is tracked and checked for anomalies. The same method is also applied in Kano et al. (2001). Note that it is not generally easy to assess which PC's in two PC models correspond to each other.

Whereas Adaptive PCA aims at accounting for changes in covariance structure, Multi-Scale PCA (MS-PCA) aims at accounting for the presence of dynamics in different time scales (Bakshi, 1998). A scheme of the method is shown in Figure 3.19. The original signals are decomposed by means of wavelet decomposition, resulting in a larger set of signals, called the *details*, d_l, and an *approximation*, a_L. The details are band-pass signals corresponding to contiguous frequency bands, while the approximation is the low-pass signal resulting from removing the information in the identified band-pass signals. For an introduction and details on wavelet decomposition, the reader is referred to Section 3.5.2. The index which (uniquely) identifies the studied bandpass signals is called the wavelet scale. Details and approximations are obtained for each measured variable. The covariance structure between the corresponding details and approximations of each variable are consequently modelled by a *scale-specific* PCA model and corresponding test statistics are constructed. Appropriate adjustment of the confidence limits is necessary to account for multiple testing at the same time and for redundancy in the split

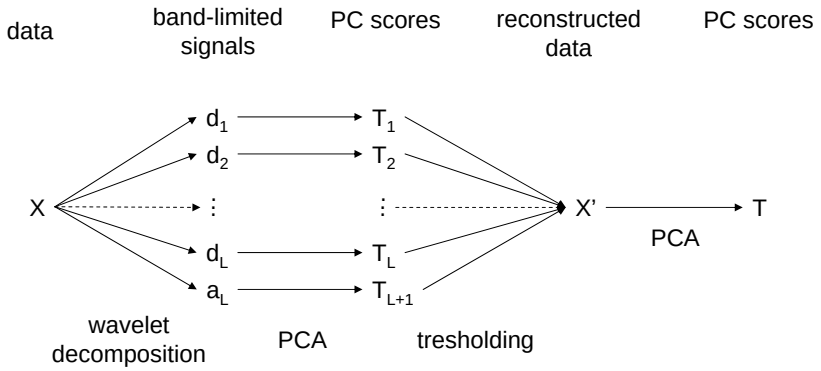


Figure 3.19: Multi-scale PCA. Wavelet decomposition results in band-limited signals which are projected onto separate scale-specific PCA models. Only significant data, detected by violation of statistics, is used to reconstruct the data prior to projection onto an overall PCA model. After Lennox and Rosén (2002).

signals, which is not accounted for by the scale-specific PCA models (Lennox and Rosén, 2002). Violation of the limits at a given scale identifies the scale as significant. Following the identification of significance, the data are reconstructed by only using the significant data at each scale. An overall PCA model is constructed on the basis of reconstructed signals. It is the latter model that is effectively used for process monitoring. Following a comparison of Adaptive PCA and MSPCA in Rosén and Lennox (2001), Lennox and Rosén (2002) propose Adaptive Multi-Scale PCA (AdMSPCA) allowing to account for both changing covariance structure and for presence of dynamic processes in different time scales.

3.3.2.3 Mixture PCA

Mixture PCA (MixPCA) is a model structure proposed to deal with the existence of so-called modes for the modelled data or process. In each mode, the process behaviour accords to a specific behaviour of the data. With MixPCA, each of those behaviours is modelled by means of a proper (linear) PCA model. As many PCA models are identified as the number of modes. When the mode to which an observation belongs is not known a priori, the monitoring task consists of projecting the

given observation on each of the identified PCA models and evaluating whether the observation is rejected or not. If the observation is rejected for all models, then the observations is classified as abnormal. Else, the observation is accepted as normal and the mode is identified by determining to which mode the observation most probably belongs. The latter can be conceived as identifying the model which delivers the lowest p-value for its Q statistic.

In the modelling step, each constituting PCA model can be constructed separately if explicit knowledge of the mode of each observation in the calibration data set is available. This approach is used in Section 5.3.2. If such knowledge is not available, then the assignment of each observation to a mode is part of the modelling exercise. The Expectation-Maximization (EM) algorithm can be used to simultaneously identify the models and the mode to which the observations most probably belong (Tipping and Bishop, 1999a). The number of modes in the data set has to be set prior to modelling however. If no knowledge on the number of modes is available then it is typical to develop PCA mixtures with different viable choices. In order to identify the appropriate number of PCA models in the mixture, Chen and Liu (1999) provide the heuristic smoothing clustering (HSC) algorithm to identify the number of modes. As such, mixture PCA is a valid clustering approach. A potential benefit of this method over Gaussian mixture modelling is that less parameters need to be identified when strong correlation exists between variables. This may reduce the variance of the mixture solutions, possibly traded off with increased bias.

3.3.3 PCA-based methods for batch processes

In the following paragraphs, adaptations of PCA methods specifically designed and studied for monitoring of batch processes are reviewed. The underlying motivation for such adaptations is that standard PCA cannot handle three-way data sets in a direct manner. Indeed, a measurement stemming from a batch process is typically referred to with three coordinates: the index of the batch in which the measurement is taken ($i = 1..I$), the measured variable ($j = 1..J$), and the time index at which the measurement was taken during the batch ($k = 1..K$). One writes the resulting three-way matrix as $\underline{\mathbf{X}}$, with dimensions $I \times J \times K$, in which $\underline{x}_{i,j,k} = \underline{X}(i, j, k)$ is a single element.

3.3.3.1 Multi-way Principal Component Analysis

Multi-way PCA (MPCA), firstly proposed by Wold et al. (1987), is a popular adaptation of PCA for batch processes. This adaption makes use of so called *unfolding* by which the three-way data matrix is transformed -*unfolded*- into a two-way matrix prior to PCA modelling. Although six ways of unfolding are theoretically available, only two of them are typical for PCA-based monitoring of batch processes (Nomikos and MacGregor, 1994). With the first of the latter, called *variable-wise unfolding*, one regards each set of measurements taken in the same batch and at the same time index as a single and independent multivariate sample. Practically, each of the available samples are put below each other as depicted in Figure 3.20 prior to performing PCA. A two-way matrix with dimensions $I \cdot K \times J$ then results ($I \cdot K$ samples with J variables). With the second approach, referred to as *batch-wise unfolding*, one regards the complete batch as a single observation. Practically, one puts all the measurements of a given measured variable at a given time instant below each other. This is depicted in Figure 3.21. The dimensions of the resulting matrix then are $I \times J \cdot K$ (I samples with $J.K$ variables). The latter approach has been used in Chapters 5 and 6 of this dissertation. Introductory notes on the use of MPCA models (among others) for batch process monitoring can be found in Wold et al. (1987), Nomikos and MacGregor (1994), Kourti and MacGregor (1995), Nomikos and MacGregor (1995), Kourti and MacGregor (1996), Kourti (2002) and Kourti (2003).

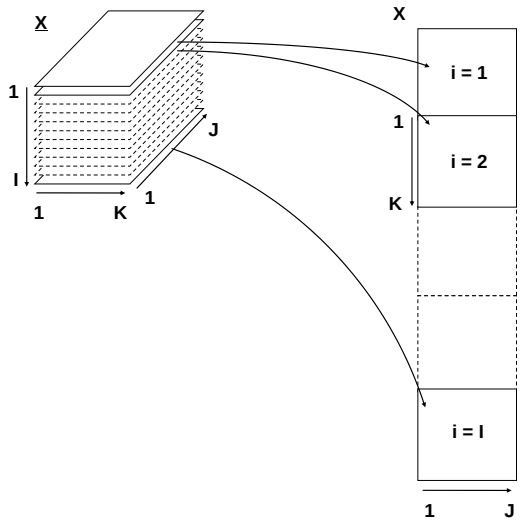


Figure 3.20: Variable-wise unfolding for MPCA.

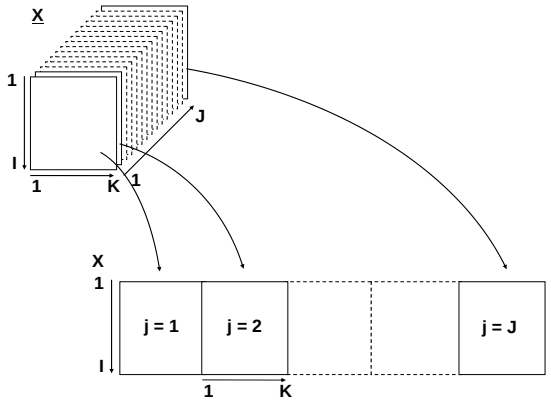


Figure 3.21: Batch-wise unfolding for MPCA.

Batch-wise unfolding results in a problem for on-line monitoring practice as the standard use of the resulting PCA model would require that a monitored batch is completed before the monitoring task can be performed. As such, problems may not be detected in time or costs related to the detected problem may be run high. Even though this problem is not studied in detail in this disseration, it is worth noting that three approaches have been developed to address this problem:

- *Zero deviation*. Assume that in all future time instants in the running batch, the measured values are equal to their mean values. If the data are centered to mean zero as is typical, this means the empty spots in the vector of (centered and scaled) measurements are filled with zeros.
- *Deviation as currently*. Assume that the deviation from the mean is the same for all future time instants in the running batch. In the case of centering to zero mean as is standard, the empty spots in the vector of measurements are filled with the last (centered and scaled) measurement of the corresponding variable.
- *Data estimation*. The missing data due to incomplete batch runs are filled (on-line) with the expected measurements according to the PCA model based on historical data. To do so, the missing data are estimated iteratively by alternating of computation of the scores (e.g. starting with zero deviation replacement of the missing data) and estimation of the missing data (based on reconstruction, equation 3.37) until convergence. A unique solution for the estimated data is theoretically possible as soon as the number of already obtained measurements exceeds the number of retained principal components in the model.

Van Sprang et al. (2002) concluded that the first two approaches are suboptimal based on a study including 5 different applications. Based on the latter study, the third approach, i.e. data estimation, is to be preferred.

The centering and scaling steps were already discussed for standard PCA but require an additional treatment for MPCA. As discussed before, it is typical to calculate a separate mean for each measured variable and time instant in the batches to center data. While using an overall mean value over the complete batch runs (for each of the measured variables) is not excluded a priori in data-driven batch modelling (see Wold et al. (1998)), it has been shown to lead to low monitoring performance (van Sprang et al., 2002; Aguado et al., 2007). As for scaling, three options are reported by Gurden et al. (2001):

- *Column scaling or auto-scaling.* With column scaling, each measurement is divided by the standard deviation of all (centered) measurements for the corresponding measured variable and time instant (see Figure 3.22(a)). J standard deviations are thus computed to do so. The measurements in each column of the 3-way matrix are thus scaled with their proper standard deviation. This is practically the same as application of scaling to unit variance on the unfolded matrix.

- *Single-slab scaling or group scaling.* A single standard deviation is calculated for each measured variable (see Figure 3.22(b)). This means that each measurement is divided by the standard deviation of all (centered) measurements for the corresponding measured variable (and all time instants). The measurements corresponding to one measured variable are thus scaled to have overall unit variance. $J \cdot K$ standard deviations need to be computed. This approach is said not to distort the relationships between the measurements of the same variable (whereas column scaling may do so).

- *Double-slab scaling.* A drawback of single-slab scaling is that the amount of variance in the data are not scaled to unity for each time instant (only for each measured variable). When the variance in one or more of measurements (systematically) changes over time in a batch run, this will result in a higher influence of certain periods over other periods in the resulting PCA model. Column scaling does not exhibit this problem but may distort the underlying relationships between the measured variables. Double-slab scaling avoids both problems by scaling the data so that overall unit variance is obtained for each time instant as well as for each measured variable. This is done by scaling the measurements to unit variance corresponding to each measured variable first (as in single-slab scaling). Then, the measurements corresponding to a single time instant are scaled to unit variance. This is done similar to the single-slab scaling, only the orientation of the *slab* is different (see Figure 3.22(c)). The scaling for the two given orientations needs to be repeated until the overall variances for each time instant and the overall variances for each measured variable have converged to one. Ultimately, $J \cdot K$ standard deviations result in this case as well. Despite the overcoming of the problems related to column scaling and single-slab scaling, the use of the latter scaling has not yet been reported for monitoring practices based on Multiway PCA.

The first and second option have been applied and compared in this work.

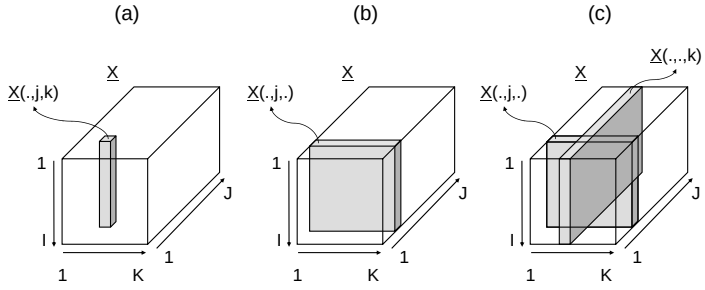


Figure 3.22: Reported scaling procedures in the context of 3-way data sets, according to Gurden et al. (2001). (a) Column scaling or auto-scaling, (b) single-slab scaling or group scaling and (c) double-slab scaling.

In Aguado et al. (2007), MPCA models with variable-wise unfolding and batch-wise unfolding are critically compared in terms of (in-batch) monitoring performance and diagnostic capabilities. The choice between variable-wise and batch-wise unfolding is shown to be of minimal influence on the process monitoring performance while a more straightforward diagnosis on the basis of batch-wise unfolded models is reported. In Ramaker et al. (2002), a method to incorporate external information (i.e. relevant information that is available prior to the start of the batch) into the MPCA-based monitoring strategy is reported.

More applications of MPCA can be found in Kosanovich et al. (1996), Gregersen and Jørgensen (1999), Lennox et al. (2000), Lennox et al. (2001), Sarolta and Kinley (2001), Bicciato et al. (2002), Aguado et al. (2005) and Chiang et al. (2006). Extensions of PCA models as provided in 3.3.2 have been applied for 3-way data as well. Dong and McAvoy (1996a) extend the Principal Curves method to multi-way Principal Curves by batch-wise unfolding of the 3-way data matrix. Yoo et al. (2006b) apply Kernel PCA to batch-wise unfolded data for monitoring of the pilot-scale SBR also studied in this dissertation, so to obtain Kernel Multi-way PCA (KMPCA). Mixture MPCA (MixMPCA) has been proposed in (Yoo et al., 2006a) to account for different modes in batch process data and has been applied to the SBR studied in this dissertation. Additional extensions have been provided to explicitly deal with within-batch and batch-to-batch variability of SBR processes (Rännar et al., 1998; van Sprang et al., 2002; Flores-Cerrillo and MacGregor, 2004; Lee and Vanrolleghem, 2004; Lee et al., 2005; Lee and Vanrolleghem, 2003) and will be reviewed in more detail below.

A simple strategy to overcome the need for data estimation, as occurring for batch-wise MPCA, is to generate a separate model for several time instants, each time modelling all data available up to the corresponding time instant. In other words, instead of constructing a model for the whole batch, one constructs a separate MPCA model for each time instant at which process monitoring is desired. (Ramaker et al., 2005) propose this so called Evolving MPCA modelling method and report satisfying results for process monitoring. In its most extreme form, K different PCA models are made, each time modelling all data available up to the corresponding time instant. However, the use of multiple PCA models may trouble straightforward interpretation when diagnosis is aimed for.

3.3.3.2 Batch-Dynamic Principal Component Analysis

Chen and Liu (2002) provide the Batch-Dynamic PCA (BDPCA) model which generalizes the classic variable-wise and batch-wise unfolding procedures Olsson (2005) to deal with dynamic processes. To practically do so, each (2-way) slice of the data matrix containing all the data of a single batch is augmented in the same way as done for DPCA (see Section 3.3.2.2), thereby delivering a new 3-way matrix, $\underline{\mathbf{X}}^a$, in which the i$^{\text{th}}$ slice (i$^{\text{th}}$ batch) is written as follows:

$$\underline{\mathbf{X}}^a_{i,j,k} = \begin{bmatrix} \underline{\mathbf{X}}_{i,.,1} & \underline{\mathbf{X}}_{i,.,2} & \cdots & \underline{\mathbf{X}}_{i,.,K-d} \\ \underline{\mathbf{X}}_{i,.,2} & \underline{\mathbf{X}}_{i,.,3} & \cdots & \underline{\mathbf{X}}_{i,.,K-d+1} \\ \vdots & \vdots & & \vdots \\ \underline{\mathbf{X}}_{i,.,d+1} & \underline{\mathbf{X}}_{i,.,d+2} & \cdots & \underline{\mathbf{X}}_{i,.,K} \end{bmatrix} \tag{3.67}$$

where:

d : delay

The dimensions of this augmented matrix are $I \times (J \cdot (d+1)) \times (K-d)$. Following this augmentation, the 3-way data matrix is unfolded by variable-wise unfolding (dimensions $(I \cdot (K-d)) \times (J \cdot (d+1))$. If $d = 0$, BD-PCA corresponds to variable-wise unfolded MPCA. If $d = K-1$, BD-PCA will correspond to batch-wise unfolded MPCA.

3.3.3.3 Function Space PCA

In Chen and Liu (2001) the so-called Function Space PCA method (FSPCA) is presented for batch process monitoring. In this method, the variable trajectories over the batch runs are approximated by means of linear combinations of orthonormal functions (i.e. for each trajectory the fitted coefficients are by default uncorrelated) prior to PCA modelling. In the work by Chen and Liu (2001), the orthonormal functions are polynomial functions of the batch running time. The PCA model is then identified onto the fitted coefficients rather than the data itself. In this respect, FSPCA is quite similar to Kernel PCA, where PCA is also used to model transformations of variables. However, KPCA does not require the transformations to be orthonormal and *expands* the original number of variables into a higher number of features whereas FSPCA requires an orthonormal basis and *reduces* the original set of variables into a set of coefficients. The authors prove that classic PCA-based monitoring charts can be constructed (provided an orthonormal basis is used). The authors provide an algorithm by which an appropriate resolution of the basis can be identified. Even though not recognized explicitly by the original authors, orthogonal wavelets or splines (rather than polynomials) may be used as well to define applicable sets of orthonormal functions.

In relation to FSPCA, it is interesting to note that mean centering of data is common practice and is reported to be the best option available (see above). However, for extensive data sets, the number of means, $J \cdot K$, may imply an overparameterized model. As a result variances of the estimated means may be larger than tolerated. The single mean approach finds itself at the other end, leaving only one parameter to estimate, and thereby leading to a bias too large. In-between solutions, i.e. with the effective degrees of freedom between the two depicted extremes are however possible by means of FSPCA. Similarly, the estimation of principal components is seldomly constrained even if the number of parameters grows large. Indeed, the solution for the c^{th} PC has $J \cdot K - c$ degrees of freedom, implying large variance of the estimated model if the number of batches is limited. FSPCA allows to search for an effective bias-variance trade-off. Nevertheless, the bias-variance concept is not widely recognized in the context of process monitoring and diagnosis as yet.

3.3.3.4 Block-structured variants of Multi-way PCA

Quite often, batch process operation consists of several constituting phases, which are characterized by specific faults and problems. To this end, one may benefit from structuring the models used for monitoring and diagnosis in such a way that the identification of the location in time of faults eases the task of fault isolation and diagnosis. In addition, data variation in different subphases may behave independent from each other, thereby supporting the use of separate models for separate phases (Camacho and Picó, 2006). Figure 3.23 shows the split operation into blocks. An extreme form of this approach, called local modelling, is pursued by Ramaker et al. (2005) by constructing a separate PCA model for each time instant (i.e. only modelling the data at the corresponding time instant). Although the latter method cannot account for -typically present- correlation between measurements taken at different time instants in a batch run, the method is reported to deliver satisfying results. As noted by the original authors, the use of a large number of PCA models may impede efficient dimension reduction. Also, inefficient diagnosis may result from this.

As it may not be known a priori which data segments in a batch run behave independently, Camacho and Picó (2006) developed the so called Multi-Phase PCA (MPPCA) method which aims at identifying meaningful subphases in a batch process based on relative improvements of captured variance. In Lu et al. (2004) the Sub-PCA method is presented to automatically identify segments of batch runs that can be modelled by the same model. Test statistics are constructed for the separate PCA models.

Both Sub-PCA and MPPCA identify separate PCA models which are not connected in any way. Contrastingly, Multi-block (Multi-way) PCA (MB(M)PCA) and Hierarchical PCA (HPCA) are two methods that provide a global (i.e. modelling the whole batch process) model, though structured according to constituting subphases. In both methods, the (3-way) data matrix is split into several (3-way) blocks of data which contain the data of the subphases of the batch process. Separate *block* models are fit to each data block as well as a *super* model which models the scores of the *block* scores. Nesting of two NIPALS (Non-linear iterative partial least squares) algorithms results in a maximization of the captured variance by the resulting *super* model. Test statistics are constructed for the *block* PCA models as well as for the *super* PCA model. Besides locating identified faults in time, this approach also allows process monitoring each time a subphase is completed without necessity to estimate missing data. The methods only differ in the definition of

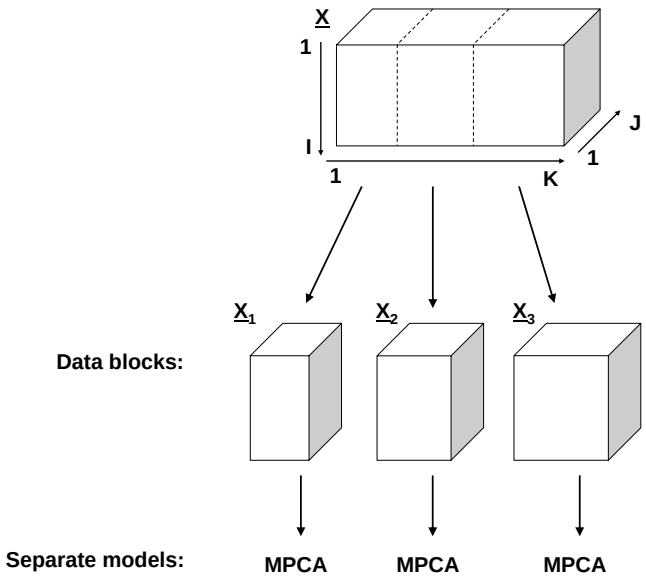


Figure 3.23: Block-structured MPCA: The data set is split into blocks to enable the construction of separate MPCA models for each block.

the *super* PCA model. In MBPCA, the *super* principal components are scaled to have unit norm (as in classic PCA). In HPCA, the *super* principal scores are scaled to unit norm (= unit variance). Both methods require that the subphases are identified a priori. Further details on the methods can be found in Wold et al. (1996) and Westerhuis et al. (1998). Applications of MBPCA can be found in Ündey and Çinar (2002) and Ruiz et al. (2004). Adaptive Multiblock PCA (AMBPCA) (Lee and Vanrolleghem, 2003) is proposed to account for gradual batch-to-batch changes in covariance structure in addition to the gradual changes in process behaviour over the batch runs.

3.3.3.5 Variants of PCA explicitly accounting for within-batch dynamics

Consensus PCA is a dynamically structured PCA model extended from Hierarchical PCA. In this method, each block is defined as a single time slice of the 3-way data matrix (i.e. K blocks are identified containing the data of 1 time instant only). In addition, the *super* scores are not calculated as a single instance for each batch but are updated at each time instant by constructing a *super* PCA model that is gradually changes over the batch run. To do so, a *super* PCA model is constructed on the basis of the *super* scores of the *super* model at the former time instant and the *block* scores for the current time instant. The relative weighing (by scaling) of the *super* scores and *block* scores in each *super* model controls the speed at which the PCA model can change over the batch run. While this weighing can be set differently at each time instant, Rännar et al. (1998) set the relative weighing to 1:1 at all time instants (the amount of variance over the previous time instants is weighted equally to the amount of variance within the data of the current time instant). This constant *relative* weighing (over the batch run) results in an *absolute* exponential weighing of the data over the batch runs. Benefits of the method are that a gradually changing covariance structure (over the batch run) is explicitly modelled as such (by exponential weighting) and that future data do not need to be estimated for monitoring of ongoing batch runs. For further details on the method, the reader is referred to Rännar et al. (1998). In Lee and Vanrolleghem (2004), consensus PCA is extended to Adaptive Consensus PCA (see Section 3.3.2 for Adaptive PCA) in order to account for (gradual) batch-to-batch changes in process behaviour and consequently applied to the SBR system studied in this dissertation. The authors report however that (the rather complex) method delivers similar performance to the (simpler, non-adaptive) MPCA method by Nomikos and MacGregor (1994). This suggests that expected benefits of the method may be minimal in practice. Results reported in van Sprang et al. (2002) do not indicate a clear improvement in performance either.

Adaptive Multiscale MPCA is proposed by Lee et al. (2005) as an extended form of PCA combining MPCA with MSPCA and Adaptive PCA (see Section 3.3.2) into one method. The method accounts for the simultaneous occurence of different processes at different time-scales by decomposing the data into constituting band-limited signals, each corresponding to a certain scale of frequency (frequency band). For each scale, an MPCA model is estimated. An overall PCA model is used to model the scale-related scores jointly. Test statistics are developed for both the scale-specific models as well as the top-level model.

3.3.3.6 Accounting for batch-to-batch dynamics

A method similar to Consensus PCA is proposed by Flores-Cerrillo and MacGregor (2004) to account for non-random batch-to-batch variation. To this end, the data of single batches are considered as separate blocks. The *block* model is then not structurally different from a regular batch-wise MPCA model. A *super* PCA model is however constructed to model the relationships between the data of the given batches and the scores of previous batches. The number of preceding batches for which the scores are included in the *super* PCA model is determined on the basis of detailed inspections of the principal components (eigenvectors) for the *super* model. Note that the overall PCA model is fixed.

Adaptive MPCA, combining batch-wise MPCA and Adaptive PCA, has been proposed (Lee and Vanrolleghem, 2003) as an alternative method to account for (gradual) changes in the covariance structure. The proposed method is based on batch-wise unfolded MPCA, extended by recursive updating of the covariance structure as in Adaptive PCA (see Section 3.3.2).

3.3.4 Concluding remarks

In the previous paragraphs, the basics of PCA model identification and applications have been explained and illustrated. On the basis of the reviewed literature, the following remarks can be made in the onctext of PCA.

While Maximum Likelihood (ML) identification of PCA models has been proposed in the context of total least squares regression to enable maximum likelihood estimation of the regression model and measurement error covariance matrix jointly, the effect of maximum likelihood estimation in the context of process monitoring has to be evaluated as yet. The maximum likelihood has not been introduced for PCA model identification for process monitoring and diagnosis purposes. Neither has it been applied to non-linear variants of PCA in the context of process monitoring or diagnosis. As such, a gap in the available literature is identified.

Bias-variance trade-off is another concept that is not widely recognized in the context of process monitoring as yet. However, given the large-dimensional data sets that typically result from batch-wise unfolding of 3-way matrices and given that case studies are not always based on equally large numbers of batches, constraining

the mean estimate as well as the PCA model itself, e.g. by FSPCA, may effectively deliver an effective bias-variance trade-off hereby reducing the expected deviation from the true model.

In the context of process monitoring, a myriad of extensions of PCA has been reported to deliver solutions to specific problems such as changing means and covariance structure of the monitored processes and the 3-way nature of SBR process data. However, extensive validation of the proposed methods for real-life applications has been lacking so far. As a result, it is the opinion of the author that future research work aimed at critical evaluation and validation of proposed methods in practice may contribute well to the research field. This is believed to be especially true for the extensions for batch processes. Related to the overwhelming number approaches available for model identification, industrial practice in process monitoring and diagnosis may be served well with the generation of model identification protocols. Indeed, protocols that for example allow to identify whether static or dynamic models should be used or whether non-linearities should be accounted for and which non-linear approach may serve the purpose best are not available in literature.

Despite numberous extensions to PCA, some concepts have not been linked or used jointly. First, in the context of process monitoring, Maximum Likelihood estimators of non-linear variants of PCA have been limited to Maxmum Likelihood Kernel PCA. ML-estimators of IT-PCA or Principal Curves have not yet been proposed. Secondly, constraining the potentially overparameterized PCA modelling problem has been proposed in a linear context in the form of FSPCA. Non-linear PCA variants have not been proposed in combination with constraints. Therefore, the following proposal may fill this gap.

Consider the Input Training PCA model that was reviewed before. In this approach, a neural network, more specifically the demapping part of a conventional artificial neural network for regression, is estimated jointly with the input values so to compress the output variables into a limited set of input variables. This approach has not been used as yet for compression of batch process data. However, after unfolding, this may be equally available so that Input Training Multiway PCA (IT-MPCA) may result. In the event of a large number of variables, which may be typical for unfolded batch process data, such a network may be difficult to train in practice without avoiding local minima or without increasing the variance of the model to a too large extent (i.e. different parameter sets will result for repeated data collections from the same (time-invariant) system due to ineffective sampling).

With respect to this problem, Function Space PCA effectively reduces the number of parameters to be estimated thereby reducing the variance in the model solutions, traded off by potential increase in bias. The FSPCA remains however linear.

A method combining non-linear PCA (as in IT-PCA) while imposing constraints on the model (as in FSPCA) has not been proposed as yet. A combined approach, Input Training Function Space PCA (IT-FSPCA) in which the coefficients of the (orthogonal) basis functions are generated by forward projection of scores through an IT-PCA model may effectively allow tackling non-linearity while providing an effective bias-variance trade-off. It is proposed that the IT-PCA using the scores as input generates the coefficients of the orthogonal basis functions which then serve to approximate the original data. In its easiest form, a set of basis functions may be determined a priori, e.g. on the basis of process knowledge or experience. The IT-PCA network needs to be calibrated with the original algorithm. Note that the complete model (from scores to original data space) needs to be estimated during the training process. Indeed the coefficients of the basis functions cannot be estimated independently from the IT-PCA part of the model. An initial training of an FSPCA model (to obtain initial guesses of the basis function coefficients) and the IT-PCA network (to obtain initial guesses of the model parameters and input scores) may however be used for initialization purposes. Algorithms that jointly determine the IT-PCA model as well as the choice of applied basis functions are yet to be developed.

As a last remark, it is noted here that methods devised to account for unequal length of SBR cycles and constituting phases have been left unattended in this chapter. The reader interested in related research is referred to Kassidas et al. (1998), Pravdova et al. (2002), Ramaker et al. (2003) and Fransson and Folestad (2006) as starting points in literature. The problem of unequal length of phases remains unattended further in this thesis. In Chapter 5 and Chapter 6, the problem is non-existent as studied cycles and phases have equal length. In Chapter 7, unequal length of constituting phases results from the implementation of a multivariate statistical controller. However, as this chapter does not deal with the batch-to-batch analysis of the corresponding data, related problems are not dealt with as yet.

3.4 Fuzzy C-means clustering (FCM)

Clustering methods aim at grouping a set of (multivariate) observations (N observations, J variables) into a number of clusters, G. Upon achieving that goal, the so called cluster model can be used to classify a new observation into the cluster that matches the observation the best. To do so, the cluster model incorporates a measure of similarity to each of the identified clusters. Independent of the definition of this similarity, classic K-means clustering always assigns a sample to the cluster with maximal similarity. The so called membership to a cluster is crisp, i.e. an observation can only be a member of one cluster at the time. By doing so, outliers may influence the obtained model to a too large extent. In view of the latter drawback, Fuzzy C-means clustering is a valid alternative (Gustafson and Kessel., 1979; Bezdek, 1981; Babuska, 1998). In this method, the membership of observations are fuzzified (i.e. not crisp) so that observations can belong in part to different clusters. The resulting cluster model is then more robust to outliers. The sum of memberships, $\nu_{i,g}$, of a multivariate observation, $\mathbf{x}_{i,.}$, to a cluster (indexed $g = 1..G$) needs to be constrained to be 1:

$$\sum_{g=1}^{G} \nu_{i,g} = 1, \qquad \nu_{i,g} \in [0,1] \tag{3.68}$$

For final classification to a unique cluster, a (new) observation is classified to a cluster with maximal membership. One may choose to assign an observation to a cluster only when a critical value is reached for this maximal membership. Not assigning an observation to any cluster may trigger a detailed investigation of the given observation. The FCM method is used in Chapter 6 to obtain such a classification model. The similarity measure between a multivariate observation, multivariate sample, $\mathbf{x}_{i,.}$, and a cluster, indexed g, is defined as follows in FCM:

$$d_{i,g}^2 = (\mathbf{x}_{i,.} - \mathbf{m}_{g,.}) \cdot \mathbf{H}_g \cdot (\mathbf{x}_{i,.} - \mathbf{m}_{g,.})^T \tag{3.69}$$

$\mathbf{m}_{g,:}$: centroid or cluster center for cluster g

$\mathbf{H}_g$: norm-inducing matrix for cluster g

The lower this distance metric, the more similar the observation is to the data in the evaluated cluster. The so called norm-inducing matrix, $\mathbf{H}_g$, defines the shape and relative magnitude of the cluster regions. In the original algorithm (Bezdek, 1981), as used in this work, this matrix is defined as the unity matrix, $\mathbf{I}$. This implies that the Euclidian distance is used as a similarity measure and that the cluster regions

are spherical. Ellipsoidal clusters, which may be more appropriate, are allowed by the following definition (Gustafson and Kessel., 1979; Babuska, 1998):

$$\mathbf{H}_g = (\rho_j \cdot det(\mathbf{F}_g))^{1/J} \cdot \mathbf{F}_g^{-1} \tag{3.70}$$

where:

$$\mathbf{F}_g = \frac{\sum_{i=1}^{N} (\nu_{i,g})^m \cdot (\mathbf{x}_{i,\cdot} - \mathbf{c}_{g,\cdot}) \cdot (\mathbf{x}_{i,\cdot} - \mathbf{c}_{g,\cdot})^T}{\sum_{i=1}^{N} (\nu_{i,g})^m} \tag{3.71}$$

ρ_g: parameter controlling the relative volume of the cluster region

m: fuzzy exponent, $m \in [1, \infty]$

The fuzzy exponent, m, controls the so called blur of the fuzzy model. The higher its value, the more blurry or fuzzy the cluster model becomes and the less influential outliers and observations with characteristics of several clusters become. For a given model, the memberships and centroids satisfy:

$$\nu_{i,g} = \frac{1}{\sum_{b=1}^{G} \left(\frac{d_{i,g}}{d_{i,b}} \right)^{\left(\frac{2}{m-1} \right)}} \tag{3.72}$$

$$\mathbf{m}_{g,\cdot} = \frac{\sum_{i=1}^{N} \nu_{i,g}^m \cdot \mathbf{x}_{i,\cdot}}{\sum_{i=1}^{N} \nu_{i,g}^m} \tag{3.73}$$

For a given number of clusters G and model parameter settings, m and ρ_g ($g = 1..G$), a clustering model is obtained by minimizing the following objective function:

$$J(c, \mathbf{m}_{1..G,\cdot}, \rho_{1..G}) = \sum_{i=1}^{N} \sum_{g=1}^{G} (\nu_{i,g})^m \cdot d_{i,g}^2 \tag{3.74}$$

The number of clusters, G, needs to be defined a priori. To support a choice for G, entropy-like measures can be evaluated or even used for automated identification. For details, the reader is referred to Bezdek (1981).

Appplications of FCM in the context of wastewater treatment are found in Marsili-Libelli and Müller (1996) and Marsili-Libelli (1998). More specifically, in Marsili-Libelli and Müller (1996) FCM is used for fault detection while in Marsili-Libelli (1998) FCM is used for the assessment of two distinct process states.

3.5 Qualitative Representation of Trends (QRT)

3.5.1 Introduction

Qualitative Analysis (QA) and Qualitative Representation of Trends (QRT) have been used as monikers for techniques in data processing that aim at the description of data series into qualitative terms. The use of qualitative information on data or the system and/or processes they stem from is often warranted and supported by the observation that an operator's reasoning or knowledge is to a large extent based on qualitative features in data series rather than their quantitative properties. Given that operators spend a large proportion of their time to the monitoring of trends in process measurements (Yamanaka and Nishiya, 1997), an automated assessment of process trends can help operators in performing the task of process supervision. Part of the potential for such a tool lies in the context of fault detection and isolation. Indeed, information about trends may be addressed to the operator only in case of (automatically identified) anomalies. As such, an operator does not need to check normal data on a regular basis and can thus focus on actual problems. This reduces the time spent on evaluation of normal data and will likely result in a faster reaction to and analysis of abnormal situations. More advanced use of such tools includes the design of automated diagnosis and/or control systems that are based on qualitative assessment of process trends. Reported applications of these methods aim indeed at process monitoring and diagnosis (Akbaryan and Bishnoi, 2001; Rubio et al., 2004; Flehmig and Marquardt, 2006) or mining of process data (Stephanopoulos et al., 1997). Other applications in biotechnology use qualitative descriptions for model identification (Vanrolleghem and Van Daele, 1994; Shaich et al., 2001). Ciappelloni et al. (2006) perform a PCA-based analysis of the location in time of break points (i.e. sudden rises) in the oxygen profile of an aerobic SBR, though without providing a method for their automated detection.

In order to deal with a multivariate context, Maurya et al. (2005) combine PCA and QRT by applying QRT to principal scores calculated by PCA, i.e. reducing the dimensionality prior to trend assessment. Flehmig and Marquardt (2006) analyze the series separately after which joint behaviour is presented as the combination of behaviour for the separate signals. It is noted that such presentation turns very complex for large numbers of series. Indeed, suppose three behaviours are possible for each univariate series (e.g. upward, steady and downward), then, for a J-variate multivariate series, the number of potential joint behaviours becomes 3^J. Interpretation of results may then become a cumbersome task.

As an example taken from the context of automated control of biological wastewater treatment plants for nutrient removal, oxidation reduction potential (ORP) signals and the points within time series thereof where transitions from accelerating to decelerating behaviour and vice versa occur (inflection points) have been recognized as key process indicators (Wareham et al., 1994; Vanrolleghem and Coen, 1995; Plisson-Saune et al., 1996; Ra et al., 1999; Andreottola et al., 2001; Fuerhacker et al., 2001; Kim et al., 2004; Li et al., 2004). As the inflection points in ORP signals can be used as indicators for the end of nitrification and denitrification processes in wastewater treatment systems (Chang and Hao, 1996), an accurate assessment of inflection points with minimal delay is desired for such signals. Figure 3.24 shows a typical ORP signal of the pilot-scale Sequencing Batch Reactor (SBR) studied in this work. Of special importance to the presented work is the fact that the signal exhibits 3 contiguous inflection points with no maxima or minima in between (respectively at minute 33, 77 and 91). The third inflection point of these is typical for the endpoint for nitrification.

Generally speaking, it is desired that techniques for QRT are generic, fast and robust (Dash et al., 2004a). More precisely, a good technique will assume the least as possible on the analyzed data, can be computed fast (for on-line applications) and delivers the same result repeatedly for the numerically identical and qualitatively identical signals, i.e. delivers consistent, unique and robust results. Different weights may be put on these targeted properties. A comparative studies of different techniques in terms of the latter qualities has not been performed as yet.

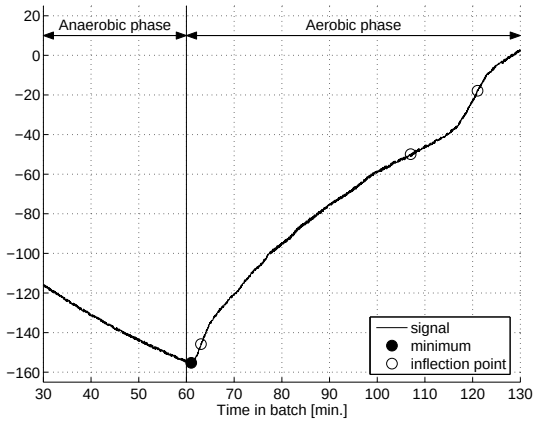


Figure 3.24: Typical ORP time series of the studied SBR.

Underlying to the latter observation may be counted:

- a lack of benchmark data sets or benchmark problems for QRT techniques.
- a lack of clear and/or unique targeted results for QRT methods. Indeed, it is often not clear what the true qualitative representation of a series is nor is there a unique definition of what qualitative descriptions should and should not include. Also, end-users may doubt their judgement as a valid target, may disagree among each other or may (consciously or unconsciously) ignore qualitative features in the analyzed data.
- arbitrary weighing of desired qualities and arbitrary definitions of qualitative similarity by researchers and/or end users.

Given the lack of unique targets, benchmark problems and arbitrary definitions, it is not a surprise that several approaches to QRT have been developed. Methods available for the assessment of qualitative representations of trends can largely be divided into three classes, based on the principles of their core methodology.

The first class is based on clustering (unsupervised) or classification (supervised learning) models. PCA-based clustering is used by Wang and Li (1999) whereas Rengaswamy and Venkatasubramanian (1995) use neural networks. An essential characteristic and potential weakness of these methods is the necessity for training of the models on historical data prior to application. This may lead to errors related to extrapolation to new data for which the models are not trained for.

Methods of the second class, based on the fitting of polynomial functions in contiguous windows, are more generic in nature. A prior assessment of one or more parameters (e.g. noise level) for acceptable fits is however essential to these methods. This requirement may render the techniques less generic then desired, limit the use of the techniques in case of changing noise characteristics and may -as a result- lead to extrapolation errors as well. The methods by Vanrolleghem and Van Daele (1994), Flehmig et al. (1998), Akbaryan and Bishnoi (2000), Dash et al. (2004a) and Charbonnier et al. (2005) belong to this group. The extrapolation problem is tackled to some extent by Dash et al. (2004a) by means of wavelet-based noise estimation on the signal itself. However, the noise estimate and the polynomial fit are estimated on the same data and in a sequential manner, possibly leading to a biased solution. Possibly, this problem can be resolved by joint (Maximum Likelihood) estimation of approximations and noise parameters.

A third class of methods contains the method explained by Bakshi and Stephanopoulos (1994) only. To the author's knowledge, the latter method is the most generic method available as yet as neither model training or assessment of statistical properties of the signal (e.g. noise level) are necessary prior to application. The method by Bakshi and Stephanopoulos (1994) is explained in Section 3.5.5. It will be shown that this method fails to adequately identify consecutive inflection points. In Section 3.5.6, the method by Dash et al. (2004a) is explained and it is shown that the method suffers in terms of robustness. An initial comparison of the two methods is given in Section 3.5.7 in view of the adoption of one of the methods for QRT and suggestions for method improvement.

QRT techniques are inductive by nature and belong essentially to the field of qualitative data mining. The deductive twin of this field is referred to as qualitative reasoning (QR) or qualitative physics (QP). In the framework of QR, inherent absence of exact or numerical knowledge on the behaviour of system is tackled by means of qualitative, i.e. non-quantitative, representation of the systems. The so-called qualitative models that result are used in view of increased understanding of systems, state prediction and process control, just as numerical models. Preliminary investigations in this field were performed by de Kleer (1977). In the 1980's, the research area experienced a large upswing of interest. The works by Forbus (1984) and Kuipers (1986) can be regarded as key entries within this time window. A gentle introduction to the field of QR can be found in Iwasaki (1997), while Kuipers (1994) delivers a more formal treatment of the subject. An extensive overview of the early developments and applications in this field can be found in Bourseau et al. (1995). Both in Patton et al. (2000) and Travé-Massuyès et al. (1997) chapters on the use of qualitative models for monitoring and diagnosis purposes can be found. Cross-breeding of quantitative and qualitative modelling techniques is studied in Berleant and Kuipers (1997). The inductive field of QRT and deductive field of QR have not been in extensive contact as yet.

As both methods for QRT that are reviewed in this chapter use wavelets as part of their methodology, an introduction to wavelets and wavelet power spectra is presented in Section 3.5.2. Afterwards, the methods for QRT by Bakshi and Stephanopoulos (1994) and Dash et al. (2004a) are reviewed and illustrated (Section 3.5.5 and Section 3.5.6). Based on the reviews made and on the limitations of these methods illustrated in the latter sections, a short discussion and a motivation for the finally selected method are presented in Section 3.5.7.

3.5.2 Wavelets - some basics

Wavelets are a class of wave-like functions that have received increasing attention since the discovery of compactly supported continuous wavelets by Daubechies (Daubechies, 1988). Among others, wavelets have been used as tools for simultaneous analysis of time series in both the time and frequency domain. Indeed, this special class of functions allows to interprete both aspects of a series jointly and, as a result, has resulted in a popular framework to do so. Applications in the context of statistical process control relate to different topics such as fault detection (e.g. Luo et al., 1998; Rosén and Lennox, 2001; Aradhye et al., 2003) and data reconciliation (e.g. Tona et al., 2005). An extensive overview of applications of wavelet-based methods for process monitoring can be found in Ganesan et al. (2004).

To illustrate how wavelets can be used, consider the signal depicted in Figure 3.25. This signal is simulated according to the following equation:

$$x_t = z_{t,1} + z_{t,2} + e_t \tag{3.75}$$

where:

$$z_{t,1} = sin\left(2 \cdot \pi \cdot \frac{t}{\tau_1}\right)$$

$$z_{t,2} = \begin{cases} sin\left(2 \cdot \pi \cdot \frac{t}{\tau_2}\right) & t \geq \frac{N}{2} \\ 0 & t < \frac{N}{2} \end{cases}$$

e_t : white noise sequence

t : time index (integers from 1 to N)

N : length of the time series (512)

τ_1, τ_2 : oscillation periods (64, resp. 16)

This signal is sampled at 512 equally-spaced points in time and features a relatively slow sinusoidal oscillation in its first half (period = 64 times the sampling period). This oscillation continues in the second half of the simulated timeframe while another sinusoidal oscillation (period = 16 times the sampling period) sets in halfway the simulated timeframe. Over the whole timeframe, white noise is

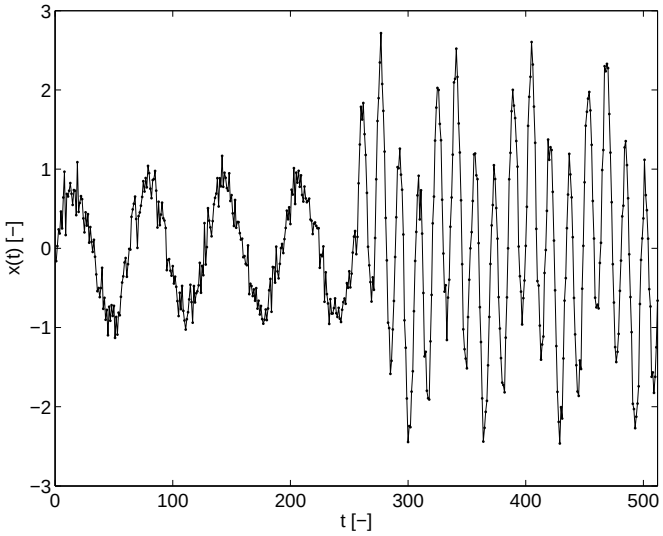


Figure 3.25: Simulated series corresponding to Equation 3.75.

added. While the change in oscillation pattern may be obvious from the figure, it is generally more difficult to discover such information by eye-sight on the basis of this type of plot. In order to characterize signals on the basis of the dominant frequencies at which a signal oscillates the Fourier transform is a common tool. In Figure 3.26 the Fourier power spectrum of the signal, obtained by means of the Fast Fourier Transform (FFT), is shown. Clearly, two distinct peaks in the Fourier spectrum emanate from such analysis. The peaks are located at periods 16 and 64 times the sampling period and thus accord to the periods of the sinusoidal signals that constitute the analyzed signal. While revealing the dominant oscillation periods, the time-local behaviour of the oscillations cannot be identified from this figure. Put otherwise, the Fourier transform leads to the loss of information in the time-domain. The framework given by wavelet theory allows to overcome this problem, as will be shown.

Fourier analysis is based on the decomposition of a signal into sinusoidal waves. According to Fourier theory, a continuous infinite-length signal can be infinitesimally approximated by the summation of sinusoidal waves. For finite-length discrete time series, the limitations are overcome by means of the Fast Fourier Transform. The squared amplitudes for each of the frequencies corresponding to the used sinusoidal waves form the Fourier power spectrum. Such a spectrum is

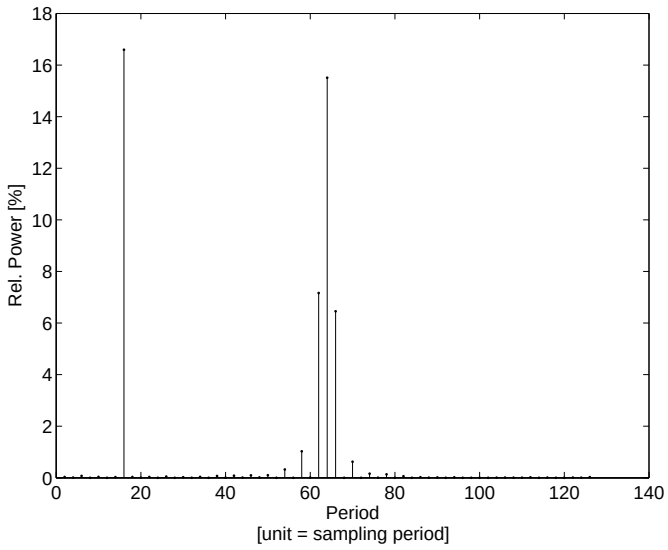


Figure 3.26: Fourier power spectrum by FFT of the simulated series as in Figure 3.25.

shown in Figure 3.26. It is due to the fact that the sinusoidal waves, which are used to approximate the analyzed signal, theoretically range from -∞ to +∞ in the time domain that the time domain information is lost by Fourier transforms. While the windowed Fourier transform has been proposed before to overcome this problem, wavelets have been recognized as a better solution and have consequently been widely adopted. In constrast to approximating a signal with sinusoidal waves, a signal is approximated by time-local waves. An example of such wave, called the Morlet wavelet, is shown in Figure 3.27. It can be seen that this wave-like function is indeed localized in time (see Figure 3.27(a)). This wavelet has -quite rarely- a real part ($\Re(\psi)$) and imaginary part ($\Im(\psi)$) in the time domain. Due to the wavelike properties of this function, such wave is also localized in the frequency domain (Figure 3.27(b)). Now, to analyze a given signal, a given original signal is approximated as a sum of time- and frequency-local functions which share the same shape. The basis shape or function is called a wavelet. The constituting functions of a signal can be achieved by compression and/or stretching of the wavelet (scaling or dilatation), by relocating it in time (time shift) and by adjusting its amplitude and sign. The resulting amplitudes multiplied by the corresponding sign are called the wavelet coefficients. Dilatation or stretching of used wavelet results in a (downward) shift in the frequency domain, denoted with the scale parameter,

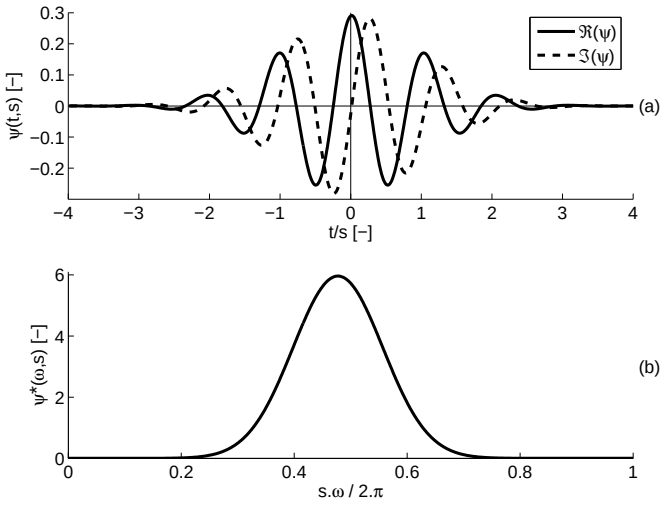


Figure 3.27: The Morlet wavelet in the time domain (a) and in the frequency domain (b). $\Re(\psi)$: real part,$\Im(\psi)$: imaginary part

s. Mathematically, one can calculate the wavelet coefficients ($w(t, s)$) for a given shift, t, and dilation and scale, s, by means of convolution of the signal with the corresponding shifted and dilated (scaled) wavelet. This operation can be written explicitly as follows:

$$w_{t,s} = \sum_{k=0}^{N} x_k \cdot \psi_{k,s,t} \tag{3.76}$$

where:

k, t : (discrete) time indices

N : length of the time series (512)

x_k : value of time series at index k

s : scale

$\psi_{k,s,t}$: daughter wavelet (in time domain) at index k for shift t and scale s

To obtain the values of the daughter wavelets $\psi_{k,s,t}$ the shift and dilation operations can be written as:

$$\begin{aligned}\psi_{k,s,t} &= \psi_{k-t,s,0} \\ &= \psi_{\frac{(k-t)}{s},1,0} \\ &= \boldsymbol{\psi}_{.,1,0} * \delta_{\frac{(k-t)}{s}} \\ &= \boldsymbol{\psi}_o * \boldsymbol{\delta}(\frac{(k-t)}{s})\end{aligned}$$

where:

k : (discrete) time indices

s : scale

t : time shift

$\boldsymbol{\psi}_o$: mother wavelet (in time domain)

$\boldsymbol{\delta}$) : Dirac delta function (in time domain) with shift t and scaling s

$*$: convolution operator

The relation between the mother wavelet function in the time domain, ψ_o, and in the frequency domain, ψ_o^*, is defined as follows by means of the Fourier transform:

$$\psi_o^*(\omega) = \int_{-\infty}^{\infty} \psi_o(t)\, e^{-i2\pi\omega t} dt \tag{3.77}$$

where:

ω : frequency

The Morlet (mother) wavelet is defined in the frequency domain as follows (Farge, 1992):

$$\psi_o^*(\omega) = \pi^{-\frac{1}{4}} \cdot e^{i \cdot \omega_o \cdot \omega} \cdot e^{-\frac{\omega^2}{2}} \tag{3.78}$$

where:

ω_o : nondimensional frequency (>5)

In all graphs and further applications, the nondimensional frequency parameter, ω_o, which controls the shape of the Morlet wavelet, is set to 6 as by Torrence and Compo (1998) and Parent et al. (2006). Figure 3.28 shows the processes of dilation, time shifting and wavelet coefficient fitting. The complete process for a given set of dilatations and shifts is referred to as wavelet decomposition. For analysis problems, i.e. where reconstruction of the signal is of no interest, any choice of shifts and dilatations can be made depending on the interest of the user. The intervals between applied shifts determine the resolution in the time domain, while the intervals between dilatations determine the resolution in the frequency

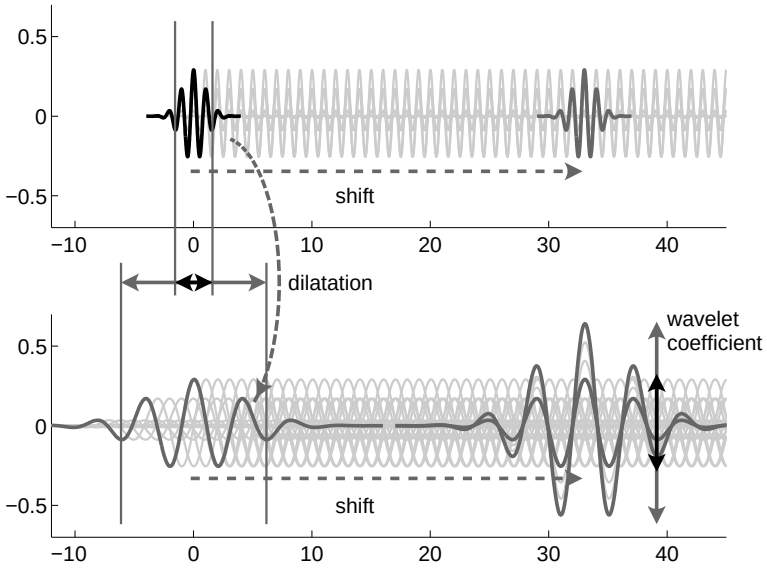


Figure 3.28: Dilation, shift and wavelet coefficient adjustment for the Morlet wavelet. Only the real part is shown.

domain. Following the calculation of the wavelet coefficients, the wavelet power for each dilatation and shift can be calculated as its square. This wavelet power is a measure for the oscillation power for a given frequency and point in time.

Figure 3.29 shows the wavelet power spectrum for the signal plotted in Figure 3.25. One can see that at periods of 64 times the sampling period a continuous high wavelet power is observed. From halfway the signal until the end another ridge can be observed in the wavelet power, with peaking power at 16 times the sampling period. The given power spectrum thus simultaneously depicts what the signal behaves like in both the time and frequency domain. As such, the observed disadvantage in Fourier spectrum analysis is overcome by wavelet spectrum analysis.

One may observe that the wavelet power spectrum for the example given is not exact, i.e. the wavelet analysis does not deliver a crisp result in neither the time domain nor the frequency domain. Indeed, some power of the constituting oscillations *bleeds* into neighbouring frequencies and time instants. This is an inherent property of wavelets – and any practically achievable filter for that matter. Crisp results can never be achieved in both domains simultaneously. Naturally, a choice

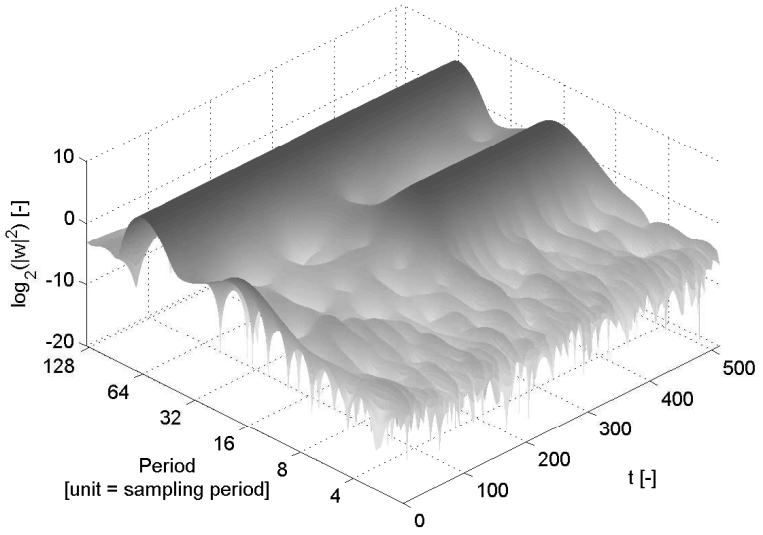


Figure 3.29: Wavelet power spectrum of the simulated series as in Figure 3.26.

for a mother wavelet, i.e. the basis shape, is often a result of a preference for crisp results in the one or the other domain and typically based on the preferences and experience of the user.

More detailed and complete introductions on wavelet theory can be found in Daubechies (1992), Meyer (1993), Strang and Nguyen (1996) and Mallat (1999). A highly accessible introduction to practical computation of the wavelet decomposition by means of the Fast Fourier Transform (FFT) is given by Torrence and Compo (1998). A new and computationally more efficient way of performing wavelet decomposition in the time domain is however been developed by Sweldens (1996). All results in this work are however computed on the basis of the FFT transform as (1) implementation was more straightforward to the author and (2) computational time was not limiting.

3.5.3 Concept of Qualitative Representation of Trends

In order to describe series of data in a qualitative fashion, the desired features that are included into such descriptions need to be identified as well as a formal way of expressing those features is necessary. In the absence of prior knowledge, typical features used in the context for qualitative representation of trends are the signs of the first and second derivative of a (processed) signal. The sign of a variable or the sign of its deviation from a setpoint may be used as well (Cheung and Stephanopoulos, 1990a). For the sign of each derivative, 3 possible states are defined (positive, zero, negative). A more refined set of states of features may be possible (e.g. extreme positive, moderate positive, zero, moderate megative, extreme negative) by use of so called landmark values (Cheung and Stephanopoulos, 1990a). The definition of such landmarks is typically -though not necessarily- based on existing knowledge or experience. In qualitative analysis, the use of landmark values to define crisp intervals has been typical so far. In the field of Qualitative Reasoning it has yet been suggested that available methods can be extended and adapted for fuzzy definitions of qualitative features (Travé-Massuyès et al., 1997). In this work, the (classic) approach, using the signs of first and second derivative has been pursued as (1) the zero value in the analyzed variables has no effective meaning and (2) no prior knowledge or method was directly available to set landmark values.

In order to describe series in terms of the sign of the first and second derivative, Cheung and Stephanopoulos (1990a) define a *triangular episode* as a segment of a series in which the first and second derivative do not change. The latter defini-

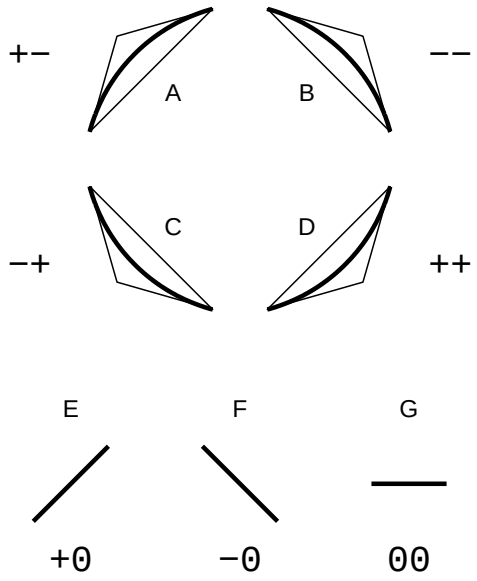


Figure 3.30: Qualitative shapes or primitives, corresponding characters and signs of the first and second derivative.

tion stems from the fact that minimal bounding triangles can be drawn around such segments. Seven generic shapes for a segment can be defined as depicted in Figure 3.30. Each of the shapes is called a *primitive* and describes a unique set of signs for the first and second derivative of a series. To each primitive a unique character is assigned. Following the definitions given so far, Cheung and Stephanopoulos (1990a) define the qualitative representation of a trend as the set of contiguous triangular episodes. In addition to the triangular episodes, *monotonic episodes* are defined in this work as segments of a series in which only the sign of the first derivative is necessarily the same. A qualitative representation of a trend can thus be defined on the basis of either monotonic or triangular primitives. In essence, both methods reviewed hereafter use the same language for qualitative description of trends. In Table 3.1, the characters used for monotonic and triangular episodes corresponding to the respective signs of the first and second derivatives are indicated.

Table 3.1: Qualitative shapes or primitives and corresponding signs of the first and second derivative.

sign		primitives	
1st derivative	2nd derivative	monotonic	triangular
	$+$		D
$+$	0	U	E
	$-$		A
0	0	G	G
	$+$		C
$-$	0	D	F
	$-$		B

3.5.4 Examples

The following signals are simulated to illustrate the reviewed methods. As the true qualitative behaviour of simulated series can be assessed by analysis of the corresponding noise-free series, the simulated series will be used as benchmarks as well.

Example 1 A single (time) series is simulated as follows. The noise-free continuous signal underlying to the simulated discrete signal is a smooth signal, i.e. the underlying signal and (all) its derivatives are continuous. The signal is sampled at equal intervals and a white noise series is added to the signal. The following equation is used:

$$x_t = z_{t,1} + z_{t,2} \cdot z_{t,3} + e_t \tag{3.79}$$

where:

$$z_{t,1} = sin\left(2 \cdot \pi \cdot \frac{t}{\tau_1}\right)$$

$$z_{t,2} = sin\left(2 \cdot \pi \cdot \frac{t}{\tau_2}\right)$$

$$z_{t,3} = -2 \cdot \left(\frac{1}{1 + e^{12 \cdot \left(\frac{2}{3} - \frac{t}{N}\right)}}\right)$$

e_t : white noise sequence

t : time index (integers from 1 to N)

N : length of the time series (12000)

τ_1, τ_2 : oscillation periods (N/2, resp. N/6)

The simulated signal is shown in Figure 3.31 together with the targeted qualitative representations of the noise-free signal. Figure 3.31(a) shows the simulated series with indications of the location of extrema and inflection points in the noise-free signal. Figure 3.31(b) and Figure 3.31(c) show the monotonic, resp. triangular presentation of the noise-free signal, thus defining the desired outcomes (Monotonic:

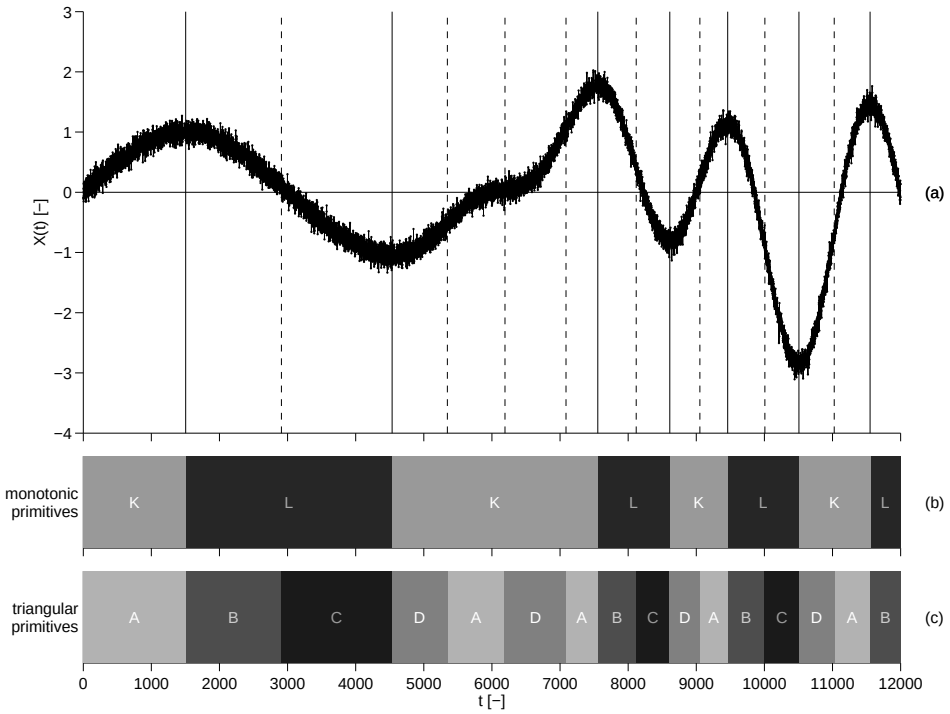


Figure 3.31: Example 1 – simulated series and its true qualitative representation: (a) Simulated series; vertical lines indicate extrema (–) and inflection points (––) of the corresponding noise-free series, (b) monotonic representation ($(KL)_4$) and (c) triangular representation ($ABD(DA)_2(BCDA)_2B$).

$KLKLKLKL \equiv (KL)_4$, Triangular: $ABCDADABCDABCDAB \equiv ABC(DA)_2(BCDA)_2B$). Of special interest in this method evaluation is the section between time indices 4535 and 7554, where the noise-free signal exhibits 3 contiguous inflection points between two extrema. When analyzing finite data series, filtering and smoothing techniques lead to border distortion in most cases. By simulation of sufficient data before and after the episode with three contiguous inflection points, border distortion is avoided in this section of special interest.

Example 2 The second example is borrowed from the work of Dash et al. (2004a) and is a Gaussian bell curve sampled at equal intervals with white noise added. The following equation applies:

$$x_t = z_{t,1} + e_t \tag{3.80}$$

where:

$$z_{t,1} = \alpha \cdot e^{-\frac{(t-\mu)}{2 \cdot \sigma^2}}$$

e_t : white noise sequence with variance $\sigma_e^2 = 1$

t : time index (integers from 1 to N)

N : length of the time series (300)

α : peak magnitude parameter (20)

μ : peak location in time (N/2=150)

σ : peak spread parameter in time (40)

The resulting data as well as the qualitative representations of the noise-free signal are shown in Figure 3.32.

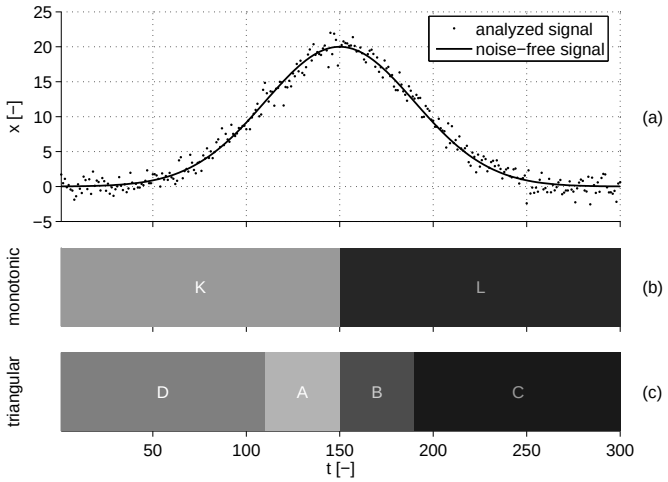


Figure 3.32: Example 2 – simulated series and its true qualitative representation: (a) Simulated series; vertical lines indicate extrema (–) and inflection points (––) of the corresponding noise-free series, (b) monotonic representation (KL) and (c) triangular representation (DABC).

Example 3 In the third example, a time series consisting of two steps is sampled at equal intervals with white noise added. The following equation is used for this signal:

$$x_t = z_{t,1} + e_t \tag{3.81}$$

where:

$$z_{t,1} = \begin{cases} 0 & t < \frac{1}{6} \cdot N \\ \alpha & \frac{1}{6} \cdot N \leq t < \frac{3}{6} \cdot N \\ 0 & t \leq \frac{3}{6} \cdot N \end{cases}$$

e_t : white noise sequence with variance $\sigma_e^2 = 1$

t : time index (integers from 1 to N)

N : length of the time series (300)

α : box magnitude parameter (15)

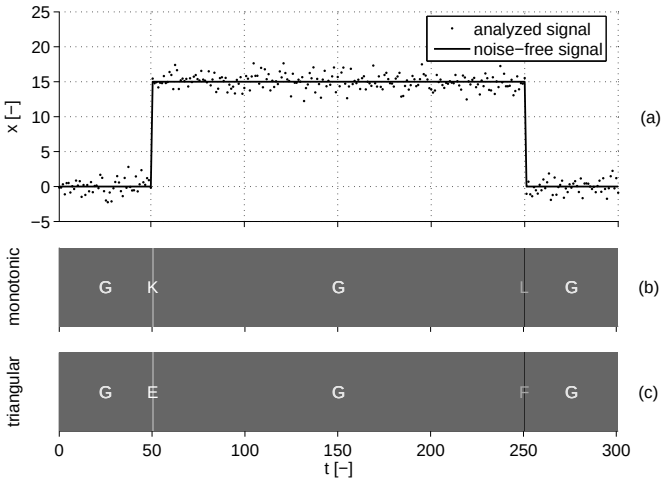


Figure 3.33: Example 3 – simulated series and its true qualitative representation: (a) Simulated series, (b) monotonic representation (GKGLG) and (c) triangular representation (GEGFG).

The resulting data as well as the qualitative representations of the noise-free signal are shown in Figure 3.33.

3.5.5 QRT by means of the cubic spline wavelet

In this section, the original method by Bakshi and Stephanopoulos (1994) is explained and illustrated. This method is improved for inflection point detection in Chapter 8 and is applied in Chapters 9 and 10. Example 1 (Figure 3.31) will be used for illustration.

Method outline Three essential steps can be discriminated within the reviewed method. The first step consists of the wavelet decomposition. The second deals with the construction of the wavelet interval tree. The third and final step encompasses the selection of relevant features such as extrema and inflection points. Figure 3.34 shows a scheme of the complete method. Each of the steps is explained here in the following paragraphs.

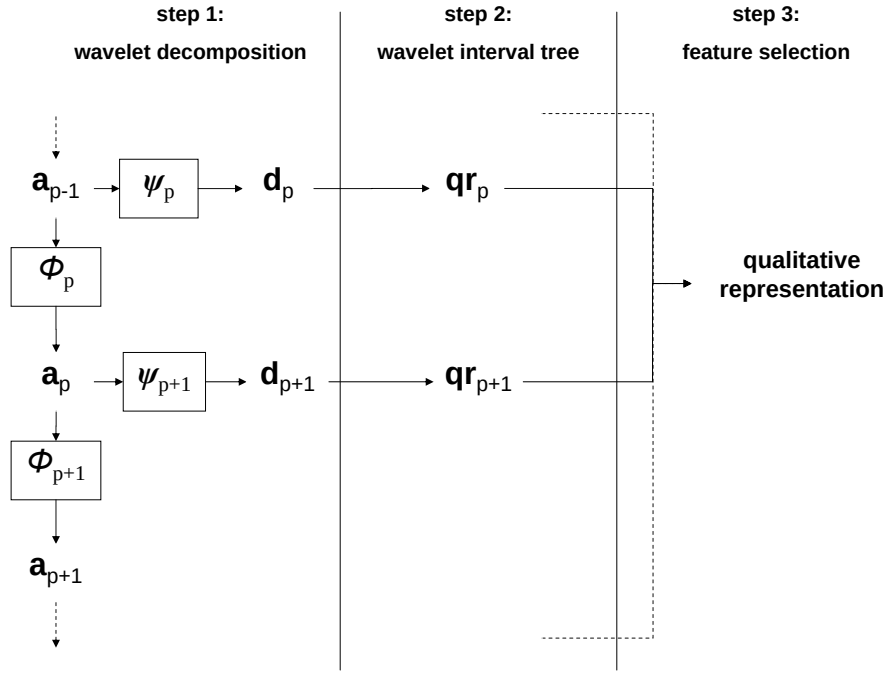


Figure 3.34: Scheme of the complete method. See text for detailed information.

Step 1: Wavelet decomposition To compute the wavelet coefficients, the signal is practically analyzed by means of a filter bank. Such a filter bank is shematically shown in Figure 3.34. Starting with the raw signal, the signal is convoluted with the cubic spline wavelet for the highest analyzed frequency band, referred to as wavelet scale index, p, equal to zero. The spline wavelet, ψ_{p}, acts here as a high-pass filter. Simultaneously, the signal is convoluted with the corresponding low-pass filter wavelet, ϕ_{p}. The signal resulting from low-pass filtering is called an approximation, denoted as a_{p}, whereas the signal resulting from high-pass filtering is called a detail, denoted d_{p}. The resulting approximation, a_{p}, is then further deomposed by means of the next set of pair-wise wavelets, $\phi_{\mathrm{p+1}}$ and $\psi_{\mathrm{p+1}}$. Each time, the cutoff frequency of the low-pass and high-pass filters is lowered and new pair-wise sets of approximations and details result. This process may continue without end for infinite-length signals. Practically, for finite-length signals (say length N), the assessment of the behaviour for frequencies lower than 2/N has little meaning. Therefore, a limited range of filters is practically feasible. The first set of filters (highest cutoff frequency) is indicated by scale index, p, 0. Call P the maximal scale index, being the index of the last pair of filters. The relation between the scale index, p, and the previously defined wavelet scale, s, is written as follows:

$$s = s_o \cdot 2^{p \cdot \delta p} \tag{3.82}$$

where:

s : wavelet scale

s_o : finest applied wavelet scale (2)

p : wavelet scale index

δp : wavelet scale resolution (1)

The parameter δp defines how many scales are used within one octave, i.e. per halving of the frequency. Setting δp to 1, as throughout this work, results in halving the cutoff frequency with each scale index increment. The low-pass filter wavelet corresponding to the mother wavelet is often referred to as the father wavelet or as the scaling function. Mallat and Zhong (1992) define the cubic spline mother wavelet for a given scale as follows in the frequency domain:

$$\psi_0^* (\omega) = \left(\frac{sin\left(\frac{\omega}{4}\right)}{\frac{\omega}{4}} \right)^4 \tag{3.83}$$

The corresponding father wavelet or scaling function is defined as follows in the frequency domain:

$$\phi_0^*(\omega) = i \cdot \omega \cdot \left(\frac{sin\left(\frac{\omega}{2}\right)}{\frac{\omega}{2}} \right)^3 \tag{3.84}$$

To compute the wavelet functions in the frequency domain according to a given wavelet scale index, p, the following equations are applicable:

$$\psi_p^*(\omega) = \psi^*(\omega, s) = \psi_o^*\left(\frac{\omega}{s}\right) = \psi_o^*\left(\frac{\omega}{2^{p \cdot \delta p}}\right) \tag{3.85}$$

$$\phi_p^*(\omega) = \phi^*(\omega, s) = \phi_o^*\left(\frac{\omega}{s}\right) = \phi_o^*\left(\frac{\omega}{s_o \cdot 2^{p \cdot \delta p}}\right) \tag{3.86}$$

The process of wavelet decomposition by means of the cubic spline wavelet filter bank is demonstrated in Figure 3.35 for the signal shown in Figure 3.31. For reasons of clarity only results for scale indices 5 to 9 (approximations) and corresponding 6 to 10 (details) are shown. The approximations at scale indices 1 to 4 retain a lot of noise leading to rapid alternations between qualitative primitives (not shown).

Step 2: From wavelet coefficients to wavelet interval tree At each scale in the wavelet decomposition the qualitative representation of an approximation can be derived on the basis of the detail at the next level (next higher scale). Indeed, the cubic spline wavelet exhibits the interesting property that the locations of zero-crossings in the detail (locations in time where the sign changes) correspond to locations of maxima and minima in the approximation where the detail is derived from. The direction of a sign change corresponds to the type of extremum (maximum/minimum). Also, extrema (maxima and minima) in a detail correspond to inflection points in the approximation the detail is derived from. Assessed extrema and inflection points are indicated in Figure 3.35 by solid and dashed vertical lines respectively. As the zero-crossings and inflection points thus indicate changes in the sign of the first and second derivative, the monotonic and triangular episodes -contiguous windows in which the sign of the first derivative, resp. the first and the second derivative does not change- can straightforwardly be assessed at each scale. From this assessment, a series of characters (an artificial word) which represents the series of monotonic or triangular episodes is obtained. Each character in this series accords to the primitive describing the behaviour in the corresponding

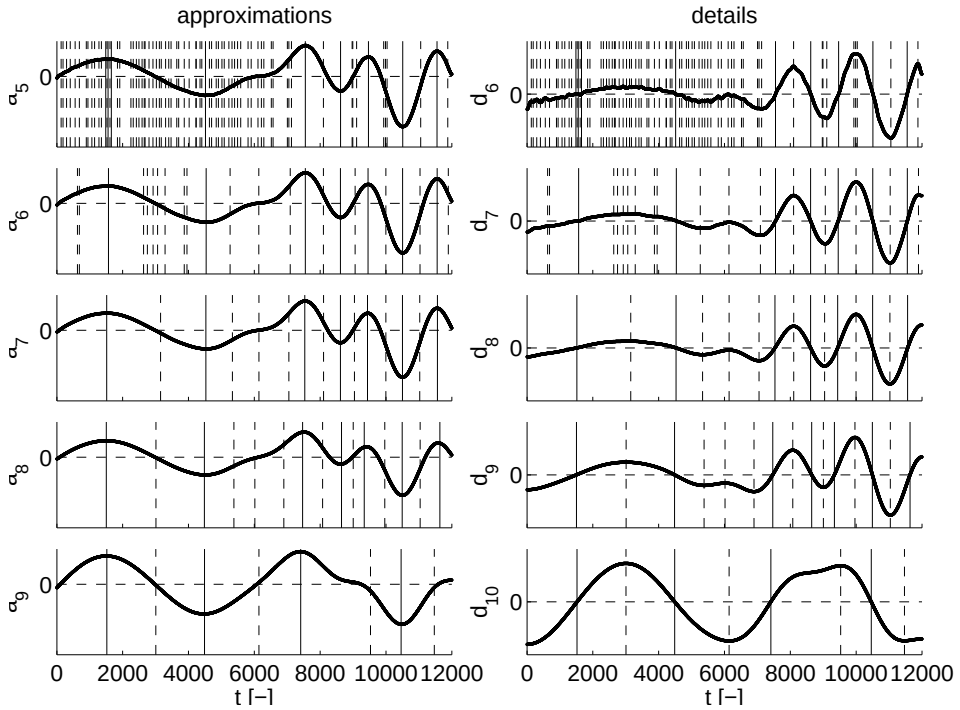


Figure 3.35: Example 1 – Cubic spline wavelet decomposition of the simulated series as in Figure 3.31(a); approximations (scale indices 5-9) and corresponding details (scale indices 6-10). Solid vertical lines indicate extrema in the approximations (zero-crossings in details). Dashed vertical lines indicate inflection points in the approximations (extrema in details).

episode. The approach followed in the original method of Bakshi and Stephanopoulos (1994) includes a simplification of the triangular presentations by removing pairs of contiguous inflection points prior to further processing. For instance, triangular presentations like ADA are automatically replaced by a single A episode. Sets of contiguous inflection points are therefore replaced by the inflection point in that set with maximal (minimal) value for the detail signal in case of an upward (downward) trend. Bakshi and Stephanopoulos (1994) report that this is acceptable for most practical applications.

When the qualitative representations are generated at each scale, the extrema and inflection points that correspond to each other at consecutive scales are linked up

with each other. Figure 3.36 and 3.37 show the so-called wavelet interval trees that result. Episodes that are not split into more episodes going from coarser to finer scales are shown jointly as a single polygon. For instance, at scale index 9 in Figure 3.36, the approximation is represented by a KLKLK sequence ($(KL)_2K$). The first three of these monotonic episodes are not split into more episodes until scale index 4, respectively 5 and 4 and are thus shown as polygons ranging over scale indices 5 to 9, 6 to 9 and 5 to 9.

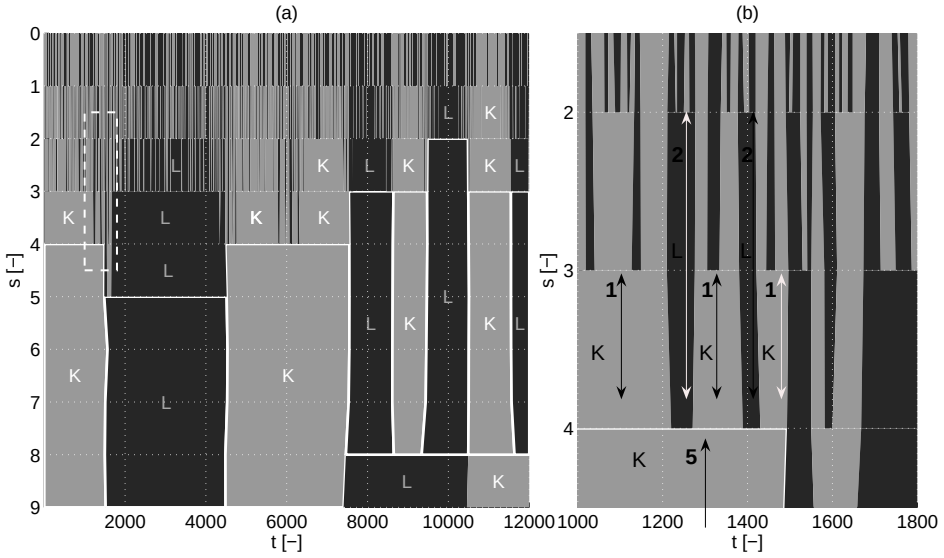


Figure 3.36: Example 1 – (a) Wavelet interval tree (monotonic). Selected episodes are indicated by white contours. Monotonic presentation: $(KL)_4$. (b) Detailed part of the wavelet interval tree, indicated in (a) by a dashed rectangle.

Step 3(a): Monotonic episode selection. Given the qualitative representations at each scale, relevant and irrelevant features need to be discriminated to arrive at a single qualitative representation. This is obtained by application of Witkin's stability criterion to the monotonic episodes. Witkin's stability criterion says that a split into more episodes (at finer scales) is acceptable only if the mean range of the episodes in the more refined presentation is larger than the range of scales over which the coarser episode exists. To understand the application of Witkin's stability criterion, consider the first episode at scale index 9 (time index 0 to 1520).

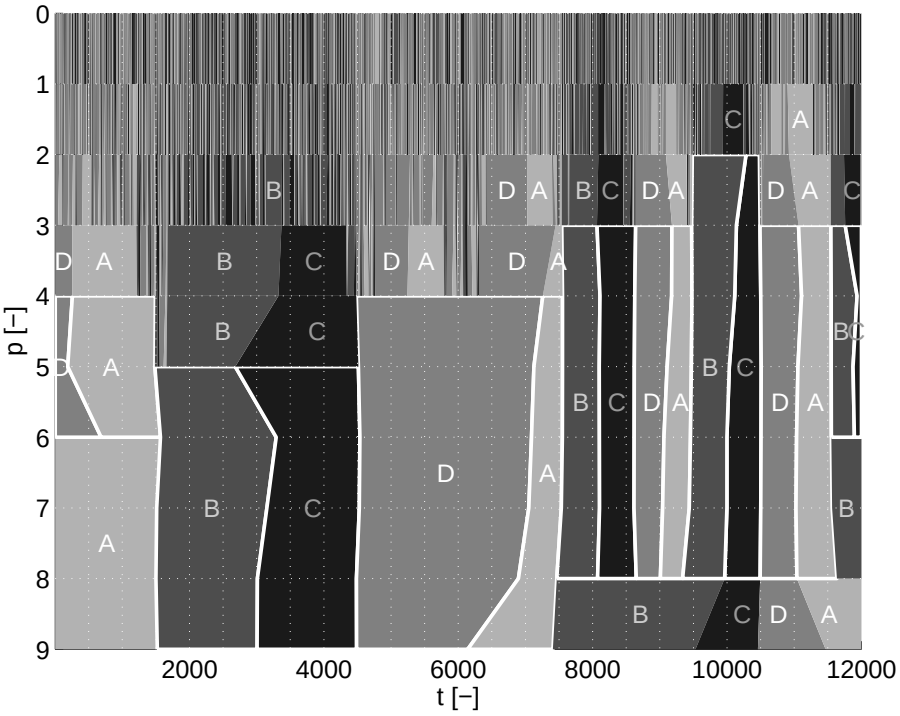


Figure 3.37: Example 1 – Wavelet interval tree (triangular) in the single inflection point approach (approach 2). Selected episodes are indicated by white contours. Triangular presentation: $(\mathsf{DABC})_4$

This episode exists over scale index 5 to 9 (range of scales = 5). At scale index 4, the K episode is split into a KLKLK sequence, in which the episodes exist over 1, 2, 1, 2 and 1 scale(s) (see Figure 3.36), giving a mean range of scale indices of 1.4. Consequently, in this case the split is not accepted ($1.4 < 5$). Consider now the fourth monotonic episode at scale index 9 (L, time index 7408 to 10460) which exists over scale index 9 only (range of scales = 1). At scale index 8, the episode is split into 3 episodes (LKL) with scale index ranges 5, 5 and 6, giving a mean range of 5.33. By application of Witkin's stability criterion, this split is thus accepted ($5.33 > 1$). By applying this criterion from coarsest scale index (9) towards the most detailed scale (1) until none of the considered episodes can be split further, the (monotonic) qualitative representation of the time series is established, being $(\mathsf{KL})_4$. Note that this matches the monotonic representation of the noise-free signal

(see Figure 3.31), thus meeting the desired goal.

Step 3(b): Triangular episode selection. Following the assessment of relevant extrema and thus of assessment of relevant monotonic episodes, the inflection points are assessed. In the approach specified by Bakshi and Stephanopoulos (1994), the triangular episodes present at the most detailed scale of each representing episode are included in the representation. Consider for instance the first selected monotonic episode in Figure 3.36. This episode exists over scales 5 to 9 (K, time index 1 to 1486 at scale 5). Consequently, the included triangular episodes at scale 5 are included in the triangular representation (DA). The next relevant monotonic episode exists over scales 6 to 9 (time index 1486 to 4536 at scale 6) and consequently the included triangular episodes at scale 6 are included (BC). The same procedure is continued for all representing monotonic episodes. Note that due to the simplification of the qualitative representations in terms of inflection points, a DA triangular sequence results for the third accepted monotonic episode. This representation is simpler than the desired sequence of the noise-free signal (see Figure 3.31).

Limitations In Figure 3.38 and 3.39, the results obtained for Example 2 and Example 3 are shown. For both signals, the obtained qualitative representations are KL (monotonic) and DABC (triangular). While the representations for Example 2 match the targeted outcome (see 3.32), this is not true for Example 3. Due to the smoothing properties of the cubic spline wavelet that the steps in the signal are assessed as inflection points. The method thus fails to discriminate between true inflection points (where the 1^{st} and 2^{nd} derivative of the underlying signal is continuous) and step changes (discontinuities). This was already observed by Bakshi and Stephanopoulos (1994) but no explicit solution to this problem was provided. Note that the assessed inflection points are located at the locations in time where the step changes occur. Despite the inability to discriminate between jumps and inflection points, the identified inflection points are thus consistent to some extent with the noise-free signal. Indeed, the jumps in the signal correspond to locations in the signal where the magnitude of change is maximal. For a differentiable function, such points are inflection points by definition.

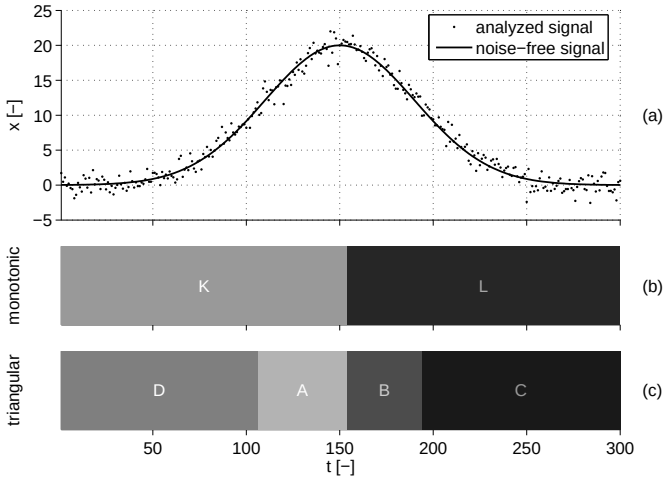


Figure 3.38: Example 2 – simulated series and obtained qualitative representation by means of the method of Bakshi and Stephanopoulos (1994). (a) Simulated series, (b) monotonic representation (KL) and (c) triangular representation (DABC).

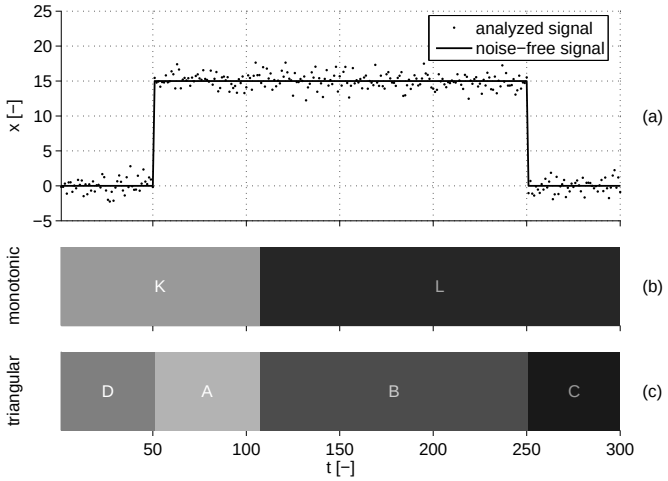


Figure 3.39: Example 3 – simulated series and obtained qualitative representation by means of the method of Bakshi and Stephanopoulos (1994). (a) Simulated series, (b) monotonic representation (KL) and (c) triangular representation (DABC).

3.5.6 QRT by means of interval-halving

In this section, the method for QTR based on interval-halving by Dash et al. (2004a) is explained.

Method outline Essential to this method is that the (least-squares) fit of a polynomial function to a data series allows to evaluate the sign of the first and/or second order derivative of that function in each point of the series. As a result, if the fitted function delivers a good fit to the data, a valid estimate for the first (second) order derivative can be obtained in each point of the section by evaluation of the first (second) derivative. In view of QRT, an evaluation of the sign of the respective derivatives is sufficient to evaluate the qualitative behaviour of the fitted series. A necessary condition for this method to work is that the fitted function fits sufficiently to the data. To this end, linear and quadratic functions are fitted to contiguous segments of the analyzed series, as opposed to increasing the polynomial order of the polynomial function. The former approach is preferred in view of computational complexity (Dash et al., 2004a). Fundamental to this approach is that *any function can be approximated to any level of detail using a sequence of piece-wise polynomials*, as stated by the original authors. The provided algorithm consists of the following two steps:

1. Identifying the set of contiguous segments to which (constrained) polynomials of first or second order fit sufficiently by means of an F-test. Such segments are referred to as unimodal regions.
2. Translation of the fitted parameters of the (constrained) polynomials into qualitative primitives.

The first step is intrinsic to the algorithm proposed by Dash et al. (2004a). Given a segment of the series over which a polynomial is to be fitted, a constant function is fitted firstly (order = 0). To test whether this function fits the data sufficiently, Dash et al. (2004a) construct a lack-of-fit test based on the F-test. This test assumes independent and identically distributed (i.i.d.) errors and needs knowledge on the noise variance. Such knowledge is generally speaking not readily available and therefore a noise estimate is provided by means of wavelet decomposition. The theoretical support for this approach is given by Donoho and Johnstone (1994). If the provided lack-of-fit test fails, i.e. the function does not fit good enough to

the data, the order of the polynomial is increased until the fit is good enough or until order 2 is reached. If the same lack-of-fit test continues to fail at order 2, the considered segment is split into two equal halves and the procedure is repeated for each of the halves. In the Matlab implementation by Dash et al. (2004b) the *left*-half of an interval is studied first so that a *left-to-right* direction of analysis results. This is practically meaningful in the context of streaming data. Indeed, as such, intervals in historical data once identified do not need to be analyzed again as new data points are generated. Upon the assessment of acceptable fits to identified segment, the method provides identification of jump changes. This is achieved by comparing the fit of two polynomials in two contiguous segments with and without the constraint that the signal is continuous in the time location in between the two segments. If the non-constrained fit is significantly better than the constrained fit (based on an F-test), then a jump change is identified in between the two segments. Note that the algorithm is especially suited for on-line use as the assessment of qualitative features is implemented in a *left-to-right* direction. The fitted parameters in former segments are used to constrain the fits in following segments. For a more detailed explanation of the interval-halving algorithm and the constrained polynomial fits, the reader is referred to the original text (Dash et al., 2004a). The algorithm requires some parameters to be set:

- Minimal time window for a segment. This was set to 4 throughout this section.
- Confidence limit for the constructed lack-of-fit tests. This was set to 95% for all segments.
- Applied wavelet and the decomposition level for noise estimation. The 3rd-order Daubechies wavelet ('D6') was used and the maximal practically computable level was used as by the original authors.
- The option to repeat noise estimation on identified segments prior to the fit of polynomial. This was set on.

The second step in the presented method is more generic of nature and consists of the evaluation of the sign of the derivatives of the fitted polynomials over the respective intervals and the assignment of the corresponding primitive. Note that the sign of the second derivative will never change over the course of one segment (linear and quadratic polynomials have a unique value for the second derivative). This is however not true for the first derivative. Indeed, a fitted quadratic polynomial may represent a curve with a maximum or minimum within the segment. As such,

a single quadratic segment may lead to two episodes (monotonic and triangular). While Dash et al. (2004a) use another assignment of characters, the same *alphabet* as for the previously presented method is used here (see Figure 3.30) for reasons of consistency.

Illustrations The qualitative representation of Example 2 on the basis of the reviewed method is shown in Figure 3.40. In the upper part of the figure (3.40(a)), one can see that four segments are identified in the series. A linear curve is fitted to the first segment (time index 1 to 75). The second segment (time index 75 to 130) begets a linear fit as well. A quadratic curve is fitted to the third and fourth segment (time index 130 to 171 and 171 to 300). The qualitative representations follow from the evaluation of the derivatives of the fitted curves.

An important advantage of the method by Dash et al. (2004a) over the method by Bakshi and Stephanopoulos (1994) is that jump changes can be handled. To illustrate this, the results for Example 3 by means of the former method are shown in Figure 3.41. As can be seen, the flat sections and jump changes in the (noise-free)

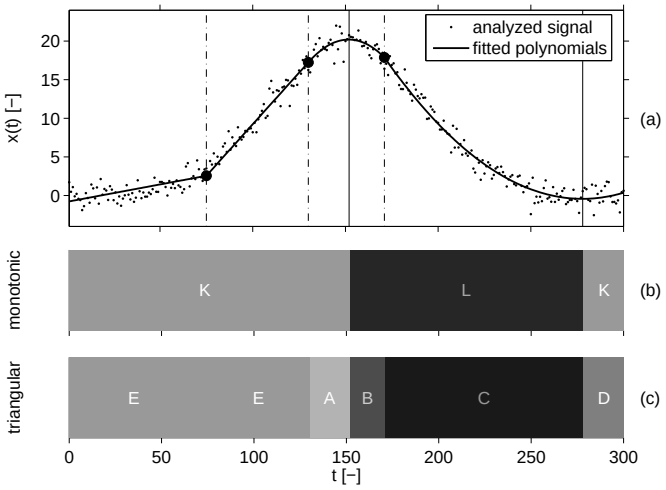


Figure 3.40: Example 2 – simulated series and obtained qualitative representation by means of the method of Dash et al. (2004a). (a) Simulated series, (b) monotonic representation (KLK) and (c) triangular representation (EEABCD).

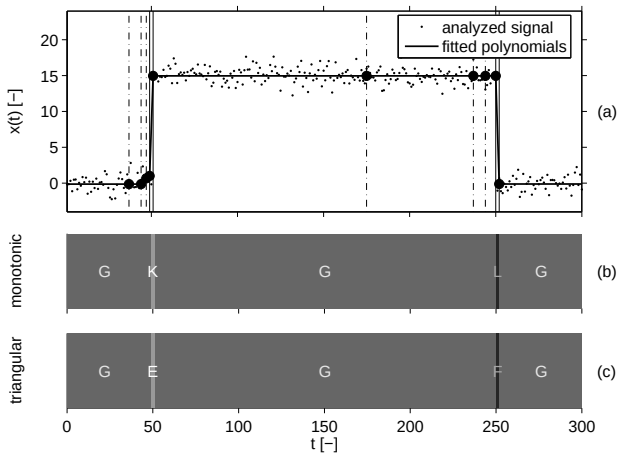


Figure 3.41: Example 3 – simulated series and obtained qualitative representation by means of the method of Dash et al. (2004a). (a) Simulated series, (b) monotonic representation (GLKGLG) and (c) triangular representation (GFEGFG).

signal are properly addressed, in contrast to the results shown for the method by Bakshi and Stephanopoulos (1994). The resulting representation is thereby different from the obtained representation of Example 2 (see Figure 3.40). The method therefore allows to discriminate signals true inflection points from inflection points.

Limitations Dash et al. (2004a) mention that the proposed method does not necessarily lead to the same solution for a same-shaped signal. This lack of robustness is paid off by a very fast assessment of the qualitative behaviour. As discussed above, the provided algorithm works (necessarily) in a *left-to-right* direction. To evaluate what the effect of this directional analysis can be, the signal in Figure 3.40 (same noise-free signal, same noise sequence), the direction of the algorithm was reversed so that the algorithm works in a *right-to-left* direction. Results are shown in Figure 3.42. As can be seen, the resulting qualitative representation is different from the previously obtained one. An influence of the direction of analysis is thus present and confirmed by the original authors (Maurya Dash, personal comm.). Such dependence may impede a valid analysis as the targeted qualitative representation is essentially not depending on the direction of interpretation. Also, for the same noise-free signal, different results (for different noise sequences) are frequently obtained (not shown), indicating a lack of robustness.

It is also worth studying the first identified segment in detail (time index to 0 to 130). To this segment of the series a quadratic function is fitted. The resulting function exhibits a minimum within the covered timeframe, resulting in a LK sequence when translated into monotonic primitives and a CD sequence when using triangular primitives. However, it may be suspected that a single L (monotonic) or C (triangular) sequence may be equally or more valid for this segment. Given that the fit of a quadratic segment may lead to 1 or 2 episodes, controlling the complexity of the fitted piece-wise polynomials by means of lack-of-fit tests does not lead to an equal control of the complexity of the qualitative representation. Possibly, this may be tackled by constraining the fitted polynomials to have unique sign of the first derivative within the respective covered timeframes. By doing so, the lack-of-fit test may result in an acceptance of a polynomial in the first segment which is then translated into a single L (monotonic) and D (triangular) primitive. If successful, the second segment should then be split into two segments leading to two separate polynomials which correspond to K, resp. L, (monotonic) and A, resp. B, (triangular) primitives. By applying such a constraint one can thus achieve direct control over the complexity of the qualitative representation. This proposed change of the method has not been tested as yet.

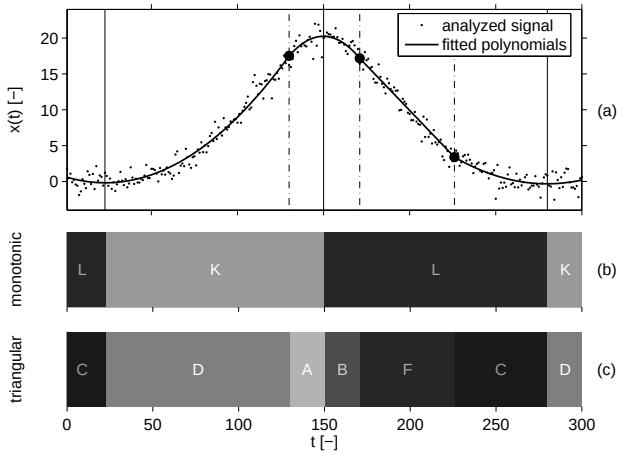


Figure 3.42: Example 2 – simulated series and obtained qualitative representation by means of the method of Dash et al. (2004a) in reverse direction. (a) Simulated series, (b) monotonic representation (LKLK) and (c) triangular representation (CDABFCD).

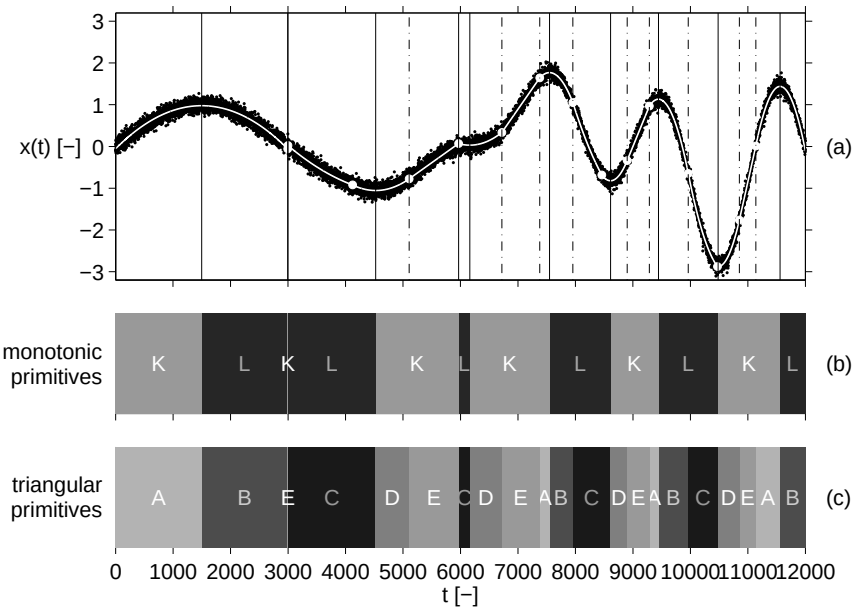


Figure 3.43: Example 1 – simulated series and obtained qualitative representation by means of the method of Dash et al. (2004a) in reverse direction. (a) Simulated series (black) and fitted polynomials (white), (b) monotonic representation ($(\mathsf{KL})_6$) and (c) triangular representation ($\mathsf{ABECDE}(\mathsf{CDEAB})_3$).

For completeness, the signal used for illustration of the method by Bakshi and Stephanopoulos (1994) is also analyzed by the method by Dash et al. (2004a). Results are shown in Figure 3.43. As can be seen, an (incorrect) K primitive is identified between time indices 2996 to 3000 (thus being of minimal allowed length). Such artefacts can easily be avoided by increasing the minimal allowed length of a segment. Another (undesired) L primitive is identified between time indices 5967 and 6164. Increasing the minimal segment length so to avoid this may however not be possible in practice (e.g. if short-term events are plausible and targeted for identification). Also, the fitted segments suggest a jump in the derivative which is not present (in the noise-free signal). When considering the identified triangular primitives, one can observe that some of the desired DA sequences are replaced by DEA sequences thus inserting a linear segment. Several solutions exist to counter this:

- Replace DEA, resp. BFC, sequences by DA, resp. BC sequences. A formal

approach to do so is provided by Cheung and Stephanopoulos (1990b).

- When processing qualitative representations further, consider DEA and DA sequences, resp. BFC and BC sequences as very similar. In other words, when comparing two time series in which one exhibits a DEA sequence and the other a DA sequence, consider the two sequences only being different to a minimal extent. An approach to do so based on the incorporation of numerical information is presented in Maurya et al. (2002).

3.5.7 Concluding remarks, method selection and suggestions

Importance of inflection points Bakshi and Stephanopoulos (1994) and Dash et al. (2004a) independently chose not to evaluate the performance of their method in terms of the identification of inflection points. Bakshi and Stephanopoulos (1994) state that the discrimination of time series on the basis of inflection points is practically of little interest. However, as explained in Section 3.5.1, inflection points in, for instance, ORP signals have been shown to indicate critical points for biological nitrification and denitrification systems. As such, the latter context provides an exception to the statement by Bakshi and Stephanopoulos (1994). The statement may be considered void as well for any ORP, resp. pH signal, stemming from processes exhibiting multiple oxidation-reduction, resp. pH, buffer systems. Dash et al. (2004a) do not focus on the identification of inflection points either and support this by stating that (1) the provided method does not allow inflection point detection easily and (2) stating that inflection points have limited effect on the discrimination of time series. Importantly, to understand the latter statement, it is worth noting that the similarity of time series is defined as in the study by Maurya et al. (2002). In the latter work a similarity measure for time series is developed and applied which is not only based on the qualitative shape of a trend in a segment of a series but also on numerical information. This numerical information is obtained by first normalizing the signal so that the start value is zero and the end value is one. Then, a shape-based similarity is calculated as the integral of the normalized signal. A A-like shape will then typically result in a larger value for the latter integral compared to a D-shaped signal (both share the same monotonic K behaviour).

As a result, information on acceleration and deceleration in identified segments of a signal is incorporated in the similarity index. The problem of inflection point identification is thus overcome by incorporation of numerical information even if not stated explicity. In conclusion, inflection points as targeted qualitative features of (time) series should not be excluded a priori.

Smoothness The method of Bakshi and Stephanopoulos (1994) is aimed explicitly at smooth signals (continuous up to second derivative). It has been shown that because of the smoothing properties of the cubic spline wavelet jump changes cannot be discriminated from regular inflection points. Also, flat sections are typically not identified as such because of these smoothing properties. Contrastingly, the method by Dash et al. (2004a) allows to discriminate flat sections and jump changes but also results in non-smooth approximations and presentations of smooth signals and the insertion of undesired primitives in the resulting qualitative behaviour.

The problem of jump detection for the cubic spline wavelet method has not been tackled in this dissertation. Yet, a solution, which is not tested as yet, is presented here. Conventional moving range detection schemes based on filters may aid to identify potential jump changes. Alternatively, already obtained wavelet coefficients in the cubic spline wavelet can be used. Indeed, jump changes are likely to lead to high values for the wavelet coefficients at the lowest scales (i.e. fastest frequency bands) compared to the coefficients not influenced by the jump change. In other words, *highly energetic* wavelet coefficients localized crisply in time may indicate jump changes. An exact cutoff for the respective coefficients may be based on knowledge of the normal noise spectrum. As a third alternative, the interval-halving method may be applied to identify jump changes.

Given the identification of potential jump changes, the signal may be split and analyzed separately, as in the interval-halving method. Each segment may be analyzed by means of the cubic spline wavelet method, as the method is suited for smooth signals. However, given the identification of potential jump changes, the complete signal may be approximated by jointly deriving the magnitude of the jump, the filtered signal (say by means of the first cubic spline wavelet scale) and the (white) noise parameters. To explain in more detail, consider that one writes the signal as follows:

$$x_t = z_t + \alpha \cdot g_t + e_t \tag{3.87}$$

where:

x_t: the original noisy signal

z_t: the underlying nosie-free signal without jump change

g_t: unit step change at time t_{jump}

α: magnitude of the jump change

e_t: random Gaussian error (white noise)

t: time index (1..N)

Having identified t_{jump} as a probable location of a jump change, an approximation of the signal can be written as follows:

$$\hat{x}_t = a_{t,1} + \hat{\alpha} \cdot g_t \tag{3.88}$$

where:

$\hat{x}_t$: estimate of the signal

$a_{t,1}$: wavelet approximation at scale 1 of the signal without jump change

$\hat{\alpha}$: estimated magnitude of the jump change at t_{jump}

Now, under the assumption of Gaussian white noise, the unbiased estimates of the parameters in the former formula can be found by minimizing the sum of squared residuals, SSR, by adjustment of $a_{1,t}$ and $\hat{\alpha}$:

$$SSR = \sum_{t=1}^{N} (\hat{x}_t - x_t) \tag{3.89}$$

This may practically be achieved by estimating $\hat{\alpha}$ and $a_{t,1}$ iteratively until convergence of SSR. Given that the approximation by the cubic spline wavelet at the first scale has $N/2 + 3$ degrees of freedom, the approximation above (with 1 extra parameter) has $N/2 + 4$ degrees of freedom. Practically, this means that one constraint of the cubic spline wavelet approximation, being continuity at time t_{jump},

is dropped. Interestingly, if using the cubic spline wavelet, the approximation $a_{1,t}$ is constrained to have smooth derivatives up to the second derivative. Splitting segments and analyzing them separately cannot guarantee such smoothness, i.e. splitting may result in fitting of jump changes in the derivatives, i.e. adding more degrees of freedom, which may not be valid. In fact by doing so, the degrees of freedom for the derived approximation amount to $N/2 + 6$, as the constraints for smooth 1st and 2nd derivatives are dropped as well. As such, edge effects at the split are avoided by not splitting the signal. In addition, upon acceptance of the jump change (based on improved fit), the approximation $a_{1,t}$ may be used for further analysis by means of the cubic spline wavelet. Indeed, the non-smooth feature in the series would be removed in a least-squares optimal sense, not affecting the cubic spline wavelet decomposition anymore. As such, the suggested approach may allow the use of the cubic spline wavelet approach, shown earlier to be more robust for smooth series, while not being limited anymore by jump changes.

Method selection In this dissertation the method by Bakshi and Stephanopoulos (1994) was chosen as a basis for the presented work for the following two reasons. First, the analyzed signals are of a smooth natsure and the problem with respect to step changes is not present. Secondly, the method by Dash et al. (2004b) was observed to lead to different results for the same (noise-free) signal depending on the direction of the analysis and upon repetitions of the noise simulation. An argument contra the chosen method is that sequences of inflection points in between extrema cannot be identified properly. The method is however consequently improved in 8 to overcome this problem.

Chapter 4

Description of studied system, obtained data and observed problems

The contradictory of a welcome probability
will assert itself whenever such an event
is likely to be most frustrating

Gumperson's Law

In this chapter, the pilot-scale SBR system, being the core subject of study, as well as the data derived from this system to do so, will be described. The SBR reactor was conceived in 2001 (Capalozza, 2001) for production of a stable microbial culture of which settling properties were studied (Govoreanu et al., 2003). The SBR itself has been used also as a study object for development of expert calibration procedures for mechanistic modelling (i.e. based on knowledge) of wastewater treatment plants (Insel et al., 2006) and model-based optimization (Sin et al., 2006). In parallel, methods for data-driven modelling have been developed and tested in the context of process monitoring and diagnosis of the SBR system as well (Lee

and Vanrolleghem, 2003; Yoo et al., 2004; Lee and Vanrolleghem, 2004; Lee et al., 2005; Yoo et al., 2006a,b; Villez et al., 2007a,b).

First, a description of the system is given after which typical data are described. In view of the monitoring and diagnosis strategies tested in Chapter 5 and Chapter 6, a detailed overview of faults identified during the studied period will be given in the last and third section of this chapter.

4.1 Description of the pilot-scale Sequencing Batch Reactor

The pilot-scale SBR has a working volume of 64 l and has been reseeded in view of the studies reported upon in this dissertation with sludge from the Destelbergen wastewater treatment plant (Aquafin, Destelbergen, Belgium) in August 2005. The targeted hydraulic retention time (HRT) and solids retention time (SRT) were 12 hours and 10 days, respectively. The starting and end volume of each cycle is 34 liter, so that a 50% volumetric exchange ratio results. Detailed information on the scheduling of the phases of the system is given in the next section.

In Figure 4.1, a scheme of the studied SBR reactor is given. Central in the scheme stands the reactor on the balance. The balance delivers the weight measurement directly to the PC by means of an RS-232 connection. Four sensor probes are in direct contact with the bulk liquid in the reactor, being probes for Dissolved Oxygen (DO), Oxidation-Reduction Potential (ORP), pH and conductivity. The DO probe also delivers the temperature measurement. All of the latter probes are connected to respective transmitters in the data acquisition box (DAQ box), which in turn connects the amplified signals to the Data Acquisition card within the PC, equipped with NI LabView 7.0.

Submerged mixers are put to guarantee mixing during non-aerated periods. Aeration of the reactor is achieved by air flow through a submerged flexible tube provided with air nozzles. The air flow rate can be controlled between 0 and 30 l/min by means of a solid gas valve.

The hydraulic parts of the system are set up around a single one-directional pump. By means of two three-way pinch valves just before and after the pump, the direc-

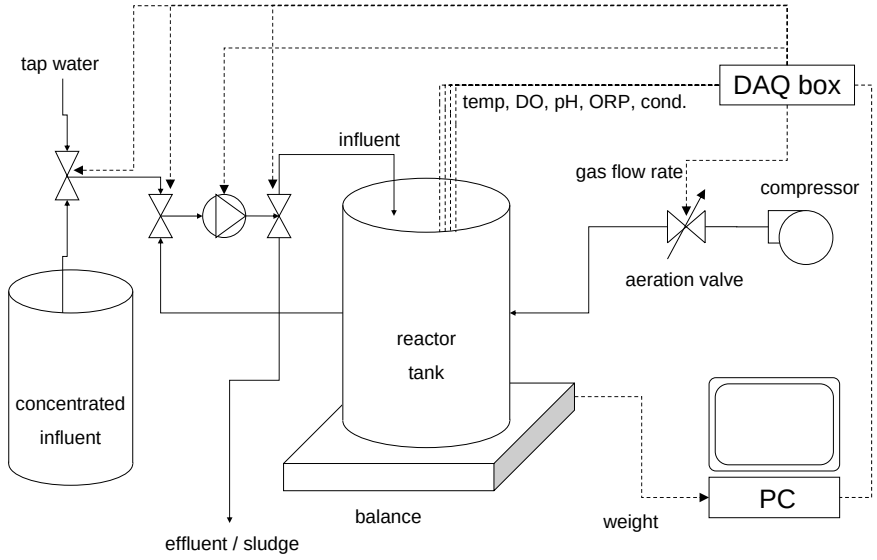


Figure 4.1: Scheme of the SBR reactor.

tion of the flow (to or from the reactor) can be controlled. An additional three-way pinch valve selecting between tap water and concentrated influent is used to obtain the desired (average) concentration of the influent. During filling, this valve selects the concentrated effluent during 2 seconds each 30 seconds so as to obtain a 1/15 of the concentrated influent. The concentrated influent is synthetic and its chemical composition is similar to real pre-settled domestic wastewater Boeije (1999). Table 4.1 summarizes the influent composition following the characterization of the (diluted) influent according to Boeije (1999).

Table 4.1: Influent wastewater characterization (adopted from Boeije (1999))

component	concentration	unit
Total COD	411	mg COD/l
Particulate Inert COD	18	mg COD/l
Soluble Inert COD	18	mg COD/l
Biodegradable COD	375	mg COD/l
Fermentable COD	95	mg COD/l
Acetate COD	70	mg COD/l
Slowly biodegradable COD	210	mg COD/l
Ortho-Phosphate (PO_4^-)	11	mg P/l
Total Kjeldahl Nitrogen (TKN)	63	mg N/l
Ammonium nitrogen (NH_4^--N)	3	mg N/l
Soluble biodegradable nitrogen	10	mg N/l
Particulate biodegradable nitrogen	50	mg N/l

4.2 Operational modes

The data used in this dissertation stem from batches logged between January, 1st of 2006 and March, 3rd of 2007. In this period, data of 1407 complete batches have been recorded. Over this period, 4 different phase schedules were applied, referred to as operational modes from here on. The applied schedules are discussed in detail first. Two different oxygen setpoints have been applied as well during this period. This is discussed in detail later on in the text.

In all applied modes, a complete cycle has a fixed length of 360 minutes (6 hours) and consists of an anaerobic phase (ANAER), a first aerobic phase (AER1), an anoxic phase (ANOX), a second aerobic phase (AER2), a settling phase (S) and a draw phase (D). A generic scheme can be found in Figure 4.2. During the first part of the anaerobic phase, the reactor is filled partially with influent. During the first part of the anoxic phase, the reactor is filled further. This provides organic carbon enabling for the denitrification reactions, which would not go on otherwise (as all biodegradable carbon in the bulk liquid is oxidized in the aerobic phase just before). At the end of the second aerobic phase, sludge is withdrawn from the reactor. The cycle ends with effluent withdrawal in the draw phase. Differences

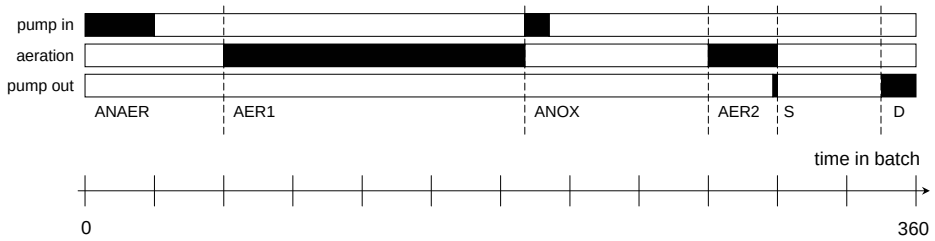


Figure 4.2: Generic operational scheme of a single SBR cycle.

in the applied modes relate to the time length of filling during the anaerobic phase and the time length of the aerobic phase. The anaerobic phase length is always 60 minutes (1 hour). The sum of the length of the first aerobic phase and anoxic phase is always 210 minutes (3.5 hours). In the first three modes, the first aerobic phase and anoxic phase lengths were fixed to 130 and 80 minutes respectively. The fourth mode comprised an on-line optimization scheme for phase length optimization of the first aerobic phase, which is explained in detail in Chapter 7. The second aerobic phase, settling phase and draw phase always take 30, 45 and 15 minutes respectively. Except for the fourth mode, the applied mode differ only in the time length for reactor filling during the anaerobic phase, being 20, 25 and 30 minutes respectively. In the fourth mode, the anaerobic fill time is also 30 minutes. The anoxic filling always starts immediately after ending the aerobic phase and always takes 10 minutes. Note that the applied pump speed setpoint is constant within each batch and is updated after each cycle. To do so, the applied pump speed was updated recursively on the basis of the difference between the desired influent addition and actual influent addition. To avoid extreme changes in pump speed, an exponential filter was applied.

Two different oxygen setpoints have been applied during the studied period for the first aerobic phase. The first 323 batches of mode 1 were operated with a 1.0 mg/l setpoint. All other batches were operated with a 2.0 mg/l setpoint for DO in both phases. Therefore, this mode is split into mode 1a and mode 21b to denote this difference. Within mode 2, 4 batches were used for test runs of the control algorithm deployed in mode 3. While the hydraulics were controlled in a fixed time fashion (hence, mode 2), the aeration system control was tested already as if in mode 3. Therefore, also this mode is split into two modes, mode 2a and mode 2b. Table 4.3 includes all operational parameters for all operational modes of the system.

Table 4.2: Different aspects of operational modes concerning the phase scheduling. (C): on-line control of length of 1st aerobic phase.

mode	# batches	Time length (min.)		
		Anaerobic filling	AER1	ANOX
1	758	25	130	80
2	560	20	130	80
3	89	30	130	80
4	37	30	60-130 (C)	80-150 (C)

Table 4.3: Different aspects of operational modes concerning all operational parameters. (C): on-line control of length of 1st aerobic phase; (T): Testing of on-line control algorithm affects aeration system.

mode	# batches	Time length (min.)			DO setpoint
		Anaerobic filling	AER1	ANOX	AER1
1a	323	25	130	80	1.0
1b	435	25	130	80	2.0
2a	556	20	130	80	2.0
2b	4	20	130 (T)	80 (T)	2.0
3	89	30	130	80	2.0
4	37	30	60-130 (C)	80-150 (C)	2.0

In Chapter 5 and Chapter 6, data from operational modes 1a, 1b, 2a, 2b and 3 are used for development of MPCA models for monitoring and diagnosis of the hydraulics of the system. Data from the operational mode 1a are used for development of MPCA models (see Section 3.3.3.1) for monitoring and diagnosis of the complete SBR system. In Chapter 7, data from mode 4 are studied as the work presented therein explicitly deals with the on-line optimization scheme applied in this mode.

4.3 Description of normal data

Figure 4.3 shows trajectories of all measurements that are recorded on-line in normal operation. Figure 4.3(a) shows the weight profile of the reactor. Clearly, the reactor is filled partly during the first anaerobic phase, whereafter the weight remains constant until the start of the anoxic phase, when additional influent is added to supply COD as electron donor for denitrification. Just before the settling phase sets in, sludge is wasted from the system (1l). At the end of the cycle, effluent is withdrawn. In Figure 4.3(b) the temperature profile of the same batch is shown. While the temperature is aimed to be 15 °C, the addition of (warmer) influent causes the temperature to rise in the reactor, especially at the beginning of the cycle when the amount of new warm influent is relatively large compared to the volume of cold bulk liquid.

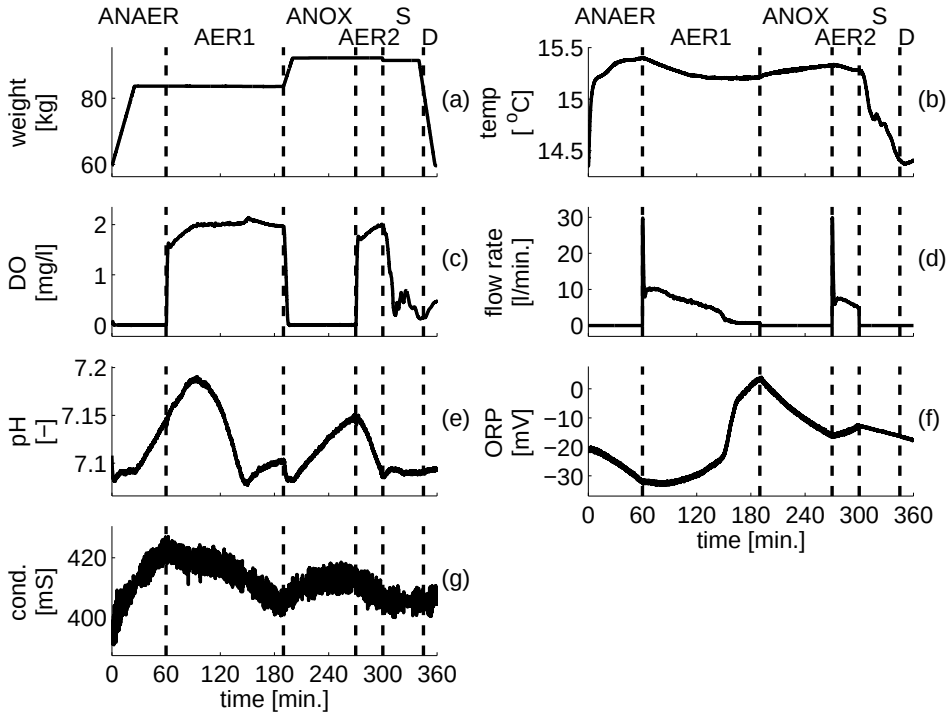


Figure 4.3: Typical profiles of on-line measurements in a single SBR cycle. Dashed lines indicate the programmed start and end of aerobic phases.

The dissolved oxygen (DO) level profile is shown in Figure 4.3(c). As desired, the oxygen level is zero in the anaerobic and anoxic phase and is close to its setpoint (2.0 mg/l in this case) during the aerobic phases. Note that a high DO level is observed in the settling and draw phase. This is explained due to the settling of the biomass. Given that the settling causes some parts of the reactor volume to be void of active biomass, the oxygen is no longer consumed in these regions of the reactor. In this case, the DO sensor finds itself in such a region. Related to the DO profile, a non-zero gas flow rate (Figure 4.3(d)) is observed during the two aerobic phases. The gas flow rate is high at the beginning of each phase, whereafter a more moderate level is reached around. In the first aerobic phase, a significant drop to a minimal level (approx. 2 l/min.) is observed around 150 minutes in the SBR cycle. This is explained as the end of exogenous respiration, i.e. the start of endogenous respiration state, in which no external substrate is oxidized anymore and nitrification is completed.

Figure 4.3(e) shows the pH profile in the studied batch. The pH profile rises during the anaerobic and anoxic phases as a result of the (proton-consuming) denitrification process. At beginning of the respective phase, the pH first decreases, increases slowly and then faster. This is common to the SBR system and is due to the fact that the influent, which is added at the beginning of the respective phases, is more acidic than the bulk liquid in the SBR. In the aerobic phase, (proton-consuming) CO_2-stripping and (acidifying) nitrification play a major role. The pH increase at the beginning of the 1^{st} aerobic phase is due to a dominating effect of CO_2-stripping. As the CO_2 concentration lowers and the nitrification process reaches maximal rate, the pH starts to decrease. When the substrate (ammonia) is completely converted to nitrate, the pH increases again as CO_2-stripping continues. A dominating CO_2-effect nor complete consumption of the substrate can be observed clearly in the 2^{nd} aerobic phase.

The ORP profile (Figure 4.3(f)) shows increasing trends in aerobic phases and decreasing trends in the other phases, as expected. In the 1^{st} aerobic phase, the breakpoint of the nitrite-nitrate buffer, i.e. when virtually all nitrite is converted to nitrate, can clearly be observed at approximately 160 minutes in the batch cycle. The nitrate apex, i.e. a drop in the ORP profile after completion of nitrate reduction, is not observed, hereby indicating that nitrate reduction is not complete. This is typical for the studied system due to COD-limitation of the influent.

The conductivity measurements are shown in Figure 4.3(g) for the studied batch. Interpretation of conductivity measurements has been experienced as difficult, in

part due to the expected influence of both nitrification-denitrification and phosphate-related processes on the conductivity. Nevertheless, the conductivity profile shows an increasing trend during the anaerobic phase explained in part by the phosphate release that typically occurs in this phase. Upon phosphate uptake by Phosphorus Accumulating Organisms (PAO's), the conductivity decreases again. The rise in the anoxic phase is unlikely to be explained by phosphate-release as this requires anaerobic conditions which are not met as oxidized nitrogen compounds such as nitrate remain present throughout the phase. The rise may be explained as the result of influent addition at the beginning of the anoxic phase. The conductivity decreases again in the aerobic phase indicating additional phosphorus uptake. During the settling and draw phase, the conductivity remains largely constant. It is noted that the conductivity measurement is very noisy compared to other measurements. This is the normal case and is not due to sensor malfunctioning.

4.4 Description of effluent quality data

Figure 4.4 shows (off-line) measurements of effluent quality variables (total ammonia nitrogen (TAN), nitrite nitrogen (NO_2^--N), nitrate nitrogen (NO_3^--N) and inorganic phosphorus (PO_4^{3-}-P)) during a single cycle of the studied batch reactor. In the anaerobic phase, (0–60 min.) nitrate is reduced to minimal levels, phosphorus concentrations increase due to phosphorus release by Phosphorus Accumulating Organisms (PAO's) and hydrolysis results in increasing concentrations of free ammonia. In the consequent aerobic phase (60–190 min.), phosphate is taken up again by PAO biomass. Simultaneously, ammonia is oxidized to nitrate via nitrite. At 140 minutes, the ammonia concentration is virtually zero, soon followed by nitrite depletion. The depletion of ammonia corresponds to the occurrence of the ammonia valley (see e.g. Figure 4.3(e)) whereas nitrite depletion give rise to the ORP nitrate break-point (see e.g. Figure 4.3(f)). In the anoxic phase (190–270 min.), the phosphorus concentration increases slightly due to the addition of influent at the beginning of the phase. Reduction of nitrate and release of ammonia due to hydrolysis occur simultaneously. No phosphorus release is observed as nitrate reduction is not complete during this phase (i.e. the necessary anaerobic conditions for phosphorus release are not met). During the second aerobic phase (270–300 min.), ammonia oxidization to nitrate occurs together with slow uptake of phosphorus. During settling a small part of the nitrate nitrogen is reduced to nitrogen gas.

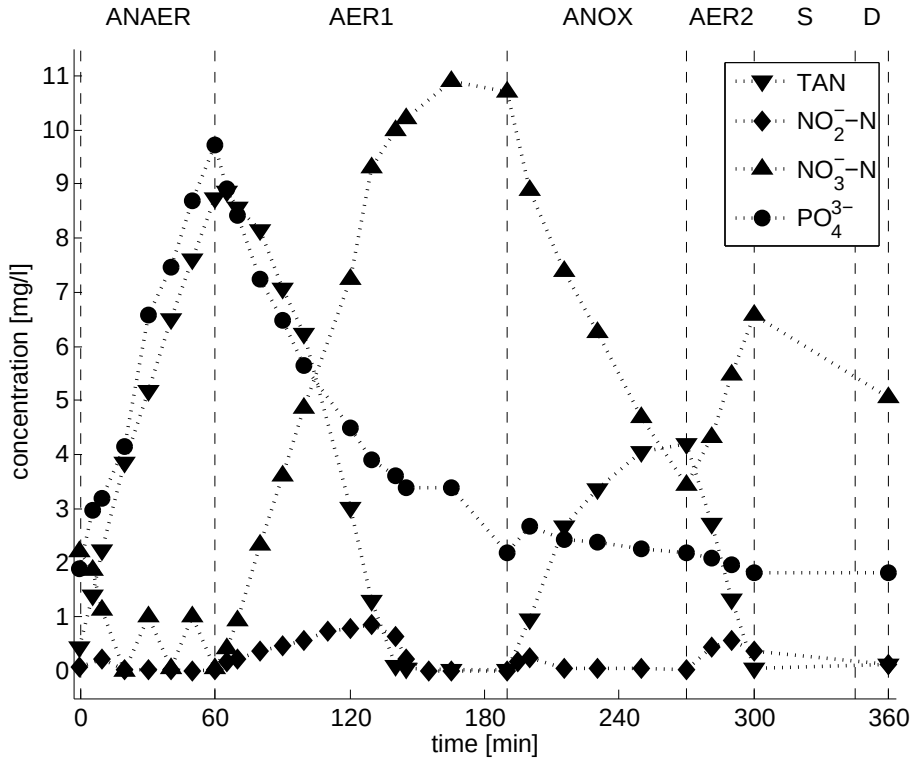


Figure 4.4: Profiles of effluent quality variables during a one-cycle measurement campaign.

4.5 Description of faults

Prior to evaluation of monitoring and diagnosis strategies for the batch process, the data were screened for faults by meticulous inspection. Here, the separate classes of faults and their presence in the used data sets are described. Faults in the hydraulic data of the system are described separately first. Thereafter, frequently occurring problems with the conductivity sensor are discussed. In a third part, other faults identified in the data set are described.

4.5.1 Faults in hydraulic parts of the system

Figure 4.5 shows the weight profiles for the 7 classes of faults identified for the hydraulic parts of the system. Fault class 1 represents batches in which no influent is added, either due to complete pump failure or due to (unintended) disconnection of influent tubes. Fault class 2 represents batches in which the added amount of influent is lower than desired, typically due to a pump or valve failure. Fault class 3 represents batches in which the added amount of influent is higher than desired. This type of failure is typical when faults of class 2 or 3 are resolved. This results from the automatic update of the pump speed after each batch. To automatically account for wear of the flexible tube in the peristaltic pump, the pump speed is updated after each batch. In the occasion of a failure of fault class 1 and 2, the pump

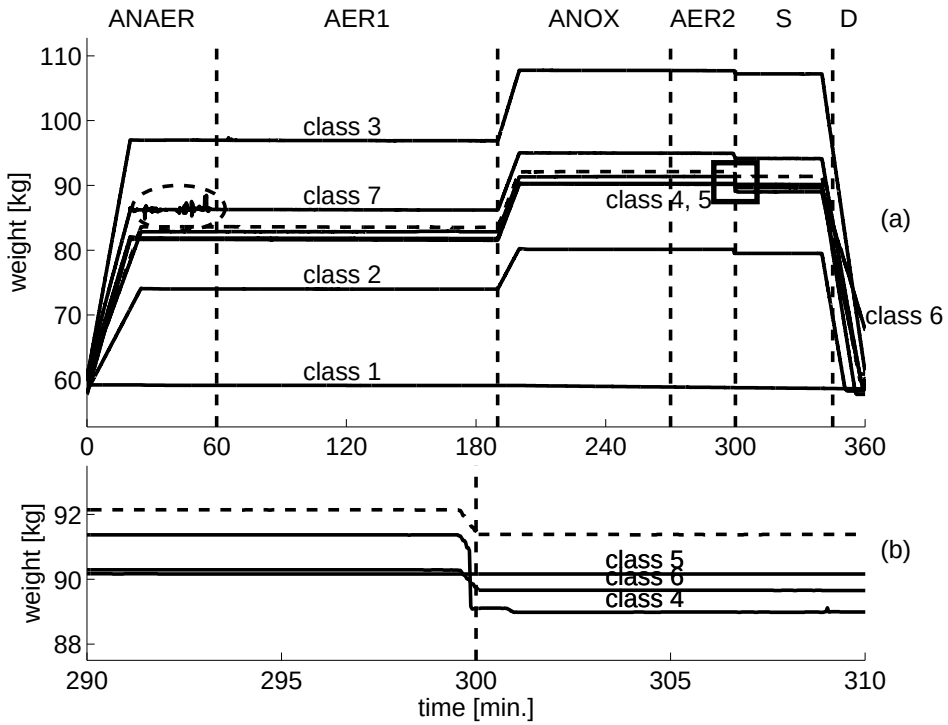


Figure 4.5: Typical weight profiles under normal (dashed) and abnormal conditions (solid). (a): complete profiles, (b) detailed view of the box indicated in (a) showing sludge wastage.

speed setpoint is consequently increased gradually with each batch. Solving the problem (i.e. remove blockage and/or reconnect tubes) without resetting of the influent pump speed then results in high influent loads. Fault class 4 is characterized by a sludge wastage which is higher than desired. This is clear from Figure 4.5(b), showing the box indicated in 4.5(a) in detail. Antagonistic to fault class 4, fault class 5 is characterized by a less than desired sludge wastage. Fault class 6 characterizes those batches in which the effluent withdrawal is less than desired. This can be seen at the end of the cycle for the corresponding profile in Figure 4.5(a). Fault class 7 groups batches in which noisy artefacts are present. The ellipsoid shown indicates the occurence of such noise for the example given. It is noted here that the operation of the system was not believed to be influenced by the latter noisy artefact, i.e. the faults only relate to data quality and not process quality. Fault class 8 (not shown) groups batches which exhibit faults in both fault class 6 and 7.

The described fault classes for the hydraulic part of the system are summarized in Table 4.4. In Table 4.5, the numbers of batches exhibiting faults are reported per fault class and mode as well as for all modes together. Given that faulty batches are a result of a single event, the number of events that triggered the faulty batches are reported as well. The number of batches is generally higher than the number of events, indicating that the occurence of a fault (one event) often affects multiple batches in the same fashion. This is due to the occurrences of faults during nighttime and weekend periods when rigorous human follow-up of the reactor was practically infeasible. Some faults like the ones of fault class 2 and fault class 6 often occur as a gradual change in pump performance. As a result, detection of those faults has often taken additional time. This explains the low ratio between events and fault batches for these classes (in all modes for fault class 3 and in mode 2 for fault class 6). Noisy artefacts in the data (fault class 7) are considered to be the result of independent events in all cases. Of all recorded batches, 73% were classified as normal. Of the abnormal batches (27%), 37% represent noisy artefacts (10% of all batches) and 63% represent serious faults (17% of all batches).

Table 4.4: Discriminated fault classes for the hydraulics parts of the studied system.

Fault class	Short description
1	Complete hydraulic failure: No influent is added nor effluent withdrawn due to failure of the hydraulic system.
2	Insufficient influent addition: A smaller than acceptable volume of influent is added to the system.
3	Excessive influent addition: A larger than acceptable volume of influent is added to the system.
4	Excessive sludge wastage: A larger than acceptable volume of mixed liquor is wasted at the end of the second aerobic phase.
5	Insufficient sludge wastage: A smaller than acceptable volume of mixed liquor is wasted at the end of the second aerobic phase.
6	Insufficient effluent withdrawal: The reactor level does not reach the desired level at the end of cycle.
7	Noisy artefacts in the data: The operation of the system is not disturbed.
8	Faults corresponding to fault class 6 and 7 occur in the same cycle.

Table 4.5: Fault classes and number of included batches.

Class	Fault class	Mode 1		Mode 2		Mode 3		all modes	
		batches	events	batches	events	batches	events	batches	events
normal		572		399		65		1036	
faulty	1	6	1	19	6	0	0	25	7
	2	64	10	43	12	5	1	112	23
	3	6	4	3	1	0	0	9	5
	4	3	1	0	0	0	0	3	1
	5	0	0	2	2	0	0	2	2
	6	4	4	75	9	1	1	80	14
	7	101	101	17	17	18	18	136	136
	8	2	2	2	2	0	0	4	4
all		758	123	560	49	89	20	1407	192

4.5.2 Faults of the conductivity sensor

During the studied period, the measurements by the conductivity sensor has been corrupted during aerobic phases of a larger part ($> 60\%$) of the recorded batches. Two problems were frequently observed. The first problem, clogging, results when sludge sticks between the two *legs* of the probe, containing the electroplates. As a result, the conductivity sensor measurement is biased (downwards). A second problem is due to the passage of gas bubbles between the *legs* of the probe. As gas bubbles pass, the conductivity measurement drops due to the low conductivity of the gas. In an attempt to solve the latter problem, a plastic shield was installed for a period of time so as to overcome this problem. While successful, the shield induced low currents in the neighbourhood of the sensor leading to increased clogging of the sensor. Tilting or repositioning of the sensor could not resolve any of the problems either. In Figure 4.6, the two reported problems are illustrated. A clogging problem occurs at the beginning of the aerobic phase (minute 60-75). Drops in conductivity due to air bubbles are frequently observed in the aerobic phases. In view of the exceptionally large proportion of corrupted data, the data of this sensor are not included in any work presented hereafter.

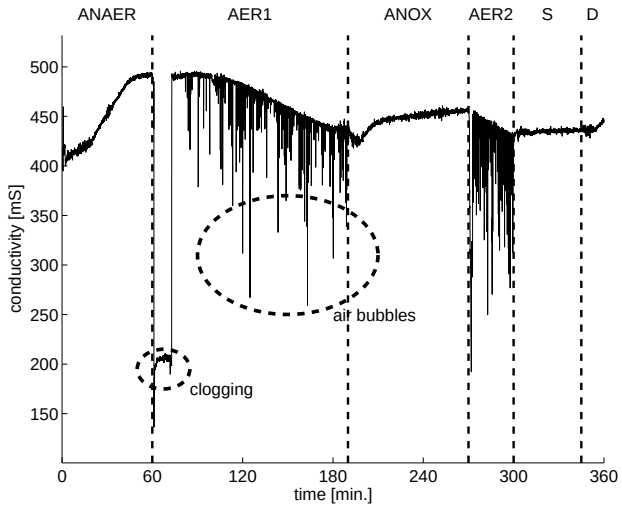


Figure 4.6: Conductivity profile of a single batch in which both clogging as well as air bubbles corrupt the measurement signal.

4.5.3 Other faults

Figure 4.7 shows the temperature profile for a normal batch as well as a profile obtained when the cooling system failed. Clearly, the temperature deviates largely from normal operation. Failure of the cooling system is identified as fault class 9.

In Figure 4.8 dissolved oxygen (DO) level profiles are shown for 5 distinct fault classes (10 to 15), characterized by problems within the aeration system. Fault class 11 is characterized by ineffective aeration of the system during the aerobic phases. Faults like these have been related to clogging of the gas diffusor in the reactor and to failures of the air compressor. Fault class 12 groups a set of batches for which the oxygen setpoint was temporarily set to a higher level. Even if this was done intentionally, they are grouped here as a fault. This is motivated by the infrequent use of this setpoint, i.e. it is not a normal situation. Fault class 13 groups batches in which the oxygen level remains high (i.e. non-zero) during large periods of time in the anaerobic or anoxic phase. This is clearly the case in the anaerobic phase for the shown example. Fault class 14 is characterized by a slow decay of oxygen during the start of the anoxic phase, as indicated in Figure 4.8. Batches in fault class 15 are characterized by (slowly) oscillating oxygen levels during the

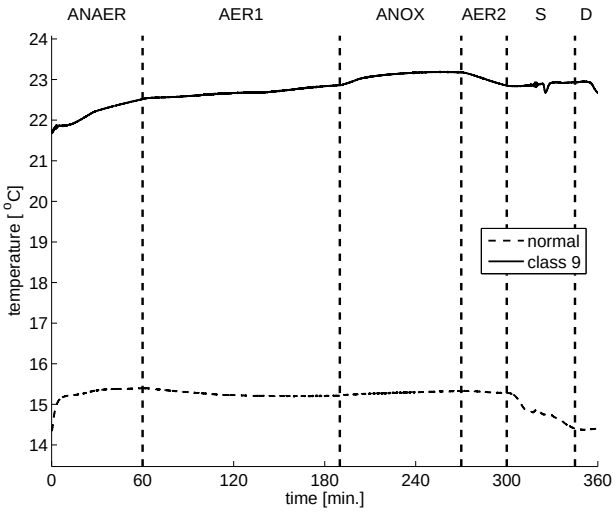


Figure 4.7: Temperature profile in a normal cycle and in a cycle characterized by failure of the cooling system (fault class 9).

aerobic phases, due to inappropriate tuning of the control for the systems behaviour at the time. Note that a linear PID controller was used to control the oxygen level, which is not supposed to be ideal in a wide operating range of a non-linear system as the studied SBR. It is also noted that even though problems with the aeration system typically lead to abnormal profiles in the DO, pH, ORP and gas flow rate signals simultaneously, they have been labeled in terms of the oxygen levels or the gas flow rate as the aeration control loop is based on the DO level and gas flow rate as the only input, resp. output of the controller.

Figure 4.9 shows profiles of DO and the gas flow rate for fault classes 16 and 17. Fault class 16 is characterized by a controller that is tuned *too aggressively* for the present situation, leading to fast oscillation in both oxygen and gas flow rate. Fault class 17 groups anomalies in the aeration-related data which could not easily be classified to the one or other class. In the example shown, the gas flow rate drops twice to its minimal value whereafter increases are seen. An irregular profile is equally seen in the DO profile data.

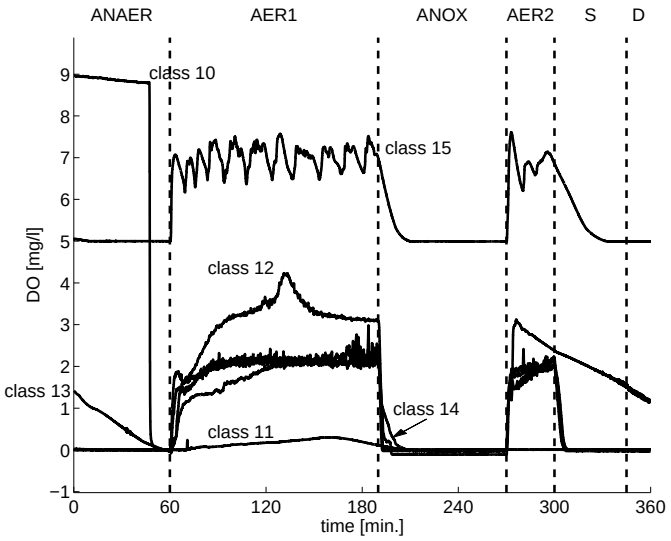


Figure 4.8: Profile of dissolved oxygen (DO) for 6 specific faults (fault class 10 to 15). 10: DO sensor maintenance, 11: ineffective aeration, 12: setpoint too high, 13: limited decay of oxygen, 14: slow decay of oxygen after aerobic phase, 15: air control tuned too weak. Data of class 15 are scaled (+5 mg/l) for reasons of visibility.

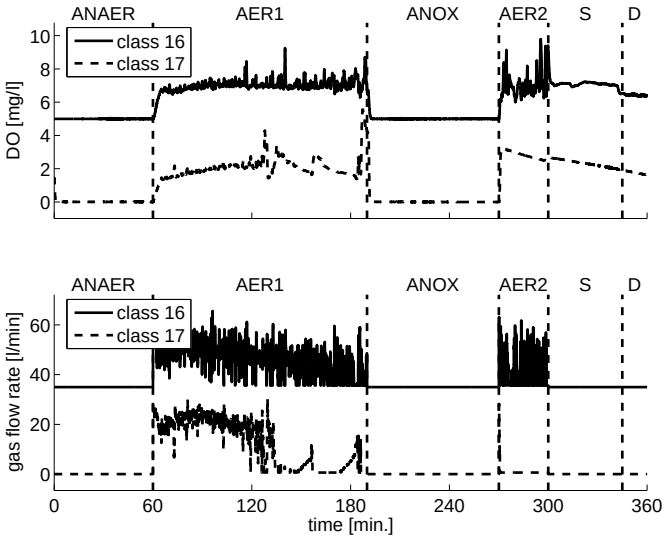


Figure 4.9: Profile of dissolved oxygen (DO) for 2 specific faults (fault class 16 and 17). 16: controller gain tuned too high, 17: anomalies. Data of class 16 are scaled for visibility (DO: +5 mg/l, gas flow rate: +35 l/min.).

Two problems related to the pH sensor were identified during data screening and are illustrated in Figure 4.10. The first problem (fault class 19) includes a calibration error of the pH probe, leading to a positive bias of the slope of the (linear) calibration between the measured voltage signal and the resulting pH signal. A second problem was identified as observational outliers. Such outliers can visually be assessed as momentary drops in the pH signal and are believed to be due to voltage sacks in the electrical parts of the data acquisition system. The same problem was also identified for the ORP signal (Figure 4.11). Note that the latter two problems never occurrred at the same time.

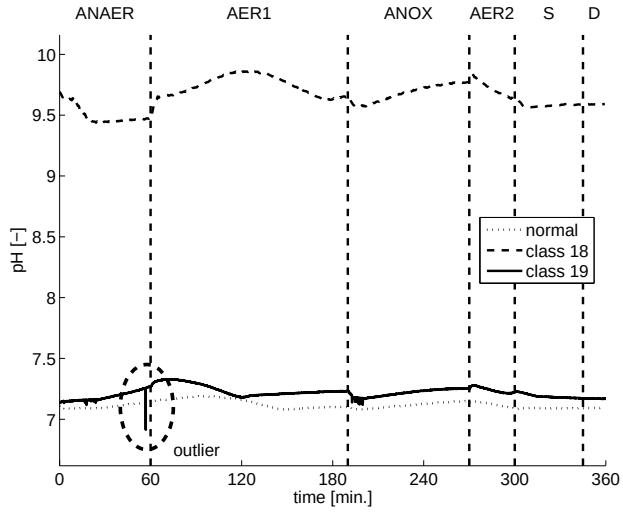


Figure 4.10: Profiles of pH for a normal batch and for fault class 18 (calibration error) and 19 (observational outlier).

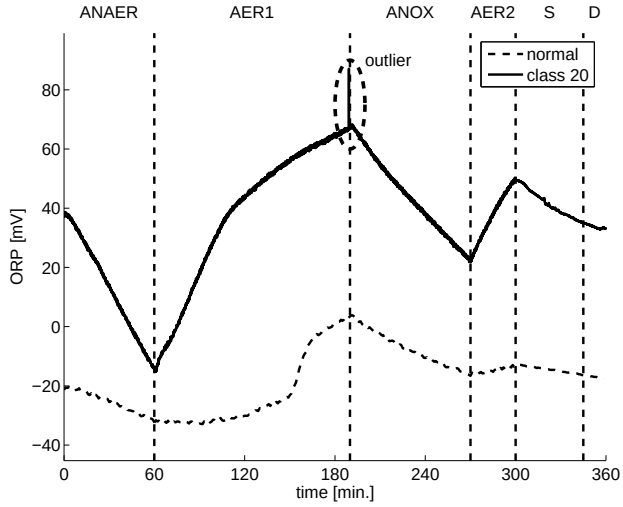


Figure 4.11: Profiles of ORP for a normal batch and for fault class 20 (observational outlier).

Identified fault types for the complete system other than those already described for the hydraulics and conductivity sensor are described in Table 4.6. Described faults classes as well as fault classes for specific combinations of presented faults are presented.

In Table 4.7, the number of faulty batches and fault events are reported without considering that faults may have occurred simultaneously. This allows to compare both types of numbers. Also in this case, it can be seen that faulty batches are often the result of a single event. This is for example true for the (single) pH calibration error (fault class 18). In Table 4.8, the effective number of batches for each fault class, hereby accounting for simultaneous faults, are reported.

Table 4.6: Discriminated fault classes (in addition to those in Table 4.4) for the complete system.

System part	Fault class	Short description					
cooling system	9	Cooler failure: the bulk liquid temperature is too high					
aeration system	10	DO sensor maintenance					
	11	DO level too low: setpoint not reached (sufficiently fast)					
	12	DO level too high in aerobic phase.					
	13	DO level high in anoxic or anaerobic phase: minimal respiration					
	14	DO level decay slow after aerobic phases: low respiration					
	15	slow oscillatory behaviour of DO measurement					
	16	fast oscillatory behaviour of DO and gas flow rate					
	17	anomalies in trajectories of DO and gas flow rate					
pH sensor	18	erroneous calibration of pH probe					
	19	outliers are observed in the pH signal					
ORP sensor	20	outliers are observed in the ORP signal					
combined faults	21	2 & 9	30	19 & 1	39	20 & 9	
	22	2 & 12	31	19 & 2	40	20 & 12	
	23	2 & 14	32	19 & 6	41	20 & 13	
	24	2 & 18	33	19 & 8	42	20 & 14	
	25	3 & 12	34	19 & 10	43	20 & 15	
	26	6 & 13	35	19 & 12	44	20 & 16	
	27	6 & 14	36	19 & 16	45	20 & 22	
	28	6 & 17	37	20 & 2	46	20 & 24	
	29	7 & 18	38	20 & 6			

Table 4.7: Numbers of batches with reported faults and the number of events giving rise to them. Note that in this table, batches with multiple faults are counted more than once.

Fault class	Batches	Events	Fault class	Batches	Events
1	16	6	10	2	2
2	42	14	11	2	1
3	3	1	12	25	8
4	0	0	13	28	9
5	2	2	14	15	12
6	44	10	15	17	10
7	34	6	16	24	6
8	2	2	17	3	3
9	60	16	18	54	1

Table 4.8: Numbers of batches with reported faults

Class	Fault class	Batches	Fault class	Batches
fault-free		226		
faulty	1	15	24	4
	2	22	25	3
	3	0	26	1
	4	0	27	1
	5	2	28	1
	6	25	29	31
	7	3	30	1
	8	1	31	2
	9	54	32	14
	10	1	33	1
	11	2	34	1
	12	11	35	1
	13	26	36	10
	14	11	37	1
	15	14	38	2
	16	11	39	5
	17	2	40	1
	18	18	41	1
	19	6	42	1
	20	5	43	3
	21	1	44	3
	22	8	45	1
	23	2	46	1
all				556

4.6 Concluding remarks

In this chapter, the SBR system used as a case study throughout this work was described. In view of the evaluation of fault detection and diagnosis strategies, batches were classified according to operational modes and identified faults. Each of the operational modes and classes was described in detail. An additional note on the identified faults is made here in view of future research.

The root cause of a larger part of the faults could be found in the hydraulic system. Clearly, the hydraulic system has proven to be vulnerable to many faults. It is the author's opinion that this vulnerability is largely due to the use of flexible tubes for the influent and effluent lines with small diameters (3 mm). Indeed, in the author's experience, such tubes often clog, resulting in reduced influent flow and eventually disconnect if pressure builds up so that zero influent flow results. In addition, the flexible tubes require the use of pinch valves to control the influent and effluent flows. Keeping the flexible tubes in position (within the pinch valves) has proven to be difficult as well, often leading to reduced flow or presumably reduced concentrations in the influent. The latter results when the tap water tube within the valve controlling the selection between tap water and concentrated influent is dispositioned. In such a case, the tap water tube is constantly open resulting in reduced suction head through the concentrated influent tube when the latter is (correctly) opened.

In order to avoid disconnection of tubes and dispositioning in pinch valves, solid tubes and solid valves (i.e. without pinched tubes) are a valid choice and peristaltic pumps should be avoided. Short tube lengths and larger diameters should be used to reduce the risk of blocking. In addition hardware should be selected in a fashion that considerable margins between required and maximal capacity are obtained. In this case, small tube diameters were found to lead to considerable head loss. It is noted here that some of the former requirements are conflicting with other requirements for research. First of all, a setup designed for long-term research may necessarily imply that the design is flexible and open for adjustments. This stands in clear contrast with paradigms in process control, where unnecessary system changes will be avoided as much as possible. Also, more robust hardware such as the suggested increase of tube diameters causes larger volumes of wastewater to stand still in the tubes, possibly causing the tubes themselves to function as (small) biological reactors in time. Especially in view of mechanistic understanding and modelling of biological systems, these volumes should be reduced when possible.

The cooling system, while found more robust, lead to a large amount of faulty batches as well. This was largely due to the fact that technicians external to the university were responsible for reparation, combined with the unavailability of a backup cooling system. Such a backup system may avoid large numbers of faulty batches. Also, the cooling system was often found to be working at maximal capacity, especially during summer. Better location, i.e. in a naturally cooler environment or use of cooling systems with higher capacity may avoid such problems.

While a considerable set of problems in the aeration system were identified, only few of them are related to technical failures. In the history of the SBR system, failures of the physical aspects of the aeration system could be narrowed down to complete compressor failures, failure of oil and water filters leading to liquid components entering the compression valve for aeration control, thereby disrupting its internal control system (i.e. the PID control law for the valve position based on an internal flow measurement and given flow setpoint becomes corrupted). If oxygen setpoint control is not a necessity for the intended research topic(s), it may be a valid option to use a fixed air flow rate during the aerobic phases hereby simplifying design and maintenance of the experimental unit.

The conductivity sensor was reported to be commonly subjected to two faults, clogging and air bubbles entering the measuring device. Pragmatic solutions, such as tilting the sensor or the provision of a shield could not effectively reduce the frequency of the reported fault effectively. The problems themselves are largely due to the design of the conductivity sensor which has two *legs* including the two electroplates. With respect to this, it is noted that new types of conductivity probes are available on the market which have a single tube-like design. Selecting such a probe may effectively resolve the reported problems.

Some additional comments can be made on the types of data that are measured. For the given system, measurements were taken during the batch runs that express supposedly meaningful information with respect to its operation, both physically and biologically. Also, measurements are taken at 2-second intervals which may suggest that information is available at large. For example, the weight variable is an indicator for the load to the system and temperature is known to have large impact on microbial growth rates and reaction speeds. It may therefore seem a paradox that this wealth of measurements may not be suitable to isolate or diagnose faults to a desired level. Consider for example a failure of the pumping system (due to disconnected tube) and a failure of the balance (e.g. being stuck). Based on the weight variable only, separation of the two faults is theoretically and practically

impossible. Additional measurements of the flow rate at the inlet and outlet would however enable the construction of a mass balance, hereby enabling separation of the two problems. Similarly, additional measurements in the cooling and aeration system may allow to construct heat and flow balances as well. Put otherwise, measurements taken so that physical balances can be assessed (e.g. by knowledge or by estimation) may allow to effectively detect, isolate and diagnose problems in time without need for advanced inferencing. Importantly for the development of process monitoring strategies for biological processes, the set-up balances may facilitate effective discrimination between deviating behaviour in the physical and biological parts of the system *a priori*, so that assessment of the performance of monitoring and diagnosis strategies especially designed for the detection of biological faults can be achieved.

Part III

Multivariate monitoring, diagnosis and control of a Sequencing Batch Reactor

Chapter 5

Monitoring of a Sequencing Batch Reactor by means of Multi-way Principal Component Analysis

5.1 Introduction

In this chapter, monitoring on the basis of PCA (Principal Component Analysis) is evaluated for the detection of process faults for SBR's for biological nutrient removal. PCA-based process monitoring is a long established approach to monitor processes for which limited information or knowledge is available but for which large data sets with many correlated variables have been collected. Given the large number of variables that are recorded in many cases, PCA allows to capture the larger part of the variability in fewer variables, then called principal scores. Separate statistics can be constructed for the behaviour of the extracted scores as well as for the non-captured part of the variance in the data. The basis of the method, Principal Component Analysis, is explained in detail in Section 3.3. Batch-wise unfolded Multi-way Principal Component Analysis (MPCA), being a conventional

extension of PCA for batch processes, is used throughout this chapter. The reader is referred is to Section 3.3.3.1 for detailed information on the method. Batch-wise unfolded MPCA is only applied for complete batches so to avoid problems with missing data. It is noted here that, given that a fault event may last for several batches, the detection of a faulty batch is still interesting in view of minimizing the number of faulty batches.

In the first section following this introduction, definitions used throughout this chapter are given. In a second section, PCA models are evaluated for monitoring of the hydraulic part of the system only. The motivation for this is that causal relationships between the hydraulic behaviour of the system and other behaviour exist, while the reverse is not believed to be true (i.e. the biological processes in the reactor do not affect the pump system). Fault detection on the basis of the hydraulics only allows to differentiate between faults in the hydraulic system and faults in the other parts of the system. As a result, detection of process faults by means of the first set of models lead directly to an isolation of the hydraulic parts of the system as the source of the identified problem. Practically speaking, the measurements of the weight variable only are included in the data set. As a result, a two-way matrix results which can be directly modelled by PCA. Note that the models are constructed on the data of complete batches so that standard use of the resulting PCA model requires that a monitored batch is complete. The inherent problem of batch-wise unfolding thus remains (see Section 3.3.3.1). Separate models are constructed for the different modes described in Chapter 4. By means of Mixture (Multi-way) PCA (see Section 3.3.2.3), it will be evaluated whether prior knowledge is essential to guarantee the monitoring performance of the models.

In the third section, all measured variables and all identified process faults are included in the study. It is evaluated whether this allows to monitor the system as a whole. The latter approach is tested for a single mode only (i.e. mode 2a). In a fourth section, the shown results are discussed.

5.2 Definitions

5.2.1 Performance measures

To evaluate the performance of any monitoring strategy, the following two measures are common:

- *Type I error rate or false alarm rate.* A type I error occurs when one rejects the null hypothesis when in fact this hypothesis is valid. In the context of monitoring the null hypothesis is that the investigated observation is a normal fault-free observation. The type I error rate is therefore the fraction of normal fault-free observations that are classified as abnormal. A typical monitoring system will activate an alarm signal when a batch is classified as abnormal. It is for this reason that the type I error rate is also referred to as the false alarm rate. The type I error rate is typically reported as a percentage.
- *Type II error rate or false acceptance rate.* A type II error occurs when one accepts the null hypothesis when in fact this hypothesis is not valid. In the context of monitoring, this means that a faulty or abnormal observation is not detected. The type II error rate is therefore the fraction of abnormal observations that are classified as normal. As this means that the operation of the monitored system is continued without further concern, this measure is also referred to as the false acceptance rate. The fraction of observations that are truly rejected, thus being correctly identified as abnormal batches is commonly referred to as the statistical power.

Both error types are generally not known exactly and therefore need to be estimated. An estimate of the type I (II) error rate can be established by calculation of the fraction of observations that are rejected (accepted) from a set of batches known to be normal (abnormal). In this study, an observation is rejected if the 95%-levels of either the Q- or Hotelling's T^2 statistics are crossed.

If observations are used to construct a (statistical) model for fault detection, then those observations cannot be part of the data set used to estimate any of the error rates. Since a set of normal data is required for model identification in the applied stratgies, an independent set of normal observations needs to be identified to estimate the type I error rate. Therefore, the set of normal observations is split into a

set for modelling, called the calibration set, and a set for estimation of the type I error rate, called the validation set.

In practice, the calibration and/or validation set may be unrepresentative for the complete set, often due to a limited set of available data. As a result, the estimated type I error may deviate from its true value. Cross-validation offers a way to tackle this problem. In this approach, the complete set of normal data is split into a number, N, of blocks by random assignment. It is typical to assign the same number of observations to each block. Each block is then used once as a validation set while the remaining blocks are used for modelling. As a result, N estimates for the type I error rate follow. The ultimate estimate for the type I error rate is then calculated by weighted averaging of these N estimates where the weights are defined as the number of observations in each valdiation block. As such, the risk for a highly deviating estimate is reduced. In other words, the variance of the type I error estimate is reduced. The reduction in variance is higher when less blocks are used, due to the fact that each of the validation blocks will become more representative for the whole set of observations. A lower number of blocks however inevitably leads to an increased (positive) bias of the estimate (because the calibration sets get smaller and thus becomes less representative for the whole set of observations). Typical choices for the balance between bias and variance are 5 and 10. In this section, 10 blocks will be used. For more details on cross-validation procedures we refer to Hastie et al. (2001).

As statistical process control strategies based on PCA do not require the use of abnormal data during the modelling stage, the type II error rate will simply be the fraction of abnormal batches that are not identified as such by the applied fault detection strategy. As several fault classes were identified in the previous chapter, the type II error can be calculated separately for each of the identified fault classes. From here, the latter type II error rates will be referred to as the fault-specific type II error rates. It is noted here that a mean type II error rate is calculated by taking the weighted mean of all type II error rates over all fault types, denoted type IIa. Each observed fault is weighed equally. In addition, a second generalized type II error rate is calculated by considering all fault classes except fault class 7. This type II error rate is denoted as type IIb. As discussed in the previous chapter, fault class 7 comprises batches for which the process was not believed to be disturbed. Only the data quality in this class is corrupted. It is for this reason that this additonal measure of performance is constructed.

As mentioned above, the type I and type II error rates are estimates of the expected probabilities for misclassification for normal data (type I) and abnormal data (type IIa, type IIb). As a result, these estimates are characterized by inherent uncertainty. To enable an effective interpretation of the estimates, 95% confidence intervals are computed for the error rates according to a binomial distribution.

Given a random process by which a two-level univariate outcome (e.g. 0/1) is assigned to each discrete run or trial and assuming a binomial distribution, the probability, Y, that at least N_1 out of N outcomes are '1' is given by the following cumulative distribution function (cdf):

$$Y = F\left(N_1|N,\beta\right) = \sum_{n_1=0}^{N_1} \binom{N}{n_1} \cdot \beta^{n_1} \cdot (1-\beta)^{N-n_1} \tag{5.1}$$

where:

$$\beta\text{: (constant) probability for outcome '1'} \tag{5.2}$$

A binomial distribution model is fitted to the acceptance/rejection results for each identified class. For each class (normal and fault classes), N_1 is the number of misclassified observations and N the number of observations in the studied class. Then, $\frac{N_1}{N}$ is the estimate of β. To compute confidence levels for the computed error rate $\frac{N_1}{N}$, 95% confidence levels are computed according to equation 5.1 by solving the equation to N_1 for $Y = 0.025$ (left-hand side confidence limit) and $Y = 0.975$ (right-hand side confidence limit) after replacement of β with its estimate $\frac{N_1}{N}$. Importantly, the binomial distribution model requires that the trials are independent and that the probability for a certain outcome is constant over all trials. This is not expected to be true for the studied data. Indeed, many of the faulty observations are the result of a single fault lasting over several batches resulting in the violation of the assumption on independency. Also, for some faults the probability of the outcome may be dependent on the magnitude of the fault. For example, for fault class 2 (partial failure of the pump) the magnitude may be thought of as the relative deviation of the added volume from the desired or mean volume. Larger magnitudes of such a fault may be expected to result in larger deviations of the resulting data from the normal operation data thus leading to decreased probability of misclassification. Also, when confidence intervals are computed for the overall type IIa and type IIb error rates it is -likely erroneously- assumed that the probability of misclassification is not depending on the type of error. As such, the reported confidence levels for the estimated error rates will only be used for indicative purposes.

As it is generally impossible to minimize the type I and type II error rates at the same time, model selection requires a compromise between the two errors. There is no general rule as to how the two errors should be weighed against each other. In practice however, this may be defined by weighing the expected cost of not detecting a fault, i.e. the cost of a continued process failure or unmet process targets, to the expected cost of a false alarm, i.e. the cost of investigating the potential problem in vein. As the type I and type II error rates are rates conditional to the occurence of a normal or, repectively, abnormal batch, the rate at which normal and abnormal batches are expected to occur needs to be accounted for when the expected cost are balanced against each other. Given this problem, the optimal model was identified for the following definitions of optimality:

1. minimal type I error rate (avoiding false alarms at all cost).

2. minimal type IIa error rate (avoiding false acceptance of any error at all cost).

3. minimal type IIb error rate (avoiding false acceptance of any error except for fault class 7).

4. minimal mean of type I and type IIa error rate, thereby equally weighing type IIa errors against type I errors. Denote the latter mean as the type I&IIa error.

5. minimal mean of type I and type IIb error rate, thereby equally weighing type IIb errors against type I errors. Denote the latter mean as the type I&IIb error.

The latter two definitions assume that the costs of type I and type II error rates are weighted equally against each other, i.e. the probability for a faulty observation times the cost of not detecting it is believed to equal to the probability of a normal batch times the cost of investigating the batch in vein. The optimal model was established for each of these definitions of optimality. Each definition of optimality will be referred to by means of the listed number. Where multiple choices for the model lead to the same (optimal) performance it is logical to choose the most parsimonious model delivering this performance (least number of PC's).

5.2.2 Evaluated model types

All evaluated monitoring models are of the batch-wise unfolded Multi-way PCA type (see Section 3.3.3.1). As discussed in Section 4.2, the used data set exhibits observations that stem from different operational modes. This results in a violation of one of the assumptions in standard PCA-based process monitoring, namely that the process data are drawn from a single multivariate (normal) distribution. To evaluate whether and how this problem can be tackled, the following strategies are compared to each other:

- Single mode modelling: a single model is linked to the (known) condition/operation of the plant. This strategy requires that the condition or operation of the system is known during on-line process monitoring.
- Mixture PCA modelling (see Section 3.3.2.3): a set of models for (known) operations of the plant are used simultaneously. This strategy does *not* require explicit knowledge of the system's condition or operation during process monitoring. As the Multi-way PCA (MPCA) approach to batch process monitoring is combined with Mixture PCA (MixPCA), the combined approach can be defined as Mixture Multi-way PCA (MixMPCA). It is noted that the automatic assignment of observations to a mode during the calibration step is not pursued. Instead, knowledge of the mode is used during the calibration steps.

Results for both approaches will be shown for monitoring of the hydraulic parts of the system. To monitor the complete system (in one mode), the single mode modelling approach will be used.

5.3 Monitoring hydraulics

In this section the results are shown for monitoring of the studied system on the basis of the weight variable only. First, results are shown for single mode MPCA modelling. Then, MixMPCA modelling is evaluated.

5.3.1 Single mode Multi-way PCA (MPCA)

In the single mode strategy, a single Multi-way PCA (MPCA) model is made for each operational mode that was identified in the historical data set (Section 4.2). To classify a new batch as faulty or normal one uses the model linked to the (known) operational mode of the new batch. This strategy is evaluated as follows. For each operational mode (see Table 4.2), separate type I and type II error rates are calculated. To do so, only data from the given operational mode are used. The type I error rate is calculated by 10-block cross-validation. Type II error rates are calculated by projection of the faulty batches onto the respective model. This is done on the basis of classification of the batches according to the prior data screening, as reported in Tables 4.4 and 4.5. In what follows, the results for each of the modes (1-3 in Table 4.2) are given. Both group scaling as well as autoscaling were tested as data preprocessing steps (see Section 3.3.3.1). The corresponding models are referred to as GS-MPCA (for group scaling) and AS-MPCA (for autoscaling) models. It is repeated that group scaling refers to mean centering of the data and division of the (centered) data with a single overall standard deviation for all measurements derived from the same sensor. Autoscaling refers to mean centering of the data and division of the (centered) data with a respective standard deviation for each time instant in the batch. Since the data sets considered here consist of a single measurement throughout the complete batch trajectories, 10800 means (6 hours at 2-second interval) and 1 (GS) or 10800 (AS) standard deviations are calculated for data processing prior to MPCA modelling.

5.3.1.1 Mode 1

GS-MPCA MPCA models with up to 100 PC's were tested for the first mode. For all evaluated models (both GS-MPCA and AS-MPCA) the type II error rates for fault classes 1, 3, 6 and 8 were zero and faults of class 5 were not present in this mode. The right hand side limit of the 95% confidence intervals for the type II error rates for fault classes 1, 3, 6 and 8 are reported in Table 5.1. The left-hand side limits are all zero. Due to the low number of observations in the respective classes, the reported confidence intervals are rather wide thereby turning the type II error uncertain as an estimate for the probability for misclassification.

The type II error rates and confidence intervals for all other fault classes are shown in Figure 5.1 as a function of the number of PC's for the GS-MPCA model series. For fault class 2 the type II error rate is low for low dimensions (1 to 3 PC's). For the number of PC's ranging from 4 up to 15, the type II error is higher (>20%). The type II error rate for this fault class slowly decreases again for higher MPCA-model dimensions. None of the (6) faults in class 4 is detected for the number of PC's ranging from 1 to 3 and from 15 to 82. For 4 to 11 PC's all faults in this class are detected. For 12 to 14 PC's, the type II error rate is 33.33% (4 out of 6 are detected). Increasing the MPCA model dimensions beyond 82 results in a decrease again of the type II error rate. The confidence interval for this fault is wide over the whole evaluated range for the number of PC's, delivering an uncertain measure of expected misclassification rate for this class of faults. The type II error rate for fault class 7 exhibit a rather smoothly decreasing trend from low to high dimensions of the MPCA models. Given the large population of observations in fault class 7, the confidence intervals for the type II error rate are much smaller compared to the other fault classes.

Table 5.1: GS- & AS-MPCA, Mode 1. Right-hand side 95% confidence limit for the type II error rates for fault classes 1, 3, 6 and 8.

Fault class	Right hand side confidence limit
1	45.9
3	45.9
6	60.2
8	84.2

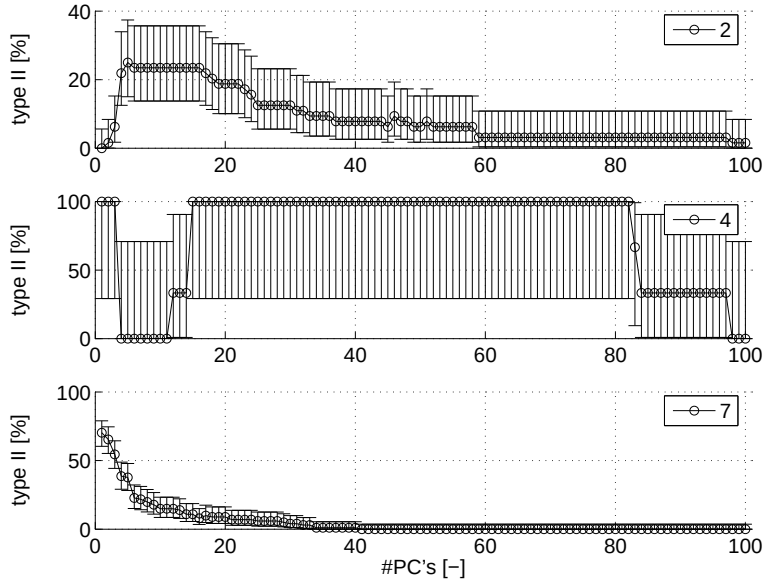


Figure 5.1: GS-MPCA, Mode 1. Type II error rates and corresponding confidence intervals for fault classes 2, 4 and 7 as a function of the number of PC's

Figure 5.2 shows the type I, type IIa and type IIb error rates as a function of the number of PC's as well as the corresponding confidence intervals. The type IIa rate confounds all the type II error rates shown so far by equal weighing of each faulty observation. The type IIb error rate does the same with exception of fault class 7. The type I error rate shows an increasing trend from about 10% at small values for the number of PC's (1, 2) up to 70% at 100 PC's and exhibits relatively narrow confidence intervals (less than 8% wide). The type IIa decreases from 40% to zero at maximal numbers of PC's and has a steep downward slope for the first numbers of PC's (up to 6 PC's). The type IIb error rate shows a similar trend but is much lower for MPCA models with low dimensions. Excluding fault class 7, thus leads to a fairly different assessment of these low-dimensional models. also, due to the large number of faults in class 7, the confidence limits computed here are much wider compared to those for the type IIa error rate.

In Table 5.2 summarized results for the models selected according to each of the optimality definitions can be found. Minimizing the type I error rate (optimality 1) leads to a low dimension of the model, namely 2, as may be expected from Figure 5.2. A type I error rate of 9.1% is achieved and paid off by a high type IIa error rate

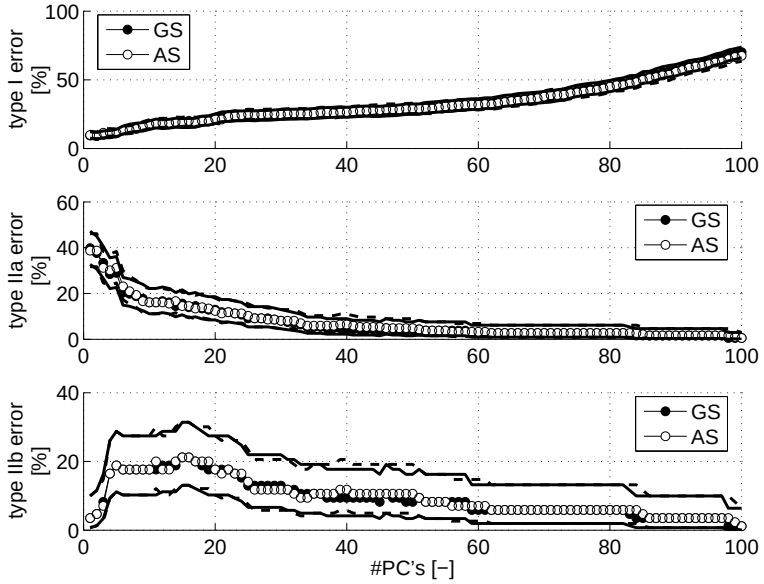


Figure 5.2: GS- & AS-MPCA, Mode 1. Type I and overall type II error rates as a function of the number of PC's. Full lines indicate confidence intervals for GS models. Dashed lines indicate confidence intervals for AS models.

(37.6%) but the type IIb error rate is not compromised as much (type IIb: 14.5%). Minimizing the type IIa error or the type IIb error rate (optimalities 2 and 3) leads to the same MPCA model with 98 PC's. The low type IIa and IIb error rates (resp. 0.2 % and 0.3 %) are paid off with a high type I error rate (67.8%). Optimality definition 4 leads to an MPCA model with fairly high dimensionality (34 PC's) and type I error rate as pay-off against the resulting low type IIa error rate (5.4%). Optimality 5 leads to a type I error rate of 9.4%, a type IIa rate of 39.8% and a type IIb rate of 3.5%.

Interestingly, the model obtained by optimality definition 1, requires a single PC only. Practically, this means the desription of reality defined by this model is a straight line in a 10800-dimensional space, which can be considered as a remarkable result. To investigate this further, the elements of the corresponding principal component, also referred to as loadings, are plotted in Figure 5.3(a). They were multiplied by the respective square root of the corresponding eigenvalue so that the original measurement unit (kg) is preserved. By means of this PC only, the

batch weight profiles are modelled to be the sum of (1) the mean profile, estimated prior to MPCA modelling and (2) the profile defined by the PC, multiplied by the respective score for the batch. This is illustrated in Figure 5.3(b), where the mean profile and the reconstructed profiles for score values equal to plus and minus 2 times the square root of the corresponding eigenvalue are plotted. The latter score values define approximate the theoretical 95% confidence region for the respective score. As can be seen, this description of reality by the one-dimensional model is at least very similar to what was shown before in Figure 4.3, hereby supporting the selection of 1 PC only. It also notable that the elements of the shown PC are smaller at the beginning and end of the cycle. As a result, obtained scores are less sensitive to the data at the beginning and the end of a cycle.

Table 5.2: GS- & AS-MPCA, Mode 1. Optimal models and corresponding performance indices.

scaling	optimality	#PC's	I	IIa	IIb	I & IIa	I & IIb
	1	2	9.1	37.6	4.7	23.4	6.9
	2	98	67.8	0.5	1.2	34.2	34.5
GS	3	98	67.8	0.5	1.2	34.2	34.5
	4	34	25.0	5.4	10.6	15.2	17.8
	5	1	9.4	39.8	3.5	24.6	6.5
	1	2	9.3	38.7	4.7	24.0	7.0
	2	100	67.5	0.5	1.2	34.0	34.3
AS	3	100	67.5	0.5	1.2	34.0	34.3
	4	34	25.2	5.9	9.4	15.5	17.3
	5	1	9.8	38.7	3.5	24.2	6.7

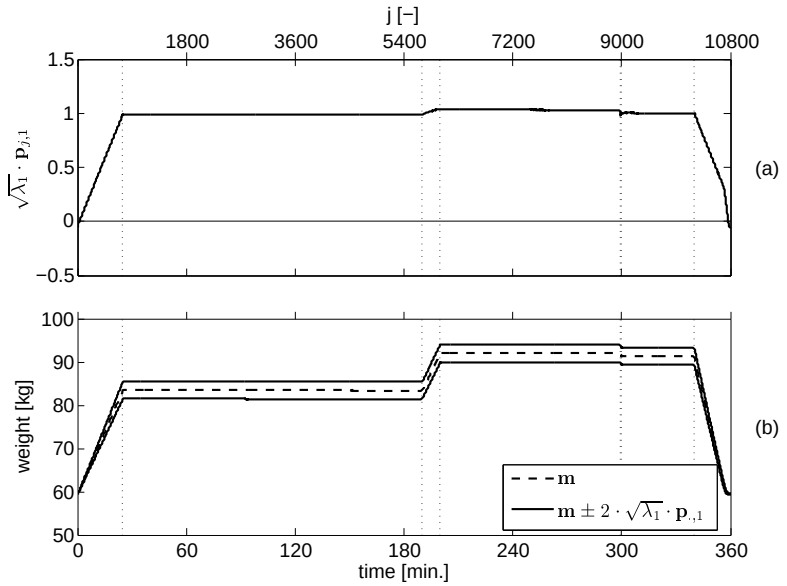


Figure 5.3: GS-MPCA, 1-PC model. (a) Loadings of PC 1 and (b) effect of the first score on the reconstructed data.

AS-MPCA The type II error rates for fault classes 2, 4 and 7 are shown in Figure 5.4 for the AS-MPCA models. Qualitatively speaking, the shown profiles are very simular to those shown for the GS-MPCA models (Figure 5.1). Differences are that models with 4 to 10 PC's or with 100 PC's (rather than 4 to 11 PC's) turn the type II error rate for fault class 2 down to zero.

The type I, type IIa and type IIb error rates for the AS-MPCA models are shown in Figure 5.2 as a function of the number of PC's. The type I error rate again shows an increasing trend from about 10% at small values for the number of PC's (1, 2) up to 70% at 100 PC's. Given that the profiles of the fault-specific type II error rates are very similar to the GS-MPCA case, it is not surprising that the type IIa and type IIb error rates are also similar in nature.

Table 5.2 summarizes the results for the models selected according to each of the optimality definitions. Minimizing the type I error rate (optimality 1) leads to the same dimensionality as for the corresponding GS-MPCA model (2). The resulting

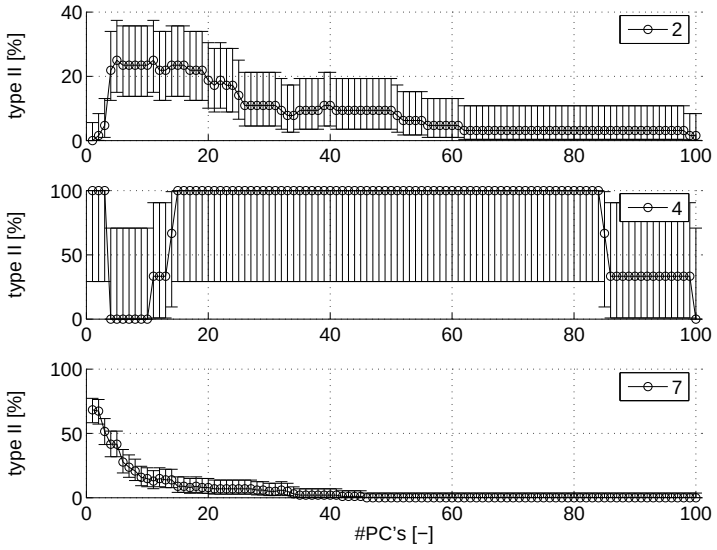


Figure 5.4: AS-MPCA, Mode 1. Type II error rates and corresponding confidence intervals for fault classes 2, 4 and 7 as a function of the number of PC's

low type I error rate (9.3%) is again paid off by a high type II(a) rate (type II(a): 38.7%), while the type II(b) error rate is similar to the previous set. Minimizing the type II(a) or II(b) error rates (optimality 2 and 3) lead to the same 100-dimensional model. This means that the maximum number of PC's that was evaluated is chosen. Low type IIa and type IIb error rates (0.5%, resp. 1.2%) are paid off here by a high type I error rate (67.5%). Minimizing the type IIa error rate (optimality 4) leads to the selection of 34 PC's (the same as for GS-MPCA) and leads to a 25.5% type I error rate, a 5.9% type IIa error rate and a 9.4% type IIb error rate. Minimizing the type IIb error rate (optimality 5) leads to a 1-dimensional model again. The corresponding type IIb error rate is the same (3.5%) while the type I error rate is slightly higher.

5.3.1.2 Mode 2

GS-MPCA MPCA models with up to 100 PC's were tested for the second mode. For any evaluated model (both GS-MPCA and AS-MPCA), the type II error rate for fault classes 1, 3 and 8 were zero, indicating that all of the faults therein are correctly identified as abnormal. The left-hand side limit of the 95% confidence interval is always zero for these classes. The right-hand side limits are given in Table 5.3. As in the case for the first mode, the resulting confidence intervals are wide due to the small number of observations in the respective classes.

In this mode, no batches were assigned to fault class 4 during screening. The type II error rates for all other fault classes are shown in Figure 5.5 as a function of the number of PC's for the GS-MPCA models. For fault class 2, the type II error rate is low ($<$3%) for models with up to 6 PC's. For 7 to 11 PC's the error rate increases and remains high ($\sim$25%) up to 30 PC's after which a slowly decreasing trend is observed. For fault class 5, none of the faults are detected at MPCA models with 1 or 2 PC's. However, because only 2 observations are present in this class, the confidence interval that is computed becomes very wide. For fault class 6, the type II error rate is non-zero for 1 to 3 PC's. Higher numbers of PC's deliver a 100% detection (zero type II error rate). Relatively small confidence intervals are found, especially for 4 PC's and higher (95% of the type II error estimates are expected between zero and 5%). Fault class 7 begets a zero type II error rates at 16 or higher numbers of PC's.

Figure 5.6 shows the type I, type IIa and type IIb error rates as a function of the number of PC's. The type I error rate shows an increasing trend from about 10% at small values for the number of PC's (1, 2) up to 90% at 100 PC's. Clearly, overspecification (too many PC's) leads to large type I error rates. Except for two relatively high values ($>$24%), the type IIa error rate shows a generally decreasing

Table 5.3: GS- & AS-MPCA, Mode 2. Right-hand side 95% confidence limit for the type II error rates for fault classes 1, 3 and 8.

Fault class	Right hand side confidence limit
1	17.6
3	70.6
8	84.2

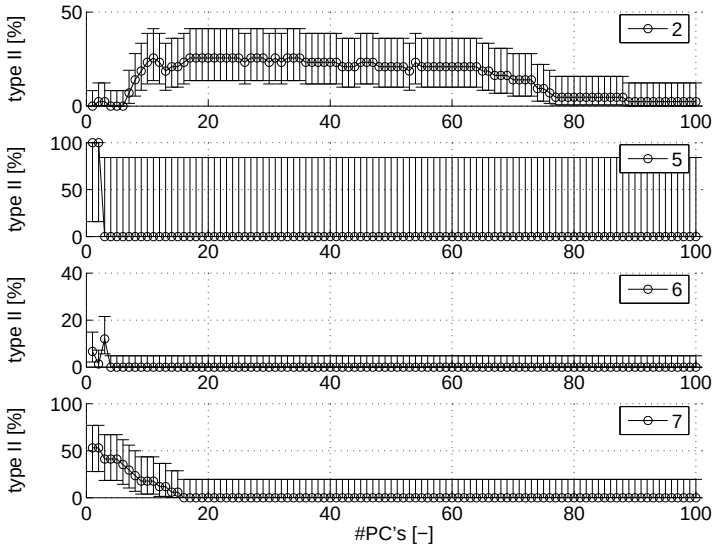


Figure 5.5: GS-MPCA, Mode 2. Type II error rates and corresponding confidence intervals for fault classes 2, 5, 6 and 7 as a function of the number of PC's.

trend from with increasing number of PC's. The type IIa and IIb error rate show fairly similar trend. At PC numbers 1 to 3, their values are relatively high (about 10%, resp. 5%). At PC's 4 to 6, the type IIa is about 5% and the type IIb error rate is zero. The widths of the confidence intervals for the type IIa and type IIb error rate are much wider than those for the type I error as a result of the much larger number of observations in the class of normal data.

Table 5.4 summarizes the results for the models selected according to each of the optimality definitions. Minimizing the type I error rate (optimality 1) leads to a low dimension of the model, namely 2. Both a low type I error rate (9.3%) and relatively low type IIa and IIb are reached by this model choice. In contrast, minimizing the type IIa error rate (optimality 2) leads to a 89-dimensional MPCA model. The resulting low type IIa error rate (0.6%) is naturally paid off by a high type I error rate (85.2%). A choice of 4 PC's leads to a zero type II(b) error rate. This choice does not lead to exceptionally high dimensions of the model (4 PC's) and does not compromise the type I error rate as much as the minimization of the type IIa error rate did. Optimality 4 and 5 lead to the same model as optimality 1, indicating that low type I and type II error rates can be achieved simultaneously.

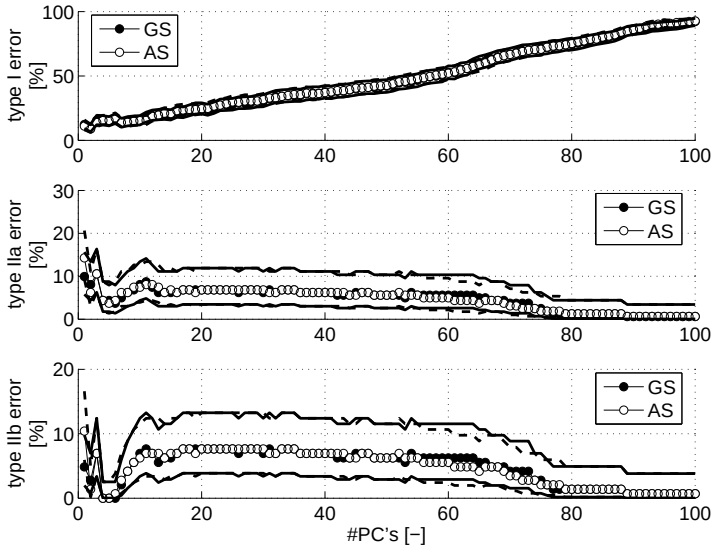


Figure 5.6: GS- & AS-MPCA, Mode 2. Type I and overall type II error rates as a function of the number of PC's. Full lines indicate confidence intervals for GS models. Dashed lines indicate confidence intervals for AS models.

Table 5.4: GS- & AS-MPCA, Mode 2. Optimal models and corresponding performance indices.

scaling	optimality	#PC's	I	IIa	IIb	I & IIa	I & IIb
GS	1	2	9.3	8.1	2.8	8.7	6.0
	2	89	85.2	0.6	0.7	42.9	43.0
	3	4	15.3	4.3	0.0	9.8	7.6
	4	2	9.3	8.1	2.8	8.7	6.0
	5	2	9.3	8.1	2.8	8.7	6.0
AS	1	2	8.8	6.2	1.4	7.5	5.1
	2	89	84.5	0.6	0.7	42.5	42.6
	3	4	15.5	4.3	0.0	9.9	7.8
	4	2	8.8	6.2	1.4	7.5	5.1
	5	2	8.8	6.2	1.4	7.5	5.1

AS-MPCA The type II error rates for the fault classes exhibiting non-zero values are shown in Figure 5.7 for the AS-MPCA models. The trends are very similar to the results shown for the GS-MPCA models. For fault class 5 a non-zero type I error rate is only found for the 1-PC model.

Figure 5.6 shows the type I, type IIa and type II b error rates as a function of the number of PC's. The type I error rate shows an increasing trend from about 10% at small values for the number of PC's (1, 2) up to more than 90% at 100 PC's. The type IIa and type IIb error rates are relatively low over the whole set of models. Minimal type IIa and IIb error rates are found at PC numbers 4 to 6. The IIb error rate is zero at PC numbers 4 and 5, as dicussed before.

In Table 5.4 the obtained results for models selected according to each of the optimality definitions are given. Minimizing the type I error rate (optimality 1) leads again to a 2-dimensional model, which is the same as for GS-MPCA. Also in this case, the low type I error rate (8.8%) does not compromise the type IIa and type IIb error rates much. Minimizing the type IIa error rate (optimality 2) leads -again- to a high dimension of the MPCA model (89 PC's). The same low type IIa error rate (0.6%) is paid off by a high type I error rate (84.5%). Minimizing the type IIb

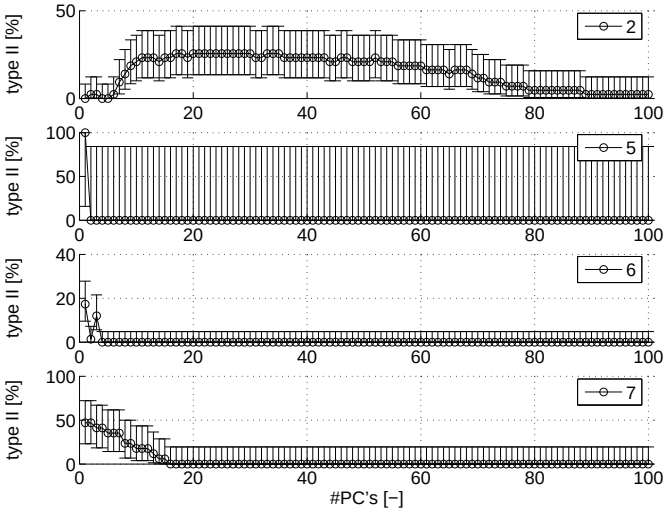


Figure 5.7: AS-MPCA, Mode 2. Type II error rates and corresponding confidence intervals for fault classes 2, 5, 6 and 7 as a function of the number of PC's.

error rate (optimality 3) warrants the choice of 4 PC's again and leads to a zero type IIb error rate as expected. Optimality 4 and 5, which minimize averages of type I and type IIa, respectively type I and type IIb, lead to the same choice as simply minimizing the type I error rate. Thus, also for the AS-MPCA models it can be concluded that, for the given mode and set of faults, both type I and type II error rates can be minimized to a satisfying extent.

5.3.1.3 Mode 3

GS-MPCA MPCA models with up to 50 PC's were tested for the third mode (the maximal number is 64 with the restricted data set). For this mode, the type II error rates for fault classes 2 and 6 were zero for any evaluated model. The right-hand side 95% confidence limits are given in Table 5.5. The left-hand side limit is always zero. The resulting confidence intervals are quite wide as a result of limited number of observations in these fault classes.

As no batches belong to fault classes 1, 3 or 5 in this mode, only the results of fault class 7 are shown in detail. Given that fault class 7 is excluded from the type IIb error rate, the latter is zero by default. As a result, the type IIb error rate is an obsolete measure for model selection in this case. It is noted here that if the type IIb error would be used anyway, optimality definition 3 would lead to an MPCA model with 1 PC by default. Optimality 5 automatically leads to the same model as derived by optimality definition 1. Only results for optimality definitions 1, 2 and 4 will therefore explicitly be reported. The type II error rates for fault class 7 in this mode are shown in Figure 5.8 for the GS-MPCA model series. The type II error rate exhibits a smoothly decreasing profile up to 25 PC's, whereafter the type II error rate becomes zero. Due to a limited number of observations in this class, the confidence intervals are fairly wide.

Table 5.5: GS- & AS-MPCA, Mode 3. Right-hand side 95% confidence limit for the type II error rates for fault classes 2 and 6.

Fault class	Right hand side confidence limit
2	52.2
6	97.5

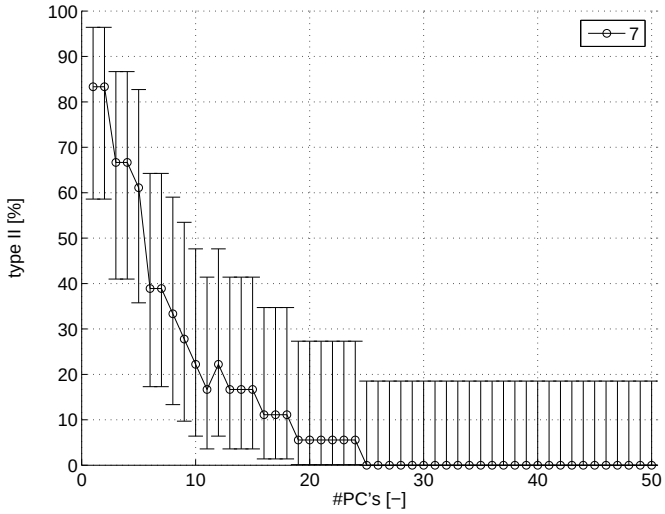


Figure 5.8: GS-MPCA, Mode 3. Type II error rates and corresponding confidence intervals for fault class 7 as a function of the number of PC's.

Figure 5.9 shows the type I and type IIa error rates as a function of the number of PC's. The type I error rate shows an increasing trend from about 5% at small values for the number of PC's (2, 3) up to 100% at 25 PC's. The type IIa rate shows a smooth decreasing trend for increasing number of PC's. This is not surprising given the smooth trend observed for fault class 7 in Figure 5.8.

Table 5.6 summarizes the results for the models selected according to the optimality definitions 1, 2 and 4. Minimizing the type I error rate (optimality 1) leads to a 2-dimensional MPCA model. A type I error rate of 4.6% is achieved and paid off by a relatively high type IIa error rate (62.5%). Minimizing the type IIa error (optimality 2) leads to a 25-PC model, a zero type II error rate and a 100% type I error rate. Practically, this means that all batches are expected to give rise to an alarm irrespective of their behaviour. Optimality 4, weighing the type I and type IIa error rates equally, leads to a 9-dimensional model. Both the type I and type IIa error rates are relatively high for this model choice.

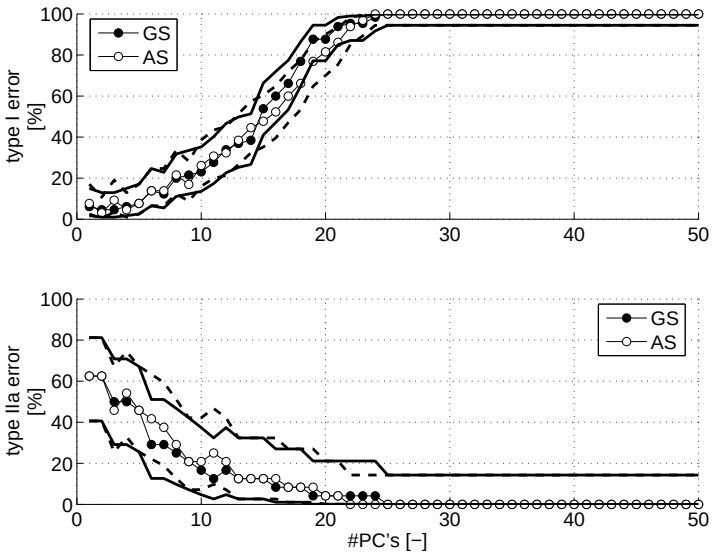


Figure 5.9: GS- & AS-MPCA, Mode 3. Type I and overall type II error rates as a function of the number of PC's. Full lines indicate confidence intervals for GS models. Dashed lines indicate confidence intervals for AS models.

Table 5.6: GS- & AS-MPCA, Mode 3. Optimal models and corresponding performance indices.

scaling	optimality	#PC's	I	IIa	I & IIa	
GS	1	1	2	4.6	62.5	33.6
	2	2	25	100.0	0.0	50.0
	4	4	10	23.1	16.7	19.9
AS	1	2	3.1	62.5	32.8	
	2	22	93.8	0.0	46.9	
	4	9	16.9	20.8	18.9	

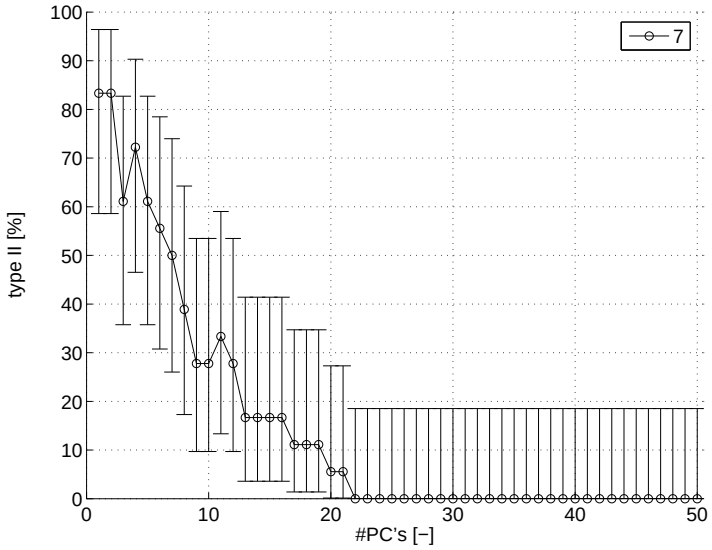


Figure 5.10: AS-MPCA, Mode 3. Type II error rates and corresponding confidence intervals for fault class 7 as a function of the number of PC's.

AS-MPCA The type II error rates for fault class 7 are shown in Figure 5.10 for the AS-MPCA models. The profiles are virtually the same as for the GS-MPCA models. One difference is that a zero type IIa error rate is obtained at 21 PC's.

Figure 5.9 shows the type I and type IIa error rates as a function of the number of PC's. The type I error rate shows an increasing trend from below 10% at small values for the number of PC's (1 to 5) up to 100% at 24 PC's. A smoothly decreasing trend is observed for the type IIa error rate, with values ranging from higher than 60% at low dimensions (1, 2) to 0% for high dimensions (22) of the model. It can be observed that the profiles are very similar to those for the GS-MPCA models.

Table 5.6 summarizes the results for the models selected according to each of the optimality definitions. Minimizing the type I error rate (optimality 1) leads to a 2-dimensional MPCA model. The resulting type I error rate of 3.1% is paid off by a relatively high type IIa error rate (62.5%). The minimization of the type IIa error rate (optimality 2 and 3) again leads to a zero type IIa error rate, balanced by a 93.8% type I error rate. By weighing the type I and type IIa error rates equally, 9 PC's are selected, delivering a type I error rate of 16.9% and a type IIa error rate of 20.8%. Again, both measures are relatively high.

5.3.1.4 Overall performance of single mode MPCA models

To compare the results of the single mode models with the MixMPCA models as applied in the upcoming section, the overall performance of the single mode models is evaluated by simple averaging of the performance measures over all modes. By doing so, equal weight is given to the performance in each mode, i.e their (Bayesian) priors are assumed to be the same. Practically, this means the overall type I error rate is calculated under the assumption that each mode has equal probability to occur in the future. By equal weighting of the performance measures, the optimal set of models is also the same as the set of individually optimal models. Note that this is only true because (1) changing the number of PC's for one model does not change the performance of the other models and (2) equal weighing was applied. In case prior knowledge is available on the respective probabilities of occurrence of each mode in the future, the weights should be changed accordingly so as to estimate the expected performance in the future.

The averaged performance measures are given in Table 5.7. Given that the analyzed operational modes are only different in terms of timing and not in their structure, it is interesting to note that optimality definition 1 leads to a 2-dimensional

Table 5.7: GS-MPCA & AS-MPCA. Averaged measures of performance.

scaling	optimality	#PC's per mode			averaged performance measures				
					I	IIa	IIb	I & IIa	I & IIb
		1	2	3					
	1	2	2	2	7.7	36.1	2.5	21.9	5.1
	2	98	89	25	84.3	0.4	0.6	42.4	42.5
GS	3	98	4	1	29.8	22.5	0.4	26.1	15.1
	4	34	2	10	19.1	10.0	4.5	14.6	11.8
	5	1	2	2	7.8	36.8	2.1	22.3	4.9
	1	2	2	2	7.0	35.8	2.0	21.4	4.5
	2	100	89	22	81.9	0.4	0.6	41.2	41.3
AS	3	100	4	1	30.2	22.5	0.4	26.3	15.3
	4	34	2	9	17.0	11.0	3.6	14.0	10.3
	5	1	2	2	7.2	35.8	1.6	21.5	4.4

model for all modes and both choices for scaling. This result suggests that two linearly uncorrelated variables are sufficient to explain naturally occurring variability, i.e. common-cause variability, in the normal operational data irrespective of the applied schedule. However, as at least two of three modes require higher dimensional models for any other definition of optimality, these low-dimensional models are performing less in terms of fault detection. Optimizing according to optimality definition 2 leads to the highest dimensionality for all models in both AS and GS model sets. Optimality 3, which excludes fault classes 7, reduces this effect for both mode 2 and 3. For mode 1, this relaxation has little effect. Optimality definitions 4 and 5 lead to models with lower dimensions. For both AS and GS models, the overall type I error rate decreases for optimality definitions going from 2 to 5 and is (naturally) lowest for optimality definition 1. In contrast, by comparison of the type IIa error rate for optimality definitions 1, 2 and 4 it can be observed that lower type IIa error rates are bargained against higher type IIb error rates. Similarly, comparion of the type IIb error rates for optimality definitions 1, 3 and 5 indicates that more stress on the type IIb error rates leads to higher type I error rates. Clearly, improved detection of faults is traded off against an increased number of false alarms.

To evaluate the choice for AS or GS models, each of the optimized performances are compared pair-wise. For AS-MPCA models the minimal type I error rate is 7.0% while it is 7.7% for the GS-MPCA models. Optimality definitions 2 and 3 lead to the same performance of the models in view of the optimized performance measure: 0.4% for both type Ia and type IIb error rates irrespective of the applied scaling. When the type I and type IIa error rates (optimality 4), resp. type I and type IIb error rates (optimality 5), are weighted equally, the AS-MPCA models perform best. All-in-all, AS-MPCA thus leads to models that are equally or more performant than the GS-MPCA models, even though the improvement of AS-MPCA over GS-MPCA models may be thought of as marginal.

5.3.2 Mixture Multi-way PCA (MixMPCA)

Mixture (Multi-way) PCA modelling serves to identify several models which describe multivariate data with different mean and covariance structure by means of several separate (Multi-way) PCA models, which jointly form an (M)PCA mixture. Data samples that accord to a model in a mixture are said to belong to the corresponding mode. Mixture Multi-way PCA (MixMPCA) as defined here inherently assumes that the membership to each of the modes is crisp, i.e. a (normal) data sample exhibit the characteristics of one single mode only and thus corresponds to one single mode as well. If knowledge is available on the membership of data samples to supposed modes, as in this case, each of the MPCA models can be identified separately. Otherwise, the assignment of data samples to a mode as well as the identification of the number of modes belongs to the model training step. Expectation-Maximation algorithms are viable strategy to train both the assignment of data samples to a given number of modes as well as the constituting MPCA models (Tipping and Bishop, 1999a). In the context of process monitoring, MixMPCA modelling is a valid approach if data samples are expected to belong to distinct distributions for which the mode is not known when the process monitoring task is pursued. Identifying the mode of a certain sample may even be the purpose of monitoring. MixMPCA allow to identify whether a sample belongs to any of the modelled modes as well as identification of the corresponding mode. The reader may note here that prior knowledge on the mode of the hydraulic system is likely to be available as the mode is the result of programmed controls of the system. In the authors experience, the recording of control actions, both human or automated, is not always pursued nor are such records managed properly by default. As such, MixMPCA may serve to reconstruct missing operational data of long-run systems.

The mixture modelling approach is defined as follows. A single model is made for each known mode present in the historical data set by construction of the MPCA model on normal batches within the given mode. The modelling step is therefore in essence not different compared to the previous approach. However, no knowledge on the mode is assumed in the monitoring step, i.e. when the models are used to detect faults. To identify an observation to be faulty or fault-free, all models are used simultaneously. If a given monitored batch violates all of their proper statistical limits, then that observation is classified as abnormal, as the data of the given observation are judged not to be in accordance with any of the known behaviours in the past. A batch is classified as normal if the batch does not violate the statistical limits (both Q as Hotelling's T^2) of at least one model. As no knowledge on the mode is presumed, the applied schedule of the batch can be assessed by

choosing the model to which the observation most likely belongs. For this second part in model application, identification of the mode, the Q statistic only was used as a measure for similarity. The assessed schedule is therefore the schedule corresponding to the model for which the p-value of the Q statistic is the largest. This is only done for batches that are accepted as normal. For batches classified as abnormal, it is assumed that the corresponding mode cannot be identified.

To evaluate the fault detection strategy on the basis of MixMPCA modelling, the overall performance measures are calculated in the same manner as the overall performance of the single MPCA model strategy in the previous section. This means that the type I, type IIa and type IIb error rates are calculated for each mode separately, equally weighting each observation in the mode. Then, each of the performance measures is weighted equally (equal priors of the modes) to calculate the overall performance measures. In addition, the second step in the application is evaluated by computation of the proportional amount of normal batches that are not assigned to their correct operational mode. Note that this fraction includes the normal batches that are not accepted. This misclassification rate, denoted as the Mode Misclassification Rate (MMR), can thus not be lower than the type I error rate. In Table 5.8, it can be seen that the given strategy leads to exactly the same model

Table 5.8: GS-MixMPCA & AS-MixMPCA. Averaged measures of performance.

scaling	optimality	#PC's per mode			averaged performance measures					
					I	IIa	IIb	I & IIa	I & IIb	MMR
		1	2	3						
	1	2	2	2	7.7	36.1	2.5	21.9	5.1	7.7
	2	98	89	25	84.3	0.4	0.6	42.4	42.5	84.3
GS	3	98	4	1	29.8	22.5	0.4	26.1	15.1	29.8
	4	34	2	10	19.1	10.0	4.5	14.6	11.8	19.1
	5	1	2	2	7.8	36.8	2.1	22.3	4.9	7.8
	1	2	2	2	7.0	35.8	2.0	21.4	4.5	7.0
	2	100	89	22	81.9	0.4	0.6	41.2	41.3	81.9
AS	3	100	4	1	30.2	22.5	0.4	26.3	15.3	30.2
	4	34	2	9	17.0	11.0	3.6	14.0	10.3	17.0
	5	1	2	2	7.2	35.8	1.6	21.5	4.4	7.2

choices and the same values for all measures of performance for either GS-MPCA as AS-MPCA models. It can therefore be concluded that the absence of knowledge on the applied mode in the data does not affect fault detection performance if approached with PCA-based mixture modelling as proposed here. In addition, the MMR is shown to be equal to the type I error rate, indicating that none of the accepted normal batches is assigned to the wrong mode.

5.4 Multisensor Monitoring

In this section the results for monitoring on the basis of the 6 selected variables are shown. These variables are weight, temperature, dissolved oxygen (DO), pH, ORP and the gas flow rate, adjusted by means of oxygen setpoint control. The measurements during the first 5 hours, i.e. including only phases with complete mixing, are taken. For all measurements this leads to 9000 samples (5 hours at 2-second interval), except for the gas flow rate for which only 4800 measurements are taken per batch, as the gas valve is only opened during non-aerated phases (160 minutes in total). Results are shown for a single operational mode in terms of hydraulics and aeration, previously indicated as mode 2a. In the first section, the type II errors for each fault class are shown separatedly first. In the second section, the type I errors and overall type II errors are discussed and model selection is pursued.

5.4.1 Class-specific Type II errors

5.4.1.1 Detection of outliers

As discussed in 4.5, the data series of the pH and ORP sensors often exhibit outliers. Their detection is evaluated separately here. To do so, results are shown for MPCA models with group scaling as preprocessing step. Only the batches in mode 2a that either are normal or either belong to fault class 19 and 20 are used here. This set has been expanded by *creating* an additional set of supposedly normal batches by reconciliation of the batches in fault class 19 and 20. To this end the identified outliers were removed by simple linear interpolation. Call this the reconciled set of batches. In the results shown here, the normal and reconciled data were used for estimation of the type I error by 10-fold cross-validation. The respective type II errors were estimated by projection onto an MPCA model calibrated by means of the normal and reconciled data. Figure 5.11 shows the type I error and the type II error for the investigated fault classes. As can be seen, the type II error rate for these fault classes is high (>40%) for any choice for the number of PC's. Due to the limited number of observations in the considered classes, the plotted confidence intervals for the type I error rate are rather wide.

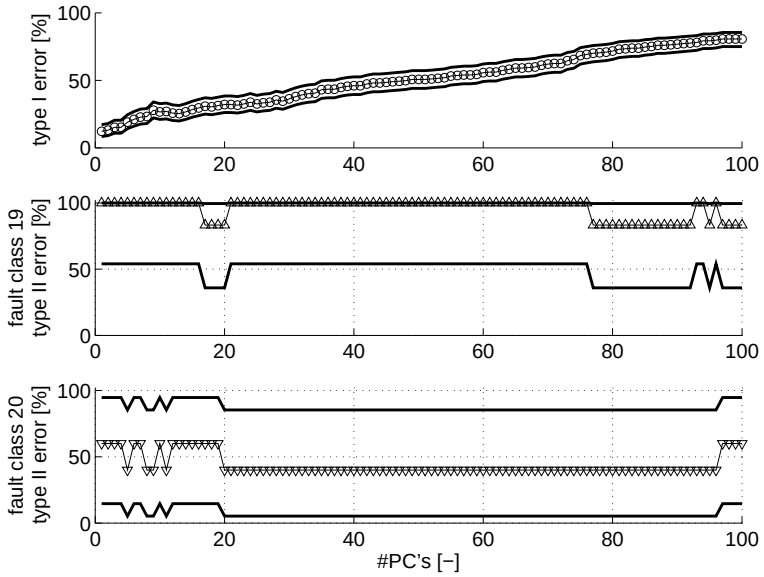


Figure 5.11: GS-MPCA. Overall type I error rate and type II error rates for fault class 19 and 20 as a function of the number of PC's. Full lines indicate confidence intervals.

Results were obtained by changing the models as follows:

- Applying autoscaling instead of group scaling
- Using only the normal data set
- Using only the reconciled data as normal set, i.e. by using only the (reconciled) batches of fault classes 19 and 20 for model calibration. The non-reconciled batches of fault classes 19 and 20 are thus projected onto the model calibrated with reconciled batches of fault class 19 and 20 only.
- Using only the respective ORP or pH signal data, i.e. an MPCA model is constructed only including the pH sensor or ORP sensor.

Neither change or combination of changes could deliver significant improvements. The constructed MPCA models are therefore shown to have limited power in view of outlier detection. This is likely be the result of the limited weight of a single

outlier on the constructed Q statistic and Hotelling's T^2 statistic, especially given the large number of variables (49800 variables). Theoretical and practical solutions to outlier detection have been studied (Fox, 1972; Stoodley and Mirnia, 1979; Chernick et al., 1982; Muirhead, 1986; Cai and Davies, 2003; Ranta et al., 2005). A more generic approach to fault detection including the detection of time-local events, such as -but not limited to- outliers, may be based on Multiscale MPCA (MSMPCA), as described in 3.3.3.5. The observed problem of outlier detection has however not been studied in more detail in this work.

In the following sections, results will be reported only for data that were reconciled for observational outliers, i.e. the outliers were removed by simple linear interpolation. The data set used for model calibration and estimation of the cross-validated type I error thus contains both the normal data set as well as the reconciled data set, counting 250 batches instead of 226 as indicated in Table 4.8. To establish type II error rates, outliers were removed as well for batches so far belonging to fault classes 30 to 46. Following this reconciliation, the reconciled batches were assigned to the corresponding class. For example, the batches in fault class 30 are assigned to fault class 1 after reconciliation for outliers (see Table 4.8).

5.4.1.2 Faults related to the hydraulic aspects of the system

Of the faulty batches exhibiting faults in the hydraulic parts of the system, those that belong to fault classes 1, 21 to 27 and 29 are all detected (no false acceptance) for any MPCA model chosen (including the applied scaling procedures and choices for selected PC's). Corresponding confidence limits for the type I error estimate are found in Table 5.9. For class 2, the confidence interval is relatively small (0-20.6%). For other classes (fault class 21, 27 and 29), related combinations of faults which do not occur as frequently, resulding in a much wider confidence interval.

Faults belonging to fault classes 2, 5 to 8 and 28 are not detected as efficiently. These 6 classes represent problems in the hydraulic system only, except for fault class 28. The latter class represents only a single batch exhibiting incomplete effluent withdrawal and an abnormal profile for the gas flow rate.

The type II error rates for fault classes 2, 5 to 8 and 28 are shown in Figure 5.12 for the GS-MPCA models and in Figure 5.13 for the AS-MPCA models. Batches in fault class 2 and 6 are most abundant among the classes shown (resp. 22 and 25 batches). For fault class 2, the type II error rate is below 10% for any choice

Table 5.9: GS- & AS-MPCA. Right hand side 95% confidence limit for the type II error rates for fault classes 1, 9 to 13, 18 and 21 to 29.

Fault class	Right hand side confidence limit
1	20.6
9	6.1
10	84.2
11	84.2
12	24.7
13	12.8
18	18.5
21	97.5
22	33.6
23	84.2
24	52.2
25	70.8
26	97.5
27	97.5
29	11.2

of model. For the GS-MPCA models, a 100% detection for models with 6 PC's or with 21 PC's and more. For the AS-MPCA models with 16 PC's or more, 100% detection is achieved. For fault class 6, a high type II error rate is observed for 1 to 6 PC's in the case of GS-MPCA models (>60%). Higher dimensional models lead to a decrease of the type II error rate, turning zero at 48 PC's. Selecting 48 or more PC's leads to a zero type II error rate except for models with 52 to 54 PC's. For AS-MPCA models, a decay in type II error rate is observed from 83% at 1 PC to zero at 17 PC's. Given the fair amount of faulty observations in fault class 2 and 6 (43, resp. 75), it is not surprising that the confidence intervals are relatively narrow. This stands in contrast to the resulting confidence intervals for other fault classes. For each of the faults classes 5, 7, 8 and 28, less than 5 batches are included. While results are thereby difficult to generalize and confidence intervals are very wide, it can be seen that for a model dimensionality sufficiently high, all faults are eventually detected. Complete detection of the batches in fault class 8 and 28 requires 10, resp. 9, PC's for GS-MPCA models and 8, resp. 5 in the case of AS-

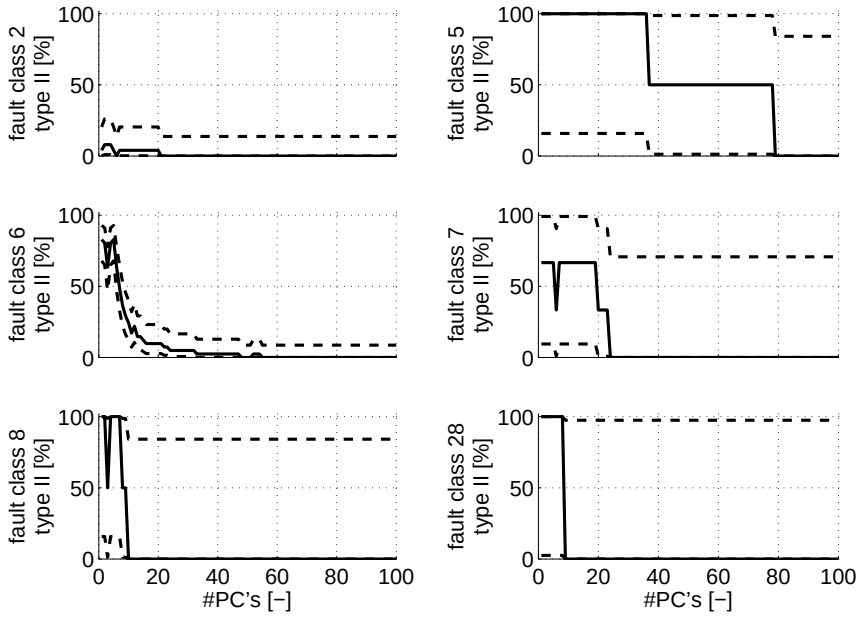


Figure 5.12: GS-MPCA. Type II error rates and corresponding confidence intervals for fault classes 2, 5 to 8 and 28.

MPCA models. Detection of batches belonging to fault class 7 requires 24 PC's for the GS-MPCA models. For the AS-MPCA models, either 47 PC's or more than 52 PC's are required. Fault class 5 requires that 79 PC's are included in the GS-MPCA model or 85 PC's in the AS-MPCA model to obtain detection of all observations.

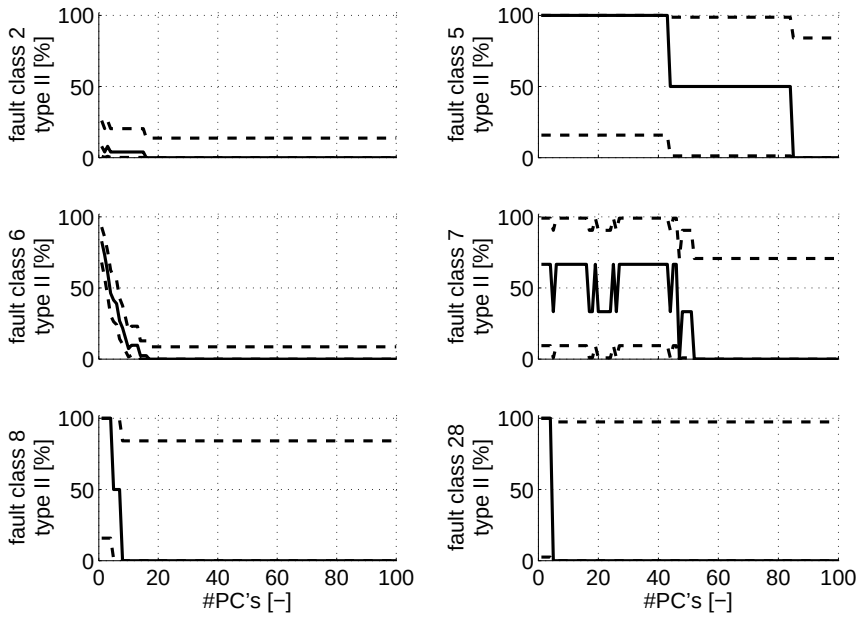


Figure 5.13: AS-MPCA. Type II error rates and corresponding confidence intervals for fault classes 2, 5 to 8 and 28.

5.4.1.3 Faults related to the cooling system

All faulty batches corresponding to a failure to control the temperature effectively were detected by means of the MPCA models for any evaluated choice for the number of PC's and for any scaling method. This may not be a surprise given the large deviations of temperature measurements from normal operation values that result from such failures (see Figure 4.7). The confidence intervals for the corresponding classes (fault class 9 and 21) can be found in Table 5.9. The confidence interval is fairly narrow for fault class 9, which results from the rather frequent occurrence of faulty observations of the corresponding type. For fault class 21, including only one observation, a very wide confidence limit is obtained.

5.4.1.4 Faults related to the aeration system

Fault classes 10 to 17 and 22 to 28 are related to anomalies in the aeration system. Of these, fault classes 10 to 13 and fault classes 22 to 27 result in a zero type II error for any modelling choice made. The situation is different for fault classes 14 to 17 and fault class 28. Figure 5.14 and Figure 5.15 show the resulting type II error rates for each of the latter classes. Fault classes 14, 15 and 16 contain the most observations. As a result the reported confidence intervals for the type II error rates are relatively narrow.

For fault class 14, the type II error rate is non-zero for GS-MPCA models with 1 or 4 PC's. A non-zero type II error is also found for AS-MPCA models with more than 2 PC's. For class 15, the type II error rate for the GS-MPCA model decreases from 47% for 1 PC to zero at 6 PC's. Beyond 6 PC's the type II error rate is only non-zero at 7 PC's. When using an AS-MPCA model, a zero type II error rate is only observed for models with 14 or more PC's. For faults of class 16, exhibiting oscillatory behaviour in the gas flow rate, a high type II error rate ($>$50%) is observed for any choice of scaling and PC's. Clearly, the reported oscillatory behaviour in the trajectories of the gas flow rate is difficult to discriminate from normal behaviour.

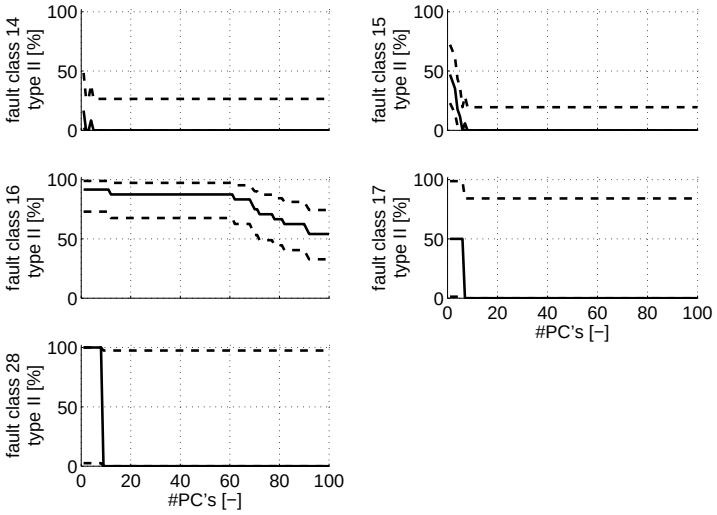


Figure 5.14: GS-MPCA. Type II error rates and corresponding confidence intervals for fault classes 14 to 17 and 28.

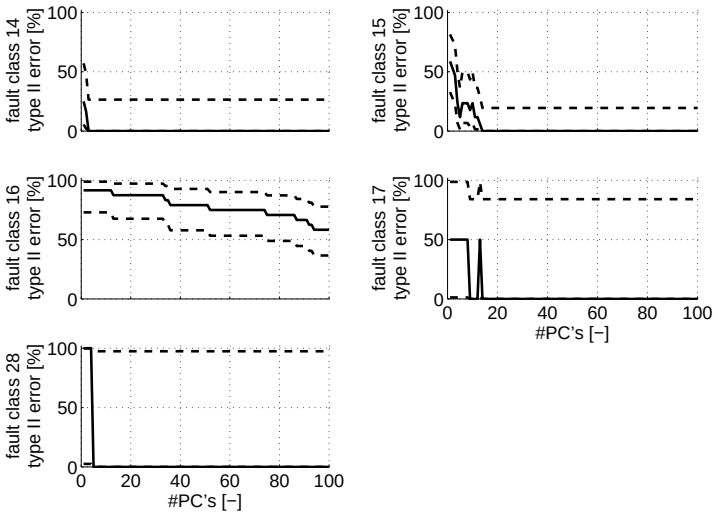


Figure 5.15: AS-MPCA. Type II error rates and corresponding confidence intervals for fault classes 14 to 17 and 28.

For fault class 17, both batches included are detected by GS-MPCA models with more than 6 PC's and by AS-MPCA models with 9 to 12 or with more than 13 PC's. As discussed already above, 9 (5) PC's are required to detect the (single) batch in fault class 28 for the GS-MPCA (AS-MPCA) model. Given the small numbers of batches included in classes 17 and 24, the latter observations are naturally difficult to generalize. This is also clear from the plotted confidence intervals for the corresponding classes.

5.4.1.5 Gross errors in pH measurements

Batches exhibiting fault symptoms in the pH profiles unrelated to problems in aeration are limited to classes 18, 24 and 29, which are all related to (the same) event, being a faulty calibration of the pH sensor. The batches belonging to these classes are all detected by means of MPCA for any evaluated choice for the number of PC's and for any scaling method. Similar to the case of the cooler failures, this event results a fairly large deviations of the pH values from their normal region (see Figure 4.10). As such, the good performance is not a surprise. The left-hand side confidence limits are all zero for these classes. The right-hand side limits can be found in Table 5.9. The confidence intervals defined by these limits are relatively narrow for fault classes 18 and 29, due to the relatively large number of observations within these classes. For fault class 24, including only 4 observations (see Table 4.8), the confidence interval is much wider.

5.4.2 Overall type I and type II error rates

In Figure 5.16, the type I, type IIa and type IIb error rates are shown for the evaluated models together with computed confidence intervals. As one can see, the type I error rate generally increases as more PC's are included. The reverse is true for the type IIa and type IIb error rates. Little difference is seen between type IIa and typeIIb error rate values. Also, little difference is observed between performance measures for the two evaluated scaling approaches. The largest decrease in the type II (a and b) error rates is observed between 1 and 10 PC's.

For each of the optimality definitions, an optimal model is selected for both types of scaling (i.e. GS-MPCA and AS-MPCA). Table 5.10 summarizes the results for the evaluated models. A 1-PC GS-MPCA model leads to a minimal type I error

rate, while 92 PC's lead to a minimal type IIa and type IIb error rate. Given the former graph, this should not be a surprise. Optimality 4 and 5, in which the type I errors and type IIa, resp. type IIb, errors are equally weighted against each other, lead both to a 3-PC model. For this choice, the type I error rate is 15.2%, the type IIa error rate is 19.7% and the type IIb error rate is 19.3%. The overall performance indices are 34.9% (type I&IIa) and 34.5% (type I&IIb).

Consider now the AS-MPCA models. Also in this case, a A 1-PC model leads to a minimal type I error rate, while 94 PC's deliver a minimal type IIa and type IIb error rate. Optimality definitions 4 and 5, lead both to a 4-PC model. For this choice, the type I error rate is 16.0%, the type IIa error rate is 16.9% and the type IIb error rate is 16.5%. The overall performance indices are 33.0% (type I&IIa) and 32.5% (type I&IIb).

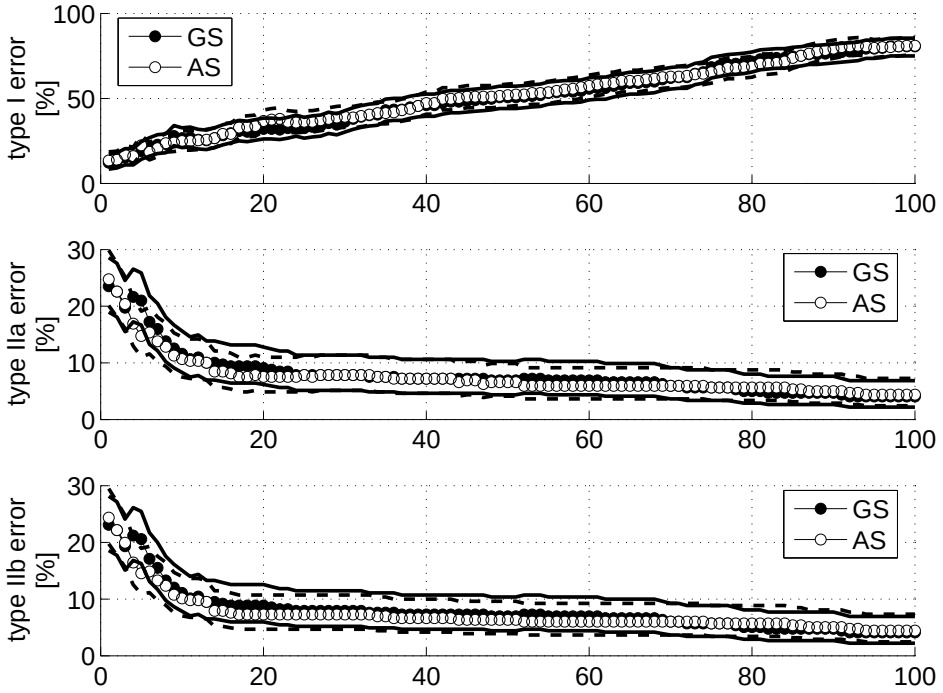


Figure 5.16: GS-MPCA & AS-MPCA. Type I and overall type II error rates as a function of the number of PC's. Full lines indicate confidence intervals for GS models. Dashed lines indicate confidence intervals for AS models.

Table 5.10: GS-MPCA & AS-MPCA. Optimal models and corresponding overall error rates.

scaling	optimality	#PC's	I	IIa	IIb	I & IIa	I & IIb
GS	1	1	12.2	23.5	23.1	35.7	35.3
	2	92	77.6	4.1	4.1	81.7	81.8
	3	92	77.6	4.1	4.1	81.7	81.8
	4	3	15.2	19.7	19.3	34.9	34.5
	5	3	15.2	19.7	19.3	34.9	34.5
AS	1	1	13.5	24.8	24.4	38.3	37.9
	2	94	80.6	4.4	4.4	85.0	85.0
	3	94	80.6	4.4	4.4	85.0	85.0
	4	4	16.0	16.9	16.5	33.0	32.5
	5	4	16.0	16.9	16.5	33.0	32.5

Table 5.11 shows the class-specific type II error rates for the GS-MPCA models. As already noted before, the detection of the batches belonging to a fault classes 1, 9 to 13, 18 to 27 and 29 is always 100%, giving rise to a zero type II error rate that is insensitive to the model choice. Except for fault class 16, the application of optimality 2 and 3 turns all class-specific type II error rates to zero percent. Optimality 4 and 5 lead to very high type II error rates ($>$90%) for fault class 5, 16 and 28 indicating that the faults in these classes are difficult to discriminate from normal behaviour. Fault class 5 contains batches for which the sludge withdrawal is too large. Given that sludge occurs in the last minute before the settling phase starts, the obtained results are not be a surprise. Note that for separate monitoring of the hydraulic parts of the system, the weight measurements during settling and draw were included. As such the weight measurement during the settling phase (45 minutes, 1350 measurements), for which the values are resulting from sludge withdrawal have a considerably larger impact on the MPCA model than the (30) samples included for MPCA modelling here. It is therefore not a surprise that the obtained model cannot discriminate these fault. In addition, fault class 5 contains only 2 faulty batches which have marginal weight in the averaged type II error rate and thereby have little impact on the model choice. The same is true for the single batch in fault class 28. Fault class 16, characterized by oscillatory behaviour of the aeration control system, contains more batches (16), yet does result in large false acceptance. It can therefore be concluded that the oscillatory behaviour of oxygen

and gas flow rate cannot be detected easily by the given models. High type II error rates (>20%) are also found for fault classes 6, 7, 15 and 17. For fault class 2 a type II error rate of 8% is obtained, which can be considered as acceptable.

Table 5.11: GS-MPCA. Type II error rates per fault class for each model choice.

optimality \ fault class	1	2	5	6	7	8	9	10	11
1	0.0	4.0	100.0	82.9	66.7	100.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	8.0	100.0	63.4	66.7	50.0	0.0	0.0	0.0
5	0.0	8.0	100.0	63.4	66.7	50.0	0.0	0.0	0.0

optimality \ fault class	12	13	14	15	16	17	18	21	22
1	0.0	0.0	16.7	47.1	91.7	50.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	54.2	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	54.2	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	35.3	91.7	50.0	0.0	0.0	0.0
5	0.0	0.0	0.0	35.3	91.7	50.0	0.0	0.0	0.0

optimality \ fault class	23	24	25	26	27	28	29
1	0.0	0.0	0.0	0.0	0.0	100.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	100.0	0.0
5	0.0	0.0	0.0	0.0	0.0	100.0	0.0

The class-specific type II error rates for the AS-MPCA models are shown in Table 5.12. Again, fault classes 1, 9 to 13, 18 to 27 and 29 lead to a zero error rate for any choice of model. Optimality 4 and 5 lead to very high type II error rates (>90%) for fault class 5, 16 and 28, confirming that MPCA-based detection of these types of faults is difficult. Similarly to the GS-MPCA case, high type II error rates (>20%) are also found for fault class 6, 7, 15 and 17. For fault class 2 a type

Table 5.12: AS-MPCA. Type II error rates per fault class for each model choice.

optimality \ fault class	1	2	5	6	7	8	9	10	11
1	0.0	8.0	100.0	82.9	66.7	100.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	4.0	100.0	46.3	66.7	100.0	0.0	0.0	0.0
5	0.0	4.0	100.0	46.3	66.7	100.0	0.0	0.0	0.0

optimality \ fault class	12	13	14	15	16	17	18	21	22
1	0.0	0.0	25.0	58.8	91.7	50.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	58.3	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	58.3	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	23.5	91.7	50.0	0.0	0.0	0.0
5	0.0	0.0	0.0	23.5	91.7	50.0	0.0	0.0	0.0

optimality \ fault class	23	24	25	26	27	28	29
1	0.0	0.0	0.0	0.0	0.0	100.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	100.0	0.0
5	0.0	0.0	0.0	0.0	0.0	100.0	0.0

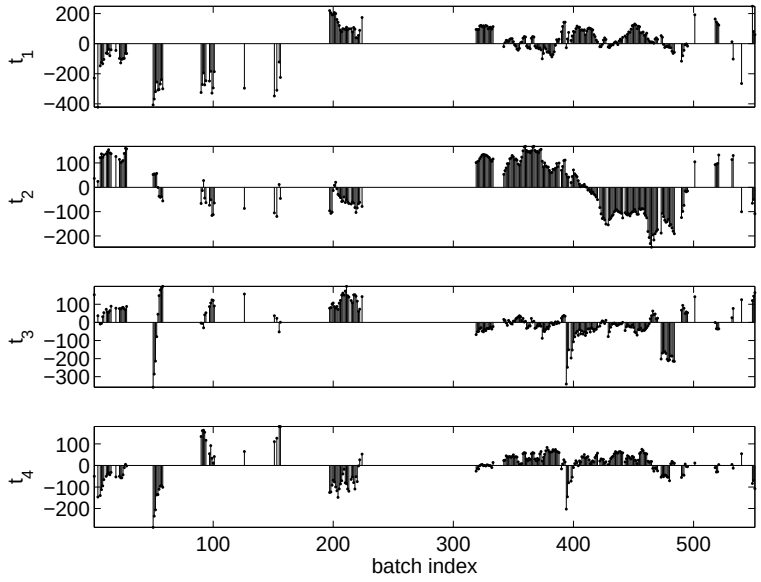


Figure 5.17: Trajectories of the (first) 4 principal scores of the AS-MPCA models. Unsteady non-random behaviour suggests that the MPCA model violates the necessary presumptions for use in monitoring.

II error rate of 4% is obtained, which is in acceptable range. The latter description of results obtained for AS-MPCA models is therefore not different from the one for GS-MPCA models. This suggests that scaling parameters have limited effects on model detection performance for the observed faults.

All-in-all, the reported performances for the models including all sensors are fairly low compared to those obtained for the hydraulic system only (see e.g. Table 5.8). To investigate this problem further, the results for best performing model according to optimality 4 are studied in more detail. Figure 5.17 shows the scores for the batches that were assessed to be normal as a function of batch index (i.e their order in the consecutive series of batches). The scores do not show a steady pattern as would be expected for a constant mean process, which is presumed in standard MPCA modelling (i.e. non-dynamic, non-adaptive). This is presumption is not generally valid for biological SBR systems. Temperature changes, influent flow rates and influent load are among the most typically reported sources of variation in wastewater treatment systems. Given a relative large proportion of changes in influent flow rates and temperature in the studied timeframe, mostly due to failures

in the hydraulic and cooling system, the studied system is equally subjected to substantial variation in temperature and influent changes.

Given that active biomass, transferred from the one batch to the next, are the bioprocess catalysts and are affected by changes in temperature and influent, the process does not return immediately to its nominal operation or state after solving an occurred problem. In fact, upon constant influent flow rates, influent concentration and temperature, the process may take months to return to a so called pseudo steady-state condition for which the biomass concentration, biomass conversion and effluent quality are in equilibrium with the named input variables. Indeed, given the implemented sludge retention time (SRT) of 15 days, i.e. the average time that a unit of active biomass remains within the system (before leaving the system by sludge wastage), this may indeed be expected. With each fault in the hydraulic system or cooling system, a disturbance, small or large, may impede that the process state converges to (pseudo) steady state. For the system and fault as described, an adaptive updating scheme for the MPCA model will however not likely be successful due to the often sudden occurrence of system failures. Indeed, the reported variations in system temperature and influent load to the system are seldom of a smooth or gradually evolving kind. Cooler failures gave rise to rather sudden temperature changes (i.e. from one batch to another batch) not to be expected in full-scale wastewater treatment systems where changes in ambient temperatures rather than cooler failures give rise to temperature fluctuations. Sudden fluctuations in influent flow rate and concentrations may occur in practice and may thus prevent the use of adaptive schemes in practice as well. Practically, initiation of the construction of new MPCA models may need to be started in the event of solving a (major) failure. While MPCA model construction may be guided automatically, e.g. on the basis of reconstruction error or captured variance, identifiability problems may arise due to limited data series available.

5.5 Discussion

In this chapter, Multi-way PCA (MPCA) models for were evaluated for monitoring of the studied SBR system. By means of a screened data set, including several faults, MPCA model selection was pursued in view of optimizing 5 monitoring performance criteria. Three of these optimalities (1 to 3) aimed at minimizing either the type I, type IIa or type IIb error rate. Optimality 4 and 5 were defined as mixed criteria, where equal weighing of type I and type IIa, resp. type IIb, errors was implied.

Two scaling approaches were tested in the presented work, being group scaling (GS) and autoscaling (AS). While AS models were shown to outperform the GS models, this is only by a marginal effect on the resulting performance indices. It thus appears that the scaling had little effect on the monitoring performance. Two distinct hypotheses are proposed here. Either (1) the scaling has little effect on MPCA models in general or (2) the reported faults are of such nature that scaling has little effect. For faults that are characterized by a mean shift, the latter hypothesis is likely to be true. To enable an effective evaluation of the two hypotheses, simulated examples may serve better so as to gain control over the assumptions underlying the MPCA models, i.e. linear behaviour of the process, constant mean process and constant covariance matrix.

Fault-specific type II error rates were evaluted for twenty-nine separate fault classes, of which 8 were identified for the hydraulic parts of the system only. Combinations of faults were considered to belong to separate classes and not to multiple classes at the same time. By doing so, the additive nature of constituting faults, i.e. when the response of the combined faults is the sum of the responses of the two individual faults, needs not to be verified.

Confidence intervals were computed for the type II error rate. It is repeated here that the approach that was used, assuming a binomial distribution model for each fault class, is considered as naive. First of all, with such binomial model it is assumed that the classification outcome for a single observation (i.e. normal/abnormal) is independent of the classification outcome for other observations. As many of the faulty observations are the result of a single event or anomaly, such independence is unlikely. Secondly, the binomial model assumes that the probability for a certain outcome is constant for all observations belonging to a given fault class. This not likely to be true either as for several classes the magnitude of the reported fault

will likely affect the probability for detection. Thirdly, combinations of faults were considered as separate classes. Due to the limited number of observations in these classes, the computed confidence limits become wide and generalization of the results is difficult for these combinations of faults. A more elaborate model for the misclassification errors may account for correlated behaviour of faulty observations, effects of magnitude on misclassification and additive or multiplying effects on the probability of misclassification. It is noted that the magnitude of faults is not known a priori for the studied data set and that additive or multiplicative effects of multiple faults are not easy to verify, let alone quantify, given the rather limited numbers of faulty observations affected by multiple faults.

Based on a data set including only the weight variable, it was shown that MPCA effectively allows to detect a majority of the identified faults. For some fault classes, including fault classes 1, 3 and 8, a 100% detection was obtained for any model chosen. As fault class 1 can be considered a so called *hard* fault class, i.e. being characterized by complete failure of the hydraulic system, this result may not be so surprising for this class. Other faults are of a *softer* nature, i.e. not complete failures. The observed insensitivity of the monitoring performance therefore suggests that for these faults the magnitude of the reported faults is so large that the chosen distance measure is not crucial to their detection. Batches in fault class 7, characterized by noisy artefacts in the data, were shown to be much more difficult to detect. Since fault class 7 is excluded in the type IIb error (and included in the type Ia error), different model choices resulted according to optimalities 2 and 3, resp. 4 and 5.

In addition to single mode MPCA models for the weight variable, the use of Mixture Multi-way PCA (MixMPCA) models was evaluated for the hydraulic system. It was shown that the performance of the MPCA models in terms of fault detection is not influenced by the absence of knowledge about the actual mode the system is in. Also, it was reported that the assignment of normal batches to their mode was not larger than the type I error rate. This indicates that MixMPCA modelling is a valid choice for process monitoring when multiple modes are available in the process history and will be so in the future. One should note here that the number of modes and the membership of the observation to each of the modes was known and used during the modelling of the MPCA mixtures. In the absence of such prior knowledge, the number of modes and assignment of observations to a certain mode is part of the modelling procedure. This remains to be tested for the given system.

MPCA-based monitoring was tested as well to monitor the complete system, i.e. including all available data. A larger part of the fault classes were shown to lead to insensitive type II error rates, being zero for any model choice. While a larger part of the faults could be detected, observed overall type I and type II error rates were rather high and may not satisfy requirements in practice. To assess the underlying cause of this, the best performing model was selected for further investigation. By means of score plots, it was shown that the normal batches did not represent a constant mean process. The difference in behaviour is believed to be the result of changed process characteristics due to past faults, such as cooler failures and unintended changes in loading. While from a theoretical point of view, such changed process characteristics may be accounted for by constructing new MPCA models whenever such a process change is found, automated fault detection or alarm generation may not be possible anymore on the basis of MPCA. Indeed, without further information, it may not be known a priori if a change in the process will be the (good) result of a correct problem-solving action or is due to an even worse situation (e.g. when the problem is not effectively solved). Clearly, when problems occur which induce changes in process characteristics that remain upon solving the problems, a straightforward MPCA model excercise without input of knowledge or information alien to the data themselves is not likely to be an optimal tool for fault detection.

Observational outliers in trajectories of pH and ORP signals as well as undesired oscillation were shown to be difficult to detect by means of the deployed MPCA models. With respect to this problem, classic tools for outlier detection (e.g by filters) and band spectra may be used. Alternatively, wavelet-based extensions to PCA (e.g. Multi-Scale MPCA) approaches may allow for improved detection of such faults while only requiring a single overall model.

Improvement in terms of MPCA model identification or PCA model identification in general may be obtained by including specific fault classes to identify the subspace spanned by the PC's of the model. Consider that certain faults can be characterized as extreme events for which the noise-free variables lie in the hyperplane defined by the noise-free normal data. A simulated example of such an event is the extreme flow event considered for Example 1 in Section 3.3.1.5. Such faults may likely be detected by means of a violated Hotelling's T^2 statistic while not violating the Q statistic (as shown in Section 3.3.1.5). However, this requires that the dimensions of the fitted hyperplane by a PCA model are correctly identified and its (angular) position lies close enough to the hyperplane defined by the error-free variables. In this context, an improved fit of the hyperplane, i.e. less

subjected to variance, may be obtained by including the considered extreme events to model the hyperplane in which the normal (and the extreme events) lie. Indeed, as the extreme event observations will be located at a considerable distance from the normal data, increased leverage of observations will be obtained by including these observations for hyperplane or subspace identification. Put otherwise, the addition of extreme events to the data set increases the relative proportion of variance within the subspace defined by the error-free variables compared to the total variance hereby improving model identification. Naturally, this assumes that the error-free data of the included extreme events do not violate the true correlation structure of the normal data, i.e. it is assumed that the extreme events truly lie in the hyperplane defined by the error-free normal data. If this assumption is valid, the identified subspace will be subjected to less variance while not increasing the bias. Angular distances between PCA models identified with and without the extreme event data may be used to evaluate (a posteriori) whether this assumption is valid. Approachs to do so date already back to Wold (1976). After subspace identification, the covariance matrix of the scores of the normal data can be calculated to construct the Hotelling's T^2 if applicable.

Another improvement of MPCA model identification -which does not exclude the former- may be achieved by constraining the identified principal components. As noted already in 3.3.3.3, large numbers of variables and limited numbers of observations may lead to unwanted variance of the obtained MPCA model. Function Space PCA allows to tackle this problem by converting the obtained data into a (lower-dimensional) set of coefficients of orthogonal basis functions, which are then consequently analyzed or modelled by means of a MPCA model. This approach is expected to lead to reduced variance of identified parameters (mean, scaling parameters and PC's) thereby giving increased generalizing power to the model. This approach may however induce bias in the identified MPCA model, which needs to be traded off with the reduced variance.

Alternatively, an orthogonal basis may also be defined by input of knowledge. Consider for example the typical normal weight profiles of the studied system (see Figure 4.3). These profiles may be approximated well by linear combinations of piece-wise linear and/or quadratic functions, e.g. splines. Practically, a spline basis may thus be used for which the coefficients are then to be analyzed by a MPCA model. The piece-wise linear fits may be constrained additionally to have zero slope (i.e. piece-wise constant) when the pump is off. By using such orthogonal basis, improved interpretability of the resulting MPCA model may be achieved in addition to reduced variance of the resulting model parameters. The resulting first

PC's will express information that closely relates to the knowledge or assumptions used during the basis choice.

When knowledge is lacking or no assumptions can be made, the generic procedure by (Chen and Liu, 2001) may be used to identify a proper basis. Also, boosting algorithms (see e.g. Hastie et al. (2001)) may be used to identify a proper resolution for the identified basis while limiting computation time. Alternatively, a given principal component calculated in a standard unconstrained fashion may be approximated a posteriori by means of wavelet shrinkage (Donoho and Johnstone, 1994) prior to computation of the next PC. It is noted that investigation of the PC's themselves such as done for Figure 5.3 may serve to indicate applicable constraints or suitable basis functions if exact knowledge is lacking.

Given a set of basis functions based on knowledge or other models, relevant information, not accounted for by the incorporated knowledge, may be lost. To evaluate or to account for this, more (non-constrained) PC's may be calculated to capture variance that is not captured by the constrained MPCA model. This can practically be achieved by MPCA analysis of the residuals calculated by means of the already obtained MPCA model. Note that a less effective dimension reduction is a potential price to pay by doing so. This is to be expected when the applied basis does not provide a sufficient fit to the data. On the positive side, if the provided basis is based on a mental or mechanistic model, anything except the constrained PC scores will be an expression of the deviation from that mental or mechanistic model. As such, the additional non-constrained PC 's may reflect missing knowledge and may therefore serve to adjust existing knowledge and/or to accomodate the monitoring strategy accordingly. Note that the latter approach for data mining is not limited to a constrained MPCA model. Any model can be used to generate residuals, i.e. information not explained by the respective model, which can be processed further by MPCA or any other viable technique for that matter. However, if the applied model implies a nonlinear transformation of the measurement errors, a simple MPCA model may not satisfy the requirements and theoretical statistical limits may not easily be computable.

5.6 Conclusions

Process monitoring by means of Multi-way Principal Component Analysis (MPCA) was pursued for the hydraulic parts of the studied SBR system first. It was shown that this leads to effective and efficient detection of a larger set of faults. However, undesired noisy artefacts in the data and other faults located in small timeframes were not detected efficiently. To overcome the latter problem, approaches effectively dealing with changes in the frequency domain such as Multi-Scale (M)PCA may provide effective solutions.

In view of monitoring of the complete system, the PCA-based strategy was evaluated for the complete system as well. This was shown to deliver effective detection for only a few types of faults. The lower performance has been attributed to a too large deviation from the assumption that the modelled process has a constant mean and covariance structure. For the system as described, including its many disturbances, adaptive monitoring schemes may be not be effective due to the sudden nature of faults and corresponding actions to resolve problems. A crucial and unsolved problem in this context is that a detected process change can be either equally be the result of a fault or a solution to a problem if no other information is provided. Future research is therefore warranted to allow the incorporation of additional information alien to the PCA model into the monitoring scheme.

In view of improved (M)PCA model identification, two approaches have been proposed. One is based on the inclusion of faulty data, which are characterized as extreme events but not as breakages of the correlation structure. This approach may allow a better subspace identification due to increased leverage of the modelled data. After subspace identification, a second MPCA model can be constructed on the scores of normal data by projection onto the first model, so to obtain a final monitoring model. A second proposal to improve PCA model identification is based on the application of constraints to the identified MPCA model so to reduce the inherent variance of the identified model solution. This is likely to result in a biased model. Function Space PCA, as found in literature is a generic method to impose such constraints. If prior knowledge is available, this may be used to define a set of basis functions that are according to the mental or mechanistic model of the system.

Chapter 6

Diagnosis of a Sequencing Batch Reactor by means of Multi-way Principal Component Analysis and Fuzzy C-means Clustering

6.1 Introduction

For a recent and extensive overview of fault diagnosis methods, the reader is referred to Venkatasubramanian et al. (2003a,b,c). In this introduction, special attention is given to PCA-based diagnosis methods. PCA-based approaches for fault diagnosis can largely be split into three categories.

The first category is based on so called contribution plots (Kourti and MacGregor, 1995). In this approach, a violated statistic (Hotelling's T^2 or Q statistic) is decomposed into contributions of each variable to the statistic. Variables that are *responsible* for the violation of the statistic will likely contibrute more to the statistic so that their identification may serve to diagnose an ongoing problem. It is noted that this technique is in its essence a fault isolation technique. Indeed, the measurements that result in a violated statistic are indicated, not the root cause of the problem. Further analysis, e.g. by investigating the magnitude of several

significant contributions, is typically required to uniquely identify the root causes. Importantly, contributions to statistics have an equal number of dimensions as the original set of variables. As such, the dimension reduction by PCA analysis for process monitoring has no equivalent reward for process diagnosis. It is also noted that, with J original variables, a problem which leads to a violation of the Q statistic can theoretically result in the location of the observation in any point in the J-dimensional space. Given the latter observations, it is no surprise that contribution plots have largely been limited for interpretation by human operators, i.e. non-automatic diagnosis.

A second framework for PCA-based diagnosis has been proposed by Dunia et al. (1996) and extended in additional papers (Dunia and Qin, 1998a,b,c; Qin and Li, 1999, 2001). Underlying to the proposed diagnosis scheme is that a complete and well-defined set of potential faults is known. The considered set of faults consists for example of sensor and actuator biases, drifts and noise, which are all explicitly parameterized and defined in the J-dimensional space. As such, faulty observations characterized by any of these fault can only lie within a priori defined subspaces of the J-dimensional space. For this reason, the latter approach is denoted as a *geometrical* approach. Upon detection of a fault, each of the potential faults is evaluated by estimation of its parameter(s). E.g. a sensor bias has one parameter being the magnitude of the bias. The fault which leads to the best of all acceptable fits is identified as the true fault. While conceptually elegant, the geometrical approach is limited to well-defined and well-understood sets of faults and is characterized by special requirements in terms of fault detectability and identifiability. As complete and well-identified sets of faults are not typically available for biotechnological processes, this approach is unlikely to be applicable without any concern. Relaxation of the need on specific knowledge can however be achieved by using other frameworks, such as Evidence Theory as adopted by Lardon et al. (2004) in the context of supervision of an anaerobic digester.

A third approach in diagnosis is based on the paradigm that similar faults will result in similar numerical behaviour of the measured data. Without requirement of knowledge on the exact relationships between root cause and numerical behaviour, a fault can be diagnosed if (numerical) characteristics of the corresponding observation are similar to the (numerical) characteristics of previously identified fault for which the root cause was identified (e.g. through detailed inspection).

Methods that rely on such approaches may be conceptualized in an *supervised* (classification) or *unsupervised* (clustering) fashion. The former requires that a

targeted fault class is identified a priori for each faulty observations in the model calibration data set. This method is applicable if such target is known, i.e. if it is known what the root cause underlying to each faulty observations is. The latter requires no a priori defined target. Similarity measures between the analyzed faulty observations is then used to group or cluster similar observations first. After such clustering, a label, identifying the corresponding fault, needs to be assigned to each cluster by detailed investigation of the behaviour of the corresponding faults in detail. This method is a valid approach if no knowledge on the root cause of faults is known or assumed a priori. Both methods are essentially used to define regions in the J-dimensional space that correspond well with recorded historical faults. As such, they are limited to faults that already have occurred in the past. In view of appropriate identification of these regions, one would paradoxically aim at the presence of as many potential types and repetitions of faults as possible in the historical data set so as to obtain good generalization properties of the obtained classification and/or clustering models. Needless to say, this is incompatible with the ultimate target of process monitoring, diagnosis and control, i.e. improved process performance. Nevertheless, the latter approach, supervised or unsupervised, is valid for supervision of biological systems as both the complexity of contribution plots and the limitations of geometrical approaches are avoided. Automated classification strategies are found for supervision of a pilot-scale SBR by Ruiz et al. (2004) and for anaerobic digestors by Steyer et al. (1997) and Steyer and Harmand (2003).

In this chapter, the unsupervised Fuzzy C-means Clustering (FCM) clustering approach will be evaluated in combination with batch-wise unfolded Multi-way Principal Components Analysis (MPCA) modelling in view of automated diagnosis. The methods used in this chapter are explained in Sections 3.3.1 (PCA), 3.3.3.1 (MPCA) and 3.4 (FCM). In analogy to the monitoring approach in Chapter 5, batches need to be complete to perform the diagnosis task. Given that the same problem may continue over several batches if not taken care of, such diagnosis is effectively still an interesting target.

Before going on, it is noted that even though an unsupervised clustering approach is used, an actual diagnosis problem is always a (supervised) classification problem. This methodological choice was motivated in the original project by the fact that the true fault classes in the then to be generated data set are not available a priori. Indeed, an intensive data screening as pursued in Chapter 4 was not foreseen at the time project inception. To encompass this problem, the unsupervised PCA-FCM approach was proposed to firstly identify the different faults present in the histori-

cal data set. To do so, each cluster would expectedly correspond to a (single) fault type by assigning the most frequent fault type within a cluster as a label to that cluster. For reasons of simplicity, the envisioned system which integrates monitoring, diagnosis and control techniques would use the same PCA-FCM model for classification of new faulty observations. Identifying the cluster to which a new faulty observation most likely belongs then automatically leads to fault identification. Given that in Chapter 4 an intensive data screening was pursued and the fault classes are thus available, the unsupervised approach may be considered void and a supervised classification approach may be considered as more appropriate. Yet, the identification of the fault classes for faulty observation now effectively allows to evaluate what the potential of the unsupervised approach really is. Without these identified classes, no reference is available which would make the evaluation of the method subjective. As such, the unsupervised PCA-FCM approach is still pursued.

In what follows, results for diagnosis of the hydraulic parts of the system are presented first. Then, complete system diagnosis is pursued. Finally, discussion and conclusions are given.

6.2 Diagnosis of hydraulics

In this section, the use of Multi-way Principal Component Analysis and fuzzy clustering for diagnosis is evaluated for the studied SBR system. Firstly, it is shown that MPCA allows explorative analysis of historical faults. In a following section, fuzzy clustering of MPCA scores is evaluated for automated fault classification. The approach is tested for the previously identified mode 1 only (see Chapter 4).

6.2.1 Explorative fault analysis by means of MPCA

Plots of the principal scores are reported to be helpful in the analysis of faults. Real-life applications can be found in (Nomikos and MacGregor, 1994; Kourti and MacGregor, 1995; Nomikos and MacGregor, 1995; Rosén and Lennox, 2001; Sarolta and Kinley, 2001; Aguado et al., 2005). For illustration, the 2 principal scores for all faulty observations in mode 1 (see Section 4.2) detected by the corresponding MPCA model based on mean centered and autoscaled data (mode 1, AS-PCA, optimality 5, see Section 5.3.1) are plotted in Figure 6.1. Call this MPCA model the common-cause variation MPCA model, denoted MPCA_C. Observations of each fault class are given the same color. It can be seen that for some classes the observations are clearly grouped together in the given plot. For instance, observations in fault class 1 are exclusively grouped in the upper-left corner of the plot. Ellipsoid A indicates this area. Within this ellipsoidal area, two major groups of dots can be identified. Each of the groups corresponds to a distinct level of the weight measurement. The first group (lower-left within the ellipsoid) represents batches for which the weight remained at its lowest level as the influent tube was disconnected. The second group (upper-right with the ellipsoid) represents batches for which the pump itself started to fail during an effluent withdrawal phase, thereby resulting in a weight level higher than the minimum level. During the data screening no attention was given to level of the weight so that the two groups were not separated as such. The ellipsoid contains a single observation which was assigned to fault class 3 during data screening. It was verified that the corresponding observation does not belong to fault class 1 and is correctly assigned to fault class 2, i.e. the hydraulic system did not fail completely during the corresponding batch. However, the added volume of influent in this batch is so low that the profile of the weight data series of this observation becomes –numerically speaking– similar to the profiles of observations corresponding to fault class 1 (complete hydraulic failure).

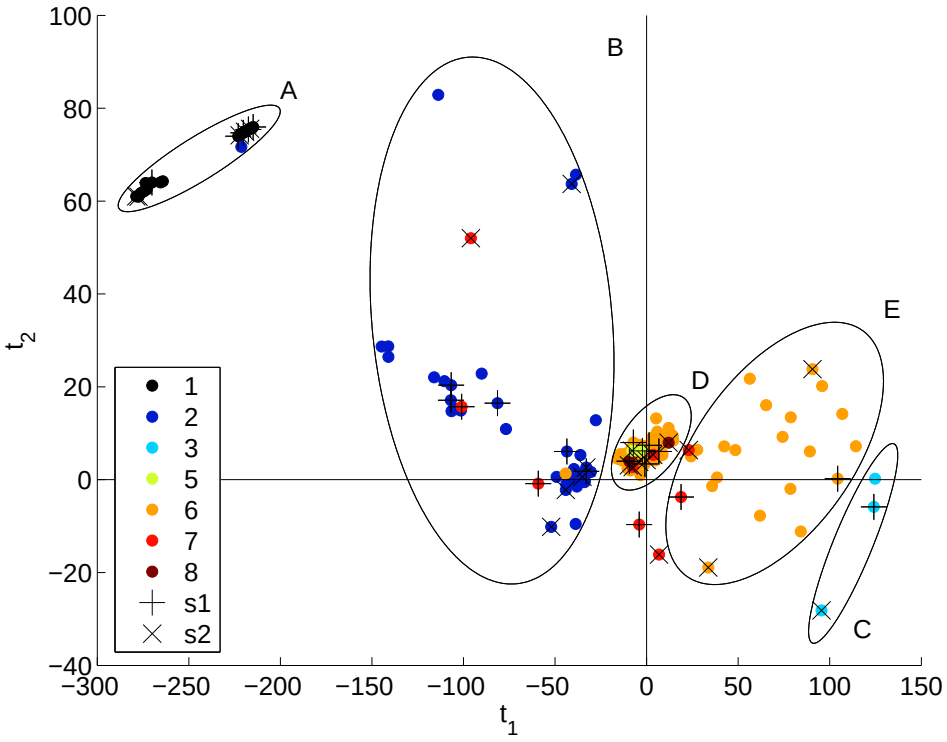


Figure 6.1: Biplot of first and second PC score for detected abnormal SBR cycles by projection on the $MPCA_C$ model. Numbers indicate fault classes assessed prior to MPCA analysis, characters indicate groupings of faults posterior to MPCA analysis. Subsets used for identification of $MPCA_R$ models (s1 and s2) are indicated by crosshairs.

Other observations belonging to fault class 2 span a large area indicated by ellipsoid B. It can be seen that all of the observations of fault class 1 and 2 have negative values for the first principal score. Observations for fault class 3 can be found in the lower-right corner of the graph. Given that this fault class consists of batches with excessive addition of influent, they can be regarded as antagonistic to the observations of fault classes 1 and 2. The shown biplot seems to confirm this antagonism given that the dots for fault class 3 exhibit positive values for the first score. The larger part of the batches assigned to fault class 6 –exhibiting insufficient effluent withdrawal– can be found in a relatively small area enclosed by ellipsoid D. A smaller set of observations of this fault class is grouped into a larger area enclosed by ellipsoid E. Revision of the batches in fault class 6 revealed that

the batches corresponding to the dots within ellipsoid D are the result of the same problem which persisted over a set of consecutive batches albeit with smaller magnitude. Those dots enclosed by ellipsoid E relate to a set of batches with the same problem, though larger in magnitude and more dispersed both in terms of time and magnitude. Ellipsoid D also encloses the (2) observations in fault class 8, i.e. the two batches in this mode with the multiple faults. Each of the two observations were confirmed to exhibit incomplete effluent withdrawal (typical for fault class 6) in addition to noisy artefacts (typical for fault class 7). The observations of other types of faults are spread out over the biplot.

Since the plotted principal scores in the former biplot are the scores corresponding to the principal components in the MPCA model for normal data, not all information contained within the data of the faulty observations may be transformed into the PC's of the constructed MPCA model. Indeed, the given MPCA model was calibrated and thus optimized to capture a maximal amount of variance of the *normal* data in a reduced set of scores. This does not necessary lead to the capturing of a maximal amount of variance and/or information from the *faulty* data.

One way to facilitate explorative analysis of faulty observations is to analyze the residuals that arise from projection on the MPCA model by means of a new MPCA model. The new MPCA model based on the residuals is denoted as the residual MPCA model, denoted MPCA_R. Since the MPCA_R model is based on the residuals for which variance captured by the MPCA_C model is removed, the resulting PC's in the MPCA_R model are by definition orthogonal to the PC's in the MPCA_C model. Note that the resulting scores are not necessarily uncorrelated to those of the MPCA_C model. To evaluate the proposed strategy, two PC's were calculated on the residuals of the faulty observations. In order to reduce the impact of one fault class on the model over the other, maximally 5 batches were taken from each fault class with 10 or more members. For fault classes with limited numbers of members ($<$10), no more than half of the observations were used for the model. The selection of the latter batches for MPCA_R model calibration was done at random and performed twice without repetitions (i.e. no batch is part of the two sets). In Figure 6.1, the respective samples taken for this calibration step are indicated by crosshairs. For each of the two sets of faulty observations an MPCA_R model was constructed (AS-PCA). Figure 6.2 and 6.3 give the respective biplots of the first and second score for these 2 subsets.

First of all, it can be seen that the biplots are very similar. This indicates that the models made on the basis of the two sets of selected batches are roughly the same

and therefore suggests that consistent information is contained in the scores for the respective PC's. In both graphs, ellipsoids are drawn to indicate meaningful regions. The respective characters assigned to the ellipsoids are the same for ellipoids indicating the same meaningful area in the plot. Ellipsoid F indicates an area in which uniquely faults from class 1 lie. As such, a more clear visual discrimination between the faults of class 1 and 2 is achieved. Also, faults of fault class 2 and 3 can be enclosed by a single ellipsoid. Interestingly, an ellipsoid, I, can be drawn around the (2) dots corresponding to fault class 5. This was not possible on the basis of the former biplot (of the scores of the common-cause variation MPCA model). The indication of an area in the biplot specifically enclosing the

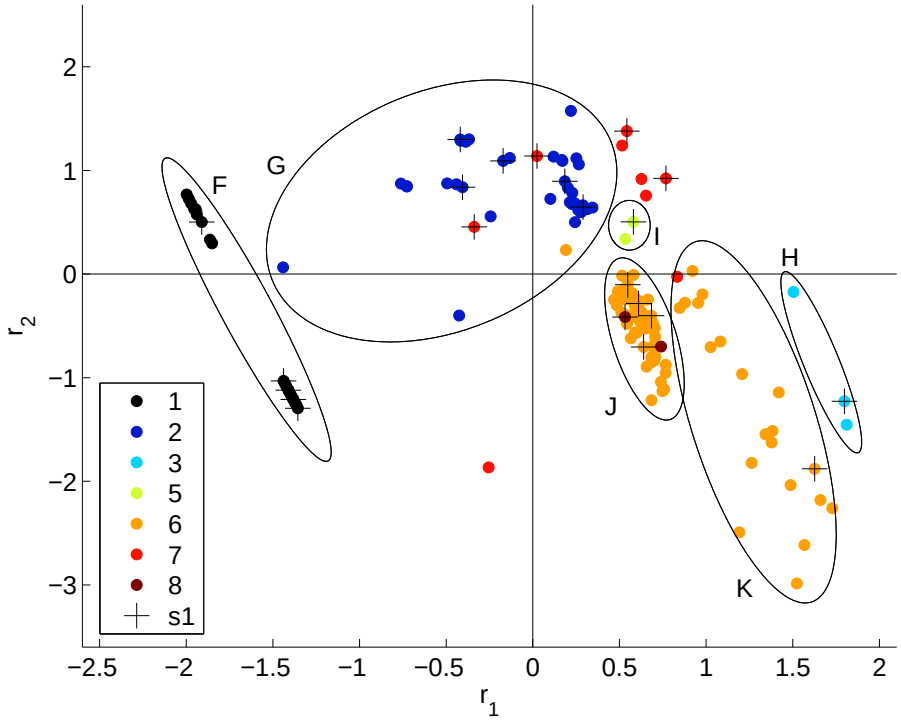


Figure 6.2: PC scores of first and second PC score for detected abnormal SBR cycles by projection onto an $MPCA_R$ model calibrated on a subset (s1) of the set of detected abnormal cycles. Numbers indicate fault classes assessed prior to MPCA analysis, characters indicate groupings of faults posterior to MPCA analysis. Crosshairs indicate the subset used for calibration.

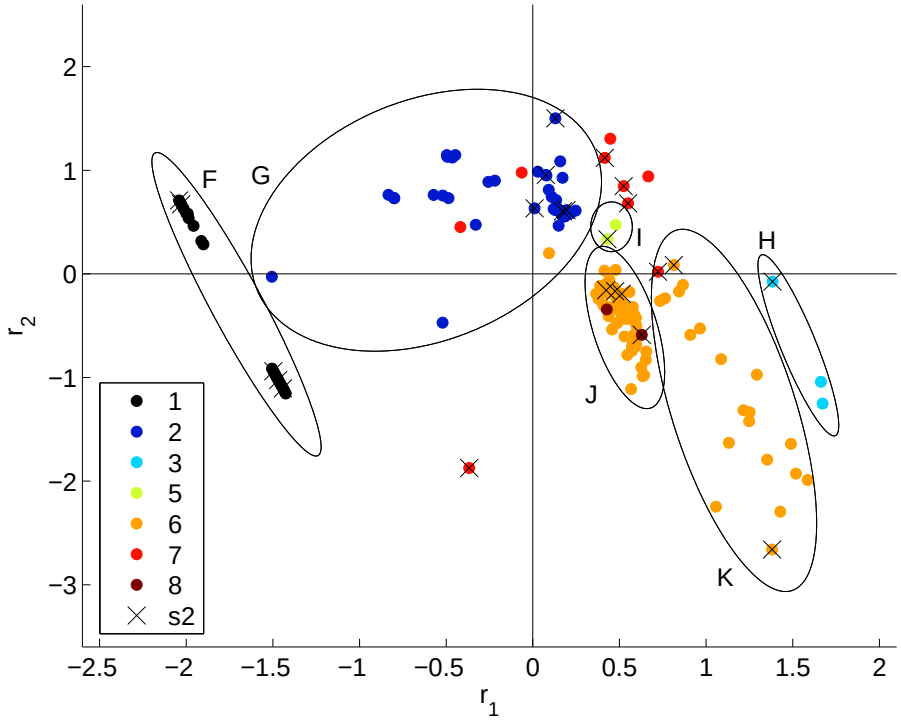


Figure 6.3: PC scores of first and second PC score for detected abnormal SBR cycles by projection onto an $MPCA_R$ model calibrated on a subset (s2) of the set of detected abnormal cycles. Numbers indicate fault classes assessed prior to MPCA analysis, characters indicate groupings of faults posterior to MPCA analysis. Crosshairs indicate the subset used for calibration.

dots corresponding to fault class 7 cannot be drawn in the biplot. Given that this class comprises a set of faults relating to data quality rather than a consistent process fault, this may not be a surprise. As the positioning of the plotted ellipsoids is roughly the same in both graphs it can be concluded that the calibration of the MPCA model to the given residuals gives rise to the capturing of consistent information. It was verified that the dots enclosed by ellipsoid D, resp. E, in figure 6.1 correspond to those enclosed by ellipsoids J, resp. K, in Figures 6.2 and 6.3. Therefore, the two distinct groups within fault class 6 are visually separable as well.

Given the observations it can be stated that:

- Explorative analysis of scores by means of visual inspection, obtained by projection onto the $MPCA_C$ model, to a large extent confirms results of data screening.
- $MPCA_R$ models, constructed on the $MPCA_C$ residuals, confirm results of data screening to a large extent as well and leads to the visual separation of fault classes which were not separated as yet on the basis of the $MPCA_C$ model.
- Both explorative analysis by means of $MPCA_C$ and $MPCA_R$ models can confirm, stress or reveal relevant differences in numerical behaviour of observations grouped into different classes.

6.2.2 Fault diagnosis by means of MPCA-based clustering

In contrast to the depicted situation discussed above, a set of faulty data may not be screened as rigorously as done for this study. Clearly, a plot of the given scores can then indicate separate types of faults present in the given set of faults. However, explorative analysis is not limited to the given biplots. Indeed, other biplots can be made by plotting $MPCA_R$ model scores against $MPCA_C$ model scores. Given the orthogonality, no special adjustments are required. For the given setting with 4 scores in total the amount of unique biplots that can be made amounts to 6. Unfortunately, this makes an explorative analysis increasingly difficult, time-consuming and cumbersome. In addition, the true multi-dimensional behaviour of the observations can seldomly be mentally depicted to its full extent by means of such biplots. While 3-dimensional plots are available, they are often difficult to interprete in the author's experience, strictly limited to three dimensions and in most practical situation limited by the effective 2-D nature of a PC screen or a physical paper. Moreover, an explorative analysis is not limited to the given number of PC's for either the $MPCA_C$ and $MPCA_R$ model. It is in view of this multi-dimensional complexity that clustering techniques can become useful tools to group data into similar observations, where similarity can be defined irrespective of the number of dimensions. As such, the limitations of conventional visualization and human interpretation abilities may be overcome.

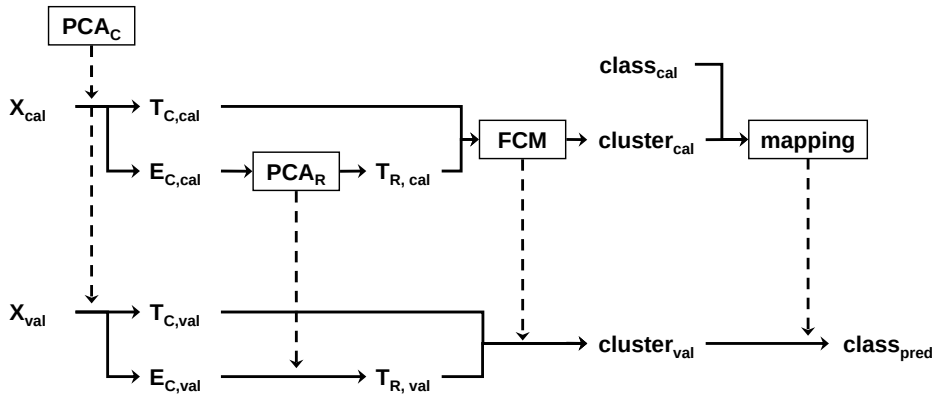


Figure 6.4: Schematic diagram for one iteration in the cross-validation procedure for identification of a classification model based on MPCA and fuzzy C-means clustering (FCM).

To evaluate the possibilities of clustering tools for diagnosis, the following strategy was tested. Each faulty observation is taken out of the data set once for validation while the other points are used for calibration (leave-one-out cross-validation). The steps that are taken for each validation are shown in Figure 6.4. First, all observations (calibration and validation sets, i.e. $\mathbf{X}_{cal}$ and $\mathbf{X}_{val}$), are projected on the already available MPCA_C model (calibrated earlier on normal data in view process monitoring) to obtain the respective scores ($\mathbf{T}_{C,cal}$ and $\mathbf{T}_{C,val}$) and residuals, ($\mathbf{E}_{C,cal}$ and $\mathbf{E}_{C,val}$). The residuals of the calibration set, $\mathbf{E}_{cal}$, are consequently used for calibration of an MPCA_R model and FCM model, resulting in corresponding scores, $\mathbf{T}_{R,cal}$ and cluster identities to which the observations are assigned, $\mathbf{cluster}_{cal}$. The observations are assigned to the (single) cluster which exhibits the maximal fuzzy membership value for the given observations (see Section 3.4 for details).

A mapping of clusters to fault classes is established as follows. For each combination of available clusters and classes, the sensitivity, defined as the fraction of calibration observations of a given fault class that are assigned to the given cluster (# observations in the given class and cluster / # observations in the given class), and the specificity, defined as the fraction of calibration observations assigned to the given cluster that belong to a given fault class (# observations in the given class

and cluster / # observations in the given cluster), are calculated for all fault classes. Now, a mapping of clusters to classes is made by linkage of a given cluster to the fault class for which maximal product of sensitivity and specificity is obtained. As such, the given cluster model is converted into a classification model by simple linkage.

Now that the complete model structure is obtained for the given calibration data set, the validation residuals, $\mathbf{E}_{C,val}$, are projected onto the MPCA_R model so to obtain the corresponding scores, $\mathbf{T}_{R,val}$. MPCA_C and MPCA_R model scores are used to project the observation onto the FCM model, hereby leading to the identified cluster, $cluster_{val}$. By means of the mapping defined earlier, the predicted class for the validation is obtained.

The given procedure is repeated for all observations so that a predicted class is obtained for all of them. Now, performance of the evaluated model is evaluated as follows. First, sensitivity and specificity for the overall model are calculated. Practically, for a given fault class, the sensitivity is the fraction of the observations truly belonging to the class that are correctly classified to that class ($N_{correct,c}/N_{class,c}$ = # observations correctly assigned to the given class, c / # observations truly belonging to the given class, c). The specificity is the fraction of the observations assigned to a given class that truly belong to that class ($N_{correct,c}/N_{assign,c}$ = # observations correctly assigned to the given class, c / # observations assigned to the given class, c).

To compute confidence intervals for the sensitivities and specificities, the binomial model is assumed as was done in Chapter 5 for the reported misclassification rates. It is repeated here that, given a random process by which a two-level univariate outcome (e.g. 0/1) is assigned to each discrete run or trial and assuming a binomial distribution, the probability, Y, that at least N_1 out of N outcomes are '1' is given by the following cumulative distribution function (cdf):

$$Y = F\left(N_1|N,\beta\right) = \sum_{n_1=0}^{N_1} \binom{N}{n_1} \cdot \beta^{n_1} \cdot (1-\beta)^{N-n_1} \tag{6.1}$$

where:

$$\beta\text{: (constant) probability for outcome '1'} \tag{6.2}$$

With respect to the confidence limits for sensitivity, $N_{correct,c}$ is the number of correctly classified observations for a given class, c, and $N_{class,c}$ the number of observations in this class. Then, $N_1/N = N_{correct,c}/N_{class,c}$, the sensitivity, is the estimate of β. To compute confidence levels for the observed error rate $N_{correct,c}/N_{class,c}$, 95% confidence levels are computed according to equation 6.1 by solving the equation to N_1 for $Y = 0.025$ (left-hand side confidence limit) and $Y = 0.975$ (right-hand side confidence limit) after replacement of β with its estimate $N_{correct,c}/N_{class,c}$. Similarly, confidence limits for the specificity are computed by estimating β in equation 6.1 by $N_1/N = N_{correct,c}/N_{pred,c}$, with $N_{pred,c}$ the number of observations predicted to be in the corresponding fault class, c. Again, it is noted that the binomial distribution model requires that the trials are independent and that the probability for a certain outcome is constant over all trials. Neither is expected to be true for the studied data as many of the faulty observations are the result of a single fault lasting over several batches resulting in the violation of the assumption on independency. Also, as discussed already in Chapter 5, the probability of the outcome may be dependent on the magnitude of the fault, thus not constant.

During the model steps, the number of PC's in the MPCA_R model and the number of clusters in the MPCA_C model need to be specified. In order to identify the optimal number of PC's and clusters, the complete procedure is repeated for all viable combinations of respective choices. In this case, the number of PC's for the residual variation MPCA model ranged from 1 to 20 and the number of clusters demanded to the clustering algorithm ranged from 2 to 20. The number of PC's for the MPCA_C model was kept to 2, as in the monitoring model that led to the detection of the set of analyzed faults. In what follows, the results (overall sensitivity and specificity) are discussed in detail only for fault class 1. Thereafter, additional results will be discussed in a more general fashion.

The sensitivity for fault class 1 is shown in Figure 6.5 for all evaluated combinations of MPCA_R and FCM models. As can be seen, the sensitivity is 100% for most of the evaluated model structures. This means that for these combinations all observations within this fault class are correctly assigned to fault class 1. In other words, batches truly belonging to fault class 1 are correctly assigned to fault class 1 by means of the evaluated MPCA-FCM combination. This is true for any combination with up to 13 PC's and 12 clusters. For 2 clusters and 19 or 20 PC's, none of the batches in fault class 1 is correctly classified. Excluding the latter two models, a general decrease in sensitivity is observed for more complex model structures, i.e. for more PC's and for more clusters included in the models. Confidence intervals

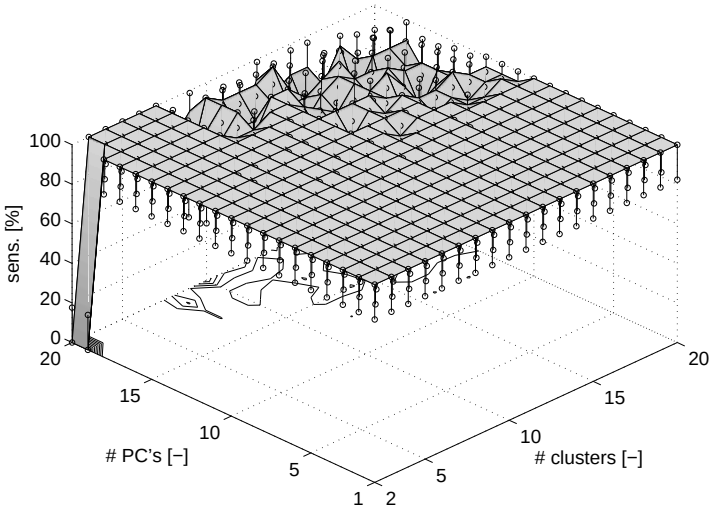


Figure 6.5: Sensitivity for fault class 1 as a function of the number of clusters and the number of PC's for the $MPCA_R$ model. The surface indicates the observed sensitivities. Dots indicate the estimated 95% confidence interval.

computed on the basis of the binomial model are relatively narrow which is largely due to the relatively high number of observations in fault class 1.

The specificity for fault class 1 is given in Figure 6.6 as a function of the number of PC's for the $MPCA_R$ model and the number of clusters for the FCM model. The specificity shown here is thus the fraction of the batches assigned to fault class 1 that are truly belonging to fault class 1. The specificity is maximally 100% and is reached for a set of models with PC's ranging between 13 and 16 and the number of clusters ranging between 8 and 18. A much larger set of evaluated choices leads to a 95% specificity. The latter value is reached already for a much simpler model, i.e. with 5 clusters and 2 PC's in the $MPCA_R$ model. It was verified that, when the 95% rate is achieved, this is due to the misclassification of the single observation of fault class 2 that was earlier observed to be visually close to those of fault class 1 (see Section 6.2.1). Only for high-dimensional models (high number of PC's, high number of clusters), this observation can be classified correctly (to fault class 2). Also here, confidence intervals for the specificity are relatively narrow, which can be explained by relatively large numbers of observations assigned to fault class 1. Both sensitivity and specificity are high for low-dimensional models. This indi-

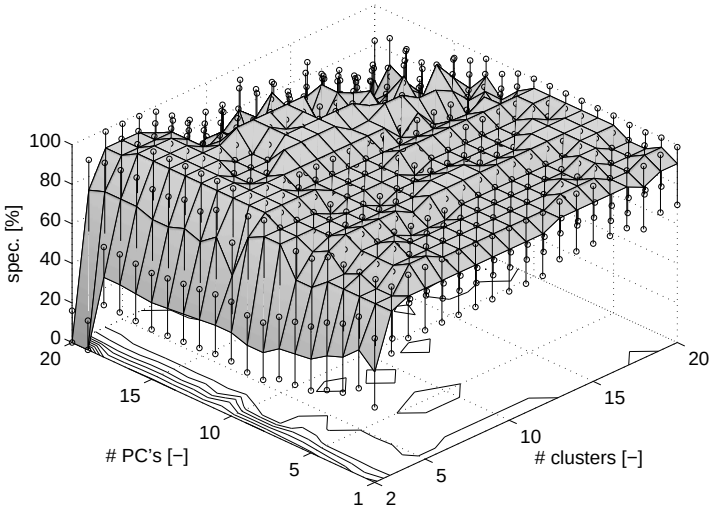


Figure 6.6: Specificity for fault class 1 as a function of the number of clusters and the number of PC's for the MPCA_R model. The surface indicates the observed sensitivities. Dots indicate the estimated 95% confidence interval.

cates that (1) a batch belonging to fault class 1 is likely to correctly be assigned to fault class 1 (high sensitivity) and that (2) a batch assigned to fault class 1 is likely to truly belong to fault class 1 (high specificity).

Plots of the sensitivity and specificity for all fault classes are given in Appendix A. The results shown already for fault class 1 are among the best obtained. High performance rates with narrow confidence intervals were also obtained for fault classes 2 and 6. Fault classes 1, 2 and 6 are thus relatively easily to separate. Generally speaking, simultaneously high values for the number of PC's and the number of clusters leads to poorer results for the latter fault classes, indicating overfit of the resulting models for these model choices. Low performances are obtained for fault classes 3, 5, 7 and 8. Also, the computed confidence intervals are very wide, indicating large uncertainty on what the expected performance for these classes is. For fault class 3, only high-dimensional models result in high sensitivity and specificity. Additional conclusions cannot be made for fault classes 3 and 5. The low performance for fault class 7 can easily be linked to the fact that the data are characterized by noisy artefacts, resulting in inconsistent numerical behaviour within this fault class. For fault class 8, the two included observations,

which exhibit a combination of faults typical for fault class 6 and 7, wre generally not separated from other faults. It was verified that for models with up to 15 PC's and up to 12 clusters, both batches in fault class 8 were assigned to fault class 6. This confirms that the constructed models are unable to appropriately characterize fault class 8, while the consistent behaviour within fault class 6 can be addressed properly.

Based on the results presented so far, it is clear that a fully automated diagnosis which allows the correct classification of all fault classes on the basis of the proposed combination of MPCA and fuzzy clustering is not possible. Nevertheless, a final model selection approach is proposed here that maximizes the classification performance for the fault classes for which good results were obtained. To do so, fault classes 1, 2 and 6 are included in this model selection. For other fault classes, poor results were obtained. The sensitivities and specificities for fault classes 1, 2 and 6 are averaged so to obtain an average sensitivity and specificity for these fault classes. An overall performance rate is obtained by computing the average of the obtained (averaged) sensitivity and specificity. Results are shown in Figure 6.7. Highest values for the latter performance index are generally found for models with the number of PC's ranging between 1 and 10 and for models with the

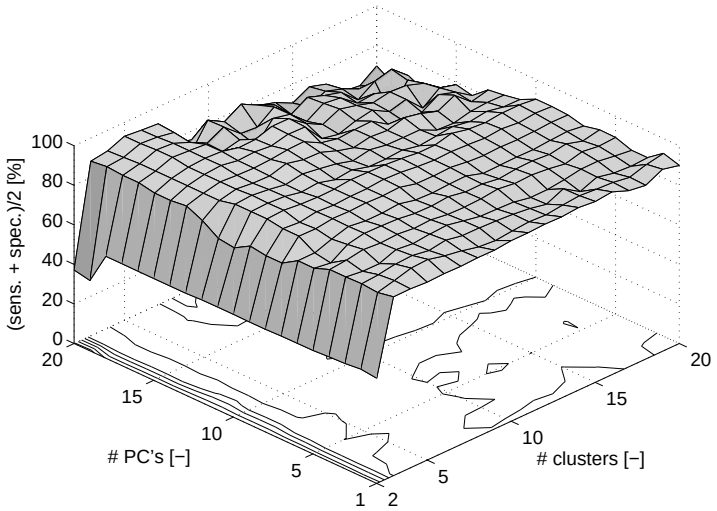


Figure 6.7: Average of sensitivity and specificity, both averaged over classes 1, 2 and 6.

number of clusters ranging between 3 and 7. High values for both the number of PC's and clusters lead to lower performance, indicating overfitting of the included models. The best performing model in terms of weighted sensitivity and specificity is the one with 2 $MPCA_R$-PC's and 6 clusters and may therefore selected for on-line application. However, given that the selected model is optimized for a selected number of fault classes only, automation of diagnosis and consequent triggers for control action is not regarded as a good strategy. An indicative use of the model outcomes during human-interfered diagnosis is suggested, parallel with data exploration tools as shown before.

6.3 Multi-sensor diagnosis

The strategies already applied for diagnosis of the hydraulic parts of the system are evaluated here for the complete system. Data of mode 2a are used exclusively (see Chapter 4). Both visual exploration and classification by means of MPCA and fuzzy clustering are evaluated.

6.3.1 Explorative fault analysis by means of MPCA

The first two scores for the abnormal data detected by means of the selected 4-PC AS-PCA model (see Section 5.4) are illustrated in Figure 6.8. In this figure, a distinct area is indicated by means of ellipsoid A. This ellipsoid groups the major part of the batches in fault class 1. In contrast to the results shown based on hydraulics only (see Figure 6.1), the spread of the batches in fault class 1 is much larger in this case. Ten of the corresponding dots represent a consecutive set of batches and are connected by lines. Starting at the most left and downward point, the plotted dots move away from the center, batch after batch. This indicates that, while the fault event itself (zero influent flow rate) and corresponding hydraulic data do not change largely in nature during this event, the other data seemingly do as a result of the fault.

Ellipsoid B groups the data corresponding to fault classes 18, 24 and 29. All of the batches included in these classes are characterized by the same calibration error, resulting from a single event. In contrast to the former event, no clear evolution of the scores can be discerned in the biplot. Given that the pH signal itself is not part of active control loops and given that the error-free pH value is not affected, this is to be expected.

In Figure 6.9, a more detailed view is given of the area in the discussed biplot which contains the larger part of the plotted samples. In this plot, ellipsoid C indicates an area dominated by batches in fault class 9, characterized by cooler failures. Ellipsoid D indicates an area dominated with problems of fault class 13, characterized by minimal oxygen consumption, leading to high oxygen levels in (supposedly) anaerobic and anoxic phases.

The indication of distinct areas corresponding to the fault class is more difficult. For example, for fault class 2 a group of included observations is found in the upper

right quadrant ($t_1 \sim 50$, $t_2 \sim 12$) and another group is concentrated in the down right quadrant ($t_1 \sim 25$, $t_2 \sim -10$). Detailed investigation (not shown) revealed that the these two groups of observations are located distantly in time. Given that the scores of the normal data were shown to vary in a non-random fashion (see Section 5.4), the same effect plays a role here. Changed process conditions result in a different location in the biplot of the same fault type. Not unlikely, the same effect will hold for other fault classes, leading to a rather scattered impression of the shown observations. As a result, only faults that affect the process performance to an extreme extent can be discriminated well in this plot.

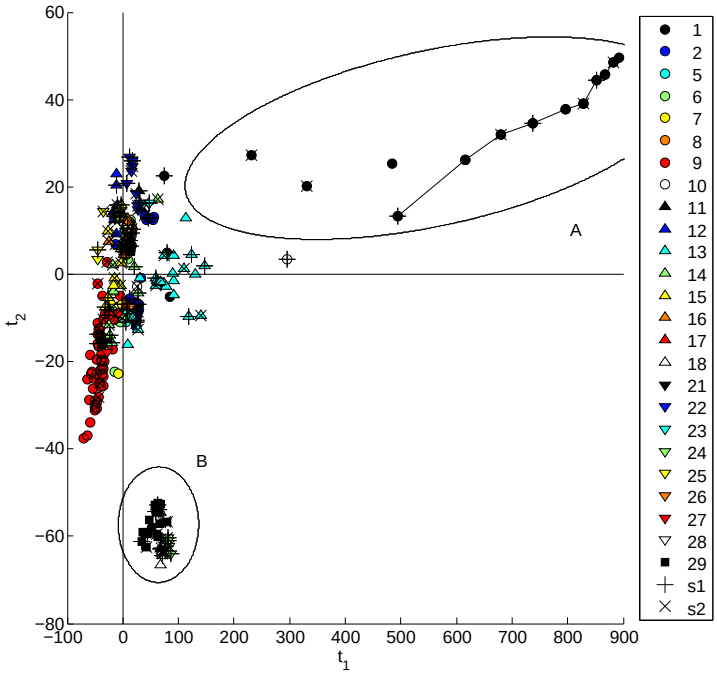


Figure 6.8: Biplot of first and second PC score for detected abnormal SBR cycles by projection on the common-cause variation MPCA model (complete). Numbers indicate fault classes assessed prior to MPCA analysis, characters indicate groupings of faults posterior to MPCA analysis. Subsets used for residual variation MPCA models (s1 and s2) are indicated by crosshairs.

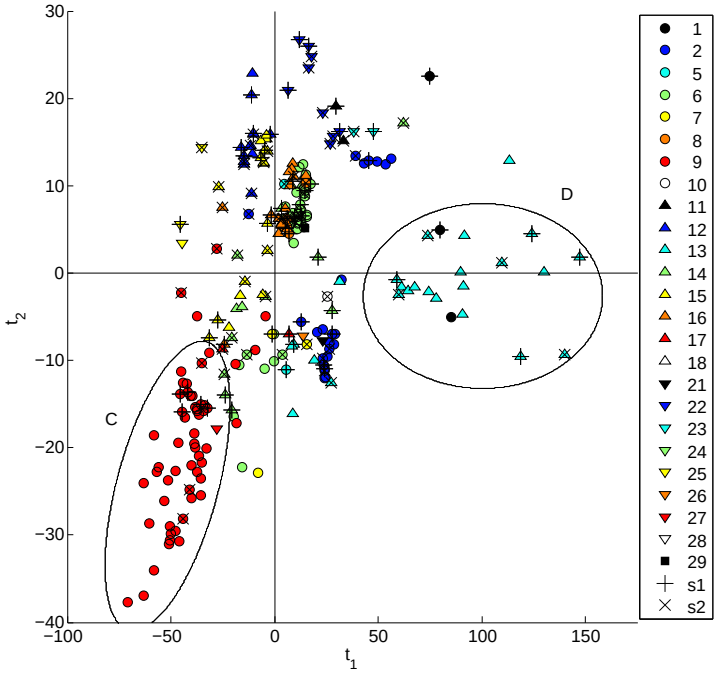


Figure 6.9: Biplot of first and second PC score for detected abnormal SBR cycles by projection on the common-cause variation MPCA model (detail). Numbers indicate fault classes assessed prior to MPCA analysis, characters indicate groupings of faults posterior to MPCA analysis. Subsets used for residual variation MPCA models (s1 and s2) are indicated by crosshairs.

The PC's contained in the common-cause variation MPCA model, denoted further as MPCA_C, may not be sufficient to explain all the variation in the data of the faulty observations. It was indeed indicated in the previous section, in which attention was focused on the hydraulic system, that the MPCA_C is based on normal data only. As a result, additional information is expected to be available in the residuals resulting from projection onto the MPCA_C model. It was suggested before that the residuals may be analyzed further by means of MPCA analysis, resulting in an MPCA_R model. To evaluate this for the complete system, two PC's were calculated on the residuals of the faulty observations. In order to reduce the impact of one fault class on the model over the other, at most 5 batches were taken from

each fault class with 10 or more members. For fault classes with limited numbers of members (2-10), no more than half of the observations were used for the model. When only 1 observation was available it was not included in the model calibration set. The batches for $MPCA_R$ model calibration were selected randomly and without repetitions, as before. The sampling procedure was performed twice, not using the same observations in the two sampled sets, so that two models could be constructed. The respective batches included in the two calibration sets are indicated in Figures 6.1 by means of corresponding crosshairs. Figure 6.10 and 6.11 give the respective biplots of the first and second score for the two models. The two plots are similar in the sense that the relative positions of the members behave similarly.

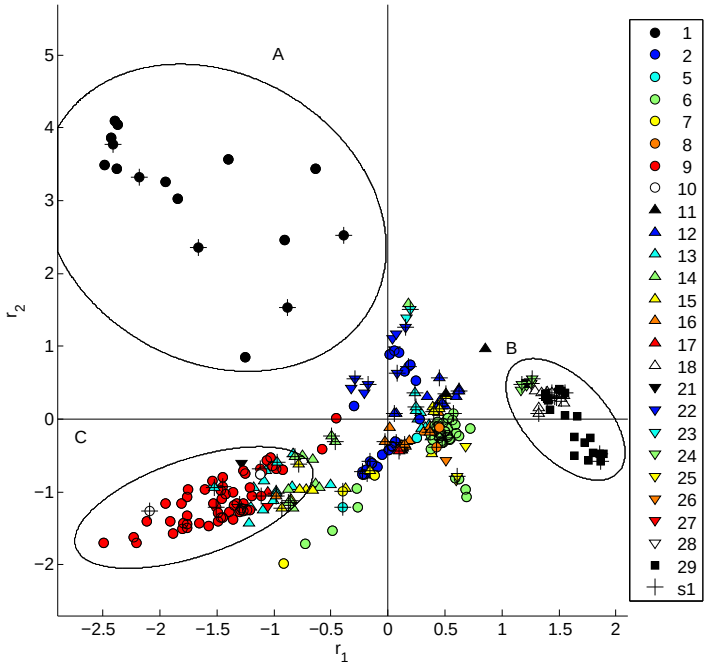


Figure 6.10: PC scores of first and second PC score for detected abnormal SBR cycles by projection onto an $MPCA_R$ model calibrated on a subset (s1) of the set of detected abnormal cycles. Numbers indicate fault classes assessed prior to MPCA analysis, characters indicate groupings of faults posterior to MPCA analysis. Crosshairs indicate the subset used for calibration.

The plots as a whole seem to be rotated versions of each other, indicating that the resulting $MPCA_R$ model itself is affected to a large extent by the sampling of the calibration data. In both biplots, distinct areas for fault class 1 (ellipsoid A), fault classes 18, 24 and 29 (ellipsoid B) and fault class 9 (ellipsoid C) can be discerned. Given that this is relatively easy to do, it stands affirmed that the $MPCA_C$ model cannot capture the behaviour of these faults to their full extent. Other than for the reported classes, straightforward interpretation of the given plots is difficult. While changing properties of the system under normal conditions may be underlying this, it is important to consider that the impact of the data belonging to fault classes 1, 9, 18, 24 and 29 is much larger on the identified $MPCA_R$ model. Indeed, due to the

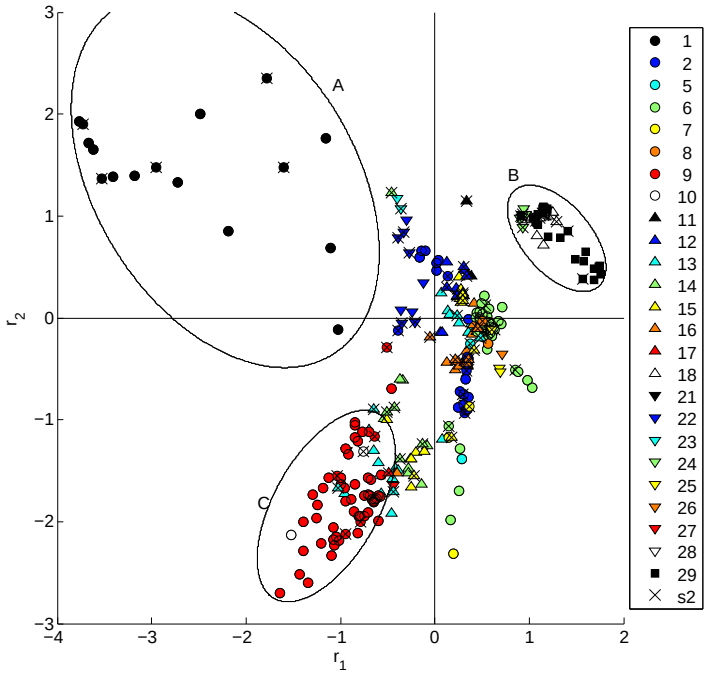


Figure 6.11: PC scores of first and second PC score for detected abnormal SBR cycles by projection onto an $MPCA_R$ model calibrated on a subset (s2) of the set of detected abnormal cycles. Numbers indicate fault classes assessed prior to MPCA analysis, characters indicate groupings of faults posterior to MPCA analysis. Crosshairs indicate the subset used for calibration. Detail of Figure 6.10

larger distance from the origin they contribute to a much larger extent to the variance maximized by means of MPCA. It is therefore not surprising that the location of other faulty observations cannot be interpreted straightforwardly.

Given the discussed results it can be stated that:

- Explorative analysis of scores by means of visual inspection, obtained by projection onto the $MPCA_C$ model, confirms results of data screening detecting major faults such as failure of pumps and the reported pH calibration error.
- $MPCA_R$ models, constructed on the $MPCA_C$ residuals, do not lead to the discovery of substantial or previously unknown information, in contrast to the results for the hydraulic system reported in Section 6.2.1.
- The MPCA models for explorative analysis have been indicated to suffer from the (invalid) assumption on a constant mean and covariance structure.
- $MPCA_R$ models are influenced most by the presence of large and frequent faults, hereby blurring the diagnosis for less obvious or less frequent faults.

6.3.2 Fault diagnosis by means of MPCA-based clustering

Even if serious limitations were identified during explorative MPCA-based analysis of the multi-sensor data, the formerly proposed approach for automated diagnosis is also evaluated here. In this approach, an $MPCA_R$ model is established for all detected abnormal observations, excluding one observation which is projected onto the model. The already obtained scores for the $MPCA_C$ model and the newly obtained $MPCA_R$ scores are clustered by means of a fuzzy C-means (FCM) clustering. To each cluster a unique fault class is linked by choosing the fault class which maximizes the product of sensitivity and specificity for the given cluster. Twelve fault classes included less than 5 detected batches. Within those, 4 classes exhibited only one detected member and 6 classes had only 2 detected members. As a result, the included batches are likely to be difficult to cluster. In order to still be able to evaluate automated assignment to fault classes the study was restricted to the fault classes exhbiting 5 or more detected members (i.e. fault classes 1, 2, 6, 9, 12, 13, 14, 15, 16, 18, 22, 24 and 29). Results for this excercise are shown here. Confidence limits were computed in the same fashion as for the pursued diagnosis

for the hydraulic parts of the system only (see Section 6.2.2). Again, the results for one fault class are shown after which results for other fault classes are discussed more generally.

In Figure 6.12, the sensitivity for fault class 1 is shown. The observed sensitivity ranges between 50% and 100% for models with PCR_R PC's up to 6 PC's and at least 4 clusters. An increasing trend can be observed for the number of clusters ranging from 4 to 20 (for the number of PC's up to 6). Including more than 6 PC's reduces the sensitivity dramatically for this fault class. More or less in accordance with the sensitivity, the specificity (Figure 6.13) for fault class 1 is generally high for a model with 6 PC's or less. The specificity also drops to minimal levels beyond 6 PC's. The range of evaluated models with more than 6 PC's deliver a mean specificity of 11.3%. Both the sensitivity and specificity are thus largely affected by the number of PC's rather than the number of clusters. The specificity is however lower for models with 6 to 9 clusters when concentrating on the model with up to 6 PC's. The fact that both measures are not influenced much by the number of clusters indicates that this fault class can rather easily be separated from

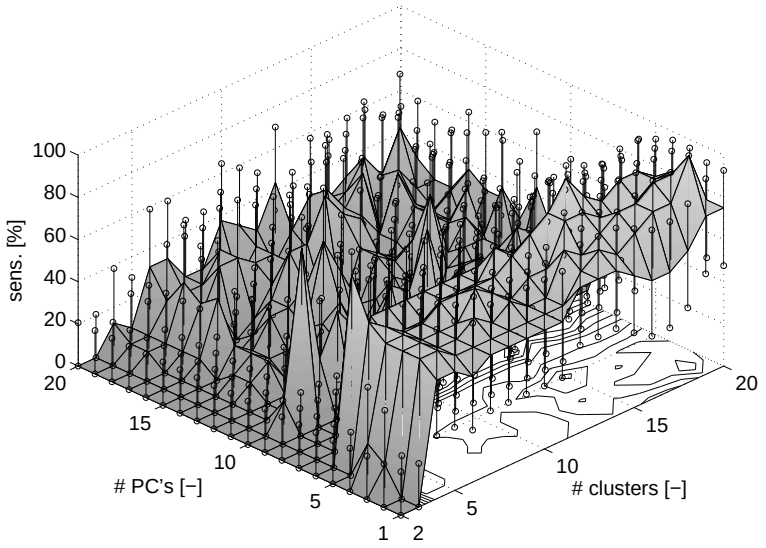


Figure 6.12: Sensitivity for fault class 1 as a function of the number of clusters and the number of PC's for the $MPCA_R$ model. The surface indicates the observed sensitivities. Dots indicate the estimated 95% confidence interval.

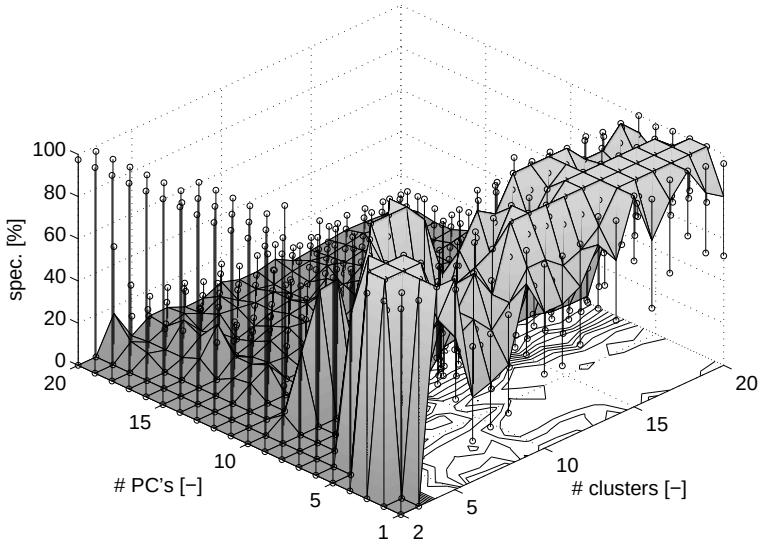


Figure 6.13: Specificity for fault class 1 as a function of the number of clusters and the number of PC's for the $MPCA_R$ model. The surface indicates the observed sensitivities. Dots indicate the estimated 95% confidence interval.

other faults. It is noted here that the sensitivities and specificities reported here are generally lower compared to those reported for the diagnosis concentrating on the hydraulics only. In addition, the confidence intervals are wider as well compared to the results shown for diagnosis of hydraulics only. This suggests that a separate model to diagnose problems in the hydraulic system is a valid approach.

Plots of sensitivity and specificity are given for all classes in Appendix B. Results are generally poorer compared with results for the diagnosis excercise focusing on the hydraulic system only. For low-dimensional models (#PC's$\leq$10, #clusters$\leq$5), high rates for both sensitivity and specificity ($\geq$70%) are recorded for fault class 6 and 13 only. Fault classes 14, 22 and 24 lead to poor results over the whole range of evaluated models. For other fault classes, reasonable rates for both sensitivity and specificity are not met or require the use of large-dimensional models.

Given the reported sensitivities and specificities, a fully automated diagnosis which allows the correct classification of all considered fault classes on the basis of the proposed combination of MPCA and fuzzy clustering is not possible in this case

either. However, a diagnosis model is selected here again so to obtain an indicative non-automated classification model. Just as above, the model is optimized in view of a selection of classes. Included classes are all classes considered so far except 14, 22 and 24 for which results were considered poor. The same procedure was used as for the diagnosis model for the hydraulic system. An average sensitivity and specificity was obtained for the considered fault classes, which were then averaged as well. Figure 6.14 shows the resulting average of sensitivity and specificity. As may be expected from the discussed results, a low number of PC's (1-6) and a high number of clusters ($>$15) delivers the highest performance. It needs however to be considered that (1) the performance is low still and (2) the reported models may be unnecessary complex and will likely overfit the data. With respect to this, it is concluded that the proposed approach does not meet desired qualities for automated diagnosis. Potential improvements are suggested further on in this text.

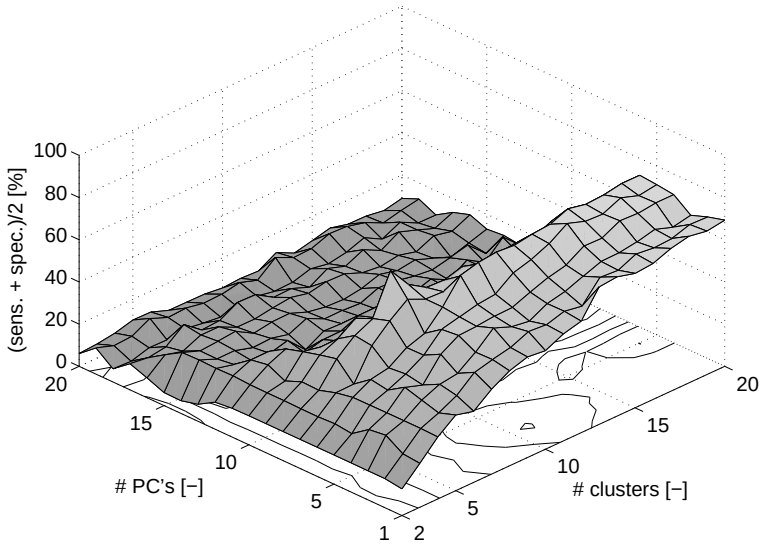


Figure 6.14: Average of sensitivity and specificity, both averaged over all classes except classes 14, 22 and 24.

6.4 Discussion

In the former paragraphs, MPCA was demonstrated to allow data exploration in view of process diagnosis of the studied SBR. To this end, a common-cause variation MPCA model, $MPCA_C$, calibrated on normal data, was used. It was also illustrated that MPCA-based analysis of residuals, delivering so called $MPCA_R$ models, reveals or at least confirms information about process faults. In an attempt to automate the diagnosis process, a joint technique based on MPCA and fuzzy clustering was proposed and tested. It was shown that for certain fault classes, this can deliver good results in terms of sensitivity and specificity. However, faults in many other fault classes cannot be discriminated well. Therefore, critical improvements of the method should be pursued or alternative methods should be evaluated prior to attempt fully automated application.

In view of improved computation of confidence intervals for the observed sensitivity and specificity, it needs to be considered that the approach on the basis of the binomial model is very naive. As discussed already in Chapter 5, the binomial model assumes (1) constant probability for the considered outcomes (correct classification/misclassification) over the studied population and (2) indepedent assignments (of the outcome). As discussed earlier, neither assumption can easily be believed valid for the studied data due to the fact that many faulty observations are the result from the same fault event (thus not independent) and that the magnitude of a fault is likely to affect the chances for misclassification.

In Section 5.5, two potential improvements were given with respect to MPCA-based monitoring. Both improvements may equally be valid for the purpose of diagnosis. Consider the first suggestion in which faulty observations are included into the calibration data set for subspace identification so as to obtain an augmented data set. By doing so, larger leverage, i.e. impact, of single observations can be expected when the deviations (within the subspace defined by the error-free data) between normal and abnormal data are substantial. Call the resulting model the $MPCA_{aug}$ model. It was indicated that an additional MPCA exercise on the resulting scores for the normal observations is then required to identify a proper confidence region for the normal data, resulting in an additional model, denoted here as $MPCA_{mon}$. To do so, the $MPCA_{aug}$ scores of the normal observations need to be centered at least. To project new data onto this model the $MPCA_{aug}$ scores should be centered in the same fashion. In view of diagnosis, a third model, $MPCA_{diag}$, may be identified on the centered and scaled $MPCA_{aug}$ scores of the augmented

observation set rather than on the normal set to identify scores that correlate well with specific faults. Given that a diagnostic model does not necessarily require orthogonality of the scores, Factor Analysis may be chosen as a more generic tool, of which MPCA can be considered a special case. Note that the $MPCA_{diag}$ model is identified within the subspace already defined by the $MPCA_{aug}$ model so that only extreme events without breakage of the correlation structure may effectively be handled by the $MPCA_{diag}$ model.

A second improvement was proposed in Section 5.5 in view of reducing the variance of identified MPCA models, irrespective of the data sets used. While a generic version of constrained MPCA, such as FS-PCA, may not directly lead to an improved diagnosic tool, the choice of a specific knowledge-driven orthogonal basis may allow identifying faults that are well understood. Consider again the hydraulic profile as discussed before (see Figure 4.3). As discussed, such normal profiles may be approximated well by piece-wise linear functions. By means of a 1^{st} order spline basis, coefficients can be computed which can further be processed by an MPCA model. In this case however, the coefficients themselves may express meaningful quantities such as the slope of the weight profile during the filling phase. The computed magnitude of this slope is thus likely to be informative for diagnosis purposes in a very direct way. For the oxygen level, a basis may be used that accomodates for the normally expected constant and zero value for DO in anoxic and anaerobic conditions. Constraining a certain number of PC's to be zero or piece-wise constant in the corresponding time frames may be effective to do so.

In the presented work, the fuzzy clustering algorithm used was based on Euclidian distance so that spherical regions for each of the clusters result (see Section 3.4). A more refined approach may include the use of the Gustafson-Kessel algorithm (Gustafson and Kessel., 1979) which allows the fitting of ellipsoidal regions with different orientation for each cluster. Simpler alternatives to clustering are K-means clustering or K nearest neighbours clustering. Alternatively, Mixture MPCA (MixMPCA) may be used to model faults that correspond to a specific correlation structure. It was noted that the presence of faults with varying deviations from the normal operating point may lead to ineffective clustering due to the large leverage of certain fault classes corresponding to large deviations. Possibly, improved results may be obtained by stacking of MPCA and/or clustering models. Indeed, a base MPCA and cluster model may be identified for all fault classes. Upon blurring of many fault classes into one cluster, the observations included in the corresponding may serve to the construction of a separate MPCA and/or cluster model, hereby avoiding the influence of faults of larger magnitude. This stacking approach may

be repeated as necessary.

Cluster models were chosen as they allow to identify groups of similar data without prior knowledge. Indeed, in the original definition of the PhD project computer-aided identification of fault classes without prior knowledge was aimed for. However, given the intensive data screening as presented in Chapter 4, classification methods are now equally available. Classification trees, though simple in nature, may facilitate interpretation of resulting diagnostic models. More advanced alternatives may be based on Discriminant Analysis (Raich and Çinar, 1996, 1997) or kernel-based methods such as Support Vector Machines (Cortes and Vapnik, 1995; Burges, 1998; Schölkopf et al., 1998a; Widodo and Yang, 2007).

With respect to the suggested model improvement by means of knowledge-based basis functions, it may be considered that the geometrical approach to diagnosis as reviewed in 6.1 may become a viable approach for at least a part of the faults, in particular those that are of a rather technical and well-understood nature, such as reported pump failures, extremely low or high hydraulic loadings and cooler failures. Combinations of geometrical approaches and alternative approaches may lead to better characteristics of diagnosis models in terms of interpretability, generalization and performance.

6.5 Conclusions

Explorative analysis of principal scores was shown to confirm the data screening that was performed prior to MPCA modelling to a large extent. Moreover, for the hydraulic data, it was shown that additional aspects of the historical performance of the system, not given attention to during screening, could be revealed. The benefits were less clear in a similar excercise for the complete system. This was explained to be due to (1) the invalid assumption on a constant mean and covariance matrix and (2) the large influence of a few types of faults characterized by large deviations from the normal operation region. While the first problem cannot readily be tackled, the second problem may be tackled effectively by model stacking. In this approach, additional models may be used to focus on a smaller subsets of a historical data base. All-in-all, automated data-driven tools for diagnosis require further investigation, including the evaluation of modifications of MPCA models, such as using extreme-event data for subspace identification and/or implying constraints, and the evaluation of alternative and more advanced clustering models.

Chapter 7

Multivariate Statistical Process Control of a Sequencing Batch Reactor

Could it be that my dream would come true,
building a machine that would actually do
what I want it to do?

Taken from *Strange Machines* by the Gathering

7.1 Introduction

In this chapter, a multivariate controller for on-line phase length optimization is evaluated. The proposed controller is applied to optimize the length of the aerobic phase of the SBR under study. Continuation of this phase beyond the point in time where all desired (bio)chemical reactions are completed is unnecessary. For the

studied system, this means that the system has reached the endogeneuous respiration state. Moreover, excessive length of this phase leads to an energy consumption for aeration without improvement of effluent quality and to reduced length of other phases of the SBR cycle. A control strategy based on the on-line reaction end-point detection by means of the Hotelling's T^2 statistic is proposed and evaluated.

On-line optimization of wastewater treatment plants is a research field that has received considerable attention since the beginning of the 90's. Yuan et al. (2003) gives an extensive overview of applied control strategies for nitrogen removal systems. An important concept for control of aerobic reactors or aerobic phases of alternating or cyclic systems, is the endogenous respiration state. This state is typical for aerated wastewater treatment systems where all (targeted) oxidation reactions are finished. This means that (1) carbon sources, typically expressed as chemical oxygen demand (COD), are oxidized, (2) bulk nitrogen compounds are oxidized, i.e. all ammonia and nitrite is oxidized to nitrate, and (3) phosphorus uptake rate (PUR) becomes minimal, i.e. the speed at which phosphorus is internalized by microbial organisms becomes small. Ideally, the phosphate concentration in the bulk liquid is then close or equal to zero. Note that the three described reactions do not necessary occur in the same system, at the same time or with equal intensity nor are the described oxidation reactions completed at the same location in time. Continued aeration beyond the point in time where the endogenous state is reached is economically uninteresting as no improvement of effluent quality can be expected from further investment of aeration energy. Consequently, the detection of the endogenous state has been an appealing research subject to many.

The (biological) oxygen consumption, referred to as the oxygen uptake rate (OUR), is known to decrease to a minimal value as the endogeneous respiration condition starts. Indeed, as the major oxidizing reactions are completed, the (biological) oxygen consumption becomes minimal (Watts and Garber, 1993). The point in time where this drop in OUR is observed is refered to as the α_{OUR} point (Watts and Garber, 1995). A fair amount of research has therefore focused on the on-line detection of this described decrease in OUR in view of aeration optimization. Most applications are based on estimation of the OUR and use its (numerical) value for inference (Demuynck et al., 1994; Johansen et al., 1997; Klapwijk et al., 1998; Third et al., 2004; Balslev et al., 2005; Bisschops et al., 2006; Corominas et al., 2006; Guisasola et al., 2006; Shaw and Falrey, 2007). When a constant air flow is applied, the decrease in OUR can be detected as an accelerating upward trend in the dissolved oxygen (DO) profile. This point is typically referred to as the $\alpha_{\mathrm{O_2}}$ point (Plisson-Saune et al., 1996; Mauret et al., 2001).

Alternatively, the entering of the endogenous respiration state can be based on the ORP and pH profiles of nitrogen removal systems. While pH signals are shown to typically exhibit a minimum in their profiles when ammonium is depleted (Al-Ghusain et al., 1995), ORP signals are shown to exhibit inflection points when the nitrite/nitrate buffer is crossed, thus indicating complete nitrification. The point is often referred to as the ammonia break point (Wareham et al., 1994). This is indeed appropriate in most cases as the rate of ammonia oxidation is typically not considerably larger than the nitrite oxidation so that ammonia depletion and nitrite depletion occur at the same location in time. In other words, the oxidation reaction speed of ammonia via nitrite to nitrate is typically limited by the reaction speed of the first step in conventional wastewater treatment systems. The detection of one or both of the described points has been used for inference in many applications (Wouters-Wasiak et al., 1994; Hao and Huang, 1996; Charpentier et al., 1998; Paul et al., 1998; Zipper et al., 1998; Ra et al., 1999; Cho et al., 2001; Kim and Hao, 2001; Yu et al., 2001; Li et al., 2004; Wang et al., 2004; Balslev et al., 2005; Cecil, 2007; Guo et al., 2007). In Cecil and Skou (2005), an explicit model for the ORP signal is proposed in view of control.

Importantly, the assessed critical points in pH and ORP profiles only reflect the status of the oxidation of nitrogen compounds. Indeed, phosphorus uptake is not described to affect the pH and ORP signals. As such, the targeted points in pH and ORP do not necessarily indicate the start of the endogeneous respiration when phosphorus is a significantly constituting pollutant in the treated water, thus confounding the identification of the targeted endogenous respiration. An application in which extracted information on OUR and ORP is combined for reaction endpoint detection can be found in Puig et al. (2005). Peng et al. (2004) combine inferences on the DO and pH measurements. Control algorithms based on information from DO, ORP and pH sensors are reported in Andreottola et al. (2001); Ma et al. (2006); Marsili-Libelli (2006).

A fourth related indirect measurement is found in the conductivity, which is related to the bulk phosphorus concentration. It has been shown for Enhanced Biological phosphorus Removal (EBPR) systems, which are aimed at treating phosphorus-rich but nitrogen-poor wastewater, that phosphorus uptake (under aerobic conditions) results in reduction of the conductivity while phosphorus release (under anaerobic conditions) results in increasing conductivity (Maurer and Gujer, 1995; Andreottola et al., 2001). The rate at which the conductivity changes approaches zero as the phosphorus uptake becomes minimal and, as such, the conductivity profile can be used for identification of the end of phosphorus release as shown in Aguado et al.

(2006). Extensive studies that aim at the interpretation of conductivity profiles in systems that both treat nitrogen and phosphorus have not been reported as yet.

Most of the control applications are based on the evaluation of a set of preset rules, which are established on the basis of system knowledge or operators' experience. Next to control algorithms based purely on (crisp) rules, (artificial) neural network ((A)NN) models can be used within the control system. Cho et al. (2001) use NN regression models to predict nutrient concentrations, which then consequently serve then as inputs for the devised control rules. Yu et al. (2001) use the same type of models to obtain a filtered value of ORP and pH which are then input for the rule-based controller. Cohen et al. (2003) use a NN classification model to identify the targeted breakpoints. Bisschops et al. (2006) use a NN regression model to identify the location of the targeted breakpoint. Fuzzy control rules are used in Wang et al. (2004). Strikingly, in none of the latter applications the respective data mining technique is used to resolve the control decision problem itself. Indeed, the output(s) of the constructed model(s) serve(s) as an input for the control algorithm, which is never the result of a data mining excercise itself. In addition, the fact that redundant information is present in DO, ORP and pH signals is never accounted for when building the reported models or controllers.

In contrast to the former methods, Marsili-Libelli (2006) delivers the only reported application of a data mining tool, i.e. fuzzy C-means clustering, for control of watewater treatment plants in which (1) redundancy is inherently accounted for and (2) the predicted cluster directly leads to the pursued control action. In the modelling step, data samples from the monitored system are grouped by the clustering algorithm into two major groups, representing the exogenous respiration state and endogenous respiration state. During active control, a new data sample is assigned to one of the clusters hereby indicating whether the system is in exogeneous or endogeneous state. Even though Marsili-Libelli (2006) indicates that the trajectories of DO, ORP and pH may differ according to whether the ammonia depletion occurs before or after phosphorus release becomes minimal, this is not explicitly accounted for during modelling. The data shown suggest that ammonia depletion occurred after phosphorus uptake was complete for the data used in the modelling stage. As a result, the obtained cluster model may not be appropriate when the reverse situation would occur, i.e. when the phosphorus uptake is completed after ammonia depletion. As such, the resulting controller may lead to an inappropriate control action.

A multivariate control strategy is proposed which does not suffer from the problem identified with the approach of Marsili-Libelli (2006). Indeed, a (statistical) model is made to derive the similarity between future data samples and data samples that are known to be sampled during endogenous respiration. As such, no explicit assumption is made on the behaviour of the data during exogeneous respiration, except for not being similar to the behaviour during endogenous respiration. How this similarity is conceived for the chosen model is shown later.

7.2 Methods

In what follows, the applied method, its underlying assumptions and the proposed integration for control are given. First, the method is described and implemented adjustments are motivated. Then, the proposed controller including the applied test is presented. Finally, the real-life case study in which this controller was tested is described.

7.2.1 Model for state detection

In order to automatically detect a determined (temporary) state of a certain process, it is considered that such a state of the system is often characterized by its process rates. In addition, process rates are likely to be reflected in the values or trends of process data. As such, assessing whether the state of a process is similar to a desired state may be reached or facilitated by assessing the values or trends of process data and comparing them to typical values corresponding to the desired state. The following multivariate strategy is proposed to do so.

7.2.1.1 Variable and sample selection

First, variables of which the values are known to describe the targeted state are chosen and data samples that reflect the desired state to be detected are selected by operators or process experts. Hence, both these steps require essential knowledge of the system under study. Secondly, a model that describes these data is established and one or more tests that allow the evaluation of similarity of future data

samples to the selected data are devised. Interestingly, the data-driven modelling approach used in this work avoids the need for exact knowledge on the numerical behaviour of the data in the targeted state. Thirdly, the constructed tests are used to classify future samples as being similar to the data described by the established model or not.

7.2.1.2 Modelling

While other models may be valid for the given purpose, the multivariate normal distribution model is used in this study for which the standard Hotelling's T^2 statistic can be constructed. Practically speaking, this is the Hotelling's T^2 statistic for a PCA model in which all PC's are retained (no dimension reduction). The Q statistic is then obsolete (as all residuals are zero). The reader is referred to Section 3.3.1.5 for detailed explanation of the latter statistics. Dimension reduction is not pursued for the following reasons. First, the presented control scheme, using one rather two statistics, becomes simpler. Secondly, data mining nor improved interpretability of the data is a target in this study. Also, the number of considered variables is already low and consequently dimension reduction is not of paramount interest.

Assuming that the modelled data stem from a multivariate normal distribution and that the samples are independent, the Hotelling's T^2 statistic follows an F-distribution for which theoretical statistical limits can be constructed (Johnson and Wichern, 2002). The given statistical limit defines a ellipsoidal region in which –theoretically– a given percentage of the data lie under the modelled conditions (i.e. the null hypothesis). Corresponding equations can be found in 3.3.1.5. To illustrate how the definition of this ellipsoidal region can be used for state detection, a bivariate process is simulated that gradually evolves to a desired state. One possible trajectory is shown in Figure 7.1. The ellipsoidal region corresponding to the 99% confidence limit is shown for data corresponding to the targeted state. It can be observed that as the process continues, the trajectory of simulated samples enters the identified region. As such, detecting that a sample lies within this region -by means of the Hotelling's T^2- indicates that the process has entered the targeted state. In other words, the Hotelling's T^2 is used as a metric to define a region in which the belief that the data belong to a targeted state is acceptable.

Underlying assumptions that are needed for the statistical test are stressed here. These include:

- that the measurement samples are independent. This means that no auto-correlation exists between the measurement samples. This requires that the mean values of the measured variables are constant over time for the given desired state and are the same for all batches (constant mean process). In most practical situations, this is accomplished when the (error-free) values of the considered variables are constant for the given desired state and the measurement errors are independent.
- the measurement samples are drawn from a multivariate normal distribution.

Since neither of these two assumptions is true by default, care should be taken when interpreting or implementing the test. In view of the latter, the test is adjusted later on in this text.

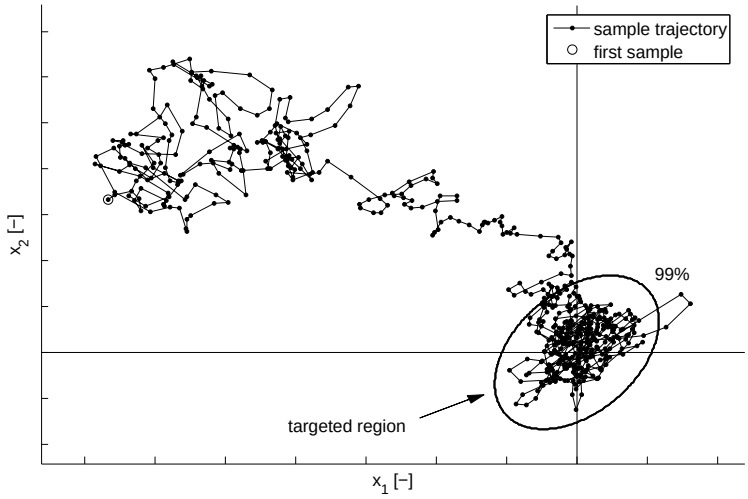


Figure 7.1: Bivariate example showing the proposed use of the Hotelling's T^2 test for state detection. The indicated ellipsoid indicates the confidence region defined by the Hotelling's T^2 statistic after estimation of the mean and covariance matrix on the basis of data corresponding to the targeted state.

7.2.1.3 On-line detection

If, for a given future data sample, the calculated statistic is below this limit, then the test is positive, i.e. the analyzed data sample is judged to be similar to the data described by the model. Consequently, the state of the system is judged to be similar to the desired state. If the statistic is above this limit, the test is negative, i.e. the analyzed data sample is considered not to be similar to the described data.

7.2.1.4 Adjustments to the test

Any deviation from the assumptions above may result in an inappropriate calculation of the statistical limit. An overestimation of this limit may result in too many positive tests for samples that are not truly similar to the data described by the model (false acceptance or type II error). An underestimation of the statistical limit may result in too many negative tests for samples that are truly similar to the data described by the model (false rejection or type I error). In this study, a type I error would mean that the phase to be optimized is continued while this is of no interest anymore (as the desired (bio)chemical conversions have been completed), therefore possibly leading to an unintended increase of economical cost. A type II error would mean that the phase is ended while the desired (bio)chemical conversions are not completed yet, i.e. the phase is ended too early, hereby leading to unmet targets for the given phase and possibly for the running and upcoming cycles as a whole. Completing the (bio)chemical conversions was considered of paramount interest in this study. A type II error was thus valued much more worse than a type I error. Given this consideration and given that the assumptions to the used statistical model are not generally valid, the following adjustments are pursued:

- A theoretical 90% limit is used. This is lower than the 95% or 99% levels which are more common in practice. As a result, the chance for a type I error is increased and the chance for a type II error is lowered compared to what common practice would have delivered.
- A set number of consecutive positive tests, N_{crit}, has to be established before the process is considered to have reached the desired state. This means that the Hotelling's T^2 test needs to remain under its limit for at least N_{crit} times the sampling interval before the phase is ended by the controller. This actual

test is thereby more restrictive and therefore leads to an increase in type I error and decrease of the type II error, as desired.

7.2.2 Integrated controller

Given the statistical test devised above, the following control strategy is proposed. The strategy is generally applicable to any optimization problem for which the detection of a temporary state is necessary. The control strategy is integrated into the control system of the studied SBR as shown in Figure 7.2. The graph shows a general scheme which is valid for optimization for any given phase of a cyclic or batch process. Denote the optimized phase as the i^{th} phase. Until a minimal length of the optimized phase, $t_{min,i}$, is reached, a counter, C, is kept to zero. As soon as this minimal time length has passed, the counter is allowed to increase. While the i^{th} phase is running, preprocessed data are obtained from the raw data. In this study, data preprocessing consisted of filtering with a second order lowpass Butterworth filter. The devised (statistical) test is used to determine whether the process is in the desired state. If the process is detected to be in this state (T: true) the counter is increased by one; if not (F: false), the counter is reset to zero. If the counter reaches the set critical level, N_{crit}, or if the running time of the phase has reached its maximal length, the i^{th} phase is ended and the next phase is started (i+1). The latter control action is referred to as the shutdown control action. As a new batch or cycle is started, the control algorithm is initialized again.

7.2.3 Case study

The studied process is a batch process for nutrient removal that consists of 5 major phases. A scheme of the standard operation (without phase length optimization) can be found in Figure 7.3. This standard operation with fixed time lenghts for the constituting phases exhibits an anaerobic phase (60 min., ANAER) including the addition of influent during the first 30 minutes, a first aerobic phase (130 min., AER1) of which the length will be optimized, an anoxic phase (80 min., ANOX) including the addition of influent during its first 10 minutes, a second aerobic phase (30 min., AER2), including sludge wastage in its last minute, a settling phase (45 min., S) and a draw phase (15 min., D). The phase that is optimized by the proposed control algorithm is the first aerobic phase (AER1). The total length of the batches is however kept the same, by extension of the length of the anoxic phase. This

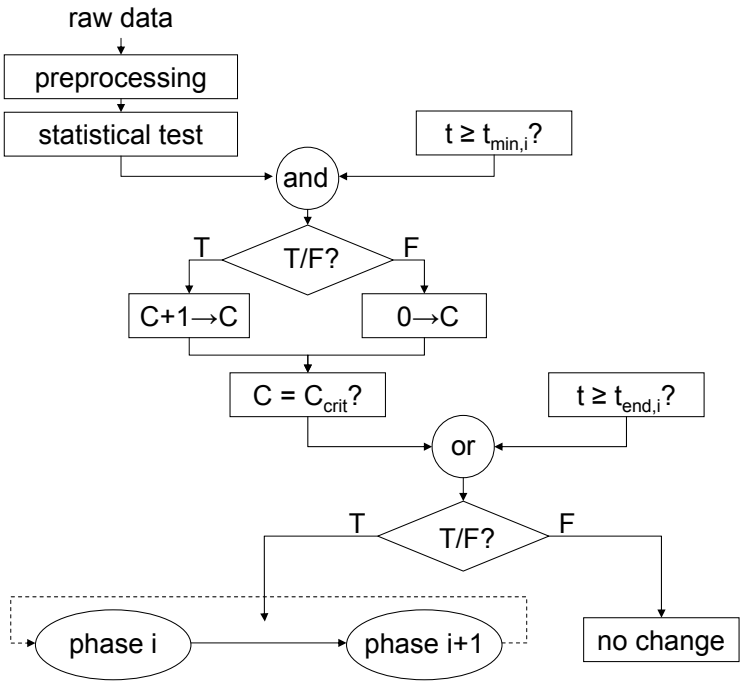


Figure 7.2: Scheme of the integrated control algorithm. A minimal time length, $t_{min,i}$, for the optimized phase is guaranteed. Upon reaching this minimal length, positive statistical tests result in incremental increases of a counter, C. Negative tests reset the counter to zero. If the counter reaches a critical number or when the maximal time length of the phase is reached, the phase is shut down and the next phase starts.

extends the time allowed for the denitrification process, while reducing the aeration time. The scheduling that results is given in Figure 7.3. The minimal length of the aerobic phase was set to 60 minutes. The maximal length of the aerobic phase was defined to be 130 minutes, as in the standard operation. The length of the anoxic phase is hereby minimally 80 minutes and maximally 170 minutes.

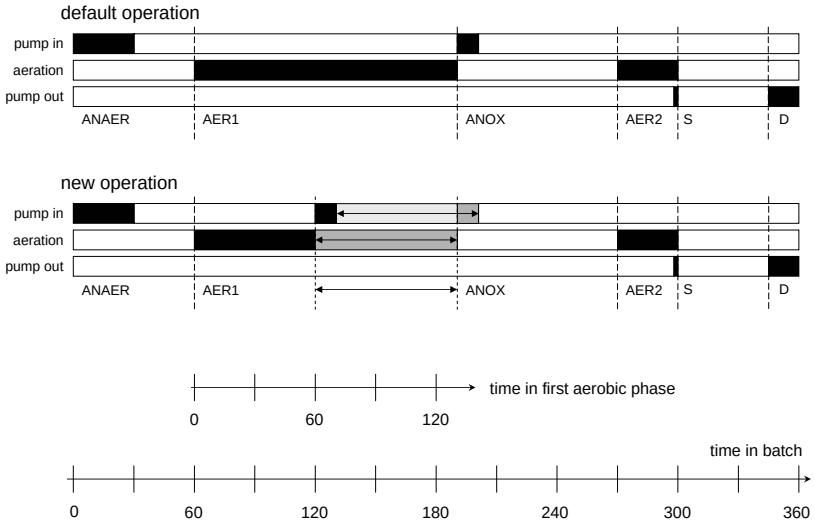


Figure 7.3: SBR phase scheduling in default and new operation

7.2.4 Selection of modelled variables

The trajectories of the filtered air flow rate and the filtered derivatives of oxygen concentration, pH and ORP during the studied aerobic phase are shown in Figure 7.4. The Butterworth filters were tuned so to obtain a cutoff period of 5 minutes (cutoff frequency = 0.0015 times the Nyquist frequency). The derivative of the DO (dissolved oxygen) is positive until minute 40. This is the point where the (controlled) dissolved oxygen level stabilizes around its setpoint level for the first time (not shown). This correlates with a stabilization of the air flow rate (Figure 7.4b), which is the actuator used to control the oxygen level. The DO level and air flow rate remain around the same level from minute 40 until minute 80 (derivative around zero). At that time, an upswing of the DO level is observed. This rise is due to a decreased oxygen cunsumption by the biomass, which is not (immediately) dealt with by the oxygen level controller. The air flow rate eventually decreases however (to get the oxygen level back at its setpoint) until about minute 100, whereafter the air flow rate is approximately constant. As a result of the described changes in the air flow rate, the derivative of the DO level becomes negative and levels off towards zero as the oxygen level stabilizes again around its setpoint. The latter series of events, starting with the upswing in oxygen level, indicate the

start of the endogeneous respiration state, which remains until the end of the aerobic phase. Continuation of the phase after minute 105 is therefore considered undesired and should therefore be stopped at that time.

From the beginning of the phase until minute 34, the pH derivative has a positive sign (Figure 7.4c). This is due to a net positive effect of the pH increase due to CO_2-stripping and pH decrease due the first nitrification step, also identified as biological ammonia oxidation or nitritation. As the CO_2-concentration lowers and as the activity of the nitrifying biomass increases, the acidifying effect of nitrification starts to dominate resulting in a negative sign of the first derivative at minute 34. The sign of the derivative remains below zero for the remainder of the phase. The derivative increases from minute 85 to 105, when the endogeneous respiration state is initialized, and remains approximately at the same level from minute 105 onwards.

The derivative of the ORP level (Figure 7.4d) shows an increase from minute 10 to 40. From minute 40 to 85 onwards the ORP level exhibits a steady increase, concurring with the steady behaviour of the air flow rate and oxygen level. This in-

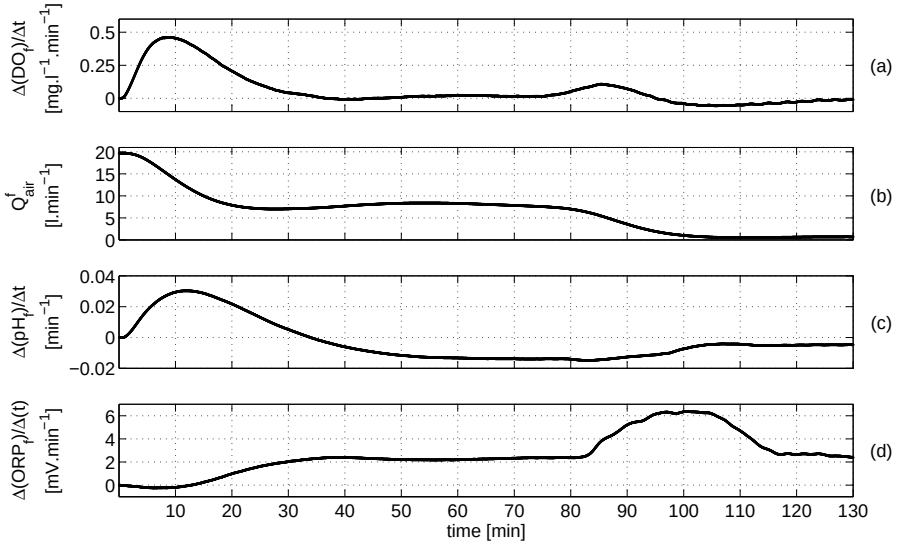


Figure 7.4: Typical profiles of the filtered air flow rate (b) and filtered derivatives of DO (a), pH (c) and ORP (d) during the aerobic phase

dicates a steady increase in the nitrate nitrogen concentration (NO_3^--N). At minute 85, a fast increase in the derivative (acceleration of the ORP level) is observed, indicating the breach of the nitrite-nitrate oxido-reduction buffer system. At minute 85, the second nitrification step (biological nitrite oxidation, nitratation) is thus completed. By minute 115 the ORP derivative reaches a level comparable to its previously stabilized level, thus indicating the reaching of a new buffer (i.e. $[O_2/H_2O]$). In summary, the on-line measurements that are shown here reach a certain level after the process has reached its endogeneous respiration state and remain close to that level.

It is noted that the conductivity measurements were available but have not been included as a variable described by the constructed model. This is supported by (1) the fact that unambiguous interpretation of conductivity profiles in wastewater treatment systems is not reported as yet and (2) the data quality of this sensor was doubted to a too large extent to include in the model.

An intensive measurement campaign was set up to measure the effluent quality variables total nitrogen (TN), total ammonia nitrogen (TAN), nitrite nitrogen (NO_2^--N), nitrate nitrogen (NO_3^--N) and inorganic phosphorus (PO_4^{3-}) during the batch for which on-line measurements were described above. Figure 7.5 shows the trajectories of these variables during the aerobic phase of this batch. As can be seen, the ammonia level lowers from about 9 mg/l at the beginning of the phase to approximately zero at minute 80, due to biological ammonia oxidation. As a result, nitrite is produced which is then further oxidized into nitrate. The nitrite level increases during the aerobic phase until minute 70, due to an apparently lower speed of the nitrate oxidation. As the production of nitrite decreases afterwards, nitrite levels decrease again to approximately zero by minute 85. Consequently, nitrate levels increase during the aerobic phase, from approximately zero at the beginning up to 11 mg/l at minute 115. Not surprisingly, the entering of the endogoneous resipration state as described in the previous paragraph occurs simultaneously with the depletion of ammonia and nitrite. Biological phosphorus uptake takes place simultaneously with the described nitrogen oxidation processes. The inorganic phosphorus is internalized by the Phosphorus Accumulating Organisms (PAO's) hereby leading to a reduction of the bulk inorganic phosphorus concentration, which is measured. This process halts at minute 85, right at the start of the endogenous respiration. The measurements suggest that inorganic phosphorus measurement lowers further from minute 105 to the end of the aerobic phase. Still, the larger part of the phosphorus uptake (approx. 80%) takes place before the endogenous respiration state is reached and at a faster rate. It can therefore be stated that the in-

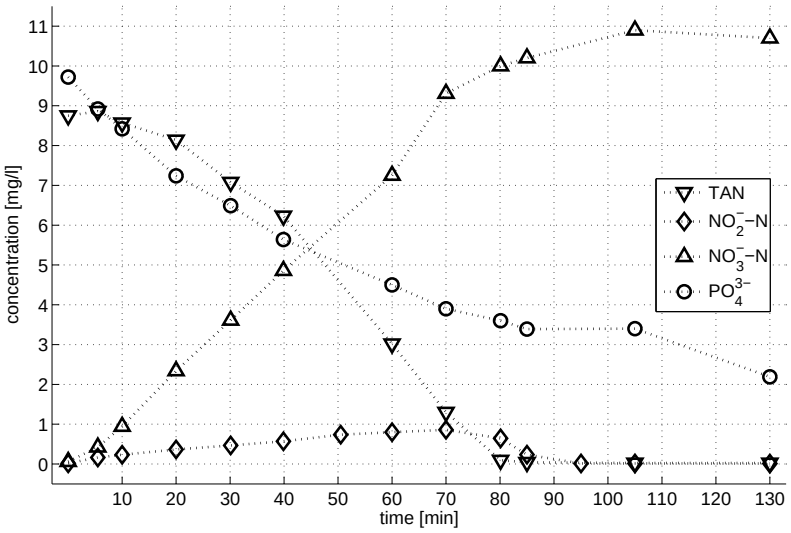


Figure 7.5: Measured profiles of total ammonia nitrogen (TAN), nitrite nitrogen (NO_2^--N), nitrate nitrogen (NO_3^--N) and inorganic phosphorus (PO_4^{3-}-P) during the aerobic phase of an intensively sampled batch

tensive measurement campaign indicates that when the endogeneous respiration is reached, the nitrogen oxidation processes are finished and a larger part of the phosphorus is taken up by the PAO biomass. As such, it is considered desirable to end the aerobic phase when the endogenous respiration state is reached. This state is shown to be detectable on the basis of the described on-line measurements as well. Therefore, the four described on-line measurable variables have been included into the statistical model as well as the derivative of the air flow rate. The latter is a valid measure for similarity to data stemming from endogenous respiration as well given that the steady behaviour of the air flow rate during endogeneous respiration results in a derivative close to zero.

7.2.5 Sample selection and applied parameters

The data samples used to construct the model were selected as follows. First, a set of 10 batches exhibiting endogeneous respiration behaviour at the end of the first aerobic phase were selected by the operators among the batches run in the week before the controller implementation. The point at which the operators believed that the variables showed endogenous respiration behaviour was indicated for all of them. All data samples after this point in time and before the end of the aerobic phase were used for modelling. As mentioned before, a 90% limit was used for a single statistical test and a cutoff period of 5 minutes was applied for data filtering. To shut down the aerobic phase, the statistical test was required to be under its limit for 30 contiguous tests (= 1 minute).

7.3 Results

7.3.1 Detection performance

In Figure 7.6 the Hotelling's T^2 statistic is shown as evaluated during the aerobic phases of two batches. Batch 1 is a batch in which the implementation of the control algorithm in LabView was verified without actually allowing the shutdown control if commanded. As can be seen, the Hotelling's T^2 statistic remains above its 90% limit until 110 minutes in the batch. In the minute following this point the Hotelling's T^2 did not rise above the set limit and the shutdown of the aerobic phase is therefore commanded by the algorithm at 111 minutes in the aerobic phase (30 contiguous positive tests take 1 minute). Since this command is not passed on to the actuators (aeration, pumps), the Hotelling's T^2 statistic is evaluated as well beyond this point. As can be seen, the Hotelling's T^2 statistic did not rise above this limit between the point in time at which the aerobic phase would be ended by the proposed controller and the default end of the aerobic phase. This was equally concluded for 3 other contiguous batches (not shown). The plant operators also confirm that endogeneous respiration was achieved when the end of the aerobic phase is commanded by the controller for these test batches. Based on these preliminary tests, it was decided to activate the control algorithm, i.e. to allow the transfer of the command by the proposed algorithm to the actuators. The Hotelling's T^2 statistic trajectory for the first batch in which the control algorithm was completely activated is also shown in Figure 7.6. In this batch, the Hotelling's

T^2 statistic remains above the 90% limit until 95 minutes in the aerobic phase. In the minute following this point in time, the Hotelling's T^2 statistic remains below the set limit. As a result, the transition from the aerobic phase to the anoxic phase at 96 minutes in the aerobic phase is commanded. Consequently, the aeration is switched off and the anoxic filling is started at minute 96, hereby gaining a precious 34 minutes.

Figure 7.7 indicates the time at which the aerobic phase was ended and the anoxic phase was started during the period in which the proposed controller was active. As can be seen, the aerobic phase is ended by the controller before the default end of the aerobic phase in each batch, hereby leading to an effective shortening of the aerobic phase in each batch. The mean time at which the aerobic phase was ended was 76 minutes, i.e. 54 minutes before the default end of the aeration time. Put otherwise, a mean reduction of 41% of length of the aerobic phase was obtained.

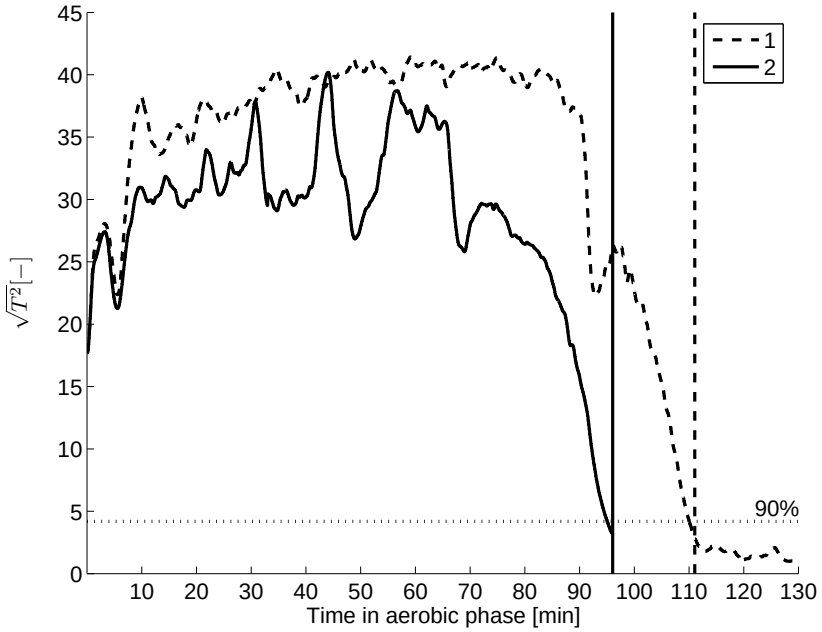


Figure 7.6: Square root of Hotelling's T^2 statistic during the first aerobic phase of a test batch (1) and the first batch with on-line phase optimization (2). Vertical lines indicate when the shutdown of the aerobic phase is commanded. The horizontal line indicates the applied 90% limit.

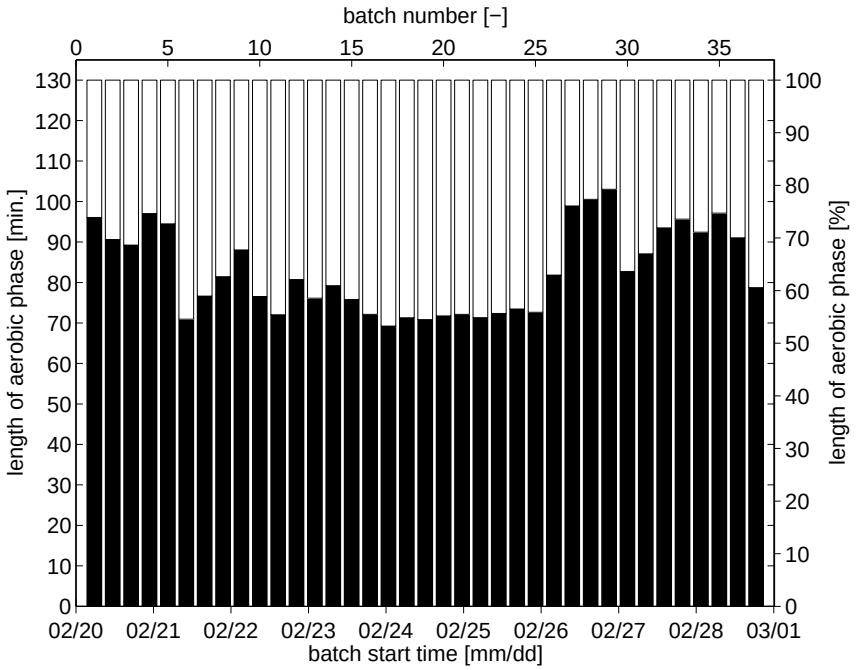


Figure 7.7: Shutdown time and proportional length of the aerobic phase during active phase length optimization.

With respect to the blown air volume, the proportional volume of air (and thus energy) that would be saved (in the respective aerobic phase) by the controller is 1.6%. This is a fairly minimal reduction and is largely due to the fact that a well-functioning PID control of the oxygen was already implemented for oxygen setpoint control. Indeed, it was shown already that the air flow rate is reduced in response to the reduced oxygen consumption by the biomass. For the period in which the controller was active, the blown air volume that would be blown without the proposed controller was estimated by assuming that the air flow rate would remain equal to the air flow rate at shutdown time until the end of the aerobic phase. As such, a proportional reduction of the blown air volume is estimated to be 5.3%. It is noted here that the same procedure, i.e. assuming a constant air flow rate beyond the shutdown time, would render the estimated reduction to be 2.1% for the test batches. The estimated reduction should therefore be regarded as an upper limit for the actually obtained reduction of blown air volume.

7.3.2 System performance

In Figure 7.8 to 7.13, the measurements of the effluent quality variabes nitrate nitrogen (NO_3^--N), nitrite nitrogen (NO_2^--N), Total Ammonia Nitrogen (TAN), Total Nitrogen (TN), Total phosphorus (TP) and Chemical Oxygen Demand (COD) are shown from 20 days before implementation of the controller until 11 days after implementation. For descriptive purposes, the medians (med) of the measurements in the periods before and after implementation are shown together with the medians plus and minus twice the median absolute deviation (mad) from these medians. The former indicate the central tendencies of the measurements during the respective periods, while the latter are indicators for the spread of the measurements. These descriptors are more robust towards outliers than the classic mean and standard deviations and were selected for this reason. To statistically evaluate the effect of the controller implementation on the process performance, the two-tailed Mann-Whitney-U test is applied to the measured effluent qualities. The two samples are defined as the considered effluent quality measurements *before*, resp. *after*, controller implementation. The null hypothesis for the test is then that there is no effect of the controller implementation on the considered measurement. This non-parametric test does not require that the measurements in each sample are normally distributed (in contrast to e.g. its parametric equivalent, the Student's t test) and is less prone to the influence of outliers. The p-values are computed as well as the so called Z-score for the sample of data before controller implementation. Practically, this score is positive (negative) in case of a raising (lowering) effect. Table 7.1 gives the p-values and corresponding Z statistic for the tests evaluated in this study. The null hypotheses will be rejected at 0.05 (5%) significance level.

In Figure 7.8, it can be observed that nitrate values are generally lower in the period after implementation compared to the period before. The two-tailed Mann-Whitney-U test delivers a p-value of 0.00018. In other words, the probability that the controller implementation delivers no effect is estimated to be less than 1 in 5000. The negative value for the corresponding Z-score indicates that the effect is negative, i.e. a lowering effect of the controller on nitrate is apparent.

Nitrite values are generally low, indicating that nitrite oxidation is complete both before and after the controller implementation (Figure 7.9). Even though visual inspection of this graph may suggest higher nitrite values in the period after implementation, no effect is statistically acceptable on the basis of the Mann-Whitney-U test. It is noted that less samples were taken for nitrite concentration measurement.

The total ammonia nitrogen (TAN) measurements made during the studied period are plotted in Figure 7.10. The measured concentrations of this species are generally low as well, which suggests that complete nitrification is equally reached during the larger part of the batches both before and after the controller implementation. This hypothesis is accepted on the basis of the two-tailed Mann-Whitney-U test (p-value = 0.28).

Visual inspection of Figure 7.11 may suggest a lowering effect of the controller on TN. The two-tailed Mann-Whitney-U test confirms that the controller has an effect on the TN values (p-value = 0.0089). Given that the corresponding Z-score is negative, it can be concluded that a lowering effect is observed.

The measured total phosphorus concentrations are plotted in Figure 7.12. Visual inspection does not suggest a change in the measured variable. The two-tailed Mann-Whitney-U test indicates this equally. Therefore, the null hypothesis –no effect of the controller on the total phosphorus level– is accepted.

The concentrations of COD (Chemical Oxygen Demand) are shown in Figure 7.13. Visual inspection suggests a lowering effect of the controller on COD. However, the Mann-Whitney-U test does not reject the null hypothesis. An effect of the controller implementation on the COD concentrations can therefore not be accepted statistically.

Given the former evaluations of effluent quality criteria, the evaluated effects of the controller inplementation on the overall performance of the system can be summarized as follows. The implementation of the controller has led to an effective reduction of the effluent nitrate nitrogen and total nitrogen while the levels of ammonia nitrogen, nitrite nitrogen, total phosphorus and COD before and after controller implementation are not significantly different. It can therefore be concluded that the implementation of the controller has unambiguously led to an overall improvement of the effluent quality.

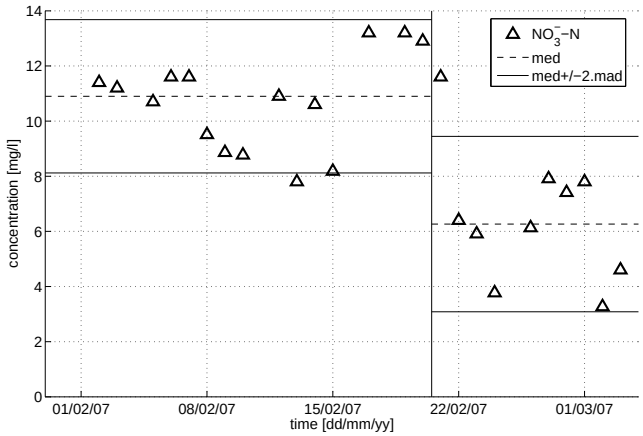


Figure 7.8: Nitrate nitrogen (NO_3^--N) concentration before and after controller implementation

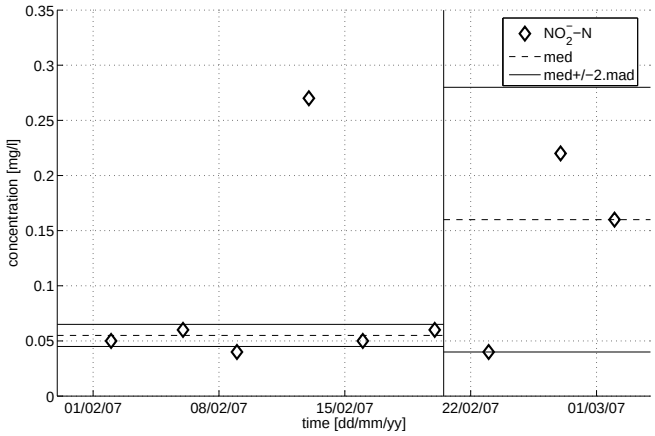


Figure 7.9: Nitrite nitrogen (NO_2^--N) concentration before and after controller implementation

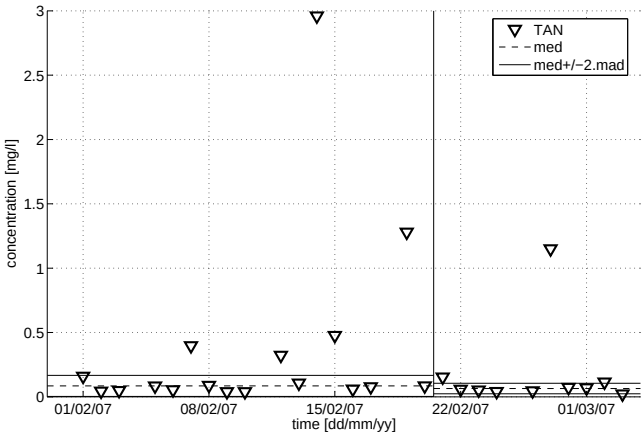


Figure 7.10: Total ammonia nitrogen (TAN) concentration before and after controller implementation

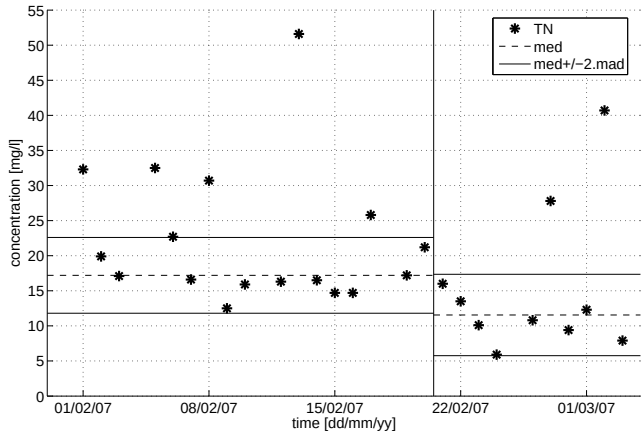


Figure 7.11: Total nitrogen (TN) concentration before and after controller implementation

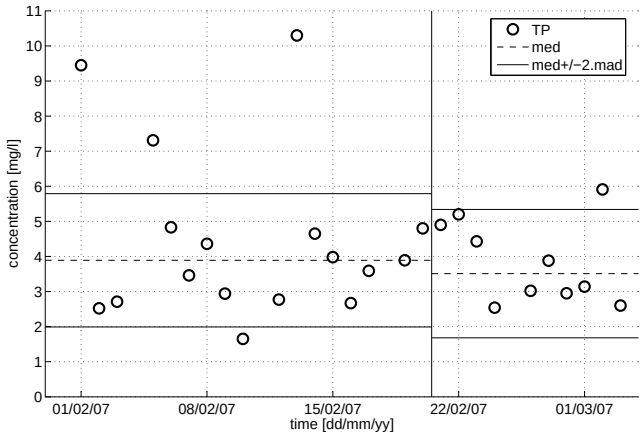


Figure 7.12: Total phosphorus (TP) concentration before and after controller implementation

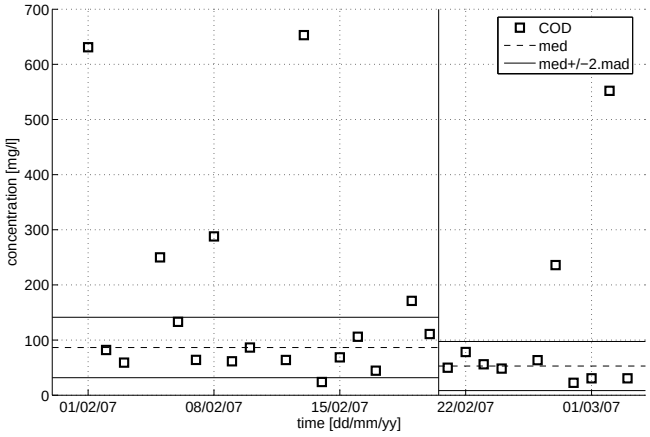


Figure 7.13: Chemical oxygen demand (COD) before and after controller implementation

species	p-values (exact)	Z
NO_3^--N	0.00018	-3.4669
NO_2^--N	0.71	0.3873
TAN	0.28	-1.1046
TN	0.0089	-2.5607
TP	0.94	-0.10042
COD	0.057	-1.908

Table 7.1: p-values for two-tailed Mann-Whitney statistic and corresponding Z-score computed for the samples taken before controller implementation

7.4 Discussion

An on-line phase length optimization strategy is proposed on the basis of the evaluation of a relatively simple multivariate statistic, being the Hotelling's T^2 statistic. To avoid large type II error rates, the test was adapted by requiring multiple contiguous positive tests and by adopting a relatively low confidence limit. The proposed algorithm was succesfully tested and led to a mean phase length reduction of 41%. The estimated proportional reduction in blown air volume is however less and estimated to be no more than 5.3%. As such, improved effluent quality is the only factor which supports the implementation of the given controller for phase optimization. It should be noted however that, when a fixed air flow rate would be used (thus for systems without DO setpoint control), the proposed control algorithm is equally viable and is expected to lead to larger improvements on the blown air volume and aeration energy costs.

In addition to the observed success of the method to reduce the phase length, the controller is shown to have led to an effective improvement of the effluent quality, thanks to an optimization of the length of the different reaction phases. In view of the fact that no effect on the effluent phosphorus concntration could be observed (Figure 7.12), it is important to note that the studied SBR suffers from nitrogen overload (especially NO_3^--N) leading to incomplete denitrification in the anoxic phase following the optimized aerobic phase. If such completion would however occur (e.g. for systems that are challenged less), the system would enter an anaerobic state in which phosphorus release can occur (Sin et al., 2006). The released phosphorus is then expected to be taken up again in consequent aerobic phases. The mechanism of phosphorus release and uptake leads to increased growth of

PAO's which are responsible for the described processes. For systems without the addressed nitrogen overload, a reduced effluent phosphorus concentration would be expected in the long term.

To be fair, it must be stated that the success of the applied strategy shown here is partly thanks to the fact that the modelling step was performed on data obtained shortly before the actual implementation of the proposed controller. Indeed, as the behaviour of on-line measurements during the SBR process cycles are generally not expected to be numerically the same over long periods of time because of changes in the microbial population and its behaviour (substrate affinity, growth rate), a larger delay between the modelling and implementation might have lead to less convincing results. Future research should therefore be aimed at the automated updating of the applied model. To this end, one may decide to delay the control decision (to end the aerobic phase) by a period of time so to obtain new data reflecting endogeneous respiration at each batch. Deactivating the controller at regular intervals or when the optimization of the phase length is of lesser importance may allow the collection of such new data as well. Model updating itself may be based on a moving window approach, as by Lee and Vanrolleghem (2003) or based on recursive updating of the covariance matrix as by Lennox and Rosén (2002).

7.5 Conclusions

A controller for on-line phase length optimization is proposed and evaluated in an on-line experiment. The newly proposed algorithm integrates the Hotelling's T^2 statistic, commonly used in the field of multivariate statistical process monitoring (MVSPC) with a simple control scheme. The resulting controller was succesfully tested on a pilot-scale sequencing batch reactor (SBR), in particular for optimization of one of its constituting aerobic phases. A clear proof of concept is given as an effective shortening of the respective phase with 41% is shown to result. In addition, the effluent quality of the system is shown to be improved. Especially nitrate nitrogen levels are shown to be reduced. In contrast, the expected energy savings are minimal, due to an earlier implementation of a DO setpoint controller.

It was noted that the underlying assumptions of the applied statistical test are not generally valid. Two adjustments, i.e. the use of a rather low confidence level and the requirement for a set of contiguous positive tests before the control action is pursued, were implemented to counteract potential problems related to the lat-

ter observation. Future research may however aim at the construction and use of (statistical) models that do not violate their underlying assumptions, in addition to adaptation of the model to changing system conditions.

Importantly, the proposed controller is general in its nature and is not limited to the reported application nor to the phase that was chosen for optimization. Future applications may therefore be aimed at the optimization of other phases that are typical for the studied SBR. The optimization of anoxic (detection of the end of denitrification) or anaerobic (detection of the end of phosphorus release) phases are relevent goals for control. More generally, the proposed controller allows to optimize any process with respect to its length given that the targeted state is uniquely described by data obtained on-line.

Part IV

Qualitative Representation of Trends: improvements and applications

Chapter 8

Improved method for Qualitative Representation of Trends

8.1 Introduction

In Section 3.5.5 the original method for Qualititative Representation of Trends (QRT) on the basis of cubic spline wavelets by Bakshi and Stephanopoulos (1994) was shown to lack the ability to identify contiguous sets of inflection points. Bakshi and Stephanopoulos (1994) argue that this is acceptable for most practical applications. It is however not obvious how relevant inflection points should be assessed in opposite cases, i.e. cases where it is of interest to discriminate sequences on the basis of apparent inflection points. Finding a successful approach for such cases is exactly the aim of the work presented in this chapter. Also, the advantages of the resulting improvements for process monitoring of SBR systems as the one studied in this work are illustrated. For a good understanding of what follows, it is recommended that the reader has captured the details provided in Section 3.5.5.

Three approaches for inflection points are evaluated in this chapter:

1. *Accept-all.* All inflection points in between included extrema are automatically accepted. This can be considered a naive approach
2. *Single inflection point.* This is the approach followed by in the original method by Bakshi and Stephanopoulos (1994). In between two extrema, a single inflection point is allowed only
3. *Witkin's criterion.* The criterion that is already used for selection of extrema is now applied for inflection points as well.

The following paragraphs deal with the further development of and evaluation of each of these approaches. Both a simulated and real-life example will be used to evaluate the different approaches.

8.2 Testing data

8.2.1 Simulated series

The series used for illustration of the original method in Section 3.5.5 is used here again for evaluation of the different approaches. Figure 8.1 repeats the figure with signal and the desired outcomes as presented already in 3.5.4. The noise-free version of this signal exhibits 3 contiguous inflection points between time indices 4535 and 7554, which cannot be identified as such by means of the original method (see Section 3.5.5). The desired qualitative representations are KLKLKLKL $\equiv$ $(\mathsf{KL})_4$ (monotonic) and ABCDADABCDABCDAB $\equiv$ $\mathsf{ABC}(\mathsf{DA})_2(\mathsf{BCDA})_2\mathsf{B}$ (triangular) respectively. The signal was simulated according to the following equation:

$$x_t = z_{t,1} + z_{t,2} \cdot z_{t,3} + e_t \tag{8.1}$$

where:

$$z_{t,1} = sin\left(2 \cdot \pi \cdot \frac{t}{\tau_1}\right)$$

$$z_{t,2} = sin\left(2 \cdot \pi \cdot \frac{t}{\tau_2}\right)$$

$$z_{t,3} = -2 \cdot \left(\frac{1}{1 + e^{12 \cdot \left(\frac{2}{3} - \frac{t}{N}\right)}}\right)$$

e_t : white noise sequence

t : time index (integers from 1 to N)

N : length of the time series (12000)

τ_1, τ_2 : oscillation periods (N/2, resp. N/6)

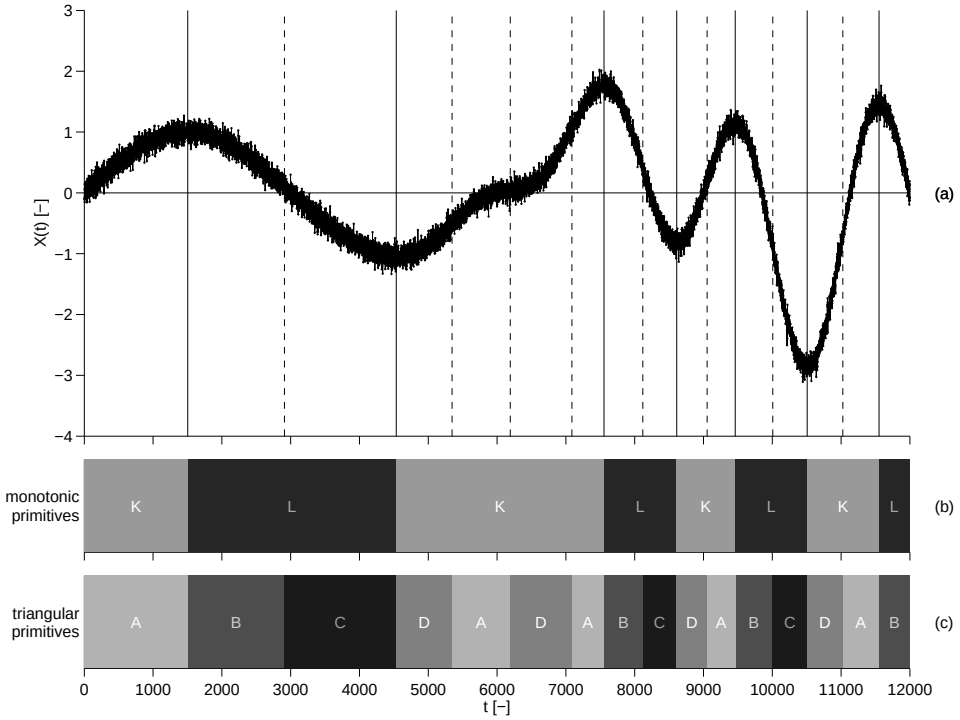


Figure 8.1: Simulated series and its true qualitative representation: (a) Simulated series; vertical lines indicate extrema (–) and inflection points (––) of the corresponding noise-free series, (b) monotonic representation ($(KL)_4$) and (c) triangular representation ($ABD(DA)_2(BCDA)_2B$).

8.2.2 Real-life series

In addition to the simulated series, a typical time series of an oxidation reduction potential (ORP) sensor signal in a Sequencing Batch Reactor (SBR) with alternating aerobic and anoxic conditions is used for benchmarking as well. The signal, already used for illustration in Section 3.5.1, is plotted in Figure 8.2. As mentioned before, this signal exhibits three contiguous inflection points (respectively at minute 33, 77 and 91) and represents therefore a challenge for inflection point identification. As already discussed in 3.5.1, inflection points in ORP signals are relevant indicators for reaction endpoints and their identification in time is relevant to process monitoring and control of wastewater treatment systems.

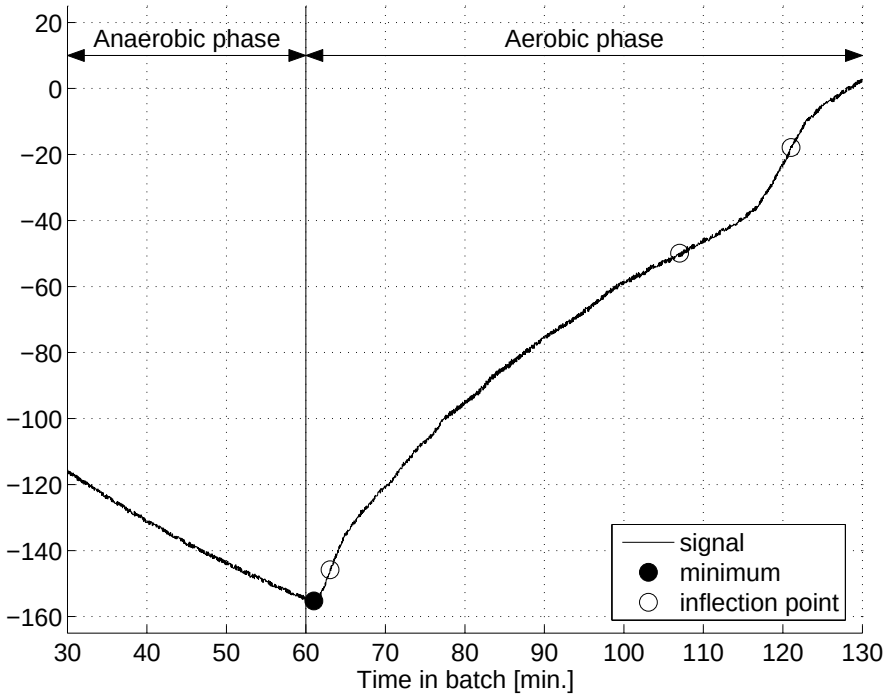


Figure 8.2: Typical ORP time series of the studied SBR.

8.3 Results

8.3.1 Simulated series

Results will be displayed and discussed in detail for the number of wavelet scales (P) set to 9 (see Section 3.5.2). At the end of this section it will be evaluated how this choice of P may affect the results. The results for the first steps in this QRT method, i.e. wavelet decomposition, construction of the wavelet interval tree and assessment of relevant monotonic episodes, were already presented in 3.5.5. Results shown here are limited to the assessment of inflection points, constituting the last step in the complete method. Results for all three approaches, including the original approach, are shown here. To improve interpretability of the next paragraphs, Figure 8.3 repeats the monotonic wavelet tree shown already in 3.5.5.

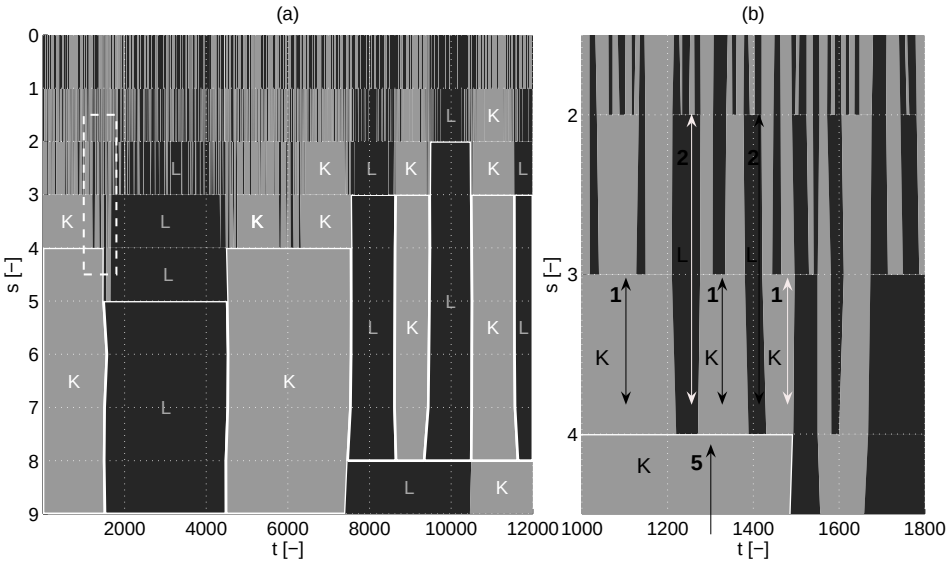


Figure 8.3: (a) Wavelet interval tree (monotonic). Selected episodes are indicated by white contours. Monotonic presentation: KL_4. (b) Detailed part of the wavelet interval tree, indicated in (a) by a dashed rectangle.

8.3.1.1 Approach 1

The wavelet interval tree with triangular episodes is shown in Figure 8.4 for the accept-all method (approach 1). In this approach, the simplification procedure as proposed by Bakshi and Stephanopoulos (1994), by which contiguous sets of inflection points are replaced by a single inflection point (see Section 3.5.5), is not applied. The relevant inflection points, and thus the triangular episodes as well, are simply obtained by selecting the triangular presentation at the most detailed scale of each monotonic episode. Consider for instance the first selected monotonic episode in Fig. 5. This episode exists over scales 5 to 9 (A, time index 1 to 1486

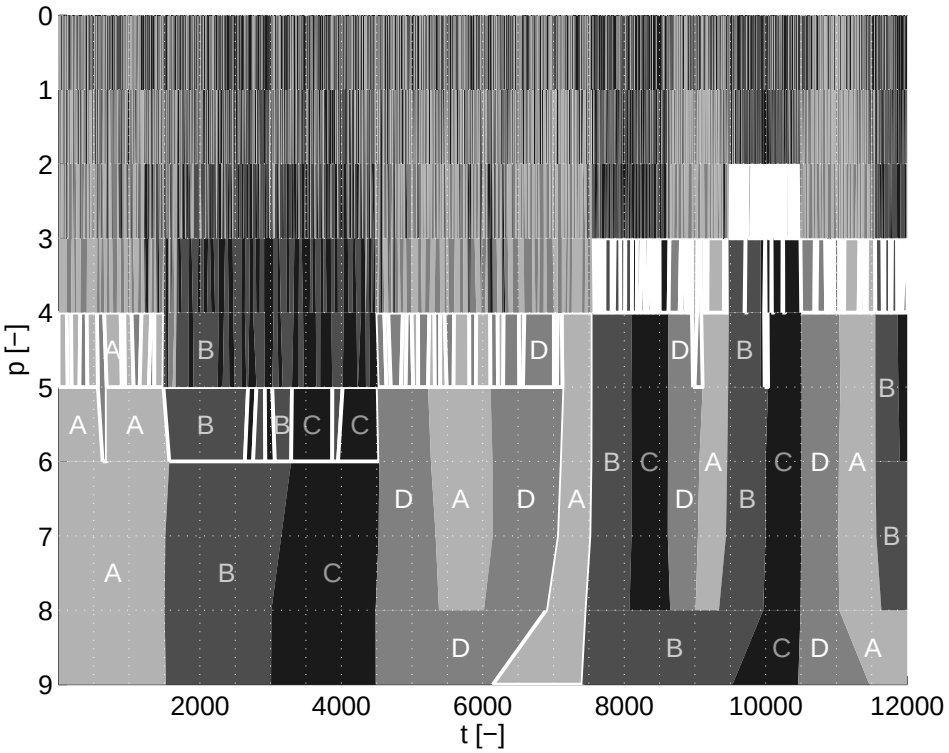


Figure 8.4: Approach 1: wavelet interval tree (triangular) in accept-all approach. Selected episodes are indicated by white contours. Triangular presentation: $A(DA)_6(BC)_4(DA)_{11}(BC)_9(DA)_6(BC)_{21}(DA)_7(BC)_4$.

at scale 5). Consequently, the included triangular episodes at scale 5 are included in the triangular representation ($\mathsf{A}(\mathsf{DA})_6$). The next relevant monotonic episode exists over scales 6 to 9 (time index 1486 to 4536 at scale 6) and consequently the included triangular episodes at scale 6 are assessed as relevant ($(\mathsf{BC})_4$). The same procedure is continued for all representing monotonic episodes. The final triangular presentation is more complex than the triangular representation of the noise-free signal, thus not matching the desired outcome. For instance, the second monotonic episode, represented by a BC-sequence in the triangular representation of the noise-free signal corresponds to a $(\mathsf{BC})_4$-sequence here.

8.3.1.2 Approach 2

In Figure 8.5 the wavelet interval tree is shown with triangular episodes for the second approach, i.e. the original approach as defined by Bakshi and Stephanopoulos (1994). It is noted that this was already shown in 3.5.5. Prior to construction of this wavelet interval tree, sets of contiguous inflection points are replaced by the inflection point in that set with maximal (minimal) value for the detail signal in case of an upward (downward) trend. This means only the inflection point with a maximal (for upward trends) or minimal (for downward trends) value for the 1^{st} derivative is kept in the representation. The steps that follow this assessment of inflection points are the same as in the previous approach. The triangular episodes at the most detailed scale of each representing monotonic episode are selected as relevant. Due to the applied modification, the representation is simpler and more similar to the desired outcome. For instance, the second monotonic episode is now represented as a single BC sequence as in the desired representation. The applied simplification also results in a single DA sequence for the third accepted monotonic episode. However, this representation is simpler than the desired $(\mathsf{DA})_2$ sequence as in the triangular representation of the noise-free signal.

8.3.1.3 Approach 3

In Figure 8.6 the wavelet tree with triangular episodes is shown for the new approach based on the application of Witkin's stability criterion. Note that the triangular episodes (polygons) are exactly the same as with approach 1 (Figure 8.4). The selection procedure is however different. In this new approach Witkin's stability criterion is applied to the triangular episodes as follows. Consider the first

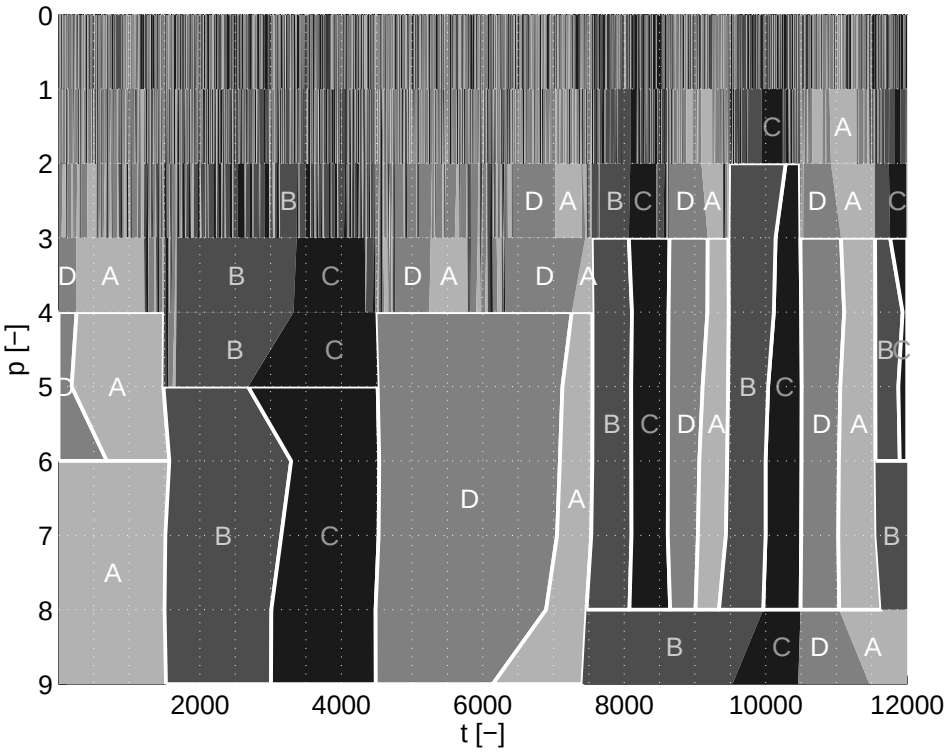


Figure 8.5: Approach 2. wavelet interval tree (triangular) in the single inflection point approach. Selected episodes are indicated by white contours. Triangular presentation: $(\mathsf{DABC})_4$.

triangular episode at scale 9 in Figure 8.6 (A, time index 1 to 1520 at scale 9). This triangular episode exists over scales 7 to 9 (range of scales=3) and corresponds to the first monotonic episode in Figure 8.3 (from scales 5 to 9). The latter means that triangular episodes at scales 1 to 5 will not be accepted in any case. Starting from scale 9, it can be seen that at scale 6, the episode is split into an ADA sequence, with scale ranges 1, 2 and 1 (mean range of scales = 1.33). By application of Witkin's stability criterion, the split is not accepted (1.33<3). Consider now the fourth triangular episode at scale 9 (D, time index 4488 to 6141). This episode exists over 1 scale only and is split into a DAD sequence at scale 8 in which all episodes have a scale range of 3. In this case, the split is thus accepted (3>1). By continuing the application of Witkin's stability criterion until none of the triangular episodes can

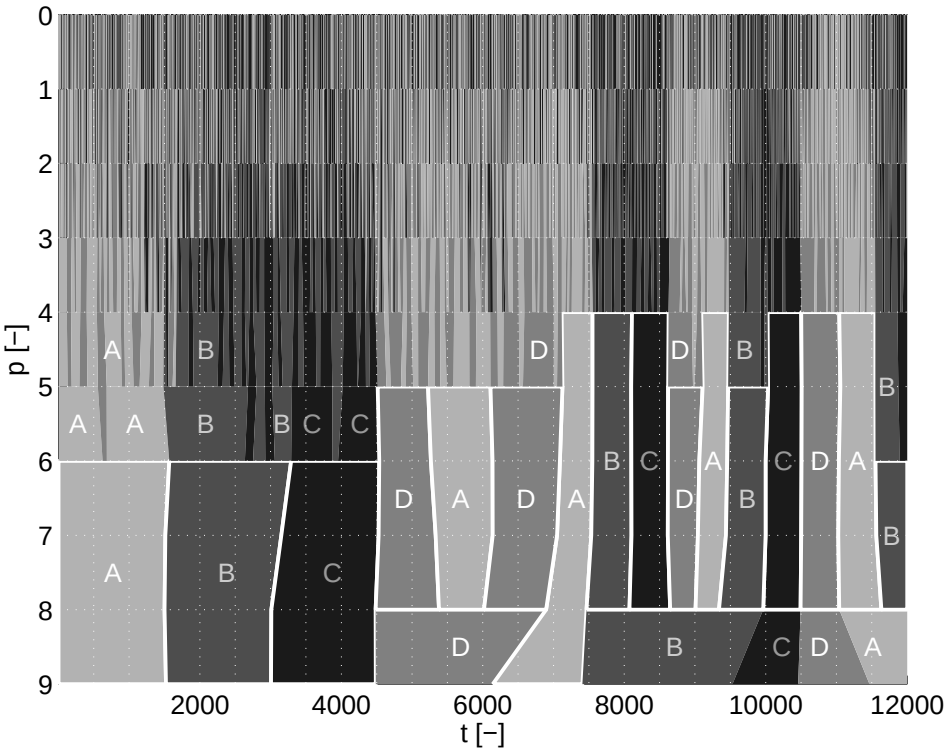


Figure 8.6: Approach 3: wavelet interval tree (triangular) with application of Witkin's stability criterion. Selected episodes are indicated by white contours. Triangular presentation: $\mathsf{ABC(DA)_2(BCDA)_2B}$.

be split or when the most detailed scale of the corresponding monotonic episode is reached, the final triangular representation is obtained: $\mathsf{ABC(DA)_2(BCDA)_2B}$. This representation is equal to the (desired) triangular representation of the noise-free signal.

8.3.1.4 Effect of number of scales on results

As mentioned above, the results were shown only for the analysis with 9 wavelet scales (P). Table 8.1 shows how the results are affected by the choice of P. The maximal number of scales for this series is 13 ($\log_2(N) = 13.55$) for $\delta p = 1$. When using 6 scales or less the monotonic representation is too detailed. This causes all triangular representations to be too detailed as well. When using 7 or more scales, the monotonic representation is correct. In the latter case, approach 1 (accept-all) leads to a too detailed representation whereas approach 2 (single inflection point) always delivers a too coarse representation. The third approach, based on Witkin's stability criterion delivers the correct representation when analyzing 8 scales or more. Given that 6 of 13 theoretical choices deliver the correct result, this approach is considered to be relatively robust.

It needs to be noted that one does not necessarily obtain the correct representation by analyzing the signal up to the coarsest scale that is practically available. To illustrate this, consider the (noise-free) signal simulated as the sum of two sinusoidal waves with distinct oscillation periods, according to the following equation:

$$x_t = z_{t,1} + z_{t,2} \tag{8.2}$$

Table 8.1: Effect of the number of scales (P) in the wavelet decomposition on the accuracy of the qualitative representation of the signal in Figure 8.1 ($\odot$: correct, $\triangle$: too detailed, $\bigtriangledown$: too coarse)

P	monotonic	triangular		
		accept-all	single inflection point	Witkin's criterion
1-6	$\triangle$	$\triangle$	$\triangle$	$\triangle$
7	$\odot$	$\triangle$	$\bigtriangledown$	$\triangle$
8-13	$\odot$	$\triangle$	$\bigtriangledown$	$\odot$

where:

$$z_{t,1} = sin\left(2 \cdot \pi \cdot \frac{t}{\tau_1}\right)$$

$$z_{t,2} = sin\left(2 \cdot \pi \cdot \frac{t}{\tau_2}\right)$$

t : time index (integers from 1 to N)

N : length of the time series (12000)

τ_1, τ_2 : oscillation periods (N, resp. N/200)

The simulated signal is plotted Figure 8.7. If all qualitative details of the signal are aimed for, including the features due to the fast oscillation, then the targeted qualitative representation is $(\mathsf{KL})_{200}\mathsf{K}$ (monotonic) or $(\mathsf{ABCD})_{200}$ (triangular). In Table 8.2, shows how the results for this signal are affected by the choice of P. As can be seen, the qualitative representations are too coarse if P is higher than 8. For all approaches, the representations are then KLK (monotonic) and ABCD (triangular) corresponding to the low-frequency oscillation. For lower numbers of P, the correct representations are found, independent of the approach used. It is to be expected that none of the representations is too complex as (1) no noise is present and (2) no sequences of multiple inflection points are present in the signal.

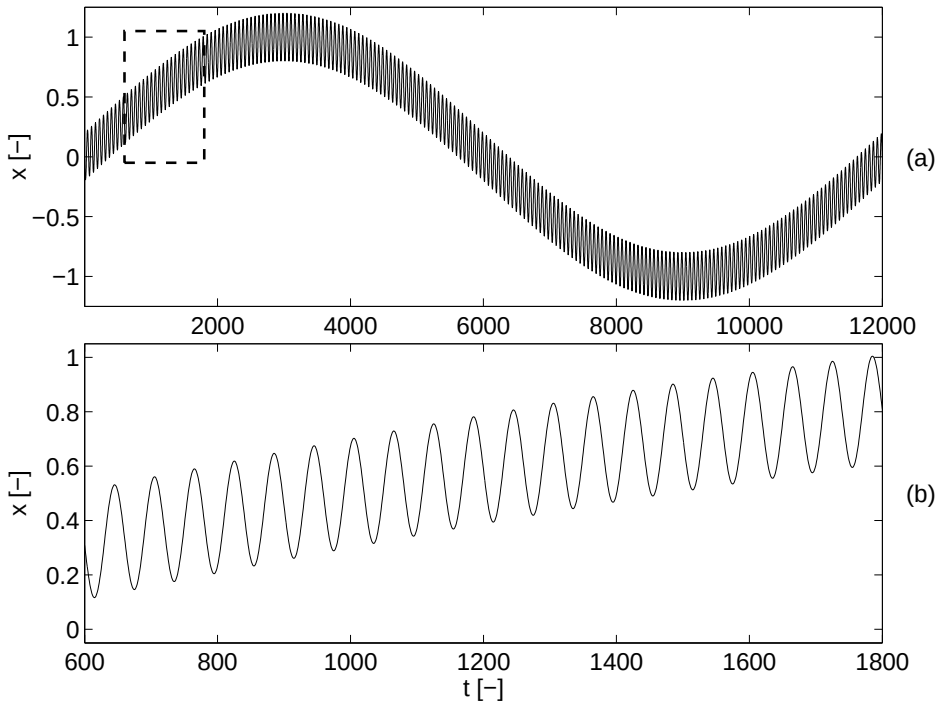


Figure 8.7: Simulated series consisting of two sinusoidal waves according to equation 8.2. (a) complete series, (b) detailed view of the signal in the box in (a).

Table 8.2: Effect of the number of scales (P) in the wavelet decomposition on the accuracy of the qualitative representation of the signal in Figure 8.7 ($\odot$: correct, $\bigtriangledown$: too coarse)

P	monotonic	triangular		
		accept-all	single inflection point	Witkin's criterion
1-8	$\odot$	$\odot$	$\odot$	$\odot$
9-13	$\bigtriangledown$	$\bigtriangledown$	$\bigtriangledown$	$\bigtriangledown$

8.3.2 Real-life series

In Figure 8.8 the studied ORP signal is shown together with the triangular qualitative representations according to each approach given in the previous section. Visual inspection of the series (8.8(a)) indicates that the signal exhibits a downward accelerating trend (primitive C) from minute 0 to approximately minute 31, followed by an upward trend until the end of the series, including four subsegments, approximately from minutes 31 to 33 (accelerating, primitive D), minutes 33 to 77 (decelerating, primitive A), minutes 77 to 91 (accelerating, primitive D) and minutes 91 to 100 (decelerating, primitive A). The qualitative representation is evaluated at 3 time instants (80, 90 and 100), as in a streaming data context which is the ultimate goal of QRT in this dissertation. The generated monotonic presentation confirms the results of the visual inspection discussed above at all evaluations (Figure 8.8(b)).

The accept-all approach for inflection points (Figure 8.8(c)) fails to discriminate true inflection points from noise artefacts, leading to a too complex presentation at each evaluation. The second (original) approach (Figure 8.8(d)) delivers a representation that is visually consistent with the data at the first evaluation (minute 80), i.e. a single inflection point (D-A transition) is assessed at 33 minutes. At the second evaluation (minute 90), a single inflection point is assessed at minute 89. The visually affirmed inflection points at minute 33 (D-A transition) and at minute 77 (A-D transition) are now ignored. As a consequence, the behaviour of the series between minutes 31 and 90 is now considered to be of type D, i.e. rising with increasing speed, in contrast to the visual inspection above. At the third evaluation, the location in time of the assessed inflection point is updated to minute 91, while the explained discrepancy between the visually affirmed and assessed qualitative behaviour remains.

The third (new) approach (Figure 8.8(e)) delivers the same qualitative representation at the first evaluation as the second approach, except for the location of the inflection point in time (minute 32). Detailed inspection can however not confirm the one or the other result. During the second evaluation (minute 80) an inflection point is added at minute 75 and another inflection point is added in the third evaluation at minute 91. This is consistent with the earlier visual inspection of the series. Even more, the assessed qualitative behaviour at any evaluation point is consistent with the assessed behaviour at earlier evaluations. Indeed, inflection points that were evaluated as relevant at a prior evaluation are not removed from the qualitative presentation at a later stage, in contrast to the first and second approach.

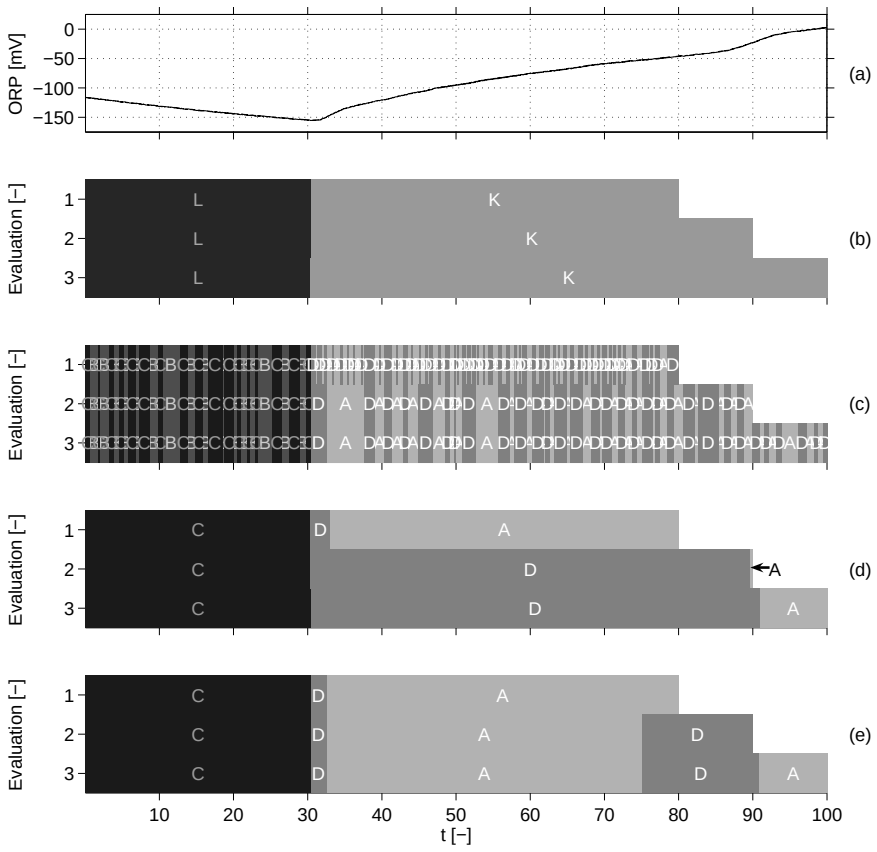


Figure 8.8: Assessment of the qualitative representation of a real-life ORP signal at 80, 90 and 100 minutes (evaluation 1, 2 and 3): (a) studied signal, (b) monotonic presentation and triangular presentations by means of (c) accept-all approach, (d) original single inflection point approach and (e) new approach based repeated application of Witkin's stability criterion.

8.4 Discussion

Three approaches to assess relevant inflection points in the framework of qualitative representation of trends by means of cubic spline wavelet filtering are applied to both a simulated and real-life time series. An interesting challenge for both series is the appearance of 3 contiguous inflection points that should be detected as such. The accept-all approach, in which all inflection points between two selected extrema are automatically selected, leads to the acceptance of many irrelevant inflection points. The single inflection point approach, i.e. the approach of Bakshi and Stephanopoulos (1994), tackles this problem by replacing contiguous sets of inflection points by a single inflection point. While the unnecessary complex results of the first approach are overcome by doing so, contiguous relevant inflection points are no longer detected individually. By applying Witkin's stability criterion for inflection points as proposed here, the correct solution has shown to be found. Given that this approach avoids unnecessary complexity in the resulting representation while allowing the assessment of contiguous relevant inflection points, this approach delivers improved results for qualitative representation of trends when compared to the method of Bakshi and Stephanopoulos (1994). It was also shown that a proper selection of the number of scales is necessary to obtain correct results.

Importantly, the first and second approaches were shown to deliver incorrect results for the 2 case studies independent of the number of analyzed scales. In addition, by using Witkin's stability criterion both for extremum and inflection point assessment, no additional burden is implied in terms of understanding or computational implementation compared to the original method.

Besides the consequences for representation of historical trends, important opportunities may arise from the proposed modification of the method to its on-line use. In the original approach, relevant inflection points are assessed by replacing contiguous sets of inflection points by a single inflection point, determined by the maximal or minimal value of the detail signal. This means that the selection of an inflection point can only be completed when the corresponding monotonic episode is finished. Indeed, the complete series of contiguous inflection points has to be known before the relevant inflection point can be determined. In the context of streaming data, the assessment of relevant inflection points is thus delayed. This delay may exceed the delay warranted by border distortion. The third approach presented in this paper does not lead to such a delay, detects inflection points consistently and allows inflection point assessment in an on-line context.

8.5 Conclusions

In this chapter, an existing method for qualitative representation of trends has been improved in view of the identification of inflection points. Three approaches for assessment of inflection points were evaluated on both a simulated time series and a real-life series. The first approach was shown to lead to a too complex representation of the series. The second approach, being the original one, tackles this problem successfully but leads to a too simple representation when contiguous sets of relevant inflection points appear in the series. This approach also exhibits practical problems for streaming data. A third and new approach was devised by applying Witkin's stability criterion to triangular episodes. It was shown that the latter approach is the only approach that allows the assessment of contiguous inflection points while avoiding unnecessary complexity in the results. As such, this new approach exhibits a considerable improvement of the original method. In addition, the approach is relatively robust to choice of the number of analyzed scales, does not result in an increased implementation burden and exhibits no practical problems with streaming data as opposed to the original method.

In Section 3.5, the original method by Bakshi and Stephanopoulos (1994) was chosen. One of the arguments for this was that the no jump changes were generally expected in the series to be analyzed. Indeed, if jumps or steps are expected, the method cannot identify them appropriately. It was shown that jumps are identified as inflection points. In order to further improve the method presented, an appropriate and separate handling of jumps is desired. Also, jump changes in the 1^{st} and/or 2^{nd} derivatives may be aimed for. Developing a method that is able to identify jump changes will likely result in a more generic method, effectively enabling the analysis of smooth and non-smooth data without or with limited prior knowledge. To this end, F-tests as used in the method of Dash et al. (2004a) to detect jump changes may be incorporated into the method adopted and adjusted in this work. Alternatively, the method by Dash et al. (2004a), which is able to handle discontinuities, may be improved for inflection point detection by fitting polynomials up to 3^{rd} order (instead of 2^{nd} order) unless discontinuities are observed (by means of F-tests).

Chapter 9

Qualitative Representation of Trends for time series data mining of urban water network flow measurements

This chapter is based on:

Villez, K., Pelletier, G., Rosén, C., Anctil, F., Duchesne, C., Vanrolleghem, P.A. (2007). Comparison of two wavelet-based tools for data mining of urban water networks time series. Wat. Sci. Technol., 56(6), 57-64.

9.1 Introduction

In this chapter, standard wavelet power spectrum analysis (see Section 3.5.2) and the new method for Qualitative Representation of Trends (QRT) (see Chapter 8) are evaluated and compared for analysis of time series of flow measurements taken

from an urban drinking water network. While classical wavelet spectrum analysis indicates where important features in a series are situated in the time and frequency domain, this does not provide information on the type or shape of the identified features. It will be shown that explicit information regarding the first and second order behaviour of a series can be extracted and leads to additional relevant information about the studied series.

9.2 Selected data

Drinking water is supplied to a residential neighbourhood of 20.500 inhabitants in the Quebec City area from five groups of wells distributing water to three pressure zones with average water use of 1.050, 4.050 and 600 m^3/d in the lower, intermediate and high pressure zone respectively. Of these groups, four are located in the lower pressure zone, while one is located in the intermediate zone. All groups of wells have a proper local distribution network, but are all connected to a booster station which fills a storage tank (6.800 m^3) located in the high pressure zone. During periods of low water demand, all excess well discharges are pumped to the latter tank. During periods of high water demand, the water tank supplies peak demand to all three zones. The data used in this study are flow measurements from the outlet of this storage tank from November 15th, 2002 to February 1st, 2003 at one minute intervals. The daily water demand pattern from such a residential neighbourhood is expected to show two peaks: one in the morning (breakfast, showers) and one in the evening (supper, dishwashing, clothes washing, baths/showers) altered with low flow rate periods in-between. Night time flow rates are the lowest.

9.3 Applied methods

9.3.1 Method 1: Wavelet power spectrum

The first method used in this chapter is wavelet power spectrum analysis as described by Torrence and Compo (1998). Details on this method can be found in 3.5.2. The Morlet wavelet was used with its shape parameter, the nondimensional frequency, ω_o, set to 6. Equation 3.78 describes this wavelet in the frequency domain. Its shapes in the time and frequency domain are illustrated in Figure 3.27.

The parameters that specify the wavelet decomposition, s_o, P and δp, were set to 2, 14 and 0.125 respectively so that the studied scales ranged from 2 times the measuring interval (period = 2 minutes or approx. 0.0014 days) to 2^{14} times the measuring interval (period = 32768 minutes or approx. 23 days) with intervals of 0.125 on a $\log_2$ scale (8 scales per octave or frequency halving). As the wavelet scale is not necessarily the same as its equivalent Fourier period, the results have been adjusted for this discrepancy to allow a correct interpretation, following the solution of Torrence and Compo (1998). Future references in this chapter to the term 'period' indicate the equivalent Fourier period, while scale remains the term for the wavelet period or scale. Torrence and Compo (1998) also provide the so-called cone of influence which defines the region in the obtained spectrum out of which edge effects distort the wavelet power spectra in such a way that interpretation becomes ambiguous.

Following the derivation of the wavelet coefficients, $w(t, s)$, the wavelet powers, $|w(t, s)|^2$, are calculated and further normalized by the overall variance of the time series, $|w(k, s)|^2/s_x^2$. The given values are then a measure of the power relative to the equivalent power of a white noise process with the same overall variance. As such, wavelet spectra become straightforward tools to assess non-stationarity, amplitude changes and dominant frequencies within time series.

9.3.2 Method 2: Analysis of qualitative trends at different scales in wavelet decomposition

The second method is the method described in 3.5.5 with adoption of the solution for inflection point detection as proposed in 8. For this analysis, s_o, P and δp, were set to 2, 8 and 1. The wavelet scales thus ranged from 2 times the sampling interval (period = 2 minutes or approx. 0.0014 days) to 2^8 times the sampling interval (period = 512 minutes or approx. 0.36 days)

9.4 Results

9.4.1 Method 1: Wavelet power spectrum

In Figure 9.1, a contour plot of the relative power spectra is shown for the data collected between November 15th, 2002 and February 1st, 2003. Over the whole period, high powers are observed in bands at periods of 1 day and ½ day respectively, suggesting regular cyclic behaviour. This concurs with the apparition of two peaks in water demand during each day. The fact that the peaks do not occur at an interval of exactly twelve hours leads to dominant powers at two distinct frequency bands. The wavelet power is generally lower at periods shorter than ½ days, but remarkably, daily peaks in the wavelet powers are seen for a couple of hours during the day. A closer look at the data (not shown) revealed that a sharp increase of

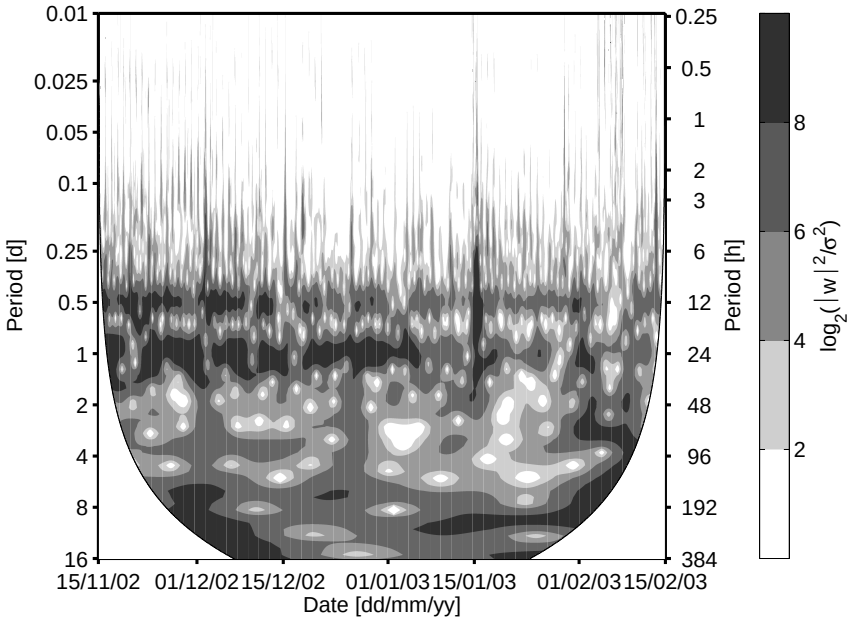


Figure 9.1: Contour plot of the wavelet power spectra from November 15th, 2005 to February 1st, 2006 and for periods from 0.01 to 16 days. The shading legend is shown at the right hand side. Wavelet powers affected by edge effects (i.e. outside the cone of influence) are masked.

the water flow typically occurs during the morning, indeed being a highly dynamic event during a short period during the day. At longer periods, patterns are rather irregular, suggesting non-stationary behaviour in these scales. Powers are rather low between periods of 1 day and 7 days. For larger periods, powers seem to be larger but larger time windows should be used to confirm this result.

9.4.2 Method 2: Analysis of qualitative trends at different scales in wavelet decomposition

With the first method, an (expected) dominant cycle of 1 day is observed in the data. Based on this expected behaviour and the latter confirmation, the data series was segmented into segments of 1 day, each starting and ending at midnight. A separate qualitative representation is assessed for each of these segments. To reduce edge effects during filtering, the data series were padded with anti-symmetric data at the left ($x_{t_0-t} = -x_{t_0+t}$) and right side ($x_{t_{end}+t} = -x_{t_{end}-t}$) of each section. In Figure 9.2 the monotonic presentations (episodes with constant sign of the first derivative) are displayed for each day in the studied period. Each horizontal bar represents the triangular episodes in a single day. The left and right ends of each rectangle in this bar are the start and end points of the monotonic episodes. It is observed that a larger part of the days have the following pattern: LKLKL, which means two minima and maxima are observed for these days. The maxima correspond with a peaking water demand in the morning and evening while minima lie in between. For a few days, the second maximum does not occur or is too weak to be accepted into the presentation. Interestingly, these days occur between 21/12 and 6/1, corresponding to Christmas Holidays.

A significant part of the studied time series exhibit frequent changes in the qualitative behaviour (e.g. 17/12). Visual exploration (not shown) proved that this behaviour is correctly identified (they are not stemming from an erroneous acceptance of noise as relevant features) and showed that these patterns exhibited many step changes. The causes of these (abnormal) step changes could however not be unambiguously linked to sensor failure (incorrect measurement) or control failure (incorrect action).

In addition to the qualitative information (chronology of up/down episodes), the location in time of the observed maxima and minima was shown to be relevant as well. It can be seen for instance that the first maximum occurs later on 1/12 and 2/12 when compared to the days just before and after. Remarkably, these two

days are a Saturday and a Sunday. The same delay of the first peak occurs for all the other weekend days in the studied period. Such a delay is not observed for the second peak (evening). In addition to the weekend-related delay of the first peak in the day, a similar effect is seen for all the days between 21/12 and 6/1, corresponding to Christmas Holidays. The observed weekend-effect thus also applies to these holidays. It can thus be concluded that the water flow data exhibit a distinct pattern in weekend days and holidays that reflects the behaviour of the city's population (Campos and von Sperling, 1996).

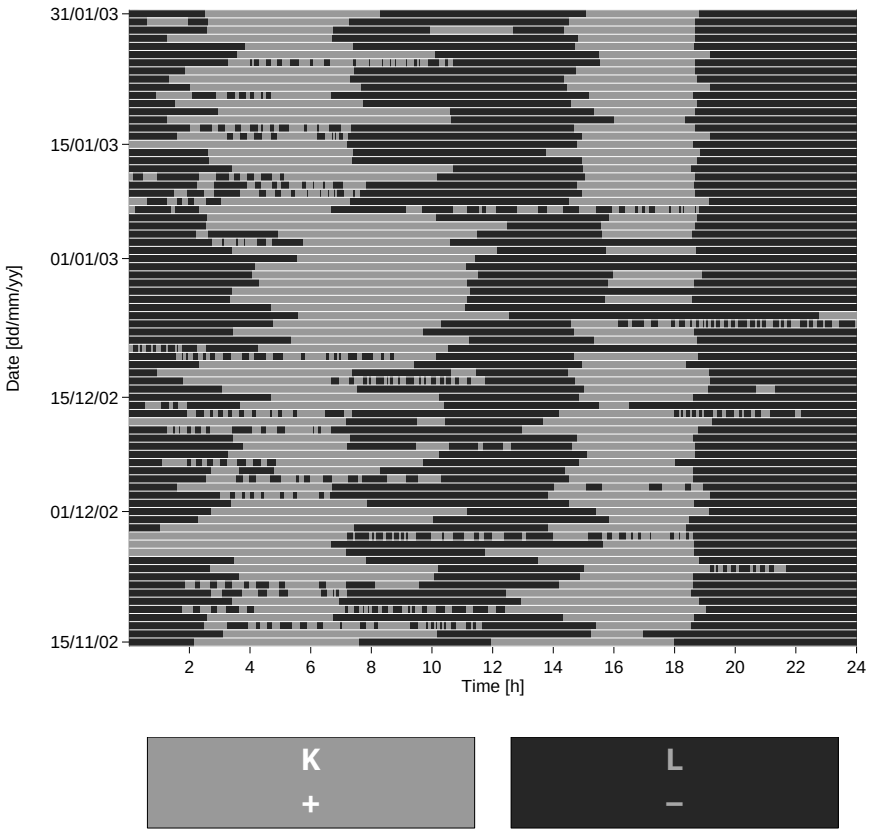


Figure 9.2: Qualitative representations of time series by means of monotonic primitives. Each horizontal bar shows the representations of the data of a single day. The shading legend is shown below the graph.

In Figure 9.3, the triangular presentations (episodes with constant sign of first and second derivatives) are given. In addition to the extrema, inflection points are thus shown as well in this graph. A single inflection point is observed between two extrema during the larger part of the data set, indicating that the acceleration of the flow is typically monotonically increasing or decreasing within each monotonic episode. In a few cases, multiple inflection points occur within one monotonic

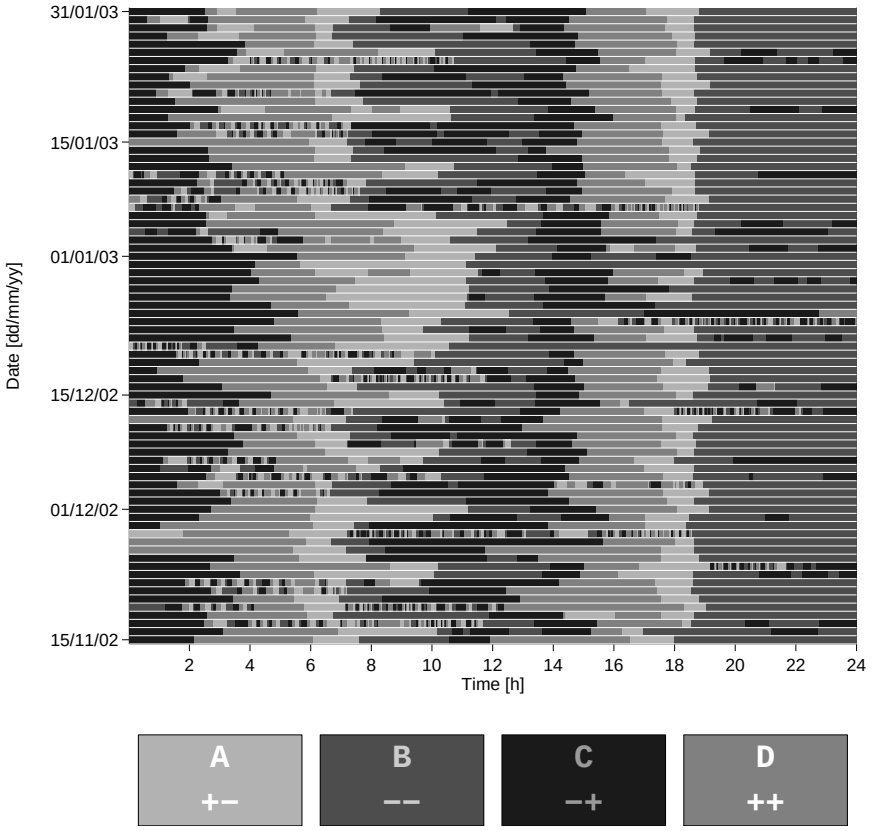


Figure 9.3: Qualitative representations of time series by means of triangular primitives. Each horizontal bar shows the representations of the data of a single day. The shading legend below the graph indicates the sign of the first and second derivative (resp. the left and right symbol for each primitive.

episode. This happens almost exclusively at night, after the second maximum (e.g. see 01/12). In other words, after the second peak demand, the speed at which the demand decreases shows maxima and minima (inflection points are the extrema of the 1st derivative) for some days.

9.5 Discussion

Two wavelet-based methods for mining of time series have been applied to a time series of hydraulic data. In the first method, the signal is transformed into a power measure over time and frequency. As such, relevant dynamics were observed primarily in the scale of days and ½ days. Coarser scales seemed to be characterized by non-cyclic behaviour, while in more detailed scales, regular peaks in power were observed, suggesting highly dynamic events during a limited time-window during most days.

In the second method, the cubic spline wavelet is used to obtain a qualitative representation of the data. As such, relevant maxima, minima and inflection points are identified to define the qualitative representation. By means of this method, typical maxima in water demand in the morning and evening were detected. In addition, the location in time was shown to be depending on the type of day (working day, weekend day and holiday). Also, inflection points were observed during some nights, indicating moments of minimal decrease of the water demand during some nights. Clearly, the qualitative presentation of the data delivers interesting information regarding the behaviour of a city's population, which is not available from wavelet decomposition only. The first method confirmed the presence of a daily cycle in the studied time series, hereby leading to a window definition for the qualitative representation. Wavelet power spectrum analysis thus functions as an excellent pre-analysis step.

9.6 Conclusions

Wavelet power analysis, allowing simultaneous analysis of series in the frequency and the time domain, and the method for QRT based on the cubic spline wavelet and improved in Chapter 8 was applied for analysis of a time series for which little a priori knowledge was available. It was shown that the first method, wavelet power analysis, indicates the location of major cycles and features in frequency and time, but not their type or shape. Since the QRT method does not aim solely at the analysis of the amplitude of the wavelet coefficients but also of their sign (1^{st} order behaviour) and changes over time (2nd order behaviour), more detailed information could be extracted by means of qualitative description of trends. In the case presented, such information may be a helpful tool for the design of measurement campaigns, modelling of the system and on-line detection of system failures (e.g. leaks).

In the case presented, wavelet spectrum analysis served as a priori analysis for QRT, i.e. defining a meaningful time window for qualitative analysis. The reverse may equally share potential in the context of data mining. Consider for example that the cubic spline wavelet method allows to reconstruct the time series from its wavelet coefficients. By selecting only the coefficients that are assessed to be qualitatively relevant during the reconstruction steps, a qualitative and adaptive (i.e. the cutoff frequency changes over time) filtering of the time series results. Spectral analysis of the reconstructed *–qualitatively cleaned-up–* time series is expected to deliver wavelet spectra that can be interpreted much easier, i.e. less blurred by noise or irrelevant artefacts in the series. Spectral analysis and/or qualitative analysis of residuals, i.e. the differences between the original and reconstructed data, may equally be valid for study of detailed patterns in (time) series.

Chapter 10

Qualitative Representation of Trends for control of a Sequencing Batch Reactor

This chapter is based on:

Villez, K., Rosén, C., Anctil, F., Duchesne, C., Vanrolleghem, P.A. (2007). Qualitative representation of trends: an alternative approach to on-line process diagnosis and control of SBR's for nutrient removal. In: Proceedings of the 3rd International IWA Conference on Automation in Water Quality Monitoring (AutMoNet2007), Ghent, Belgium, September 5–7, 2007, appeared on CD-ROM.

The potential of QRT for diagnosis and control is evaluated in this chapter. Application of QRT techniques are not widespread in the field of process monitoring and diagnosis. It will however be shown that the assessed qualitative description of trends can be coupled easily with existing process knowledge and does not demand the end user to understand the underlying method in detail, in contrast to, for instance, multivariate techniques in Statistical Process Control. The assessed links can be integrated straightforwardly into the framework of supervisory con-

trol systems by means of look-up tables, expert systems or case-based reasoning frameworks. This in turn allows the design of a supervisory control system leading to fully automated control actions. The technique is illustrated by an application to a pilot-scale SBR.

10.1 Selected data

The data set used consists of the pH time series during the aerobic phase of 100 complete batches from the studied pilot-scale SBR setup and were collected at the end of 2006 (Nov. 22 – Dec. 20). The analyzed profiles start at 10 minutes in the aerobic phase until its end. By doing so, the often complex dynamics of the start of the aeration phase, which impeded straightforward interpretation of the results. Typical time series and their interpretation are shown further in the text.

10.2 Applied method

The applied method for QRT is the one described in 3.5.5 as only the monotonic primitives are used here as a basis for qualitative presentation. For this analysis, s_o, P and δp, were set to 2, 9 and 1. The analyzed wavelet scales thus ranged from 2 times the sampling interval (period = 4 seconds or approx. 0.07 minutes) to 2^{10} times the sampling interval (period = 2048 seconds or approx. 34 minutes).

10.3 Results

10.3.1 Analysis

A qualitative representation of each of the selected pH trajectories in the first aerobic phase of the system under study was obtained. In Figure 10.1, each horizontal bar corresponds to the qualitative representation of a single trajectory. Note that only the monotonic primitives (upward/downward) were used for this study. For example, the analyzed trajectories of batches 13–17 are represented as a KL (up-

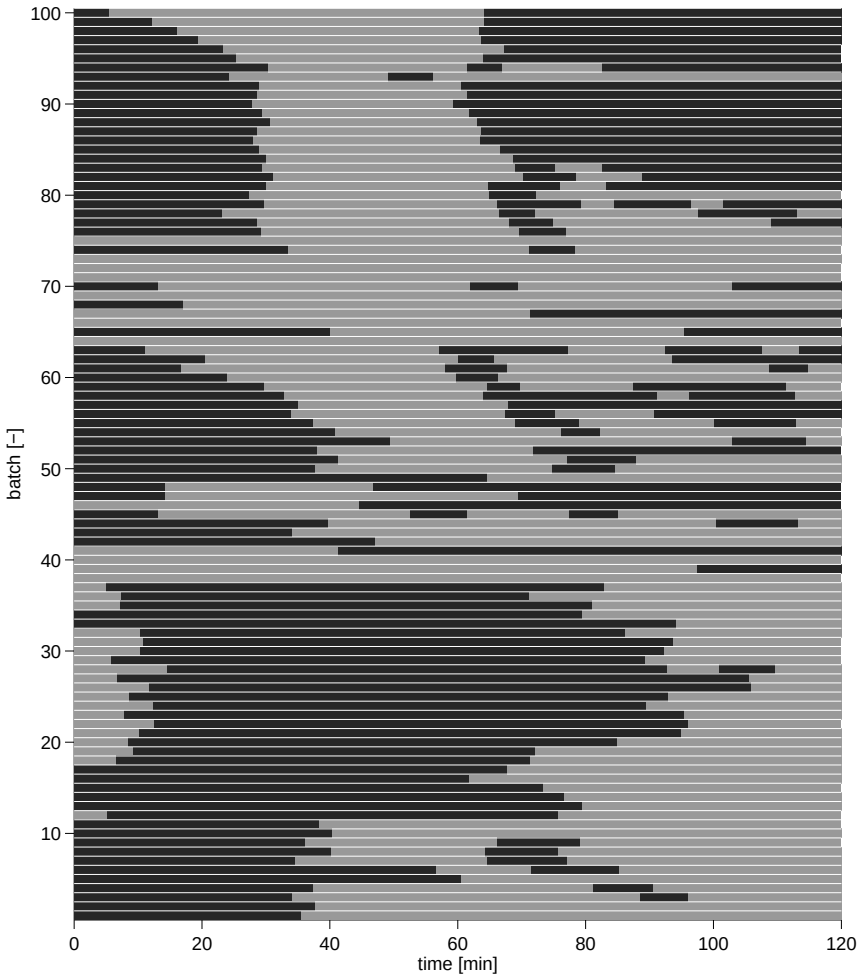


Figure 10.1: Qualitative representations of 100 pH trajectories in the 130 minute aerobic phase. Dark shading indicates upward trends, light shading indicates downward trends.

ward/downward) sequences. Batches 7–9 exhibit a KLKL sequence. By simple listing of the qualitative representations for all batches in the study, a so-called *dictionary* is automatically generated, in which a meaning is yet to be assigned to each *word*. In this case, a 10-word dictionary results. In Table 2, the numbers of

batches (cluster size) for each observed type of qualitative behaviour (cluster label) are given. Interestingly, the 4 most populated clusters (40% of the assessed representations) represent 70% of the batches. In addition, it is observed that the corresponding qualitative behaviours are relatively simple in nature (all sequences exhibit 4 characters at most). A major part of the batches thus corresponds to a limited set of relatively simple qualitative descriptions.

For each of the most popular descriptions, one pH trajectory and its corresponding qualitative representation is shown in Figure 10.2. Figure 10.2(a) shows a pH trajectory for which a KL description results. The upward trend in the beginning is explained as a net positive effect of CO_2-stripping over nitrification. From approximately 37 minutes, a decreasing trend is observed, indicating that the acidifying effect of nitrification has become larger than the effect of CO_2-stripping. This trend continues until the end of the aerobic phase. In Figure 10.2(b) a pH trajectory that is translated into a KLK description is shown. The first two episodes (upward and downward) are indicating the same effects as for the previous example. However, the pH increases again after a period of decrease, namely from approximately 109 minutes. This indicates that the conversion of ammonia to nitrite is complete and that the remaining active process is CO2-stripping only. A KLKL description is found for the pH trajectory shown in Figure 10.2(c). The explanation of this behaviour is the same for the first 3 episodes (upward, downward and upward) as for the previous example. At approximately minute 104, a decreasing trend sets in however. Due to a (delayed) reduction in the flow-rate (by means of PID control of the oxygen level), the CO_2-stripping now becomes minimal at this point and a net –though minimal– decrease of the pH results. Figure 10.2(d) shows a pH trajectory described as a LKL sequence. This pattern is interpreted in the same way as the KLKL description, except for the absence of an initial upward trend in the latter description. A dominating effect of CO_2–stripping at the beginning of the aerobic phase is indeed not observed and is the result of a minimal loading of the system (as the influent is the major contributor to the CO_2 present in the system).

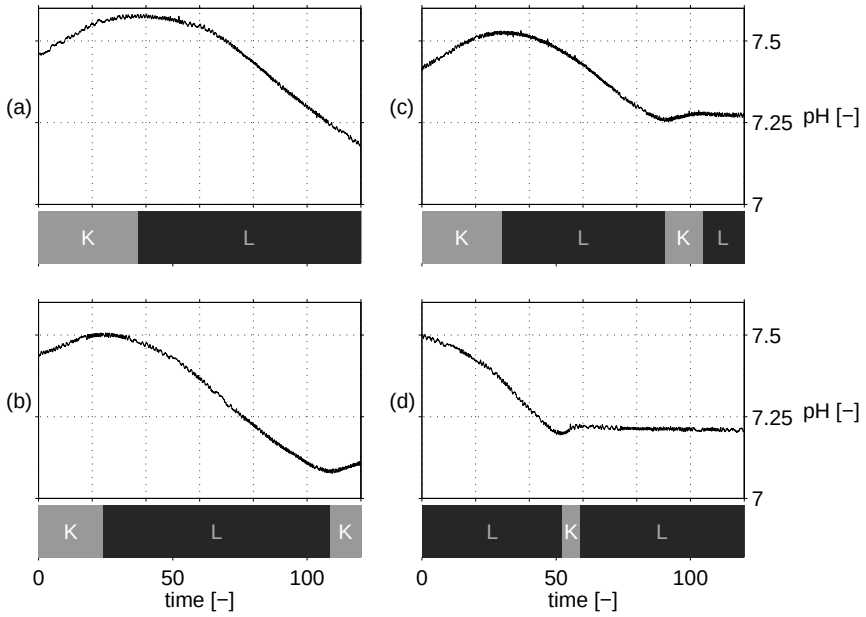


Figure 10.2: Examplary pH trajectories and corresponding qualitative representations.

10.3.2 Diagnosis

Let us now try to provide diagnostic information to each of the clusters (i.e. adding a meaning to the entries in the dictionary). According to the operators, a KL presentation (e.g. batches 17–19) corresponds to a high load situation with an incomplete aerobic phase (incomplete nitrification, i.e. the pH did not stop decreasing). An KLKL presentation (e.g. batches 7–9) is related to a completed aerobic phase (complete nitrification) under high load conditions (i.e. after nitrification is completed the pH starts to rise because of CO_2 stripping). Interestingly, a high load is diagnosed by the operators if the pH trajectory starts with an upward trend (K), while downward trends (L) at the start of the aerobic phase were related to low load conditions. The underlying reasoning to this diagnosis is that the influent is the major contributor to the amount of CO_2 stripped during the aerobic phase. A decreasing trend of the pH indicates a low CO_2 concentration in the system, hence indicating a low load to the system. It is thus possible to diagnose the system under study on the basis of qualitative representations of the pH trajectories.

Table 10.1: Qualitative representations, associated diagnostics and occurrence

pattern		diagnostic information		occurrence (%)
observation	dictionary entry	load	completion	
–	K	1 (high)	2 (incomplete)	0
KL	KL	1	2	16
KLK	KLK	1	1 (complete)	20
KLKL	KLKL	1	1	16
KLKLK	KLKLK...	1	0 (unknown)	8
KLKLKL		1	0	6
KLKLKLK		1	0	2
L	L	2 (low)	2	9
LK	LK	2	1	4
LKL	LKL	2	1	18
LKLKL	LKLK...	2	0	1

In Table 10.1, the diagnosis given by the operators is given for each observed representation together with the frequency at which the observed representations occurred. This table functions as the dictionary of the qualitative representations of pH trajectories. As discussed above, an accurate diagnosis with respect to the load is possible for all observed behaviours. For some qualitative representations, no unambiguous diagnostics concerning the completion of the biological processes could be assessed. This was either due to the operator not being familiar with the observed pattern or due to the fact that different diagnoses (complete and incomplete) were possible within the set of batches with the same qualitative behaviour. Still, a complete diagnosis was possible for 83% of the batches.

10.3.3 Incorporation of knowledge

The framework of qualitative representation of trends allows one to also consider imaginative representations of trajectories and the assessment of corresponding diagnostics and control actions, even if data of such sequences are not available. Such an injection of knowledge into this data-driven methodology reveals that the coupling of deductive and inductive methods is practically feasible. In this study,

relatively simple sequences, such as L and LKLK were not observed within the one month of SBR monitoring (see Table 10.1). However, the operators of the studied system are able to complete the diagnosis dictionary with unencountered sequences without the necessity of factual observations of these. Table 10.1 gives the completed dictionary for the intended diagnosis and control system. Note that complex sequences (starting with KLKLK or KLKL) were grouped into a single entry in the table so that an entry exists in the table for all possible patterns that may ever appear.

10.3.4 Control

The proposed methodology for diagnosis of a batch can be combined with the selection of an appropriate control action. This combination leads to a generic diagnosis and control system based on the assessment of qualitative representation of trends (Figure 10.3). The raw signal is processed to obtain the qualitative representation as proposed. The resulting qualitative representation is then looked up in the developed dictionary (as defined in Table 10.1) which relates the qualitative representations with the corresponding diagnoses. The descision on the control action to exerted can be made straightforwardly by means of a table in which the entries are defined as a possible combination of premises and a corresponding (set of) control

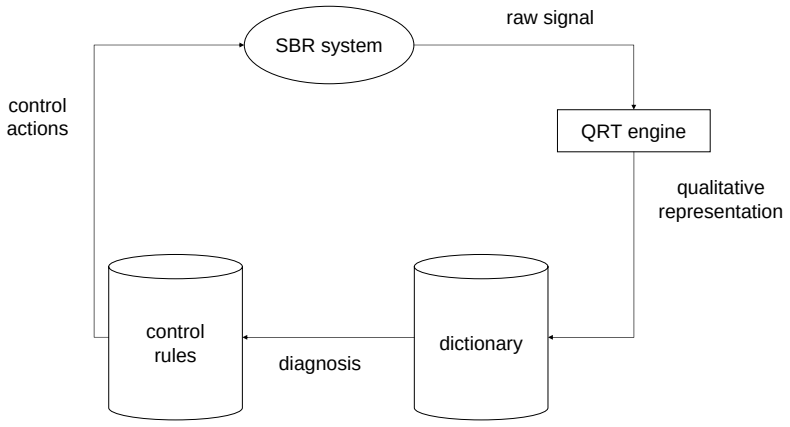


Figure 10.3: Integrated control scheme.

action(s). While this appears feasible for simple control problems, more complex problems may require the use of rule-based systems or case-based reasoning.

To start up and update any of these inferencing systems, operators can assess the relations between qualitative representations and diagnostics in a straightforward manner as qualitative presentations of trends are often concurring with their mental models in many cases. It is especially interesting that (1) operators do not need explicit knowledge on the mathematical details of the applied technique to make such a control system work and (2) the implementation of the technique does not require explicit process insight. These aspects stand in contrast with the use of statistical models in process monitoring and diagnosis, e.g. PCA. Given the complex relation between the outcomes of such models (scores, statistics, cluster membership) and the original data, interpretation of the model outcomes and diagnostics is often difficult and requires a good understanding of the modelling technique and of the process. Hence, qualitative presentations of trends offer a straightforward way to avoid such difficulties.

Given the assessed diagnostics over the 1-month historical data set, the operators of the studied system were asked for an appropriate action to be taken in order to optimise operation in terms of effluent quality and plant economy while safeguarding acceptable operation. In Table 3, these are presented together with the percentage of batches for which they would be taken. As can be seen, the operators would increase the load to the system under low load conditions, regardless of the explicit assessment of completion (i.e. all ammonia is oxidized). This is not so surprising since none of the low load observed conditions corresponded to the diagnosis of an incomplete process. For the high load conditions a more refined set of actions was suggested by the operators. In case the biological processes

Table 10.2: Diagnostics, associated control actions and occurrence

diagnostic information		action description	occurrence (%)
load	completion		
1 (high)	0 (unknown)	no change, call operator	16
1	1 (complete)	reduce air supply	36
1	2 (incomplete)	increase air supply	16
2 (low)	0/1	increase load, equal air supply	23
2	2	increase air supply	9

are finished, operators would reduce the aerobic phase length, while they would extend the aerobic phase length if the biological processes are incomplete in the past aerobic phase. In case the load is high and no accurate assessment of the process completion is available, the operators suggested not changing the operation without further analysis. As a result, an automated adjustment of the load and aerobic phase length is possible for 84% of the batches. Put otherwise, only 16% of the batches need to be diagnosed by means of a more detailed investigation. We note here that the location in time of the identified characteristics (e.g. extrema) were not included as criteria for diagnosis in this preliminary study.

10.3.5 From raw data to supervisory control: complete procedure

In Table 10.3 the complete procedure by which the supervisory controller can be established is given together with the expected interaction with the operator in each step. Quite interesting for implementation of such a supervisory controller is that no interaction is required in the complex step involving the qualitative representation of trends. In other words, the end user does not need to understand the mathematical and computational aspects of the underlying technique.

Table 10.3: From historical data to supervisory control: procedure

step	description	operator interaction
1	data selection and/or screening	optional
2	generate qualitative representation of trends	no
3	generate/update dictionary	no
4	add diagnostics to (new) entries	yes
5	complete dictionary with unobserved entries	no
6	add diagnostics to unobserved entries	yes
7	link control actions with diagnostics	yes

10.4 Discussion

The usefulness of qualitative representations for diagnosis and control of an SBR for nutrient removal was evaluated. Even though the trends of only one variable (pH) and only one reaction phase of the SBR cycle were studied, it could be shown that for a major part of the batches an accurate diagnosis was possible on the basis of the presented methodology. Control actions could be associated with all possible diagnoses.

As every part of the running system can be automated, a closed-loop diagnosis and control system for the SBR plant under study is possible. Using temporal information regarding the identified episodes (start time, end time, time length) in the inferencing steps may however further improve the performance of the intended control system. It is important to note that the behaviour of the pH variable in the first aerobic phase of the system was well understood, which is believed to be essential to the straightforward development of the proposed control system from data-driven process analysis.

Future studies may focus on or include sensor data and phases for which the understanding is less complete to evaluate whether qualitative representation of trends allows (1) the retrieval of new knowledge about the biological system and (2) the assessment of diagnosis and control strategies in situations where only limited process knowledge is available. While the qualitative representation of trends is essentially an inductive method, the qualitative nature of its results can be coupled easily with deductive approaches to diagnosis (e.g. expert systems, case-based reasoning). Behaviours imagined by experts but not part of the studied data set may then take part in the premises of certain rules in an expert system or may define a set of artificial cases in a case-based reasoning framework. The possibility to diagnose future faults that show little similarities with faults in historical data sets and the straightforward link between the outcome of the technique and existing process knowledge, are considered major strengths in comparison with quantitative methods, e.g. PCA, which generally do not provide this opportunity.

10.5 Conclusions

The previously reviewed method for QRT was also applied to the SBR system studied throughout this dissertation. In a preliminary design of a targeted control system, the QRT method was coupled with tables to make inferences on the system status, possible problems and to be exerted control actions. The control system aims at the control of both the load of the system and the length of the aeration phase. Most importantly, the design stage of the control system proved to be straightforward as the outcomes of the QRT methods, the resulting *words*, could easily be coupled with available knowledge on the qualitative behaviour of the time series under different conditions by means of a *dictionary*. In addition, the dictionary could be completed for words not observed within the analyzed period.

Part V

Conclusions and perspectives

Chapter 11

Conclusions and Perspectives

If you do not ask the right questions,
you do not get the right answers.
A question asked in the right way
often points to its own answer.
Asking questions is the A-B-C of diagnosis.
Only the inquiring mind solves problems.

Edward Hodnett.

In this thesis, data analysis techniques have been developed, tested and evaluated for the purpose of monitoring, diagnosis and control of cyclic processes. To this end, a pilot-scale SBR was used as a case study throughout the major parts of the presented work. The reported techniques were essentially data-driven, i.e. deep knowledge or mechanistic understanding of the analyzed systems was not required for the automated steps in data analysis. Two so far largely separated concepts were used as a basis for data-driven analysis. The first concept is based on Principal Component Analysis (PCA). Given that SBR process data are of a three-dimensional nature (i.e. batch number, time-in-batch and recorded variable), the conventional Multi-way PCA (MPCA) extension was chosen as an effective way to

enable PCA modelling. As a second, also data-driven framework for data analysis, qualitative representation of trends (QRT) has been used. In what follows, the most important conclusions derived of this dissertation as well as perspectives, including improvements of existing methods and newly proposed methods are presented.

11.1 Design of real-life biological experimentation systems for development and validation of strategies for monitoring, diagnosis and control

A pilot-scale Sequencing Batch Reactor system for biological nutrient removal was used as a study object throughout the larger part of this dissertation. The data sets derived from the SBR system proved to be challenging for state-of-the-art techniques in monitoring, diagnosis and control. In this respect, the collection of data from the experimental system may be considered to have been fruitful.

Many of the reported faults proved to be of a technical, rather than of a biological nature. However, originally the biological aspects of wastewater treatment plants, e.g. the evolving nature of microbial communities, were identified as challenges for monitoring, diagnosis and control. Due to occurrence of faults related to the physical parts of the system (pumps, aeration system, cooling system), proper identification of expected natural or common-cause evolving behaviour of the biological part of the studied process was impeded.

In view of future research, opportunities for better design of the physical parts of the system have been proposed in Chapter 4. Suggestions included (1) the acquirement of sensors better suited for the difficult conditions met in wastewater, (2) the use of more robust hydraulic hardware such as solid tubes, solid state valves and avoidance of peristaltic pumps and (3) inclusion of additional measurements in the system design. The latter suggestion was made in view of the construction of balance checks, which may turn the discrimination between physical and biological faults practically feasible. The latter suggestion is considered of special importance in view of the assessment of the performance of monitoring and diagnosis strategies that are designed or aimed at biological faults in particular.

11.2 Multivariate approaches to monitoring, diagnosis and control

11.2.1 Principal Component Analysis

In Chapter 3.3, an introduction to standard PCA was given as well as an overview of state-of-the-art approaches to tackle issues with non-linearity, evolving natures of processes and the three-way nature of batch process data. It was observed that the concept of maximum likelihood (ML), which has relatively recently been formulated in the context of PCA modelling, has not yet received much attention in the context of PCA-based process monitoring and diagnosis. Also, it was discussed that the often extremely large dimensionality of batch process data may result in unwanted variance of the PCA model solution, especially if the number of batches is small. A trade-off between bias (which generally increases by removing estimated parameters) and variance (which generally increases by adding estimated parameters) is only possible on the basis of one contribution in the context of process monitoring so far, called Function Space PCA (FSPCA), which indicates another gap in related literature. As such, investigation of benefits and drawbacks of maximum likelihood estimators and bias-variance trade-off for PCA modelling remain to be evaluated in the context of process monitoring and diagnosis. Given that (1) ML-estimation of PCA models deals with the optimal assessment of scaling parameters and that (2) the scaling procedures used in Chapter 5 did not affect the monitoring performance, ML-estimation seems a lesser subject for research at first. However, in Chapter 5 it was not clear whether the insensitivity of the monitoring performance was due to an insensitivity inherent to the method or due to the nature of the faults. A theoretical or simulation study may reveal whether ML-estimation can effectively improve monitoring performance.

11.2.2 Monitoring by means of Principal Component Analysis

In Chapter 5, a PCA-based monitoring strategy was tested for the hydraulic parts of the SBR system separately as well as for the complete system. High detection performance was reported for a larger part of the identified fault classes. However, the contrary was true for some peculiar faults, like noisy artefacts in the weight measurements, observational outliers in ORP and pH sensor trajectories and oscillatory behaviour of the aeration system. This has lead to conclude that standard

PCA does not permit to detect faults that cannot be characterized as a shift or drift but rather by frequency-specific behaviour. Time-local events (such as outliers) and frequency-bound events (such as noise or oscillatory behaviour) thus require techniques specifically tuned for proper detection. Wavelet-based tools, such as the Multiscale PCA extension (MSPCA), or (wavelet) spectral analysis define frameworks by which this is likely to be possible.

Two conventional scaling approaches were tested in view of process monitoring – group scaling and autoscaling. On average, slightly better results were obtained by means of autoscaling. This improvement may be considered as marginal, indicating that scaling has a seemingly minimal influence on detection performance. Two hypotheses, which could not be falsified as yet, are that (1) a larger part of the reported faults are of such deviating nature that their detection is not sensitive to the scaling approach and (2) scaling does indeed not affect process monitoring performance. It was suggested that simulated examples are likely to be of better use to evaluate the hypotheses.

In view of the presence of several operational modes in the historical data set, it was evaluated whether mixture PCA models, i.e. separate models corresponding to each of the operational modes, could be used without prior knowledge of the actual operational mode. It was concluded that prior knowledge of the mode does not affect monitoring performance for the hydraulic parts of the system.

Suggestions for further research were aimed at improvement of model identification procedures. One approach is based on the inclusion of specific abnormal data, i.e. so called extreme event data that lie within the subspace defined by the normal data. This is likely to result in an increased leverage or impact of observations on the subspace identification. A second approach to be tested makes use of the framework of Function Space PCA, in which the PCA model is constrained by means of orthogonal basis functions. Such an approach is likely to reduce the variance of the estimated parameters, hence leading to improved generalization of the obtained models.

11.2.3 Diagnosis by means of Principal Component Analysis

In Chapter 6, both explorative analysis as well as automated diagnosis on the basis of PCA models and fuzzy C-means clustering (FCM) was attempted. Explorative analysis allowed to identify regions in the modelled subspaces corresponding to specific faults observed in the system and also revealed information that was not observed or given special attention during data screening. The pursued automated approach was however shown only to work well with a limited number of faults. Possible explanations for this are that (1) the FCM model that was used is not appropriate enough, in the sense that only spherical regions are identified, and (2) some of the classes have much larger impact on the PCA models than other classes, hereby blurring the obtained PCA scores for classes with a lower impact on the models.

Straightforward improvements may be expected by (1) using the FCM model structure identified by the Gustafson-Kessel algorithm, which allows to identify elliptical regions, and by (2) model stacking. Other suggestions consisted of adding extreme-event data for subspace identification and FSPCA-like approaches in which knowledge-based or meaningful constraints are implied on the model, as already proposed in view of process monitoring. Additional clustering or classification methods may be tested as well, depending on objectives for interpretability, robustness and performance.

11.2.4 Multivariate Statistical Process Control

A control algorithm aimed at on-line phase length optimization for the studied SBR was proposed and succesfully tested on-line in Chapter 7. The Hotelling's T^2 statistic was combined with a simple decision scheme to allow the shutdown of the running aerobic phase of the SBR cycle as soon as it could be detected that exogenous biological reactions were completed, i.e. when the process entered the endogenous respiration state. The method provided does not require that anything but the endogenous respiration state is explicitly modelled. As such, the *path* by which the process arrives in the endogenous state is not crucial to the development and application of the proposed control scheme, contrasting with classification schemes proposed in literature. The use of the developed test for detection of a targeted state of the monitored system was not proposed before. The reported study gives a clear proof of concept as (1) a considerable shortening of the respective phase and (2) a

significant improvement of the effluent quality of the system could be reported. Especially nitrate nitrogen levels were shown to be reduced. In contrast, the expected energy savings are minimal, due to an earlier implementation of a well-functioning DO setpoint controller.

It was noted that the underlying assumptions of the applied statistical test are not generally valid. Two adjustments, i.e. the use of a rather low confidence level and the requirement for a set of contiguous positive tests before the control action is pursued, were implemented to counteract potential problems related to the latter observation. Future research may however aim at the construction and use of (statistical) models that do not require the underlying assumptions mentioned above, in addition to adaptation of the model to changing numerical characteristics of the targeted state in time.

Importantly, the proposed controller is general in nature and is not limited to the reported application nor to the phase that was chosen for optimization. Future applications may therefore be aimed at the optimization of other phases that are typical for the studied SBR. The optimization of anoxic (detection of the end of denitrification) or anaerobic (detection of the end of phosphorus release) phases are potential and valid goals for wastewater treatment systems. More generally, the proposed controller allows to optimize any process with respect to its run length given that the targeted state is uniquely and sufficiently described by data obtained on-line.

11.3 Qualitative Representation of Trends

11.3.1 Method development

The reader is reminded of the formalism of qualitative analysis, by which a unique character is assigned to contiguous episodes in a time series which are characterized by a unique sign of the 1^{st} or 1^{st} and 2^{nd} derivative of the (approximated) signal. Each possible set of signs for the derivatives corresponds to a single character (see Figure 3.30 and Table 3.1 in Section 3.5). As such, qualitative analysis results in a word for each analyzed series.

In Section 3.5, two existing methods for qualitative representation or analysis of trends have been compared by means of a few examples. It was shown that none of the methods presents a solution-to-all. Indeed, the interval-halving method, was shown to be sensitive to direction of analysis and was reported not to be as robust as one would generally desire. In contrast, the cubic spline wavelet method was shown not to allow (1) the discrimination between jump changes and smooth inflection points and (2) the identification of (piece-wise) constant behaviour. Also, the latter method was not able to identify consecutive inflection points in between extrema. A suggested solution for improvement of the latter method in terms of jump detection was suggested in Section 3.5.7 but is not pursued as yet given that studied data series were not characterized by such jump changes.

In Chapter 8, the cubic spline wavelet method was improved in view of detecting inflection points in time series. To this end, three approaches for inflection point detection were tested. It was shown that the approach for extremum detection could simply be repeated for inflection point detection. Next to demonstration of the improvements on a simulated example, the improved method was demonstrated by a preliminary application to a real-life ORP signal exhibiting multiple inflection points.

11.3.2 Applications

Two applications of QRT were investigated. In the first of these, QRT was shown to reveal information on the qualitative behaviour of time series not perceived through -more classic- wavelet spectrum analysis. Moreover, it was shown that the provided QRT method allows to identify meaningful behaviour of a city's population. Dominant cycles observed with wavelet cycles lead to a window definition for QRT. Reverse interaction, either by analysis of reconstructed signals (i.e. reconstruction on the basis of the qualitatively relevant wavelet coefficients) or by analysis of residuals (i.e. qualitatively irrelevant behaviour), was suggested as a further development to get the most out of time series data.

The second application was aimed at demonstrating the potential of QRT in wastewater treatment control. It was shown that for the studied SBR, diagnostics can straightforwardly be linked with assessed qualitative behaviour in the form of words. The established links between those words and the diagnostics were set up as a dictionary. Effective automated control on the basis of the established diagnostics has been suggested to enable closed-loop control on the basis of qualitative ana-

lysis. The resulting words, i.e. the qualitative behaviour without information on the time-location of the identified features, may however not be enough in practice to enable efficient system control. Indeed, the location of extrema and inflection points in time is information on its own and may be useful for effective diagnosis and/or control. As such, incorporation of temporal information, i.e. location of extrema and inflection points, as well as the (numerical) values of the signal and/or derivatives in these points may allow the construction of improved inference systems.

11.3.3 Multivariate Qualitative Representation of Trends

QRT has been applied only for one measurement at a time, i.e. in an univariate context. Given that more sensors are available for the studied SBR system, improved data mining or diagnosis results may also be established by simultaneously analyzing data from multiple sensors. As already reported in literature (see Section 3.5.1), the number of characters can be extended so that every possible combination of qualitative behaviour in multiple variables can be uniquely identified. If L characters are used for an univariate series, then for the J-variate case, L^J characters will be necessary to capture all theoretically possible joint qualitative behaviours of the J variables. Given that the resulting alphabets may lead to ineffective abstraction of the data, data dimension reduction prior to qualitative analysis may allow to reduce the complexity of the alphabet. Data dimension reduction by means of PCA prior to qualitative analysis was proposed in literature. This however requires that the (numerical) correlation between the data stemming from different sensors effectively allows data dimension reduction with minimal loss of information.

While extensions of QRT for multivariate contexts have been proposed, qualitative analysis itself remains a univariate technique, i.e. the qualitative analysis itself remains applied to a single univariate series at the time. Indeed, the first approach (extending the alphabet) accounts for the multi-sensor context after obtaining the single presentations. The second reduces the number of single presentations by reducing the number of trajectories before qualitative analysis.

Methods that express qualitative behaviour of multiple trajectories jointly have not been proposed so far. As perspectives, a few proposals are therefore made here:

- *Qualitative Representation of Surfaces* (QRS) or *Multi-way Qualitative Analysis* (MQA). In the context of spectral data taken from a process, one is faced with the two-dimensional nature of the matrix which composes all the data taken within a cycle or a given time window (dimension of time and of wavelength). Rather than analyzing the trajectories for each wavelength separately, one may wish to exploit correlation with measurements at nearby wavelengths. Thus, rather than treating the separate columns in such matrix as trends, the matrix as a whole can be interpreted as a surface composed of absorbance measurements as function of time and wavelength. Qualitative descriptions of such surfaces and thus of the underlying data underlying (by extrema, lines connecting points with steepest slopes and saddle points) may help data interpretation. The envisioned use is for visualization purposes at first. More advanced use may consist of putting a categorical value in each spot of the two-dimensional matrix corresponding to the qualitative abstraction of the image, thereby enabling further analysis of the abstracted data.
- *Errors-In-variables QRT* (EIV-QRT). So far, the QRT methods have been implemented for series where a variable is measured as function of an error-free independent variable, i.e. the qualitative representation is a function of that independent variable. This is typical for the context of time series. However, when the (qualitative) relationship *between* two (or more) variables is targeted, the order or time index at which the corresponding measurements were taken may not be of real interest. In such situations, a description of the qualitative behaviour of one variable against the other could be an identified target in data analysis. As such, this may be considered as the qualitative equivalent of the reviewed principal curves method (see Section 3.3.2). A typical problem may concern the qualitative description of the difference between a measured input variable and a measured output variable of a system. To assess the behaviour of the system, a qualitative description of the trajectory in the biplot of the two variables is considered. As a result, the time dimension is then lost. Note that strictly one dimension is removed by doing so. Extensions to larger dimensions are theoretically possible, even though not avoiding a more complex alphabet.

11.3.4 Matching, warping and the concept of maturity

The data-driven results in the context of qualitative analysis presented in this dissertation were limited to the assessment of the qualitative description of given time series. Unsupervised comparison or matching of the resulting qualitative descriptions to each other has only been covered recently in literature (Maurya et al., 2005; Balasko et al., 2006). Balasko et al. (2006) focus explicitly on the comparison of qualitative descriptions with unequal length, i.e. unequal number of identified episodes, and uses fuzzy measures of similarity between qualitative shapes. In Maurya et al. (2005), the qualitative shape formalism is omitted in part as numerical measures based on integration of normalized shapes are used to define similarity between different time series.

So far, fuzzification of the qualitative descriptions themselves has not been proposed in the (inductive) field of qualitative analysis (QA). However, an episode with a clear upward trend but with marginal (positive) acceleration may simultaneously be both a linear upward trend and an accelerating upward trend in the fuzzy sense. Such fuzzification has not been proposed as yet, contrasting to works in the (deductive) field of Qualitative Reasoning, where fuzzy presentations have already received attention (Travé-Massuyès et al., 1997). Improved and generalized techniques may therefore emerge from fuzzification of qualitative analysis results.

Reconsider that a set of qualitative descriptions of time series are available after qualitative analysis, e.g. as for the pH trajectories studied in Chapter 10. Many of such time series represent a similar shape, e.g. a sequence of upward, downward and upward trends, while the location in time at which the changes in trend direction may vary (due to e.g. variation in loading and changing microbial activity). With respect to further analysis of the trajectories it may be of interest to align the different trajectories (with the same shape) so that trajectories match each other as much as possible. For example, by alignment of batch trajectories, the effects of varying location in time of key events on PCA-based models may be cancelled out so that remaining characteristics of the batch data can be investigated in detail.

It is repeated here that alignment or warping of the data of the studied SBR was not technically necessary for MPCA modelling. Indeed, as the SBR cycles and the constituting phases had equal length, classic MPCA models could be used in Chapters 5 and 6. In contrast, alignment of key events in time series may allow to discriminate between temporal variation and other variation in the studied data. An inspiring treatment of this subject is given in Ramsay and Silverman (2005).

Dynamic time warping (DTW) is a technique that has been developed for series alignment (Keogh, 2002). This technique aims to define a transformation of the time axis, i.e. a warped time axis, so that an analyzed series and a prototype series become numerically similar. The mapping of the original time axis to the warped axis is always monotonic (i.e. the chronological order is preserved) and is established for each analyzed series separately (except for one prototype series). DTW was conceived originally in the context of speech recognition, e.g. to match the same words spoken by different persons. Correlation Optimized Warping (COW) is an alternative method originally proposed in the context of synchronization of chromatographic data (Tomasi et al., 2004). Synchronization of batch process time series by both methods has already been evaluated by Kassidas et al. (1998), Pravdova et al. (2002), Ramaker et al. (2003) and Fransson and Folestad (2006). Contrasting with the unparameterized DTW and COW methods, Eilers (2004) proposes the so-called Parametric Time Warping (PTW) method.

Given the assessed qualitative behaviour of a set of time series, warping of the time axis of a series on the basis of the location of qualitatively defined events (e.g. extrema, inflection points) may be equally possible. Practically, this would mean that the time axis of each batch trajectory is transformed in such a way that key events, e.g. the ammonia valley in a pH time series, are located at the same point on a newly defined axis. The latter axis may be redefined as a maturity index. Indeed, consider that the detection of the ammonia valley indicates 100% process maturity given that the ammonium oxidation reaction is then completed. As such, each pH trajectory may be transformed in such a way that all ammonia valleys lie at 100 on the newly defined maturity axis. Also, a special requirement may be that the continuity of derivatives (e.g. 1^{st} and 2^{nd}) is preserved after transformation. Note that the transformations themselves, identified separately for each analyzed series, define maturity as a monotonic function of time and may contain valuable information as well.

Time warping on the basis of qualitative descriptions of trends has so far been studied only by Balasko et al. (2006), based on a pair-wise sequence alignment method adopted from the bio-informatics field. Note that these latter ultimately align the qualitative presentations and not the time series themselves. Future research may therefore investigate further whether time warping of numerical series on the basis of their qualitative descriptions is meaningful and/or beneficial for advanced mining of time series data-bases.

11.4 Combining Qualitative Representation of Trends and Principal Component Analysis

As pointed out in Section 3.5.1, PCA-based analysis of the locations in time of so-called breakpoints has already been proposed in literature. However, the provided methods so far do not include the automated detection of these breakpoints. Given that a method has been presented in this dissertation which allows the automated identification of inflection points, such PCA-based analysis -or any other kind of analysis for that matter- becomes possible in a fully automated fashion. Future research may therefore aim at the coupling of QRT, delivering the time location of key qualitative events such as extrema and inflection points, and numerical post-processing. PCA-based analysis will however require that the same number of time locations is available for each sample. As such, time series with varying numbers of (qualitatively defined) key events in time may offer an interesting challenge to deal with.

If a certain and strict order exists in the occurrence of key events, then the assessed time locations can be assigned to the respective column in the data matrix, while the remaining missing time locations may be treated as missing data. Interestingly, if certain (linear or non-linear) relationships are found to hold between the location in time of key events, then the (PCA) model that describes these relationships may be used to estimate the location in time of future key events. Such ability may present an important shift in qualitative analysis from *assessing* the occurrence of an event to *predicting* the occurrence of an event. Such predictions may be well suited for planning of future operations and phases or may serve to adjust control laws for the ongoing process if predicted endpoints are estimated to occur too early or too late. In addition to the analysis of the location in time of key events, the (filtered) numerical values of the monitored variable in extrema and its (filtered) derivative in inflection points, may serve to improve post-processing of qualitative descriptions as well.

11.5 Break-point detection versus complete trajectory analysis

In addition to splitting the presented results into PCA-based or QRT-based results, the work can also be split into (1) approaches conceived as general MPCA modelling of complete trajectories of the studied cyclic processes (Chapters 5 and 6) and (2) approaches that were focused on specific parts of batch trajectories, in particular break points or reaction end points (PCA-based in Chapter 7 and QRT-based in Chapter 10). Generally speaking, the results presented for the more focused approach were reportedly better or more promising. For the general-purpose approach, benefits were less outspoken and more critical remarks were necessarily made. It is therefore useful to evaluate as to why such a distinction in obtained results is observed.

A first consideration to make is that the studies focusing on specific aspects of the studied cyclic process were based on data sets that are smaller than those presented for the general-purpose approach. By doing so, effects resulting from the changing nature of the studied processes have been avoided or limited compared to the general-purpose studies. It can therefore be hypothetized that larger data sets could have lead to worse results for the focused approaches. This comment was already made for the PCA-based controller studied in Chapter 7. This consideration is also valid for the (single) real-life ORP trajectory example used in Chapter 8 and for the proposed control system in Chapter 10.

Secondly, it may be considered that the focused approaches as reported for the SBR system (Chapters 7, 8 and 10) are closely related to break point detection schemes reported in the reviewed literature (see Chapter 2). As such, the focused approaches are related to popular approaches in wastewater engineering. The general-purpose approach, however corresponding to specific literature as well, has had limited coverage in terms of full-scale application to wastewater treatment systems. As such, the reported results correlate well with the fact that break-point detection approaches are more popular than complete trajectory analysis.

As a third and last observation to be made, it may be considered that the focused approaches deal with rather specific questions about the behaviour of the data. Indeed, in Chapter 7 a well-defined and well-understood target behaviour was described statistically and in Chapter 8 and Chapter 10 meaningful key points in the data trajectories were searched for. These explicit searches for targeted behaviour

stand in contrast to the MPCA-based approaches where no specific behaviour is implied or searched for a priori. Put otherwise, the success of the focused approaches may as well be due to the fact that the answers that were searched for already lay in the questions that were posed.

The latter consideration may be of special interest for the design of supervisory control systems. Given that the latter remark suggests that improved results can be obtained when more specific questions are asked, the provision of an environment in which specific meaningful questions are likely to be generated and posed, may be considered a portal to progress. As such, the generation of such an environment, eventually including tools for formulation and handling of generated questions, is a valid purpose of research on its own. For example, artificial intelligence techniques, possibly framed around ontological and/or semantic presentations of (real-life) systems, may eventually allow that computers can generate questions that turn to be crucial for engineering successes in process supervision.

11.6 Data versus knowledge

The work presented in this thesis has in essence been limited to the use of data-driven methods for process analysis, monitoring, diagnosis and control. Underlying to the reported choice is the consideration that knowledge of biological systems is inherently incomplete. Following this assumption, it was hypothetized that the expected behaviour cannot be formulated accurately by means of mechanistic modelling or deep knowledge.

Somehow contrasting to the statement above, at several times throughout the presented work, links with available knowledge could be assessed. They proved useful or were suggested to aid in the improvement of data-driven modelling. Indeed, in Chapters 5 and 6 it was suggested that improved PCA modelling may be achieved by adding constraints to the identified model, following known or hypothetized mechanistic relationships or following interpretation of a priori non-constrained PCA models. Also, the suggested addition of extreme events to calibration data sets for subspace identification requires that it is known that the extreme events correspond to the correlation structure of the normal data. In Chapter 7, process knowledge was used to identify data samples that correspond to a targeted state, consequently modelled in a data-driven fashion. In the work related to qualitative analysis, a so-called dictionary was set up to automatically assess qualitative

behaviours with diagnostic information, in turn connected to proposed control actions. To enable the design of this dictionary, the expertise of the system's operators proved to be useful, even essential.

The original proposal (silently) assumed that the status of available (mechanistic and deep) knowledge of the system is largely non-existing and would remain to be so. In retrospect, this is far from the now historical truth. Indeed, mechanistic knowledge of the physical and electro-mechanical parts of the system (e.g. hydraulics, aeration, cooling, automated data acquisition and control) was obtained (i.e. within months). Also, rigorous screening of a large historical data set has resulted in extensive and fairly detailed descriptions of faults typical to the studied system (see Chapter 4). In addition, interpretation of measurement campaigns such as the one presented in Chapter 7 and intensive discussions have lead to the formulation and assessment of deep knowledge of the biological parts of the system. The knowledge acquisition process, as described here, was not accounted for as such in the original project. As a result, knowledge-driven methods, such as mechanistic modelling or qualitative reasoning, were largely excluded from the proposed developments.

Given the reported increase in process understanding and given that links between data-driven results and process understanding were pointed out throughout the work, it is the author's opinion that guaranteeing flexibility to incorporate growing knowledge into the design and planning of supervisory control systems is likely to deliver benefits for both the development of elements of such supervisory control systems as well as for process performance as a whole. Consider for example that assessed knowledge can be *stripped* from on-line process data. For example, the use of basis functions corresponding to that knowledge, as discussed already in this chapter, may allow to create residuals that express deviations from the knowledge-based presentation of reality. Also, the provision of data which enables balance checks as discussed above may support such *knowledge stripping* as well. Analysis of the residuals obtained in this way may serve to better understand the system, possibly leading to improved strategies for monitoring, diagnosis and/or control. In the same fashion, incorporation of knowledge may serve to separate well-understood or well-described problems, e.g. failures of the physical parts of the system, from rather uncommon or vaguely understood problems. Thereby, well-understood situations or problems may be addressed faster and/or more effectively. Also, as time spent on well-understood situations is reduced by doing so, an increased amount of time and energy can be spent on other problems which require more detailed inspection.

In a possibly utopic proposal, it may be considered that repeated iterations of data-driven knowledge extraction (by residuals analysis) and the knowledge stripping (generating new residuals with acquired knowledge) may eventually lead to the complete understanding of the studied process and turn the data-driven approach obsolete for process supervision. Such an attempt may be utopic indeed given that (1) the process lifetime should be large enough compared to the number of knowledge extraction-stripping loops that are needed to obtain a complete knowledge-based system presentation and (2) process knowledge may become invalid in due time as a result of intended and unintended changes of the behaviour of the studied system. As such, providing both knowledge-based and data-driven approaches as well as their coupling into supervisory control systems is likely to generate the best of possibilities for process monitoring, diagnosis and control in practice.

11.7 Final thoughts

With respect to the task of supervision, many more techniques than those used and presented in this work are available in literature, using different assumptions, formalisms and perspectives. With respect to individual techniques for process supervision and control, it has been shown that many opportunities for further improvement exist for individual techniques, both generally as well specifically for biological processes. Possibly more important for practice, the parallel use and integration of knowledge-based and data-driven approaches presents a new direction of research in process monitoring, diagnosis and control also identified in literature (Venkatasubramanian et al., 2003c). In the same line of thought, interaction between qualitative and quantitative methods may pave a path to improved extraction of information, process supervision and understanding of biotechnological processes. Ultimately, no possible angle of view should be dismissed when looking at biological systems.

Bibliography

Aguado, D., Ferrer, J., and Seco, A. (2007). Multivariate SPC of a sequencing batch reactor for wastewater treatment. *Chem. Intell. Lab. Syst.*, 85:82–93.

Aguado, D., Montoya, T., Ferrer, J., and Seco, A. (2006). Relating ions concentration variations to conductivity variations in a sequencing batch reactor operated for enhanced biological phosphorus removal. *Environ. Modell. Softw.*, 21:845–851.

Aguado, D., Zarzo, M., Ferrer, J., and Seco, A. (2005). A multivariate methodology for detecting operational shifts: application to a sequencing batch reactor. In *IWA Conference on Nutrient Removal in Wastewater Treatment Plants and Recycle Streams (BNR2005), Krakow, Poland, September 19-21, 2005*, pages 755–764.

Akbaryan, F. and Bishnoi, P. (2000). Smooth representation of trends by a wavelet-based technique. *Comput. Chem. Eng.*, 24:1913–1943.

Akbaryan, F. and Bishnoi, P. (2001). Fault diagnosis of multivariate systems using pattern recognition and multisensor data analysis technique. *Comput. Chem. Eng.*, 25:1313–1339.

Al-Ghusain, I., Huang, J., Hao, O., and Lim, B. (1995). Using pH as real-time control parameter for wastewater treatment and sludge digestion processes. *Wat. Sci. Technol.*, 30(4):159–168.

Alewell, C. and Manderscheid, B. (1998). Use of objective criteria for the assessment of biogeochemical ecosystem models. *Ecol. Modell.*, 107:213–214.

Andreottola, G., Foladori, P., and Ragazzi, M. (2001). On-line control of a SBR system for nitrogen removal from industrial wastewater. *Wat. Sci. Technol.*, 43(3):93–100.

Aradhye, H., Bakshi, B., Strauss, R., and Davis, J. (2003). Multiscale SPC using wavelets: theoretical analysis and properties. *AIChE J.*, 49:939–958.

Artan, N., Wilderer, P., Orhon, D., Morgenroth, E., and Özgür, N. (2001). The mechanism and design of sequencing batch reactor systems for nutrient removal - the state of the art. *Wat. Sci. Technol.*, 43(3):53–60.

Babuska, R. (1998). *Fuzzy modelling for control*. Kluwer, Amsterdam, The Netherlands.

Bakshi, B. (1998). Multiscale PCA with application to multivariate statistical process monitoring. *AIChE J.*, 44:1596–1610.

Bakshi, B. and Stephanopoulos, G. (1994). Representation of process trends – part III. multiscale extraction of trends from process data. *Comput. Chem. Eng.*, 18:267–302.

Balasko, B., nemeth, S., and Abonyi, J. (2006). Qualitative analysis of segmented time-series by sequence alignment. In *HUCI2006*. Appeared on CD-ROM.

Balslev, P., Pedersen, P., Kjeldsen, J., and Brøkner, P. (2005). Online control of nitrogen removal at small treatment plants using simple sensors. In *IWA Conference on Nutrient Removal in Wastewater Treatment Plants and Recycle Streams (BNR2005), Krakow, Poland, September 19-21, 2005*. Appeared on CD-ROM.

Berleant, D. and Kuipers, B. (1997). Qualitative and quantitative simulation: bridging the gap. *Artif. Intell.*, 95:215–255.

Bezdek, J. (1981). *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, NY, USA.

Bicciato, S., Bagno, A., Soldà, M., Manfredini, R., and Di Bello, C. (2002). Fermentation diagnosis by multivariate statistical analysis. *Appl. Biochem. Biotechnol.*, 102-103:49–62.

Bisschops, I., Spanjers, H., and Keesman, K. (2006). Automatic detection of exogenous respiration end-point using artifical neural network. *Wat. Sci. Technol.*, 53(4-5):273–281.

Boeije, G. (1999). *Chemical fate prediction fur use in geo-reference environmental exposure assessment*. PhD thesis, Ghent University.

Bourseau, P., Bousson, K., Dague, P., Dormoy, J., Evrard, J., Guerrin, F., Leyval, L., Lhomme, O., Lucas, B., Missier, A., Montmain, J., Piera, N., Rakoto-Ravalontsalama, N., Steyer, J., Tomasena, M., Travé-Massuyès, L., Vescovi, M., Xanthakis, S., and Yannou, B. (1995). Qualitative reasoning: A survey of techniques and applications. *AI Commun.*, 8:119–192.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2:121–167.

Cai, Y. and Davies, N. (2003). A simple diagnostic method of outlier detection for stationary gaussian time series. *J. Appl. Stat.*, 30:205–223.

Camacho, J. and Picó, J. (2006). Multi-phase principal component analysis for batch processes modelling. *Chem. Intell. Lab. Syst.*, 81:127–136.

Campos, H. and von Sperling, M. (1996). Estimation of domestic wastewater characteristics in a developing country based on socio-economic variables. *Wat. Sci. Technol.*, 34(3-4):71–77.

Capalozza, C. a. (2001). Design, startup and monitoring of a pilot sequencing batch reactor for breeding stable nutrient removal sludge. Master's thesis, Ghent University.

Cecil, D. (2007). The control of denitrification time in full scale by the automatic detection of the low nitrate bend in the redox curve. In *Proceedings of the 3rd International IWA Conference on Automation in Water Quality Monitoring (AutMoNet2007), Ghent, Belgium, September 5-7, 2007*. Appeared on CD-ROM.

Cecil, D. and Skou, E. (2005). A model of the redox measurement in aerated activated sludge. *Wat. Sci. Technol.*, 53(4-5):465–472.

Chang, H. and Hao, O. (1996). Sequencing batch reactor system for nutrient removal: ORP and pH profiles. *J. Chem. Technol. Biotechnol.*, 67:27–38.

Charbonnier, S., Garcia-Beltan, C., Cadet, C., and Gentil, S. (2005). Trends extraction and analysis for complex system monitoring and decision support. *Eng. Appl. Artif. Intell.*, 18:21–36.

Charpentier, J., Martin, G., Wacheux, H., and Gilles, P. (1998). ORP regulation and activated sludge: 15 years of experience. *Wat. Sci. Technol.*, 38(3):197–208.

Chen, J. and Liu, J. (1999). Mixture principal component analysis models for process monitoring. *Ind. Eng. Chem. Res.*, 38:1478–1488.

Chen, J. and Liu, J. (2001). Derivation of function space analysis based PCA control charts for batch process monitoring. *Chem. Eng. Sci.*, 3289-3304:3289–3304.

Chen, J. and Liu, K.-C. (2002). On-line batch process monitoring using dynamic PCA and dynamic PLS models. *Chem. Eng. Sci.*, 57:63–75.

Chen, M., J.-H., K., Kishida, N., Nishimura, O., and Sudo, R. (2004). Enhanced nitrogen removal using C/N load adjustment and real-time control strategy in

sequencing batch reactors for swine wastewater treatment. *Wat. Sci. Technol.*, 49:309–314.

Chernick, M., Downing, D., and Pike, D. (1982). Detecting outliers in time series data. *J. Am. Stat. Assoc.*, 77:743–747.

Cheung, J.-Y. and Stephanopoulos, G. (1990a). Representation of process trends – part I. a formal representation framework. *Comput. Chem. Eng.*, 14:495–510.

Cheung, J.-Y. and Stephanopoulos, G. (1990b). Representation of process trends – part II. the problem of scale and qualitative scaling. *Comput. Chem. Eng.*, 14:511–539.

Chiang, L., Leardi, R., Pell, R., and Seasholtz, M. (2006). Industrial experiences with multivariate statistical analysis of batch process data. *Chem. Intell. Lab. Syst.*, 81:109–119.

Cho, B., Liaw, S.-L., Chang, C., Yu, S., and Chiou, B.-R. (2001). Development of a real-time control strategy with artificial neural network for automatic control of a continuous-flow sequencing batch reactor. *Wat. Sci. Technol.*, 44(1):95–104.

Choi, S., Lee, C., Lee, J.-M., Park, J., and Lee, I.-B. (2005). Fault detection and identification of nonlinear processes based on kernel PCA. *Chem. Intell. Lab. Syst.*, 75:55–67.

Chou, C. and Verhaegen, M. (1997). Subspace algorithm for the identification of multivariable dynamic errors-in-variables models. *Automatica*, 33:1857–1869.

Ciappelloni, F., Mazouni, D., Harmand, J., and Lardon, L. (2006). On-line supervision and control of an aerobic SBR process. *Wat. Sci. Technol.*, 53(1):169–177.

Cohen, A., Hegg, D., de Michele, M., Song, Q., and Kasabov, N. (2003). An intelligent controller for automated operation of sequencing batch reactors. *Wat. Sci. Technol.*, 47(12):57–63.

Corominas, L., Sin, G., Puig, S., Traore, A., Balaguer, M., Colprim, J., and Vanrolleghem, P. (2006). Model-based evaluation of an on-line control strategy for SBRs based on OUR and ORP measurements. *Wat. Sci. Technol.*, 53(4-5):161–169.

Cortes, C. and Vapnik, V. (1995). Support-vector network. *Mach. Learn.*, 20:273–297.

Dash, S., Maurya, M., and Venkatasubramanian, V. (2004a). A novel interval-halving framework for automated identification of process trends. *AIChE J.*, 50:149–162.

Dash, S., Maurya, M., and Venkatasubramanian, V. (2004b). A novel interval-halving framework for automated identification of process trends, CIPAC technical report (CIPAC-03-3). Technical report, Purdue University.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 41:909–996.

Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM Press, Philadelphia.

Dayal, B. and MacGregor, J. (1997). Recursive exponentially weighted PLS and its applications to adaptive control ad prediction. *J. Process Control*, 7:169–179.

de Kleer, J. (1977). Multiple representations of knowledge in a mechanics problem solver. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence, M.I.T., Cambridge, MA, August 22-25, 1977.*, pages 286–291.

Demuynck, C., Vanrolleghem, P., Mingneau, C., Liessens, J., and Verstraete, W. (1994). NDEBPR process optimization in SBRs: reduction of external carbon-source and oxygen supply. *Wat. Sci. Technol.*, 30(4):169–179.

Dong, D. and McAvoy, T. (1996a). Batch tracking via non-linear principal component analysis. *AIChE J.*, 42:2199–2208.

Dong, D. and McAvoy, T. (1996b). Nonlinear principal component analysis – based on principal curves and neural networks. *Comput. Chem. Eng.*, 1:65–78.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455.

Dunia, R. and Qin, S. (1998a). Joint diagnosis of process and sensor faults using principal component analysis. *Control Eng. Practice*, 6:457–469.

Dunia, R. and Qin, S. (1998b). Subspace approach to multidimensional identification and reconstruction. *AIChE J.*, 44:1813–1831.

Dunia, R. and Qin, S. (1998c). A unified geometric approach to process and sensor fault identification and reconstruction the unidimensional fault case. *Comput. Chem. Eng.*, 22:927–943.

Dunia, R., Qin, S., Edgar, T., and McAvoy, T. (1996). Use of principal component analysis for sensor fault identification. *Comput. Chem. Eng.*, 20:Suppl., S713–S718.

Eilers, P. (2004). Parametric time warping. *Anal. Chem.*, 76:404–411.

Farge, M. (1992). Wavelet transforms and their applications to turbulence. *Annu. Rev. Fluid Mech.*, 24:395–457.

Flehmig, F. and Marquardt, W. (2006). Detection of multivariable trends in measured process quantities. *J. Process Control*, 16:947–957.

Flehmig, F., Watzdorf, R., and Marquardt, W. (1998). Identification of trends in process measurements using the wavelet transform. *Comput. Chem. Eng.*, 22:S491–S496.

Flores-Cerrillo, J. and MacGregor, J. (2004). Multivariate monitoring of batch processes using batch-to-batch information. *AIChE J.*, 50:1219–1228.

Forbus, K. (1984). Qualitative process theory. *Artif. Intell.*, 24:85–168.

Fox, A. (1972). Outliers in time series. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 34:350–363.

Fransson, M. and Folestad, S. (2006). Real-time alignment of batch process data using COW for on-line process monitoring. *Chem. Intell. Lab. Syst.*, 84:56–61.

Fuerhacker, M., Bauer, H., Ellinger, R., Sree, U., Schmid, H., Zibuschka, F., and Puxbaum, H. (2001). Approach for a novel control strategy for simultaneous nitrification/denitrification in activated sludge reactors. *Wat. Res.*, 34:2499–2506.

Ganesan, R., Das, T., and Venkataraman, V. (2004). Wavelet-based multiscale statistical process monitoring: A literature review. *IIE Trans.*, 36:787–806.

Govoreanu, R., Seghers, D., Nopens, I., De Clercq, B., Saveyn, H., Capalozza, C., Van der Meeren, P., Verstraete, W., Top, E., and Vanrolleghem, P. (2003). Linking floc structure and settling properties to activated sludg population dynamics in an SBR. *Wat. Sci. Technol.*, 47(12):9–18.

Gregersen, L. and Jørgensen, S. (1999). Supervision of fed-batch fermentations. *Chem. Eng. J.*, 75:69–76.

Gruber, G. and Bretrand-Krajewski, J.-L. (2005). Practical aspects, experiences and strategies by using UV/VIS sensors for long-term sewer monitoring. In *Proceedings of the 10th International Conference on Urban Drainage, Copenhagen, Denmark, August 21-26, 2005*. Appeared on CD-ROM.

Guisasola, A., Pijuan, M., Baeza, J., Carrera, J., and Lafuente, J. (2006). Improving the start-up of an EBPR system using OUR to control the aerobic phase length: a simulation study. *Wat. Sci. Technol.*, 53(4-5):253–262.

Guo, J., Yang, Q., Peng, Y., Yang, A., and Wang, S. (2007). Biological nitrogen removal with real-time control using step-feed SBR technology. *Enzyme Microb. Technol.*, 40:1564–1569.

Gurden, S., Westerhuis, J., Bro, R., and Smilde, A. (2001). A comparison of multiway regression and scaling methods. *Chem. Intell. Lab. Syst.*, 59:121–136.

Gustafson, E. and Kessel., W. (1979). Fuzzy clustering with a fuzzy covariance matrix. In *Proceedings of the IEEE Conference on Decision Control (IEEE CDC), San Diego, CA*, pages 761–766.

Hao, O. and Huang, J. (1996). Alternating aerobic-anoxic process for nitrogen removal: process evaluation. *Water Environ. Res.*, 68(1):83–93.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *J. Am. Stat. Assoc.*, 84:502–516.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning. Data Mining, Inference, and Prediction*. Springer, NY, USA.

Insel, G., Sin, G., Lee, D., Nopens, I., and Vanrolleghem, P. (2006). A calibration methodology and model-based systems analysis for SBR's removing nutrients under limited aeration conditions. *J. Chem. Technol. Biotechnol.*, 81:679–687.

Irvine, R., Wilderer, P., and Flemming, H. (1997). Controlled unsteady state processes and technologies - an overview. *Wat. Sci. Technol.*, 35(1):11–18.

Iwasaki, Y. (1997). Real world applications of qualitative reasoning: Introduction to the special issue. *IEEE Expert*, 12:16–21.

Jackson, J. (1991). *A user's guide to principal components*. Wiley-Interscience, New York, USA.

Jackson, J. and Mudholkar, G. (1979). Control procedures for residuals associated with principal component analysis. *Technometrics*, 21:341–349.

Jia, F., Martin, E., and Morris, A. (1998). Non-linear principal component analysis for process fault detection. *Comput. Chem. Eng.*, 22:Suppl., S851–S854.

Johansen, N., Andersen, J., and la Cour Jansen, J. (1997). Optimum operation of a small sequencing batch reactor for BOD and nitrogen removal based on on-line OUR calculation. *Wat. Sci. Technol.*, 35(6):29–36.

Johnson, R. and Wichern, D. (2002). *Applied multivariate statistical analysis*. Prentice-Hall Inc., Upper Saddle River, NJ, USA, 5th edition.

Joliffe, I. (2002). *Principal component analysis*. Springer, New York, USA, 2nd edition.

Kano, M., Hasebe, S., Hashimoto, I., and Ohno, H. (2001). A new multivariate statistical process monitoring method using principal component analysis. *Comput. Chem. Eng.*, 25:1103–113.

Kano, M., Nagao, K., Hasebe, S., Hashimoto, I., Ohno, H., Strauss, R., and Bakshi, B. (2000). Comparison of statistical process monitoring methods: application to the Eastman challenge problem. *Comput. Chem. Eng.*, 24:175–181.

Kassidas, J., MacGregor, J., and Taylor, P. (1998). Synchronization of batch trajectories using dynamic time warping. *AIChE J.*, 44:864–875.

Keogh, E. (2002). Exact indexing of dynamic time warping. In *Proceedings of the 28th International Conference on Very Large Data Bases 2002 (VLDB2002)*, pages 406–417.

Kim, D. and Lee, I.-B. (2003). Process monitoring based on probabilistic PCA. *Chem. Intell. Lab. Syst.*, 67:109–123.

Kim, H. and Hao, O. (2001). pH and oxidation-reduction potential control strategy for optimization of nitrogen rmoval in an alternating aerobic-anoxic system. *Water Environ. Res.*, 73:95–102.

Kim, J.-H., Chen, M., Kishida, N., and Sudo, R. (2004). Integrated real-time control strategy for nitrogen removal in swine wastewater treatment using sequencing batch reactors. *Wat. Res.*, 38:3340–3348.

Kishida, N., Kim, J., Chen, M., Tsuneda, S., Sasaki, H., and Sudo, R. (2004). Automatic control strategy for biological nitrogen removal of low C/N wastewater in a sequencing batch reactor. *Wat. Sci. Technol.*, 50(10):45–50.

Klapwijk, A., Brouwer, H., Vrolijk, E., and Kujawa, K. (1998). Control of intermittently aerated nitrogen removal plants by detection endpoints of nitrification and denitrification using respirometer only. *Wat. Res.*, 32:1700–1703.

Kosanovich, K., Dahl, K., and Piovoso, M. (1996). Improved process understanding using multiway principal component analysis. *Ind. Eng. Chem. Res.*, 35:138–146.

Kourti, T. (2002). Process analysis and abnormal situation detection: from theory to practice. *IEEE Control Syst. Mag.*, 22(5):10–25.

Kourti, T. (2003). Abnormal situation detection, three-way data and projection methods; robust data archiving and modeling for industrial applications. *Annu. Rev. Control*, 27:131–139.

Kourti, T. and MacGregor, J. (1995). Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chem. Intell. Lab. Syst.*, 28:3–21.

Kourti, T. and MacGregor, J. (1996). Multivariate SPC methods for process and product monitoring. *J. Qual. Technol.*, 28:409–428.

Kruger, U., Zhou, Y., and Irwin, G. (2004). Improved principal component monitoring of large-scale processes. *J. Process Control*, 14:879–888.

Ku, W., Storer, R., and Georgakis, C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chem. Intell. Lab. Syst.*, 30:179–196.

Kuipers, B. (1986). Qualitative simulation. *Artif. Intell.*, 29:289–338.

Kuipers, B. (1994). *Qualitative Reasoning: Modeling and simulation with incomplete knowledge*. MIT Press, Cambridge, MA, USA.

Kwok, J. T.-Y. and Tsang, I. W.-H. (2004). The pre-image problem in kernel methods. *IEEE Trans. Neural Netw.*, 15:1517–1525.

Langergraber, G., Fleischmann, N., and Hofst adter, F. (2003). a multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater. *Wat. Sci. Technol.*, 47(2):63–71.

Langergraber, G., Gupta, J., Pressi, A., Hofstaedter, F., Letti, W., Weingartner, A., and Fleischmann, N. (2004). On-line monitoring for control of a pilot-scale sequencing batch reactor using a submersible UV/VIS spectrometer. *Wat. Sci. Technol.*, 50(10):73–80.

Lardon, L., Punal, A., and Steyer, J.-P. (2004). On-line diagnosis and uncertaintly management using evidence theory – experimental illustration to anaerobic digestion processes. *J. Process Control*, 14:747–763.

Lee, D., Jeon, C., and Park, J. (2001). Biological nitrogen removal with enhanced phosphate uptake in a sequencing batch reactor using single sludge system. *Wat. Res.*, 35:3968–3976.

Lee, D., Park, J., and Vanrolleghem, P. (2005). Adaptive multiscale principal component analysis for on-line monitoring of a sequencing batch reactor. *J. Biotechnol.*, 116:195–210.

Lee, D. and Vanrolleghem, P. (2003). Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis. *Biotechnol. Bioeng.*, 82:489–497.

Lee, D. and Vanrolleghem, P. (2004). Adaptive consensus principal component analysis for on-line batch process monitoring. *Environ. Monit. Assess.*, 92:119–135.

Lee, H., Min, Y., Park, C., and Park, Y. (2004a). Automatic control and remote monitoring system for biological nutrient removal on small wastewater treatment plants in Korea. *Wat. Sci. Technol.*, 50(6):199–206.

Lee, J.-M., Yoo, C., Choi, S., Vanrolleghem, P., and Lee, I. (2004b). Nonlinear process monitoring using kernel principal component analysis. *Chem. Eng. Sci.*, 59:223–234.

Lennox, B., Hiden, H., Montague, G., Kornfeld, G., and Goulding, P. (2000). Application of multivariate statistical process control to batch operations. *Comput. Chem. Eng.*, 24:291–296.

Lennox, B., Montague, G., Hiden, H., Kornfeld, G., and Goulding, P. (2001). Process monitoring of an industrial fed-batch fermentation. *Biotechnol. Bioeng.*, 74:125–135.

Lennox, J. and Rosén, C. (2002). Adaptive multiscale principal components analysis for online monitoring of wastewater treatment. *Wat. Sci. Technol.*, 45:227–235.

Li, W. and Qin, S. (2001). Consistent dynamic PCA based on errors-in-variables subspace identification. *J. Process Control*, 11:661–678.

Li, W., Yue, H., Valle-Cervantes, S., and Qin, S. (2000). Recursive PCA for adaptive process monitoring. *J. Process Control*, 10:471–486.

Li, Y., Peng, C., Peng, Y., and Wang, P. (2004). Nitrogen removal from pharmaceutical manufacturing wastewater via nitrite and the process optimization with on-line control. *Wat. Sci. Technol.*, 50(6):25–30.

Lu, N., Gao, F., Yang, Y., and Wang, F. (2004). PCA-based modeling and on-line monitoring strategy for uneven-length batch processes. *Ind. Eng. Chem. Res.*, 43:3343–3352.

Luo, R., Misra, M., Qin, S., Barton, R., and Himmelblau, D. (1998). Sensor fault detection via multiscale analysis and nonparametric statistical reference. *Ind. Eng. Chem. Res.*, 37:1024–1032.

Ma, Y., Peng, Y., Yuan, Z., Wang, S., and Wu, X. (2006). Feasibility of controlling nitrification in predenitrification plants using DO, pH and ORP sensors. *Wat. Sci. Technol.*, 53(4-5):235–243.

Mallat, S. (1999). *A wavelet tour of Signal Processing*. Academic Press, San Diego, CA, USA, 2 edition.

Mallat, S. and Zhong, S. (1992). Characterization of signals from multiscale edges. *IEEE Trans. on Pattern Anal. and Machine Intel.*, 14:710–721.

Maribas, A., da Silva, M., Laurent, N., Loison, B., Battaglia, P., and Pons, M.-N. (2007). Monitoring of rain events with a submersible uv/vis spectrophotometer. In *Proceedings of the 3rd International IWA Conference on Automation in Water Quality Monitoring (AutMoNet2007), Ghent, Belgium, September 5-7, 2007*. Appeared on CD-ROM.

Marsili-Libelli, S. (1998). Adaptive fuzzy monitoring and fault detection. *Int. J. COMADEM*, 1(3):31–37.

Marsili-Libelli, S. (2006). Control of SBR switching by fuzzy pattern recognition. *Wat. Res.*, 40:1095–1107.

Marsili-Libelli, S. and Müller, A. (1996). Adaptive fuzzy pattern recognition in the anaerobic digestion process. *Pattern Recognit. Lett.*, 17:651–659.

Maurer, M. and Gujer, W. (1995). Monitoring of microbial phosphorous release in batch experiments using electric-conductivity. *Wat. Res.*, 29:2613–2617.

Mauret, M., Ferrand, F., Boisdon, V., Sperandio, M., and Paul, E. (2001). Process using DO and ORP signals for biological nitrification and denitrification validation of a food-processing industry waste-water treatment plant on boosting with pure oyxgen. *Wat. Sci. Technol.*, 44(2-3):163–170.

Maurya, M., Rengaswamy, R., and Venkatasubramanian (2002). A signed directed graph and qualitative trend analysis-based framework for incipient fault diagnosis, technical report cipac-02-1. Technical report, Purdue University.

Maurya, M., Rengaswamy, R., and Venkatasubramanian, V. (2005). Fault diagnosis by qualitative trend analysis of the principal components. *Comput. Chem. Eng.*, 83:1122–1132.

Meyer, Y. (1993). *Wavelets and operators*. Cambridge University Press, UK.

Montgomery, D. (2005). *Introduction to Statistical Quality Control*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 5th edition.

Muirhead, C. (1986). Distinguishing outlier types in time series. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 48:39–47.

Narasimhan, S. and Shah, S. (2007). Model identification and error covariance matrix estimation from nosiy data using PCA. *Control Eng. Practice*, In Press.

Nomikos, P. and MacGregor, J. (1994). Monitoring batch process using multiway principal component analysis. *AIChE J.*, 40:1361–1375.

Nomikos, P. and MacGregor, J. (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37:41–59.

Olsson, G. (2006). Instrumentation, control and automation in the water industry – state-of-the-art and new challenges. *Wat. Sci. Technol.*, 53(4-5):1–16.

Olsson, G. and Newell, B. (1999). *Wastewater Treatment Systems. Modelling, Diagnosis and Control*. IWA Publishing, London, UK.

Olsson, R. (2005). *Batch control and diagnosis*. PhD thesis, Lund University.

Parent, A.-C., Anctil, F., and Parent, L.-É. (2006). Characterization of temporal variability in near-surface soil moisture at scales from 1 h to 2 weeks. *J. Hydrol.*, 325:56–66.

Patton, R. J., Frank, P. M., and Clark, R. N. (2000). *Issues of Fault Diagnosis for Dynamic Systems*. Springer-Verlag, London, UK.

Paul, E., Plisson-Saune, S., Mauret, M., and J., C. (1998). Process state evaluation of alternating oxic-anoxic activated sludge using ORP, pH and DO. *Wat. Sci. Technol.*, 38(3):299–306.

Peddie, C., Mavinic, D., and Jenkins, C. (1990). Use of ORP for monitoring and control of aerobic sludge digestion. *J. Environ. Eng.*, 116:461–471.

Peng, Y., Chen, Y., Peng, C., Liu, M., Wang, S., Song, X., and Cui, Y. (2004). Nitrite accumulation by aeration controlled in sequencing batch reactors treating domestic wastewater. *Wat. Sci. Technol.*, 50(10):35–43.

Peng, Y., Gao, J., Wang, S., and Sui, M. (2002). Use pH and ORP as fuzzy control parameters of denitirifcation in SBR process. *Wat. Sci. Technol.*, 46(4-5):131–137.

Plisson-Saune, S., Capdeville, B., Mauret, M., Deguin, A., and Baptiste, P. (1996). Real-time control of nitrogen removal using three ORP bending-points: signification, control strategy and results. *Wat. Sci. Technol.*, 33(1):275–280.

Pravdova, V., Walczak, B., and Massart, D. (2002). A comparison of two algorithms for warping of analytical signals. *Anal. Chim. Acta*, 456:77–92.

Puig, S., Corominas, L., Vives, M., Balaguer, M., and Colprim, J. (2005). Development and implementation of a real-time control system for nitrogen removal using OUR and ORP as end points. *Ind. Eng. Chem. Res.*, 44:3367–3373.

Qin, S. and Li, W. (1999). Detection, identification, and reconstruction of faulty sensors with maximized sensitivity. *AIChE J.*, 45:1963–1976.

Qin, S. and Li, W. (2001). Detection and identification of faulty sensors in dynamic processes. *AIChE J.*, 47:1581–1593.

Ra, C., Lo, K., and Mavinic, D. (1999). Control of a swine manure treatment process using a specific feature of oxidation reduction potential. *Bioresource Technol.*, 70:117–127.

Raich, A. and Çinar, A. (1996). Statistical process monitoring and disturbance diagnosis in multivariable continuous processes. *AIChE J.*, 42:995–1009.

Raich, A. and Çinar, A. (1997). Diagnosis of process disturbances by statistical distance and angle measures. *Comput. Chem. Eng.*, 21:661–673.

Ramaker, H., van Sprang, E., Gurden, S., Westerhuis, J., and Smilde, A. (2002). Improved monitoring of batch processes by incorporating external information. *J. Process Control*, 12:569–576.

Ramaker, H., van Sprang, E., Westerhuis, J., and Smilde, A. (2003). Dynamic time warping of spectroscopic batch data. *Anal. Chim. Acta*, 498:133–153.

Ramaker, H.-J., van Sprang, E., Westerhuis, J., and Smilde, A. (2004). The effect of the size of the training set and number of principal components on the false alarm rate in statistical process monitoring. *Chem. Intell. Lab. Syst.*, 73:181–187.

Ramaker, H.-J., van Sprang, E., Westerhuis, J., and Smilde, A. (2005). Fault detection properties of global, local and time evolving models for batch process monitoring. *J. Process Control*, 15:799–805.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York, USA.

Rännar, S., MacGregor, J., and Wold, S. (1998). Adaptive batch monitoring using hierarchical PCA. *Chem. Intell. Lab. Syst.*, 41:73–81.

Ranta, R., Louis-Dorr, V., Heinrich, C., and Wolf, D. (2005). Iterative wavelet-based denoising and robust outlier detection. *IEEE Signal Process. Lett.*, 12:557–560.

Reddy, V. and Mavrovouniotis, M. (1998). An input-training neural network approach for gross error detection and sensor replacement. *Comput. Chem. Eng.*, 76:478–489.

Rengaswamy, R. and Venkatasubramanian, V. (1995). A syntactic pattern-recognition approach for process monitoring and fault diagnosis. *Eng. Appl. Artif. Intell.*, 8:35–51.

Rieger, L.and Langergraber, G. and Siegrist, H. (2006). Unvertainties of spectral in situ measurements in wastewater using different calibration approaches. *Wat. Sci. Technol.*, 53(12):187–197.

Rieger, L., langergraber, G., Kaelin, D., Siegrist, H., and Vanrolleghem, P. (2007). Long-term evaluation of a spectral sensor for nitrite and nitrate. In *Proceedings of the 3rd International IWA Conference on Automation in Water Quality Monitoring (AutMoNet2007), Ghent, Belgium, September 5-7, 2007*. Appeared on CD-ROM.

Rosipal, R. and Girolami, M. (2001). An expectation maximization approach to nonlinear component analysis. *Neural Comput.*, 13:505–510.

Rosén, C. and Lennox, J. (2001). Multivariate and multiscale monitoring of wastewater treatment operation. *Wat. Res.*, 35(14):3402–3410.

Rubio, M., J., C., Ruiz, M., J., C., and Mélendez, J. (2004). Qualitative trends for situations assessment in SBR wastewater treatment process. In *Proceedings of the 4th ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence (BESAI), August 2004, Valencia, Spain.*

Ruiz, M., Colomer, J., Rubio, M., Meléndez, J., and Colprim, J. (2004). Situation assessment of a sequencing batch reactor using multiblock MPCA and fuzzy classification. In *4th ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence (BESAI), Valencia, Spain, August 22-23, 2004*.

Sarolta, A. and Kinley, R. D. (2001). Multivariate statistical monitoring of batch processes: an industrial case study of fermentation supervision. *Trends Biotechnol.*, 19:53–62.

Schölkopf, B., Smola, A., and Müller, K. (1998a). Support vector methods in learning and feature extraction. *Australian Journal on Intelligent Information Processing Systems*, 1:3–9.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998b). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10:1299–1319.

Schuermans, M., Markovsky, I., Wentzell, P., and Van Huffel, S. (2005). On the equivalence between total least squares and maximum likelihood PCA. *Anal. Chim. Acta*, 544:254–267.

Serralta, J., Borrás, L., Blanco, C., Barat, R., and Seco, A. (2004). Monitoring pH and electric conductivity in an EBPR sequencing batch reactor. *Wat. Sci. Technol.*, 50(10):145–152.

Shaich, D., Becker, R., and King, R. (2001). Qualitative modelling for automatic identification of mathematic models of chemical reaction systems. *Control Eng. Practice*, 9:1373–1381.

Shaw, A. and Falrey, A. (2007). Hitting the mark. development and field testing of a new intelligent sequencing batch reactor control system. *Water Environment and Technology (WE&T)*, July 2007:75–80.

Sin, G., Villez, K., and Vanrolleghem, P. (2006). Application of a model-based optimisation methodology for nutrient removing SBRs leads to falsification of the model. *Wat. Sci. Technol.*, 53(4-5):95–103.

Spagni, A., Buday, J., Ratinin, P., and Bortone, G. (2001). Experimental consideration on monitoring ORP, pH, conductivity and dissolved oxygen in nitrogen and phosphorus biological removal processes. *Wat. Sci. Technol.*, 43(11):197–204.

Stephanopoulos, G., Locher, G., Duff, M., Kamimura, R., and Stephanopoulos, G. (1997). Fermentation database mining by pattern recognition. *Biotechnol. Bioeng.*, 53:443–452.

Steyer, J., Bernard, O., and Batstone, D. (2006). Lessons learnt from 15 years of ICA in anaerobic digestion processes. *Wat. Sci. Technol.*, 53(4-5):25–33.

Steyer, J., Rolland, D., Bouvier, J., and Moletta, R. (1997). Hybrid fuzzy neural network for diagnosis – application to the anaerobic treatment of wine distilliery wastewater in a fluidized bed reactor. *Wat. Sci. Technol.*, 36(6-7):209–217.

Steyer, J.-P. and Harmand, J. (2003). *Biotechnology for the Environment: Wastewater Treatment and Modeling, Waste Gas Handling*, chapter Fault detection and isolation in wastewater treatment plants: comparison of different approaches and experimental results, pages 87–100. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Stoodley, K. and Mirnia, M. (1979). The automatic detection of transients, step changes and transient changes in the monitoring of time series. *The Statistician*, 28:163–170.

Strang, G. and Nguyen, T. (1996). *Wavelets and Filter Banks*. Wellesley-Cambridge Press, UK.

Surmacz-Gorska, J., Gernaey, K., Demuynck, C., Vanrolleghem, P., and Verstraete, W. (1996). Nitrification monioring in activated sludge by oxygen uptake rate (our) measurements. *Wat. Res.*, 30:1228–1236.

Sweldens, W. (1996). The lifting scheme: A custom-design construction of biorthogonal wavelets. *Appl. Comput. Harmon. Anal.*, 3:186–200.

Tan, S. and Mavrovouniotis, M. (1995). Reducing data dimensionality through optimizing neural-network inputs. *AIChE J.*, 41:1471–1480.

Third, K., Sepramaniam, S., Tonkovic, Z., Newland, M., and Cord-Ruwisch, R. (2004). Optimisation of storage driven denitrification by using on-line specific oxygen uptake rate monitoring during SND in a SBR. *Wat. Sci. Technol.*, 50(10):171–180.

Tipping, E. and Bishop, C. (1999a). Mixtures of probabilistic principal component analysers. *Neural Comput.*, 11:443–482.

Tipping, E. and Bishop, C. (1999b). Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 61:611–622.

Tomasi, G., van den Berg, F., and Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemometr.*, 18:231–241.

Tona, R., Benqlilou, C., Espuña, A., and Puigjaner, L. (2005). Dynamic data reconciliation based on wavelet trend analysis. *Ind. Eng. Chem. Res.*, 44:4324–4335.

Torrence, C. and Compo, G. P. (1998). A practical guide to wavelet analysis. *Bull. Amer. Metereol. Soc.*, 79:61–78.

Torres, A. and Bertrand-Krajewski, J. (2007). PLS local calibration of a UV-visible spectrometer used for in situ measurements of COD and TSS concentrations in urban drainage systems. In *Proceedings of the 3rd International IWA Conference on Automation in Water Quality Monitoring (AutMoNet2007), Ghent, Belgium, September 5-7, 2007*. Appeared on CD-ROM.

Tracey, N., Young, J., and Mason, R. (1992). Multivariate control charts for individual observations. *J. Qual. Technol.*, 24:88–95.

Traoré, A., Grieu, S., Puig, S., Corominas, L., Thiery, F., Polit, M., and Colprim, J. (2005). Fuzzy control of dissolved oxygen in a sequencing batch reactor plants. *Chem. Eng. J.*, 2005:13–19.

Travé-Massuyès, L., Dague, P., and Guerrin, F. (1997). *Le raisonnement qualitatif*. Editions Hermes, Paris, France.

Ündey, C. and Çinar, A. (2002). Statistical monitoring of multistage, multiphase batch processes. *IEEE Control Syst. Mag.*, 22:40–52.

van den Broeke, J., Langergraber, G., and Weingartner, A. (2006). On-line and in situ UV/vis spectroscopy for multi-parameter measurements: a brief review. *Spectroscopy Europe*, 18(4):15–18.

van den Broeke, J., Ross, P., van der Helm, A., Baars, E., and Rietveld, L. (2007). Use of on-line uv/vis-spectrometry in the measurement of dissolved ozone and AOC concentrations in drinking water treatment. In *Proceedings of the 3rd International IWA Conference on Automation in Water Quality Monitoring (AutMoNet2007), Ghent, Belgium, September 5-7, 2007*. Appeared on CD-ROM.

van Sprang, E., Ramaker, H.-J., Westerhuis, J., Gurden, S., and Smilde, A. (2002). Critical evaluation of approaches for on-line batch process monitoring. *Chem. Eng. Sci.*, 57:3379–3991.

Vanrolleghem, P. and Coen, F. (1995). Optimal design of in-sensor-experiments for on-line modelling of nitrogen removal processes. *Wat. Sci. Technol.*, 31(2):149–160.

Vanrolleghem, P. and Van Daele, M. (1994). Optimal experimental design for structure characterization of biodegradation models: on-line implementation in a respirographic biosensor. *Wat. Sci. Technol.*, 30(4):243–253.

Venkatasubramanian, V., Rengaswamy, R., and Kavuri, S. (2003a). A review of process fault detection and diagnosis Part I: Quantitative model-based methods. *Comput. Chem. Eng.*, 27:293–311.

Venkatasubramanian, V., Rengaswamy, R., and Kavuri, S. (2003b). A review of process fault detection and diagnosis Part II: Qualitative models and search strategies. *Comput. Chem. Eng.*, 27:313–326.

Venkatasubramanian, V., Rengaswamy, R., and Kavuri, S. (2003c). A review of process fault detection and diagnosis Part III: Process history based methods. *Comput. Chem. Eng.*, 27:327–346.

Villez, K., Pelletier, G., Rosén, C., Anctil, F., Duchesne, C., and Vanrolleghem, P. (2007a). Comparison of two wavelet-based tools for data mining of urban water networks time series. *Accepted to Wat. Sci. Technol.*

Villez, K., Rosén, C., Anctil, F., Duchesne, C., and Vanrolleghem, P. (2007b). Qualitative representation of trends: an improved method for multiscale extraction of trends from process data. *Submitted to IEEE Trans. Signal Process.*

Vives, M., Balaguer, M., García, S., García, R., and Colprim, J. (2003). Textile dyeing wastewater treatment in a sequencing batch reactor system. *J. Environ. Sci. Health Part A-Toxic/Hazard. Subst. Environ. Eng.*, A38:2089–2099.

Wang, S., Gao, D., Peng, Y., Wang, P., and Yang, Q. (2004). Nitrification-denitrification via nitrite for nitrogen removal from high nitrogen soybean wastewater with on-line fuzzy control. *Wat. Sci. Technol.*, 5-6:121–127.

Wang, X. and Li, R. (1999). Combining conceptual clustering and principal component analysis for state space based process monitoring. *Ind. Eng. Chem. Res.*, 38:4345–4358.

Wareham, D., Hall, K., and Mavinic, D. (1993). Real-time control of wastewater treatment systems using ORP. *Wat. Sci. Technol.*, 28(11-12):273–282.

Wareham, D., Mavinic, D., and Hall, K. (1994). Sludge digestion using ORP-regulated aerobic-anoxic cycles. *Wat. Res.*, 28:373–384.

Watts, J. and Garber, W. (1993). On-line respirometry: a powerful tool for activated sludge plant operation and design. *Wat. Sci. Technol.*, 28(11-12):389–399.

Watts, J. and Garber, W. (1995). Respirometric control of the activated sludge process. In *Proceedings of the IAWQ Specialized Conference on Sensors in Wastewater Technology, Copenhagen, Denmark, 1995*.

Wentzell, P., Andrews, D., Hamilton, D., Faber, K., and Kowalski, B. (1997). Maximum likelihood principal component analysis. *J. Chemometr.*, 11:339–366.

Wentzell, P. and Lohnes, M. (1999). Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations. *J. Chemometr.*, 45:65–85.

Westerhuis, J., Kourti, T., and MacGregor, J. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemometr.*, 12:301–312.

Widodo, A. and Yang, B.-S. (2007). Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors. *Expert Syst. Appl.*, 33:241–250.

Wiesmann, U., Choi, I., and Dombrowski, E.-M. (2006). *Fundamentals of biological wastewater treatment*. John Wiley and Sons Ltd, Germany.

Wilderer, P., Irvine, R., and Goronszy, M. (2001). Sequencing Batch Reactor Technology. Technical report, IWA Scientific and Technical Report No:10.

Winkler, S., Bertrand-Krajewski, J., Torres, A., and E., S. (2007). Benefits, limitations and uncertainty of in-situ spectrometry. In *Proceedings of the 3rd International IWA Conference on Automation in Water Quality Monitoring (AutMoNet2007), Ghent, Belgium, September 5-7, 2007*. Appeared on CD-ROM.

Wold, S. (1976). Pattern recognition by means of disjoint principal components models. *Pattern Recognit.*, 8:127–139.

Wold, S. (1994). Exponentially weighted moving principal component analysis and projections to latent structures. *Chem. Intell. Lab. Syst.*, 23:149–161.

Wold, S., Geladi, P., Esbensen, K., and Öhman, J. (1987). Multi-way principal components and PLS-analysis. *J. Chemometr.*, 1:47–56.

Wold, S., Kettaneh, N., Fridén, H., and Holmberg, A. (1998). Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chem. Intell. Lab. Syst.*, 44:331–340.

Wold, S., Kettaneh, N., and Tjessem, K. (1996). Hierarchical multi-block PLS and PC models, for easier interpretation, and as an alternative to variable selection. *J. Chemometr.*, 10:463–482.

Wouters-Wasiak, K., Heduit, A., Audic, J., and Lefevre, F. (1994). Real-time control of nitrogen removal at full-scale using oxidation reduction potential. *Wat. Sci. Technol.*, 30(4):207–210.

Yamanaka, F. and Nishiya, T. (1997). Application of the intelligent alarm system for the plant operation. *Comput. Chem. Eng.*, 21:S625–S630.

Yoo, C., Lee, D., and Vanrolleghem, P. (2004). Application of mulitway ICA for on-line process monitoring of a sequencing batch reactor. *Wat. Res.*, 38:1715–1732.

Yoo, C., Villez, K., Lee, I.-B., C., R., and Vanrolleghem, P. (2006a). Multi-model statistical process monitoring and diagnosis of a sequencing batch reactor. *Biotechnol. Bioeng.*, 96:687–701.

Yoo, C., Villez, K., Lee, I.-B., and Vanrolleghem, P. (2006b). Multivariate nonlinear statistical process control of a sequencing batch reactor. *J. Chem. Eng. Jpn.*, 1:43–51.

Yu, R., Liaw, S., Cho, B.-C., and Yang, S. (2001). Dynamic control of a continuous-inflow SBR with time-varying influent loading. *Wat. Sci. Technol.*, 43:107–114.

Yuan, Z., Keller, J., and Lant, P. (2003). *Biotechnology for the Environment: Wastewater Treatment and Modeling, Waste Gas Handling*, chapter Optimization and control of nitrogen removal activated sludge processes: a review of recent developments, pages 3c: 187–227. Kluwer Academic Publishers, Dorecht, The Netherlands.

Zhang, J., Martin, E., and Morris, A. (1997). Process monitoring using non-linear statistical techniques. *Chem. Eng. J.*, 67:181–189.

Zipper, T., Fleishmann, N., and Haberl, R. (1998). Development of a new system for control and optimization of small wastewater treatment plants using oxidation-reduction potential (ORP). *Wat. Sci. Technol.*, 38(3):307–314.

Summary

This thesis deals with the development, application and validation of techniques for data analysis in view of supervisory control of cyclic systems, including and integrating aspects of monitoring, diagnosis and control. Two so far largely separated tools for data mining of process data are used as a basis for the presented developments. These are Principal Component Analysis (PCA) and Qualitative Representation of Trends (QRT). A pilot-scale sequencing batch reactor (SBR) for wastewater treatment is used as a case study for the major parts of the work presented. Another application is pursued regarding the analysis of flow measurement time series derived from an urban drinking water network.

The pilot-scale SBR setup studied throughout the work has served as a valid source of data-rich information-poor data sets, which are consequently used for evaluation of several data analysis tools for process supervision and control. It is observed that the design of the given experimental unit is suboptimal in view of the desire to discriminate between physical and biological failures of the system. In fact, an impending need exists to detect and diagnose biological faults, and the intent was to develop and evaluate techniques for this purpose. Several remarks and suggestions for improved design of experimental setups are given.

The first of the evaluated approaches for data analysis, Principal Component Analysis, is popularly reported as a straightforward method to deal with correlated measurements in the context of process monitoring and diagnosis. In view of fundamental limitations of the original (linear) PCA technique, a myriad of extensions and adjustments that deal with non-linearity, dynamics, natural changes of processes and the typical three-dimensional nature of batch process data (i.e. batch index, time-in-batch and measured variable), are already presented in literature. However, a lack of proper evaluation, validation and comparative studies is observed. In addition, the concepts of Maximum Likelihood and bias-variance trade-off, already theoretically presented for PCA modelling excercises, have so far been left unattended at large by the process monitoring research community.

A conventional extension for batch process data, Multi-way PCA (MPCA), is evaluated for process monitoring of both the hydraulic parts of the system only as well as for the multivariate data of the complete SBR system. For many types of faults, good performance rates are obtained by means of the MPCA models. In this, the performance proves to be rather insensitive to choices in the modelling stage such as the chosen approach for scaling of the data. Peculiar types of faults, which are of a time-local or frequency-local nature, are not likely detected by the presented approach. It is therefore concluded that process monitoring should integrate techniques that allow the detection of such events, e.g. by making use of the wavelet framework already presented and applied for process monitoring. Also, suggestions for improvement of PCA modelling practice, not limited to the case study, are added.

Modelling by means of MPCA is also used as a basis for process diagnosis. Explorative treatment of the diagnosis problem showed that MPCA is a straightforward tool to visualize and recognize faults. Despite this result, automatisation of this recognition process by means of combination of MPCA and fuzzy clustering is found to be applicable to a limited extent and therefore remains suitable for further research. Suggested modifications may lead to increased interpretability of the models, more generalized and appropriate representation of reality and, overall, improvement of diagnosis performance.

SBR systems are characterized by their cyclic operation and the recognition of several phases that constitute the cycles. The time length of cycles and phases is not necessarily the same for each cycle which allows large flexibility in design and operation. Generally speaking, the determination of the optimal lengths is not an easy task. Therefore, a new control scheme for phase length optimization is proposed, applied in real-time and evaluated as successful. In the proposed scheme, the Hotelling's T^2 statistic, which is also defined in the context of PCA, is used to define a region in the data space, corresponding to the targeted completion of biochemical reactions. If sufficient consecutive multivariate data samples are found within this region, the necessary reactions are judged to be complete signifying that the running phase can be shut down safely. The control scheme proved technically successful and is shown to lead to significant improvements in effluent quality of the studied SBR system, hereby representing a clear proof of concept.

The second approach used for automated data analysis is taken from the field of qualitative analysis. Qualitative analysis aims at the qualitative description of data and has so far largely been concentrated on the qualitative representation of time

series. Following a review and evaluation of available techniques, it is concluded that proper identification of inflection points is not possible on the basis of the most generic techniques available in literature. One technique, based on the cubic spline wavelet decomposition, is selected and improved in such a way that inflection point detection is possible in a more reliable and consistent way. Suggestions for other observed problems, such as the inability to appropriately detect jump changes by means of the cubic spline wavelet method and the lack of robustness associated with the interval-halving approach found in literature, are proposed.

The qualitative analysis framework is succesfully evaluated for data mining of time series typical for urban drinking water networks. Qualitative analysis proved to reveal details of the studied time series that are not obvious from the more classic wavelet spectrum analysis technique. In an application to the studied SBR system, the qualitative analysis technique is integrated into a preliminary design for closed loop supervisory control on the basis of qualitative analysis. To this end, two simple tables, one connecting the qualitative analysis results with corresponding diagnostics and another connecting the diagnostics with proposed control actions, are devised.

In the closing chapter *Conclusions and perspectives*, the decision to develop monitoring schemes in the absence of knowledge, which was fundamental to the earlier choice for data-driven approaches, is evaluated critically. It is indicated that the presumed absence of knowledge proves not to correspond to the now historical reality. Based on the latter observation, mixed approaches, taking the best of knowledge-driven and data-driven approaches, are motivated for future practice in process analysis and supervision. In addition, several suggestions are given with respect to multivariate and qualitative analysis, including improvements and extensions of currently available techniques as well as new applications and opportunities for data analysis.

Samenvatting

Deze thesis handelt over de ontwikkeling, toepassing en validatie van technieken voor data-analyse met het oog op superviserende regeling van cyclische systemen waarbij de integratie van opvolging, diagnose en regeling niet uit het oog wordt verloren. Twee tot dus ver grotendeels gescheiden technieken voor data mining van procesdata zijn gebruikt als basis voor de ontwikkelingen tijdens het doctoraatsonderzoek. Deze zijn Principale Component Analyse (PCA) en Kwalitatieve Representatie van Trends (Qualitative Representation of Trends, QRT). Een pilootschaal Sequentiële Batch Reactor (SBR) voor waterzuivering dient tot een gevalstudie voor de grootste delen van het werk. Een andere toepassing betrof de analyse van tijdreeksen van debietmetingen die in een stedelijk drinkwaternetwerk werden verzameld.

De pilootschaal SBR opstelling die gebruikt is in dit werk heeft gediend tot de winning van data-rijke informatie-arme data sets die gebruikt zijn voor de evaluatie van verscheidene technieken voor data-analyse in functie van procesopvolging en -regeling. Er is vastgesteld dat het ontwerp van de experimentele eenheid suboptimaal was gegeven de wens om fysisch-technische en biologische fouten van elkaar te kunnen onderscheiden. De bedoelde ontwikkeling en evaluatie van technieken was echter bedoeld om te voldoen aan de dringende nood om biologische fouten te kunnen detecteren en herkennen. Verschillende bemerkingen en suggesties voor verbeterd ontwerp van experimentele opstellingen zijn meegegeven in dit werk.

De eerste geëvalueerde benadering voor data analyse, Principale Component Analyse, is vaak vermeld als een eenvoudige techniek om met gecorreleerde data om te gaan in de context van procesopvolging en -diagnose. Gezien een aantal fundamentele beperkingen van de originele (lineaire) PCA techniek zijn een overvloed aan uitbreidingen en aanpassingen beschreven in de literatuur, onder meer om met niet-lineariteit, dynamiek, natuurlijke procesveranderingen en de typische driedimensionele aard van data van batchprocessen (i.e. batch index, tijd-in-batch en gemeten variabele) te kunnen omgaan. Effectieve evaluatie, validatie en vergelijkende studie van deze uitbreidingen bleef echter tot zover uit. Meer nog, bepaalde concepten, zoals Maximale Waarschijnlijkheid (Maximum Likelihood) en de af-

weging van bias en variantie, die reeds theoretisch uitgewerkt zijn voor PCA zijn tot zover onaangeroerd door de onderzoeksgemeenschap rond procesopvolging.

Een conventionele uitbreiding voor data van batch processen, Multi-way PCA (MPCA), is geëvalueerd voor procesopvolging van zowel de hydraulische delen als het volledige SBR systeem. Voor vele fouttypes is een goede performantie vastgesteld door middel van MPCA modellering. Er is vastgesteld dat de performantie ongevoelig is voor geëvalueerde keuzes in voorbehandeling van de data, zoals herschaling van de data. Bijzondere types fouten, die een tijdsgebonden of frequentiegebonden karakter hebben, worden niet gauw gedetecteerd met de gekozen benadering. Daarom is besloten dat de integratie van technieken die de detectie van dergelijke gebeurtenissen toch mogelijk maken noodzakelijk is voor procesopvolging, bij voorbeeld door gebruik te maken van het wavelet-raamwerk dat reeds in andere toepassingdomeinen is voorgesteld en toegepast voor procesopvolging. Ook zijn suggesties gemaakt met betrekking tot verbeterde PCA-modellering die niet beperkt zijn tot de beschreven gevalstudie.

Modellering door middel van MPCA is ook gebruikt als basis voor procesdiagnose. Exploratieve benadering van het diagnose-probleem toonde aan dat MPCA een handige manier is om fouten te visualizeren en te herkennen. Automatisatie van dit herkenningsproces door combinatie van MPCA en fuzzy clustering is echter in beperkte mate mogelijk en blijft daarom nog voor verder onderzoek geschikt. Gesuggereerde wijzigingen kunnen leiden tot een betere interpretatie van de resulterende modellen, beter passende en meer algemeen geldende beschrijvingen van de realiteit en, algemeen, betere performantie.

SBR systemen worden gekarakteriseerd door hun cyclische procesvoering waarbij verschillende fasen in de cycli kunnen herkend worden. De lengte in de tijd van de cycli en fasen is niet noodzakelijk dezelfde voor elke cyclus wat een grote flexibiliteit toelaat. Het vaststellen van optimale lengtes is echter geen sinecure. Daarom is een nieuw regelschema voor optimalisatie van faselengtes voorgesteld dat vervolgens is toegepast in real-time en succesvol geëvalueerd. In het voorgestelde schema wordt de Hotelling's T^2 statistiek –eveneens gedefinieerd is in het kader van PCA– gebruikt om een regio in de multivariate data-ruimte te definiëren die overeenkomt met de na te streven voleindiging van biochemische reacties. Wanneer voldoende opeenvolgende waarnemingen in deze regio zijn gevonden wordt geoordeeld dat de nodige reacties beëindigd zijn en dat als gevolg daarvan de lopende fase kan afgesloten worden. Het is aangetoond dat het regelschema met succes faselengtes kan verkorten en dat dit in de gedemonstreerde toepassing

leidde tot een significante verbeteringen van de effluent kwaliteit, waarmee een duidelijk conceptbewijs is geleverd.

De tweede benadering die voor automatische data-analyse is aangewend is in de discipline van de kwalitatieve analyse gevonden. Kwalitatieve analyse doelt op de kwalitatieve beschrijving van data en bleef tot zover geconcentreerd rond de beschrijving van tijdreeksen. Door middel van een overzicht en evaluatie van de beschikbare technieken is besloten dat gepaste identificatie van buigpunten in tijdreeksen niet mogelijk is op basis van de meest algemeen toepasbare technieken die beschreven zijn in de literatuur. Eén techniek, gebaseerd op de kubische spline wavelet ontleding, is geselecteerd en verbeterd zodat detectie van buigpunten mogelijk is op een betrouwbare en consistente manier. Suggesties voor andere vastgestelde problemen, zoals de onmogelijkheid om stapvormige veranderingen te identificeren door middel van de kubische spline wavelet en het gebrek aan robuustheid dat gekopped is aan de alternatieve techniek op basis van halvering van tijdsintervallen, zijn gesuggereerd.

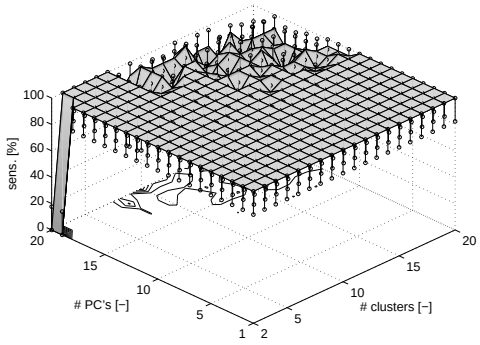
Het raamwerk voor kwalitatieve analyse is succesvol geëvalueerd voor data mining van tijdreeksen typisch voor stedelijke drinkwaternetwerken. Het is aangetoond dat kwalitatieve analyse details in de bestudeerde tijdreeksen belicht die niet door middel van de meer klassieke wavelet spectrum analyse duidelijk worden. De techniek voor kwalitatieve analyse werd ook geïntegreerd in een preliminair ontwerp voor gesloten-kring superviserende regeling van het bestudeerde SBR systeem. In dit initieel ontwerp zijn eenvoudige tabellen gebruikt om enerzijds de resultaten van de kwalitatieve analyse te verbinden met de overeenkomstige diagnostiek en anderzijds de diagnostiek te verbinden met de remediërende regelacties.

In het afsluitende hoofstuk *Conclusies en perspectieven* werd een kritische evaluatie gemaakt van de beslissing om schema's voor opvolging te ontwikkelen zonder kennis-input, wat een fundamentele reden was voor de keuze voor data-gedreven technieken in dit werk. Het is aangetoond dat de veronderstelde afwezigheid van kennis niet overeenstemt met de nu historische werekelijkheid. Gebaseerd op deze vaststelling, worden gemengde benaderingen, die het beste van kennis-gedreven en data-gedreven benaderingen bundelen, gemotiveerd voor toekomstige toepassing in proces-analyse en -supervisie. Meer nog, verschillende suggesties worden meegeven met betrekking tot multivariate en kwalitatieve analyse, onder meer verbeteringen en uitbreidingen van beschikbare technieken alsook nieuwe toepassingen en opportuniteiten voor data-analyse.

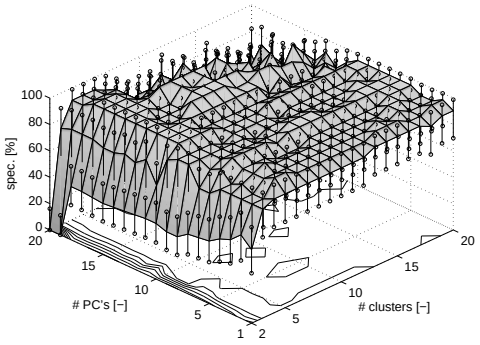
Appendix A

Sensitivity and specificity plots with respect to diagnosis of the hydraulic parts of the SBR system

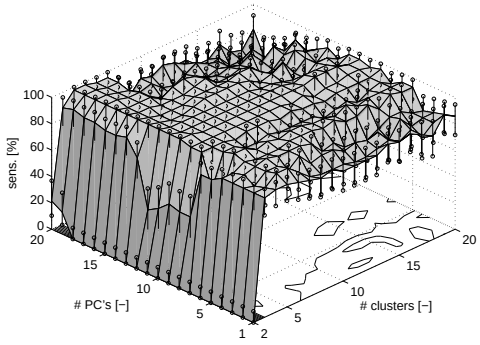
In the following graphs, computed sensitivities and specificities are plotted as surfaces. The respective confidence intervals, based on the binomial model, are indicated with dots.



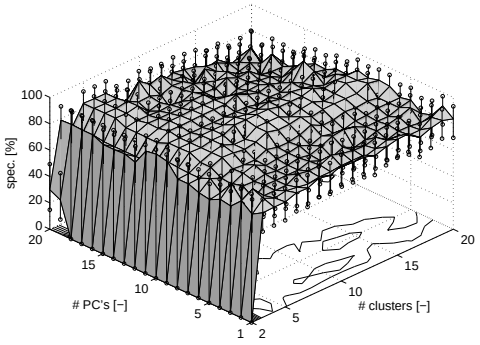
Sensitivity for fault class 1



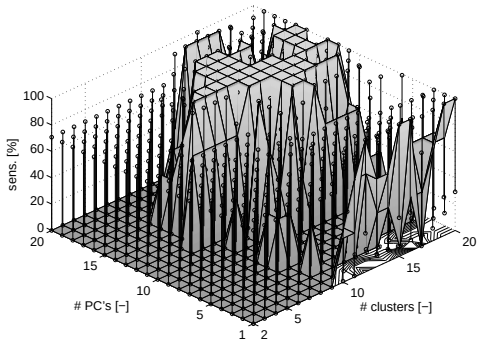
Specificity for fault class 1



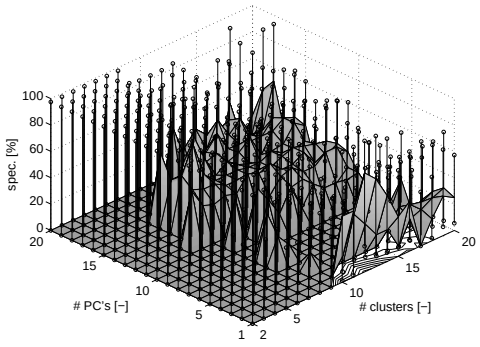
Sensitivity for fault class 2



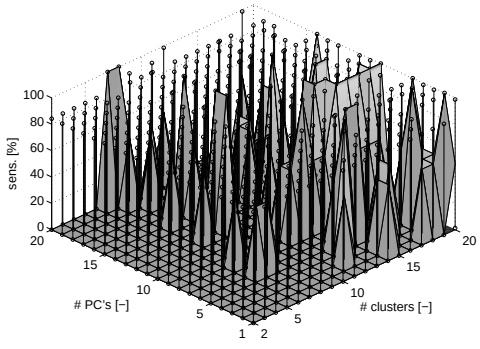
Specificity for fault class 2



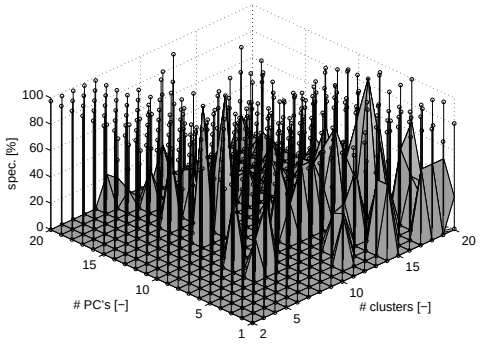
Sensitivity for fault class 3



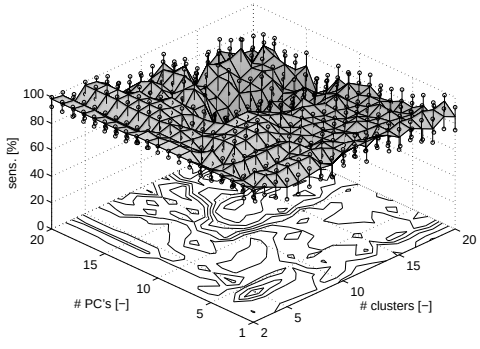
Specificity for fault class 3



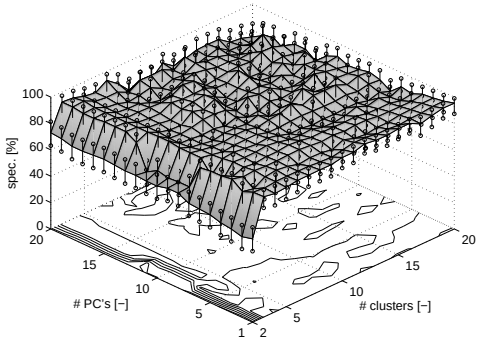
Sensitivity for fault class 5



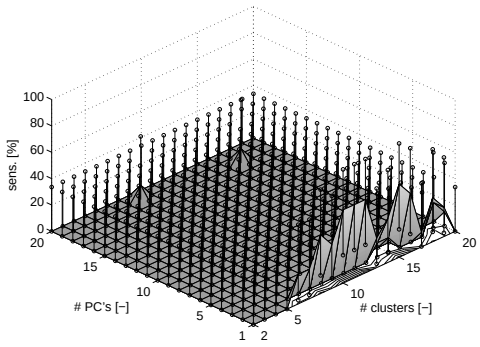
Specificity for fault class 5



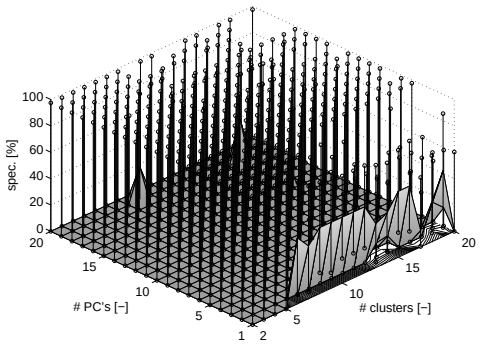
Sensitivity for fault class 6



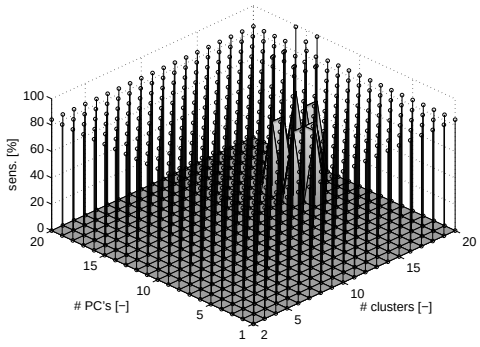
Specificity for fault class 6



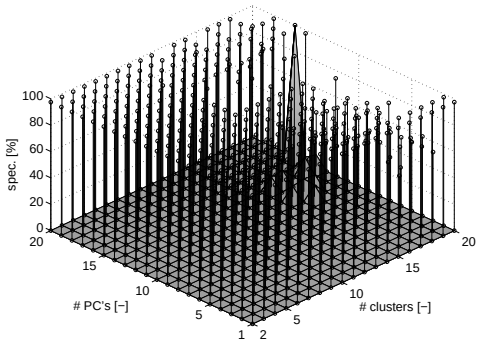
Sensitivity for fault class 7



Specificity for fault class 7



Sensitivity for fault class 8

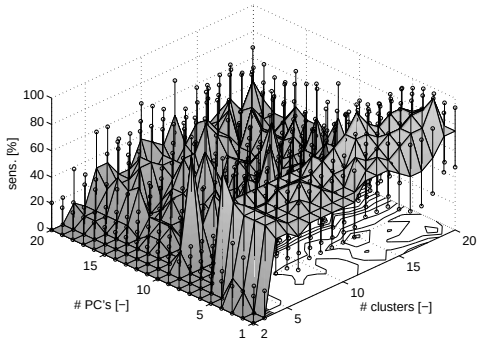


Specificity for fault class 8

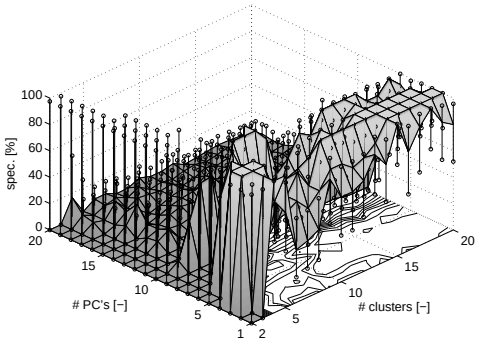
Appendix B

Sensitivity and specificity plots with respect to diagnosis of the complete SBR system

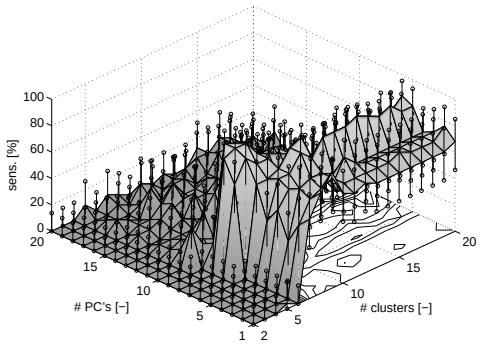
In the following graphs, computed sensitivities and specificities are plotted as surfaces. The respective confidence intervals, based on the binomial model, are indicated with dots.



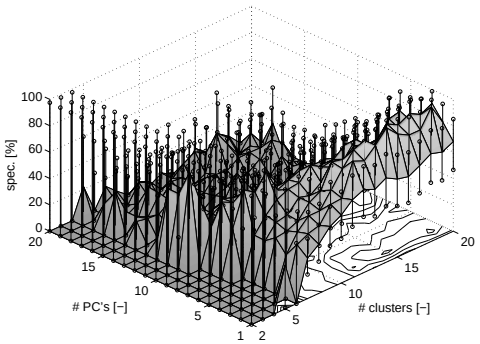
Sensitivity for fault class 1



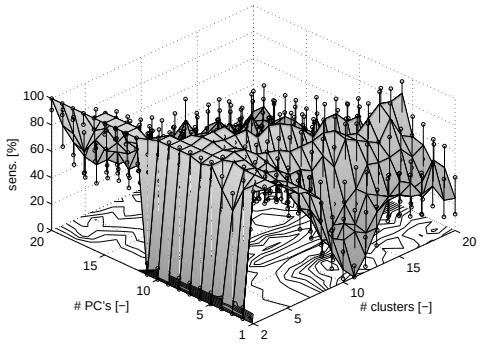
Specificity for fault class 1



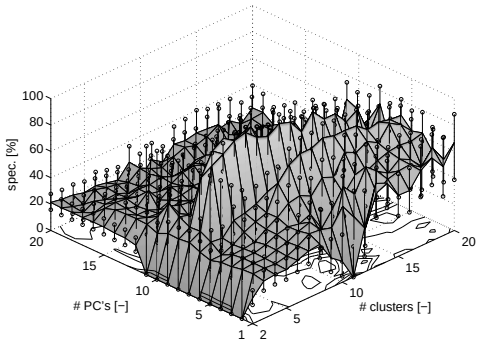
Sensitivity for fault class 2



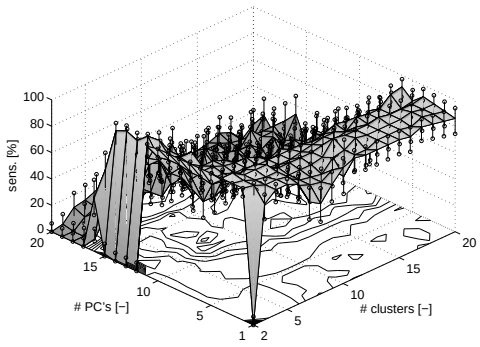
Specificity for fault class 2



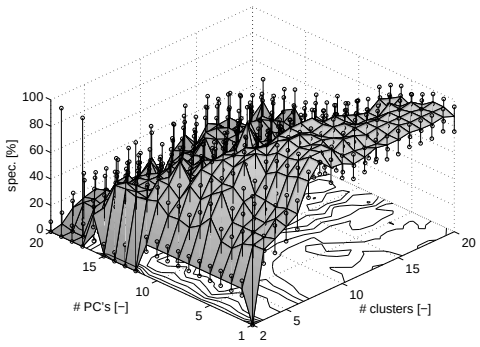
Sensitivity for fault class 6



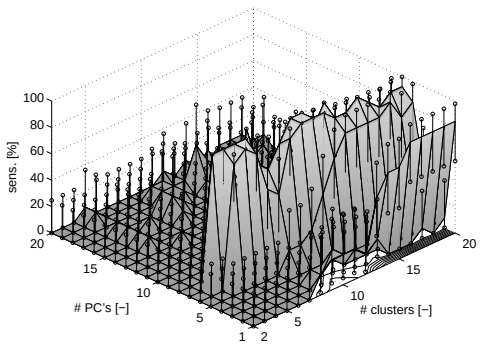
Specificity for fault class 6



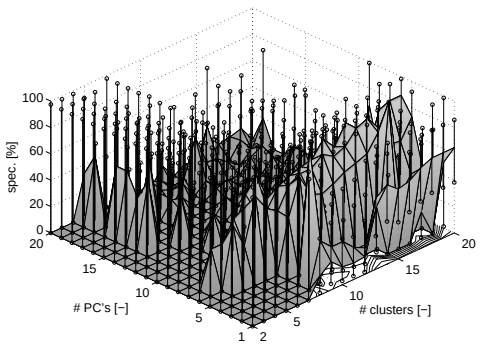
Sensitivity for fault class 9



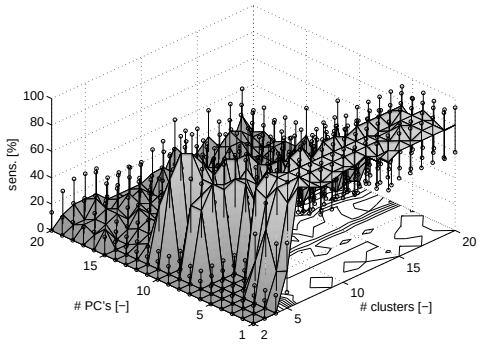
Specificity for fault class 9



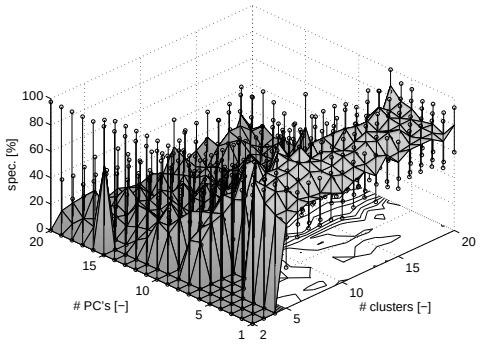
Sensitivity for fault class 12



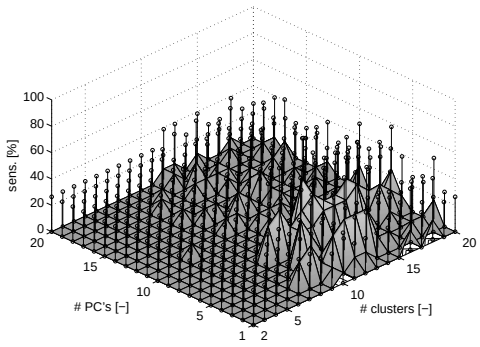
Specificity for fault class 12



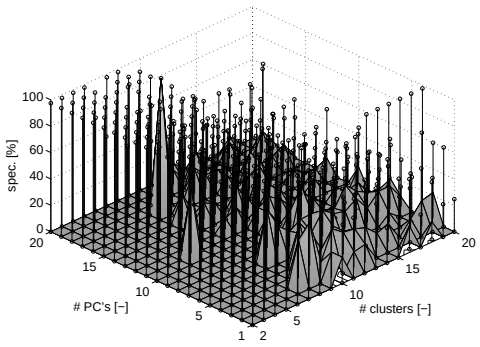
Sensitivity for fault class 13



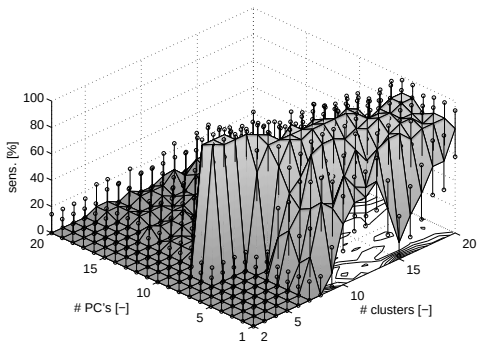
Specificity for fault class 13



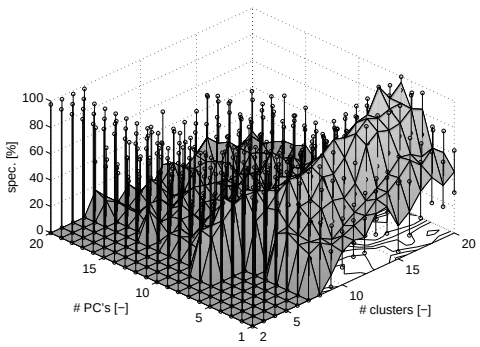
Sensitivity for fault class 14



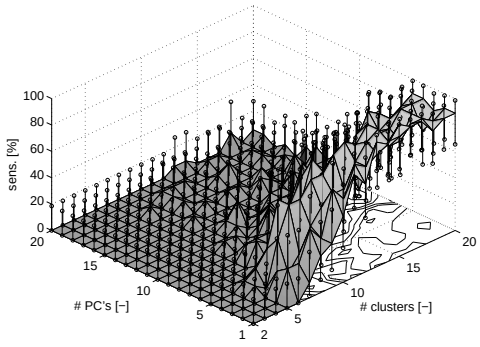
Specificity for fault class 14



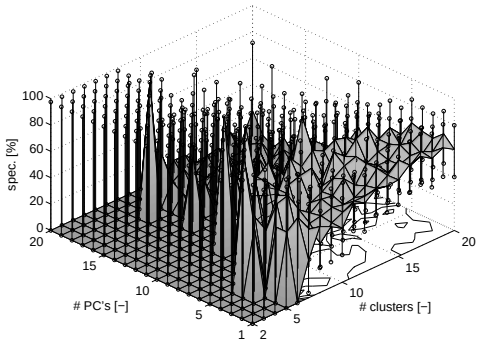
Sensitivity for fault class 16



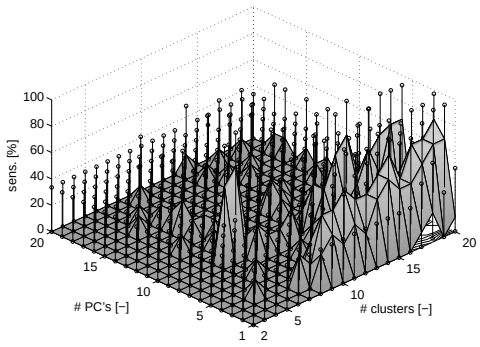
Specificity for fault class 16



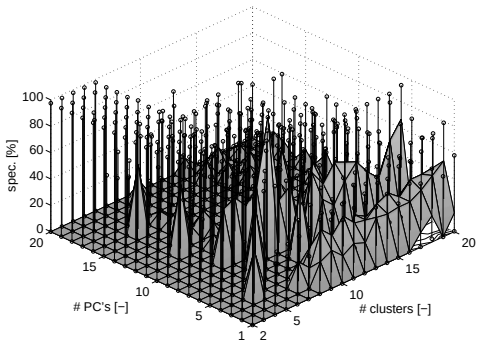
Sensitivity for fault class 18



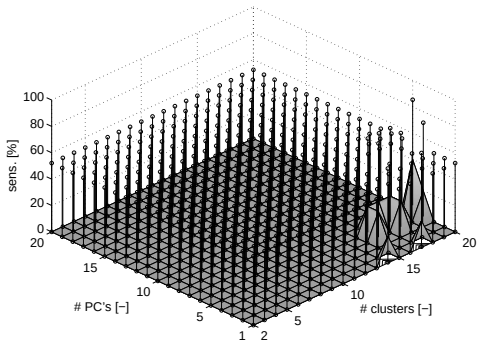
Specificity for fault class 18



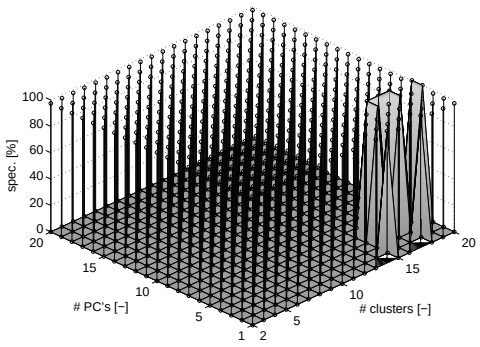
Sensitivity for fault class 22



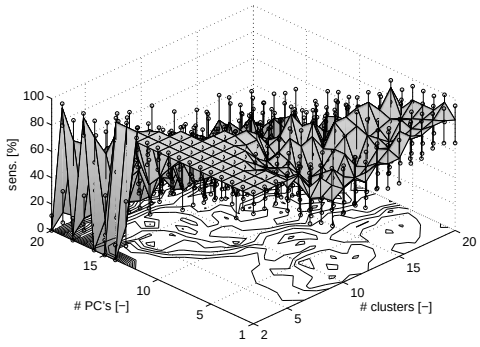
Specificity for fault class 22



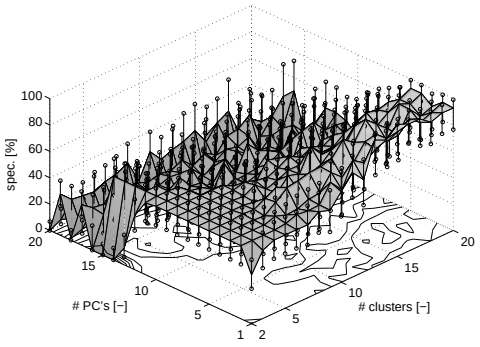
Sensitivity for fault class 24



Specificity for fault class 24



Sensitivity for fault class 29



Specificity for fault class 29

Curriculum vitae

Personalia

Name	Kris Villez
Sex	man
Nationality	Belgium
Birth date	01/10/1980
Birth place	Izegem, Belgium
e-mail adress	kris.villez@BIOMATH.UGent.be

Education

10/98 – 07/03	Bio-engineer in environmental technology (distinction), Ghent University **Thesis**: Simulatie- en experimentele studie van het SHARON-proces voor koppeling met een Anammox-eenheid *Simulation and experimental study of the SHARON-process for coupling with an Anammox unit*, Promotor: Prof. Dr. ir. P. Vanrolleghem

Career and work experience

12/03 – 12/07	IWT-funded PhD student at the Department of Applied Mathematics, Biometrics en Proces Control (BIOMATH), FBW, Ghent University Research regarded computer-integrated monitoring, diagnosis and control of biological processes such as the SHARON process, SBR's for nutrient removal and living crop cultures.

Papers in journals with peer review and listed by Web of Science

- Sin, G, Villez, K. and Vanrolleghem, P.A. (2006). Application of a model-based optimisation methodology for nutrient removing SBRs leads to falsification of the model. *Water Science and Technology* 53(4–5), 95-103.
- Yoo, C.K., Villez, K., Lee, I.B., Van Hulle, S. and Vanrolleghem, P.A. (2006). Sensor validation and reconciliation for a partial nitrification process. *Water Science and Technology* 53(4–5), 513–521.
- 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.K., Villez, K., Lee, I.B., Rosén, C. and Vanrolleghem, P.A. (2007). Multi-model statistical process monitoring and diagnosis of a sequencing batch reactor. *Biotechnology and Bioengineering* 96(4), 687–701.
- Yoo, C.K., Villez, K., Van Hulle, S. and Vanrolleghem, P.A. (2007). Enhanced process monitoring for wastewater treatment systems. Accepted to *Environmetrics*.
- Van Hulle, S.W.H., Van den Broeck, S., Maertens, J., Villez, K., Donckels, B.M.R., Schelstraete, G., Volcke, E.I.P., Vanrolleghem, P.A. (2005). Construction, start-up and operation of a continuously aerated laboratory- scale SHARON reactor in view of coupling with an Anammox reactor. *Water SA* 31, 317–334.

Papers in other journals with peer review

- Villez, K., Pelletier, G., Rosén, C., Anctil, F., Duchesne, C., Vanrolleghem, P.A. (2007). Comparison of two wavelet-based tools for data mining of urban water networks time series. *Water Science and Technology* 56(6), 57–64.
- Villez, K., Rosén, C., Anctil, F. and Duchesne, C. and Vanrolleghem, P.A. (2007). Qualitative representation of trends: an alternative approach to online process diagnosis and control of SBR's for nutrient removal. Accepted to *Water Science and Technology*.
- Villez, K., Ruiz, M., Sin, G., Rosén, C., Colomer, J. and Vanrolleghem, P.A. (2007). Combining Multiway principal Component Analysis (MPCA) and clustering for efficient data mining of historical data sets of SBR processes. Accepted to *Water Science and Technology*.

Papers in journals without peer review

- Nopens, I., Villez, K., Rieger, L. and Vanrolleghem P.A. (2004). Monitoring the water cycle - state of the art and future needs. *IWA Yearbook 2007*, 33–36.
- Volcke, E.I.P., Villez, K., Van Hulle, S. en Vanrolleghem P.A. (2004). Wat met rejectiewater? *What with rejection water? Afvalwaterwetenschap*, 3(4), 297–318.

Proceedings

- Villez, K., Rosén, C., Van Hulle, S., Yoo, C.K., Nopens, I. and Vanrolleghem P.A. (2005). On-Line Dynamic Monitoring of the SHARON Process for sustainable nitrogen removal from wastewater. In: *Proceedings of the 15th European Symposium on Computer Aided Process Engineering (ESCAPE15)*, Barcelona, Spain, May 29–June 1, 2005, 1297-1302.
- Villez, K., Lee, D.S., Rosén, C., Vanrolleghem, P.A. (2006). Comparison of linear and non-linear PLS methods for soft-sensing of an SBR for nutrient removal. In: *Proceedings of the 3rd Biennial meeting of the International Environmental Modelling and Software Society (iEMSs2006)*, Burlington, Vermont, USA, July 9–12, 2006, appeared on CD-ROM.

Conferences and symposia (oral contribution)

- Watermatex 2004, 6th International Symposium on Systems Analysis and Integration Assessment, Beijing, China, November 3–5, 2004.
- BNR2005: IWA Specialized Conference on Nutrient Management in Wastewater Treatment Processes and Recycle Streams, Krakow, Poland, September 19–21, 2005.
- Watermatex 2007: 7th International IWA Symposium on Systems Analysis and Integrated Assessment in Water Management, Washington DC, USA, May 7–9, 2007.
- AutMoNet2007: 3rd International IWA Conference on Automation in Water Quality Monitoring, Ghent, Belgium, May 7–9, 2007.

Conferences, seminars and symposia (poster contribution)

- PLS05: 4th International Symposium on PLS and Related Methods, Barcelona, Spain, September 7–9, 2005.
- ICA2005: 2nd IWA Conference on Instrumentation, Control and Automation, Busan, Republic of Korea, May 29-June 2, 2005.

Conferences, seminars and symposia (workshop contribution)

- iEMSs2006: 2006 Summit on Environmental Modelling and Software, Burlington, Vermont, USA, July 9–13, 2006.

Conferences, seminars and symposia (participation)

- Chimiométrie 2006, Paris, France, November 30–December 1, 2006.
- HIW07: International Workshop on Advances in Hydroinformatics 2007, Niagara Falls, Canada, June 4–7, 2007.
- BIRA Symposium On-line Afvalwateranalyses *(BIRA Symposium On-line Wastewater Analyses)*, Antwerp, Belgium, November 17, 2005.

- APCRE'05: 4th Asia-Pacific Chemical Reaction Engineering Symposium, Gyeongju, Republic of Korea, June 12–15, 2005.

Conferences, seminars and symposia (organisation)

- AutMoNet2007: 3rd International IWA Conference on Automation in Water Quality Monitoring, Ghent, Belgium, May 7–9, 2007. Member of the organising committee.
- STIC & Environnement 2007: Sciences et Techniques de l'Information et de la Communication pour l'Environnement. Member of the scientific committee.
- WWTmod2008: Wastewater Treatment Modelling Seminar. Mont-Sainte-Anne, Quebec, Canada, June 1–3, 2008. Member of the scientific committee.
- WWC2008: IWA World Water Congress and Exhibition 2008. Vienna, Austria, September 7–12, 2008. Member of the scientific committee.

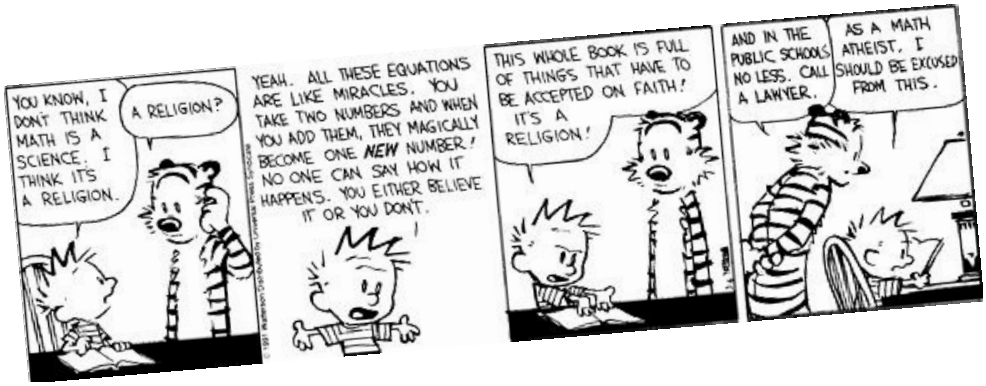
Reviewed papers of International journals

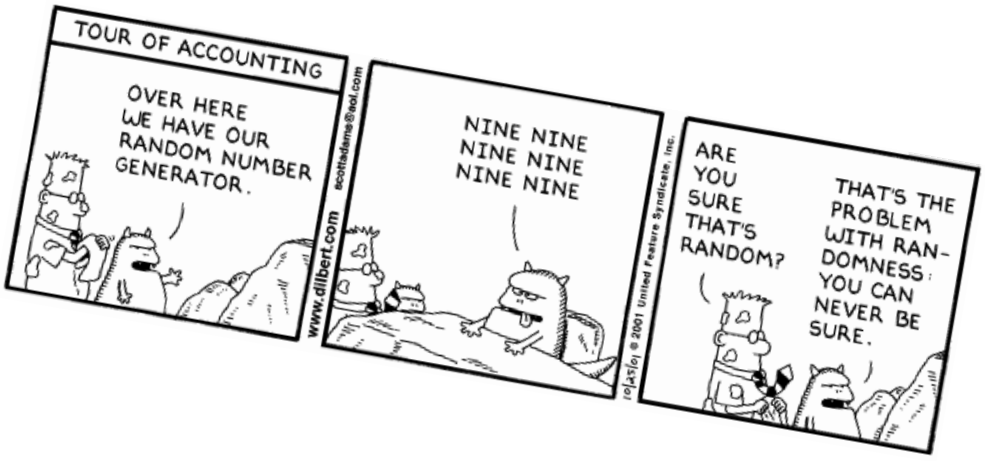
- Control and Engineering Practice: 1 paper
- Process Biochemistry: 1 paper
- Water Research: 1 paper

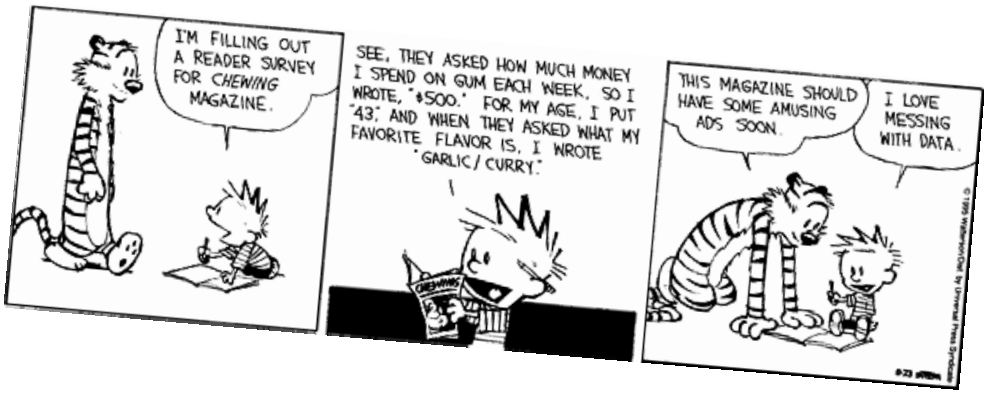
Didactic activities

2004–2005 Guidance of Roman Raaymakers while writing his thesis *On-line monitoring van een Sequencing Batch Reactor (SBR) (On-line monitoring of a Sequencing Batch Reactor (SBR))* in fulfilment of the requirements for the degree of Licenciado en Ciencias químicas (Graduate in Chemical Sciences), FBE, Ghent University.

2005–2006 Guidance of Eline d'Hooge while writing her thesis *On-line multivariate statistische procescontrole van een Sequencing Batch Reactor (SBR) voor biologische nutriënt-verwijdering (On-line multivariate statistical process control of a Sequencing Batch Reactor (SBR) for biological nutrient removal)* in fulfilment of the requirements for the degree of Bio-engineer in environmental technology, FBE, Ghent University.


YOU KNOW, I DON'T THINK MATH IS A SCIENCE. I THINK IT'S A RELIGION.
A RELIGION?
YEAH. ALL THESE EQUATIONS ARE LIKE MIRACLES. YOU TAKE TWO NUMBERS AND WHEN YOU ADD THEM, THEY MAGICALLY BECOME ONE NEW NUMBER! NO ONE CAN SAY HOW IT HAPPENS. YOU EITHER BELIEVE IT OR YOU DON'T.
THIS WHOLE BOOK IS FULL OF THINGS THAT HAVE TO BE ACCEPTED ON FAITH! IT'S A RELIGION!
AND IN THE PUBLIC SCHOOLS NO LESS. CALL A LAWYER.
AS A MATH ATHEIST, I SHOULD BE EXCUSED FROM THIS.


TOUR OF ACCOUNTING
OVER HERE WE HAVE OUR RANDOM NUMBER GENERATOR.
NINE NINE NINE NINE NINE NINE
ARE YOU SURE THAT'S RANDOM?
THAT'S THE PROBLEM WITH RANDOMNESS: YOU CAN NEVER BE SURE.
www.dilbert.com scottadams@aol.com
10-27-01 © 2001 United Feature Syndicate, Inc.


I'M FILLING OUT A READER SURVEY FOR CHEWING MAGAZINE.
SEE, THEY ASKED HOW MUCH MONEY I SPEND ON GUM EACH WEEK, SO I WROTE, "$500." FOR MY AGE, I PUT "43." AND WHEN THEY ASKED WHAT MY FAVORITE FLAVOR IS, I WROTE "GARLIC/CURRY."
THIS MAGAZINE SHOULD HAVE SOME AMUSING ADS SOON.
I LOVE MESSING WITH DATA.