A Chemometric Approach to Process Monitoring and Control

With Applications to Wastewater Treatment Operation

Christian Rosen

LUND UNIVERSITY

Doctoral Dissertation in Industrial Automation
Department of Industrial Electrical Engineering
and Automation

Department of
Industrial Electrical Engineering and Automation
Lund University
Box 118
SE-221 00 LUND
SWEDEN

http://www.iea.lth.se

ISBN 91-88934-20-9
CODEN:LUTEDX/(TEIE-1022)/1-290/(2001)

Printed in Sweden by Universitetstryckeriet
Lund University
Lund, 2001

To Magdalena

Abstract

In this work, various aspects of multivariate monitoring and control of wastewater treatment operation are discussed. A number of important difficulties face operators and process engineers when handling online measurements from wastewater treatment processes. These include, for instance, a high number of correlated measurement variables, non-stationarities, nonlinearities and multiscale process behaviour. A systematic way to handle and analyse data is needed to effectively extract relevant information for monitoring and control. In this work, a chemometric approach is taken. Principal component analysis (PCA) is used to reduce both the dimensionality of the problem and the noise level in data. However, it is shown that basic PCA is not sufficient to describe the process adequately. There are mainly two reasons for this. First, the processes display a non-stationary behaviour due to the diurnal, weekly and seasonal variations in the composition of the wastewater. Second, disturbances and events occur at different time scales making basic PCA less suitable.

The problem of non-stationary data is overcome using adaptive PCA in terms of updating of the scale parameters as well as the covariance structure. It is shown that adaptive PCA significantly improves the monitoring results as the model adapts to new process conditions without losing its ability to detect deviating process behaviour. To solve the problem of disturbances that occur at different time scales multiscale PCA is used. Multiscale PCA is a combination of multiresolution analysis and PCA. Measurement signals are decomposed into several time scales, and PCA models at each scale are identified. By doing so, the sensitivity to small process deviations that otherwise are obstructed by the diurnal variation is considerably increased. By omitting the lowest time scale from the analysis, the remaining time scales will inherently be (practically) sta-

tionary, since this corresponds to using a highpass filtered version of the data. Another solution, where the PCA models at each scale are made adaptive is also presented.

Using the monitoring results to adjust the process in a supervisory control manner is discussed. Two different methods are presented. The first is based on a multistep procedure. The current operational state is detected and classified using clustering in the principal component space. This information is used to determine appropriate setpoints for local controllers so that the process returns to what is considered normal operation. In the setpoint determination step, both static and dynamic models are used. The dynamic models are used within the framework of model predictive control (MPC). The multistep approach is best suited for extreme event control, since nonlinear and discrete control actions easily can be incorporated. The second method to integrate monitoring and control is based on PCA. Here, the inverse PCA model is used to directly calculate appropriate setpoints for the local controllers so that the process can be controlled to attain specified output requirements. The controller can be seen as a multivariate feedback controller implemented on top of the local control system. It is shown by simulation studies that both methods for supervisory control can successfully be used to control the process according to the control objectives.

Preface

This thesis constitutes the second part of the work I have carried out in my pursuit of a PhD in Industrial Automation at the Department of Industrial Electrical Engineering and Automation, Lund University, Lund, Sweden. The first part is presented in my Licentiate's thesis entitled 'Monitoring Wastewater Treatment Systems' in 1998.

The topic of this thesis is multivariate monitoring and control of wastewater treatment processes. I have come to realise that this topic is somewhere between a number of different research fields, such as process control, chemometrics, statistics and wastewater engineering. To be 'somewhere in between' has been a fascinating experience and I have learnt a lot from the different fields. However, it has also been difficult, since each has its own favourite methods and ways of expressing things. Sometimes, even a kind of hostility between the areas has been noticeable. This is sad, since they certainly complement each other. I have tried to pick suitable parts from each area to achieve the objectives of the work. However, as the title suggests, most ideas are taken from chemometrics, which also can be considered an interdisciplinary field. Thus, the notation is mostly according to that of chemometrics.

The thesis is a compilation thesis, consisting of an introduction to wastewater treatment and multivariate monitoring and control, followed by seven included papers. The papers are not organised chronologically. Instead, they are organised so that a common thread should be possible to follow. Since some of the

papers are old and since the paper format only provides limited space, each paper is followed by an addendum. In the addenda, comments on various aspects of the papers are given.

Acknowledgements

First, I would like to thank my informal supervisor and friend, Dr. Ulf Jeppsson, Department of Industrial Electrical Engineering and Automation (IEA), Lund University, for all the support, encouragement and help that he has given me. This work owes a lot to the many hours of discussions and collaborations with Ulf.

I would also like to thank my supervisor, Prof. Gustaf Olsson, IEA, who has given me the opportunity to be where I am today. He has been an essential source of inspiration during the years at IEA and his ability to put things into perspective has certainly helped me to navigate between the different research diciplines. I would also like to thank him for trusting me and letting me go my own way.

Many people at the department have had some part in the work I have done. I would especially like to express my gratitude to: Per Karlsson for all the proof reading and the interesting discussions throughout the years; Dr. Olof Samuelsson and Morten Hemmingsson for enduring my never ending stream of questions about everything from eigenvalues to LaTeX code; Dr. Mats Larsson for helping me with the MPC implementation. I would also like to thank everyone at IEA for providing a friendly and interesting atmosphere at the department.

I had the chance to go to the Advanced Wastewater Management Centre (AWMC), the Univeristy of Queensland, Brisbane, Australia for one year. I will probably remember this year as one of the best, both from an academic as well as a personal point of view, and I would like to thank all the people at the AWMC for giving me this opportunity. I would especially like to thank James Lennox and Dr. Zhiguo Yuan, who I now consider close friends. They have both contributed substantially to my work and I have learnt a great deal from them. To collaborate with them has certainly broadened my perspectives.

I would also like to thank the staff at the Ronneby wastewater treatment plant. They provided the process data, which made the first part of this work possible.

There is a life besides work. I wish to thank all my good friends for their support and their (at least pretended) interest in my work.

My family has always supported me, and for this I would like to thank my mother, father and sister. My wife, Magdalena, has sacrificed a lot during the years, including giving up her job to go abroad for a year. I am truly greatful for all her love, support and patience. This thesis is dedicated to her.

Lund, October 18, 2001
Christian Rosen

This work was partially supported by the Swedish Water and Wastewater Association (VAV) as part of the VA-forsk research programme, the Swedish National Board for Industrial and Technical Development (NUTEK) and Stiftelsen Sigfrid och Walborg Nordkvist.

Contents

I Introduction **1**

1 Introduction **3**
1.1 Motivation 4
1.2 Objectives 5
1.3 Contributions 7
1.4 Outline of the thesis 7
1.5 Publications 8

2 Wastewater treatment processes **13**
2.1 Process description 13
2.2 Automation in wastewater treatment operation 16
2.3 Modelling of wastewater treatment processes 19

3 Multivariate monitoring **29**
3.1 Challenges 29
3.2 Statistical process control 31
3.3 Multivariate statistical process control 32
3.4 Discussion on the applicability to WWTP operation 44
3.5 Multivariate monitoring in wastewater treatment 47

4 Multivariate feedback adjustment for control **49**
4.1 Challenges 49
4.2 Extreme event control and disturbance rejection 51
4.3 Product design 55

5 Summary of work **59**
5.1 Introduction 59

5.2 Univariate monitoring 60
5.3 Basic multivariate monitoring 61
5.4 Advanced multivariate monitoring 63
5.5 Control adjustments 66

6 Concluding remarks **69**
6.1 Summary of results 69
6.2 Comments on implementation 73
6.3 Topics for future research 74

7 Populärvetenskaplig sammanfattning **79**

II Included Papers **83**

A Disturbance detection in wastewater treatment systems **85**

B Monitoring wastewater treatment operation. Part I: Multivariate monitoring **103**

C Monitoring wastewater treatment operation. Part II: Multiscale monitoring **131**

D Adaptive multiscale principal component analysis for online monitoring of wastewater treatment **165**

E Supervisory control of wastewater treatment plants by combining principal component analysis and fyzzy c-means clustering **183**

F A framework for extreme-event control in wastewater treatment **203**

G A chemometric approach to supervisory control of wastewater treatment operation **227**

III Bibliography **259**

Part I

Introduction

Chapter 1

Introduction

Over the last centuries the human effects on the hydrologic cycle have increased. In order to establish convenient environments for living as well as agricultural and industrial production, artificial recycles have been created. So, in addition to the necessity to life, water is used for numerous purposes, for example irrigation, transport of material and energy as well as cleaning. Whatever the purpose is, processing and use normally result in pollution of water. Enormous amounts of water need to be treated each day, and although only a fraction is actually treated, the wastewater treatment industry constitutes the world's largest industry in terms of treated mass of raw material.

The methods to treat wastewater have, during the last century, gradually been refined, from simple grids and aerated ponds to highly complex processes including many separate steps. The requirements on the water being discharged to the recipient bodies have become more demanding as the environmental awareness has increased on both individual and governmental levels. The stricter requirements as well as the increased complexity of the involved processes emphasise the need for more knowledge on and better control of the processes.

The level of automation has risen somewhat from a very basic level in the beginning of the 1970s to approach a level in parity with the complexity of the involved processes at the most modern installations. However, there is much left to do, and a lot can be learnt from other industrial fields, such as chemical process, pharmaceutical and paper and pulp industries. In this work, a systematic approach to information extraction from wastewater treatment data is presented, using ideas taken from work carried out in especially the chemical process industry. Further, a methodology to use the extracted information for improvement of the overall control of the wastewater treatment plant is discussed.

1.1 Motivation

In most process industries, monitoring of the process and the process output is performed to achieve conformity with quality, safety and economic requirements imposed on the production. The level of monitoring differs from various fields and pioneering efforts are found in, for instance, the petrochemical and pharmaceutical industries. Wastewater treatment industries cannot be considered to be among the most diligent and systematic users of monitoring. Up to today, monitoring in wastewater treatment has mostly focused on a few key effluent entities upon which regulations are enforced by governments or other authorities. However, as more entities are regulated and the regulations become more rigid, the demands on the operation of the processes increase. Minimising the use of resources, for instance, energy, chemicals and manpower, and decreasing the amount of sludge products produced, have also become important issues in order to adapt the wastewater treatment processes to the ideas of sustainability. The development towards more resource efficient and sustainable systems has led to an increased need of process and operation knowledge. Thus, new and upgraded wastewater treatment plants are equipped with measurement systems for collecting data on a large amount of entities. In large wastewater treatment plants, the data collecting system may include hundreds or even thousands of measured entities. The measurements are used for monitoring the process and the quality of the process output. Measurements are also used for control directly in control loops or indirectly as a basis for manual control actions.

Due to the varying operational conditions, process and quality variables need to be monitored continuously to ensure a reliable and efficient operation and, thus, daily average values are not sufficient to get early detections or warnings of deviating or abnormal conditions. Consequently, this calls for techniques to handle large data sets online. The methods for monitoring used today are normally based on time series charts, where the operator can view the different variables as historical trends. It is difficult to track more than a few variables and when the number of monitored variables increases, it is difficult to draw any conclusions. Moreover, collective effects cannot be assessed by individual investigation of variables. Therefore, the methods must handle the difficulties involved in extracting information from multivariable data from the processes. These difficulties include large data sets, collinear data, data with nonlinear relationships, non-stationary data, data with dynamic relationships, noisy or

unreliable data and missing data. Further, to be useful in the operation, the information must be presented in an understandable and easily interpretable way.

Ideally, the information gained from process data is used to operate the process in the most efficient way possible. However, it is not always obvious how the information can be utilised to counteract a process deviation or disturbance. With a control system involving many local controllers, the cause-effect relationships may be complex and difficult to assess in time for possible corrections to have an effect on the disturbance. Some disturbances arise quickly, so an operator support in the decision making is desirable.

Wastewater treatment plants are not always manned. This means that if a severe disturbance occur, the time for reaction may be long. In these situations, a method to automatically derive and implement changes to the control system could provide a remedy. Automatic derivation and implementation of control system changes, typically setpoint changes or invoking new control handles, can be obtained by integrating the control and the monitoring system. This has been done in many industrial fields, but in wastewater treatment, these supervisory or plant-wide control systems are still uncommon. However, a recent study shows that the wastewater community now are beginning to show an increased interest in these issues (Jeppsson et al.; 2001). Also, researchers have increased their efforts to develop control architectures that better suit the difficulties encountered in wastewater treatment operation (Sànchez et al.; 1996; Katebi et al.; 1998; Roda et al.; 2001).

A systematic way to multivariate monitoring and supervisory (or plant-wide) control of wastewater treatment plants may provide more efficient and safer operation with a higher effluent water quality as a result. Moreover, it may also allow for new process designs and techniques with higher demands on the level of surveillance and control.

1.2 Objectives

The primary objective of this work is to investigate the applicability of multivariate statistics for online monitoring and control of wastewater treatment operation of a continuous biological nutrient removal plant. A number of secondary objectives can be stated, objectives to achieve the primary objective: 1) Identify challenges and difficulties encountered in online handling of multivariate process data generated by the wastewater treatment processes; 2) Adapt and

combine already known technologies, mainly from the field of chemometrics and statistical process control, to suit wastewater operational data; 3) Develop extensions to existing techniques to improve online monitoring performance in those cases when existing techniques do not suffice; 4) Investigate and assess the complexity level of the monitoring techniques required to obtain reliable and useful process information; 5) Investigate how the monitoring information can be used for wastewater treatment control; 6) Integrate monitoring and control to form a framework for monitoring and supervisory control.

The above listed objectives are complemented by some objectives of earlier presented work (Rosen; 1998a): a) Provide techniques for measurement data validation and quality improvement; b) Highlight and demonstrate methods to extract information from single variables and show the applicability of these methods for detection of disturbances.

The objectives should be seen from an engineering perspective rather than a scientific perspective. Thus, some statistical issues have been put aside to make room for solutions that are not the most elegant from a statistical science point of view, but feasible in practice. It should also be stated that the objectives 5 and 6 are natural extensions of the first four objectives. As will be seen, objectives 5 and 6 are somewhat driven by the curiosity of investigating how far one can extend the use of multivariate statistics for purposes originally not intended. Thus, no exhaustive investigation of other plant-wide or supervisory control approaches has been made.

It is appropriate with a comment on the use of terms. In the context of this work, the term 'multivariate control' is used for the task of coordination and/or optimisation of a set of local controllers in a multivariate system. The term 'multivariate system' is here used for a system with multiple inputs or outputs. In control theory, this is often referred to as a 'multivariable system'. Therefore, multivariate control is used here to avoid confusion with methods normally associated with 'multivariable control' (Glad and Ljung; 2000). Moreover, the term 'non-stationary' is here used in a practical sense. This means that a signal is said to be 'stationary' if the mean and variance over a period of interest are constant (within the expected statistical fluctuations). This means that use of 'weakly stationary' would be closer to a correct terminology (Åström and Wittenmark; 1997; Schreiber; 1997; Kennel; 1997). Further, non-stationary is also used for the multivariate case when the covariniance structrure changes over a time period of interest (Kano et al.; 2000a,b; Yoo et al.; 2001; Lennox; 2001).

1.3 Contributions

The main results of this work is summarised in Chapter 6. The major contributions of this work can be summarise as:

> The applicability of multivariate statistics for online monitoring of wastewater treatment operation is shown by examples on real process data. Techniques for isolation of deviating variables are proven useful and different methods for visualisation of the process state are shown successful;
>
> The dominating difficulties in monitoring wastewater treatment operation are identified and analysed;
>
> It is shown that many of the difficulties are circumvented by implementing extensions of the basic algorithms;
>
> New multiscale approaches are developed to solve the problems associated with processes displaying a wide range of time constants;
>
> A framework for integration of monitoring and control is developed;
>
> A new use of chemometric methods for supervisory control is presented.

A summary of available techniques for multivariate statistical monitoring is given and the bibliography includes a major part of the recent work done within the field. Finally, the work may provide a basis for engineers interested in applying the discussed methods and techniques to wastewater treatment monitoring and control.

1.4 Outline of the thesis

This thesis is organised in three parts: Part 1 includes a general introduction, some background information and conclusions, Part 2 consists of the included papers and in Part 3 a bibliography is given. Part 1 consists of a number of chapters: In Chapter 1, motivation, objectives and contributions of the work are given. Moreover, in Chapter 1, the author's publications are listed. Chapter 2 includes some basic information on wastewater treatment operation. In Chapter 3, an overview of multivariate statistical process control is given and in Chapter 4 some ideas on how multivariate statistics can be used for control of wastewater treatment operation is outlined. Chapter 5 is a summary of the

included papers and Chapter 6 contains some conclusive remarks on the results, implementation aspects and topic for future research. In Chapter 7, a summary of the work is given in Swedish.

1.5 Publications

Included papers

Paper A: *Detection of disturbances in wastewater treatment systems*,
Christian Rosen and Gustaf Olsson

Paper B: *Monitoring of wastewater treatment operation. Part I: Multivariate monitoring*,
Christian Rosen and James A. Lennox

Paper C: *Monitoring of wastewater treatment operation. Part II: Multiscale monitoring*,
Christian Rosen and James A. Lennox

Paper D: *Adaptive multiscale principal component analysis for online monitoring of wastewater treatment*,
James A. Lennox and Christian Rosen

Paper E: *Supervisory control of wastewater treatment plants by combining principal component analysis and fuzzy c-means clustering*,
Christian Rosen and Zhiguo Yuan

Paper F: *A framework for extreme-event control in wastewater treatment*,
Christian Rosen, Mats Larsson, Ulf Jeppsson and Zhiguo Yuan

Paper G: *A chemometric approach to supervisory control of wastewater treatment operation*,
Christian Rosen and Ulf Jeppsson

Author's contribution to included papers

The included papers are the result of collaboration with other researchers. Therefore, a few comments on the contribution of the author to each paper are appropriate.

Paper A: Ideas, implementation, analysis and writing are attributed to the author with support from Olsson.

Paper B: Collaboration with Lennox resulted in the ideas of the paper. Implementation, analysis and writing are attributed to the author with support from Lennox.

Paper C: See Paper B.

Paper D: The author's main contribution to Paper D is in the idea stage of the work. Some of the underlying ideas originate from the collaboration with Lennox in Papers B and C.

Paper E: Collaboration with Yuan resulted in the ideas of the paper. Implementation, analysis and writing are attributed to the author with support from Yuan with the exception of the SBH controller, which is attributed to Yuan.

Paper F: Paper F is a continuation of the ideas in Paper E. Implementation, analysis and writing are attributed to the author with two exceptions: coding and testing of the MPC algorithm were done in collaboration with Larsson; derivation of the reduced order model was done in collaboration with Jeppsson.

Paper G: Ideas, implementation, analysis and writing are attributed to the author with support from Jeppsson.

International journal publications

Rosen, C. and Olsson, G. (1998). Detection of disturbances in wastewater treatment systems, *Wat. Sci. Tech.* **37**(12): 197-205.

Rosen, C. and Yuan, Z. (2000). Supervisory control of wastewater treatment plants by combining principal component analysis and fuzzy c-means clustering, *Wat. Sci. Tech.* **43**(7): 147-156.

Rosen, C. and Lennox, J. A. (2001). Multivariate and multiscale monitoring of wastewater treatment operation, *Wat. Res.* **35**(14): 3402-3410.

Lennox, J. A. and Rosen, C. (2001). Adaptive multiscale principal component analysis for online monitoring of wastewater treatment, *Wat. Sci. Tech.* (accepted).

Rosen, C., Larsson, M., Jeppsson, U. and Yuan, Z. (2001). A framework for extreme-event control in wastewater treatment, *Wat. Sci. Tech.* (accepted).

Rosen, C. and Jeppsson, U. (2001) A chemometric approach to supervisory control of wastewater treatment operation, *J. Chemometr.* (submitted).

International conference publications

Rosen, C. and Olsson, G. (1997). Detection of disturbances in wastewater treatment systems, *7th IAWQ Workshop on Instrumentation, Control and Automation of Water and Wastewater Treatment and Transportation Systems*, 6-9 July, 1997, Brighton, UK.

Lennox, J. A. and Rosen, C. (2000). Using wavelets to extract information from wastewater treatment process data, *1st World Water Congress of the International Water Association (IWA)*, 3-7 July, 2000, Paris, France.

Rosen, C. and Yuan, Z. (2000). Supervisory control of wastewater treatment plants by combining principal component analysis and fuzzy c-means clustering, *5th IWA International Symposium on Systems Analysis and Computing in Water Quality Management (WATERMATEX)*, 18-20 Sept., 2000, Gent, Belgium.

Lennox, J. A. and Rosen, C. (2001) Adaptive multiscale principal component analysis for online monitoring of wastewater treatment, *1st IWA Conference on Instrumentation, Control and Automation (ICA2001)*, 3-7 June, 2001, Malmö, Sweden.

Rosen, C., Larsson, M., Jeppsson, U. and Yuan, Z. (2001). A framework for extreme-event control in wastewater treatment, *1st IWA Conference on Instrumentation, Control and Automation (ICA2001)*, 3-7 June, 2001, Malmö, Sweden.

Yuan, Z., Bogaert, H., Rosen, C. and Verstraete, W. (2001). Sludge blanket height control in secondary clarifiers, *1st IWA Conference on Instrumentation, Control and Automation (ICA2001)*, 3-7 June, 2001, Malmö, Sweden.

Rosen, C. and Jeppsson, U. (2001). Supervisory control of wastewater treatment operation by PC-space control, *7th Scandinavian Symposium on Chemometrics*, 19-23 Aug., 2001, Copenhagen, Denmark.

Other publications

The author of this thesis has earlier presented several reports and articles, which are all in line with the work presented here. Rosen and Olsson (1997b) treats the transformation from data to information and discusses some practical aspects of data collection and information extraction. Rosen and Olsson (1997a) is a report on analysis of online data from Pt Loma wastewater treatment plant in San Diego, USA. The incentive of the Pt Loma study was optimisation and increased knowledge of the chemical precipitation at the plant in order to meet more stringent requirements from the government of California. Time delay related issues and fault propagation in multilevel flow models (graph-based diagnosis) are discussed in Rosen (1998b). A comprehensive report on both univariate and multivariate detection of disturbances in wastewater treatment operation is given in Rosen (1998a).

Chapter 2

Wastewater treatment processes

The first ideas of recovery of water quality were based on physical means, such as dilution and sedimentation. However, this became precarious as cities grew larger and the importance of hygienic issues increased. Chemical precipitation was introduced to increase the settleability of the particulate matter in the wastewater to increase the settling efficiency. Biological treatment of wastewater dates back to the late 19th century (Orhon and Artan; 1994). It started with the trickling filter or biological bed, which was developed in the early 20th century (Hammer; 1986). Another breakthrough in biological treatment of sewage was the discovery that supplemental aeration of wastewater resulted in higher level of purification. In the beginning of the 20th century, experiments were carried out on what was to be called the activated sludge process. During the last few decades, wastewater treatment has become an industry of high complexity. Increasing requirements on efficiency in terms of effluent water quality and economics are important reasons. More knowledge on the physical, chemical and biological processes involved has been obtained, which has resulted in more advanced and efficient configurations. The ability to measure, analyse and control certain substance concentrations, flow rates and other entities is beginning to influence the design and operation of treatment plants considerably.

2.1 Process description

Wastewater treatment processes aim at removal of pollutants in the wastewater by transformation and separation processes. This is achieved in various ways, depending on the characteristics of the wastewater, the desired effluent quality

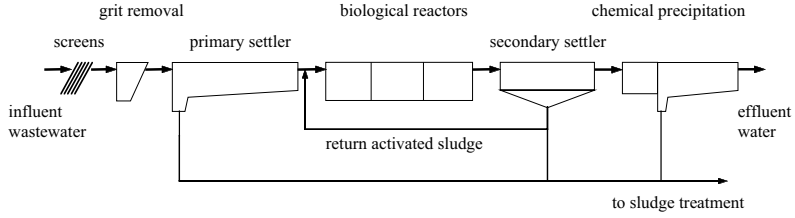

Figure 2.1: Principle layout for a continuous wastewater treatment plant (water phase).

and other environmental and social factors. Traditionally, the wastewater treatment processes are divided into physical, chemical and biological treatment, which are used in many different combinations. Figure 2.1 shows the principal layout of a typical treatment plant with physical, biological and chemical treatment.

Physical treatment

Physical treatment involves, for instance, screens, sedimentation, flotation, filters and membrane techniques. Sedimentation implies that particles heavier than water are settled in tanks and separated from the water phase. In flotation, or dissolved air flotation (DAF), particles are separated from the water phase by using dissolved air in pressurised water. When the pressure decreases, the dissolved air is released as small air bubbles, which attach to and lift the particles to the surface of the tank, where they are removed.

Chemical treatment

Chemical treatment involves coagulation and flocculation of colloidal and suspended matter as well as precipitation of some dissolved matter, such as phosphorous. Typical chemicals used are ferro, ferri and aluminium salts as well as lime. To further increase the efficiency of the process, coagulation aids such as polymers are often used. The chemical treatment also includes separation of the flocculated matter as a chemical sludge by means of sedimentation, flotation, etc.

Biological treatment

Biological processes are based on biological cultures, consisting of bacteria, unicellular life forms and even some multicellular life forms. The organic pollutants in the wastewater serve as food and energy sources for the microbiological culture as it grows. The microbiological culture can either grow suspended in the water phase or in a fixed position on surfaces as a biofilm. Suspended growth is used in so called activated sludge (AS) reactors, while the fixed growth is used in fixed bed reactors. A combination thereof is, for instance, suspended carriers, where the biofilm grows on small carriers, which are suspended in the water phase. Biological treatment aims at having a certain amount of microbiological culture in the process. In an AS reactor this is achieved by separating the sludge from the water phase in a sedimentation unit and returning it into the biological reactor. The excess sludge created in the process is removed and treated in sludge treatment processes, which stabilise and dewater the sludge. Stabilisation of sludge makes it biologically safe and often usable as fertiliser. The reduction of organic matter in a biological treatment plant is typically 90% or more. There are also processes for biological removal of phosphorous, but phosphorous removal is not further discussed in this work.

Predenitrification AS process

Many modern treatment plants utilising AS have biological nitrogen removal. Biological nitrogen removal relies on nitrifying and denitrifying bacteria for removal of nitrogen in two steps: nitrification and denitrification. Two different types of bacteria cultures are used to achieve nitrification and denitrification: autotrophic bacteria use inorganic carbon as carbon source whereas heterotrophic bacteria use organic carbon as carbon source. In the nitrification step, ammonium is oxidised to nitrite and then nitrate (nitrification) by autotrophs. In the second step, nitrate is reduced to nitrogen gas (denitrification) by heterotrophs. A difficulty with this procedure is that the two steps require different ambient conditions to function effectively. The nitrification step needs dissolved oxygen, whereas the denitrification step requires an oxygen free environment. A solution to this is to divide the reaction volume into separate compartments in which the conditions are different. A relatively common configuration for nitrogen removal is the predenitrification process. The first reactor is anoxic, that is no dissolved oxygen is present, and is followed by an aerated volume. This may appear somewhat backwards as the nitrification

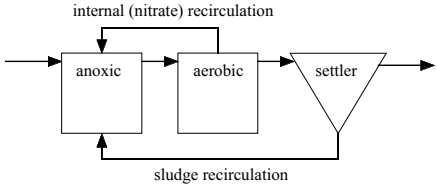

Figure 2.2: Basic principle of the predenitrification configuration.

is done after the denitrification. However, the denitrification process requires readily biodegradable organic substrate and this is normally present in the influent wastewater. If the denitrification was to take place after the nitrification (i.e. a post-denitrification configuration), most of the organic substrate would have been consumed and external carbon would have to be added. Thus, to provide the anoxic reactor with nitrate, a recirculation stream is introduced from the last reactor to the first reactor (sometimes the sludge recirculation is sufficient and no internal recirculation is needed). This configuration puts a limit on how far the nitrogen removal can be driven, but it provides an, in many cases, economical solution and has become popular. The basic principle of the predenitrification configurations is shown in Figure 2.2.

2.2 Automation in wastewater treatment operation

Online measuring and data collection systems

The number of measurable entities increases as research on and development of instrumentation and sensors progress. A difficulty in online measuring is the aggressive environment in which the sensors must function. Another problem is that many of the interesting entities must be derived from reaction analysis in batch (or continuous) experiments. Development of automatic systems for this type analysis is progressing all the time and today a number of bio-chemical variables can be measured online. Interesting development in the sensor area involves new types of sensors such as sensor arrays or soft sensors, where variables are deduced from a number of measurements, and biosensors that utilise (immobilised) cultures of bacteria. Many of these techniques are on the fringe of being commercially available and will play an important role in wastewater char-

Table 2.1: On-line measurement in predenitrification process

Type	Measured variable	Comment
Physical		
	infl. flow rate	continuous
	air flow/pressure	continuous
	temperature	continuous
	suspended solids	continuous
	sludge blanket height	discrete/continuous
Physio-chemical		
	pH	continuous
	redox potential	continuous
	dissolved oxygen (DO)	continuous
	conductivity	continuous
Bio-chemical		
	respiration	delay 15-30 mins
	ammonia	delay 5-30 mins
	nitrate	delay 5-30 mins

acterisation and sensing within a few years. In Table 2.1, the most important measurements that are readily (but sometimes costly) available (Vanrolleghem; 1994; Jeppsson et al.; 2001).

The data collecting systems differ from plant to plant and from supplier to supplier but common sampling rates (in Sweden) are 10 and 12 per hour, i.e. every sixth and fifth minute, respectively. The sample values are often an average over the sampling period, during which some sensors continuously deliver values and others perhaps only once a minute. All sensors are afflicted with time lags, but normally these are short in comparison to the dominant time constants of the process.

Control handles

There are limitations to what can be controlled in a wastewater treatment plant. This is due to a lack of powerful control handles in comparison to the relatively severe disturbances that varying influent wastewater characteristics impose on the system. Major available control handles for a predenitrification process are

Table 2.2: Manipulated variables in predenitrification process.

Type	Manipulated variable	Controls	Controller
Flow rates			
	infl. flow rate	hydraulic load	OO/FF/FB
	internal recirculation	nitrate to first reactor	FF/FB
	internal recirculation	nitrate in last reactor	FF/FB
	sludge recirculation	sludge to first reactor	FF/FB
	sludge recirculation	sludge blanket level	FB
	waste flow	tot. amount of sludge	Man/OO/FB
Chem. addition			
	ext. carbon addition	access to substrate	FF/FB
	polymer addition	sludge settling prop.	FF
	lime addition	pH	FF/FB
	P-precipitants	effl. phosphate	FF
Aeration			
	air flow/pressure	DO conc.	FB
	air flow/pressure	redox potential	FB
	air flow/pressure	respiration	FB
	DO setpoint	ammonia conc.	FB
Other			
	step feed	flow distribution	Man/FF

FF=feed-forward (incl. ratio control); FB=feedback;
Man=manual; OO=on-off (incl. time control and alarm triggered)

listed in Table 2.2 (Vanrolleghem; 1994; Jeppsson et al.; 2001). The table reflects state of the art in practice and it is not likely that there are many plants that have access to all of the listed manipulated variables.

A majority of the manipulated variables are macro variables (DO and some of the chemical additions excepted) whereas some of the major mechanisms that drive the processes are on the micro level. Moreover, these mechanisms are often coupled. Thus, most control handles must be considered rather blunt and often a combination of control handles is required to reach a certain control objective.

Process dynamics

A wastewater treatment process consists of many subprocesses with dynamics of various time scales. Some variations are slow, for instance sludge dynamics and temperature, with time scales of days, week and even months. The daily variation in influent flow rate and substance concentrations is perhaps the most dominant variation. However, there are even faster dynamics present, such as dissolved oxygen (DO) dynamics and hydraulic shocks. The different time scales make it difficult to analyse the cause-effect relationships, especially when recirculation and other feedback loops are present. Therefore, it is important to establish the dynamic behaviour of the involved processes and adapt the analysis methods in accordance to the dynamics. An overview of cause-effect relationships and the corresponding (qualitative) time constants is given in Olsson and Jeppsson (1994).

2.3 Modelling of wastewater treatment processes

In Papers E to G, different supervisory control approaches are discussed. To investigate their performance, simulation studies of wastewater treatment operation have been used. Although the configurations of the plants differ, the same models for both the biological reactions as well as for settling have been used in all three papers. For the readers not accustomed with wastewater treatment modelling and simulations a short description of the models is given here.

The Activated Sludge Model No.1

The Activated Sludge Model No.1 (ASM1) is the result of a task group work, initiated by the International Water Association (IWA, formerly IAWQ and IAWPRC) in 1983, and published in 1987 (Henze et al.; 1987, 2000). It must be pointed out that the model owes a lot to earlier work carried out by a number of researchers. The perhaps most important work was carried out in South Africa during the late 1970s and early 1980s (Ekama and Marais; 1979; Dold et al.; 1980; Van Haandel et al.; 1981). Since the introduction of ASM1 the task group work has continued and in 1995 the Activated Sludge Model No.2 (ASM2) was presented. ASM2 includes new compounds and biological processes that describe biological phosphorus removal and is a more complex model that ASM1 (Henze et al.; 1995, 2000). It was soon followed by the ASM2d, a minor extension to the ASM2 model (Henze et al.; 1999, 2000). In 2000,

the ASM3 (Gujer et al.; 1999; Henze et al.; 2000) was presented, returning to the structure of the less complex ASM1, but with a number of extension and changes.

Although new models have been introduced, the ASM1 is still very much in use due to the extensive knowledge and experience that have been obtained in the research community. An example of this is the recently developed benchmark for control of biological wastewater treatment (will be shortly described later), which relies on the ASM1.

State variables

The state variables included in the ASM1 are listed in Table 2.3. The state variables differ somewhat from the ones measured and observed at a plant. Organic matter and dissolved oxygen have the unit mg COD/l. The nitrogen fractions have the unit mg N/l, and the unit for alkalinity is moles HCO_3^-/m^3.

Table 2.3: The state variables of the ASM1 model.

Symbol	Variable
S_I	Inert organic matter
S_S	Readily biodegradable substrate
X_I	Particulate inert organic matter
X_S	Slowly biodegradable substrate
$X_{B,H}$	Active heterotrophic biomass
$X_{B,A}$	Active autotrophic biomass
X_P	Particulate product from biomass decay
S_O	Dissolved oxygen
S_{NO}	Nitrate and nitrite nitrogen
S_{NH}	Ammonia nitrogen
S_{ND}	Biodegradable organic nitrogen
X_{ND}	Particulate biodegradable organic nitrogen
S_{ALK}	Alkalinity

At the plant the total suspended solids (TSS) is normally measured. Therefore, the particulate matter must be converted to TSS. Henze et al. (1995) proposed following conversion:

$$TSS = 0.75(X_I + X_P + X_S) + 0.9(X_{B,H} + X_{B,A}) \tag{2.1}$$

Reaction dynamics

There are eight different dynamic processes in the ASM1 model describing the dynamics.

> Aerobic growth of heterotrophs—readily biodegradable substrate, dissolved oxygen, ammonia and alkalinity are consumed and heterotrophic biomass is produced. The growth rate is modelled by a Monod expression.
>
> Anoxic growth of heterotrophs—readily biodegradable substrate, nitrate and ammonia are consumed and heterotrophic biomass and alkalinity are produced. The growth rate is modelled by a Monod expression.
>
> Aerobic growth of autotrophs—dissolved oxygen, ammonia and alkalinity are consumed and autotrophic biomass and nitrate are produced. The growth rate is modelled by a Monod expression.
>
> Decay of heterotrophs—heterotrophic biomass is decomposed into slowly biodegradable substrate and other particulate products.
>
> Decay of autotrophs—autotrophic biomass is decomposed into slowly biodegradable substrate and other particulate products.
>
> Ammonification of soluble organic nitrogen—biodegradable organic nitrogen is transformed to ammonia. Alkalinity is produced.
>
> Hydrolysis of entrapped organics—slowly biodegradable substrate is transformed to readily biodegradable substrate.
>
> Hydrolysis of entrapped organic nitrogen—particulate biodegradable organic nitrogen is transformed to biodegradable organic nitrogen.

The complete dynamic modell is described elsewhere (Henze et al.; 1987, 2000).

Parameters

The kinetic and stoichiometric coefficients of the ASM1 model must be given values. The task of determining these values is known as model calibration. If the model is used to simulate a specific plant, the calibration must be carried out for this plant. This implies extensive experiments at pilot and bench-scale

plants. However, a set of parameter values suggested by the IWA task group is presented in Henze et al. (1987). These values may be used when no further information is available, or when the task is to model a plant in general.

Settler Model

The settler model used in the simulation model is known as a one-dimensional layer model. One-dimensional models only describe the settling process along the vertical axis, leaving only cross-sectional area and depth as design parameters. The one-dimensional layer model is thoroughly described in, for instance, Ekama et al. (1997) and Jeppsson (1996) and only the basic ideas behind the model will be presented here.

Layers

The idea behind one-dimensional layer models is that the settler is divided into a number of layers. The mass balance for each layer is calculated, assuming complete mixing within each layer. The sludge transport between the layers is assumed to depend on two mechanisms; bulk movement and gravity settling. The bulk movement is caused by the hydraulic flow and are, hence, directed both upwards and downwards. A feed layer must be determined. Above the feed layer the bulk flow is directed upwards, corresponding to the effluent flow (Q_e); below the feed layer the bulk flow is directed downwards, corresponding to the underflow (Q_u). The gravity settling is always directed downwards due to the gravity action on the sludge. In Figure 2.3, the principle of the one-dimensional layer model is shown.

Settling Velocity Functions

Many different settling velocity functions are found in the literature. Traditionally, the settling velocity function is based either on an exponential or a power function, where the settling velocity only depends on the local concentration (Kynch; 1952). In the model used in this study, the double-exponential settling velocity function is used. In this function, consideration is taken to the fact

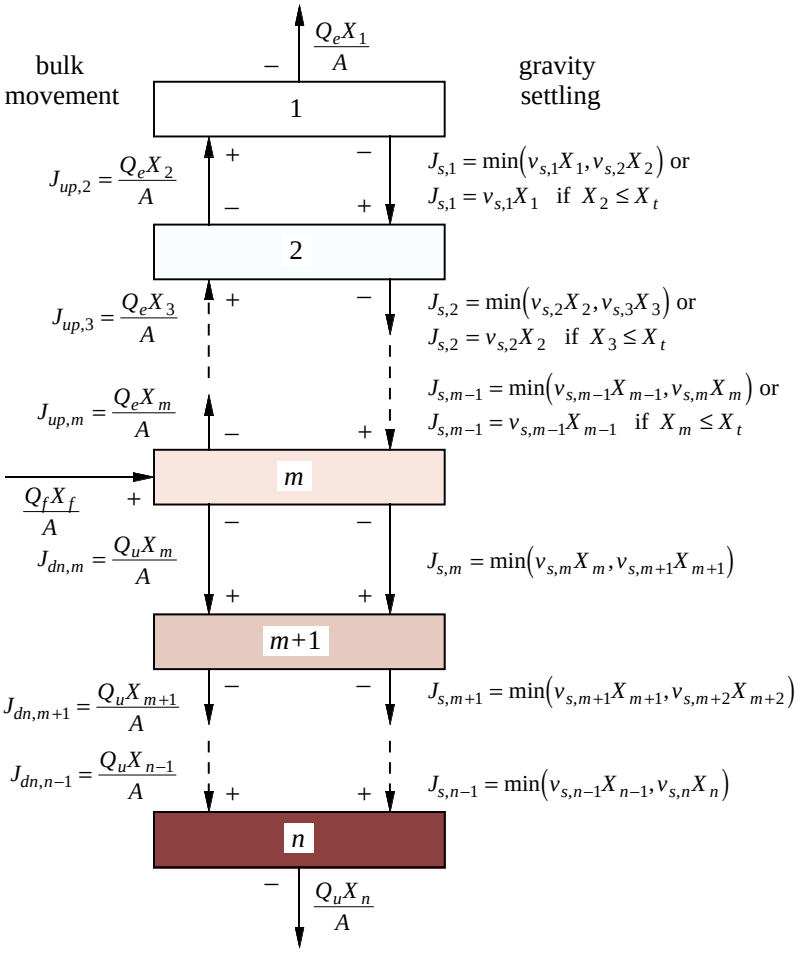

Figure 2.3: General description of the traditional one-dimensional layer settler model (Jeppsson; 1996).

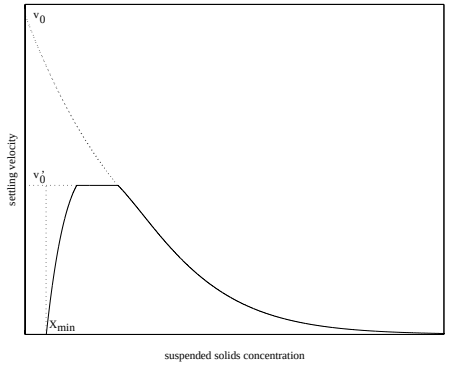

Figure 2.4: Schematic description of the double-exponential settling velocity model at a constant X_f (Jeppsson; 1996).

that low concentrations do not imply extremely high settling velocities. The function is defined as (Takács et al.; 1991):

$$v_s = max\left(0,\, min\left(v_0',\, v_0\left(e^{-r_h(X-X_{min})} - e^{-r_p(X-X_{min})}\right)\right)\right) \quad (2.2)$$

where v_0' and v_0 is the maximum practical and theoretical settling velocity, respectively. r_h is a settling parameter characterising the hindered settling zone and r_p is a parameter associated with the settling behaviour at low solids concentrations. X_{min} is calculated as a fraction of X:

$$X_{min} = f_{ns} X_f \quad (2.3)$$

where X_f is the concentration into the feed layer. Figure 2.4 shows the double-exponential settling function.

The benchmark simulation model

The idea to develop a simulation benchmark for wastewater treatment control was first evoked in the mid 1990s. The development of the benchmark was then carried out in parallel (and in cooperation) by the European Cooperation in the field of Scientific and Technical Research (COST) Actions 682/624 and the first IAWQ Task Group on Respirometry-based Control of the Activated Sludge Process (later the second IWA Respirometry Task Group) (Spanjers et

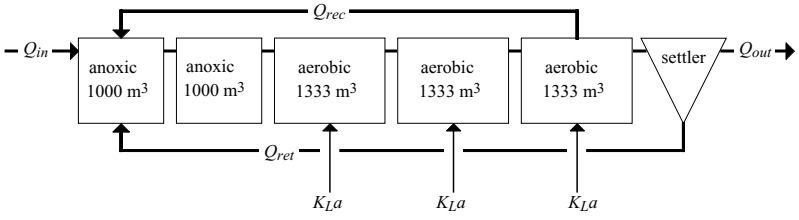

Figure 2.5: The principal layout of the COST benchmark simulation plant.

al.; 1998b,a; Pons et al.; 1999; Copp; 2000). The motivation for the work is that there exists a need for a way to evaluate different control strategies in an objective way, i.e. without plant specific requirements or limitations. A standardised simulation protocol makes this possible. However, it was decided that the protocol should be platform independent so that no specific requirements on software are required. A look at the last few conferences in the areas of control and simulation of wastewater treatment processes (e.g. WATERMATEX (2000) and ICA (2001)) reveals that the approach is successful; many researchers around the world use the 'COST benchmark' for simulation and evaluation of control strategies.

Plant configuration

The simulation plant comprises five reactors, of which the first two are anoxic and the following three are aerobic. The reactors are followed by a sedimentation unit and the plant is, consequently, a predenitrification plant. The physical dimensions of the plant are: anoxic reactors: 1000 m^3; aerobic reactors; 1333 m^3; sedimentation unit: 6000 m^3 (area 1500 m^2). Two internal recycle streams are included: internal (nitrate) recirculation from tanks 5 to 1, and sludge recirculation, from settler to tank 1. In the default configuration, values and properties of certain variables, parameters and max/min values are specified. The average influent flow rate used for design of the plant is 18446 m^3/day. The layout of the simulated plant is shown in Figure 2.5.

Process models

The reactors are modelled using the ASM1 model as completely mixed reactors and the settler is modelled using a ten-layer Takács model.

Influent characteristics

Three different influent files have been developed. The files contain data at 15-minute fifteen intervals, and display a significant diurnal as well as weekly pattern. The files are developed to mimic real wastewater characteristics typical for a plant of the chosen size. The influent data includes values for S_S, $X_{B,H}$, X_S, X_I, S_{NH}, S_I, S_{ND}, X_{ND} and Q_{in} with S_O, $X_{B,A}$, X_P and S_{NO} set to zero. The influent files are:

Dry weather—The dry weather data display a diurnal and weekly pattern corresponding to that of dry weather. Thus, no major upsets are present. The file contains data for a 14-day period (so do the other files).

Storm weather—The storm weather data include two major upsets. A short storm (high influent flow rate) at the end of day 8 and a longer storm at the start of day 11. The first disturbance includes a flush-out, i.e. particulate matter said to be present in the sewer system is flushed out due to a sudden increase in the flow rate. This implies that the particulate matter concentration during the event is high. During the second disturbance, however, the particulate matter is diluted, resulting in low influent concentrations. During both disturbances, the soluble compounds are diluted. The maximum flow rate during both disturbances is about three times the average influent flow rate.

Rain weather—The rain weather data contain a prolonged period of rain. The rain starts at day 8 and diminishes at day 10. During the rain event, both soluble and particulate matters are diluted.

The influent flow rates for the separate cases are displayed in Figure 2.6.

Performance measurements

To evaluate different control strategies, a performance index has been developed. This index includes a process assessment and a controller assessment. The process assessment is divided into an effluent quality index, effluent violations and

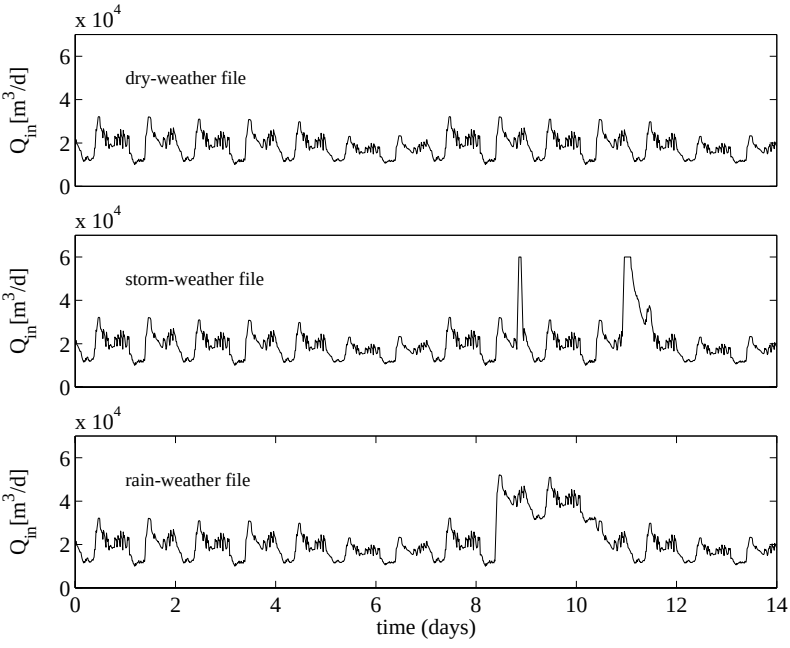

Figure 2.6: Influent flow rate for the influent files developed within the COST simulation benchmark work.

operational costs (sludge production, energy for pumping and aeration). The controller assessment includes controlled variable performance and manipulated variable performance. Consequently, the index is not just one measure, but a set of measures useful for investigation and assessment of the dynamic properties of new control strategies.

Other aspects

The benchmark platform is in constant development. The most updated information is available on the COST Action 624 benchmark web site (http://www.ensic.u-nancy.fr/COSTWWTP/).

Chapter 3

Multivariate monitoring

This chapter outlines some basic aspects on multivariate statistical process control (MSPC) of industrial processes in general and wastewater treatment process in specific. The basic methods are described together with extensions needed to adapt MSPC to the requirements for wastewater treatment operation monitoring.

3.1 Challenges

Common to most industrial processes is that the number of measured variables is high and that the number tends to increase even more as developments in sensor technology and process control progress. The incentive for this increase is stricter requirements on product quality, process efficiency, process safety, etc. In wastewater treatment, this has resulted in that we today have access to many online measurement signals from the process. In a well-equipped treatment plant the number may exceed hundred (or even thousand), ranging from binary equipment signals and alarms to flow rates, pressures and nutrient concentrations. It has been established that the human being is capable of handling just a few inputs simultaneously (seven is often mentioned) before a degrading analysis capacity is recorded. This makes combinatorial effect of many variables difficult to anticipate and comprehend. MacGregor (1997) state a few difficulties or challenges that need attention in handling and analysing industrial data.

> Data quality—Measurements are normally afflicted by noise arising from various sources in the sampling and measurement procedure. Moreover, sometimes data are not available at all due to sensor malfunction or com-

munication problems within the data collection system. Noise and missing data points make it difficult to extract and interpret information from data.

Data size—The high number of measured variables is a result of the increase in accessibility due to the introduction of computers in the data collection procedure. Expressions like 'data overload' and 'data rich, information poor' arise from the fact that although the information is in the vast amount of data, humans simply do not have the ability to analyse and interpret high dimensional problems. Not seldom is the result that only a few 'key' variables are monitored and, consequently, a lot of information is lost (Wise and Gallagher; 1996b; MacGregor; 1997).

Collinear data—The fact that a process contains many measured variables does not necessary imply that the process inherently is high dimensional. On the contrary, most industrial processes display a behaviour that can be captured in a few 'true' dimensions. This is because there are normally only a few main mechanisms that drive the process. The collinearity problem does not only provide difficulties for human interpretation, but also for conventional statistical analysis methods, since they rely on a high degree of independence among the variables (MacGregor; 1997).

For any type of measurement monitoring system, the above discussed challenges must be met. However, there are further difficulties to overcome before a monitoring system can successfully be applied to wastewater treatment operation (or processes with similar behaviour).

Non-stationary data—The conditions in which wastewater treatment processes are operated are normally of a varying nature. Diurnal, weekly and seasonal patterns are normally found in the influent wastewater characteristics. These disturbances must be considered as normal and is in practice seen as state of things rather than disturbances. It is often difficult to discern other process disturbances from those caused by the varying influent conditions, which tend to have a dominant effect on the process behaviour.

Multiscale data—A difficulty related to the dynamic properties of the disturbances as well as of the process is that disturbances occur in many

different time scales. By this, we mean that some disturbances affect the process in a short time frame, whereas others have a much slower response. Apart from that this fact complicates the discernment of disturbances in a similar way to that of non-stationarity, it also deteriorates the performance of many monitoring techniques (Bakshi; 1998). Moreover, information on the time scale (or 'speed') of a disturbance may prove crucial for a decision on counteractive actions. The multiscale nature of data is, however, not only a problem; it can also be used to decouple the process in time.

Nonlinearities—Wastewater processes display a nonlinear behaviour and relationships between variables cannot always be approximated by a linear function. Consequently, if this is the case, nonlinearities must be taken into account when developing a monitoring model.

Dynamic data—Almost all data from dynamic processes (such as wastewater treatment) are autocorrelated, which means that each observation is not independent of the previous observation. This may have a great impact on statistical properties of the monitoring output (Negiz and Çinar; 1997) and, consequently, caution must be taken when interpreting the result.

3.2 Statistical process control

Monitoring process operation using univariate time series is often referred to as statistical process control (SPC). The first ideas of SPC for quality improvement go back as far as to the beginning of the century when, for instance, Vilfredo Pareto and Walter Shewart made some important contributions to SPC (Thompson and Koronacki; 1993). The ideas were further developed during the 1950s, but it is not until the 1970s that SPC has become a standard tool for quality improvement in the process industry. SPC involves many methods for monitoring and presenting measurement variables, but perhaps the most common ones are:

'x'-charts—measurement values plotted against the time;

MA charts—moving average of the measurement series plotted against time;

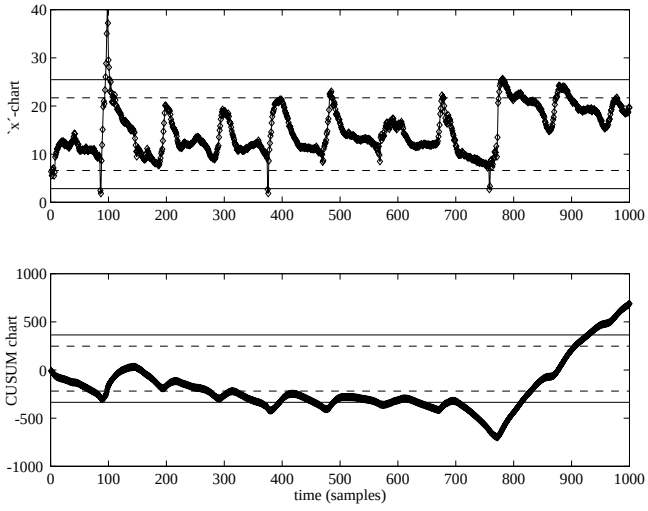

Figure 3.1: Examples of univariate monitoring charts. 'x'-chart (top) and CUSUM chart (bottom).

EWMA charts—exponentially weighted moving average filtered measurements plotted against time;

CUSUM charts—cumulative sum of the difference between the measurement and a target value plotted against time.

These methods have great similarities to conventional signal processing techniques. There are many references to SPC in the literature, such as Bissel (1994), Thompson and Koronacki (1993) and Box and Luceno (1997). SPC in wastewater treatment applications is described in Chapman (1998). An example of univariate monitoring of the influent ammonia concentration to a wastewater treatment plant is shown in Figure 3.1. The 'x'-chart and CUSUM chart give complementary information on the current ammonia load.

3.3 Multivariate statistical process control

Multivariate statistical process control comprises a number of methods that often is referred to as projection methods (Davis et al.; 1996). The basic idea of projection methods is that a high dimensional space, spanned by a number of measured variables, is projected onto a model space of fewer dimensions. The

model space is spanned by linear[1] combinations of the original variables to form 'pseudo variables', often referred to as principal components or latent variables. The identification of a projection method, thus, involves finding the pseudo variables that best describe the major features of the data set constituted by the measured variables and where the pseudo variables span only the 'true' dimension of the process. Thus, correlated data are not a difficulty but a necessity for projection methods.

The basis of the multivariate statistical methods is principal component analysis (PCA). PCA was first described by Pearson (1901) as a method to find the closest fit of lines and planes to points in space. From this, a number of people have contributed to make PCA what it is today (Fisher and MacKenzie; 1923; Hotelling; 1933; Wold; 1966). Another cornerstone of multivariate statistics is the partial least squares (PLS), introduced by Wold in the 1970s (Geladi; 1988). PLS is a regression method closely related to PCA.

The history of multivariate statistics is closely linked to progress within areas such as econometrics, psychometrics and chemometrics (Geladi; 1988), the latter involving the area of process monitoring. With a starting point in the early 1970s, chemometrics developed along with computational power (Geladi and Esbensen; 1990; Esbensen and Geladi; 1990). During the 1980s, chemometrics received an increased attention and multivariate process monitoring was included as an application of the techniques. During the 1990s a number of developments and extensions of the basic methods were introduced and chemometrics is today an important tool in many process industries. The parallel development of the methods has resulted in an abundance of redundant terminology. For instance, PCA goes under different names such as singular value decomposition (SVD), Karhuen-Loéve expansion, eigenvector analysis, characteristic vector analysis and Hotelling transformation, depending on what research area it is used in (Wold et al.; 1987a). Factor analysis may also be mentioned, but it is slightly different from PCA.

Standard methods

Principal component analysis

In mathematical terms, PCA is obtained by singular value decomposition of the covariance or correlation matrix of the process data. By doing so, a sub-

[1] Also nonlinear projection methods exist and will be discussed later.

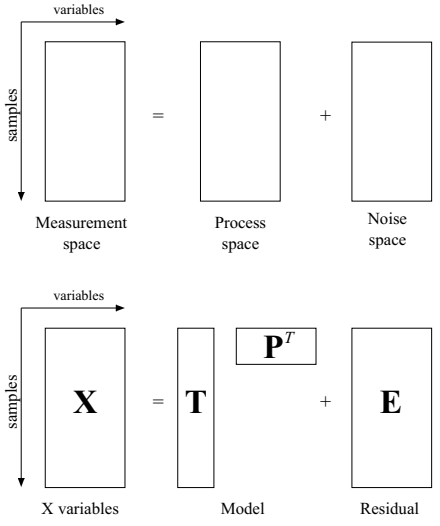

Figure 3.2: Decomposition of X into a process subspace and a noise subspace.

space (process subspace) containing the true (non-random) variation is identified. Complementary to this subspace is the noise subspace, which ideally contains only noise (Figure 3.2).

An alternative way of describing PCA is that a line (or component) is fitted in the direction of greatest variability of the measured variable space. Next, a line is fitted in the second greatest direction of variability orthogonal to the first line and, thus, a plane is obtained. The next line is fitted in the third greatest direction, orthogonal to the plane. This is continued until it is established that no systematic variability is left (Figure 3.3).

In matrix form, PCA is written as:

$$\mathbf{X} \quad = \quad \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{3.1}$$

where $\mathbf{X}$ is the original data set of size $[m \times n]$, $\mathbf{T}$ is called scores $[n \times a]$, $\mathbf{P}$ is called loadings $[n \times a]$ and $\mathbf{E}$ is the model residual (or noise subspace). If $a = n$ then $\mathbf{E} = 0$, as all the variability directions are described. However, if $a < n$, i.e. less principal components than original variables are retained, then $\mathbf{E}$ describes the variability not described by the sum of the $\mathbf{T}\mathbf{P}^T$ matrices. In general, $a << n$ is true for industrial applications.

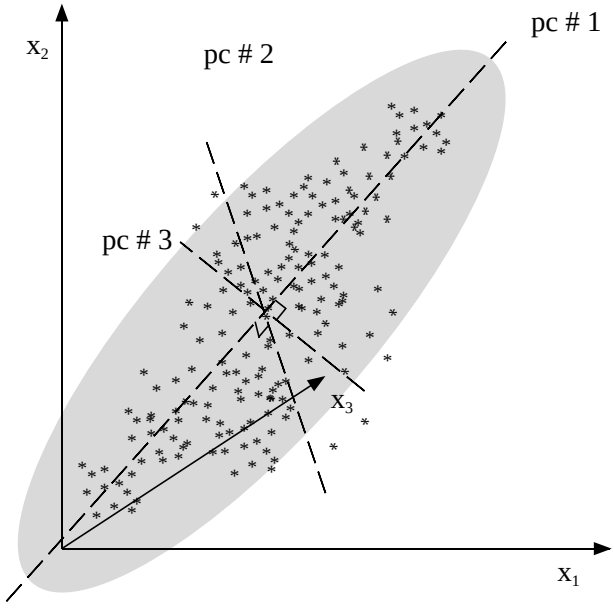

Figure 3.3: PCA as a successive fitting of components in the directions of greatest variability.

Prior to decomposition, variables are mean centred. Further, variables are normally scaled to give them equal influence on the model (variables are expressed in different units and display substantially different numerical ranges). There are situations when this pre-treatment is not appropriate, but all through this thesis data are both mean centred and scaled to unit variance (autoscaled) unless otherwise stated.

The basic principle for process monitoring using PCA is that a training set of data representing normal operational conditions is decomposed and a process subspace is identified. When new process data are obtained, they are projected onto the process subspace and noise space, respectively. By investigating the projected data (**T**) and the residual (**E**), process deviations and disturbances can be detected utilising various techniques (Jackson and Mudholkar; 1979; Kresta et al.; 1991; Yoon and MacGregor; 2001). By looking at the sum of the squared prediction error (SPE), the current model fit is investigated. If

the current operation display poor fit, the current operational state is obviously different from that of the training set. Hotelling's T^2 is a summarised way of surveying the scores and is used to assess the variations within the model[2]. Moreover, when a deviation is established, backtracking through the model is done to isolate what variables contribute to the deviation (MacGregor et al.; 1994; Teppola et al.; 1998c; Rosen and Olsson; 1998).

A decision that is crucial for the performance of the PCA model is when to stop including more principal components (PCs) and this is where modelling experience and process knowledge comes in. However, there are different tests that can be applied (Wold; 1978; Himes et al.; 1994; Qin and Dunia; 2000) to support the decision.

A more comprehensive description of PCA and its applications is found in, e.g. Jackson (1980), Jackson (1981), Joliffe (1986), Wold et al. (1987a) and Eriksson et al. (2001). Piovoso et al. (1992), Zullo (1996), Kourti et al. (1996), Wikström et al. (1998), and Teppola et al. (1998c) provide many interesting examples of application of PCA in the process industry in general (for wastewater treatment applications see Section 3.5).

PCR

Principal component regression is a simple extension of PCA. An output or quality variable is regressed on the scores instead of on the measured variables as is the case in ordinary least squares regression.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{3.2}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{B} \tag{3.3}$$

where $\mathbf{Y}$ is the output variable vector/matrix and $\mathbf{B}$ is the regression vector/matrix. Now, we have a model with predictive power, which are advantageous in many cases. The monitoring properties are obviously the same as for PCA, but if the model is developed for prediction, the number of PCs is determined as the number that gives the best prediction of the variables in $\mathbf{Y}$.

PLS

In PCR, the decomposition of $\mathbf{X}$ is done to maximise the captured variability in $\mathbf{X}$. This is generally not optimal for prediction purposes. In partial least squares

[2]A more elaborate discussion on SPE and T^2 is given in Paper B.

(PLS) regression, the decomposition of $\mathbf{X}$ and $\mathbf{Y}$ is carried out iteratively. By exchanging information between the two blocks in each step, the principal components (or latent variables, which is a more common term in PLS modelling) of the X-space are rotated so that the predictive power of the X-space with regard to the Y-space is enhanced. There are different algorithms to calculate PLS, but a common algorithm for this is the NIPALS algorithm (Geladi; 1988; Lindgren et al.; 1993; Kaspar and Ray; 1993; Dayal and MacGregor; 1997a; Phatak and de Jong; 1997). The equations of PLS are:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{3.4}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{3.5}$$

$$\mathbf{X} = \mathbf{T}\mathbf{B}\mathbf{Q}^T \tag{3.6}$$

where $\mathbf{U}$ and $\mathbf{Q}$ are the scores and loadings for the Y-block. $\mathbf{B}$ describes the inner relation between the latent variables of $\mathbf{X}$ and $\mathbf{Y}$ space.

As was the case for PCA (and PCR), the choice of the number of latent variables is crucial for the monitoring and prediction outcome and typically cross-validation is used to select the appropriate number (Wold; 1978; Wakeling and Morris; 1993; Messick et al.; 1997).

Process monitoring using PLS follows the same procedure as for PCA. However, since the model is developed with one or several specific output variables in mind, process deviations affecting these variables will be emphasised. Thus, if PCA is considered as an unsupervised monitoring model and only dependent on the choice of variables in X space, PLS can be seen as a supervised model with the possibility to tailor the model to detect deviations of certain interest.

Examples of process monitoring application utilising PLS are given in e.g. Kresta et al. (1991) and MacGregor and Kourti (1995). For the reader with a background in control theory, an interesting interpretation of PLS regression is given by Di Rusco (1998).

Extension of standard methods

Dynamic methods

An often encountered objection to MSPC for process monitoring is that MSPC does not consider the dynamics of a system. In its basic configuration, MSPC is a static modelling technique.

Time-lags are present in all dynamic processes. By time-lag we mean the time it takes for a change in the X-block to propagate to the Y-block. In basic multivariate monitoring the time lag between the X-block and the Y-block is not addressed. One way of dealing with this, assuming there is just one quality variable in the Y-block, is to investigate the cross-covariance function between every input variable and the output variable and calculate the suitable lag (Åström and Wittenmark; 1997). A second way is to use an a priori model for the time lag of every relation, for example, depending on the retention time. By doing so, the time lag between each process variable and the quality variable will change dynamically as the flow rate changes and we will obtain a quasi-dynamic representation of the flow rate dynamics (Röttorp and Jansson; 2001).

However, it is straightforward to extend MSPC so that dynamics are accounted for. By simply introducing lagged duplicates of each variables in the X or Y-block, dynamic relations can be modelled. Thus, the lagged X-block is written:

$$\mathbf{X} \quad = \quad [\mathbf{X}_k \; \mathbf{X}_{k-1} \; ... \; \mathbf{X}_{k-l}] \tag{3.7}$$

where $\mathbf{X}_{k-l}$ denotes the data matrix lagged l samples. Thus, ideas taken from time-series modelling, such as finite impulse response (FIR), auto-regressive exogenous input (ARX), auto-regressive moving average (ARMA), etc., can easily be used within the framework of multivariate statistics (Ricker; 1988; Wise and Ricker; 1993; Ku et al.; 1995; Baffi et al.; 2000; Tsung; 2000; Kano et al.; 2001; Li and Qin; 2001). The augmentation of the X and Y-block increases the number of variables and especially in stiff systems (such as wastewater treatment processes) this may become cumbersome. An approach to decrease the number of variables by excluding overlapping samples is reported in Luo et al. (1999).

Nonlinear methods

In their basic form, PCA, PCR and PLS are linear methods and, consequently, there are limitations to what can be achieved when they are applied to nonlinear systems. Nonlinear pre-treatment of data is an often suggested method. This typically involves using the squared or logarithmic value of a variable. This is appropriate if the relation between variables is known to be nonlinear. Also, physical knowledge can be built into the model by using cross terms, e.g. for mass flows etc.

Different nonlinear PCA algorithms have been proposed in the literature, e.g. Kramer (1991), Malthouse et al. (1995), Dong and McAvoy (1996b) and Jia et al. (1998). Common to all algorithms is that they describe the relation between original variables and scores with nonlinear functions, identified by a neural network. The relation between $\mathbf{X}$ and $\mathbf{T}$ is:

$$\mathbf{X} = f(\mathbf{T}) + \mathbf{E} \tag{3.8}$$

Nonlinear PLS regression was proposed by Wold et al. (1989). Here, the relationships between X-scores and Y-scores are modelled in a nonlinear fashion.

$$\mathbf{U} = f(\mathbf{T}) + \mathbf{E} \tag{3.9}$$

The inner relationship f is typically a polynomial or splines (Wold; 1991). Other methods based on neural networks describing the inner relation are also found in the literature, e.g. Qin and McAvoy (1992), Malthouse et al. (1997) and Baffi et al. (1999). An interesting alternative to the neural network approach is reported by Berglund and Wold (1997).

Nonlinear multivariate monitoring algorithms have been applied to process monitoring in, for instance, Dong and McAvoy (1996a), Zhang et al. (1997), Jia et al. (1998), Shao et al. (1999), Fourie and de Vaal (2000) and Lin et al. (2000).

Adaptive methods

As mentioned earlier, many industrial processes do not display a stationary behaviour. Operational conditions change due to reasons such as varying raw material quality, surrounding temperature, varying process load and equipment wear. This is not an ideal situation for the methods described above. They all rely on the assumption that data are stationary in the time scale of interest. Consequently, extensions to the basic algorithms must be implemented to overcome this difficulty.

The way to address this problem depends on the nature of the process drift and two major cases can be distinguished. The first case originates from univariate changes in mean and variance, that is mean and variance are varying, but the qualitative relations between variables stay the same. In this case, it is sufficient to update the scaling parameters (mean and variance) of the data as shown in Rosen and Lennox (2001).

The second case involves changes in the relations between the variables (covariance structure) in addition to changes in the mean and variance. Here, the covariance structure of the model must be updated. A straightforward way is to use a moving (rectangular) time window, on which the model is based. A more sophisticated way is by recursive means (Helland et al.; 1991; Wold; 1994; Dayal and MacGregor; 1997a; Qin; 1998; Stork and Kowalski; 1999; Ouyang et al.; 2000; Li et al.; 2000). The principle of the updating schemes is that when new data are available they are included in the data matrix according to certain weights. For recursive methods, these weights are exponentially decreasing so that the history is increasingly disregarded as the monitoring progresses. There is usually a need for an updating criterion, to ensure that only data that are representative for the process are used in the updating of the model. Recursive models do to some extent reduce the problem of nonlinearities, as a recursive model can be regarded as a linearisation of the system at the current operational point.

Multiscale methods

It has been pointed out that multivariate statistics does not take into account the multiscale nature of process data (Bakshi; 1998). When deviations occur on multiple scales it is difficult to discern small but important deviations since they may 'drown' in the residuals of, perhaps, less important but large deviations.

Wavelets and multiresolution analysis (MRA) provide a solution to this obstacle. Wavelets and MRA constitute a framework for decomposition of signals into separate time scales. The theory of the framework is described elsewhere, see e.g. Vetterli and Herley (1992), Strang and Nguyen (1996), Alsberg et al. (1997) and Torrence and Compo (1998), and in this context, MRA can be seen as a specific group of filterbanks. A signal is decomposed into several scales (Figure 3.4). The highest scale contains the high frequency information of the signal. Successively, lower scales contain lower frequencies until a predefined depth. The remainder of the signal constitutes the lowest scale and is a low pass filtered version of the original signal. MRA may be expressed as continuous or discrete (Shensa; 1992; Rioul; 1993), and for the purposes of this work, only the discrete form is used. An appealing feature of MRA is that perfect reconstruction of a decomposed signal is possible. This means that all separate scales add up to the original signal.

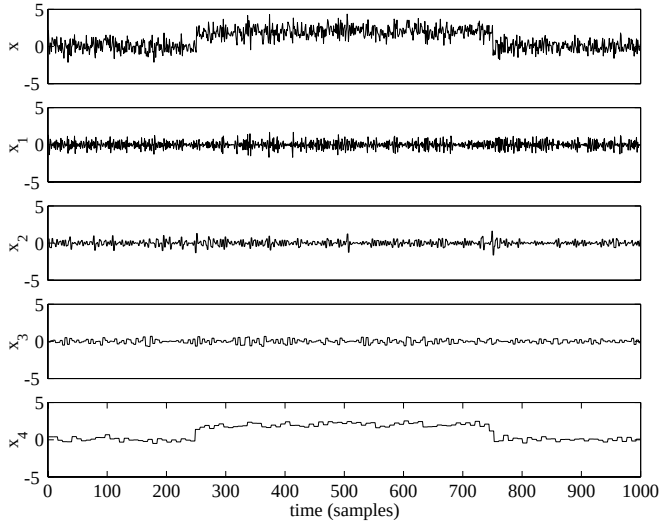

Figure 3.4: Decomposition of signal into four scales. Original signal in the most upper panel.

For monitoring purposes, MRA is used to decompose each variable into a number of scales as a data pre-treatment step, and a monitoring model is identified for each scale (Kosanovich and Piovoso; 1997; Shao et al.; 1999; Rosen and Lennox; 2001). By monitoring each time scale separately, the models are specialised on certain features, which means that the covariance structure may differ significantly between the scales. The result is an increased sensitivity to small, but significant, deviations. Moreover, MRA on the measurements provide solutions to a few other difficulties. Since the MRA decomposes signals into separate time scales, the autocorrelation of the signals is reduced (Bakshi; 1999) and one can partly justify the use of static monitoring models. Furthermore, since the scales will have zero mean (except for the lowest scale), models need not be updated if the variance and the covariance structure are approximately constant.

The price comes at a higher complexity level; many scales must now be monitored. By combining scales into fewer and more physically interpretable scales, the complexity is somewhat reduced (Rosen and Lennox; 2001). However, Bakshi (1998) proposes a multiscale PCA in which data are decomposed into several scales and PCA models are used to determine whether the scales are sig-

nificant at each sampling instance. Through reconstruction of only significant scales and monitoring by a uniscale PCA, the advantages of MRA are combined with a low dimensional PCA model (Bakshi; 1998; Rosen and Lennox; 2001). In Lennox and Rosen (2001), an adaptive algorithm based on the multiscale PCA is proposed. Other applications where MRA is combined with multivariate process monitoring are found in Trygg and Wold (1998) and Trygg et al. (2001).

Hierarchical and multiblock methods

When the number of variables is high, the interpretation is complicated. By organising the data in blocks and perform multiblock or hierarchical PCA/PLS (see e.g. Wangen and Kowalski (1988), MacGregor et al. (1994), Wold et al. (1996) and Rännar et al. (1998)) the interpretability may be increased. Data are organised in layers where the scores or latent variables of the lower level are used to form models at higher levels (Figure 3.5). This may prove advantageous in systems where the variables are generated from different parts of the process, each constituting separate process units. An example from a wastewater treatment plant could be that the data from the biological process, chemical precipitation and sludge treatment would form separate blocks that are unified in a model on a superlevel.

Batch process methods

For batch process monitoring, the situation becomes somewhat different to that of continuous process monitoring. Here, the structure of data is increased by yet another dimension—the batch. Thus, the data matrix structure is three-dimensional with variables, time and batches representing one dimension each. There are some different ways to address this problem. In multiway PCA (Wold et al.; 1987a) data are unfolded into a two-dimensional structure and then PCA is applied before the data are folded back. Other useful methods include parallel factor analysis (PARAFAC) and Tucker3 (see e.g. Dahl et al. (1999), Louwerse and Smilde (2000) and Bro et al. (2001)). In fact, batch processes constitute a large share of industrial processes and significant amounts of work have been carried out to adapt multivariate statistics to such processes. However, this is outside the scope of this work and readers are referred to the cited publications and the references therein.

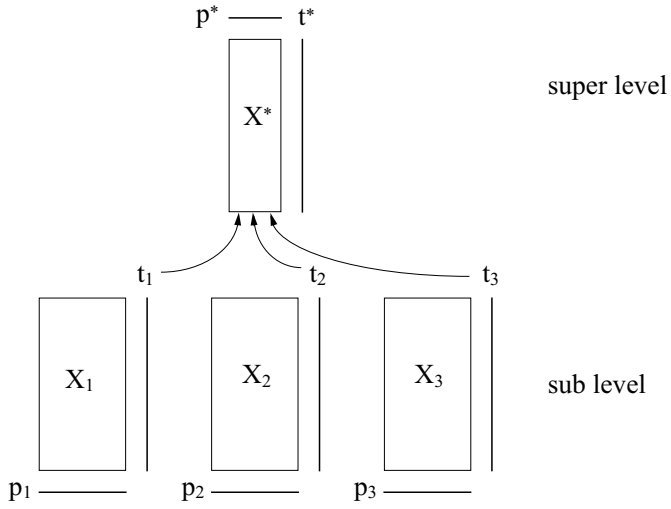

Figure 3.5: Principle of hierarchical PCA with a sub- and superlevel.

Some comments on statistical issues

Multivariate statistical methods raise some statistical issues. Most of these issues are outside the scope of this work. However, it is appropriate to discuss a few of the most important statistical difficulties encountered when dealing with large data sets from industrial processes.

It is possible to determine confidence limits for the scores, SPE and T^2. Usually it is assumed that data are normally distributed and the observations are independent. However, in the monitoring case this assumption can be relaxed. The only important assumption needed is that the data used for deriving the models are representative for the common cause variations in the process (MacGregor; 1997).

Non-parametric confidence limits can also be used (see e.g. Rosen (1998a)), and more importantly, limits can be derived from experience, especially the operator's own experience. If it is established that a confidence limit is too sensitive/insensitive the limit is simply changed to suit the operation.

3.4 Discussion on the applicability to WWTP operation

In this chapter, multivariate statistics have been discussed together with a number of different extensions to adapt the techniques to process monitoring. Which of these extensions must be incorporated in an algorithm for wastewater treatment operation? The best choice of method is often a balance between performance and complexity. The increase in performance must motivate the increase in complexity. In this work, the stance is that the major problems that need to be addressed are the non-stationary and multiscale nature of data, whereas dynamics and nonlinearities may be of less importance. The reasons for this stance will be discussed in this section[3].

Static or dynamic

At a first glance, it is tempting to claim that a monitoring technique for wastewater treatment operation must be dynamic. We certainly know that the wastewater treatment processes are dynamic and, hence, so should the model be. However, a closer investigation yields that this claim becomes weaker due to reasons discussed below.

The cause-effect relationships in a wastewater treatment plant (only continuous treatment is discussed here) are rather complex due to recirculation streams, generally the internal recirculation and sludge recirculation. Thus, the course of events in the first reactor is not only dependent on what happens in the influent flow, but also on what happens in the other reactors and the settler. This will severely complicate a cause-effect relationship analysis based on, for instance, cross-covariance function studies. Consider, as a simple example, a predenitrification plant with a reactor volume of 5000 m^3. Let the average influent flow rate be 10,000 m^3/day, which means that the plant would have a hydraulic retention time of 12 hours. From an input-output point of view, the retention time does not depend on the internal recirculation stream but looking at the actual flow rate between the reactors the effective retention time is decimated to 4 hours if the internal recirculation is 200 % of the influent flow rate (the same reasoning is valid for sludge recirculation if the volume of the settler is included in the example). If the recirculation flow rate is increased further (which is not exceptional), we approach a situation that can be considered as a completely

[3]The following discussion is based on the experience of the author and may not be supported by, especially, method developers.

mixed reactor. Now, if the sampling time is significantly faster than the effective retention time and the aim of the monitoring is to detect correspondingly fast disturbances, a dynamic approach is adequate. However, if this is not the case, a static approach is probably sufficient.

A second objection to dynamic models in wastewater treatment is related to the stiffness of the system; to model a stiff system with time constants ranging from minutes to months involves many lagged duplicates of especially the slow varying variables. From a situation of tens or hundreds of variables, we may end up with thousands or even more variables. This increases the complexity considerably and the interpretation task becomes cumbersome.

Linear or nonlinear

Many subprocesses within wastewater treatment display a nonlinear behaviour and nonlinear methods may provide a remedy if these are to be modelled. However, from a macro point of view, which is normally the view of the operators, wastewater treatment displays a surprisingly linear behaviour. There are important exceptions, for example sludge loss, but for a plant in a normal operational state the nonlinearities are often well behaved. By well behaved we mean nonlinearities that display smooth and monotonic behaviour. Monitoring is often a case of classification of the current operational state into one of two classes: normal or abnormal. For example, let a linear model approximately describe the normal region of the operational space. Deviations outside this region are driven by linear and/or nonlinear mechanisms. When a deviation is established it is often of less importance 'how much' abnormal the current state is. The very fact that a deviation has been established is serious enough to invoke actions (Figure 3.6). Thus, it is not obvious that a nonlinear model increases the practical monitoring quality sufficiently to motivate the increase in complexity.

The situation is different when a prediction model is the objective or when there is obvious nonlinear behaviour in the normal operational region. In such cases, the only feasible way is nonlinear modelling. However, this is outside the scope of this work and will not be discussed further.

Constant or adaptive

The non-stationary nature of wastewater treatment data surely calls for adaptive models. Models that cover large regions of operation (typically identified from

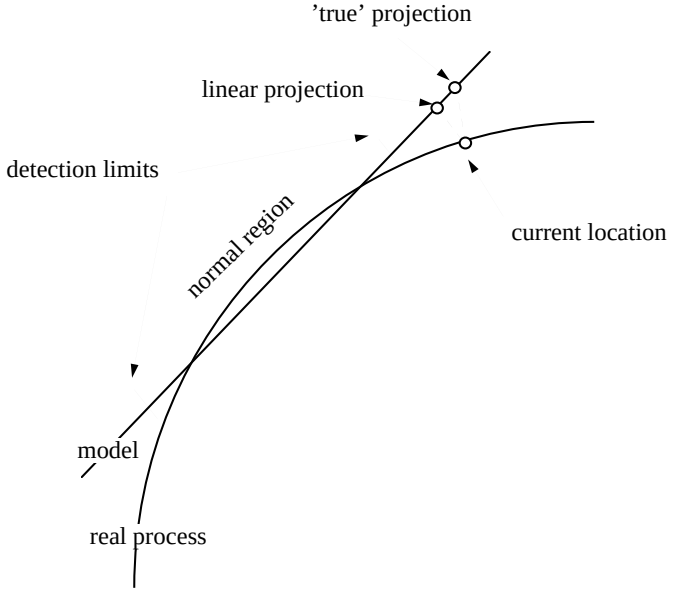

Figure 3.6: Illustration of a linear model representing a smooth, monotonic nonlinear process. If the normal region is adequately described by a linear model, the discrepancy between the linear projection and the 'true' projection is of less importance if a violation of the limits has already been established.

months or even years of data) can be used for supplementary information on the 'absolute' location of the current operational state. However, for day-to-day monitoring, adaptive models is an appealing alternative to frequent identification of new models, which of course is costly (and basically the same as adaptive models) or to have a library of monitoring models representing different operational states. Another important advantage of adaptive models was mentioned earlier; adaptive models can be seen as a linearisation at the current operational state and, thus, mitigate the need for a nonlinear process representation.

Uniscale or multiscale

Wastewater treatment data display a multiscale nature. Disturbances occur in many different time scales and it is often difficult to discern small but significant changes in the 'background' variation caused by the varying influent conditions.

By multiscale monitoring, the sensitivity for these types of disturbances is increased. The multiscale approach is also partly a solution to the previously discussed issue of dynamic models. Each scale represents a 'snapshot' in a limited range of frequencies and, thus, each scale model will intrinsically only describe the variable relations of that range. The increased complexity can be hidden from the end users by means of adaptive multiscale models, as demonstrated in Paper D or by combining adaptive models with the approach of interpretable scales discussed in Paper C. However, for smaller systems with less wide ranges of time constants, a dynamic model may be a more adequate choice.

3.5 Multivariate monitoring in wastewater treatment

There are several examples of applications of multivariate statistical monitoring (and modelling) to wastewater treatment operation in the literature. An early application of PLS modelling of wastewater treatment plants is given in Aarnio and Minkkinen (1986). Monitoring of the effluent water from a paper and pulp wastewater treatment plant using PCA is reported in Kim Oanh and Bengtsson (1995). In Teppola et al. (1997), it is reported that PLS modelling seems to be a promising tool for detection of shifts in variables. Isolation of deviating variables is also possible. Classification of the operational state in principal component space to obtain a decision support system is discussed in Sànchez et al. (1997). Rosen and Olsson (1998) use PCA and PLS for disturbance detection and prediction of wastewater treatment operation. Contribution plots are used to isolate deviating variables. A comprehensive discussion on both single variable and multivariable monitoring of wastewater treatment operation is found in Rosen (1998a). Modelling of wastewater treatment processes using PLS and PCR are reported in Mujunen et al. (1998) and Teppola et al. (1998b). Fuzzy clustering and PLS is combined in Teppola et al. (1998a) to predict the sludge volume index and to interpret the results. Teppola et al. (1999) utilise adaptive fuzzy clustering combined with PCA for process monitoring to handle seasonal variations. Yet another combination of two clustering techniques and PLS is reported by Teppola and Minkkinen (1999). Detection of process disturbances by examination of the trajectories in the score space is reported in Weiss et al. (1998) and Pons et al. (1999). Some results from basic PCA monitoring of a wastewater treatment plant are reported in Bendwell (2000). Rosen and Yuan (2000) use clustering and PCA to identify different operational states in a supervisory control framework. In Rosen and Lennox (2001), adaptive PCA and

a combination of wavelet analysis and PCA is used to overcome problems associated with non-stationary and multiscale data. A slightly different approach based on dissimilarity indices are used for disturbance detection in Yoo et al. (2001) and in Lennox and Rosen (2001) an adaptive multiscale PCA is proposed for wastewater treatment monitoring.

Chapter 4

Multivariate feedback adjustment for control

Multivariate statistical techniques can be integrated with process control to form a supervisory level in the control system. Information from the monitoring system is used for the control of the process to derive appropriate control strategies or actions suitable for the current operational state. In this chapter, a discussion on how this can be done and which control task may be solved by doing so, is presented.

4.1 Challenges

The increase in the number of measured and controlled variables improves the observability and controllability of the process. Instead of passively observe the effects of disturbances and process changes, it is possible to react to them and adjust the control so that disturbances are attenuated. The monitoring information plays an essential role in this work. Although multivariate monitoring methods are used for both disturbance detection and isolation, it is not obvious how to adjust the process so that a desired result is obtained. This becomes especially obvious when the number of manipulated variables is high. The problem of process adjustments are here divided into three subproblems:

> Extreme event control—bringing the process from an abnormal state back to the normal state.

Disturbance rejection—finding appropriate control actions that attenuate disturbances.

Product design—finding the operational region to achieve a certain output product or quality.

The control objective in extreme event control is generally to drive the process back to the normal operational state. This may involve completely different control actions than what is normally used. Discrete actions like step feed, use of equalisation tanks, etc. are typical examples of such actions in wastewater treatment operation. Further, it may involve a shift in the overall control objective, from process output quality to process safety (e.g. sludge inventory control). It may also involve process shutdown, if the safety of the process cannot be ensured. In wastewater treatment, this would correspond to bypassing. Disturbance rejection or attenuation is here defined as minimising the effects of disturbances that must be considered as normal. These are, for example, disturbances due to changes in the influent conditions or temperature variations due to seasonal effects. It is, thus, an issue of adapting the process to the changes in the operational conditions. In process industry, a product is generally characterised by certain composition of different compounds, material, etc. Product design implies determining how the process should be operated to obtain this product. An example of this in wastewater treatment is controlling the process to obtain desired sludge properties in the secondary settler. Here, the product is the sludge, and its properties are not generally directly measurable (online). However, sometimes the desired output quality is directly measurable. In these cases, the process is operated to attain a setpoint on the measured quality variable. Controlling the process to certain output setpoints can be considered as a special case of product design. In wastewater treatment, controlling the process to attain certain setpoints on effluent key variables, such as nitrogen, phosphorous, suspended solids, organic matter, etc., is an example of the special case.

The three problems constitute supervisory control. Supervisory control is typically implemented on top of the local control layer, but with longer time constants than that of the local controllers. Examples of the levers available to exercise supervisory control are coordination of several local controllers, invoking new control actions and shifting the control objectives. In the remainder of this chapter, a discussion will be given on how supervisory control can be implemented in wastewater treatment, utilising ideas from multivariate monitoring.

4.2 Extreme event control and disturbance rejection

Extreme event control and disturbance rejection are similar problems. An extreme event can, of course, be regarded as a severe disturbance. It is therefore not surprising that the methods to solve these problems are similar. In this section, two alternative, and as we will see later, complementing methods to integrate multivariate monitoring and control are discussed. First, methods based on detection and classification of the current operational state are considered. Here, the current operational state is first determined and then the information is used to apply appropriate controller setpoints. Second, a multivariate monitoring model is inverted and used directly to calculate suitable controller setpoints.

In the first approach, the determination of appropriate setpoints, e.g. by models or look-up tables, is done independently of the detection/classification and the approach can be seen as a multistep approach. This means that the setpoint determination does not rely upon the same assumptions made for the monitoring model. This is useful when the operational state is outside the valid range of the monitoring model and additional control handles are used. In this work, this approach is taken in the case of extreme event control. In the second approach, the determination of setpoints rests on the same ground as the monitoring model. This means that the current state must be within the region that is adequately represented by the monitoring model. This direct approach is used for 'normal' operational state control.

Multistep approach

Step 1—detection/classification

The basis of this methodology is that different operational states can be represented by different locations in a multivariate space. By identifying these locations beforehand, the current operational state is analysed and classified in accordance to previously encountered states. The multivariate measurement space is reduced by projecting it onto a smaller space defined by, for instance, a PCA model. In the reduced space, regions corresponding to different operational states (disturbances) are identified. The identification of regions can be done manually or by using an unsupervised clustering algorithm. When new data are projected onto the model space, the current locations are classified using clustering. From the clustering, a membership function that describes to what

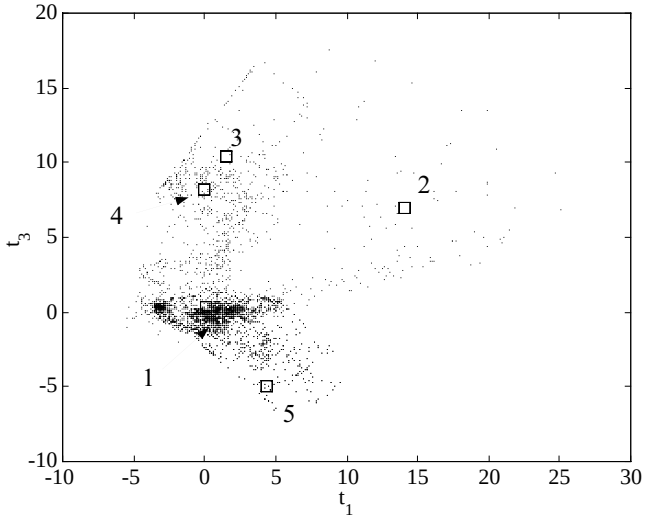

Figure 4.1: Five clusters representing different disturbances projected on a low-dimensional score space (squares indicate cluster centres).

extent the current location belongs to the predefined classes is obtained. The membership function can be crisp, the location only belongs to one class at the time, or it can be fuzzy, the location may belong to several classes. Fuzzy clustering is useful in the case when the boundaries between regions, corresponding to different states, are uncertain or when some regions are not covered by a class. Fuzzy boundaries also provide a means to seamless transitions between classes.

In Figure 4.1, an example of a reduced space (score space of a PCA) with a number of clusters corresponding to different operational states, is shown. Even though the clusters are somewhat hard to see, the clusters are easily discerned using clustering algorithms (there are more than two dimensions in the reduced space).

It is desired that as many known disturbances as possible are included in the training data. To check the performance, the algorithm needs some kind of surveillance. By monitoring the squared sum of distances to all clusters, a measure for evaluation of the classification is obtained. In this way, a limit is set on the maximum allowed deviation from the known classes so that the result also may be classified as 'unknown'. When an unknown disturbance occurs

the classes need to be updated. This can be done by simply adding a new class. However, if the new disturbance is established to be only a variation of an already known class the centre of the location old class can be shifted manually or by use of adaptive clustering (Marsili-Libelli and Müller; 1996; Teppola et al.; 1999).

Step 2—setpoint determination

From step 1, the current operational state is obtained. To determine appropriate setpoints, various methods may be used. The most basic way is to have a set of pre-defined controller setpoints for each type of disturbance. Such a look-up table is typically based on experiences from previous encounters with the same type of disturbance. This is an intuitive method, but may imply that too large safety margins are used. A more refined way is to calculate the setpoints using process models with the current process state as initial condition. The models can be of various level of sophistication, ranging from steady state models to fully dynamical models implemented within the model predictive control (MPC) framework (see e.g. Garcia et al. (1989), Camacho and Bordons (1999), Morari and Lee (1999) and Mayne et al. (2000)). Whatever method is used, the advantage of using state classification prior to setpoint determination is that it reduces the complexity of the model, since knowledge of the disturbance is available. Disturbances affect the process in different ways and different time scales. Consequently, each model does not have to describe all the dynamics in the process; it only has to describe the dynamics that are influenced by the disturbance in the time scale of interest. This fact is utilised to tailor models for specific purposes.

Step 3—post-treatment

When new controller setpoints have been calculated, using look-up tables or models, the final controller setpoints are weighted according to the membership function determined in step 1. This means that if crisp classification is used, the weights are either 1 or 0, but in the case of fuzzy classification the weights will be between 1 and 0. Thus, parallel computations of setpoints are required at transitions between classes. In Figure 4.2, the framework structure is shown.

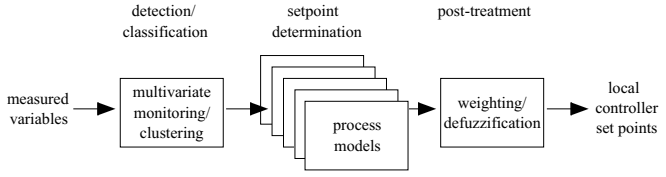

Figure 4.2: Principle of multistep control, including determination of the current operational state, determination of appropriate set points and post-treatment.

Direct approach

In the direct approach to disturbance rejection using multivariate monitoring, the model is inverted so that appropriate controller set points are computed directly from the model. In its most basic form, this is done when loading plots are investigated and control actions are derived from the analysis. From loading plots, which are the column vectors of $\mathbf{P}$ (see Equation 3.1) plotted versus each other, a lot of information are obtained. A loading plot describes the relations between the real variables and the principal components (or latent variables). Moreover, such a plot also describes the mutual relations of the real variables. Thus, a loading plot can be used to investigate possible contributing variables to a deviation along a principal component but it can also be used to find variables that would drive the process back to normal (disturbance rejection). A loading plot is shown in Figure 4.3.

Using loading plots to find appropriate control adjustments is a manual procedure; the operator investigates the plots and draws conclusions on which adjustments of the control are based. However, the same way of thinking is extended to an automatic derivation of adjustments to the local control. The difference between the current and desired location in score space is mapped to a difference in the original variables (Figure 4.4). Assuming that some of the variables are manipulated variables, it is straightforward to calculate the change in the manipulated variables so that the process moves towards the desired location. For data that are mean centred this means that controller adjustments strive to force the process to the origin in the score space. This approach to control the process in the score space was proposed by Piovoso and Kosanovich (1994)[1]. The original method is afflicted by a weakness: it does not address the

[1]The authors briefly indicated this approach in an earlier publication (Piovoso et al.; 1992).

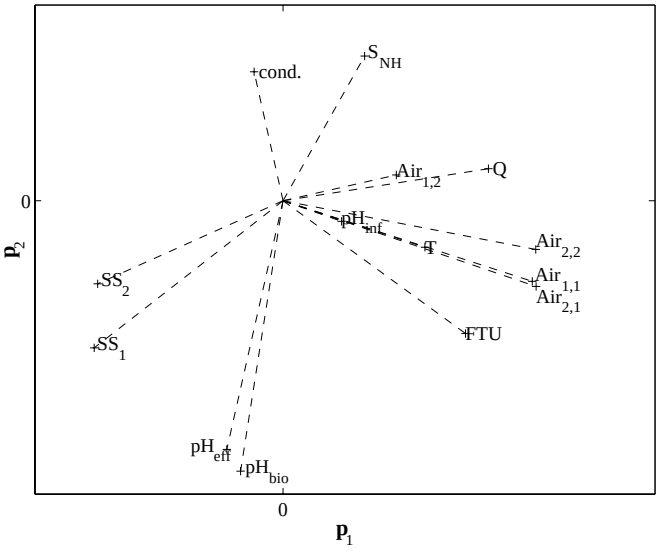

Figure 4.3: The relations between variables and the principal components are visualised by a loading plot. The plot are used to find suitable controller adjustments to drive the process in a certain direction.

change in the system characteristics introduced by the closing of the loop; when feedback information is used to derive control adjustments, the system cannot be considered as an open loop system anymore. Other researchers have proposed extensions to the methodology, which address this problem (Chen and McAvoy; 1996; Chen et al.; 1998).

4.3 Product design

The problem of product design is different from that of disturbance rejection in the sense that here the supervisory control system must be able to control the process to certain output specifications instead of simply controlling the process with a minimum of deviation from 'normal' operation. A quality requirement must be defined and is typically one or several setpoints imposed on certain variables. It can also be expressed as a certain composition of the end product.

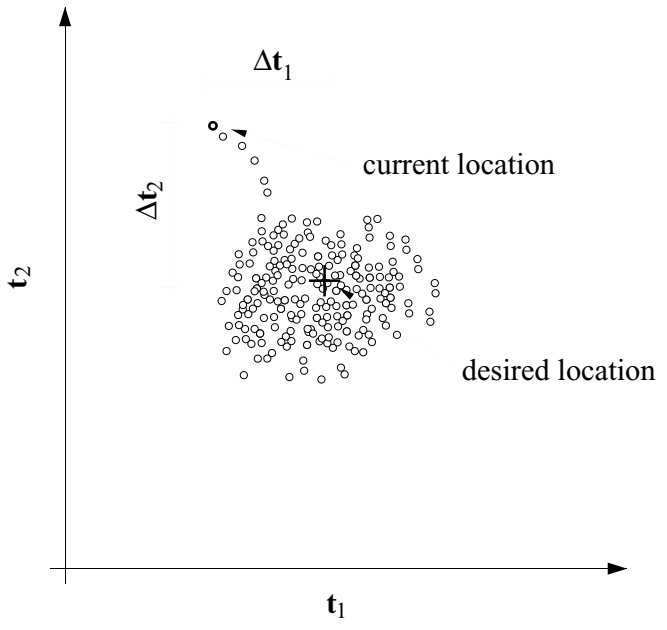

Figure 4.4: Principle of score space control. The differences in the score space ($\Delta \mathbf{t}_1$, $\Delta \mathbf{t}_2$, $\Delta \mathbf{t}_3$,...) are mapped to differences in the original variable space.

Extended direct approach

Assume that a certain product quality is desired. This product quality is specified using one or several measurable variables. The problem consists of finding the values of manipulated variables that will yield the desired quality and, hence, the relations between the process variables and the quality variables must be known. If the quality variables are regressed onto the process variables, using for example PCR or PLS regression, such relations can be found. Now, since the desired quality is known, the inverse of the model will give us process variable values that will fulfil the quality requirements. Let the model be a latent variable model, that is with a reduced process space, and the number of latent

variables be a. Furthermore, let the number of quality variables be k, then three different cases may occur (Jaeckle and MacGregor; 1996):

> $k > a$, that is the number of latent variables is smaller than the number of quality variables. The system is overdetermined (generally no exact solution) and least squares is used to find the best approximation.
>
> $k = a$, that is the number of latent variables is equal to the number of quality variables and there is a unique solution.
>
> $k < a$, that is the number of latent variables is higher than the number of quality variables. Here, the system is underdetermined (infinite number of solutions) and the pseudo-inverse[2] is used to find one possible solution.

In an industrial process, the last case is the most realistic. The use of the pseudo-inverse is justified since it will yield the solution with the smallest Euclidean norm. Consequently, the solution is the one with the smallest variation in the latent variables, or put differently, the solution is the one closest to the origin in the score space. For mean centred data, this is a reasonable choice. However, since there is an infinite set of solutions, it is shown in Jaeckle and MacGregor (1996) and Jaeckle and MacGregor (2000) that a 'window' of solutions can be obtained, from which suitable solutions are chosen, based on process knowledge and operational history. This is advantageous when the operational space is limited by physical or control related limitations.

Model errors are introduced by a number of causes. First, the closing of the loop will introduce a distortion of the covariance structure. Second, controller saturations or physical limitations will also alter the structure. Third, disturbances due to external variations or system changes affect the model agreement and, fourth, model mismatch due to nonlinearities, insufficient excitation in the identification phase, etc., will see to that there is a discrepancy between the model and the process. If this is not compensated for, the controller will not yield the desired result; the process location in the model does not agree with the real process location. In Rosen and Jeppsson (2001a), a possible solution to this difficulty is presented for the case when the quality variables are available

[2] A more elaborate discussion of the use of different pseudo-inverses is given in the addendum of Paper G.

online. A compensation term is introduced to correct the model errors and the method is capable of compensating errors online.

The direct method, as implemented in Rosen and Jeppsson (2001a), can be considered as a multivariate steady-state feedback approach. Therefore, it is best suited for supervisory control in longer time scales, controlling the average value of the quality variables. Of course, the suitable time scale is dependent on the controlled system and for a wastewater treatment plant, this is daily or weekly average control. The long control perspective may, or may not, be advantageous. In many countries, the effluent quality requirements are policed by use of average values and in this case the approach would be adequate. However, other countries use grab samples or discharge fees. Here the approach is less suitable. However, it is shown in Rosen and Jeppsson (2001a) that the short term performance can be considerably improved by simple feed-forward terms. Another possible improvement to the direct method would be to incorporate the ideas from multiscale monitoring. The supervisory controller consists of several time scales on which control models are identified for each scale. As in the monitoring case, this may provide a way to circumvent the problem of addressing process dynamics. The multiscale framework does also allow for temporal decoupling of otherwise coupled variables or states. Multiscale procedures apply to both the multistep and direct methods and some aspects of multiscale control are discussed in Stephanopoulos et al. (1997), Stephanopoulos and Ng (2000) and Alsberg (1999).

Chapter 5

Summary of work

5.1 Introduction

The driving forces behind the often substantial collection of online data in industrial processes are generally related to quality, safety and economic requirements on the processes and their outputs. The introduction of computers and instrumentation in the operation allows for a vast amount of data to be collected. The data may provide an important source of information but one should not confuse data and information. The information is generally hidden in the data and more or less sophisticated techniques are necessary to extract adequate and reliable information.

The quality of the obtained information is strongly dependent of the quality of the measurements. The devise 'garbage in - garbage out' is certainly applicable to information extraction and, therefore, methods to test and improve the quality are important tools to ensure reliable information. Another important aspect is the appreciation of information. Remember that the information must be interpretable and understandable for the users before it is truly informative. This is especially imperative in the case when process data include many variables of different engineering units.

Online data monitoring is normally carried out to investigate the current process status; is the process behaving normally and are the output variables meeting their requirements. Monitoring data online can be carried out in a univariate or multivariate fashion. In univariate monitoring, variables are monitored separately and no consideration is taken to the mutual relationships between variables. In contrast, multivariate monitoring utilises these relationships to ex-

tract more information and reduce the dimensionality of the monitoring problem.

Monitoring information can be used to increase the knowledge of the process and the major mechanisms that drive the process. However, the primary incentive for extracting information is control. To be able to react quickly and firmly when process deviations are detected may be the difference between failure and success. By combining the monitoring task with the task of determining appropriate setpoints for the local controllers, a fast response to disturbances is made possible. The integration of monitoring and control, thus, provides a framework for automatic supervisory control.

In this work, both univariate and multivariate techniques for extraction of information from wastewater treatment operation data are discussed, with a clear focus on the multivariate techniques[1]. The multivariate approach has proven successful in many industrial applications and in this work, it is shown that wastewater treatment is no exception. Further, alternative ways to use monitoring information in the control system to obtain automatic supervisory control are presented. The methods used are based on recent developments in the field of chemometrics.

5.2 Univariate monitoring

Measurement quality

Before any analysis of process data is carried out, it is important to validate the quality of the data. In Paper A, a number of obstructions that complicate the information extraction process are briefly discussed. These obstructions include missing values, noise and outliers. Different digital filtering techniques provide solutions to these problems but one needs to be careful, especially in the case of outliers, so that valuable process information is not discarded.

Statistical process control

Statistical process control (SPC) is a framework that was originally developed for univariate monitoring. Here, single measurement signals are analysed online with respect to a number of (mostly) statistical measures. Typical measures are the variable's amplitude, mean, variability, rate of change, trends and de-

[1] Single variable analysis is discussed more thoroughly in Rosen (1998a).

viation from normal situation. Generally, the measures are monitored as time series with parametric or non-parametric confidence intervals representing the normal region. It is shown in Paper A that these methods are applicable to wastewater treatment operation and an example of detection of sensor failure using variability monitoring is given.

Two major disadvantages of univariate monitoring for industrial processes must be mentioned. First, when the number of measured variables is high, the interpretability of the whole monitoring problem decreases. Monitoring many variables is a cumbersome task and there is an imminent risk of variables being neglected. Second, univariate monitoring does not account for collective effects. A multivariate process may be outside its normal region with all variables still within their respective limits. In such situations, univariate monitoring is not sufficient.

5.3 Basic multivariate monitoring

Multivariate monitoring techniques

A preliminary study of the applicability of multivariate statistical process control (MSPC) for process monitoring is presented in Paper A. Through a number of examples, it is shown that principal component analysis (PCA) can successfully be used to detect deviating process performance. PCA is a method to reduce the dimensionality of a problem, by projecting a high-dimensional space onto a space of a lower dimension using the fact that most industrial data display high degree of correlation. The principal components (PCs) define a space in which the scores (the projected data) are linear combinations of the original measurements and constitute pseudo variables that capture the major mechanisms of the process. By reducing the space, information is disregarded. However, this does not mean that adequate information is wasted. As a matter of fact, the essence of PCA is that the adequate information is represented by the principal components, whilst non-adequate information is wasted (or considered as noise). Normally, data are mean centred and scaled so that measurements of different units and numerical amplitudes may be compared. If this is not done, a variable with high numerical amplitude may affect the model in a disproportional way. More detailed information on PCA is given in Paper B.

In Paper A, partial least squares (PLS) regression is also discussed. PLS can be considered as a development of PCA and the important difference is that in

PLS, data are divided into two blocks; one block represents the independent variables (X-block) and the other represents the dependent variables (Y-block). By reducing both blocks simultaneously, a regression model with monitoring as well as predictive power is obtained. PLS is generally used for prediction of one or several variables, but can also be used to focus the detection effort on the variables most influential on one or several output (or quality) variables. Thus, if PCA can be considered as a technique for general process monitoring, PLS (and other regression methods) is a technique for specific process monitoring techniques[2].

Basic online monitoring using PCA and PLS involves offline identification of a model from data that represents normal operational conditions. New data are projected onto the model and the scores and/or the model residuals are then monitored as new samples are obtained. Note that it is imperative that new data are scaled and mean centred in the same way as training data.

Information visualisation

The scores can be monitored using conventional univariate SPC techniques with calculated confidence. However, since the dimensionality of the task may still be substantial, this may be cumbersome. Instead, inherent characteristics of PCA (and PLS) are utilised. The first component describes the direction of the largest variability of the X-block and the second component describes the second largest direction and so on. This means that by plotting the first PC versus the second PC, a significant part of the variability in the X-block is covered by a so-called score plot (in the PCA example in Paper A, the first two PCs cover more than 75% of the variability of the X-block). A confidence region is calculated and whenever the current location is outside this region, a disturbance is established. This is an intuitive and comprehensive way of displaying high dimensional data.

The number of adequate and information carrying scores may be too high to make score plots applicable. Then, two summarising measures may be used. The first is the Hotelling's T^2, which describes the summarised variations within the monitoring model. T^2 can be monitored using confidence limits. However, T^2 must normally be complemented by the second measure: the sum of the squared prediction error (SPE). SPE expresses the summarised dis-

[2]General and specific methods are sometimes referred to as unsupervised and supervised methods, respectively (Davis et al.; 1996).

tance from the model, or put simply: how well new data fit the model. The reason they complement each other is that when a disturbance is manifested in 'coordinated' variable changes (i.e. the relations between the variables or covariance structure remains the same), this will be detected in the T^2 and the SPE will stay low. However, when the disturbance manifests itself as a disturbance in the covariance structure, the SPE will become high, while the T^2 stay low. In the general case though, a disturbance will cause both the T^2 and SPE to increase. A more comprehensive discussion on the T^2 and SPE measures is given in Paper B.

In Paper A and B, the above-discussed measures for visualisation of the operational state are used. It can be concluded that the measures provide a much more compact way of presenting process information compared to univariate time series plots. It should be noted that no (adequate) information is generally lost in the MSPC procedure.

Variable isolation

An appealing feature of most MSPC methods is the ability to isolate deviating variables. This is used when a disturbance has been observed. By backtracking through the model, the variables responsible for deviations in the score plots, the T^2 and the SPE can be identified. So-called contribution plots are used to analyse the contributions from individual variables. In Papers A and B, a few examples show that these plots can successfully be applied to wastewater treatment data.

5.4 Advanced multivariate monitoring

Handling non-stationary data

Basic MSPC using PCA assumes that data are stationary, i.e. the variable mean and variance are approximately constant. This is seldom the case in wastewater treatment operation due to diurnal, weekly and seasonal variations. The ever changing operational conditions either makes the monitoring model too insensitive to smaller changes (e.g. if a whole year's operational data are used to identify a model) or it makes the monitoring model less useful due to lack of fit. In Paper B, this problem is addressed by implementing adaptive monitoring models.

The adaptation of the model can be carried out in different ways. If the mutual relationships between the variables (the covariance structure) are believed to be approximately constant, the adaptation can be obtained by only updating the scaling parameters. A moving (historical) window can be used for updating. Here, the historical data have the same influence on the parameters. However, a more sophisticated way is to update the parameters recursively, so that exponential weighting is obtained. The forgetting factor, i.e. the speed with which historical values are disregarded, may vary depending on the aim of the model and the nature of data.

When the covariance structure is believed to change, the whole model has to be updated. The same approach as for the parameters may be applied to the covariance structure. A major disadvantage of updating the covariance structure is that one loses the possibility to use score plots for visualisation. This is due to the rotation of the space defined by the PCs caused by the continuous updating of the covariance structure.

It is shown in Paper B that adaptive models are applied successfully to real wastewater treatment data where a static model does not suffice. A period of more than 100 days, spanning from late summer to early winter and with significantly changing operational conditions, is investigated and the adaptive monitoring models detect deviations from normal operation. It is also shown that isolation of contributing variables is achieved to facilitate process diagnosis.

Handling multiscale data

Wastewater treatment data display a multiscale nature. Events and disturbances appear in many different time scales, from long term (months) to short term (minutes or hours). This is a problem for the MSPC techniques discussed here. Multivariate monitoring is generally carried out in one time scale. This time scale contains frequencies ranging from the Nyquist frequency to the lowest frequencies present in the process. The presence of different time scales introduces an error in the monitoring model and this error degrades the sensitivity and, consequently, the ability to detect small, but significant, changes in data. Small deviations are 'drowning' in the variations caused by, for instance, the varying influent conditions.

In Papers C and D, a framework for multiscale multivariate monitoring is presented. It is based on recent techniques for multiresolution analysis (MRA). In MRA, data are split into separate time scales using the wavelet transform.

The decomposed data can be evaluated by multivariate monitoring, for instance PCA, to obtain a multiscale monitoring methodology. Multiscale monitoring has some important advantages. The sensitivity of the monitoring model is increased as every scale is monitored separately. Moreover, the separation of data into multiple time scales implies that the higher scales will have approximately a constant mean and only the lower and/or lowest scale will display trends or long term variations. Consequently, by omitting the lowest scale from the monitoring, the problem of monitoring data from changing process conditions is partly solved. Also, information on the scale on which a disturbance or event appears, may be used in the interpretation to find the physical cause of the event or disturbance. An example of process monitoring using a combination of PCA and MRA is given in Paper C. The sensitivity is increased and it is also seen that the non-stationarity problem is solved.

Monitoring many separate time scales introduces an increased complexity; although PCA provides a reduction of dimensions we now have scores on many scales, and the total number of scales may be larger then the original number. There is, fortunately, a way out of this dilemma. By combining scales into physically interpretable scales, the number of scales is reduced and the interpretability is increased. In Paper C, such as approach is utilised to monitor process data. Compared to the case, in which no recombination is done, the interpretability is increased both due to a decreased number of scales and the fact that the scales better corresponds to the major time scales of the process.

A third multiscale approach is presented in Paper C. In this, monitoring models are used on each scale to determine whether a certain scale displays significance. The signals are then reconstructed using only significant scales. The reconstructed signals are monitored using a uniscale PCA model. Thus, the ability of the multiscale approach to detect small disturbances is combined with the dimension reduction of the uniscale approach. The result is a sensitive monitoring model generating information that is displayed in a compact way. It is shown in Paper C that the three methods have similar performance.

Paper D outlines an adaptive multiscale PCA (AdMSPCA) for process monitoring where the adaptive capabilities are combined with multiscale feature extraction. In analogy with the techniques described above, data are decomposed into several time scales using MRA. An adaptive model, similar to the one discussed in Paper B, is identified on each scale. Thus, each scale model follows the evolution of the process. In the paper, the AdMSPCA algorithm is compared with adaptive PCA. The AdMSPCA shows a greater ability to adapt to a wide

range of changes. Moreover, the AdMSPCA appears to be more sensitive to slower changes. This may be beneficial when sufficiently slow adaptation must be weighted against persistent violation of control limits.

5.5 Control adjustments

To integrate monitoring and control to form an automatic supervisory control scheme is a natural extension to monitoring. This enables the control system to react to process changes without operator intervention.

Open loop adjustments

When a disturbance has been detected and the current operational state is considered abnormal, it is desired to force the process to return to the normal state. This may involve invoking new control handles, normally not used when the process is in-control. It may also involve a shift in the control objective, from, for instance, low effluent nutrient concentrations to retaining the sludge in the system. In Paper E and F, a framework for extreme event control of wastewater treatment operation is proposed. The framework consists of several subtasks. Monitoring and classification of the operational state is performed by combining PCA and fuzzy clustering. Several regions corresponding to different types of extreme events (disturbances) as well as normal operation are identified in the PC space. Using the fuzzy clustering algorithm, a membership function that describes to what region the current operational state belongs is obtained. When the current operational state is identified, an algorithm determines appropriate setpoints for the local controllers. The output from the setpoint determination is weighted according to the membership function. The last step is carried out since fuzzy clustering allows a state to belong to more than one region. This enables seamless transition between different operational states and yields faster control response and smoother control actions. The proposed framework is outlined in Paper E. Here, look-up tables are used to determine appropriate setpoints. Consequently, the setpoint need to be established a priori based on operational experience.

A number of different test cases are studied by simulation. The test cases are based on influent data developed by the COST 624 benchmark group. The results show that the proposed scheme is capable of detecting and classifying different extreme events and that the implemented controller setpoint changes

improve the performance of the plant according to the control objectives.

The setpoint determination of Paper E is further developed in Paper F. Look-up tables, steady-state and dynamic models are compared to assess if an increased complexity yields improved performance. A reduced order dynamic model is developed. The model relies only on measurements that are practically possible to obtain in a real situation. The steady-state controller (SSC) control law is obtained by linearisation of the reduced model at the current operating point. The inverse steady state relationship between inputs and outputs is calculated. A more sophisticated model based setpoint determination is obtained by introducing model predictive control (MPC). In MPC, the reduced model is used to simulate various control setpoints at each sample. The 'optimal' setpoints are chosen, using an optimisation criterion.

The different setpoint determination strategies are tested in a simulation study. It is shown that both SSC and MPC yield better performance in terms of control costs and flexibility than the look-up table. An important and appealing feature of MPC is the ability to design a cost function and to account for controller saturation. This may be valuable when a cost-benefit assessment is desired. However, it is not obvious that the increased complexity can be justified in all cases. The look-up table strategy is simple and robust to measurement disturbances and the choice between the strategies is a balance between performance and simplicity.

Feedback adjustment

During normal operation, the situation is somewhat different to that of extreme events. Here, the task of the supervisory control system can be said to be twofold: disturbance rejection and product design. In normal operation, disturbance rejection implies minimising the effect of disturbances that must be accepted as normal. Such situations include the diurnal pattern in the influent characteristics, temperature changes due to weather variations, inhibition effects, etc. It is not an issue of returning the process back to normal; it is rather an issue adapting the process to the changes in the operational conditions. In contrast to disturbance rejection, product design involves changes to the control system to achieve specific requirements on the effluent quality. In wastewater treatment operation, this generally implies determining what controller setpoints are required to meet the effluent standards imposed on the operation.

In Paper G, a chemometric approach to supervisory control of wastewater treatment is proposed. The main objective of the work is to control the mean effluent nitrogen concentration (product design) from a biological stage, configured as a predenitrification process. As a secondary objective, the supervisory control system should minimise the variation in the effluent concentration (disturbance rejection). A PCA model is used to monitor the current operational state. By inverting the model, local controller setpoints that drive the process to a desired location in the score space are determined. To compensate for model errors due to, for instance, identification difficulties, process nonlinearities, controller saturations, etc., a compensation term is introduced. The controller can be seen as a multivariate feedback controller.

Using the COST benchmark simulation model, the controller performance is evaluated. The first objective is achieved and it is shown that the controller can control the process to an arbitrary (within reasonable values) effluent setpoint for the effluent nitrogen concentration. It is also shown that by introducing a feed-forward term in the controller, the effluent concentration variation is significantly reduced. The supervisory controller approach poses two major drawbacks. First, the identification of the controller must be carried out during constant influent conditions. Second, the controller does not consider the costs of the proposed control actions. However, in the addendum some comments and possible solutions to these difficulties are given.

Chapter 6

Concluding remarks

The amount of data collected at industrial sites today is significant. Data are collected for many reasons, of which process monitoring and control are two important purposes. In process monitoring and control, data need to be treated online, which makes it more demanding than other data handling problems. Relevant information must be extracted from the data and presented and interpreted adequately within a short time span. The large amount of data puts special requirements on the methods used for process monitoring and control. In this work, a chemometric approach to meet these requirements is presented.

6.1 Summary of results

Process monitoring

In wastewater treatment process operation, the operators and process engineers face a number of difficulties and challenges to transform the vast amount of data into information. Faulty measurements due to erroneous sensors or noise arising from various sources in the sampling and measurement procedure lead to poor data quality. The high number of measured variables implies that there is a risk for 'data overload'; humans simply do not have the ability to analyse and interpret high dimensional problems. A process where many variables are measured does not necessary imply that the process inherently is high dimensional. Instead, data often display a high degree of redundancy, i.e. data are collinear. This does not only create difficulties for human interpretation, but also for conventional statistical analysis methods, since they rely on a

high degree of independence among the variables. Furthermore, the conditions in which wastewater treatment processes are operated are normally of a varying (non-stationary) nature. It is often difficult to discern process disturbances different from those caused by the varying influent conditions, which tend to have a dominant effect on the process behaviour. The fact that disturbances occur in many different time scales, which is another difficulty, complicates the distinction of disturbances in a similar way to that of non-stationarity and it also deteriorates the performance of many monitoring techniques. In addition to the difficulties already mentioned, the fact that the involved processes are nonlinear and well as dynamic must be accounted for.

In this work, multivariate statistical process control (MSPC) is investigated as a remedy for these difficulties. The potential of MSPC as a tool for monitoring wastewater treatment processes is shown. Moreover, standard MSPC techniques are extended and adapted to suit the requirement of monitoring of wastewater treatment operation.

Principal component analysis (PCA) is one of the MSPC methods used. PCA accounts for collective effects, as it allows for simultaneous analysis of all included variables. It also reduces the dimensionality of the data and compress it into information. PCA provides different ways to visualise the process in an interpretable and intuitive manner, helping the user to extract relevant information and make sensible decisions. A PCA model is identified using data from normal or desired process operation, and then used to detect deviations from this behaviour.

However, due to changing conditions, for instance, diurnal variations, seasonal changes and long term trends, the monitoring model must be updated. This can be achieved by making the PCA model adaptive. Several levels of adaptation may be used. It is shown that adaptive scaling parameters are an option when the relationships between the variables do not change. This approach has advantages since it allows intuitive graphical representations, such as score plots. When the relationships between the variables change, the covariance structure of the model must also change. Adaptive PCA (i.e. adaptive covariance structure) together with updated scaling parameters, provides us with a powerful tool for monitoring non-stationary processes in faster time scales.

Due to the multiscale nature of events and disturbances, a multiscale approach to online monitoring of wastewater treatment measurement data is proposed. Decomposition of data into separate time scales is combined with principal component analysis to extract significant features in different time scales

and to reduce data dimensionality. The advantages of such an approach are an increased sensitivity to small but significant changes and a way to approach the problem of monitoring of data from varying conditions. The scales can be recombined to represent physically interpretable scales. By doing this, two things are achieved. First, the number of scales that has to be monitored is smaller and second, the scales are chosen to match dominant time scales of the process, resulting in a more intuitive interpretation. A more sophisticated way to simplify the interpretation is also presented in the multiscale principal component analysis (MSPCA) methodology, which involves feature extraction from data on each scale and then recombination using a uniscale PCA. An extension of MSPCA to include adaptation of the scale models is proposed in the adaptive MSPCA (AdMSPCA). The results show that the AdMSPCA sometimes yields a faster response to slower disturbances, whereas the results are similar for cases involving faster changes.

Process control

When a disturbance or process deviation is observed, the task for the operators and process engineers is to make the process return to the normal operational state. However, although information on the disturbance is available through the monitoring system, it is not obvious how to adjust the process so that a desired result is obtained. This becomes especially obvious when the number of manipulated variables is high. Consequently, a systematic approach to adjust the process so that the requirements imposed upon it are fulfilled is desirable. This is often referred to as supervisory control and involves coordination of local controllers, invoking new control handles and shifting the control objectives. Supervisory control is here divided into a few subproblems: extreme event control, disturbance rejection and product design. The control objective in extreme event control is generally to force the process back to its normal state. In disturbance rejection or attenuation, the goal is to minimise the effects of disturbances that must be considered as normal. Product design implies determining how the process should be operated to achieve a certain output quality, for instance to achieve certain quality variable setpoints. In this work, a framework for integrating MSPC and control is proposed. By two different approaches, all three subproblems are addressed.

In the first approach, it is shown that by integrating PCA, clustering and setpoint determination, automatic supervisory control of wastewater treatment processes is achieved. The ability of PCA to represent the underlying mechanisms in a few components is combined with clustering to determine the current operational process state. The information on the operational state is used to derive appropriate setpoints for local controllers. Both static and dynamic setpoint determination models are used. The most basic model consists of a look-up table, which is an intuitive alternative and yields robust results. However, the fact that different disturbances affect the system in different ways allows for a reduction of a setpoint determination model. By implementing a reduced order dynamic model in a model predictive control (MPC) framework, a flexible method for process recovery is obtained. Likewise, a steady state controller based on continuous linearisation of a reduced order model is shown capable of driving the process back to its normal operational state. This multistep approach is best suited for extreme event control since new or discrete control handles are be incorporated into the procedure.

The second approach can be seen as a multivariate feedback controller. By inverting the monitoring model (a PCA model), the controller outputs required to reach a certain point in the model space, are calculated. To compensate for model errors due to closing of the loop, process changes, local controller saturations, etc., a compensation term is added to the controller. The main objective of the supervisory controller is to control the average effluent quality to certain setpoints. However, a secondary objective of the controller is to minimise the effluent quality variation during varying influent wastewater characteristics. The results show that the controller is able to meet setpoints imposed on the effluent nitrogen concentration, both for constant and varying influent concentrations. Moreover, the variation in the effluent concentration is reduced significantly by the introduction of a feed-forward term in the controller. It is also shown that the controller compensates for controller saturation or actuator loss if the loss or saturation occurs in a PC direction covered by other actuators and that it is relatively insensitive to measurement disturbances. Due to the limitation inherent of the linear approximation, this approach is best suited for disturbance rejection and product design.

6.2 Comments on implementation

It is appropriate to discuss some aspects of implementation of the methods described in this thesis. There are many factors to consider, and the best solution for disturbance detection and isolation may vary considerably from plant to plant.

Data screening

An important issue for both monitoring and control is the quality of the data. Low quality of data limits its use considerably. Therefore, the measurement system, including sensors, devices and computers must be properly and continuously maintained and checked. However, measurement related disturbances will always occur, especially in an industrial environment. Digital filtering is a straightforward, and yet, flexible way to improve the data quality. Median-based filters have proven to be effective. The frequently occurring step changes and discontinuities in data are preserved (which is not the case when using linear filters), while noise is reduced. The main drawback is the unavoidable time delay, which may cause problems in some applications where a fast response is prioritised.

Process monitoring

There is an intricate balance between complexity and performance when industrial processes are to be monitored. When everything is functioning properly, the level of complexity is of less importance. However, when this is not the case, simplicity is a desired feature. The methods used should be insensitive to some process changes and at the same time they should detect others. A rule of thumb is to keep everything as simple as possible. This rule would, in many situations, disqualify parts of what have been discussed in this work. However, it is the author's opinion that one extension to standard MSPC cannot be neglected in wastewater treatment operation: the ability to adapt to new operational conditions. The conditions vary considerably, and the only alternative would be to constantly identify new models. Making a MSPC model adaptive is a relatively straightforward task and involves only a few extra parameters in addition to those of static monitoring. The advantages do in this case balance the increased complexity. Moreover, if used in parallel with a less sensitive, static model, the operator is provided with tools for both detection (the adapt-

ive model) and process analysis (the static model). Whether this is achieved by use of time-scale decomposed monitoring (where the higher scales are adaptive in terms of mean values and the lowest scale constitute the static model) or by use of a recursive model for adaptation and a conventional model for static monitoring, is dependent on the situation.

Process control

An important feature of the multistep approach, based on the combination of a detection/classification unit and a setpoint determination unit, is that it is intuitive. The reasoning follows similar paths to that of an operator; first, the current operational state is assessed and then appropriate setpoints are determined. A look-up table is simple and the setpoints may be chosen conservatively so that a safety margin is retained. Further, a look-up table can easily be verified off-line and function as an 'expert system'. Thus, when a disturbance is observed, the operator uses the system as an advisor and compare its advise to that of him/herself. When there is sufficient confidence in the system, the supervisory controller can be applied online. Model based setpoint determination is probably more difficult, although not impossible, to use as an advisor, since the controller setpoints are updated more frequently.

In a real application, it is important that when there is a failure on the supervisory level, the local controllers are provided with suitable setpoints, for instance the ones used for normal operational control. Consequently, surveillance of the supervisory controller is imperative. This surveillance must be carried out in both the hardware and software domains. A possible integration of the methods discussed in this work is outlined in Figure 6.1.

6.3 Topics for future research

There is a clear trend towards an increasing number of sensors and signals in the operation of industrial processes. This leads to large amounts of data, and often, redundant data. Redundant data call for methods to extract relevant information, even in areas that we today consider univariate. Thus, there are many interesting topics within the areas of multivariate monitoring and control. A few of them are mentioned below.

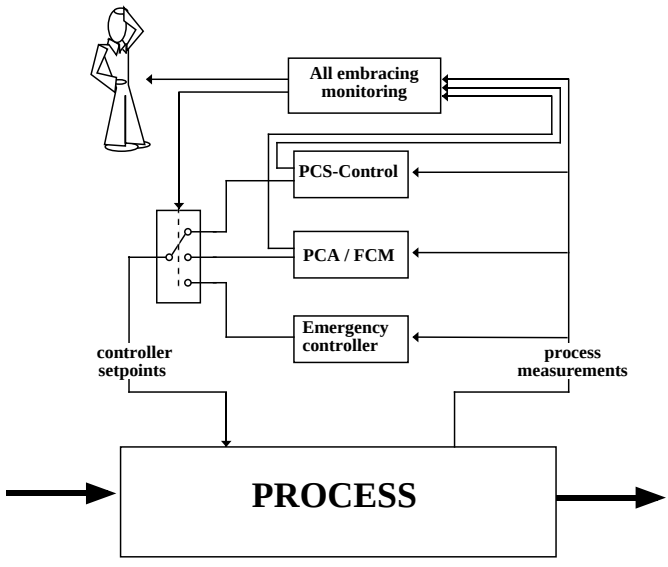

Figure 6.1: Possible integration of monitoring, extreme event control and normal operational state control. Information from both supervisory control systems are sent to the all embracing monitoring system, which determines which control structure should be used. An emergency controller is used when the other controllers are not applicable.

Process monitoring

An interesting area of research is the application of monitoring and control of large, complex processes with a high number of different subprocesses. The information paths in such systems need to be structured. Multiblock or hierarchical methods (briefly described in Chapter 3) may then provide an alternative. For wastewater treatment systems, a number of subunits may be identified, e.g. biological stage, precipitation stage, sludge handling and sewer network (or parts of it). An MSPC model is used for monitoring each subunit. However, some of the information is sent to a higher level (superlevel). On this level, another MSPC model is used to monitor the coordination between the different subunits. Such a system would not be significantly more complex than a monitoring system that includes all measurements in one MSPC model, but it would certainly produce more easily interpretable results.

Image analysis by means of multivariate analysis may be a 'hot' topic in the future. In a presentation at the 7th Scandinavian Symposium on Chemometrics (2001), MacGregor showed how he and his group used PCA to monitor a combustion flame. For wastewater applications, automatic sludge characterisation using wavelets and PCA may prove successful in the future. The data matrix is constituted by the pixels of a digital image.

Process control

Preliminary studies for extending the principal component space (PCS) controller discussed in Paper G to include the ideas of multiple time scales seem promising. Measurements are decomposed into a number of time scales (preferably corresponding to the dominant time scales of the process) and models are identified for each scale (the same procedure as for multiscale monitoring). Using the fact that the scales add up to the original data, the control signals from each scale are added. Each scale model is 'optimised' for a certain frequency band and the complete controller, including all scales, may prove better than a 'uniscale' controller. Further, the dynamic properties of such a controller ought to be better than a controller based on a uniscale model.

One step further from supervisory control is plant-wide control. The same ideas as for hierarchical monitoring ought to be applicable to plant-wide control. The superlevel model provides target values for the supervisory controllers, which in turn produce local controller setpoints. However, it is probable that nonlinear techniques will be required, since the relations between the subunits may be far from linear.

Process diagnosis

Although the area of process diagnosis is outside the scope of this work[1], it is closely linked to process monitoring and detection. An approach to diagnosis analysis, which has proven successful, is graph-based diagnosis. Graph-based diagnosis implies that the causal relations are described by nodes and connections in networks (Larsson; 1994). By describing the functionality of a wastewater treatment plant in such a way, a diagnostic tool could provide valuable help in the cause-effect analysis of a disturbance.

[1] Diagnosis has been discussed by the author in Rosen (1998b) and Rosen (1998a).

Data transfer systems

In information transfer systems, multivariate statistics may be useful as a data compression and coding method. Assume that instead of transferring the actual data, the scores from a PCA model are transferred. When the receiver wants to explore the data, the PCA model is used to inflate the data. Thus, the amount of data that need to be transferred is reduced at the same time as it is made useless unless the receiver has access to the original model.

Chapter 7

Populärvetenskaplig sammanfattning

En kemometrisk metodik för processövervakning och -styrning, tillämpningar på avloppsvattenrening

För att styra, övervaka och utvärdera processer inom processindustrin, samlas stora mängder data från mätgivare i olika delar av processen. Den tekniska utvecklingen inom mätteknik- och datorområdet har gjort det möjligt att mäta många olika storheter. Inom processindustrin kan antalet vara mycket stort; hundratals och ibland tusentals värden loggas kontinuerligt med intervall avpassade för den enskilda processen.

Att analysera och i viss mån förutsäga skeenden i dagliga driften av avloppsreningsverket vilar i stor utsträckning på operatörerna och driftspersonalen vid avloppsreningsverken. Informationen från driften måste vara pålitlig, lättillgänglig och uppdaterad för att analyser skall kunna utföras och korrekta beslut tas. Idag finns på många verk runtom i landet en omfattande mängd mätningar som görs i realtid. På ett avloppsreningsverk kan dessa uppgå till mer än hundra signaler som måste bearbetas och analyseras innan informationen som de bär kan göras tillgänglig för användaren. Vidare måste informationen presenteras på ett lättbegripligt sätt, vilket är speciellt viktigt om förestående störningar i systemet skall kunna undvikas.

För att utvinna användbar information ur stora mängder data krävs en systematisk hantering av data. Att undersöka varje signal eller variabel individuellt är tidskrävande. Detta leder ofta till att endast ett antal 'nyckelvariabler' används. Detta är olyckligt av flera skäl: dels går information förlorad eftersom

endast ett fåtal variabler övervakas, dels förbises sammansatt information som beskrivs av samspelet mellan flera variabler. För att utnyttja datamängden maximalt krävs alltså metoder som kan hantera och analysera ett stort antal variabler samtidigt. Vidare måste metoderna kunna reducera datamängden till en gripbar mängd utan att relevant information går förlorad för att sedan presentera informationen på ett intuitivt och lättbegripligt sätt.

Multivariat statistik utgör en grupp av metoder för hantering och analys av stora mängder data. Metoderna bygger på empirisk modellering, d.v.s. historiska data används för att identifiera en modell som beskriver relationerna mellan olika variabler i en datamängd. I multivariat statistik reduceras stora datamängder väsentligt; man kan säga att dimensionaliteten på problemet reduceras. Detta kan åstadkommas genom att man utnyttjar redundansen i datamängden för att producera att antal 'pseudovariabler' som bär information från alla variabler. Dessa pseudovariabler kan sedan övervakas antingen var och en för sig, eller i de fall då de fortfarande är många, i form av gemensamma mått. Det är viktigt att man inte ser denna teknik som en 'svart låda'. Det är fullt möjligt att använda analysen 'baklänges' så att när en avvikelse i en pseudovariabel har konstaterats är det möjligt att direkt isolera den eller de verkliga variabler som avvikit från sitt normala uppträdande.

I denna avhandling presenteras ett systematisk tillvägagångssätt att övervaka processen med hjälp av idéer från multivariat statistik. Ett antal utvidgningar av allmänna metoder diskuteras som möjliga kandidater att lösa de problem som uppstår vid driften av avloppsreningsverk. Dessa svårigheter har främst att göra med de föränderliga förhållanden som råder vid ett verk. Avloppsvattnets sammansättning skiljer sig avsevärt över dagen, veckan, månaden och året. Detta gör att driften hela tiden förändras på ett sätt som få andra processindustrier upplever. Övervakningsmetoderna för en sådan process måste klara av att anpassa sig till nya situationer, utan att det sker driftstörningar eller utlöses onödiga larm.

Adaptiv multivariat statistik har visat sig vara en bra lösning på problemet. Övervakningsalgoritmen uppdateras när processen förändras, och uppdateringshastigheten kan varieras beroende på målet för övervakningen. I avhandlingen visas att adaptiva algoritmer klarar av att anpassa sig till nya driftsförhållanden, utan att förlora kapaciteten att upptäcka avvikande beteenden i processen.

Ett annat problem med övervakning av avloppsvattenrening är att störningar och skeenden uppträder i olika tidsskalor. Med detta menas att vissa förlopp är snabba medan andra är långsamma. En process med dessa egenskaper brukar kallas för en 'styv' process. Detta gör det svårt att modellera relationerna mellan olika variabler. En lösning på detta problem föreslås. Den bygger på s.k. 'wavelets', en relativt ny teknik från matematiken, som används för tidsskaleuppdelning av mätvariblerna. Genom att dela upp variabler i olika tidsskalor, kan övervakningsmodeller för varje skala identifieras och problemet med styva processer undviks i viss mån på detta sätt.

När övervakningssystemet har upptäckt ett avvikande beteende, bör processen styras på ett sådant sätt att konsekvenserna av störningen minimeras. I denna avhandling presenteras två metoder hur informationen från övervakningssystemet direkt kan användas i styrsystemet för processen. Ett typiskt styrsystem i ett avloppsreningsverk består av ett antal lokala styrenheter. Dessa styrenheter har som uppgift att styra en (eller några få) variabler i processen. Genom att mäta och korrigera kan ett s.k. 'börvärde' (önskade värdet på variabeln) upprätthållas. För att tillhandhålla börvärden till de lokala styrenheterna finns ett överordnat styrsystem. I allmänhet utförs detta manuellt på avloppsreningsverk, d.v.s. operatörerna sätter lämpliga börvärden för det aktuella processtillståndet. I avhandlingen diskuteras hur detta överordande system kan automatiseras. Genom att återkoppla information från övervakningen, kan styrsystemet korrigera för störningen utan att processoperatören behöver ingripa. Detta är viktigt då avloppsreningsverket i allmänhet är obemannat merparten av tiden eftersom driften pågår hela dygnet.

Den första metoden kan beskrivas som en flerstegsmetod. Först används multivariat statistik för att beskriva det nuvarande processtillståndet. Information om vilken typ av processtillstånd som råder skickas till styrsystemet, som i sin tur reagerar på tillståndet. Denna metod lämpar sig väl för styrning under extrema processtillstånd, då processen kan sägas vara långt från sitt normala tillstånd. För styrning under normala förhållanden beskrivs en metod för hur övervakningsmodellen och den överordade styrmodellen integreras fullkomligt. Genom att använda och styra pseudovariablerna från övervakningssystemet kan de lokala styrenheterna koordineras så att ett överordnat styrmål uppnås. Genom simuleringsstudier visas att båda metoderna är potentiellt mycket intressanta för styrning av avloppsreningsverk.

Multivariata metoder kommer att spela en viktig roll inom processövervakning och -styrning i framtiden. Utvecklingen inom området sker snabbt, och

inom vissa industriella grenar har dessa metoder visat sig vara mycket kompetenta. Det är författarens åsikt att vi hittills bara sett början på en utveckling som kommer möjliggöra nya tekniker och processer i framtiden.

Part II

Included Papers

Paper A

Disturbance detection in wastewater treatment systems

C. Rosen and G. Olsson

Wat. Sci. Tech. **37**(12): 197-205, 1998

Abstract: *The development in sensor technology has made many wastewater treatment systems data rich but not necessarily information rich. To extract the adequate information from several sensors is not trivial, and it is not sufficient to consider only the time series. Different tools for detecting unusual online measurement data and deviating process behaviour are discussed. In this paper various dimension reduction as well as advanced filtering methods are considered in order to extract adequate information for fault detection and diagnosis. Both the operator and the process engineer can take advantage of such methods for proper monitoring of the plant, in particular extreme events and their causes.*

Keywords: Data analysis; detection; diagnosis; monitoring; multivariate analysis; principal component analysis (PCA); projection to latent structures (PLS).

Introduction

As the number of measured variables in a wastewater treatment plant is increasing and the need and possibility to control the process is becoming greater, the monitoring and diagnosis of the measurements are gaining in importance to obtain knowledge of the process performance and a higher product quality.

The environment for the measurement equipment is often very hostile and consequently the measurements are often defective, such as having low signal-to-noise ratio, or missing values and outliers. This implies that, before any analysis can be performed, the measurements must be pre-processed, for example filtered. The basic information in a measurement is the amplitude, but the measurement signal often contains a lot more information. This information, e.g. rate of change, trends and variability, can be used to gain additional knowledge of the process and measurement equipment performance.

The methods for monitoring and detection used today are normally based on time series charts, where the operator can view the different variables as historical trends. It is hard to keep track of more than a few variables and when the number of monitored variables are increasing it is difficult to draw conclusions. To be able to monitor the process behaviour effectively, an extraction of important information must be performed from the large number of measured variables. The information must be presented in an understandable and interpretable way. Powerful methods are available for the reduction of the high dimensionality of the information. Methods for monitoring and detection based on dimension reduction methods such as Principal Component Analysis (PCA) and Projection to Latent Structures or Partial Least Squares (PLS) have been proposed by, for example, MacGregor et al. (1994), to deal with situations with many, collinear and sometimes redundant, variables. In applications for wastewater treatment, Krofta et al. (1995), have applied the analysis techniques for dissolved air flotation.

In this paper we will discuss the element of detection and show some examples on different detection methods. We have decided to leave the mathematics out and wish to concentrate on the basic ideas and principles. For a more thorough treatment of the methods the interested reader is referred to the specialised literature and the proposed references.

Single variable analysis

Before any analysis of the measurements can be carried out data screening is crucial. Corrupted measurements must be found and dealt with, so that false conclusions based on the measurements are avoided. Almost every measurement series is affected by:

- missing values: they can be dealt with in several ways. Extrapolation in online situations or interpolations in off-line situations can be done, if the missing values are few and not succeeding each other. If the there is a extended period of missing values, the information is lost and the measurements must be disregarded (Bergh; 1996).

- noise: digital filters can be applied. Any filtering will cause some information loss, but digital filters allow a smart compromise between signal information and noise corruption. More reading on filtering can be found in Åström and Wittenmark (1997) or Olsson and Piani (1992).

- outliers: this is a delicate problem. Depending on the measurement equipment different conclusions can be drawn. Algorithms for detection of outliers based on the statistical properties of the measurements can be found in the literature. Detection of outliers can also be handled by redundant sensors or digital filtering (Åström and Wittenmark; 1997). However, one must be careful when dealing with outliers and a highly unexpected value may sometimes be true and significant (Bergh; 1996).

The preliminary data screening aims at finding adequate signals for further analysis. Now, each individual signal can be analysed with respect to a number of characteristic features:

- amplitude, the basic information in the measurement. Usually, a normal range with high and low limits is defined to be able to make qualitative comparisons.

- mean, the deviation from the mean can be used to relate the current amplitude to the normal value.

No.	Variable	No.	Variable
1.	influent temperature	11.	influent flow rate
2-3.	sludge concentration	12.	pH after biol. treatment
4-7.	air valve position	13.	dispersion flow rate (DAF)
8.	influent conductivity	14.	sludge level (DAF)
9.	influent ammonia	15.	sludge concentration (DAF)
10.	influent pH		

Table A.1: Available online measurements at the Rustorp Treatment Plant.

- deviation from normal situation, in some cases with variables varying periodically, e.g. flow rate, it can be informative to investigate the deviation from the normal variation.

- rate of change, gives information on the dynamic features of the measurement.

- trends, are useful information on the long term variations.

- variability, also reveals dynamic properties of the signal. Poor sensor performance can be detected by examining the variance and the frequency content of the signal.

The primary data analysis discussed so far gives the first pieces of information about the process operational state. The primary signals can be combined in various ways to calculate or estimate other variables. It is outside the scope of this paper to further discuss model based estimation. This has been done elsewhere (Olsson; 1989).

The data used in this paper is collected at the Rustorp wastewater treatment plant, Ronneby, Sweden. Thus is no simulated data used for the examples. The Rustorp plant is a municipal nutrient removal activated sludge plant serving about 25000 p.e. The process is operated with predenitrification, and dissolved air flotation (DAF) as a final step. The acquired data are sampled every 5 minutes from online instrumentation, as shown in Table A.1.

In addition to the process variable measurements in Table A.1, the plant outlet quality variables pH, phosphate and turbidity are measured.

Example 1—poor sensor performance

In Figure A.1(upper) the influent ammonia is plotted for a period of about 3.5 days. The first 500 samples are displaying a normal behaviour in terms of variation and noise characteristics. At about sample 3070 there is an abrupt change in the signal characteristics. The noise level increases significantly, which might indicate poor sensor performance. To more systematically detect this change in signal characteristics, an analysis of the variance of the high-passed filtered signal is carried out. The filter retains only the high frequency content of the signal. To examine the variance of the raw signal would not be enough, since there are variations in the signal, apart from the noise. Figure A.1 (lower) shows how the variance (calculated from a moving window of 72 samples) of the high-pass filter output suddenly increases after about sample time 3070.

Detection of operational states

In many cases the investigation of individual signals is insufficient and can not reveal the true state of the process. Variables influence each other and one must often look at several variables simultaneously. Multiple process data can be analysed in many different ways and Davis et al. (1996) suggest that the analysis methods can be divided into three distinct components:

- numeric-numeric, including time and frequency domain analysis;
- numeric-symbolic, including dimension reduction and distribution functions based methods;
- symbolic-symbolic, including knowledge based systems.

Here we want to detect a measurement pattern, which may be regarded as an operational state, determined by the measurements and observations. Thus, as the process conditions change the plant could be said to be in different operational states. We define an operational state as a multidimensional region, where all the process states and parameters are located. Thus, if some of the states or parameters drift away from this region, the process is said to move into another operational state. A general operational state is here defined as a region that

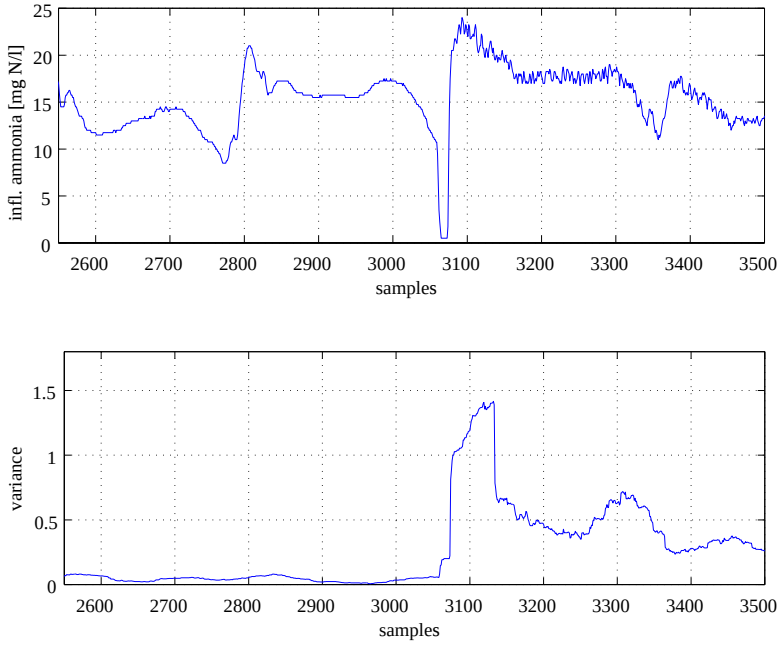

Figure A.1: Detection of poor sensor performance. The upper part of the figure shows the influent ammonia concentration. The lower part shows the variance of the high-pass filtered signal (sample time = 5 minutes).

includes influent parameters and process state variables, without any specific output quality variable in mind. It can also be defined as a specific operational state, including adequate effluent quality variables.

There are (at least) two fundamental ways of describing a specific plant:

- Physically based definition, using mass balances of substrates and organisms, including reactions. Combining the a priori models with observations it is possible to derive some of the unknown parameters (grey-box models).
- Input-output based definition, based on the correlation or other relations between measured or observed variables. The parameters of the model do not necessarily have any physical interpretation (black-box models).

The parameters obtained from any of the above model description can be used to determine the operational state of the process. In any of the models various uncertainties can be represented by stochastic processes. There are many references in the literature to grey-box modelling and we will not discuss this further, even though grey-box modelling might give us a considerable contribution to the understanding of the processes (Carstensen; 1994; Lindberg; 1997). In this paper, we consider input-output models to classify the cause-effect relationships, in particular some dimension-reduction methods. Some examples are shown to demonstrate the methods for detection purposes. There are also other techniques available, which are not further explored here, such as artificial neural networks, clustering and classification algorithms.

Dimension-reduction based methods

Depending on the aim for the detection model, it is important to group the available variables (or indirect variables calculated or estimated from other variables) into independent (X) and dependent (Y) variables. When the whole plant is considered, the process variables, including the influent characteristics variables, typically are defined as the independent variables, while the effluent quality variables are defined as dependent. However, when the modelled system does not comprise the whole plant, it has to be derived from the system boundaries which variables are the dependent and independent ones respectively. This implies that a variable can belong to the X-block in one model and to the Y-block in another one. Normally causality will help us to decide, but when recycling loops or feedback are present in the system, the causality is not trivial. For example, the DO level at the outlet of the aerated basin can be a independent variable (X-variable) in a model monitoring the anoxic zone in a predenitrification plant. Scaling must sometimes be done, to be able to compare changes in different variables to each other. When the units and amplitudes differ, it is convenient to normalise to zero mean and unit variance.

PCA

Principal Component Analysis is a way to investigate large data sets with many process variables. However, many of these variables are often highly correlated,

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	2.53e+00	50.63	50.63
2	1.24e+00	24.82	75.45
3	6.98e-01	13.95	89.41
4	3.81e-01	7.62	97.03
5	1.49e-01	2.97	100.00

Table A.2: Percent variance captured by PCA model

since most variables only reflect a few underlying mechanisms that drive the process in different ways (Kourti and MacGregor; 1994). The true dimension of the process space is often a lot smaller than the dimension of the data matrix space. The aim for the PCA is to project the high dimensional space into a more visual low dimensional space and by doing this finding the key variables. This is achieved by transforming the measurements of the original coordinate system in such a way that a maximum of the variance is explained by the new coordinate system. Thus, there will be a number of new latent variables, called principal components, which describe most of the variance in the process in a space of fewer dimensions than the original space. A model can be built from a set of training samples, and then used to detect deviations from the normal model space.

Example 2—monitoring the general operational state in two dimensions

In this example we will show the applicability of PCA in monitoring the influent wastewater characteristics. Available online measurements of the influent wastewater are; temperature (1), conductivity (2), ammonia (3), pH (4) and flow rate (5), which are variables of the so-called X-block. The detection model is built and trained from 6000 samples ($\approx$ 21 days) of normal operating conditions. Table A.2 shows that already two principal components describe 75 percent of the variance in the X-block.

A new period of measurements are transformed by the model into the first two principal components (PC #1 and PC #2). The elliptic boundaries in Figure

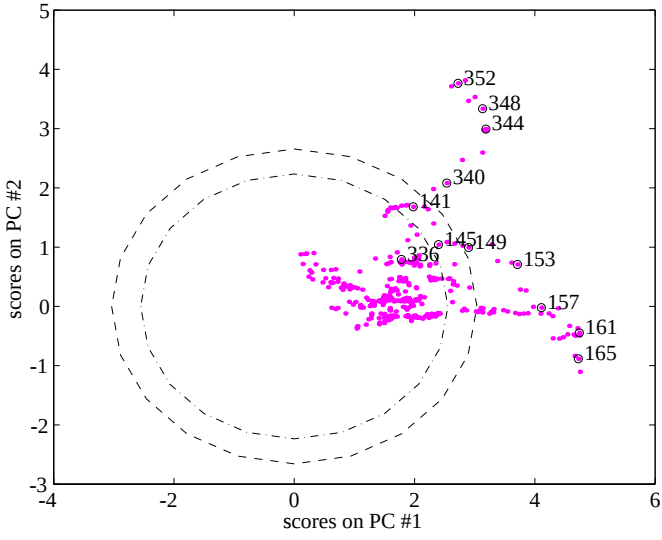

Figure A.2: Monitoring the influent characteristics with PCA (sample time = 5 minutes).

A.2 correspond to the 95 and 99 percentage significance level of the original training data.

It can be seen that the process is not exactly situated in the centre, but it is mostly inside the boundaries. However, two major deviations from the normal situation can be observed. The first deviating situation starts at about sample 141 and propagates outside the normal operating region to its maximum deviation at sample 165. The other situation is noted in samples 336 to 352. What has caused these deviations? As discussed by MacGregor et al. (1994), the variable contributions to the changes in the PC #1 and PC #2 directions, can be shown as in Figure A.4. The first two bar charts in Figure A.4 show the event between samples 141 and 165, and the last two bar charts show the event at samples 336 to 352. Bar chart 1 and 2 show that it is primarily the change in variable #3, i.e. ammonia, that has caused the plot to exceed the boundaries in the PC #1 direction while the change in variables #1-3, i.e. temperature, conductivity and ammonia, has caused the movement in the PC #2 direction. This is confirmed by the plots in Figure A.4 (right). Bar chart number 3 and

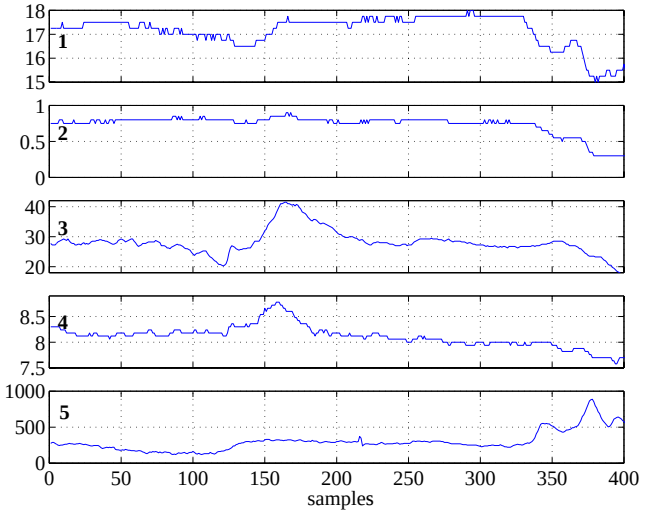

Figure A.3: Monitoring the influent characteristics with traditional time series charts (sample time = 5 minutes).

4 show that the second event is caused by variable #5, i.e. flow rate, along the PC #1 axis and variable #1,2 and 5 along the PC #2 axis. (Note that the sign of the bars in the bar charts can not directly be used to decide in what direction changes occur, but it can easily be derived.)

Observations of the time series in Figure A.3 do not easily reveal what are the most significant changes of the process. Thus, the PCA plot can help the operator to focus on the right causes of an event.

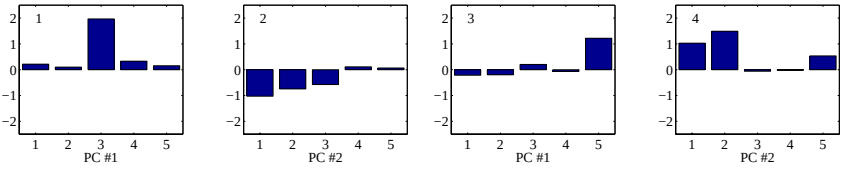

Figure A.4: The contribution of every X-variable to the change in PC #1 and #2 direction at the first (chart 1 and 2) and second event (chart 3 and 4).

	X-Block		Y-Block	
LV #	This LV	Total	This LV	Total
1	26.58	26.58	34.75	34.75
2	18.82	45.40	8.55	43.31

Table A.3: Percent variance captured by the model.

PLS

Partial Least Squares or Projection to Latent Structures is also, as the PCA, a dimension reduction method, but in PLS the latent variables (LV) of the X-space are calculated to maximise the correlation between the input matrix X and an output matrix Y, containing the product quality variables, such as effluent turbidity or ammonia. In this way the detection effort can be applied on the variables most influential on one or several specific quality variables.

Example 3—monitoring a specific operational state considering the output turbidity

In Example 2 the PCA method only tries to explain the variance in the X-block. Using PLS the X-block can be linked to the Y-block, containing the quality variables. In this example the X-block contains all the variables in Table A.1 and the Y-block contains the effluent turbidity. The model is built from 2500 samples ($\approx$ 9 days) of normal operating conditions. Table A.3 shows the variance explained by the first two latent variables (LV).

The new measurements are transformed into the two-dimensional space defined by the first two latent variables.

Figure A.5 (left) shows that the process variables are well inside the boundaries until a disturbance occurs at about sample 415. In a few samples, the process has drifted far outside the boundaries until a maximum deviation is reached at sample 439. Using the same method as in Example 2, three phenomena are detected. They appear in the bar-chart of Figure A.6. Four air-valve measurements (#4-7) are apparent contributors to the observed deviation from the normal operating range. There is an obvious change in the oxygen demand, since the dissolved oxygen is controlled. Therefore, the air valve positions are

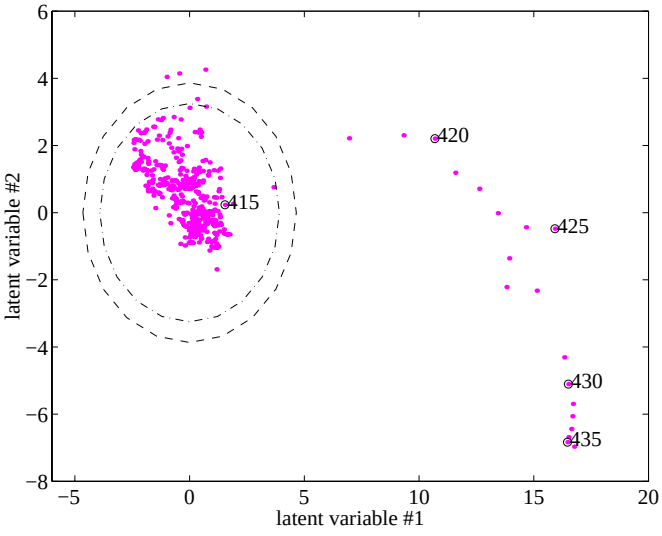

Figure A.5: Monitoring process with PLS (sample time = 5 minutes).

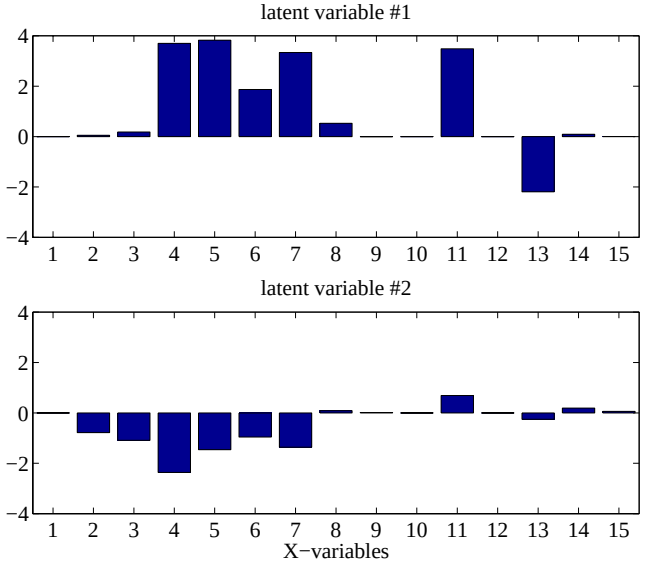

Figure A.6: Diagnosis of the event at sample 415-435 (right) (sample time = 5 minutes).

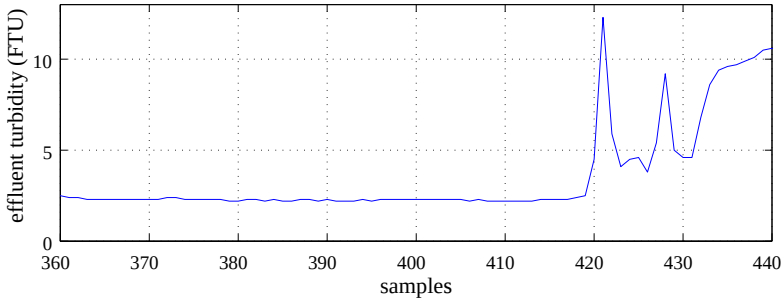

Figure A.7: The measured effluent turbidity (sample time = 5 minutes).

an indirect measurement of the activity in the aerated basins. Secondly, there is a significant change in both the influent flow rate and in the dispersion flow rate to the DAF. Each one of these changes can contribute to a change of the effluent turbidity by (1) an overload situation reflected in an increased aeration, (2) hydraulic overload (washout) or (3) an excess of small air bubbles at the turbidity meter.

It should be noted that the effluent turbidity is not explicitly used to generate the data of Figure A.5, instead it is only used to build the model. Therefore, it is highly interesting to verify its behaviour by direct measurements. This is shown in Figure A.7. The plot in Figure A.5 is exceeding the boundaries between samples 417 and 418. The increase of the effluent turbidity at the first peak starts at sample 419 or 420. This delay of two samples (10 minutes) corresponds well to the expected delay if there is a case of wash out (cause 1). The turbidity beyond the first peak may most probably be explained by a combination of causes (1) and (2).

Nonlinearities

There are more ways to further improve the performance of the dimension reduction methods. When the modelled system is believed to be nonlinear, which is often the case, nonlinear PLS can be used. It can be achieved by expressing the inner relation between the X-block and the Y-block in a nonlinear

manner (Wise and Gallagher; 1996a). A second way to deal with nonlinearities is to take the variables in the X-block believed to be nonlinear and raise them to a desired power, e.g. $x^{1/3}$ or x^2.

Dynamics

The dimension reduction based methods are static, and the dynamics in the system are not represented. For example, the time lag between the input block X and the output block Y is not addressed. One way of dealing with this, if there is just one quality variable in the Y-block, is to investigate the cross-covariance function between every input variable and the output variable and calculate the suitable lag (Åström and Wittenmark; 1997; Wise and Gallagher; 1996a). A second way is to use an a priori model for the time lag of every relation, for example, depending on the retention time. By doing this the time lag between each process variable and the quality variable will change dynamically as the flow rate changes and we will have a quasi-dynamic representation of the flow dynamics.

Example 4 - predicting the pH at the outlet of the aerated basin with PLS

The time lag between the inputs and the output variable has been computed using the cross-covariance techniques indicated above. PLS is now used as a prediction model for detection purposes, for instance as a tool for sensor failure detection. When the residual, i.e. the difference between the predicted and measured value, exceeds a certain threshold, poor sensor performance can be detected and countermeasures be taken.

In the earlier examples, we focused on the first two latent structures, but this is to leave information out. By cross-validation, the optimum number of latent structures can be found (Wold; 1978). The model is built from 5000 samples ($\approx$ 17 days) of normal operating conditions with nonlinear PLS.

Figure A.8 shows the residual of the prediction over a certain time period. The limits correspond to 95 percent significance level of the total prediction residual. There is a sensor failure at sample 7200, which is easily detected.

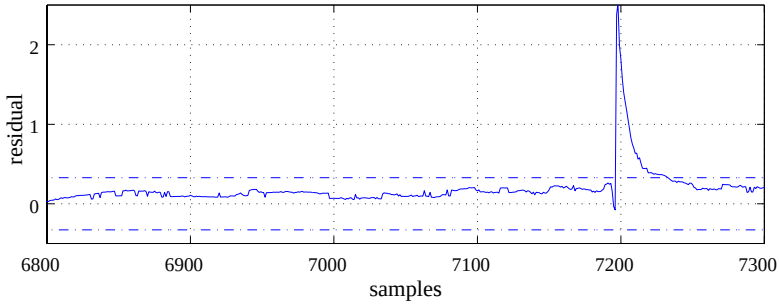

Figure A.8: The prediction residual of pH predicted and measured.(1 sample = 5 minutes).

Conclusions

With an increasing instrumentation in wastewater treatment systems, there is certainly more information available. At the same time, there is a risk for too complex information, and the need to condense data for the operator and for the process engineer will increase. In this paper, some powerful data reduction methods are described. We have shown the applicability of single variable analysis and dimension reduction based methods for the detection of process deviations. This paper does not claim to be exhaustive and there are other powerful analysis methods available, such as Multiple Linear Regression (MLR), Principal Component Regression (PCR) and Continuum Regression. From full-scale plant data it has been shown, that advanced detection methods can be successfully applied in wastewater treatment systems. The methods described in this paper can be combined with classical statistical process control (SPC). In SPC, warning and action limits are derived from the statistical properties of the measurement signals and are used to advice the operator when to take action. Commonly used control sharts, such as CUSUM, X-sharts and EWMA (Bissel; 1994; Chapman; 1998) are all applicable on the new variables gained from the methods discussed above.

Acknowledgements

This work was financially supported by the Swedish Water and Wastewater Association (VAV) under the VA-forsk research program. Appreciation is also extended to the staff of the Rustorp treatment plant, Ronneby, for all the help in collecting the online measurements. Dr Ulf Jeppsson at IEA is always a valuable and appreciated source of knowledge and critical thoughts.

Paper A

Addendum

This paper presents some of the author's first results of univariate and multivariate detection in wastewater treatment operation. At the time when the results were first presented (at the ICA conference in Brighton, 1997), it was one of the first attempts to approach the complexity of wastewater treatment monitoring from a multivariate perspective. Later it was shown (see Papers B - D) that the somewhat naive approach was not sufficient to be practically applicable. A number of shortcomings could be listed, but most of them will be discussed and solved in later papers by the author.

Two important shortcomings should be mentioned. The use of PLS for monitoring of specific operational states turned out to be rather problematic. It is clear that the input-output relationships are complex, and generally not possible to model with linear models. In Rosen (1998a), PLS modelling for prediction of the effluent nitrate concentration proved successful using simulation models. However, the same results have not been achieved using real data. The reason for this is mainly twofold. First, the relations between input and output are often nonlinear and in some cases quite severely so (e.g. effluent turbidity). Second, the measurement quality is too low for reliable predictions. It should be noted that no exhaustive effort was made to overcome these difficulties, and with today's improved sensor technology, more successful results may be obtainable.

The second shortcoming of the paper is its failure to stress the ability of multivariate statistics to improve data quality. Robust noise reduction and missing data replacement can be achieved with, e.g. PCA (Nelson et al.; 1996; Grung

and Manne; 1998; Walczak and Massart; 2001a,b). Thus, multivariate statistics can be incorporated in the data pre-treatment as well as in the monitoring stages.

Although the results are basic, the paper shows that especially multivariate techniques display a promising potential for detection and monitoring of wastewater treatment operation.

Paper B

Monitoring wastewater treatment operation. Part I: Multivariate monitoring

C. Rosen and J.A. Lennox

Published together with Paper C in a combined form in
Wat. Res. **35**(14): 3402-3410, 2001.

Abstract: *In this work, principal component analysis (PCA) for wastewater treatment process monitoring is discussed. The basic approach for PCA monitoring is presented together with more advanced approaches to handle changing process conditions. Adaptive PCA in terms of updating of the scale parameters as well as the covariance structure is discussed. Problems encountered using these techniques for wastewater treatment monitoring are pointed out and ways to overcome some of the difficulties are proposed. The methodology is illustrated with examples using real process data.*

Keywords: Adaptive; detection; monitoring; PCA; wastewater.

Nomenclature

α	confidence percentile or forgetting factor
λ	vector of eigenvalues
Λ	diagonal matrix of eigenvalues
Λ^{-1}	diagonal matrix of the inverse of the eigenvalues
a	number of PCs retained in a PCA model
$\mathbf{c}(k)$	contribution vector
$\mathbf{e}(k)$	error vector at time k
$\mathbf{E}$	error matrix
m	number of samples in a data matrix
M	effective memory length
$\mathbf{M}_j$	general matrix of rank 1
n	number of variables in a data matrix
$\mathbf{p}_j$	jth loading vector
$\mathbf{P}$	loading matrix
$\mathbf{P}_j$	diagonal matrix of the jth loading vector
r	number of dimensions in a data matrix
SPE	sum of squared prediction error
$\mathbf{t}_j$	jth score vector
$\mathbf{T}$	score matrix
T^2	Hotelling's T^2 statistics
$x_j(k)$	value of jth variable at time k
$\hat{x}_j$	estimated value of jth variable at time k
$\mathbf{x}(k)$	data vector at time k
$\mathbf{X}$	matrix of measurement variables

Introduction

In this paper we discuss multivariate statistical process monitoring and show its applicability to wastewater treatment monitoring. We focus on principal component analysis (PCA) as it serves, through its simplicity, as a good introduction to multivariate statistics (MVS) for process monitoring (MacGregor and Kourti; 1995; Wise and Gallagher; 1996b). PCA-based monitoring as such is not new, but we believe there is need for a thorough discussion of how it can be implemented, what the advantages and shortcomings are and what we can expect to achieve by using PCA in wastewater treatment monitoring. One of the shortcomings is that PCA, in its basic configuration, is not suited for monitoring

changing processes. This is addressed in part I of this work. In part II, PCA monitoring is extended to involve multiple time scales.

Data are collected from the process at most industrial plants. The driving forces behind the data collection are normally related to quality, safety and economic requirements imposed on the operation. The use of data and the methods to acquire data may differ significantly depending on what the purpose is for the measured entity. As new sensor techniques allow more and more entities to be measured affordably, more sensors are installed. Thus, at many industrial processes, data from hundreds or thousands of sensors may be collected. Most of the sensors operate online (in real-time), giving operators the ability to observe events and changes in the process as they occur. This is a potentially significant source of process knowledge, if the data can be transformed into information and interpreted in a correct and adequate way. To do this, a systematic approach to monitoring of the data is needed.

In most process industries, monitoring of the process and its outputs is an important part of the operation. Monitoring can be said to consist of three phases:

1. detection—recognising that there is a deviating event or that the process is not operating at its normal point;
2. isolation—finding the deviating measurement variables that have triggered the detection;
3. interpretation—finding the physical causes of the deviation and assessing its impact on the process.

The first phase is well suited for computers, as it is a monotonous and quantitative task. The second phase is normally be integrated with the first task and, thus, also be carried out by computers. However, the third phase requires process knowledge and is mainly a task for the operator relying on his experience, although attempts have been made using knowledge based systems, such as expert systems, for interpretation. In order to facilitate the third phase, the methods used in the first two phases must extract and organise the information appropriately.

The extent and sophistication of monitoring differs in various fields of application. The wastewater treatment industry cannot be considered to be among

the more diligent and systematic users of monitoring. Up to now, monitoring in wastewater treatment has mostly focused on a few key effluent entities, for which regulations are enforced by governments or other authorities. However, as more entities are regulated and the regulations become stricter, the demands on the operation of the processes increase. Minimising the use of resources (e.g. energy, chemicals and human resources) and decreasing the amount of sludge products produced, have also become important issues if wastewater treatment processes are to be adapted to the principles of sustainability. All these factors have led to an increased need for process knowledge and better control of process performance.

The methods for monitoring used today in wastewater treatment are normally based on time series charts where the operator can view the different variables as historical trends. These methods are often referred to as conventional statistical process control (SPC) (see e.g. Thompson and Koronacki (1993) and Bissel (1994)). Limits representing normal operation are defined, and as long as each variable stays within these limits, the process is said to be in control. These conventional methods have two important shortcomings. Since the effects on the process are considered using each variable in isolation, the collective effects of several variables are not accounted for. For example, several variables still within their individual normal regions may together drive the process from its normal operational state. Secondly, when the number of variables is large we run the risk of obtaining more data than we can assess and use for decisions. Interpretation becomes difficult: we are 'data rich but information poor'. Consequently, methods for handling large data sets as well as collective effects online are needed.

Within the field of chemometrics, the issue of transforming multivariate data into interpretable information has been addressed during the last three decades (Geladi; 1988). Attempts to develop methods that extract relevant information from process data have resulted in a number of different methods based on MVS. One of the simplest methods is PCA (Jackson; 1980; Wold et al.; 1987b). Many further developments of PCA now exist, including multiway PCA (Wold et al.; 1987a), multiblock PCA (Wold et al.; 1996; Westerhuis et al.; 1998) and dynamic PCA (Ku et al.; 1995). Regression techniques, such as principal component regression (PCR) and projection to latent structures (PLS) are also based on the same or similar ideas (Geladi and Kowalski; 1986; Höskuldsson;

1988). During the last decade, MVS-based methods have been applied to process monitoring and modelling, which has resulted in a multivariate approach to statistical process monitoring (Wise et al.; 1990; Kresta et al.; 1991; MacGregor and Kourti; 1995; Wise and Gallagher; 1996b). Recently, multivariate analysis, monitoring and modelling have been used in wastewater-related applications (Krofta et al.; 1995; Rosen and Olsson; 1998; Rosen; 1998a; Mujunen et al.; 1998; Teppola et al.; 1998a).

The paper is organised as follows. In the next section, we explain PCA. The following section deals with PCA as a monitoring method. In order to handle changing process conditions data, adaptive PCA is presented and different types of adaptation are discussed. Then we apply these methods to monitoring wastewater treatment data. This is followed by a discussion. Finally, the work is summarised and concluded in the last section.

Principal component analysis

PCA can be described as a method to project a highly dimensional measurements space to a space with significantly fewer dimensions. Often, for industrial process data, many variables are highly correlated, since they reflect relatively few underlying mechanisms that drive the process. In PCA, we use this correlation between the variables to find principal components (PCs) that represent the underlying mechanisms and, thus, reduce the data. The number of PCs is often much smaller than the number of original variables.

Model generation

Let $\mathbf{X}$ be an autoscaled (i.e. mean-centred and scaled to unit variance) $[m \times n]$ matrix of measurement values for n variables at m number of samples defining a variable space of r dimensions. The r-dimensional matrix $\mathbf{X}$ can be decomposed into a sum of matrices $\mathbf{M}_j$, each of which represents the variability in the jth dimension, i.e. the jth PC:

$$\mathbf{X} = \mathbf{M}_1 + \mathbf{M}_2 + \ldots + \mathbf{M}_a + \mathbf{E} \tag{B.1}$$

The matrix $\mathbf{M}_j$ can be written as the outer product of two vectors $\mathbf{t}_j$ and $\mathbf{p}_j^T$. Thus:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_a\mathbf{p}_a^T + \mathbf{E} \quad \text{(B.2)}$$

or

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad \text{(B.3)}$$

where $\mathbf{E}$ is the residual (or error) matrix and $a \leq r$. If $a = r$ then $\mathbf{E} = 0$, as all the variability directions are described. However, if $a < r$, i.e. less PCs than original variables are retained, then $\mathbf{E}$ describes the variability not described by the sum of the $\mathbf{T}\mathbf{P}^T$ matrices .

The column vectors of $\mathbf{T}$ ($\mathbf{t}_j$) are called the score vectors or scores and the column vectors of $\mathbf{P}$ ($\mathbf{p}_j$) are the loading vectors or loadings.The matrix $\mathbf{P}$ can be determined by singular value decomposition of the covariance matrix of $\mathbf{X}$:

$$cov(\mathbf{X}) = \mathbf{U}\Sigma\mathbf{V}^T \quad \text{(B.4)}$$

where Σ is the diagonal matrix of the singular values $s_1, s_2, \ldots, s_n$ in decreasing order of magnitude. However, since $cov(\mathbf{X})$ is a square matrix and $\mathbf{X}$ is autoscaled, $\mathbf{U} = \mathbf{V}$ and Equation B.4 can, thus, be written:

$$cov(\mathbf{X}) = \mathbf{P}\Lambda\mathbf{P}^T \quad \text{(B.5)}$$

where Λ is the diagonal matrix of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ in decreasing order of magnitude. $\mathbf{P}$ has some interesting properties. Since it is a unitary matrix $\mathbf{P}^T\mathbf{P} = \mathbf{P}\mathbf{P}^T = \mathbf{I}$ and $\mathbf{P}^T = \mathbf{P}^{-1}$. If tools for eigenvalue and singular value analysis are not available, an alternative way of identifying a PCA model ($\mathbf{P}$) is to use the iterative NIPALS algorithm. The algorithm is found in, for instance, Geladi and Kowalski (1986).

Scaling and components

As mentioned above, the data matrix is scaled prior to analysis. This makes it possible to compare variables with different amplitude and variability. Autoscaling is a simple way to scale the data but is not the sole option. Figure B.1 illustrates some different options for scaling and centring. A word of caution is

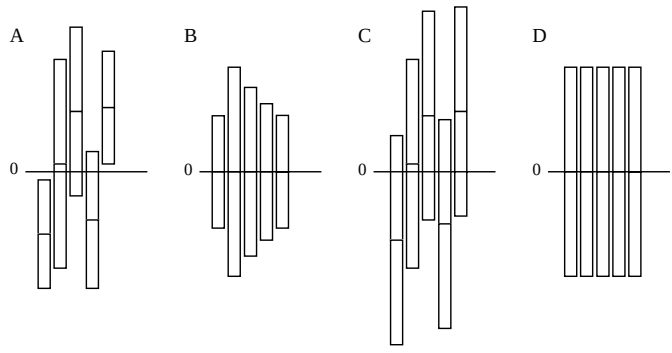

Figure B.1: Data pre-processing. The data for each variable are represented by a variance bar and its centre. Unmanipulated data (A), mean centred data (B), variance scaled data (C) and mean centred and variance scaled data, i.e. autoscaling (D).

justified when dealing with nearly constant variables with a low signal-to-noise ratio. If such variables are scaled to unit variance, the noise contribution to the variability will be high (Kresta et al.; 1991). Other methods for scaling can be used when it has been established that certain variables have more influence on the process than others. Giving higher weights to these variables may improve the monitoring performance.

The number of components retained in the PCA model is crucial to the model performance. Too few components implies that there are not enough dimensions to represent the process variability, while too many components implies that measurement and process noise will be modelled. There are a few different ways to determine the optimum number of components. The simplest is to plot the eigenvalues associated with each PC (see the scree plot in Figure B.2). The presence of a jump or a knee often indicates an appropriate number of components. Parallel analysis involves the calculation of the intersection in a scree plot between the curves of the eigenvalues from the process data and the eigenvalues from a set of uncorrelated data (Figure B.2). A simple version of this is sometimes referred to as the 'quick-and-dirty' method, which means that all eigenvalues smaller than one are excluded. Alternatively, a cross-validation procedure can be used. The number of components that yields the smallest prediction error is chosen. More information on the choice of the number of components is found in, for instance, Wold (1978) and Himes et al. (1994).

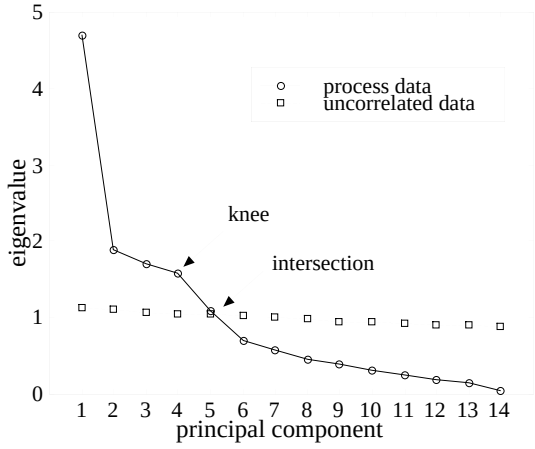

Figure B.2: The eigenvalues can be used to determine the number of components. The knee and the intersection indicate good choices. The data are the same as used for identification in the examples presented later.

Online monitoring using PCA

The most basic way of using PCA for monitoring involves identification of a model from data representing normal or desired operation. New data are then projected onto the model and the scores and/or the model residuals are then monitored as new samples are obtained.

$$\hat{\mathbf{t}}(k) \quad = \quad \mathbf{x}(k)\mathbf{P} \tag{B.6}$$

It is important to note that the new data must be scaled in the same manner as the data used for identification.

Monitoring scores

Monitoring of the scores is carried out using either conventional univariate SPC techniques or scatter plots. By plotting, for instance, the first score vector against the second, process changes can be viewed as movement of a point in the plane as new samples are added. Points that cluster generally represent similar process behaviour whilst points in different regions in the PC space gen-

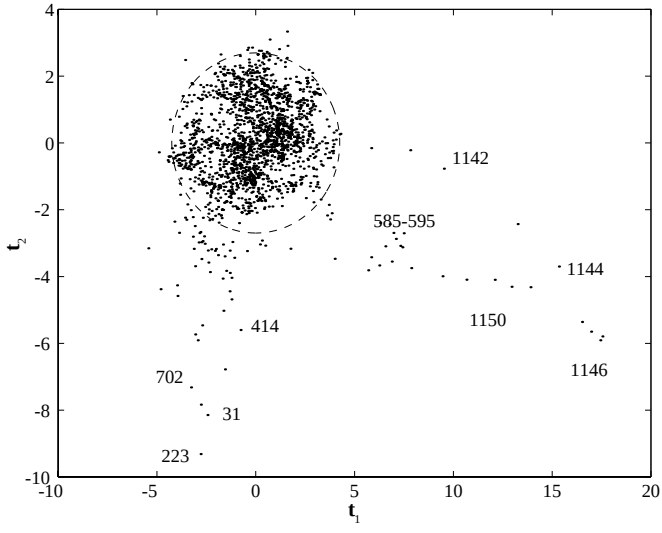

Figure B.3: The scores of the first and second principal component with approximate 99 % confidence limits. The data are the same as those used for identification in the examples presented later.

erally represent different operational states. A score plot is shown in Figure B.3 (the data are the identification data used in the examples presented later). Each point represents the values of the first and second scores at a certain point in time. Under normal operating conditions, the centre of gravity of the points should be close to the origin, due to the mean centring of data. Points far from the origin indicate a disturbance and, consequently, the operational state is no longer classified as normal. Confidence or control limits are used to discern disturbances. The ellipse in Figure B.3 represents the 99 % confidence limits of the first and second score vectors of the training data.

Model residuals

In addition to monitoring the scores, the statistical fit of the model can be monitored (Jackson and Mudholkar; 1979; Kresta et al.; 1991). Two commonly used measures of fit are sum of squared prediction error (SPE) and Hotelling's T^2. SPE and T^2 are convenient as they summarise the multivariate process

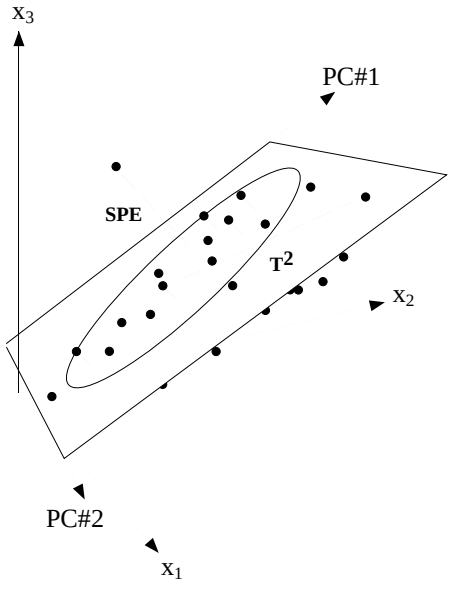

Figure B.4: The geometrical interpretation of the SPE and T^2 measures, respectively. x_1, x_2 and x_3 are the original variables and PC#1 and PC#2 are the principal components. The points represent samples or observations.

information of the variables by single values. Figure B.4 gives a geometrical interpretation of SPE and T^2. SPE is a measure of the distance from the model plane to an observation. T^2 is a measure of the distance within the model plane from an observation to the origin. Consequently, a disturbance that involves a change in the relations between the variables, will increase the SPE since the model does not cover the direction of the disturbance. A disturbance in the model plane, i.e. of the same nature as the identification data, will turn up as an increase in T^2.

SPE and T^2 are normally monitored using conventional SPC charts with in-control limits defining the normal or desired operation region. SPE is calculated as (Kresta et al.; 1991):

$$SPE(k) \quad = \quad \sum_{j=1}^{n} \left(x_j(k) - \hat{x}_j(k)\right)^2 = \mathbf{e}(k)\mathbf{e}^T(k) \qquad \text{(B.7)}$$

where n is the number of variables of $\mathbf{X}$ and $\hat{x}_j(k)$ is the model prediction. $\mathbf{e}(k)$ is the kth row of $\mathbf{E}$. Since $\hat{x}_j(k) = \mathbf{x}(k)\mathbf{P}\mathbf{p}_j^T$, Equation B.7 is rewritten as:

$$SPE(k) \quad = \quad \sum_{j=1}^{n} \left(x_j(k) - \mathbf{x}(k)\mathbf{P}\mathbf{p}_j^T \right)^2 \tag{B.8}$$

Jackson and Mudholkar (1979) showed that the confidence limit for SPE from a PCA model can be approximated as:

$$SPE_{lim} \quad = \quad \Theta_1 \left[\frac{c_\alpha \sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0 (h_0 - 1)}{\Theta_1^2} \right]^{\frac{1}{h_0}} \tag{B.9}$$

where

$$\Theta_i \quad = \quad \sum_{j=a+1}^{n} \lambda_j^i \tag{B.10}$$

for $i = 1,\ 2,\ 3$ and

$$h_0 \quad = \quad 1 - \frac{2\Theta_1 \Theta_3}{3\Theta_2^2} \tag{B.11}$$

c_α in Equation B.9 is the standard normal deviate corresponding to the upper (1-α) percentile and a in Equation B.10 is the number of PCs retained. In Figure B.5 (top) an SPE plot of the same data as used in Figure B.3 is shown.

Hotelling's T^2 statistics can be interpreted as the normalised sum of squared scores. The T^2 at time k is calculated as (Jackson and Mudholkar; 1979):

$$T^2(k) \quad = \quad \mathbf{t}(k)\Lambda^{-1}\mathbf{t}^T(k) = \mathbf{x}(k)\mathbf{P}\Lambda^{-1}\mathbf{P}^T\mathbf{x}^T(k) \tag{B.12}$$

where $\mathbf{t}(k)$ are the scores at time k and Λ^{-1} is the diagonal matrix of the inverse of the eigenvalues associated with the retained PCs (see Equation B.5). Confidence limits for SPE and T^2 can be calculated (see e.g. Jackson and Mudholkar (1979) or Wise et al. (1990)). These limits are used to determine whether the process is in control or not. In Figure B.5, SPE (top) and T^2 (bottom) plots are shown together with their confidence limits. The confidence

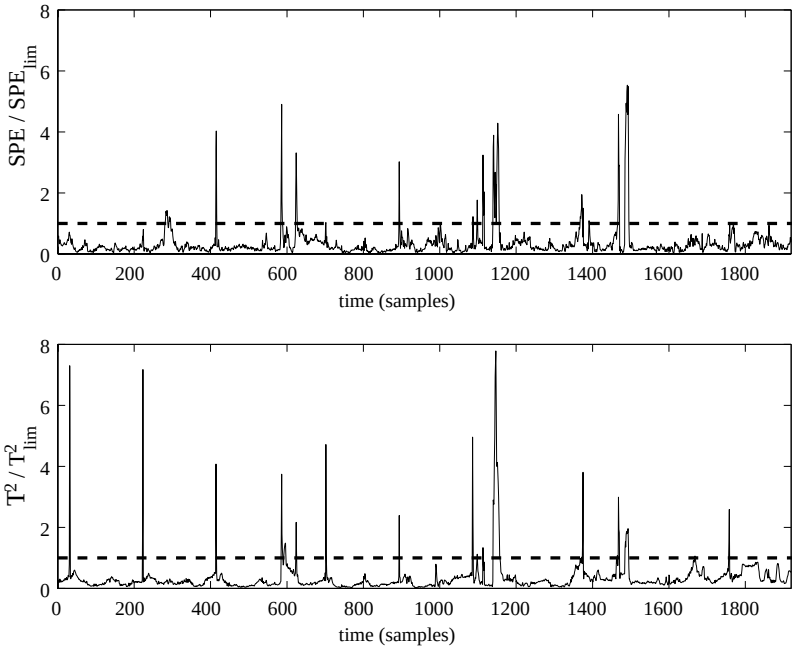

Figure B.5: The relative SPE and T^2, i.e. the ratio between the residual and its limit. The data are the same as used for identification in the examples presented later.

limits for T^2 are obtained using the F-distribution (Wise; 1991):

$$T^2_{lim} = \frac{a(m-1)}{m-a} F_{k,\, m-a,\, \alpha} \tag{B.13}$$

where m is the number of samples in the model and a is the number of PCs.

It is important to note that SPE and T^2 statistics assume that the data are normally distributed. This is usually not the case for wastewater treatment data, so it is important that confidence limits are not blindly trusted. In most cases though, the distributions of the model residual are approximately normal due to the central limit theorem. Note that the SPE and T^2 measures as such do not depend on a certain distribution, only their confidence limits. Robust limits can be determined empirically using, for instance, percentiles (Rosen; 1998a).

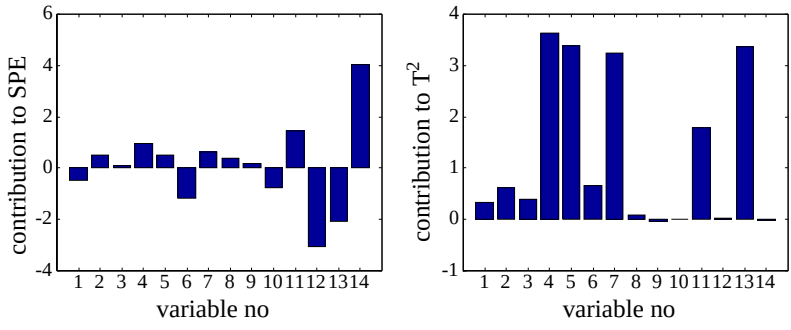

Figure B.6: The contribution of every variable to the prediction error at sample 1490 (left). The contribution to the deviation along the first PC at sample 1146 (right)

Isolation of deviating variables

There is more information to be gained from a deeper investigation of the PCA model. When a disturbance is detected in score plots or in the SPE and T^2 plots, a physical interpretation is found by transforming the model output back into the original process space. Contribution plots indicate the variables that have caused the deviation. By plotting $\mathbf{e}(k)$ as a bar graph, the contributions to $SPE(k)$ are seen (Figure B.6 (left)). The relative size of the bars indicates the contribution of each variable to the prediction error. Similarly, when T^2 exceeds its limits, the variables that have forced the score away from zero along the jth PC can be found by inspecting:

$$\mathbf{c}(k) \quad = \quad \mathbf{x}(k)\mathbf{P}_j \tag{B.14}$$

where $\mathbf{c}(k)$ is a vector with the contributions from each process variable at time k. $\mathbf{P}_j$ is the diagonal matrix of the column vector $\mathbf{p}_j$. A bar graph of the contribution to the score changes along the first PC is shown in Figure B.6 (right).

In conclusion, PCA provides means to extract and summarise the process information into a small number of measures, which are then monitored and presented in a graphical way. The scores represent the main mechanisms driving the process, the SPE describes the deviations from the usual relations between

variables and the T^2 represents the magnitude of the scores. PCA also allows investigation of the contribution to deviations, which makes it possible to isolate the deviating original variables.

Adaptive PCA monitoring

PCA monitoring, as described above, assumes that the process conditions do not change significantly. This is rarely the case for wastewater treatment processes. Diurnal and seasonal changes affect the variables so that their mean, variability and correlation change accordingly. Small but important events tend to be obscured within the residuals related to the normal variation.

Updating the scaling parameters

A way to reduce the problem of changing conditions is to make the monitoring model adaptive, i.e. update the model on an adequate time scale. If the covariance structure of the monitored variables is believed to be static, i.e. the relations between the variables remain the same, regardless of their mean values and variability, the adaptation may be limited to data preprocessing (Hwang and Han; 1999). This involves, for instance, highpass filtering or updating of the scaling parameters. A simple way of updating the scaling parameters is to calculate them from a moving window of appropriate length. A more sophisticated method is to update the parameters for the jth variable recursively as:

$$\bar{x}_j(k) \quad = \quad \alpha\, \bar{x}_j(k-1) + \frac{x_j(k)}{M} \tag{B.15}$$

and

$$\hat{\sigma}_j^2(k) \quad = \quad \alpha\, \hat{\sigma}_j^2(k-1) + \frac{(x_j(k) - \bar{x}_j(k))^2}{M-1} \tag{B.16}$$

where $\bar{x}_j(k)$ and $\hat{\sigma}_j^2(k)$ are the estimated variable mean and variance at time k, respectively, and α ($0 \leq \alpha \leq 1$) is a forgetting factor with $\alpha = 1$ corresponding to no discounting of past data. M is the effective memory length and can be calculated as (Åström and Wittenmark; 1989):

$$M \quad = \quad \frac{1}{1-\alpha} \tag{B.17}$$

When combining Equations B.15 and B.17 it becomes clear that the updating is equivalent to an exponential filter. This means that the scaling parameters will depend on the exponential discounting of historical data.

Updating the covariance structure

When the covariance structure is believed to change over time, the model itself must be updated. A new model, based on a moving window, is identified for each new sample. The most recent score value and model residual are then the model output. Consequently, a sample will have a constant influence on the model until it leaves the window. The model at time k is based on the covariance matrix

$$(\mathbf{X}^T\mathbf{X})(k) \quad = \quad \sum_{i=0}^{W}(\mathbf{x}(k-i))^T(\mathbf{x}(k-i)) \tag{B.18}$$

where W is the the length of the window and $\mathbf{x}(k-i)$ are the values at i samples back in time. The scaling parameters are also updated from the moving window.

In analogy with the updating of the scaling parameters, the model can be updated using exponential weights. The covariance matrix is then updated recursively (Dayal and MacGregor; 1997b):

$$(\mathbf{X}^T\mathbf{X})(k) = \alpha\,(\mathbf{X}^T\mathbf{X})(k-1) + (\mathbf{x}(k))^T\mathbf{x}(k)) \tag{B.19}$$

where α is the forgetting factor from Equations B.15-B.17. The chosen value of the updating parameter α varies significantly depending on the aim of the monitoring. When the focus is on fast changes, the updating speed is high, whereas when slower variation, or trends, are also of interest the choice of the updating speed is a trade-off between model accuracy and the lowest detectable frequencies.

Adaptation of the covariance matrix introduces some difficulties. Firstly, if no precautions are taken, the model will adapt to disturbances and failures atypical of the normal process behaviour. Secondly, if there is not sufficient excitation in the process data, process information will be lost as old data are discounted. It is therefore wise to test the information content of the data before it is used to update the model, especially if the forgetting factor is small, i.e. the effective

No	Variable	Symbol
1	temperature	T
2	sludge concentration in line 1	SS_1
3	sludge concentration in line 2	SS_2
4	air valve position of blower 1 in line 1	$Air_{1,1}$
5	air valve position of blower 2 in line 1	$Air_{2,1}$
6	air valve position of blower 1 in line 2	$Air_{1,2}$
7	air valve position of blower 2 in line 2	$Air_{2,2}$
8	influent conductivity	cond.
9	influent ammonia	S_{NH}
10	influent pH	pH_{inf}
11	influent flow rate	Q
12	pH in effluent from biological treatment	pH_{bio}
13	effluent turbidity	FTU
14	effluent pH	pH_{eff}

Table B.1: Measured variables at Ronneby WWTP.

memory is short. These issues are discussed in e.g. Wold (1994) and Dayal and MacGregor (1997b). A third problem caused by an adapting covariance matrix is that the coordinate system defined by the principal component will rotate. This makes it hard to use scatter plots for interpretation of the process performance, as the coordinate system differs for each sample. An alternative method for adapting monitoring models and handling changing process conditions are discussed in Teppola et al. (1998a).

Results

In this section a few examples are presented and discussed to illustrate the above discussed methodology. The data used are real data from Ronneby wastewater treatment plant, Sweden. The Ronneby plant is operated as a biological nutrient removal plant with additional chemical treatment. The sampling period is 15 minutes. Table B.1 lists the available measurements from the online measurement system. The data display obvious changes in the process conditions, which is not surprising since they span several months—from summer to late autumn. We therefore present a few examples with different degrees of adaptation. However, we begin with a fixed monitoring model as basis for comparison.

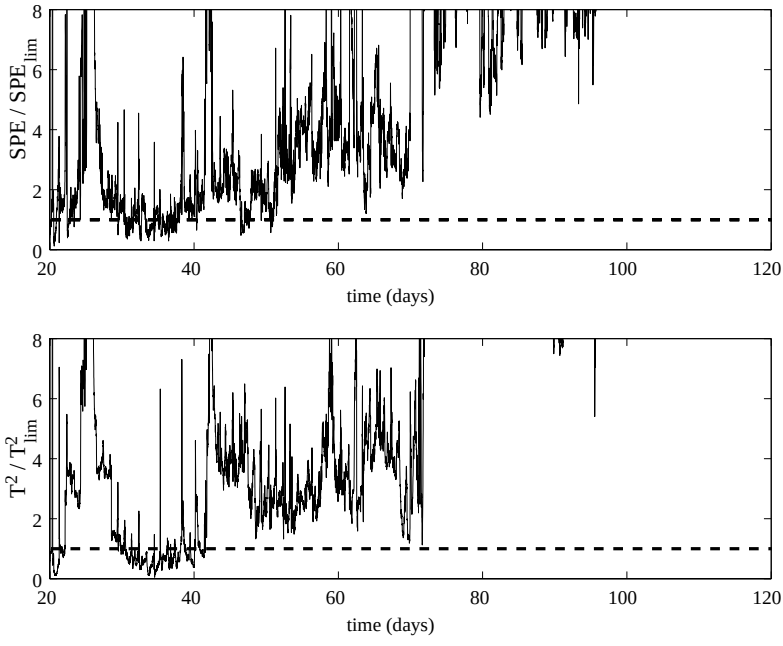

Figure B.7: No adaption. The relative SPE and T^2, i.e. the ratio between the residual and its limit.

PCA—no adaptation

A PCA model is identified from a 20-day period when the operation is considered to be normal, except for a few brief upsets (see Figure B.7). No preprocessing is carried out. Six components are chosen, using a scree test (see Figure B.2).

New data are projected onto the model and the result is seen in Figure B.7. It is obvious that the changing conditions cannot be covered by the fixed model. Almost from the beginning of the new period, both SPE and T^2 violate their limits and from day 40 the performance of the model deteriorates significantly. Thus, the model is not useful for monitoring and this illustrates the problem of monitoring data from changing process conditions.

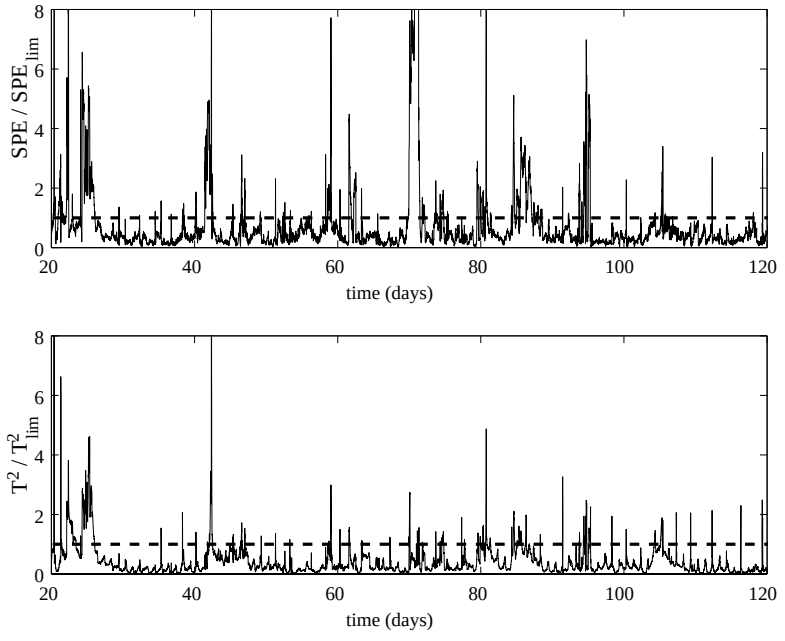

Figure B.8: Adaptive scaling parameters. The relative SPE and T^2, i.e. the ratio between the residual and its limit.

Adaptive scaling parameters

To overcome the problem of changing process conditions, a monitoring model with adaptive updating of the scaling parameters according to Equations B.15 and B.16 is identified from the same data used in the previous example. Six principal components are chosen using the 'quick-and-dirty' method. The forgetting factor, α, is set to 0.9995, which corresponds to 2000 samples $\approx$ 21 days. This is a reasonable choice, allowing detection of slower changes as well as fast ones. In Figure B.8 the SPE and T^2 are shown. They are plotted as the ratio to their limits and, hence, the limits are unity. The first 20 days are the identification period. It can be said that the model is generally valid (i.e. the SPE is below its limit) during the whole period, indicating a relatively constant covariance structure. However, there are some periods with high SPE, for instance, at days 20-21, 24-25, 42, 60-61 and 70 (Figure B.8).

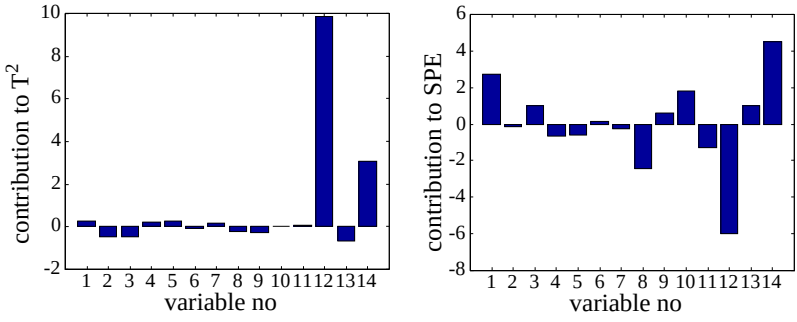

Figure B.9: The variables contributing to the deviation along the second PC at day 42.4 (left). The variables contributing to the deviation in SPE at day 42.4 (right).

Which variables have caused the deviation at, for instance, day 42? The loading plot indicates that pH_{bio} and pH_{eff} are strong contenders, but S_{NH} as well as $cond.$ may be responsible. A contribution plot, showing each variable's contribution to T^2 is shown in Figure B.9 (left). There are two variables contributing to the deviation along the second PC: pH_{bio} (no 12) and pH_{eff} (no 14). This confirms the information from the loading plot. In Figure B.9 (right), the variables contributing to the deviation in SPE at the same point in time are shown. Now, we are certain that pH_{bio} and pH_{eff} are responsible for the disturbance[1]. The cause of the deviations in these variables remains to be found: this is a task for the operator.

When looking at the T^2 measure, it is seen that it hardly exceeds its limit, not even when the SPE far exceeds its limit. The reason for this is possibly that the deviations are mainly caused by changes in the relations between the variables. Thus, the model does not capture the variability in a correct way, and most of the variability is outside the model plane. This indicates that the covariance structure of data is changing. For this reason, the information obtained from the score plots must be used with caution.

[1]Note that the directions in the contribution plots cannot be used directly.

Since the covariance matrix of the model remains constant, score plots can be used to visualise the process. In Figure B.10, the first and second score vectors are plotted against each other. It can be seen that there are a few upsets that force the process to deviate along the first and second PCs. Figure B.11 indicates which variables may be responsible for these deviations. The loadings associated with the first and second PC are plotted and the direction of a positive influence of each variable are indicated.

Adaptive PCA

The example presented above suggests that a fixed covariance structure is not fully capable of capturing the important variability of the process. Therefore, a model with adaptive scaling parameters as well as adaptive covariance structure is now used to monitor the same variables as in the example above. The result is shown in Figure B.12. The SPE is here mostly well inside its confidence region, while T^2 violates its limit more distinctively (compare with the fixed model in Figure B.8). This implies that the model captures the variability and that most of the variability will show up in T^2. The reason for this is that when the covariance structure adapts to new conditions, most of the variability will be in the model plane (T^2) and not orthogonal to the model plane (SPE). Put simply, some of the variability from the SPE chart in Figure B.8 has now been transferred to the T^2 chart in Figure B.12. This highlights the importance of monitoring both SPE and T^2 as they provide complementary information.

An example of the complementary relation between SPE and T^2 is seen if we take a closer look at the event at days 41-42 (Figure B.13). Initially, there is a change in the relation between the variables (i.e. a detection by SPE at day 41.4). As the variable relation changes and/or the model adapts, the SPE decreases again at day 41.6. However, T^2 is now large, indicating high variable magnitudes. It stays high until day 42.2, when SPE again becomes high, now indicating a severe disturbance at day 42.4. The same methods for isolation of the variables causing the deviations in SPE and T^2 used in the previous example can be applied here. It is important, though, that when the contribution to T^2 is considered, not only the first two score vectors are investigated.

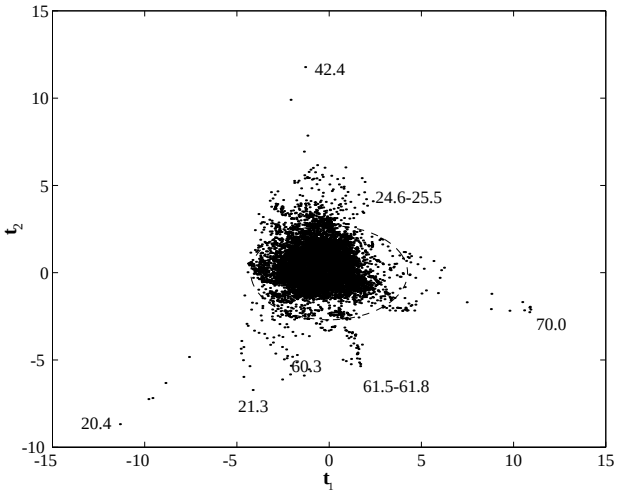

Figure B.10: The first and second score vector plotted as a scatter plot. The dashed ellipse indicates the 99 % confidence region.

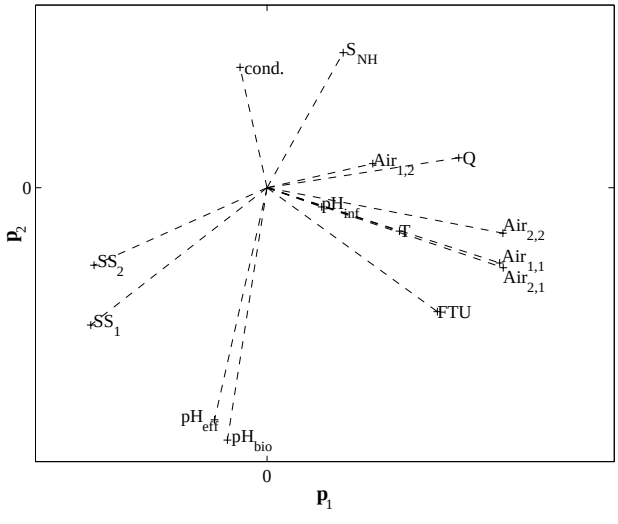

Figure B.11: The first and second load vector. An increase in the a variable will drive the process in the direction indicated. The distance from the origin indicates the relative influence of each variable. Variables located close, consequently, have similar influence on the process.

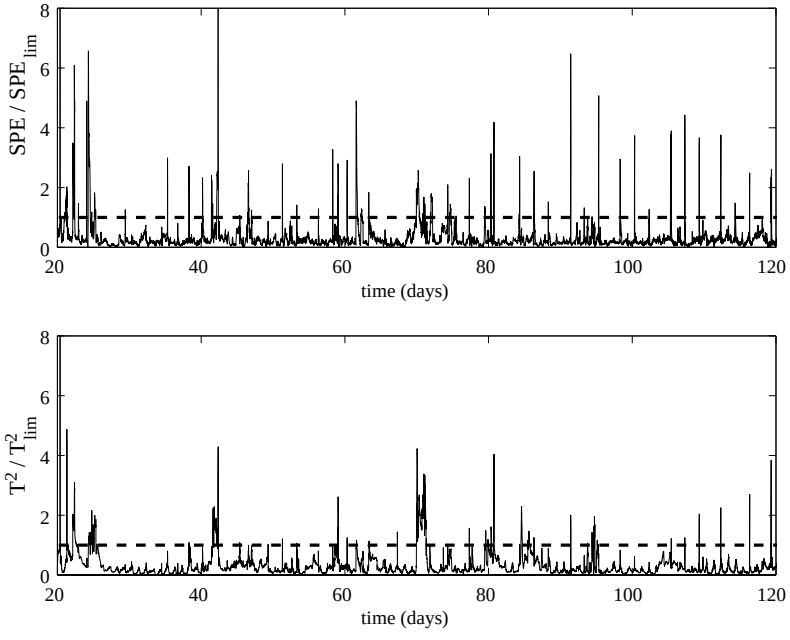

Figure B.12: Adaptive PCA. The relative SPE and T^2, i.e. the ratio between the residual and its limit.

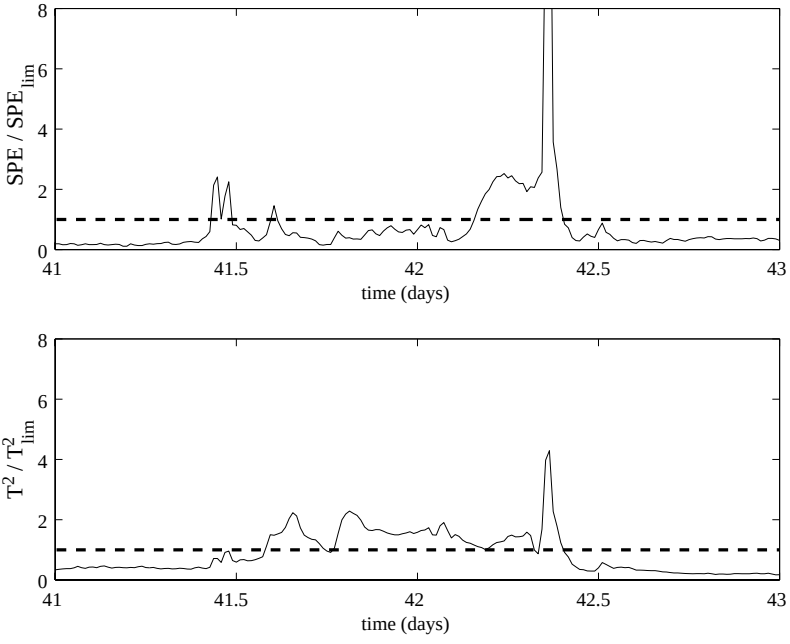

Figure B.13: Adaptive PCA. A closer look of SPE and T^2 at days 41-42.

General remarks

A comparison between the methods exemplified here shows the importance of addressing the changing process conditions. The model using fixed scaling parameters and covariance structure cannot handle the changing conditions in the process. It makes no sense to use such a model for monitoring data from changing conditions. When adaptive scaling parameters are used, the model can be used with some confidence, without loosing the capability to use score plots. This is an advantage of a fixed covariance structure since score plots are very intuitive.

Unfortunately, the model with adaptive scaling parameters is not fully capable of representing the process. The model based on adaptive scaling parameters and covariance structure handles the changing conditions well, but at a price: we can no longer use score plots, as in Figure B.10, to visualise the process. The choice of method becomes a trade-off between visualisation and accuracy.

Discussion

The authors' experience is that monitoring models ought to be kept as simple as possible. This implies that the number of parameters introduced is kept at a minimum. The reason for this is that the monitoring models should be as transparent as possible for the users. Therefore, one may start with a static model. If this is insufficient, updating of the scaling parameters may be enough for the model to handle changing conditions. If not, fully adaptive PCA (i.e. adaptive scaling parameters as well as covariance structure) is used.

The price we pay by omitting slow changes using adaptive PCA is that the monitoring becomes relative. Only changes are detected; the absolute state of the process is unknown. Depending on the updating speed, the changes possible to detect are of different speed. Here, the focus is on long term adaptation of the monitoring model (N corresponding to 21 days) which leaves the diurnal variation unaffected. However, if the focus is on the daily operation, where small deviations from the normal pattern must be detected, the updating speed must be high (e.g. N corresponding to 0.1 days). Whatever adaptive method and updating speed is used, it is a good idea to complement relative monitor-

ing with some form of absolute monitoring, e.g. sum of squares from normal operational point.

A few important objections to PCA as a monitoring technique for industrial processes are found in the literature. In addition to the ones raised on PCA and changing process conditions, it is often pointed out that PCA is a static technique, not suitable for dynamic processes. However, this is addressed by including a time lag or history in the analysis, i.e. the old measurements are included in the $\mathbf{X}$ matrix (e.g. Ku et al. (1995) and Luo et al. (1999)). This may further improve the monitoring model, but at a price of model complexity and a high number of variables in the $\mathbf{X}$ matrix. Another objection, related to both changing conditions and dynamics, is that PCA models a system in one time scale. Most industrial processes display multiple time-scale behaviour, i.e. events and disturbances occur at different scales. This introduces an error in the model, making it more difficult to separate the stochastic and deterministic components of data. This has been pointed out by, for instance, Kramer and Mah (1994) and Bakshi (1998). Therefore, in part two of this work, we discuss how more information is gained by extending the basic monitoring approach to multivariate analysis at multiple scales.

Conclusions

We have shown the potential of principal component analysis (PCA) as a tool for monitoring wastewater treatment processes. PCA accounts for collective effects, as it simultaneously analyses all variables. It also reduces the dimensionality of the data and extract the important information. PCA also provides different ways to visualise the process in an interpretable and intuitive manner, helping the user to extract information and make sensible decisions.

A PCA model is identified using data from normal or desired process operation, and then used to detect deviations from this behaviour. However, due to changing conditions, for instance, diurnal variations, seasonal changes and long term trends, the monitoring model must be updated. This is achieved by making the PCA model adaptive. Several levels of adaptation can be used. We have shown that adaptive scaling parameters are an option when the relationships between the variables do not change. This approach has advantages since

it allows intuitive graphical representation, such as score plots. When the relationships between the variables change, the covariance structure of the model must also change. Adaptive PCA, i.e. adaptive covariance structure, together with updated scaling parameters, provides us with a powerful tool for monitoring non-stationary processes in faster time scales.

Paper B

Addendum

It is mentioned in the paper that an updating of the monitoring model must be done in such a way that the model does not adapt to a behaviour atypical for the process. Moreover, if the process during some periods is more or less stable, this will also affect the updated model negatively. Only the first problem is discussed here, since the second problem is less serious in wastewater treatment operation.

Spikes and fast transients may be excluded from the model since this is what we generally want to detect. Thus, we want the model to remain sensitive to these disturbances. An updating rule based on the difference between the SPE and the median filtered SPE could be used.

An example: Let the SPE of a static (non-adaptive) model be according to the dashed line in Figure B.14 (top). From samples 1 to 90, the model seems to represent the process in an adequate way, except for the short disturbance at sample 40. After sample 90, it is evident that there has been a (more or less) permanent change in the process. It is clear that we want the model to adapt to the process change after sample 100, but not to the short disturbance at sample 40. Also, it is not desirable that the transient behaviour between samples 90 and 100 are used for updating, since this is typically the behaviour we want the model to be sensitive to. By looking at the squared difference between the SPE and the median filtered SPE (Figure B.14 (bottom)), the two periods that should be excluded from the updating are clearly discernible. Using a limit on the difference, a simple updating rule is obtained. Note that slow changes do not affect the updating decision.

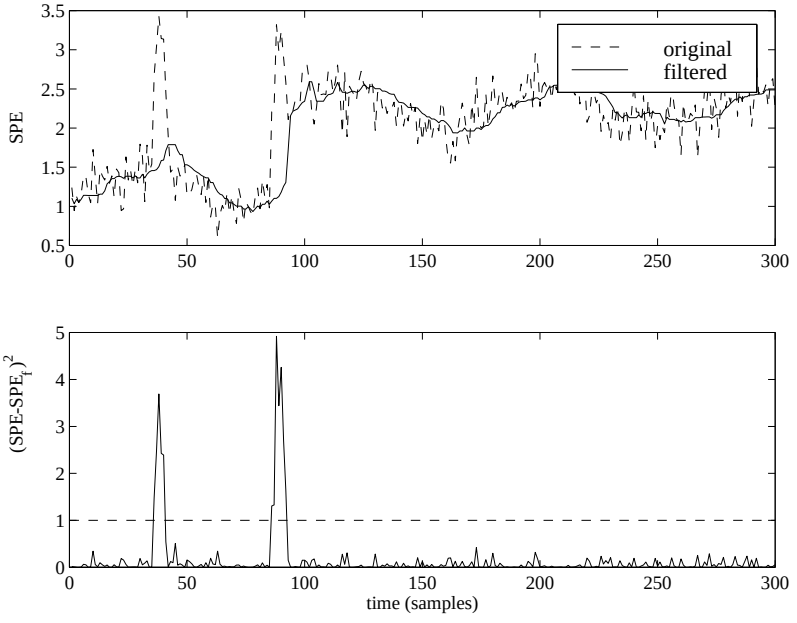

Figure B.14: A hypothetical SPE from a static model, (--) SPE, (—) median filtered SPE (top). The squared difference between SPE and median filtered SPE with a decision limit (bottom).

This approach is simple and intuitive since it is based on the same measure as used for monitoring. The only parameter of the updating rule is the length of the filter. It is important that a median filter is used if the rule is crisp (either include or exclude), since it preserves fast transients. If a 'fuzzy' rule is used, that is the model is updated using weights between 0 and 1, a linear digital filter is more appropriate. The limit itself may also be updated, using historical SPE values.

Paper C

Monitoring wastewater treatment operation. Part II: Multiscale monitoring

C. Rosen and J.A. Lennox

Published together with Paper B in a combined form in
Wat. Res. **35**(14): 3402-3410, 2001.

Abstract: *In this work a multiple time scale approach to multivariate monitoring of wastewater treatment processes is presented. The methodology involves multiresolution analysis (MRA) in combination with multivariate principal component analysis (PCA) modelling. MRA provides a tool for investigation and monitoring of process measurements at different time scales by decomposing measurement data into separate time scales. This makes it possible to increase the sensitivity of the monitoring and to detect small but significant events in data displaying large variations. The decomposition into separate time scales results in an increased number of data. Therefore, MRA is combined with the PCA to reduce the dimensionality of data. The multiscale approach is also used to overcome the problem of monitoring changing process conditions. The changes often appear as slow variations, i.e. at low frequencies. Thus, time scales that display stationary behaviour may be modelled by PCA, whereas scales that are not stationary have to be monitored by other means. The time scale information is important information in the interpretation of a disturbance and in finding its physical cause.*

Keywords: Monitoring; multiresolution; multiscale; PCA; wastewater; wavelets.

Nomenclature

$\mathbf{a}_j$	vector of scaling coefficients on scale j
$\mathbf{d}_j$	vector of wavelet coefficients on scale j
$\mathbf{E}$	error matrix
γ_j	significance factor on scale j
$\mathbf{g}$	highpass filter
$\mathbf{g}^*$	highpass reconstruction filter
$\mathbf{G}_j$	transformation matrix corresponding to $\mathbf{g}$
$\mathbf{G}'_j$	submatrix of $\mathbf{G}_j$
$\mathbf{G}^*_j$	reconstruction matrix corresponding to $\mathbf{G}_j$
$\mathbf{h}$	lowpass filter
$\mathbf{h}^*$	lowpass reconstruction filter
$\mathbf{H}_j$	transformation matrix corresponding to $\mathbf{h}$
$\mathbf{H}'_j$	submatrix of $\mathbf{H}_j$
$\mathbf{H}^*_j$	reconstruction matrix corresponding to $\mathbf{H}_j$
L	number of scales
N	number of variables
$\mathbf{P}$	loading matrix
$\mathbf{P}_j$	loading matrix on scale j
SPE	sum of squared prediction error
$\mathbf{t}_j(k)$	scores at time k on scale j
$\mathbf{T}$	score matrix
T^2	Hotelling's T^2 statistics
$\mathbf{W}$	transformation matrix
$\mathbf{x}_j(k)$	data vector at time k on scale j
$\mathbf{X}$	matrix of measurement variables
$\mathbf{X}_j$	matrix of measurement variables on scale j

Introduction

In part I of this work, we pointed out some limitations with multivariate monitoring using principal component analysis (PCA). The problem of changing process conditions was addressed by making the monitoring model adaptive. A second related limitation is caused by the fact that PCA monitoring is carried out in one time scale only—that of the sampling frequency. Here, we present and discuss a multiscale approach to multivariate monitoring of wastewater treatment operation, utilising wavelet analysis, or more specifically multiresolu-

tion analysis (MRA) to decompose each variable into separate time scales, before multivariate monitoring is carried out. This allows us to focus the monitoring on phenomena that occur in different scales. It also allows for decomposition of the data into time scales that have a physical interpretation, such as, hydraulic dynamics, concentration dynamics and population dynamics. The time scale information can be used in the interpretation stage to find the physical cause of a disturbance.

Process operation monitoring is carried out to ensure that the process outputs comply with requirements on product quality, process safety and efficient use of resources. In most industrial applications, including wastewater treatment, process performance and operation is measured continuously. Often, the number of measured variables is high, demanding a structured approach to monitoring and analysis of the process. Multivariate statistics (MVS) provide a methodology to extract and structure information from large amounts of data. MVS have been used to monitor industrial processes for several decades. There are many examples in the literature on applications of MVS for process monitoring in general (see e.g. Kresta et al. (1991) or MacGregor and Kourti (1995)) and for monitoring of wastewater treatment operation (Rosen and Olsson; 1998; Mujunen et al.; 1998).

Normally, MVS-based monitoring is carried out at a single time scale, defined by the sampling interval. This time scale contains frequencies from the sampling frequency (f_s) (or more accurately the Nyquist frequency $f_N = 0.5 f_s$) down to the lowest frequencies present in the process. There are some limitations to what can be achieved with this approach. Since only one scale is monitored, uniscale MVS is most appropriate when the data contain events occurring at one scale, i.e. in a narrow frequency band. This is not the case in most industrial applications, and certainly not in wastewater treatment, where both fast and slow deviations occur (in this work, terms as fast and slow are used to describe how fast a signal changes). The presence of different time scales introduces an error to the monitoring model, as it becomes difficult to separate the stochastic and deterministic components of the data (Bakshi; 1998). This error degrades the sensitivity and, consequently, the ability to detect small, but significant, changes in data. Another shortcoming of the basic MVS approach is encountered when data are from periods of changing process conditions. Wastewater processes change continuously due to diurnal, weekly and seasonal changes. This intro-

duces two problems. Due to the changes, an MVS-based monitoring model does not remain valid for long since it irequires that the mean of the data are approximately constant, i.e. no trends are present. Secondly, small changes are not recognised as they tend to be obscured by the normal variations of the process.

Preprocessing of data sometimes improves the separation between the stochastic and deterministic components of the data. Linear filters, such as mean and exponential filters, may be used. However, in the case of multiscale data, nonlinear filters, such as median filters (Rosen; 1998a) and finite impulse response median hybrid (FMH) filters (Kosanovich and Piovoso; 1997), cause less distortion due to their multiscale nature (Bakshi; 1998). Lately, wavelet-based processing of data has been used to identify events in different time scales from process data (Lang et al.; 1996; Alsberg et al.; 1997; Flehmig et al.; 1998). Wavelet analysis has also become an important tool in analysis of environmental data (Percival and Mofjeld; 1997; Dohan and Whitfield; 1997; Whitfield and Dohan; 1997). The problem of monitoring changing processes was addressed in part I of this work, utilising different methods of adaptive monitoring to solve the problem. Another approach is to identify the trends and remove them from data (Champely and Doledec; 1997). Yet another way, based on clustering techniques, is described by Teppola et al. (1998a).

A framework within which both the problem of events in different scales and the problem of monitoring changing processes can be solved, is based on a combination of MRA and MVS. In MRA, data are split into separate time scales, using the wavelet transform. The decomposed data can be evaluated by MVS-based monitoring, for instance PCA, to achieve a multiscale monitoring methodology. Multiscale monitoring has some important advantages. The sensitivity of the monitoring model is increased, as every scale is monitored separately. Moreover, the separation of data into multiple time scales implies that the higher scales (high frequencies) will have approximately a constant mean and only the lower and/or lowest scale (low frequencies) will display trends or long term variation. Consequently, by omitting the lower and/or lowest scale from the monitoring, the problem of monitoring data from changing process conditions is partly solved. Also, the information on which scale a disturbance or event appears, may be used in the interpretation to find the physical cause of the an event or disturbance. The result is a monitoring methodology that com-

bines the ability of MRA to extract both time and frequency (scale) information on an event or disturbance with the ability of PCA to reduce the dimensionality of data and present information in an interpretable manner. This methodology for process monitoring has been proposed by Kosanovich and Piovoso (1997) and Bakshi (1998). Work related to multiscale monitoring is also found in Luo et al. (1999). In this paper, we apply and further develop this methodology to increase the flexibility of the monitoring so that the scales are linked to physically interpretable time scales. We also use the multiscale framework to solve the problem of monitoring wastewater treatment data during periods of changing process conditions.

This paper is organised as follows. In the next section, multiscale decomposition using MRA is presented as an interpretation of filterbanks. The following section discusses how multivariate monitoring may be carried out in multiple scales and three different methods are presented. The methods are applied to wastewater treatment data and illustrated by examples. This is followed by a discussion and some concluding remarks.

Multiscale decomposition

Wavelet analysis is a relatively recent technique for simultaneous analysis of the time and frequency contents of a signal. Conventional frequency analysis based on the Fourier transform consists of breaking up a signal into sine waves of various frequencies. The frequency content of the signal is found at the cost of time information. Wavelet analysis is similar in that it also decomposes the original signal using waves. The major difference is that where Fourier analysis uses sine waves of infinite length ($-\infty \leq t \leq \infty$), multiresolution analysis uses waveforms of finite length—wavelets (wavelet means 'little wave'). The finite length of the wavelets allows them to describe a local event in both time and frequency. Wavelets have proven useful in many different fields, from image processing to model identification (see e.g. Strang and Nguyen (1996) or Alsberg et al. (1997)).

For the purpose of monitoring and detection, wavelets can be used to decompose a signal into different scales with decreasing level of detail or resolution. This is sometimes referred to as multiresolution analysis (MRA). A signal, $\mathbf{x}$,

is filtered with a highpass, $\mathbf{g} = [g_m\ g_{m-1}\ \dots\ g_2\ g_1]$, and a lowpass filter, $\mathbf{h} = [h_m\ h_{m-1}\ \dots\ h_2\ h_1]$, respectively. The result is a set of coefficients describing the details of the signal, $\mathbf{d}$, and another set describing the approximation of the signal, $\mathbf{a}$ (Figure C.1). This decomposition is carried out to a desired number of scales, by recursively applying the highpass and lowpass filters to the approximation coefficients of the previous level. As can be seen in the figure, after a filter is applied, the result is downsampled. This means that the total number coefficients in the wavelet domain is the same as the number of samples in the original signal. The coefficients still contain all the information carried by the original signal.

For both the highpass and the lowpass filter, there are corresponding reconstruction filters, $\mathbf{g}^*$ and $\mathbf{h}^*$, respectively. Using these filters for the reconstruction, together with upsampling, will result in a perfect reconstruction of the original signal in one scale (see Figure C.1).Alternatively it is possible to reconstruct, separately, the components corresponding to each scale. The number of coefficients in the time domain will then be the original number of samples multiplied by the number of scales (Figure C.2).The procedure of transformation to the wavelet domain and then back to separate scales in the time domain can be seen as a bandpass filter. The result is the original data decomposed into scales with decreasing detail (Figure C.3).

Wavelet domain

Matrix algebra gives us a compact way of representing the calculations. Let the data signal ($\mathbf{x}$) be a vector of size $[m \times 1]$. In matrix form the wavelet coefficients on the highest level are found as:

$$\mathbf{d}_1 = \mathbf{G}_1\mathbf{x}; \tag{C.1}$$

where $\mathbf{G}_1$ is a transformation matrix of size $[m/2 \times m]$.

$$\mathbf{G}_1 = \begin{bmatrix} g_m & g_{m-1} & \cdots & g_2 & g_1 & 0 & 0 & \cdots \\ 0 & 0 & g_m & g_{m-1} & \cdots & g_2 & g_1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{C.2}$$

Similarly, the scaling coefficients on the first level are:

$$\mathbf{a}_1 = \mathbf{H}_1\mathbf{x}; \tag{C.3}$$

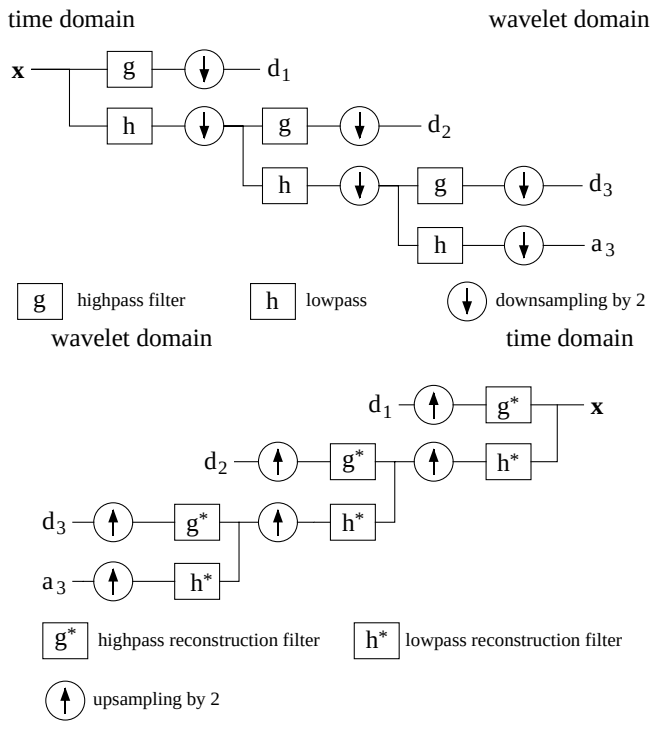

Figure C.1: Multiscale decomposition interpreted as bandpass filtering (top). Transformation from a uniscale time domain to a multiscale wavelet domain. Multiscale reconstruction (bottom). Transformation (reconstruction) from a multiscale wavelet domain to uniscale time domain.

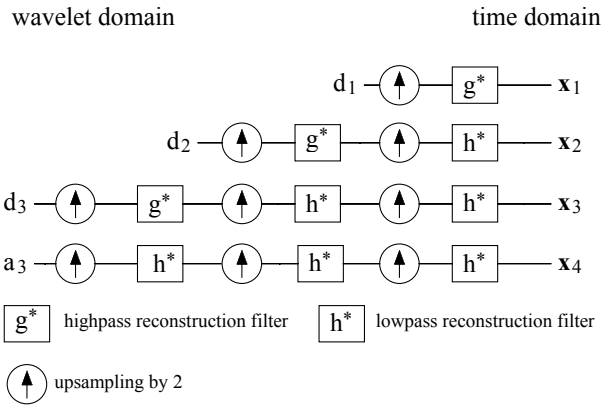

Figure C.2: Transformation from a multiscale wavelet domain to a multiscale time domain.

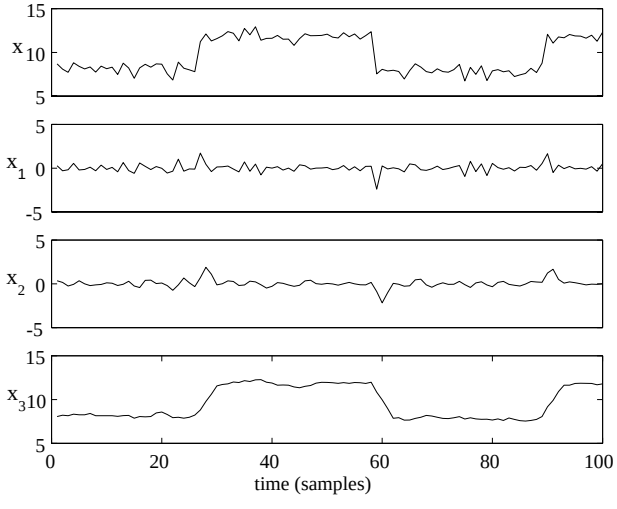

Figure C.3: Decomposition of signal into scales. Detail 1 (x_1), detail 2 (x_2) and approximation 2 (x_3) all sum up to the original signal ($\mathbf{x}$).

with

$$\mathbf{H}_1 = \begin{bmatrix} h_m & h_{m-1} & \dots & h_2 & h_1 & 0 & 0 & \dots \\ 0 & 0 & h_m & h_{m-1} & \dots & h_2 & h_1 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix} \tag{C.4}$$

The wavelet and scaling coefficients on the next level are calculated recursively from the coefficients on the previous level. The transformation matrices on the next level are given by:

$$\mathbf{G}_j = \mathbf{G}'_{j-1}\mathbf{H}_{j-1} \tag{C.5}$$

and

$$\mathbf{H}_j = \mathbf{H}'_{j-1}\mathbf{H}_{j-1} \tag{C.6}$$

where $\mathbf{G}'_{j-1}$ and $\mathbf{H}'_{j-1}$ are submatrices of size $[m/2^j \times m/2^j]$ of $\mathbf{G}_{j-1}$ and $\mathbf{H}_{j-1}$, respectively. For L scales a transformation matrix $\mathbf{W}$ can be constructed from the matrices of Equations C.5 and C.6 so that transformation on all scales is:

$$\begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \vdots \\ \mathbf{d}_L \\ \mathbf{a}_L \end{bmatrix} = \begin{bmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \\ \vdots \\ \mathbf{G}_L \\ \mathbf{H}_L \end{bmatrix} \mathbf{x} = \mathbf{W}\mathbf{x} \tag{C.7}$$

Note that due to the dyadic downsampling the total number of coefficients are the same as the number of data points in $\mathbf{x}$. The above transformation is easily extended to involve a multivariate data matrix $\mathbf{X}$ of size $[m \times n]$, where n is the number of variables.

Time domain

As mentioned before, wavelet decomposition allows for perfect reconstruction. This means that the transformation matrix $\mathbf{W}$ has some special properties. $\mathbf{W}$

is unitary, that is $\mathbf{W}^T\mathbf{W} = \mathbf{I} = \mathbf{W}\mathbf{W}^T = \mathbf{I}$ and $\mathbf{W}^T = \mathbf{W}^{-1}$. So, perfect reconstruction of the signal $\mathbf{x}$ is achieved:

$$\mathbf{x} = \mathbf{W}^{-1}\mathbf{W}\mathbf{x} \tag{C.8}$$

Here, the reconstruction is a transformation from a multiscale wavelet domain to a uniscale time domain. However, it is also possible to make a transformation from the multiscale wavelet domain to a multiscale time domain. The multiscale time domain transformation is obtained by:

$$\mathbf{x}_j = \mathbf{G}_j^*\mathbf{G}_j\mathbf{x} \tag{C.9}$$

where the reconstruction matrix $\mathbf{G}_j^*$ is found as a subspace of $\mathbf{W}^{-1}$:

$$\mathbf{W}^{-1} = \begin{bmatrix} \mathbf{G}_1^* & \mathbf{G}_2^* & \dots & \mathbf{G}_L^* & \mathbf{H}_L^* \end{bmatrix} \tag{C.10}$$

The number of data points in $\mathbf{x}_j$ is the same as the number of points in the original data $\mathbf{x}$. Moreover, the sum of all scales is now equal to the original signal:

$$\mathbf{x} = \sum_{j=1}^{L+1} \mathbf{x}_j \tag{C.11}$$

MRA for process analysis

Two tasks are achieved by using MRA on measurement data: the separation of events in both time and frequency (scale) and characterisation of the different events. The appearance of an event in different scales may be useful in the determination of physical causes. An important advantage of analysing the measurement variables in separate scales is the increase in sensitivity for small, but significant, events. When data include diurnal variations, small changes can sometimes be obscured by the larger variations. However, by separating the data into scales, small events are localised in both time and frequency. For instance, a slight change in the mean of a signal is difficult to detect since daily variations or noise may be of much higher amplitude.

Before MRA is carried out on a signal, a wavelet function must be chosen. The Haar wavelet was one of the first wavelets, proposed by Haar in 1909 (Alsberg et al.; 1997). Since then, the mathematicians have developed many different wavelets, with properties that make them suitable for various applications. The choice of wavelet is subject to many influential factors. In monitoring of wastewater treatment data, the measurement signals often display discontinuities. If the monitoring aims at detecting these types of features (e.g. steps and spikes) the square shaped Haar wavelet is suitable, due to its good localisation in time. If the task, however, is to detect smooth and slow changes, a higher order wavelet may be more suitable. The interested reader can consult the literature for more information on the choice of wavelet (see e.g. Torrence and Compo (1998) or Misiti et al. (1996)).

Here, we use the Haar wavelet. The Haar wavelet has good localisation in time but poor localisation in frequency. This can be seen from the gain-frequency plot of its filters (Figure C.4). The poor localisation in frequency results in a significant leakage between the time scales. Higher order filters have better localisation in frequency but at the cost of deterioration in time localisation. The filters corresponding to the Haar wavelet are:

$$\begin{aligned} \mathbf{h} &= \left[\tfrac{1}{\sqrt{2}},\ \tfrac{1}{\sqrt{2}}\right] \qquad \mathbf{h}^* = \left[\tfrac{1}{\sqrt{2}},\ \tfrac{1}{\sqrt{2}}\right] \\ \mathbf{g} &= \left[-\tfrac{1}{\sqrt{2}},\ \tfrac{1}{\sqrt{2}}\right] \qquad \mathbf{g}^* = \left[\tfrac{1}{\sqrt{2}},\ -\tfrac{1}{\sqrt{2}}\right] \end{aligned} \tag{C.12}$$

A practical decision is the number of scales that the original data set will be decomposed to. This depends significantly on the application and the modelled process. However, some general rules can be given. The number of scales should ideally be chosen so that the separation between stochastic and deterministic components of a variable is sufficient. This means that, under the assumption that the high frequency content of a signal is stochastic (noise) and that the low frequency content is deterministic, the lowest or lower scales should predominately contain deterministic features. Another aspect of the choice of scales has to do with changing process conditions. As mentioned before, the multiscale approach may be used to overcome the problem of monitoring changing processes. This is done by omitting the lowest scale approximation from the monitoring. The number of scales is then a compromise between the desire to

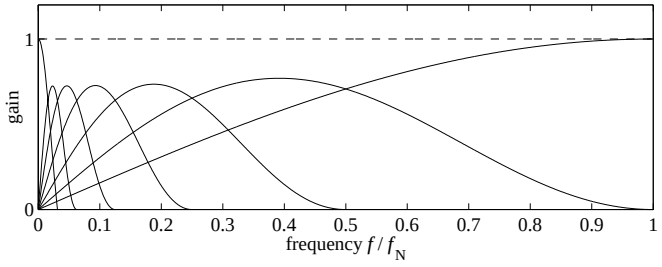

Figure C.4: The gain as function of the frequency of the filters of a six scale decomposition based on the Haar wavelet. The alias effects have been removed for clarity.

remove the changing components of the data while being able to detect slower changes. There is also a practical limit on the number of scales, as the filter (window) becomes longer the more scales are included.

MRA provides a tool for decomposing a measurement signal into different time scales. This can be utilised in process monitoring for detection of and extraction of information on deviating events. Decomposition into separate time scales is a natural way to investigate measurement data from processes with a wide dynamic range.

Monitoring in multiple scales

The reason for online monitoring of multivariate data in multiple scales is not only to determine whether the process is within normal operation, but also to determine the scale in which the disturbance appears. From this, information on the characteristics of the disturbance is inferred and used for interpretation.

New measurements are continuously added according to the sampling rate and in order to use MRA online, a moving window must be used on which Equation C.9 is applied. The most recent value is the output of the decomposition. Again, using the filter interpretation, the decomposition can be interpreted as a number of finite impulse response (FIR) filters—one for each scale. Monitoring a single variable is a straight-forward task and univariate statistical process

control (SPC) techniques are utilised. However, if MRA is carried out on N original variables, decomposing the variables into L scales, we will end up with $(L+1) \times N$ new variables. This increase in the number of new variables calls for multivariate methods that are capable of both reducing the dimension of data and handling collinear variables. A natural choice is PCA. Methodologies, combining MRA with PCA have been proposed by Kosanovich and Piovoso (1997) and Bakshi (1998). The methods presented in this section differ slightly to the ones of Kosanovich and Piovoso (1997) and Bakshi (1998). The main difference is that here the monitoring is carried out in the time domain on each scale as opposed to the wavelet domain. This gives us a flexibility to recombine scales, which otherwise is not possible. Another difference is that to overcome the problem of non-stationary data, the lowest scale is omitted from the monitoring.

Multivariate monitoring

PCA has become a popular method for multivariate monitoring. Descriptions of PCA-based monitoring are found in the literature (e.g. Wise and Gallagher (1996b) or part I of this work). PCA can be viewed as a coordinate transformation, where the new coordinate system is rotated in such a way that a minimum of orthogonal directions cover as much as possible of the variability in the data matrix ($\mathbf{X}$). The transformation is expressed as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{C.13}$$

To identify a PCA model, $\mathbf{P}$, is to find the transformation from the original coordinate system to the new lower dimensional coordinate system defined by the principal components (PCs). $\mathbf{T}$ is the transformed data or the scores. If the number of dimensions of the new coordinate system is chosen wisely, $\mathbf{E}$ contains mostly noise (see part I for choice of the number of components). As new data are projected onto the model identified from training data, the scores can be monitored using SPC techniques. The scores can also be visualised in scatter plots, providing the user with a graphical representation of the process performance. A more compact way is to monitor the statistical fit of the model, for instance, the sum of the squared prediction error (SPE) and Hotelling's T^2 at each sample. For a more detailed explanation of PCA, see part I.

Combination of MRA and PCA

A combination of the capabilities of MRA to decompose data into scales and of PCA to reduce the dimension of the data provides a powerful tool for monitoring processes with several time scales. First, PCA models are identified for each scale from data representing normal or desired behaviour. This is done by calculating the reconstructed data on each scale by applying a moving window the length of the longest filter ($\mathbf{h}_L$ or $\mathbf{g}_L$) to the training data. Monitoring new data involves decomposition and projection of data onto the scale PCA models. A window is used in the same manner as when the models were identified. The transformed data from the window on scale j are:

$$\mathbf{X}_j = \mathbf{H}_j^* \mathbf{H}_j \mathbf{X}_w \tag{C.14}$$

Let the last sample of $\mathbf{X}_j$ be $\mathbf{x}_j$ and project it on the PCA model:

$$\mathbf{t}_j(k) = \mathbf{x}_j(k)\mathbf{P}_j \tag{C.15}$$

where $\mathbf{t}_j(k)$ are the scores at time k and $\mathbf{P}_j$ is the loading matrix of the PCA model at scale j. The scores are monitored using SPC techniques or scatter plots. The SPE can be calculated using:

$$\hat{\mathbf{x}}_j(k) = \mathbf{x}_j(k)\mathbf{P}_j\mathbf{P}_j^T \tag{C.16}$$

and

$$SPE_j(k) = \sum_{i=1}^{n} (\hat{x}_{i,j}(k) - x_{i,j}(k))^2 \tag{C.17}$$

The residual is compared to the predefined limit for that particular scale model. If the residual exceeds the limit, a detection is triggered. In the same manner, Hotelling's T^2 is calculated as:

$$T^2(k) = \mathbf{t}(k)\Lambda^{-1}\mathbf{t}^T(k) = \mathbf{x}(k)\mathbf{P}\Lambda^{-1}\mathbf{P}^T\mathbf{x}^T(k) \tag{C.18}$$

where $\mathbf{t}(k)$ are the scores at time k and Λ^{-1} is the diagonal matrix of the inverse of the eigenvalues associated with the retained PCs.

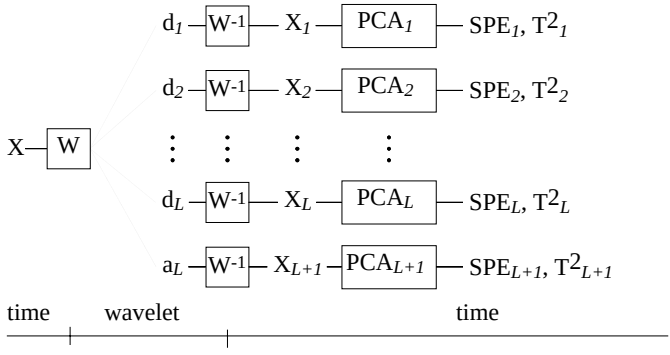

Figure C.5: Decomposition of the original data matrix into scales. The scale data are monitored using a PCA model on each scale.

The number of data points used for detection at a certain point in time is $(L + 1) \times N$, where L is the number of scales and N is the number of variables. This increased number is, however, based on only the N original data points. If a certain overall confidence limit for all the scales together is to be obtained, the confidence limits on each scale must be adjusted. Bakshi (1998) suggests that the confidence limit for each scale model is adjusted according to:

$$C_L \quad = \quad 100 - \frac{1}{L+1}(100 - C) \tag{C.19}$$

where C is the desired overall confidence limit and C_L is the confidence limit used on each scale.

The original variables may not be of similar amplitude. Therefore, it is important that the data are scaled in respect to amplitude. In multivariate analysis, mean centring and scaling to unit variance (autoscaling) is often used. Scaling is done in order to give all variables the same influence on the multivariate model, regardless of their amplitude. This is a good starting point when no a priori knowledge is available. When using multiple scales, there are different options as to where the scaling is applied. Applying the scaling before the decomposition into separate scales means that the data on the different scales are not necessary scaled equally. This is because the variables have different frequency content. However, this may be advantageous since different variables do in fact have different importance at different scales, and therefore ought to

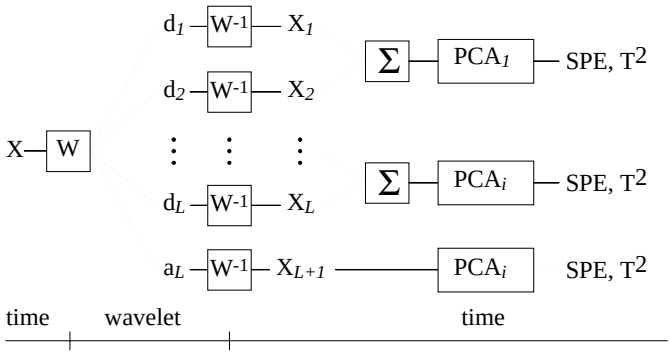

Figure C.6: Recombination of scales into fewer and more physically interpretable scales.

be more or less influential on that scale. A second option is to scale the decomposed data and then rescale it before the reconstruction. This gives every variable the same influence at each scale.

Recombination of scales

A difficulty with the multiscale method described above is that when the number of scales increases, the number of variables to monitor increases as well. PCA will reduce the number, but with a high number of scales, we might end up with more variables than in the original data set. One way to solve this problem is to recombine some of the scales so that the effective number of scales to monitor decreases. For instance, the first scales are recombined to a 'fast' scale (e.g. hydraulic dynamics), the middle scales are combined to constitute a 'medium fast' scale (e.g. concentration dynamics) and the lower and/or lowest scale (e.g. population dynamics) represents the 'slow' changes (Figure C.6). This makes it possible to choose the number of scales that is monitored. There is another benefit of this approach. The choice of dyadic scales has no physical justification, but is a property of MRA. Therefore, recombining the MRA scales into physically interpretable scales, approximately corresponding to true time scales present in the process, may simplify the interpretation of the results.

Multiscale PCA

Multiscale PCA (MSPCA) was originally proposed by Bakshi (1998). The idea is that the PCA models at each scale are used to determine whether a scale contains significant information or not at a certain point in time. If a scale is significant, the data on that scale is used in the reconstruction of a uniscale estimate of the original data. The reconstructed uniscale data are then monitored using a unifying uniscale PCA, which is based on the scales included in the reconstruction. The recombination of scales solves the problem of many scales, while the feature extraction regards each scale separately to extract important events.

In the version of MSPCA presented here, data from a period of normal or desired process behaviour are decomposed into scales to the wavelet domain. A PCA model is identified at each scale except the lowest (approximation). New data are decomposed and projected onto the scale models. If a model residual at any scale exceeds its value, the scale is said to be significant, and is used to reconstruct coefficients into a uniscale set of data. The reconstructed data are monitored using a uniscale PCA, which is based on training data of the scales that display significance. The rule for significance can be expressed:

$$\gamma_j(k) \quad = \quad \begin{cases} 1 & \text{if } SPE_j(k) > SPE_{lim,j} \text{ or } T_j^2(k) > T_{lim,j}^2 \\ 0 & \text{otherwise} \end{cases} \tag{C.20}$$

The reconstruction of a uniscale data set is then:

$$\mathbf{x}(k) = \sum_{j=1}^{L} \gamma_j(k)\mathbf{x}_j(k) \tag{C.21}$$

This is possible since the reconstruction of data is the sum of all scales (see Equation C.11). The reconstructed data sample at time k is projected onto a PCA model constructed from the covariance matrix:

$$\mathbf{X}_0^T\mathbf{X}_0 = \sum_{j=1}^{L} \gamma_j(k)\mathbf{X}_{0,j}^T\mathbf{X}_{0,j} \tag{C.22}$$

where $\mathbf{X}_{0,j}$ is the training data (coefficients in the wavelet domain) and $\gamma_j(k)$ is the current significance factor on scale j. This means that the uniscale PCA

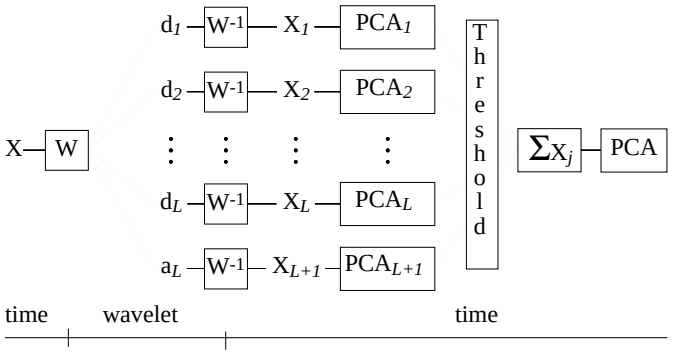

Figure C.7: The MSPCA procedure.

model will change according to the scales included in the reconstruction. The procedure is summarised as:

Identification of model:

1. Identify a period of normal process operation;
2. Decompose data into L number of scales (Equation C.9);
3. Identify a PCA model on each scale.

Online monitoring:

1. Decompose new data using a moving window (Equation C.9);
2. Project new data onto the PCA models at each scale;
3. Determine whether a scale is significant (Equation C.20);
4. Reconstruct data from significant scales (Equation C.21);
5. Calculate the corresponding uniscale PCA model and project the reconstructed data on the model (Equation C.22);
6. Calculate model residuals and limits for uniscale model.

The procedure is illustrated in Figure C.7.

Isolation of deviating variables

Whatever method is used to detect deviations, identification of the deviating variables may be carried out by investigating the contributions to the residuals. The approach used in the uniscale case, can here be applied at each scale. Contribution plots have been described in the literature (e.g. MacGregor et al. (1994) or part I of this work). In MSPCA, the variable contribution plot can be complemented by a scale contribution plot, which provides information on which scale is contributing the most to the residuals. This is done by investigating each term in Equation C.21. The information on which scales are contributing the most is used to find the cause of a disturbance.

Results

In this section we demonstrate and discuss the three methods presented earlier, i.e. PCA at each scale, PCA at a few recombined scales and MSPCA. The data used for the examples are actual process data from the Ronneby wastewater treatment plant in Sweden. The Ronneby plant is operated as a biological nutrient removal plant with additional chemical treatment. The data sampling period is 5 minutes. Table C.1 lists the available measurements from the online measurement system. The data span several months, from summer to late autumn implying significant changes in operating conditions over the period. In all examples the data are decomposed into six different scales, i.e. six detail scales and one approximation scale, using the Haar wavelet (see Equation C.12).

PCA monitoring of scale decomposed data

A data set from the first 20 days are decomposed and used as training set for a PCA model at each scale. An investigation of the eigenvalues using the screeplot method (see part I of this work) of each scale model suggests that six components should be retained at each scale. No preprocessing of data is carried out.

No	Variable	Symbol
1	temperature	T
2	sludge concentration in line 1	SS_1
3	sludge concentration in line 2	SS_2
4	air valve position of blower 1 in line 1	$Air_{1,1}$
5	air valve position of blower 2 in line 1	$Air_{2,1}$
6	air valve position of blower 1 in line 2	$Air_{1,2}$
7	air valve position of blower 2 in line 2	$Air_{2,2}$
8	influent conductivity	cond.
9	influent ammonia	S_{NH}
10	influent pH	pH_{inf}
11	influent flow rate	Q
12	pH in effluent from biological treatment	pH_{bio}
13	effluent turbidity	FTU
14	effluent pH	pH_{eff}

Table C.1: Measured variables at Ronneby WWTP.

New data are decomposed and projected to each scale PCA model and the resulting model residuals can be seen in Figure C.8 and Figure C.9. Looking at the SPE charts, we see that the SPE is mostly below its limit on each scale except for scale 7—the approximation scale. This means that the scale models are considered valid through the period. Some major events are found at, for instance, day 25, 42, 62, 70 and 95. Here, the deviations are detected in all scales. In most cases, the higher scales (scale 1 and 2) detect deviations before the lower scales, meaning that most of the disturbances are fast. When investigating original data, these disturbances manifest themselves as step changes or spikes in one or several variables. However, it is interesting to note that at, for instance, day 118 the disturbance is detected in the fourth and fifth scales first, which means that the disturbance is slow. This is seen on the original data as a relatively slow increase in flow rate and all air valve positions. The T^2 charts display similar behaviour. Most disturbances are fast, i.e. detected in the higher scales first and, thus, commenced by a rapid change in one or several variables. The disturbance at day 118 is most obvious in scale six, which is consistent with the SPE chart.

How can the scale information be used? The most straightforward way to use the scale information is to localise the scale in which a disturbance first is detected. Fast disturbances are detected in the higher scales whereas slow disturbances

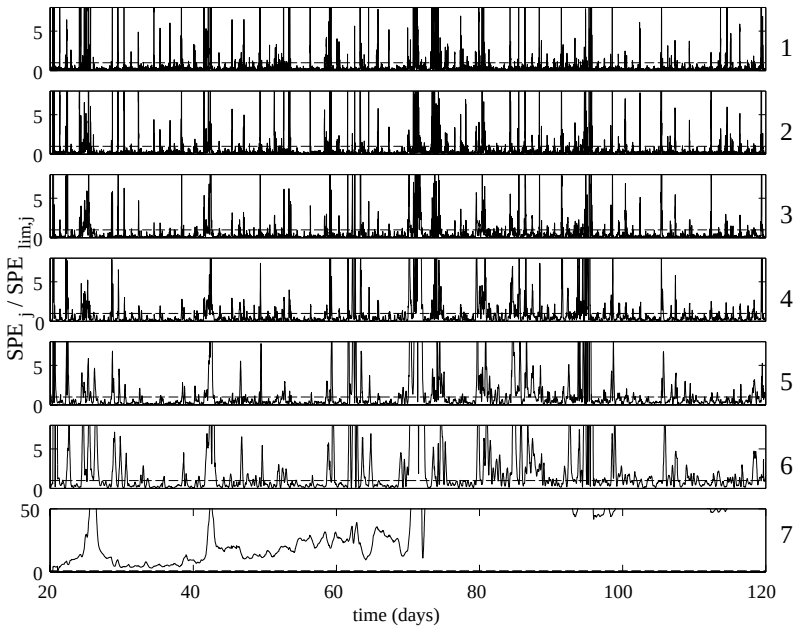

Figure C.8: The SPE residuals for scale 1 to 7 during a time period of 100 days, with confidence limits of 99.9 % according to Equation C.19. The lowest scale (7) should only be used as an indication on the absolute distance from operation defined by training data.

are detected in the lower scales. There is more information to gain, though. The way a disturbance appears across scales can reveal information on the disturbance characteristics. For instance, a spike will be strong in the higher scales but barely visible in the lower scales. A step will appear clearly in all scales whereas a ramp will appear most strongly in the lower scales.

A problem with multiscale monitoring is that the increased sensitivity leads to more detections that have to be evaluated. Moreover, events that occur in several scales will be detected with a slight delay from higher scales to lower scales. For instance, the disturbance at day 42 is first detected in the highest scale and is then propagated through the scales to the lowest scales, even though an investigation of original data reveals that it is the same disturbance. The increased

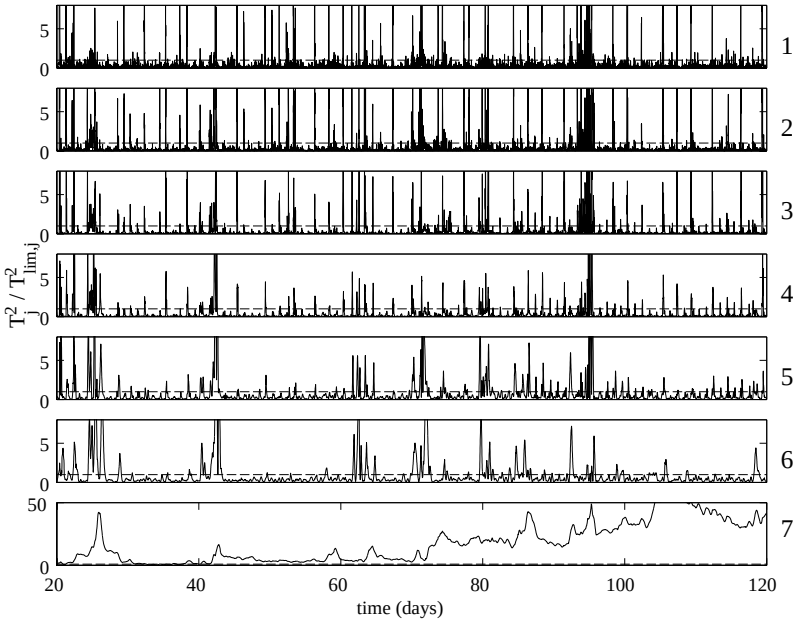

Figure C.9: The T^2 residuals for scale 1 to 7 during a time period of 100 days, with confidence limits of 99.9 % according to Equation C.19. The lowest scale (7) should only be used as an indication on the absolute distance from operation defined by training data.

sensitivity and redundancy make the chart difficult to interpret. Therefore, when the number of scales is more than a few, we need a way to structure all the available information.

PCA monitoring of physically interpretable scales

When the number of scales exceeds a few, it is hard to interpret the monitoring result. As mentioned before, a simple way to solve this problem is to recombine the scales into a small number of reconstructed scales before applying the monitoring models. Ideally, this is done in such a way that the resulting scales have a physical interpretation. Here, we will not focus on how to obtain the different time scales present in a system. We define fast dynamics as the first three scales.

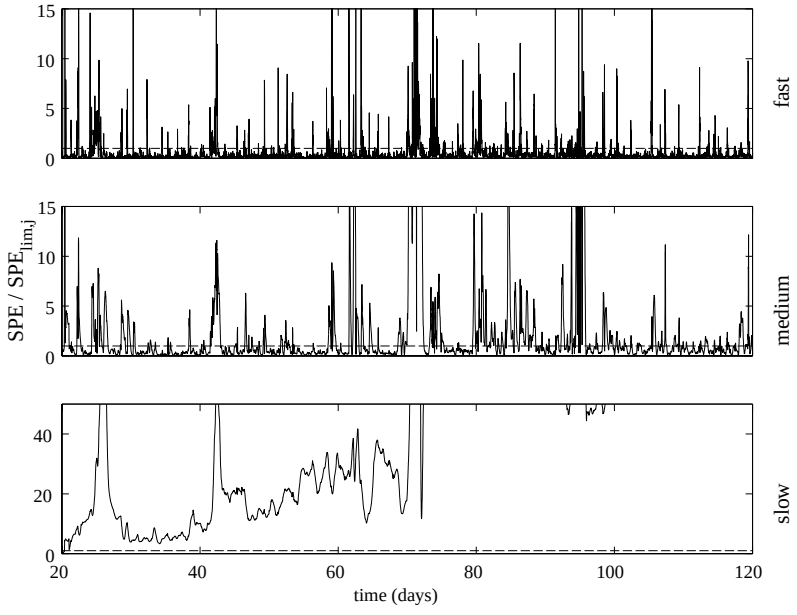

Figure C.10: The SPE residuals for the recombined scales defined as 'fast', 'medium fast' and 'slow'. The slow scale should only be used as an indication on the absolute distance from operation defined by training data.

This approximately corresponds to variations from 0 to 4 hours. The medium time scale consists of scales 4 to 6, which corresponds to 4 to 32 hours. The slow dynamics are represented by the approximation scale, i.e. approximately everything slower than the diurnal variation. The SPE and T^2 charts for the recombined scales are seen in Figures C.10 and C.11. The charts are easier to comprehend, as the data are split into only three components, i.e. fast, medium and slow. The charts show that the detail scale models (fast and medium) are valid, whereas the approximation scale (slow) model is not. This means that the slow scale can only be used as an indication on the absolute distance from normal operational conditions. Some of the disturbances occur in both the fast and the medium scale. This can roughly be interpreted as continuous disturbances. When a disturbance only appears in the fast scale, the disturbance is spike shaped.

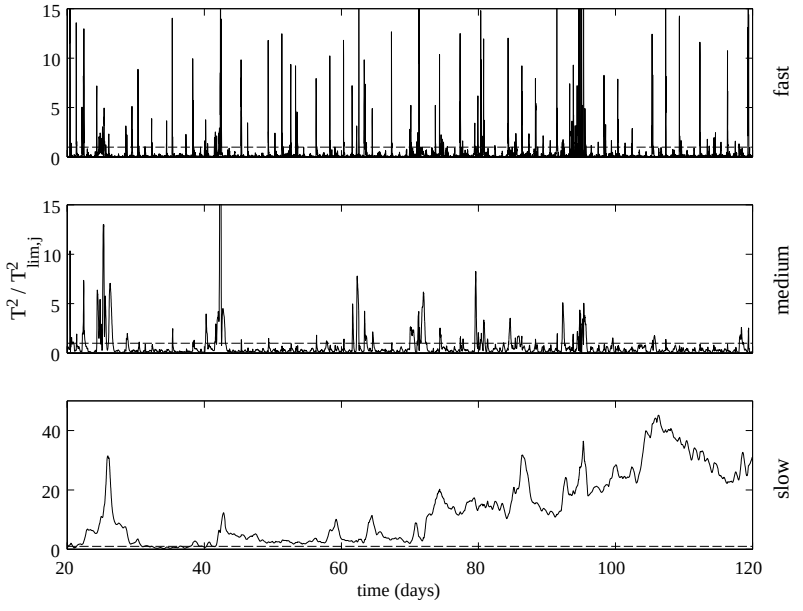

Figure C.11: The T^2 residuals for the recombined scales defined as 'fast', 'medium fast' and 'slow'. The slow scale should only be used as an indication on the absolute distance from operation defined by training data.

MSPCA monitoring

The MSPCA method is applied to the data used in the previous examples. The identification period is the first 20 days of operation. After identification of the scale models, the original training data set is stored in order to calculate the uniscale PCA model. The number of scales is seven, i.e. six detail scales and one approximation scale. Both SPE and T^2 are used on each scale to determine whether a scale is significant. The limits for the model residuals are corrected according to Equation C.19 in order to obtain an overall confidence of 99 %.

The resulting uniscale SPE and T^2 can be seen in Figure C.12. The confidence limit on the uniscale residuals are also 99 %. The SPE chart indicates that the model is valid throughout the whole period. Consequently, the covariance

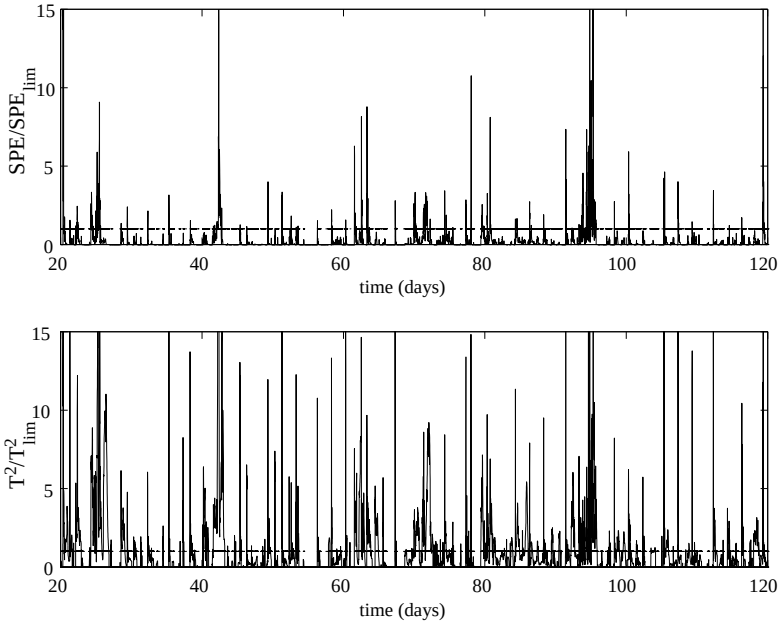

Figure C.12: SPE (top) and T^2 (bottom) from the MSPCA during a time period of 100 days. The confidence limit for both residuals are 99 %. The lowest scale (approximation) is omitted from the analysis.

structure of the reconstructed data does not change significantly. This is because the lowest scale or the approximation scale has been omitted from the analysis. There are some upsets in SPE, but the major deviations can be seen in the T^2.

There is an interesting and important feature of MSPCA, which distinguish it from the other two methods. Consider a step change in one or several variables at a certain point in time. All the methods will detect this step at the same time. However, if the variable or variables return to normal, both the method involving PCA on each scale and the method involving prior recombination of scales, will continue to detect a deviation due to the delay of the filters. This is not the case with MSPCA. When the variables return to normal, the number of significant scales will increase making the confidence level to increase and generally compensate for the delay. This is an appealing feature of the MSPCA.

An example of this is seen after the disturbance at day 42, where the SPE quickly returns to values below its limit.

Discussion

Qualitatively, the three methods presented here have similar performance and they detect the same events equally well. However, there are some differences. Using a PCA model on each scale results in a lot of redundant information making the interpretation cumbersome. This is partly overcome by recombining the scales. If it is possible to discern some dominant time scales of the studied process, the recombination of scales into physically interpretable scales ought to simplify the interpretation considerably. MSPCA's capability to unify all the scales into one scale is desirable and provides a compact way of monitoring. Moreover, MSPCA has an important advantage in that it returns to normal as soon as a disturbance has ended.

It is worth mentioning that the multiscale methods only partly solve the problem with changing process conditions. This is because the changes occur in all scales, and not only the lowest scale. Omitting the lowest scale will not remove changes in the relationship between the variables, i.e. in the covariance structure. Removing the lowest scale is, thus, similar to the method of updating the scaling coefficients, presented in part I.

In part I of this work, we stated that the simplest possible model should be used for monitoring. The multiscale approach includes far more degrees of freedom than the uniscale approach. Is it worth taking this extra step? The decomposition into scales provides a way to discern small changes in data with, for instance, large diurnal variations. Moreover, the multiscale approach provides information on the scale at which a disturbance occurs, which may be used for diagnostic purposes to find the physical cause. However, the advantages come at a price of higher complexity of the monitoring model. The choice of approach used depends on the complexity of the process. A simple process, or perhaps a section of a process, where most events occur in one scale, can be adequately monitored at one scale. In a more complex process, with different subprocesses and different time constants, a multiscale method is more suitable.

An interesting topic not discussed in this paper is related to integration of monitoring and control. Information of the operational state is important in control strategy design and implementation. In, for instance, PCA monitoring the operational state may be determined from the process location in the PC space and loading plots are used to find suitable control actions. In multiscale monitoring one additional piece of information is obtained: time scale information. So in addition to the information on the present operational state, information on how fast the state has changed may be used to determine the strength of corrective measures to drive the process back to desired operation. Integration of time information in the control strategy is one of the more interesting topics for further studies.

Conclusions

This paper has presented a multiscale approach to monitoring of online wastewater treatment measurement data. Multiscale decomposition of data into separate scales is combined with principal component analysis, in order to extract significant features in different scales and to reduce data dimensionality. The advantages of such an approach are an increased sensitivity to small but significant changes and a way to approach the problem of monitoring of data from changing conditions.

After decomposition of data into separate scales, a PCA model is used on each scale to determine whether the operational status is considered normal or not. However, if the number of scales is more than a few, the result becomes cumbersome to interpret using SPC charts. Therefore, the scales can be recombined to represent physically interpretable scales. By doing this, two things are achieved. Firstly, the number of scales that has to be monitored is smaller and secondly, the scales are ideally chosen to match observed time scales in the process, resulting in a more intuitive interpretation.

A more sophisticated way to simplify the interpretation is presented in the multiscale principal component analysis (MSPCA) methodology, first suggested by Bakshi (1998). The methodology involves feature extraction from data on each scale and then recombination using a uniscale PCA.

Applying the methods for multiscale monitoring on wastewater treatment data shows that they are more sensitive to small changes than uniscale monitoring. They can also provide a solution to the problem of monitoring during changing process conditions, since most of the changes occur in lower scales, which are omitted from the monitoring.

Paper C

Addendum

Comments on the MSPCA algorithm

The MSPCA algorithm used in the paper involves a high degree of freedom in the sense that there are many, slightly different, ways to implement it. This may be conceived as a disadvantage, but if the user understands the algorithm, it can be turned into an advantage. The algorithm is flexible and can be adjusted to suit the particular problems at a certain plant. Thus, a more elaborate discussion on the algorithm is needed.

In the paper, the scale PCA models are applied in the decomposed time domain. This requires that the online discrete wavelet transform (ODWT) is used to decompose identification data. The ODWT means retaining the last set of coefficients on each scale from a window of data decomposed using the discrete wavelet transform (DWT) (Lennox; 2001). At each new sampling instance, the window is moved one step and new coefficients are retained. Thus, the number of coefficients at each scale will be the same as the number of samples, i.e. no downsampling. The wavelet coefficients can be transformed to the decomposed time domain by the inverse DWT and the relation between the wavelet and time domain coefficients is expressed as a scaling factor (Nounou and Bakshi; 1999; Lennox; 2001). It is, thus, equivalent to apply the PCA models in the wavelet or time domain. However, as Lennox (2001) has pointed out, it is important to note that the decomposition of variance and covariance is inexact (hence the approximation in Equation C.22).

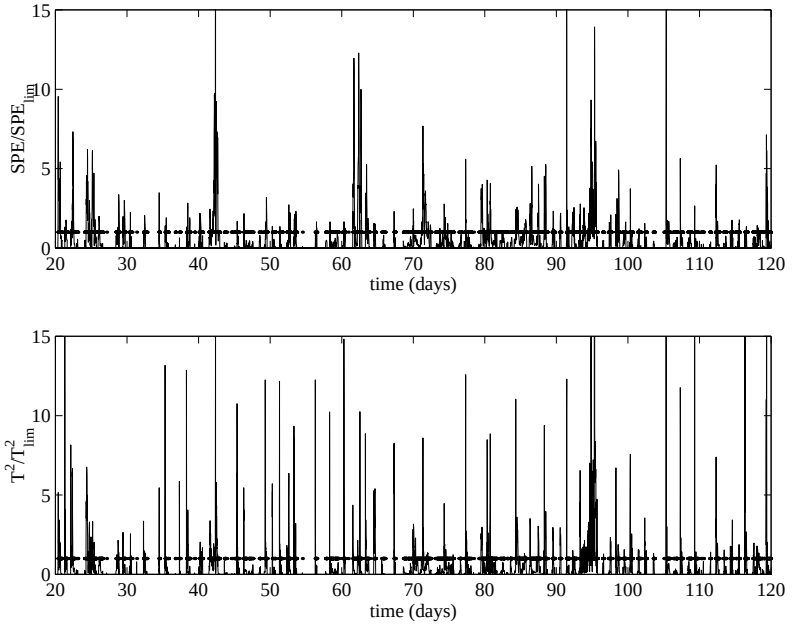

Figure C.13: Results when using $\hat{\mathbf{X}}_j$ for reconstruction of data

However, it is also possible to use the exact relation of Equation C.22. If the models are identified using data decomposed with DTW, the equality will stand. This means that each PCA will be based on coefficient matrices of different size. When creating the covariance matrix for each scale matrix, a compensation factor must be introduced, since the coefficients carry information from original data that consist of higher number of samples. The performance of the two different methods does only differ marginally. In Figure C.13, the result of a DWT-based MSPCA monitoring algorithm is shown. It can be seen that the results are similar to the ODWT-based algorithm.

A second choice involves the number of PCs at each scale. Bakshi (1998) argues that the same number of PCs should be used for each scale. Lennox (2001), however, advocates a choice based on the characteristics of the scale data, and this seems to be a wiser, albeit more cumbersome, choice. Generally, the scale

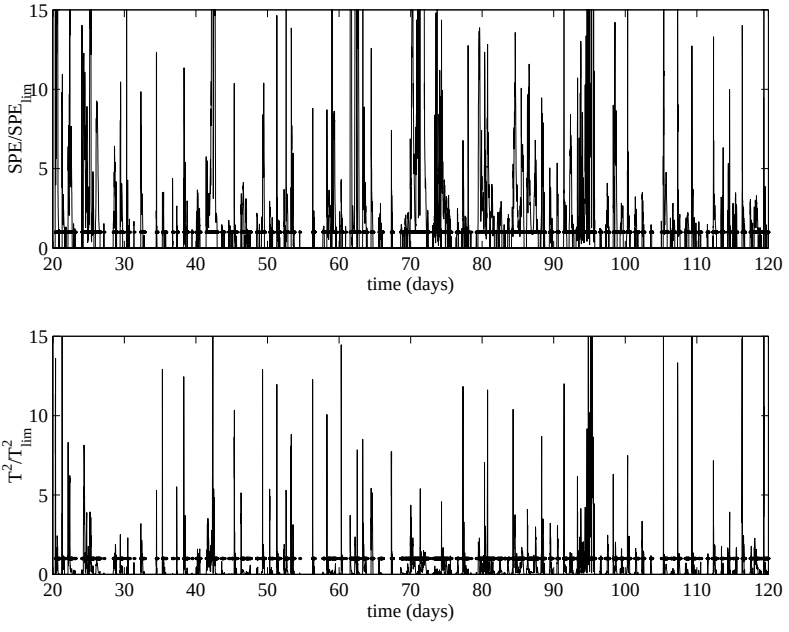

Figure C.14: Results when using $\mathbf{X}_j$ for reconstruction of data. The original data is the same as in Figure C.13.

models are only identified once, so the extra effort may prove advantageous. In the data used for this work, the effect of different choices of PCs is not significant.

Another freedom in the algorithm is related to the reconstruction of the unified data. The scales that have indicated significance are added to create the unified data. However, here we may use both the estimated data of the scale PCAs ($\hat{\mathbf{X}}_j = \mathbf{T}_j\mathbf{P}_j^T$) or the actual scale data ($\mathbf{X}_j$). Using the PCA estimated data will remove all variation not modelled by the particular scale model, and the unified SPE measure will decrease since it will only express the non-modelled variation between the scales. Since this measure has a tendency to be rather high in the unified model, this may be an appropriate 'fix' as long as one is aware of the information lost. This is done in Paper C. In Figure C.14, $\mathbf{X}_j$ was used for the reconstruction. There is an evident spikiness in the SPE that may obstruct the interpretation. However, one should not forget that the figure shows 100

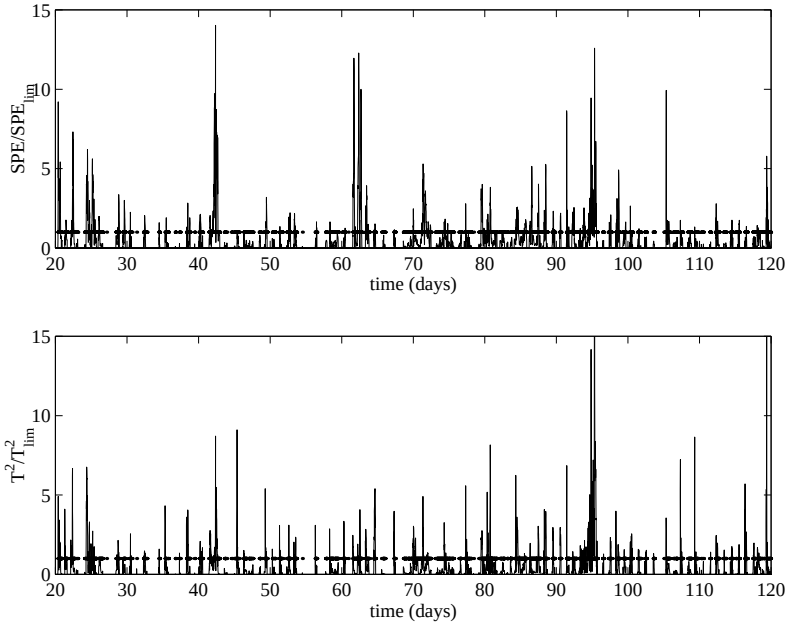

Figure C.15: Results when using $\hat{\mathbf{X}}_j$ in a weighted (values other than 1 or 0 in Equation C.20) reconstruction of data. The original data is the same as in Figure C.13.

days of operation. Using $\mathbf{X}_j$ for reconstruction may be advantageous for short term (day-to-day) monitoring.

The significance weighting ($\gamma_j(k)$ in Equation C.20) is limited to 1 or 0 in the paper. As pointed out by Lennox (2001), the weighting can be refined further. For instance, if it is established that higher scales are too dominant in the unified model, these can be given lower weights. In this way, the unified model is tailored to suit the type of disturbances that are of particular interest. This idea can be taken further, using different weighting for different users. The operators may be interested in small and fast changes, whereas the process engineer's interest may be on a slightly different time scale. The management is normally interested in the long-term effects and has less interest in the short-term changes. The flexibility of the MSPCA algorithm can, consequently, be used to present information suitable for each user. This idea is also applicable on

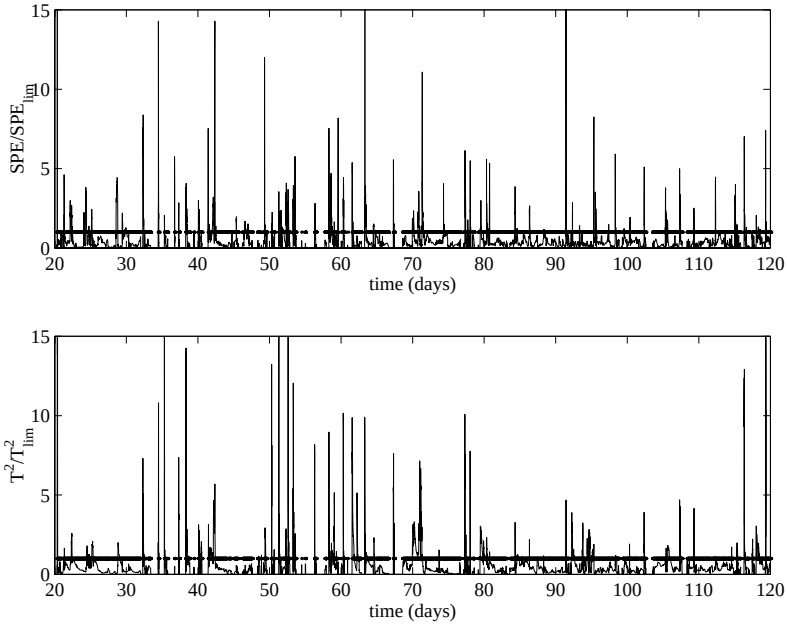

Figure C.16: Monitoring results using a (semi) adaptive MSPCA algorithm. The original data is the same as in Figure C.13.

the other algorithms presented in the paper. A weighted reconstruction using $\hat{\mathbf{X}}_j$ is used to produce Figure C.15. The number and size of the spikes are reduced, especially in the T^2, but the algorithm still detects the same events.

Adaptive MSPCA

In the paper, the lowest scale, or the approximation, was omitted from the algorithm to allow for monitoring of non-stationary data. Another simple extension to the MSPCA algorithm's ability to handle non-stationary data would be to use the 'windowed data', instead of the training data, to identify the unified model. This would yield a moving window adaptive model, similar to what was discussed in Paper B (see Equation B.18). It is important to note that the scale models are not adaptive. However, the wavelet decomposition corresponds to updated scaling parameters (mean value only). If the lowest scale (approxima-

tion) is included, this matrix must be mean centred, since it will not inherently have zero mean. If this is not done, there is a considerable risk that the lowest scale will be significant all the time. A decision that will affect how fast the model adapts is how long the window used for DWT is, since more than only the last coefficients will be used to build the scale covariance matrices. In Figure C.16, the (semi) adaptive MSPCA algorithm is used to monitor the same events as in the previous examples. A window of length 512 samples is used. It is seen that the algorithm has similar detection properties as the algorithms discussed above. However, the spikiness is less dominant, even though $\mathbf{X}_j$ is used instead of $\hat{\mathbf{X}}_j$. The inclusion of the lowest scales can be seen as the slow transients (e.g. at day 42). Compared to the window used for the adaptive PCA in Paper B ($\approx$2000 samples), it should be noted that the window of 512 samples is probably too short.

A more sophisticated adaptive multiscale algorithm is proposed in Paper D. The adaptation is there carried out on each scale, and, thus, it allows for changing covariance structures in all frequencies.

The author would like to emphasise the further investigations carried out by Lennox, based on the work presented in Paper C. In Lennox (2001), a deeper analysis of the MSPCA algorithm is discussed and illustrative examples are given. Interested readers are, thus, referred to the text of Lennox for more information.

Paper D

Adaptive multiscale principal component analysis for online monitoring of wastewater treatment

James Lennox and Christian Rosen

Wat. Sci. Tech. 2001, (accepted).

Abstract: *Fault detection and isolation (FDI) are important steps in the monitoring and supervision of industrial processes. Biological wastewater treatment (WWT) plants are difficult to model, and hence to monitor, because of the complexity of the biological reactions and because plant influent and disturbances are highly variable and/or unmeasured. Multivariate statistical models have been developed for a wide variety of situations over the past few decades, proving successful in many applications. In this paper, we develop a new monitoring algorithm based on Principal Components Analysis (PCA). It can be seen equivalently as making Multiscale PCA (MSPCA) adaptive, or as a multiscale decomposition of adaptive PCA. Adaptive Multiscale PCA (AdMSPCA) exploits the changing multivariate relationships between variables at different time-scales. Adaptation of scale PCA models over time permits them to follow the evolution of the process, inputs or disturbances. Performance of AdMSPCA and adaptive PCA on a real WWT data set is compared and contrasted. The most significant difference observed was the ability of AdMSPCA to adapt to a much wider range of changes. This was mainly due to the flexibility afforded by allowing each scale model to adapt whenever it did not signal an abnormal event at that scale. Relative detec-*

tion speeds were examined only summarily, but seemed to depend on the characteristics of the faults/disturbances. The results of the algorithms were similar for sudden changes, but AdMSPCA appeared more sensitive to slower changes.

Keywords: Fault detection and isolation; multivariate statistical process monitoring; adaptive PCA; multiscale PCA; confidence limits.

Introduction

Multivariate statistical models have become increasingly popular for online process monitoring over the last few decades. Principal Components Analysis (PCA) is the simplest of them, modelling static covariance relationships in a lower-dimensional subspace. While this simplicity is appealing, ordinary PCA models are often insufficient when dealing with the complexities of real data. In biological wastewater treatment (WWT) plants, influent flowrate and composition are often highly non-stationary, varying on time-scales ranging from hours (e.g. toxic shocks) to months (seasonal effects). The process itself evolves over time, as the biomass adapts to different conditions. Highly nonlinear biological reactions make linearized relationships between variables dependent on the operating point. Finally, the wide range of dynamics of the biological and physical processes involved makes it difficult to look at correlations on a single time-scale.

Adaptive PCA (Wold; 1994; Li et al.; 2000) updates the PCA model online as new data are obtained, allowing changes in the measurements' mean, variance and correlation to be followed. The model will adapt not only to process evolution but also to dynamics or nonlinearities on time-scales slower than that of the adaptation (i.e. the model is locally linear). Gallagher and Wise (1997) show adaptivity to be of great benefit to long-term monitoring performance for a manufacturing process. Multiscale PCA (MSPCA) (Bakshi; 1998) uses a wavelet transform to decompose measurement data into different time-scales. A separate PCA model is used to monitor each scale. Scale monitoring is used to adaptively pre-filter the data, which are then monitored by a single PCA model based on only the most important scales at a given time. MSPCA is especially useful for processes such as WWT that have a wide dynamic range. MSPCA differs from the earlier combination of wavelet transforms and PCA in Kosanovich and Piovoso (1997).

Rosen and Lennox (2001) apply both Adaptive PCA and MSPCA to WWT data and develop a modified MSPCA algorithm to accommodate mean non-stationarity. In this paper, adaptivity and multiscale decomposition are combined to create a new method: Adaptive Multiscale PCA (AdMSPCA). It is hoped to overcome individual limitations of Adaptive PCA and MSPCA in the context of WWT monitoring (Rosen and Lennox; 2001). AdMSPCA makes the scale models of MSPCA adaptive and the unifying PCA model implicitly adaptive, since it is constructed from the adaptive scale covariance matrices. Performance of AdMSPCA and adaptive PCA on a real WWT data set is compared and contrasted.

Monitoring algorithms

PCA monitoring

Process measurements are usually cross-correlated, with an effective dimension less than the number of variables (m). Consequently, univariate monitoring is inefficient and can even be misleading. PCA monitoring exploits cross-correlation and redundancy, identifying a multivariate 'process subspace' containing significant (i.e. non-random) variation. In standard PCA, identification is performed off-line using historical data representing normal operation. The first a principal components (PCs) in a PCA model define the process subspace. The remaining $m - a$ components define a complementary 'noise subspace'.

New data can be projected onto each subspace, yielding their 'process' and 'noise' components, yielding two summarising statistics (Jackson and Mudholkar; 1979; Kresta et al.; 1991). The T^2 statistic is a (weighted) squared distance from the multivariate mean on the model hyperplane. The squared prediction error (SPE) is the squared distance of a measurement from the model hyperplane. Statistical confidence limits can be defined for both quantities (Jackson; 1980). Useful review and tutorial articles on PCA monitoring are Wise and Gallagher (1996b) and Kourti and MacGregor (1995). SPE or T^2 alone do not give information concerning the cause/s of abnormal operation. Contribution plots (MacGregor et al.; 1994) are useful diagnostic tools, isolating individual variables making an important contribution to large SPE or T^2 values.

Adaptive PCA

In the earliest adaptive PCA algorithm, the model is identified online from an exponentially-weighted moving data window (Wold; 1994). Recursive PCA (Li et al.; 2000) is a more computationally efficient approach, where the covariance matrix, or even the PCA model, is directly updated with each new sample. In this work, the PCA model is computed by singular value decomposition (SVD) of the recursive covariance matrix R_k. Although covariance-based recursive PCA is used, constant variance scaling is first applied to each variable, using variances calculated from initial identification data. The additional complication of adaptive variances is generally unnecessary in practice, and precludes the theoretical derivation of confidence limits (Li et al.; 2000). The number of PCs in the adaptive model is fixed in the initial identification step, using the variance of the reconstruction error (VRE) method (Qin and Dunia; 2000). It could also be determined online (Li et al.; 2000) using this, or any other method. Making the model dimension adaptive increases flexibility to model process changes, but also increases online computation and makes the results harder to interpret when the number of PCs changes. Here, a fixed number of PCs also made comparison between Adaptive PCA and AdMSPCA simpler.

The following recursive equations are applied after each sample (i.e. $nk = 1$), followed by SVD of R_k:

$$\mathbf{b}_k = \mu \mathbf{b}_{k-n_k} + (1-\mu)\, 1_{n_k}^T \mathbf{X}_k \tag{D.1}$$

$$\hat{\mathbf{X}}_k = \mathbf{X}_k - 1_{n_k} \mathbf{b}_k \tag{D.2}$$

$$\Delta \mathbf{b}_k = \mathbf{b}_k - \mathbf{b}_{k-n_k} \tag{D.3}$$

$$\mathbf{R}_k = \mu \left(\mathbf{R}_{k-n_k} + \Delta \mathbf{b}_k^T \Delta \mathbf{b}_k \right) + (1-\mu)\, \hat{\mathbf{X}}_k^T \hat{\mathbf{X}}_k \tag{D.4}$$

where $\mathbf{X}_k$ is the kth raw data block ($\mathbf{X}_k \in \Re^{n_k \times m}$), $\mathbf{b}_k$ is the adaptive mean and $\mathbf{R}_k$ is the recursive covariance at sample $N_k = \sum_{i=1}^{k} n_i$. The forgetting factor $0 < \mu \leq 1$ defines the weight given to past observations. It relates to the effective window length (defining a time-scale of adaptation) as:

$$N_{eff} = \frac{n_k}{(1-\mu)} \tag{D.5}$$

A practical difficulty with adaptive models is that they may adapt not only to normal process evolution, but also to abnormal changes. To prevent this, updating can be paused whenever the SPE is exceeding a certain limit (Li et al.; 2000). Changes that ought to be detected in the first instance, but later accepted as normal, are harder to deal with. Such model 'resetting' requires information additional to that contained in the sensor data, and is beyond the scope of the current work.

MSPCA

MSPCA is described in detail in Bakshi (1998). We summarise the algorithm here. Each variable is first decomposed into a number of scales using an on-line wavelet transform (Bakshi; 1998). Here the Haar wavelet is used. It is the simplest wavelet and has the shape of a step. The transform splits the signal into J *details* and an *approximation*. Details are simply successive band-pass versions of the signal, while the approximation is a low-pass version. A defining characteristic of wavelets is that adding up detail coefficients from every detail scale, plus the low scale approximation coefficients, recovers the original signal. Another defining characteristic is that the band-width of the scales is proportional to their frequency. Not only can scale coefficients be added, but their covariance matrices can be also added to give the data covariance matrix (Bakshi; 1998). Note that while 'high scales' will be taken to mean those with the highest frequency content, scales will be numbered form 1 to J with 1 being the highest. Figure D.1 shows the temperature time series (see Table D.1) and its decomposition into five scales.

Separate PCA models are identified for the detail coefficients of each scale and the approximation coefficients of scale J. Each scale is then monitored independently using SPE and/or T^2. The scale monitoring results are not used directly. This would increase the number of entities to monitor and could also lead to confusion, since the same event can appear at different times in different scales, because of different filtering delays. The scale monitoring is instead used as a criterion for reconstructing the signals and the covariance matrix online, including only significant scales at each instant. Scales are deemed significant whenever their SPE or T^2 exceeds its confidence limit.

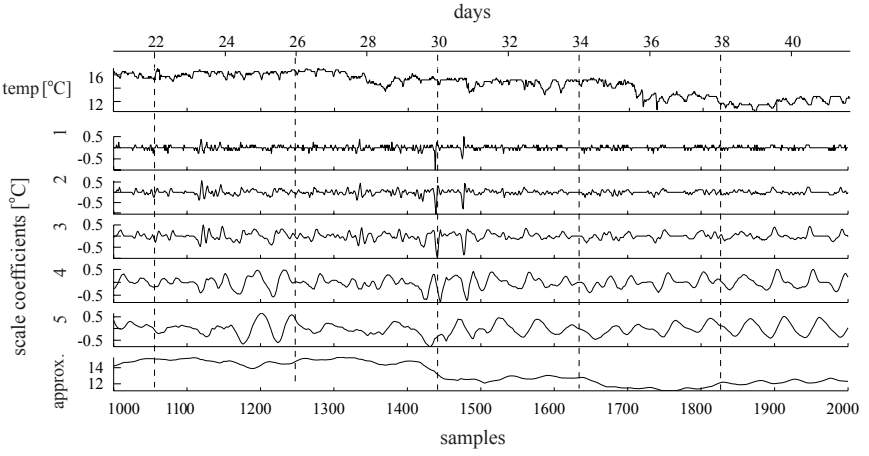

Figure D.1: Decomposition of the temperature time series by the online Haar wavelet transform with $J = 5$.

A final 'unifying' PCA model is identified from the reconstructed covariance matrix. The reconstructed sample is then projected onto this model and its SPE and/or T^2 monitored. Figure D.2 illustrates the steps in the MSPCA algorithm. Because the online wavelet representation is redundant ($J+1$ times as many coefficients as samples), scale limits must be adjusted as:

$$p_j \quad = \quad \sqrt[J+1]{p_R} \tag{D.6}$$

Unlike the limit of (Bakshi; 1998), Equation D.6 accounts for the probability of simultaneous detections at different scales—i.e. it is the probability of no detection at any scale. The process of reconstruction tends to cancel out the different detection delays across scales (Bakshi; 1998). Sometimes though, the unified SPE will have a 'spiky' appearance, as because of the inclusion and exclusion of scale over the course of an event.

Adaptive MSPCA (AdMSPCA)

AdMSPCA can be seen in two ways: either as making MSPCA scale models adaptive or as a multiscale decomposition of recursive PCA. For each variable, the order of wavelet transformation and recursive updating operations is interchangeable (i.e. Equations D.1-D.3 can be applied to the wavelet coefficients

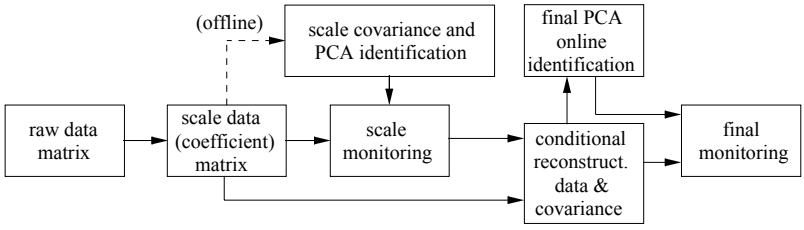

Figure D.2: Schematic of MSPCA identification and monitoring.

at each scale, or to the data directly). In the latter case, $\Delta\mathbf{b}_k$ and $\hat{\mathbf{X}}_k$ must also be decomposed into scales. Equation D.4 is then applied to recursively identify the covariance matrix for each scale in exact analogy to the scale PCA model identification in MSPCA.

In adaptive PCA, the adaptation rate implicitly defines a single time-scale of interest. Identification of adaptation parameters for each scale is impractical, but the problem of a single time-scale can be overcome by relating adaptation parameters at different scales. It is often reasonable to assume that high-frequency phenomena involve faster changes than low-frequency ones. In any case, sampling theory tells us that more samples are required to identify lower frequencies than higher ones. This suggests making the effective window length (N_{eff}) proportional to the length of the wavelet at that scale (i.e. doubling each scale).

As with all PCA-based models, it is important to correctly identify the number of PCs to retain. MSPCA permits a different number of PCs at each scale while recursive PCA permits an adaptive determination of the number of PCs. Both these options could be implemented in AdMSPCA, but here a fixed number of PCs is used at each scale—both for simplicity and ease of comparison with adaptive PCA. Equation D.6 is used to find the p-value (p_j) for each scale model, giving the desired overall p-value (p_R). The confidence limit of the unifying PCA model is calculated as if it were an ordinary PCA model, except that the degrees of freedom is defined empirically as the average of the window-lengths of the reconstructed scales.

An updating rule such as that described above can be applied at each scale. This is not possible if the recursive Equations D.1-D.3 are applied first and the wave-

No.	Variable	No.	Variable
1	Infl Temp (C)	7	Infl Ammonia (mg/L)
2	Air A1 (% opening)	8	Infl pH
3	Air B1 (% opening)	9	Infl Flow m^3/d)
4	Air A2 (% opening)	10 Biol Effl pH	
5	Air B2 (% opening)	11	Plant Effl Turbidity
6	Infl Conductivity	12	Plant Effl pH

Abbreviations: Infl = influent; Effl = effluent;Biol = biological;
Temp = temperature; Cond = conductivity; Turb = turbidity

Table D.1: List of variables used from Ronneby dataset.

let transform and Equation D.4 second. Updating control introduces a nonlinear element, destroying the equivalence of the two approaches to AdMSPCA. Here, the wavelet transform is taken first, and then the recursion equations are applied at each scale. Since the scale details by definition have an expected mean of zero, mean-centring is applied only on the low-scale approximation. The updating rule cannot be based on the unifying PCA SPE, because the same event may appear at different times on different scales. Controlling updating of individual scale models affords great flexibility, since the band-pass nature of the detail scales means that most events cause transient rather than persistent exceedences of the confidence limit, as is observed in the case study.

Application to WWT data

Case study

The data are from the Ronneby WWT plant in Sweden (Rosen; 1998a). The selected variables are listed in Table D.1. Obvious features of the data include diurnal patterns in most variables and frequent abnormalities in single or multiple variables. The original data were decimated from 288/d to 48/d. This was to simulate sampling at the latter rate, where the mostly uninformative higher frequencies are not present. In practice, down-sampling should be performed with appropriate pre-filtering to avoid either aliasing effects or destroying multivariate relationships. Nonlinear filters may be useful to satisfy the latter requirement.

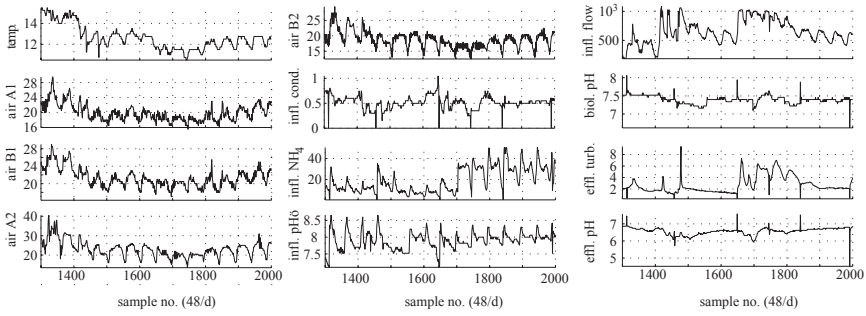

Figure D.3: 14.6 days of the raw data record (samples 1300-2000).

Two distinct periods can be seen in the raw data (see Figure D.3). Up to sample 1400 (samples 1-1200 are not illustrated) the data are quite normal. Unusual events appear to be of short or moderate duration. A second period can be defined beginning at the major qualitative changes in most variables around 1400. The temperature and flowrate variables become significantly lower and higher respectively, suggesting a prolonged influx of melt-water. Other major changes are seen later on, for example in effluent turbidity from 1650. We concentrate on the transition to the second period and some events within it.

Five scales were used in AdMSPCA, the fifth scale detail containing frequencies around 1/d. The adaptation rate was chosen to give an effective window-length of three days (144 samples). This was also the window-length used for the fifth scale detail and approximation of AdMSPCA. As reported also by Gallagher and Wise (1997), the results were relatively insensitive to the window-length. Difference between N_{eff} = 3d and N_{eff} = 4d were barely noticeable, while N_{eff} = 7d gave mainly quantitative differences except when updating was halted. Note that the algorithms were used in their basic form as outlined in previous sections, with a simple threshold on SPE (and arbitrary 1.2 times the 95 % limit) to halt updating.

Results

Many significant features appeared in the SPEs of the two algorithms. Features were sometimes qualitatively similar and/or simultaneous and sometimes not. Sharp peaks and spikes in particular tended to be detected similarly by each al-

gorithm, whereas many (but not all) gradual changes were detected differently—or sometimes by only one of the algorithms. Differences were particularly marked in the latter half (from 1646 to 1912) of the illustrated data, where adaptive PCA had continuously large SPE. Over this same period AdMSPCA gave many significant detections, but these were broken up into a large number of peaks and sub-peaks, separated by several significant periods of non-detection. For example where adaptive PCA gave a major peak around 1765-1770, there was no detection at all by AdMSPCA. Where there were such major differences, it can be assumed that the algorithms were detecting—and hence probably modelling—quite different things. Without complementary information about the data set, it cannot be directly concluded that either method performed better or worse. The results must be examined in detail to try and determine how the two algorithms differ and their relative merits.

Qualitative and quantitative comparisons of detections are facilitated by plotting detections as points along an axis (Figure D.4 - bottom). A notable qualitative feature of the plot is that while the SPE of AdMSPCA generally appears more variable than that of adaptive PCA, there were times when AdMSPCA gave longer periods of detection. Quantitatively, it is important that the algorithms detect events as quickly as possible. While the data used in this work did not lend themselves to a detailed or objective study of detection speeds, the results can nevertheless give indications about performance differences. Returning to the above-mentioned period 1646-1912: AdMSPCA first detected seven samples before adaptive PCA (and continuous detection began eight samples earlier). However, the algorithms simultaneously detected the spike at the very beginning (1649) of the period. The large peaks from adaptive PCA with maxima at 1662 and 1712 have clear counterparts in the SPE of AdMSPCA, as does the spike at 1745. The latter was delayed by one sample with AdMSPCA. Some other peaks of adaptive PCA had zero SPE with AdMSPCA. These observations suggest that large, sudden events are detected with equal delay, whereas gradual changes are detected with differences in delay that depend on the nature of the changes.

Beyond 1726, AdMSPCA briefly returned to normal before two short peaks (1736-1762) that correspond to a trough in adaptive PCA (although the latter remains above the confidence limit). The next cluster of AdMSPCA peaks corresponds to the beginning of a broader adaptive PCA peak. It is seen that

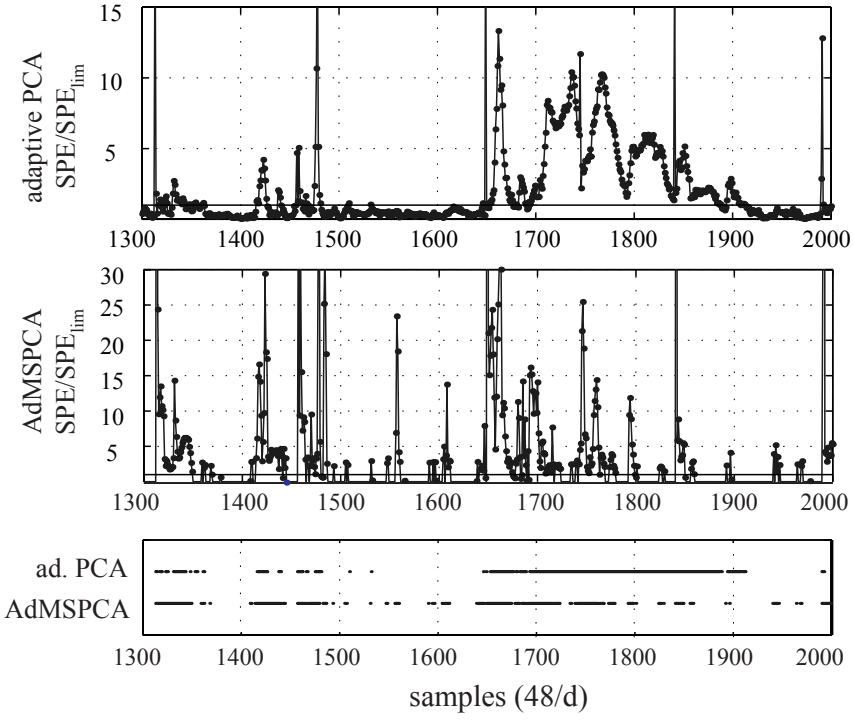

Figure D.4: Adaptive PCA monitoring (top); AdMSPCA monitoring (middle); and detection location plot (bottom).

maxima of AdMSPCA tend to match sudden increases *or decreases* in the SPE of adaptive PCA. This is intuitively reasonable, because the Haar wavelet coefficients are basically differences of the data. The large and sustained peaks of adaptive PCA seem to mirror periodic patterns in the data themselves. Given that the adaptive PCA updating is paused during most of this period, it is likely that significant model mismatch develops. Consequently, the adaptive PCA algorithm cannot necessarily distinguish between normal and abnormal changes at this point. AdMSPCA does not suffer from this problem, because the individual adaptation of scale models and the faster adaptation rates at higher scales make it considerably more flexible. The shorter peaks in AdMSPCA signal rapid changes in conditions that the scale models then manage to adapt to. A less generous interpretation is that AdMSPCA involves broader definition of 'normal'.

Scale SPEs (Figure D.5) can help determine the nature of an event. Note that the low-scale approximation SPE is omitted from Figure D.5, since there were no detections in this period. High scales are sensitive to sudden changes and can more quickly adapt if the covariance structure of their coefficients alters. Low scales are sensitive to slower changes and will react gradually to covariance changes. As a general rule, features (but not necessarily detections) at higher scales will be shorter and those at lower scales longer. While short features at a low scale are unlikely given the smoothness of their wavelet coefficients, long features at high scales can be caused by a (permanent or continuing) change in covariance structure not followed by the scale model. This theory is borne out by the features observed at different scales in the case study.

Features at scales 1-3 generally appear similar, except that their length tends to increase with scale. The incidence of scale detections (Figure D.5, bottom) shows a pattern of detections occurring slightly later at lower scales, consistent with the longer filtering delays. By scale five, the peaks in SPE are distinctly fewer, longer and smoother. As previously noted, no detections occurred at the low-scale approximation. Comparing the size of features between scales can give useful information about events. Around 1760 for example, there was a detection almost entirely restricted to scale three (frequencies around 5/d). This was therefore a slower change and/or one that did not cause a departure from the correlation structure at the higher scales. Just beforehand, there was a larger detection involving all detail scales. However, the highest scales were

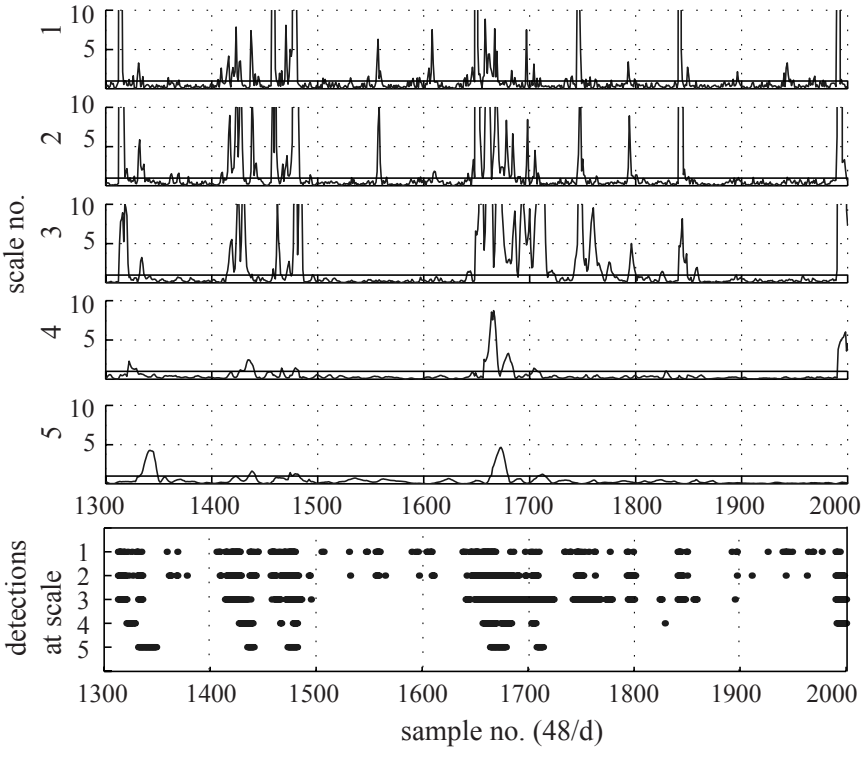

Figure D.5: AdMSPCA detail scale SPEs (axes 1-5) and detection location plot for detail scales (bottom).

dominated by a very sharp peak (1646), whereas scale four had a flat peak. This suggests the superposition of a sharp change with a slower one, the latter perhaps continuing into the smaller peak around 1760.

Using more complicated PCA-based methods, it can be difficult to establish a relationship between generated features and the patterns in the raw data—and ultimately with the underlying physical phenomena. Contribution plots can be generated for both algorithms and are a useful tool for isolating significant variables. It should be remembered that in periods where there is a continuing high SPE, isolation results may be of limited value as they rely on an invalid model. This is the case for a large section of the adaptive PCA results. Variable 'contributions' to SPE are the individual squared deviations of each variable from the model hyperplane. The contribution plots here are modified to normalise the proportion of each variable falling in the noise subspace. This is useful when isolation is performed based on SPE alone (without T^2).

Contributions were calculated for two samples corresponding to SPE peaks of AdMSPCA (Figure D.6) . The earlier time was chosen to avoid the large spike described above. Such gross phenomena are generally evident in the raw data so sophisticated isolation procedures are of less interest. The contributions illustrated should then pertain to slower underlying variations, visible in the lower scales of AdMSPCA. Both methods show that Infl NH4 (7) and Turb (11) were significant at 1750, then Turb and Effl pH (12) at 1760. Adaptive PCA additionally showed Infl pH (8) and Flow (9) as significant at both times, although pH only marginally so at 1760. AdMSPCA showed no additional variables to be significant at either time. Inspection of the raw data suggests that the change in Effl pH does not occur until after 1750, as suggested by AdMSPCA. On the other hand, there is a large, step-like change in both Flow and Infl pH at 1750. These correspond better to the results of adaptive PCA than of AdMSPCA. The regular periodic pattern in Flow does not recommence until some time later, so it is reasonable that adaptive PCA still indicates it to be important at 1760. The importance of Infl Ammonia is not clear from examination of the variable in isolation; however, it is likely that there is a genuine reason for this, given both methods isolated it. Such problems of verifying results are common when using these multivariate monitoring methods to real data, about which little is known. Simulations or detailed, well-equipped pilot studies are often the only realistic means of acquiring appropriate data for comprehensive testing and verification.

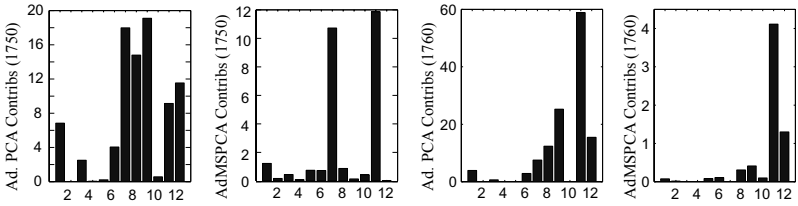

Figure D.6: Weighted contributions (see text) to adaptive PCA and AdMSPCA SPEs at 1750 (left) and 1760 (right).

Conclusions

A new algorithm, AdMSPCA, was developed and applied to WWT measurements. It combines elements of adaptive PCA and MSPCA, in order to monitor evolving processes with a wide range of dynamics. Comparison of detection speeds gave mixed results: AdMSPCA was distinctly faster in one case involving a slower change, but the results were similar for cases involving faster changes or spikes. The band-pass nature of the wavelet detail coefficients was reflected in the tendency of AdMSPCA peaks to match sharp changes in adaptive PCA SPE. Major qualitative differences between the SPEs of the two algorithms were attributed to the updating strategy used. Unlike the adaptive PCA model, AdMSPCA scale models, did not persistently exceed their updating SPE limits. Significantly different (and perhaps improved) results might therefore be obtained for either algorithm using more sophisticated updating rules.

While AdMSPCA still provides concise overall SPE and T^2 statistics for detection, isolation can be more complicated than for adaptive PCA, especially if contributions by scale and variable are considered simultaneously. A difficulty common to both algorithms is that changing model directions can diminish interpretability. However, exploiting these changes for detection and/or isolation might be an interesting direction for future research. Only SPE (not T^2) was used as a detection criterion here because of periodic variation within the process subspace. This reduced isolation potential, since information was not available in the process subspace directions. To avoid biasing isolation towards variables contributing mainly to the noise subspace, variables' contributions were weighted according to their projections on the process and noise sub-

spaces. Future research might consider using only SPE for detection, but giving T^2 a complementary—possibly qualitative—role in isolation. Non-parametric confidence limits on scores and residuals are a more sophisticated, but also more complicated, alternative.

Paper D

Addendum

As was pointed out in Chapter 1, the author's contribution to this paper is mainly in the idea stage of the work. Thus, implementation, analysis and writing are attributed to Lennox (Lennox and Rosen; 2001; Lennox; 2001) with support from the author. However, the paper fits well into the framework of the thesis and constitutes an 'ultimate' algorithm for multivariate wastewater treatment monitoring (nonlinear relations excepted). Whether it is worth going this far remains to see. The relatively high complexity of the algorithm must be weighed against the possible performance improvements.

Paper E

Supervisory control of wastewater treatment plants by combining principal component analysis and fyzzy c-means clustering

C. Rosen and Z. Yuan

Wat. Sci. Tech. **43**(7): 147-156, 2000

Abstract: *In this paper a methodology for integrated multivariate monitoring and control of biological wastewater treatment plants during extreme events is presented. To monitor the process, online dynamic principal component analysis (PCA) is performed on the process data to extract the principal components that represent the underlying mechanisms of the process. Fuzzy c-means (FCM) clustering is used to classify the operational state. Performing clustering on scores from PCA solves computational problems as well as increases robustness due to noise attenuation. The class-membership information from FCM is used to derive adequate control setpoints for the local control loops. The methodology is illustrated by a simulation study of a biological wastewater treatment plant, on which disturbances of various types are imposed. The results show that the methodology can be used to determine and coordinate control actions in order to shift the control objective and improve the effluent quality.*

Keywords: Fuzzy clustering; multivariate monitoring; PCA; setpoint control; supervisory control; wastewater treatment.

Introduction

In this paper an approach to an integrated multivariate monitoring and control system for wastewater treatment operation during extreme events (disturbances) is proposed. The method is based on multivariate statistics combined with clustering analysis to determine the operational state of the process. Information on the operational state is then used to determine appropriate setpoints for local control loops. The methodology is illustrated by a simulation study of a biological wastewater treatment plant on which disturbances of various types are imposed.

Automatic process control is today an important part of the operation of most biological wastewater treatments plants. The dissolved oxygen concentration is an example where automatic control has been successfully applied (Olsson and Newell; 1999). A local control system is typically concerned with a sub-process (called a unit process) of the whole system. Normally, these unit processes are controlled using a local feedback loop where the output of the process is measured and compared with a certain setpoint from which an appropriate control action is derived. Feed-forward control can also be utilised. Here, the control action is derived from a model describing the dependency of a process variable on a manipulative variable. An example is flow-rate proportional control of the return sludge flow rate.

The low-level control constituted by feedback and feed-forward control is usually sufficient under normal conditions when the characteristics of the influent wastewater are reasonably constant. However, as the operational conditions change, the control setpoints often have to be changed accordingly to obtain the desired operation. There are some different reasons why a supervisory control level is needed. Firstly, since the process is broken down into unit processes, there is a need to coordinate different control actions so that they do not have a counteractive effect. Secondly, a process may display nonlinear behaviour when the operational conditions are far from the normal operating point, requiring changes to control set points. Thirdly, during extreme operational conditions such as hydraulic shocks or toxicity, the aim of the operation may shift significantly. Thus, a higher-level control system is needed to determine the control setpoints or control structure of the low-level control systems. This level is often referred to as the supervisory control level.

Supervisory control of wastewater treatment plants is typically performed by operators of the plants. That is, they take the responsibility for changing the control setpoints based on the information extracted from the online instrumentation data by an automatic monitoring algorithm or simply by directly visualising the process data. A natural step towards higher degree of automation in wastewater treatment plants is to close the loop by an algorithm, which automatically feeds the control setpoints to the local control loops. This has not been feasible in the past, but as the number of measured entities increase, the measurement quality improves and monitoring tools become more sophisticated, it is time for a discussion on how automatic supervisory control can be implemented in wastewater treatment.

Multivariate monitoring and classification

The objectives of process monitoring are to gather data and extract useful information from the measurements. Process data can be monitored using a wide range of different monitoring tools. Stochastic process control (SPC)[1] (see e.g. Bissel (1994)) is a set of commonly used tools. In SPC, each variable is presented as a time series to the operator and control and alarm limits are used to define normal and abnormal operation. Multivariate techniques have become popular within many industrial fields as they can account for collective effects and reduce the dimensionality of the monitored data (see e.g. Wise and Gallagher (1996b)). Principal component analysis (PCA) and other multivariate statistics (MVS) based techniques have proven useful for monitoring of wastewater treatment operation (Rosen and Olsson; 1998; Teppola et al.; 1998a).

The information obtained in the monitoring phase can be used to classify the current operational state. Knowledge based system or expert systems have been used with varying success (Davis et al.; 1996). Clustering techniques represent a different approach. Here, clustered data in the measurement space are said to represent similar process behaviour. Fuzzy c-means (FCM) clustering has been used to recognise clusters in wastewater treatment data (Marsili-Libelli and Müller; 1996). A combined approach of multivariate statistics and clustering to wastewater data monitoring can be found in Teppola et al. (1998a). The approach presented here goes one step further. PCA and FCM are combined

[1] Should be statistical process control.

to determine the operational state of the process. The fuzzy information is then used to derive appropriate setpoints for local control loops in the process.

Principal component analysis

PCA can be described as a method to project a highly dimensional measurement space onto a space with significantly fewer dimensions. Often, several variables are highly correlated, since most variables only reflect a few underlying mechanisms that drive the process in different ways. This correlation is used in PCA to represent the underlying mechanisms as principal components (PCs). Let $\mathbf{X}$ be an autoscaled (i.e. mean-centred and scaled to unit variance) $[m \times n]$ matrix of measurement values for n variables at m number of samples defining a variable space of r dimensions. The r-dimensional matrix $\mathbf{X}$ can be decomposed into a sum of the outer product of vectors $\mathbf{t}$ (scores) and $\mathbf{p}$ (loadings):

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_a\mathbf{p}_aT + \mathbf{E}$$

or

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{E.1}$$

where $\mathbf{E}$ is the residual matrix and ar. If $a = r$ then $\mathbf{E} = 0$, as all variability is described. However, if $a < r$, i.e. less PCs than original variables are retained, then $\mathbf{E}$ describes the variability not described by $\mathbf{T}\mathbf{P}^T$. Ideally, when a is chosen adequately, $\mathbf{T}\mathbf{P}^T$ describes the underlying mechanisms and $\mathbf{E}$ represents the noise in matrix $\mathbf{X}$. Often, in industrial systems $a << r$, which implies a significant reduction of the number of dimensions (Wise and Gallagher; 1996b).

PCA in its simplest form is a static modelling technique. However, there are ways to incorporate dynamics in the model, by including old measurement values in the analysis. The matrix $\mathbf{X}$ is then constructed as:

$$\mathbf{X} = [\mathbf{X}_k\ \mathbf{X}_{k-1}\ \ldots\ \mathbf{X}_{k-l}] \tag{E.2}$$

where $\mathbf{X}_{k-l}$ means matrix $\mathbf{X}_k$ translated l steps back in time. In this way, dynamic relationships between variables can be modelled.

The basic way of using PCA for monitoring involves identification of a model from data representing normal or desired operation. New data is then projected onto the model and the scores and/or the model residuals are then monitored as new samples are obtained:

$$\mathbf{t}_k = \mathbf{x}_k\mathbf{P} \tag{E.3}$$

It is important to note that new data are scaled in the same manner as data used for identification. Other available MVS-based methods include adaptive PCA (Wold; 1994; Dayal and MacGregor; 1997a), principal component regression (PCR) and projection to latent structures (PLS) (see e.g. Wise and Gallagher (1996b)).

Fuzzy c-means clustering

Fuzzy c-means (FCM) clustering is a method that allows a certain instance to be member of several classes at the same time, i.e. it is possible to be between two or more classes. Assuming that the cluster centres are known, the membership u to a certain cluster (class) i of an instance at time k can be calculated by:

$$u_{k,i} = \left(\sum_{j=1}^{C} \left(\frac{d_{k,i}}{d_{k,j}} \right)^{\frac{2}{m-1}} \right)^{-1} \tag{E.4}$$

where C is the number of classes and $\mathbf{d}_{k,i}$ and $\mathbf{d}_{k,j}$ are the distances from the instance to the centres of clusters i and j, respectively. The parameter m determines the fuzziness of the classification. An m that is close to unity yields a crisp classification, whereas an increasing m makes the classification fuzzy. The distances can be calculated as:

$$d_{k,i}^2 = (\mathbf{t}_k - v_i)(\mathbf{t}_k - v_i)^T \tag{E.5}$$

where $\mathbf{t}_k$ is the instance and v_i is the cluster centre. The cluster centre can be determined manually. However, FCM is an unsupervised method, i.e. it can be used to find clusters in data. This is done iteratively for all instances N using (E.6) to find a minimum:

$$J_m(C, m) = \sum_{i=1}^{C} \sum_{k=1}^{N} (u_{k,i})^m \, d_{k,i}^2 \tag{E.6}$$

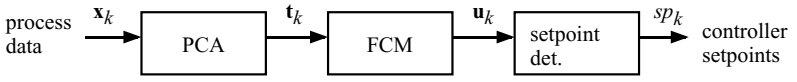

Figure E.1: Supervisory control by combining PCA, FCM and setpoint determination.

The cluster centres are calculated from:

$$v_i = \frac{\sum_{k=1}^{N} (u_{k,i})^m \mathbf{t}_k}{\sum_{k=1}^{N} (u_{k,i})^m} \tag{E.7}$$

The algorithm described above assumes that process data do not change significantly during the period of interest. Algorithms for adaptive FCM can be found in the literature (Marsili-Libelli and Müller; 1996; Teppola et al.; 1998a)and will not be discussed here.

Supervisory control by combining PCA and FCM

In this paper an approach to automatically determine controller set points (supervisory control) by means of combining PCA and FCM is proposed. This means that measurement data are projected as scores onto a smaller space defined by the principal components. Then, clustering analysis is carried out to locate the process in this space and, thus, determine the current operational state. This procedure has some advantages. Firstly, a reduced space through PCA implies reduced computational time for clustering analysis since the dimensionality of the problem is reduced. This is particularly beneficial for complex problems. Secondly, only variations represented by the model is projected as scores. Thus, measurement and process noise are, at least in the ideal case, not present in the scores. This means less misclassifications and more robust setpoint determination. Thirdly, the convergence of the clustering algorithm is improved, as the scores are orthogonal. Teppola et al. (1998a) have reported convergence problem in the clustering algorithm with highly collinear data. The supervisory control scheme is shown in Figure E.1.

To translate the class membership to a control output or a setpoint, the classification need to be defuzzified. The centre of area method is adopted here. The setpoint corresponding to a certain class is multiplied with the membership function for that class:

$$sp_k = \frac{\sum_i u_{k,i} sp_i}{\sum_i u_{k,i}} \tag{E.8}$$

Thus, the final setpoint, sp_k, is calculated from the setpoints for all classes. Compared to the simple approach in which the class with the highest membership function is chosen, the centre of area approach utilises the fuzzy information on the class memberships. It is worth noting that when FCM is used, the denominator in Equation E.8 is always unity. The setpoint for each class, sp_i can be predefined, which is the case in this work, or derived using models containing process knowledge.

Case study: extreme event control

A simulation case study with application to wastewater treatment operation is reported here. In this study, a supervisory control component is designed to coordinate a couple of local control loops. The control objective is to protect the process under extreme conditions and, whenever possible, improve the effluent quality. The primary objectives of the study are to illustrate the proposed methodology and to evaluate its applicability to wastewater treatment operation.

Simulated plant

The simulated plant comprises two biological reactors and a secondary settler (Figure E.2). Both reactors are aerated and, consequently, only carbon reduction and nitrification are of interest. The controlled variables considered in the design of the supervisory control system include dissolved oxygen (DO)-concentrations in both reactors, influent feed ratio between the two reactors and sludge-blanket level in the settler. The latter is controlled by manipulating the return-sludge flow rate. The IAWQ Activated Sludge Model No 1 (Henze et al.; 1987) and a ten-layer one-dimensional settler model (Takács et al.; 1991), are used to simulate the biological reactions and the settling process, respectively. Influent data developed by a working group on benchmarking of wastewater

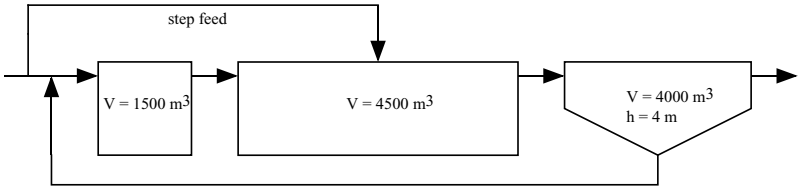

Figure E.2: Principal layout of the simulated plant.

treatment plants within the European scientific research exchange programme COST 624 are used in the simulation (Vanhooren and Nguyen; 1996). The principal layout of the simulated plant is shown in Figure E.2, together with important physical dimensions.

Monitoring algorithm and local control systems

The integrated control system consists of a number of different elements:

- monitoring of the influent water characteristics including flow rate, ammonia and suspended solids (SS) concentration measurements with a sampling rate of 4 h^{-1};
- classification of the present operational state into one out of five different states;
- determination of setpoints using the newly proposed strategies (see below);
- a local control loop for the sludge-blanket level by means of return-sludge flow rate; local control loops for the DO-concentrations in the two reactors and a local control loop for the step-feed. The latter two control loops are assumed to be ideal, and hence their dynamics is not simulated.

In COST data there are three extreme events present: storm with sewer flush-out, i.e. high flow rate with high concentration of SS, storm with dilution and rain, also with diluted water. In addition to these disturbances an extreme ammonia load disturbance is included to mimic events that may occur in systems with anaerobic sludge treatment. Five operational states were defined ac-

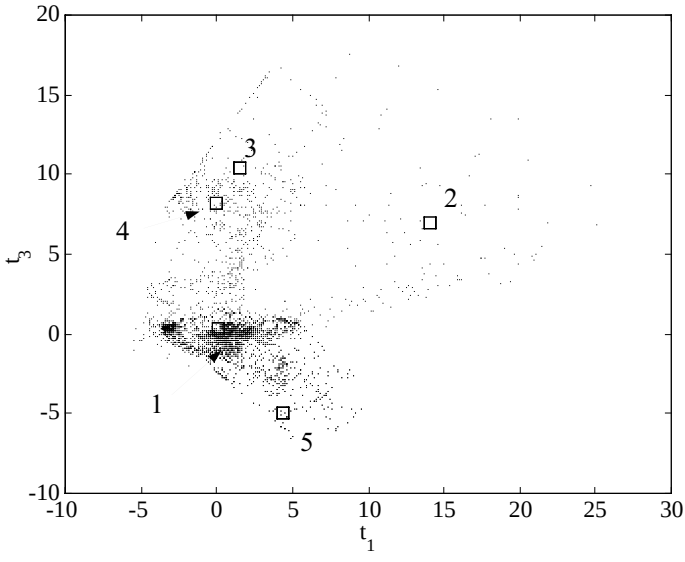

Figure E.3: Clusters with cluster centres, representing the five different operational states. Normal conditions (1), storm with sewer flush-out (2), storm (3), rain (4) and extreme ammonia load (5).

cordingly: normal operation, storm with sewer flush out, storm, rain and high ammonia load.

A dynamic PCA model (see Eq. (2)), including one time-lagged value for each variable, is identified from influent data representing normal operation. From the PCA, three principal components are used to classify the operational state by means of FCM clustering. For each defined disturbance, a training data set is constructed with similar disturbances to those of original COST data. Each training data set is projected onto the PCA model and two clusters are identified using FCM: normal and extreme event, resulting in a total of eight clusters. The clusters representing normal conditions coincide and, thus, five distinctive clusters are obtained. These five clusters with cluster centres are shown in Figure E.3. Look-up tables with predefined setpoints are used to determine the setpoints for each operational state. Two values on m in (5) are evaluated: $m = 1.01$, which corresponds to a crisp (non-fuzzy) classification and $m = 1.4$, which yields a fuzzy classification.

Supervisory control strategies

A control strategy has been developed for each operational state. Each event strategy represent a shift in control objective from that of the normal operation.

1. Normal operation strategy. During normal operation, the sludge-blanket level is controlled to 0.68 m. DO-concentrations are 1.0 mg/l in both reactors and 100 % of the influent flow is directed to the first reactor.

2. Storm/flush-event strategy. During flush-out, the sludge-blanket setpoint is raised to 2.0 m to lower the hydraulic load to the settler. Influent flow is directed to the first reactor.

3. Storm strategy. The strategy during storm events is to raise the sludge-blanket setpoint (1.5 m) for the same reason as mentioned above. The influent flow is redirected to the second reactor. In this way, sludge is accumulated in the first reactor with lower sludge load to the settler as a result. The idea is that this decreases the sludge loss to the effluent. The DO-concentration is lowered to 0.1 mg/l in the first reactor and raised to 2.0 mg/l in the second.

4. Rain strategy. The rain event strategy is similar to the storm strategy. The sludge-blanket setpoint is set to 1.0 m.

5. Extreme ammonia load strategy. The strategy during high ammonia load is to lower the sludge-blanket setpoint (0.2 m) in order to get a higher concentration of biomass in the reactors. The setpoint for the DO-concentrations is raised to 3.0 and 2.0 mg/l in the first and second reactor, respectively. 100 % of the influent flow is directed to the first reactor.

A reference control case is defined to evaluate the performance of the controlled case. The return-sludge flow rate is proportionally controlled to 80 % of influent flow rate. This rate yields the same sludge-blanket height as for the controlled case during normal operation. The excess sludge removal rate is controlled so that the reference and controlled case has the same sludge age. The DO-concentration is kept constant at 1 mg/l and no step feed is utilised.

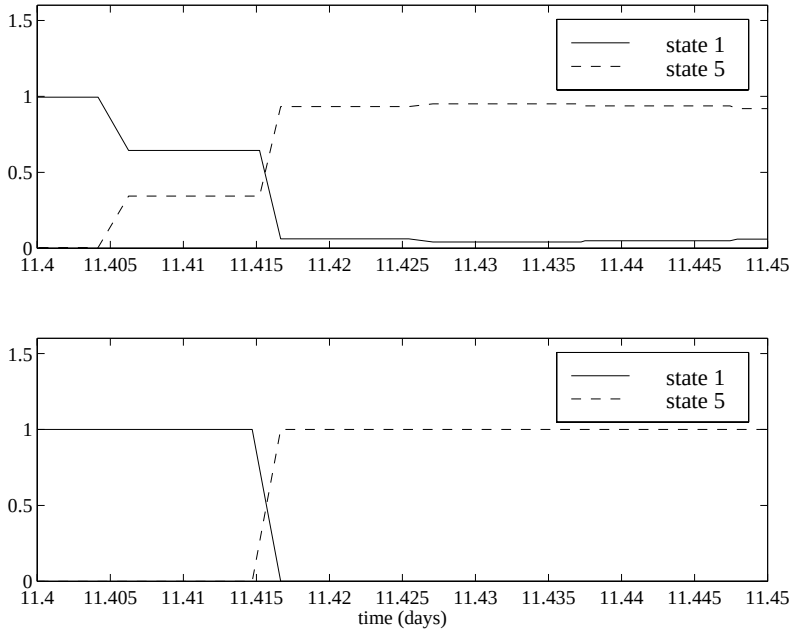

Figure E.4: Detection and classification of the extreme ammonia load. Fuzzy (top) and crisp (bottom) classification with $m = 1.4$ and $m = 1.01$, respectively.

Results and discussion

The results of the simulation study are discussed from two separate viewpoints: performance of the monitoring and control system and performance of the process.

Monitoring and control system

The main objective for the supervisory control system is to identify disturbances and to change the setpoints accordingly. Detection and an accurate classification are achieved during all disturbances. It is worth noting that the fuzzy classification ($m = 1.4$) could detect a disturbance earlier and change the setpoints accordingly (Figure E.4). Another advantage with the fuzzy classification is that

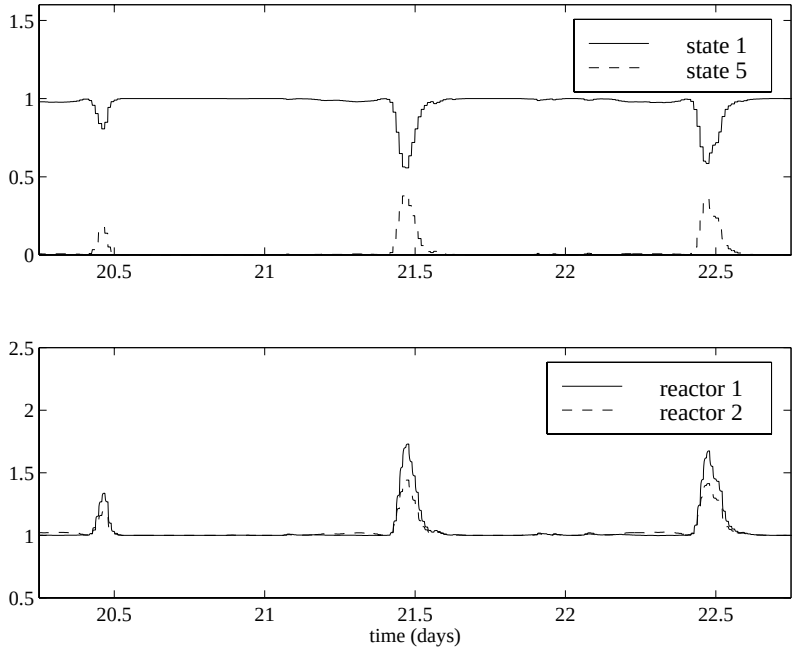

Figure E.5: Fuzzy classification during normal operation (top). During peak load, the state is classified between class 1 and 5, resulting in a varying DO-concentration (bottom).

the DO setpoints were changed during ammonia peak loads, even though the operational state is considered normal (Figure E.5). The fuzzy information also resulted in a smoother control with less saturation of control signals.

In Figure E.6 the differences between the crisp and the fuzzy classification for generating setpoints for the sludge blanket level are shown. The gradual transition between two operational states implies a smoother setpoint change and, consequently, a smoother change in the manipulated variable, the return-sludge flow rate (Figure E.7). This has the advantage that hydraulic shocks are not introduced by the control system and that unnecessary wear on mechanical equipment is avoided.

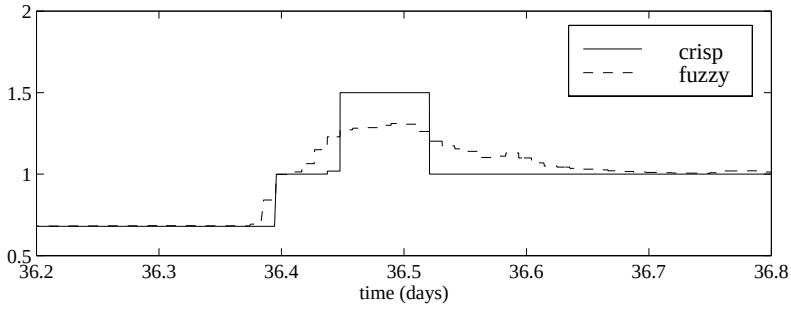

Figure E.6: Setpoints for the sludge-blanket level at the start of the rain event using crisp and fuzzy classification.

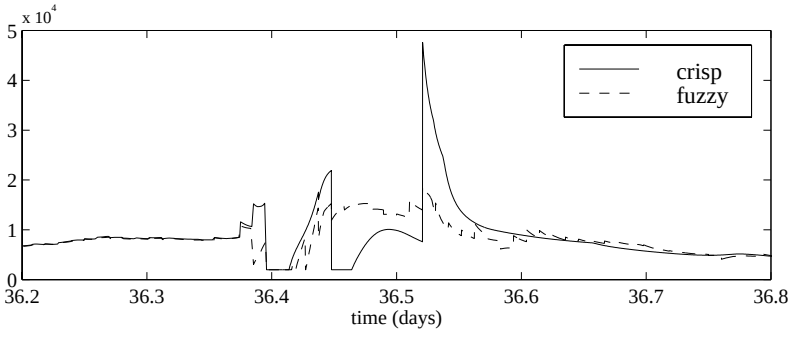

Figure E.7: Resulting return-sludge flow rate at the start of the rain event using crisp and fuzzy classification from Figure E.6.

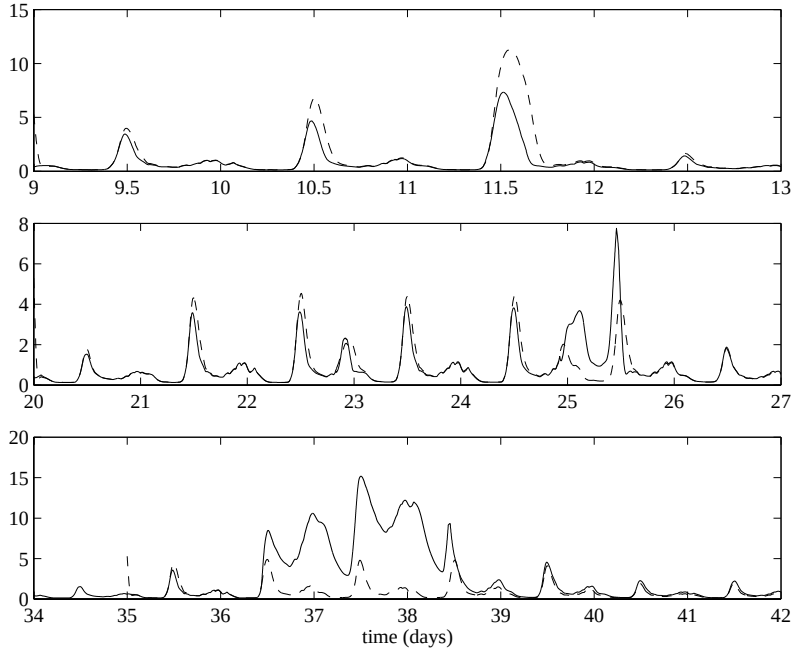

Figure E.8: Effluent ammonia concentration during high ammonia load (top), storm events (middle) and rain event (bottom). controlled case (—), reference case(--)

Process performance

It can be seen from Figure E.8 that the controlled (according to the proposed method) case yielded lower or as low effluent ammonia for the period of high ammonia load, but higher effluent concentrations during the periods of storms and rain. This is in compliance with the shift in the control objective-from carbon reduction and nitrification to prevention of sludge loss. This strategy shift is visible in the effluent SS concentrations (Figure E.9). During the high ammonia load period, the effluent SS is in parity with the reference case, but during storm and rain periods, the effluent SS is decreased compared to the reference case.

It is important to note that the case study is carried out mainly to illustrate the applicability of the integrated monitoring and control strategy to wastewater

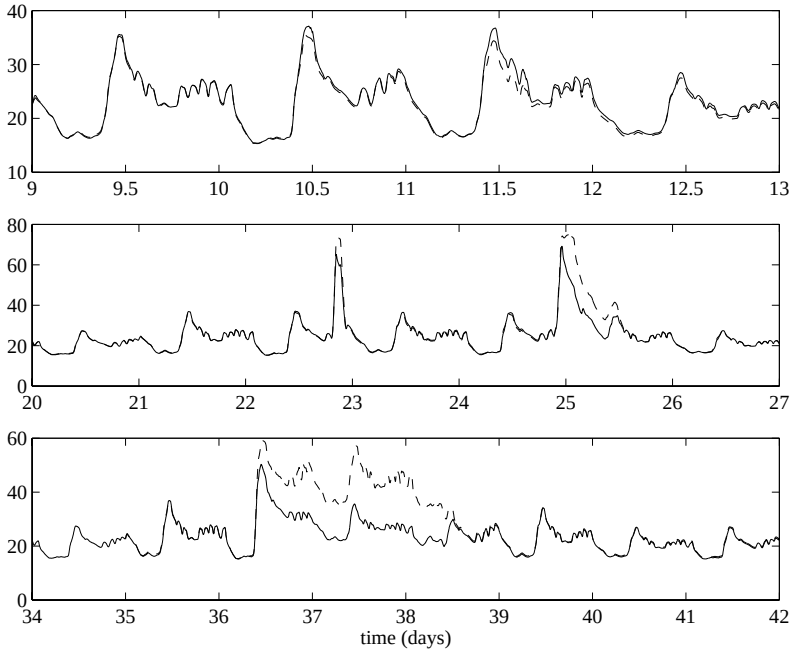

Figure E.9: Effluent SS concentration during high ammonia load (top), storm events (middle) and rain event (bottom). controlled case (—), reference case(--)

treatment processes. There are a number of possible improvements. For instance, only influent data are used to monitor and classify the operational state. There is more information to gain from including process measurements in the analysis. Moreover, PCA-based monitoring as presented here cannot handle changing process conditions. That is, the mean and variance of data have to be approximately constant over longer periods of time. Also, no consideration is taken to the fact that events occur in different time scales. There are solutions to these problems and they are discussed in Rosen and Lennox (2001). Another important improvement could be to incorporate the nonlinear process behaviour in the monitoring model. This can be done by pre-processing of data or by using nonlinear projection methods (Zhang et al.; 1997).

An issue that has not been addressed here is the ratio between the sampling/ updating time of the low-level controllers and the supervisory control scheme.

Supervisory control together with low-level controllers can be seen as cascade control. This means that the outer control loop (supervisory control) must have considerably slower dynamics than the inner loop (low-level control) to avoid oscillatory or unstable behaviour in the system. In this work, a sample/update rate of 4 h^{-1} was used for the supervisory control. This is perhaps too short in most real systems, especially if sludge dynamics are included. Therefore, in the general case the supervisory control may have to be separated into several levels with decreasing sampling/update rate depending on the dynamics of the controlled variables.

No strong conclusions with regard to the impact of the supervisory control system on the performance of the plant can be drawn from this preliminary study. However, the study indicates that the approach to supervisory control proposed in this work may be used to coordinate local control loops and to determine appropriate setpoints for the current operational state.

Conclusions

In this paper an approach to automatic supervisory control of wastewater treatment operation is proposed. By integrating online monitoring and control, appropriate low-level controller setpoint and structures for the current operational state of the process can be determined. The simulation study indicates that the proposed approach to automatic supervisory control is applicable to wastewater treatment operation. Principal component analysis (PCA) is a powerful method to extract relevant information from measurement data since it is capable of representing the underlying mechanisms by means of principal components (PCs). Fuzzy c-means (FCM) can be used to classify the operational state of the process and this is preferably done in the PC-space. A comparison of a fuzzy and crisp classification shows that fuzzy classification gives faster detection and smoother control than crisp classification.

Paper E

Addendum

The return sludge flow rate controller deserves some further explanation. The controller comprises one feed-forward component and two feedback loops: one of which corrects the control error, the other imposes an upper limit on the sludge retention time in the settler ($SRTiS$) to prevent excessive denitrification or phosphorus release from taking place in the settler (Figure E.10).

When the load to a settler increases, Q_{ret} needs to be increased to maintain the sludge blanket height (SBH) at the setpoint, leading to the design of the following feed-forward control law:

$$Q_{ret,ff} \quad = \quad k_{ff} Q_{in}$$

where $Q_{ret,ff}$ is the return sludge flow rate, Q_{in} is the influent flow rate and k_{ff} is the feed-forward gain. k_{ff} is also determined by the supervisory controller and is, consequently, a function of the current operational state. A proportional feedback controller,

$$Q_{ret,fb} \quad = \quad k_{fb} \left(SBH_{sp} - SBH \right)$$

is used to reduce the control error that is produced by the feed-forward controller.

Assuming that the maximum allowable $SRTiS$ is $SRTiS_{max}$, the sludge re-

turn flow rate (Q_{ret}) should satisfy,

$$\begin{aligned} Q_{ret} &\geq Q_{ret,SRTiS} \\ Q_{ret,SRTiS} &= \frac{SBHA_sX_{s,ave}}{SRTiS_{max}X_{ret}} \end{aligned}$$

where A_s is the sectional area of the settler, $X_{s,ave}$ is the average solids concentration below the sludge blanket and X_{ret} is the solids concentration in the recycled sludge. Thus, in this work, the X_{ret} is assumed to be measured or estimated.

The controller is also used in Yuan et al. (2001a). However, some changes are introduced to improve the performance. An increasing Q_{ret} causes a dilution of the recycled sludge, reducing the effectiveness of Q_{ret} in maintaining the sludge blanket height at the setpoint. Therefore, k_{ff} is designed as shown in Figure E.11, i.e. k_{ff} is a function of Q_{in}. $k_{ff,0}$, the feed-forward gain for $Q_{in} = Q_{in,ave}$ (the average dry weather influent flow rate), is designed so that the controller is able to maintain the sludge blanket at the ordinary SBH setpoint without the presence of the error correction feedback loop. Slope α is chosen large enough such that,

> a negative control error (i.e. $SBH_{sp} < SBH$) is produced for the feedback loop to correct, when $Q_{in} < Q_{inf,ave}$, so that the controlled SBH will be higher than the setpoint;
>
> a positive control error (i.e. $SBH_{sp} > SBH$) is produced for the feedback loop to correct, when $Q_{in} > Q_{inf,ave}$, so that the controlled SBH be lower than the setpoint.

In dry weather, the nitrogen load is typically in phase with the hydraulic load. Then the above feed-forward controller, together with an appropriately designed feedback controller (see below), will result in a desirable variation of the sludge inventory in the settler so that both nitrification and denitrification are enhanced (Yuan et al.; 2001a). Moreover, if Q_{in} could be measured upstream from the treatment plant (e.g. one hour before it arrives at the treatment system), the feed-forward controller would have time to react in advance so that the sludge is recycled to the reactor before the high load arrives.

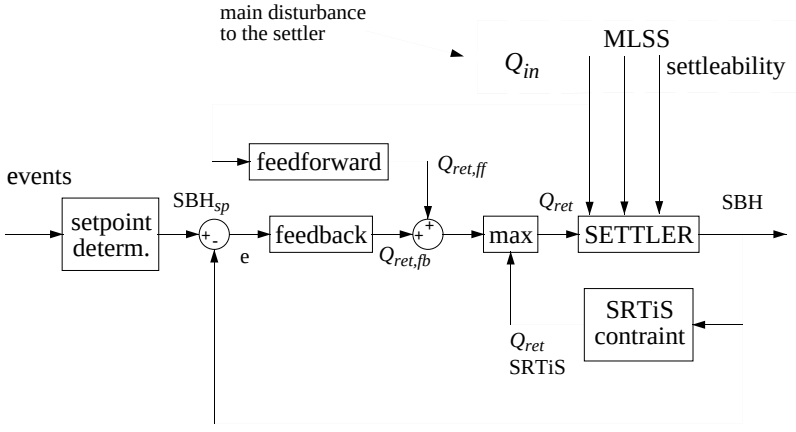

Figure E.10: Sludge return control system structure.

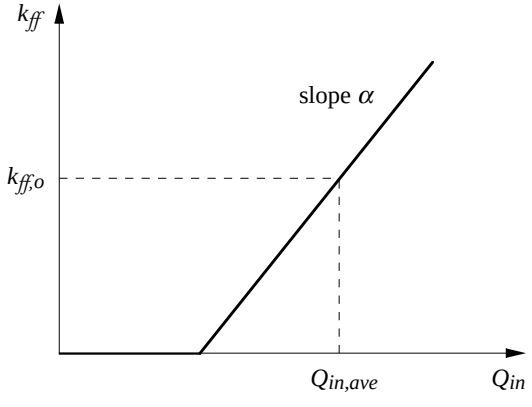

Figure E.11: Feed-forward gain design.

Paper F

A framework for extreme-event control in wastewater treatment

C. Rosen, M. Larsson, U. Jeppsson and Z. Yuan

Wat. Sci. Tech. 2001, (accepted).

Abstract: *In this paper an approach to extreme event control in wastewater treatment plant operation by use of automatic supervisory control is discussed. The framework presented is based on the fact that different operational conditions manifest themselves as clusters in a multivariate measurement space. These clusters are identified and linked to specific and corresponding events by use of principal component analysis and fuzzy c-means clustering. A reduced system model is assigned to each type of extreme event and used to calculate appropriate local controller setpoints. In earlier work we have shown that this approach is applicable to wastewater treatment control using look-up tables to determine current setpoints. In this work we focus on the automatic determination of appropriate setpoints by use of steady-state and dynamic predictions. The performance of a relatively simple steady-state supervisory controller is compared with that of a model predictive supervisory controller. Also, a look-up table approach is included in the comparison, as it provides a simple and robust alternative to the steady-state and model predictive controllers. The methodology is illustrated in a simulation study.*

Keywords: Model predictive control; principal component analysis; supervisory control; wastewater.

Introduction

Operation of wastewater treatment plants has become increasingly automated during the last decades. This has been made possible due to a significant increase in the number of process variables that can be reliably measured online together with an increased knowledge of the biochemical processes. Today, there are automatically controlled aeration, return-sludge pumping and internal nitrate recirculation to mention a few simple examples (Olsson and Newell; 1999; Yuan et al.; 2001b). In addition to a process control system, there is normally a system for process surveillance or monitoring. Generally, different regions in the multidimensional process space constituted by process measurements represent different operational conditions of the treatment system. By recognising these regions, by means of multivariate process monitoring and linking them to certain operational states, the current operational state can be classified using online measurement data. Normally, there is an operating region in which the process is considered to be normal or in-control. Within this region no, or limited, corrections of the local controllers are needed. However, during extreme events, control strategies completely different from those carried out during the in-control case may be necessary to meet the requirements on effluent quality or process safety.

To work out strategies for the operation and determine setpoints for local controllers is often referred to as supervisory control. In most plants the operators or process engineers carry out the supervisory control. However, a step towards more automated operation is to automatically determine the operational state and derive appropriate control actions based on the monitoring information. In Rosen and Yuan (2000) an approach to integrate monitoring and control to form an automatic supervisory control scheme is presented (Figure F.1). Principal component analysis (PCA), e.g. Piovoso and Kosanovich (1994), is combined with Fuzzy C-Means (FCM) clustering, e.g. Marsili-Libelli and Müller (1996), to identify clusters and classify the current operational state in an orthogonal space of reduced dimensionality compared to the original measurement space. The work focuses on the monitoring/classification task, and the setpoint determination is limited to an a priori defined control sequence designed to drive the process back to its normal state and to decrease the negative effects of the disturbance. In this paper, model-based alternatives to predefined control sequences (rule based or look-up tables) are discussed. These alternatives are

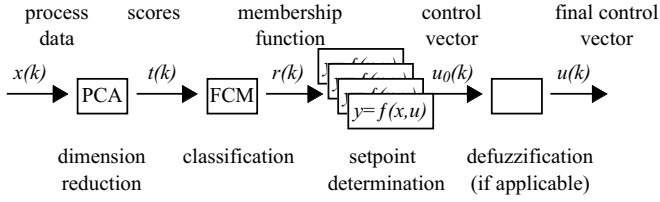

Figure F.1: The structure of the supervisory control system.

based on a reduced system model that is used to derive the control sequence either by steady-state control (SSC) or by predicting the effects and optimising the control actions according to a loss function, i.e. model predictive control (MPC). This work, together with the work presented in Rosen and Yuan (2000), aims at describing a framework for extreme event control in wastewater treatment operation. In this work, normal operational control is not considered. Thus, when the process is considered to be in the normal state, the controller set points are set to predefined values. This is, however, a task that could be handled within the presented framework as yet another setpoint determination model. Supervisory control by use of inverse PCA during normal operational conditions has been suggested in Rosen and Jeppsson (2001a).

The paper is organised as follows: First, the extreme event control scheme is described. This is followed by a description and definition of a simulation case study of high ammonia load control. The results of the study are then presented, together with a discussion and comparison of different methods for controller setpoint determination. The paper is concluded with a summary and a discussion on future research.

Extreme event control scheme

Dimension reduction and classification of current operational state

The classification of operational states by use of a combination of PCA and FCM clustering has been described earlier (Rosen and Yuan; 2000). The approach is based on the fact that different operational states (caused by disturbances) generally manifest themselves as clusters in a space defined by the online measurements. However, with a high number of measured and monitored vari-

ables, the measurement space will be high-dimensional. To allow for robust classification, a multivariate analysis technique is used to reduce the dimensionality and to decrease the noise level. In the reduced space, clusters representing normal operation as well as different disturbance types are defined using clustering techniques. As new samples are collected, the clustering algorithm classifies the current operational state.

Setpoint determination

When the current operational state is known, the setpoint can be determined in various ways. In Rosen and Yuan (2000) look-up tables with a set of predefined setpoints were used. In this paper the method is refined to compute the setpoints using steady-state and dynamic predictions from a reduced system model.

To use an all-embracing model, such as the ASM1 (Henze et al.; 1987), for the dynamic prediction is difficult due to the large number of parameters that need to be identified and updated and several non-measurable states. Instead, the information on the operational state is used to select a reduced-order model (Jeppsson; 1996) tailored to suit the type of disturbance currently encountered. This can be justified since different disturbances affect the process in different ways. Some disturbances affect the slow dynamics of the process, whereas others affect the fast. This makes it possible to decouple the system and use simplified models to describe the effects of each disturbance. The decoupling is generally temporal, which implies that the prediction horizon may vary for different types of disturbances. In some cases a static model is adequate to assess the effects of a certain disturbance. In many other cases, however, a dynamic approach may be required.

Look-up table. The look-up table is a set of pre-defined set points for each type of disturbance. When a disturbance is detected and classified, the values in the table are used until the classification results in normal (or another) type of operation. In this approach, no consideration is taken to the specific current process knowledge obtained by measurements and, thus, the controller setpoints need to be set so that a "safety margin" is obtained. This may lead to extreme controller setpoints and, consequently, often high operational costs, but the method is appealing due to its robustness and simplicity.

Steady-State Control (SSC). In this approach, a reduced model of the system is linearised at the current operational point at each sample. The linear approximation of the reduced model can be expressed in state-space form. If the derivatives of the state-space model are set to zero, the steady-state relationships between inputs (setpoints for locally controlled variables), u, and outputs, y, can be computed. Then, the inverse relationship yields the control change needed to reach a desired output. Assuming a setpoint, y_r, for output variable y, the SSC control vector, u, for the linearised model is calculated as:

$$u^+ \quad = \quad u^0 - \left(CA^- B\right)^\dagger \left(y_r - y\right) Q \tag{F.1}$$

where A, B and C are the model matrices in a state-space representation and $(\,)^\dagger$ denotes the pseudo inverse and u^+ and u^0 are the calculated and current control states, respectively. It should be pointed out that the inverse relationship is not always attainable. However, in this application the inverse relationship can generally be computed. The calculated control action may be overly aggressive, since it is based on a steady-state assumption. Therefore, a relaxation factor can be employed so that only a fraction of the calculated control input is implemented (Piovoso and Kosanovich; 1994). Here the factor is implemented as a diagonal matrix, Q, to be able to differentiate between control actions required to meet the respective output setpoint. Apart from the obvious limitation of SSC to account for process dynamics, another limitation must be mentioned; SSC does not handle limits put on the control signals. This may lead to loss of controllability when the controller yields non-applicable control actions.

Model Predictive Control (MPC). The principle of MPC (e.g. Morari and Lee (1999) and Camacho and Bordons (1999)) is that a system model is used to predict the future output trajectories for a set of possible control inputs (Figure F.2). The initial state of the system model is the current process state including estimated state variables, control variables and disturbances (if measurable or possible to estimate). A cost function is defined based on the deviation of each predicted trajectory from the desired trajectory. The optimal control sequence, in the sense that it minimises the cost function, is then obtained by solving the optimisation problem online each time a new control output is to be determined. The optimisation is carried out in the control variable space limited by the available control region and, thus, MPC can take controller limitations into

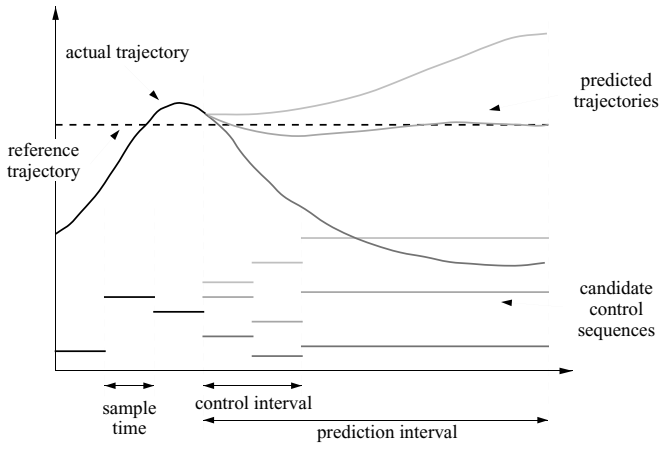

Figure F.2: Principle of model predictive control.

account. The cost function is expressed as:

$$
\begin{aligned}
J\left(u^{+}, x^{*}\right) \quad = \quad & \int_{t^{*}}^{t^{*}+t} \Big[\left(\hat{y} - y_r\right)^T Q \left(\hat{y} - y_r\right) + \\
& \left(u^{+} - u^{0}\right)^T R_1 \left(u^{+} - u^{0}\right) + \\
& \left(u^{+} - u_{min}\right)^T R_2 \left(u^{+} - u_{min}\right) \Big] \, dt
\end{aligned} \tag{F.2}
$$

where $\hat{y} = f(t, x^{*}, u^{+})$ is the predicted output trajectory using information up to time , with piecewise constant control sequence u^{+} during the prediction interval. y_r is the reference trajectory, u^{0} is the current control state and Q, R_1 and R_2 are the weighting matrices for output errors, control variations and absolute control outputs, respectively. R_1 is used to penalise the changes in the control output to obtain a smoother control signal, whereas R_2 is used to penalise the absolute control cost. This is useful in achieving a balance between effluent quality and operational costs. MPC has been applied to many applications, from chemical industry (Garcia et al.; 1989) and power distribution (Larsson et al.; 2000) to wastewater treatment (Weijers; 2000). Here, linearised models are used to find the optimal control vector. However, nonlinear MPC can also be used at the cost of more computationally intense solutions.

Control Scheme

Before the calculated setpoints are applied on the local controllers, they are weighted according to the membership function from the classification stage (the defuzzification step in Figure F.1). Thus, the controller setpoints are, thus, the weighted sum of the results of more than one setpoint determination model. This means that the crispness of the classification is important so that seamless transitions between setpoints are obtained.

Case study: extreme ammonia load control

A simulation case study with application to wastewater treatment operation is reported below. In this study, a supervisory control component is designed to coordinate local control loops. The control objective is to minimise effluent ammonia concentration given an ammonia load disturbance and, if possible, operational costs. The primary objectives of the study are to illustrate the proposed methodology and to evaluate its applicability to wastewater treatment operation.

Simulated plant

The simulated plant is based on the benchmark model developed within the cooperation of COST action 624 (Pons et al.; 1999). All model and physical parameter values are chosen according to the official benchmark model, and the configuration is illustrated in Figure F.3. High ammonia load is chosen as the disturbance for this study. This is a fast disturbance in the range of hours to one day and mimics a hypothetical increase in the ammonia load caused by, for instance, a temporary discharge. The disturbance was introduced into the influent data file for dry weather conditions designed within the COST benchmark programme. To make the detection task challenging, the disturbance is located in time so that it coincides with the normal diurnal influent ammonia peak (see Figure F.4 in the Results and discussion).

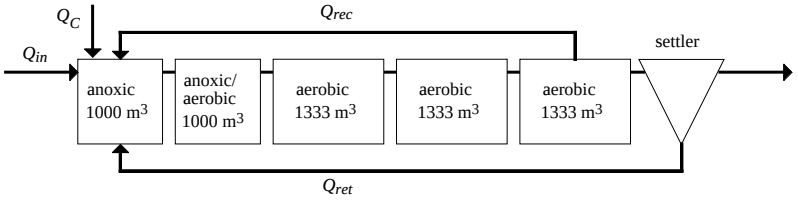

Figure F.3: Simulated plant configuration.

Reduced model for nitrogen control

The reduced system model given in the appendix is implemented as a three-reactor model-one anoxic, one anoxic/aerobic and one aerobic reactor. No settler was modelled, since only soluble states are considered. Controlled variables are the dissolved oxygen (DO) levels in reactors 2-5 (where reactors 3-5 are given the same value), internal recirculation flow rate (Q_{rec}) and carbon addition flow rate into reactor 1 (Q_C). To ensure the plant is operated as a pre-denitrification plant and avoid unrealistic control outputs, minimum and maximum levels are defined for each controlled variable (Table F.1).

One important factor in developing a reduced model is to minimise the number of parameters that need to be identified in a real application. Also, the number of states that has to be estimated should be kept at a minimum. The aim of the model is to represent dynamics in a time constant range of hours to one day, which implies that a number of states, normally seen in ASM models such as ASM1, can be omitted from the model. The result is a model with three states in each modelled reactor: readily biodegradable substrate, S_S, nitrate/nitrite, S_{NO}, and ammonium/ammonia, S_{NH}. The model assumes that total suspended solids, TSS, are constant for the prediction interval and that

Controlled variable	Minimum value	Maximum value
DO in reactor 1	0 mg/l	2 mg/l
DO in reactors 3-5	1 mg/l	4 mg/l
Q_{rec}	1×18466 m^3/d	5.4×18466 m^3/d
Q_C	0 m^3/d	10 m^3/d

Table F.1: Controlled variables and their limits.

dissolved oxygen, S_O, changes instantaneously. This is a fair assumption, since TSS changes slowly compared to the frequencies of interest and S_O can be considered controlled by local controllers.

State estimation

It is assumed that S_{NO} can be measured in reactors 1 and 5 and that S_{NH} is measured in the influent and reactor 5. These measurements are used to update the reduced model states for nitrate and ammonia. The remaining states and influent concentrations are roughly estimated based on mass balances and empirical knowledge (see appendix). The influent estimates are based on the fact that the correlation between influent S_S and S_{NH} is strong in the COST influent data. This is certainly not as clear-cut in real data. However, it is not unrealistic to assume that some correlation exists and that this, together with other measurements, may constitute a basis for influent S_S-estimation. TSS and S_O are assumed measurable, whereas the model parameters listed in Table F.3 in the appendix are assumed known.

Controller parameters

Apart from the model parameters, a number of controller parameters need to be set. Firstly, the effluent nitrate and ammonia setpoints, used for both the SSC and the MPC, are set to slightly lower values than those obtained during normal operation. This implies $S_{NH,ref} = 1$ mg/l and $S_{NO,ref} = 11$ mg/l. For the SSC, the relaxation factor in Equation 1 is set to $Q_{SSC} = diag[0.07\ 0.3]$, i.e. outputs are weighted so that emphasis is put on keeping the effluent ammonia concentration low. For MPC, QMPC is set to $Q_{MPC} = diag[0.5\ 2.0]$ for the same reasons. $R_1 = diag[1\ 1\ 1\ 1]$ to reduce too fast controller setpoint changes. $R_2 = diag[0.75\ 0.5\ 0.375\ 0.0625]$ with the values set to reflect the costs of QC, DO_2, DO_{3-5} and Q_{rec}, respectively. However, since neither the look-up table nor the SSC strategies incorporate cost penalties, a simulation with $R_2 = 0$ is carried out for comparison reasons. The MPC prediction horizon used is 40 samples, that is 5 hours, which is adequate for most of the nitrogen dynamics to settle. The control horizon is set to 3 samples or 45 minutes, which is a compromise between flexibility and computational time. The MPC

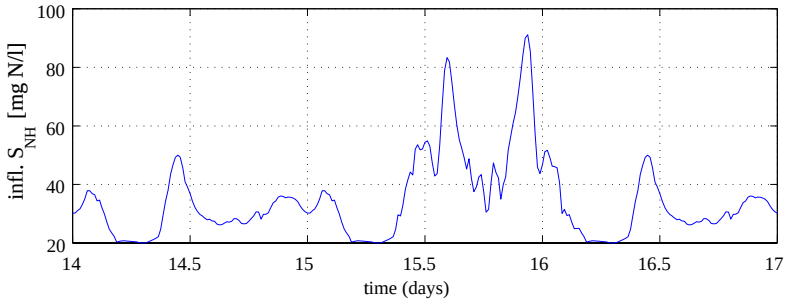

Figure F.4: Influent ammonia disturbance. There are similar, but not as dominant, disturbances in influent S_S, X_S, X_I, X_P, S_{ND} and X_{ND}.

is implemented in a standard way according textbooks on model predictive control. Crisp classification was used in order to simplify the comparison between the different methods.

Results and discussion

Simulation results

The plant is simulated during 15 days prior to the ammonia disturbance using dry weather influent data to achieve quasi-steady state conditions. The normal operation setpoints are set to be $Q_C = 2$, $DO_2 = 1$, $DO_{3-5} = 1.5$ and $Q_{rec} = 3.5$. The disturbance is introduced at day 15.4, and if no changes are made to the controller setpoints the result can be seen in Figure F.5. It is clear that the disturbance has a significant effect on the effluent quality of the plant.

Look-up table. The look-up table values for the ammonia load disturbance is set to the maximum for each control variable (see Table F.1). Looking at the effluent concentrations during the disturbance event (Figure F.6), it is evident that the look-up table approach decreases the effluent nitrogen discharge compared to the constant control case. The effluent ammonia concentration does not exceed 10 mg/l. However, the cost is high in terms of energy and carbon consumption. An interesting observation is that at the end of the disturbance event, the look-up table yields far too high controller setpoints. This is due to

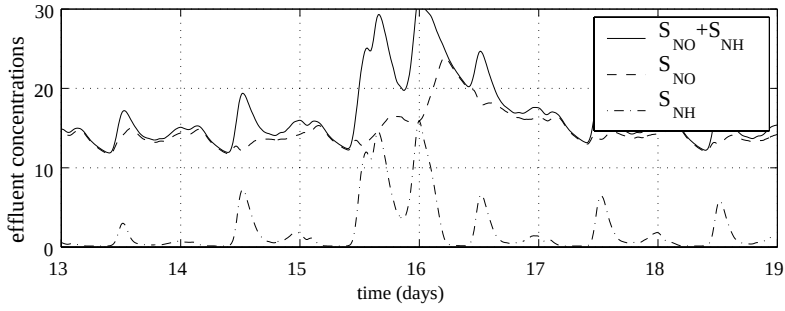

Figure F.5: Effects of the high ammonia load disturbance entering the plant at day 15.4. The setpoints are kept at the same values as for normal operation.

that the detection algorithm is somewhat slow to detect that the operational state has returned to normal. This can be considered as a failure of the detection algorithm, but as will be seen later, this failure has only minor impact on the two model-based controllers as both use updated information. Hence, the look-up table approach is more sensitive to accurate state classification.

SSC. According to Figure F.7, the SSC attenuates the disturbance. Here, it can clearly be seen that the control signals change during the disturbance event as the steady-state model is continuously updated. This is especially evident for the aeration of the second reactor, which is set to zero during the period of lower influent ammonia load around day 15.8. Also the internal recirculation flow rate changes during the course of the event. The continuous updating, however, implies that the model will always be one step behind, since it assumes that the conditions will stay constant at each updating instance. The effluent concentrations are kept at reasonable levels and the operational cost is lower than in the look-up table case.

MPC. The MPC also suffers from the problem of being one step behind and the performance is similar to that of SSC (Figure F.8). However, the peaks in ammonia concentration are lower resulting in a lower discharge of nitrogen and the effluent nitrate concentration also displays a somewhat calmer behaviour. Except for the external carbon addition, the MPC yields lower control setpoints. In Figure F.9, the behaviour of the weighted MPC is displayed. Here,

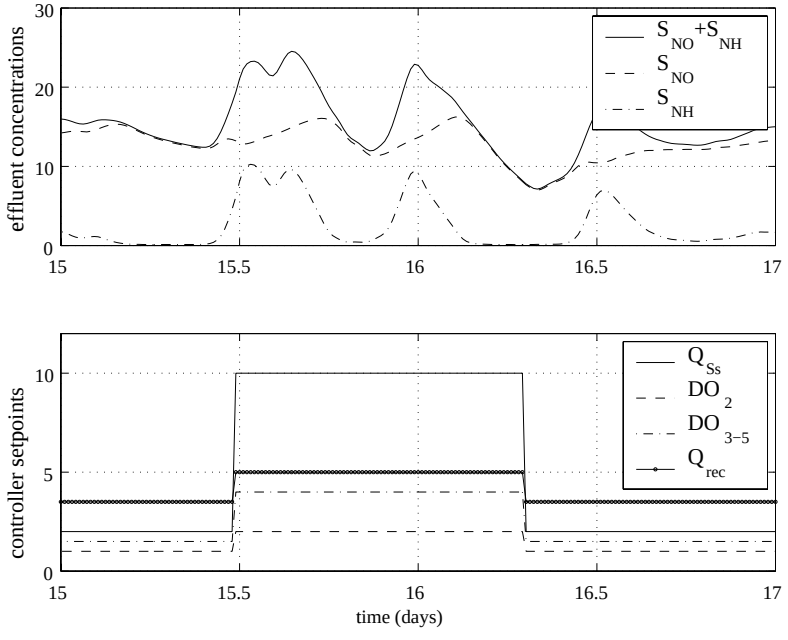

Figure F.6: Set-points determined using a look-up table.

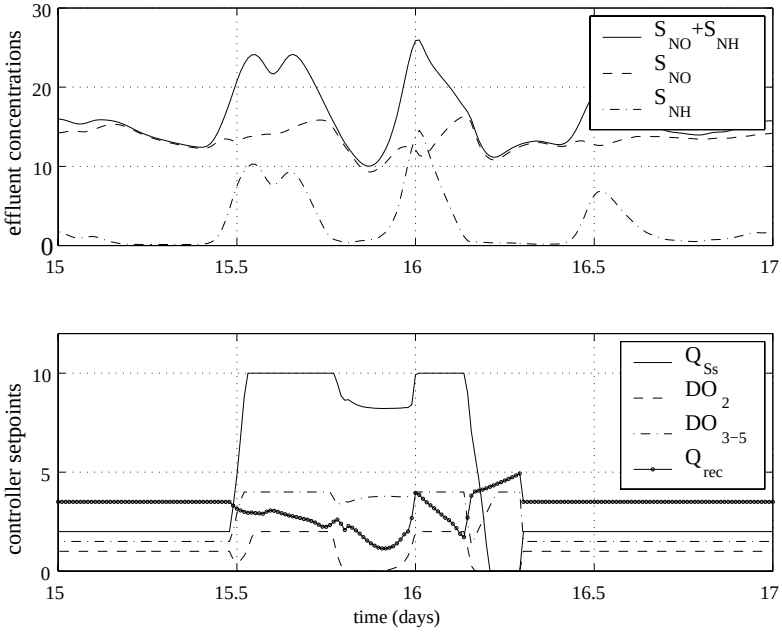

Figure F.7: Set-points determined by using SSC.

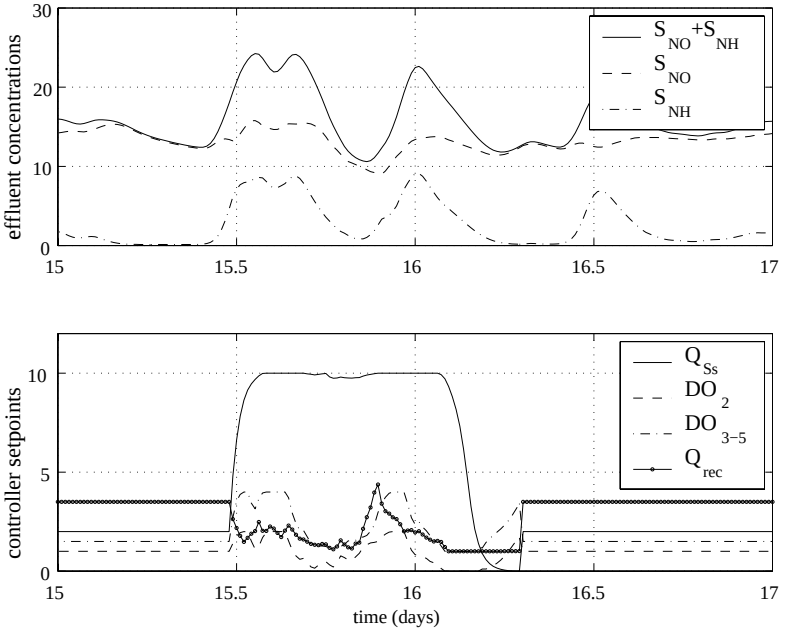

Figure F.8: Set-points determined by MPC without weights on u.

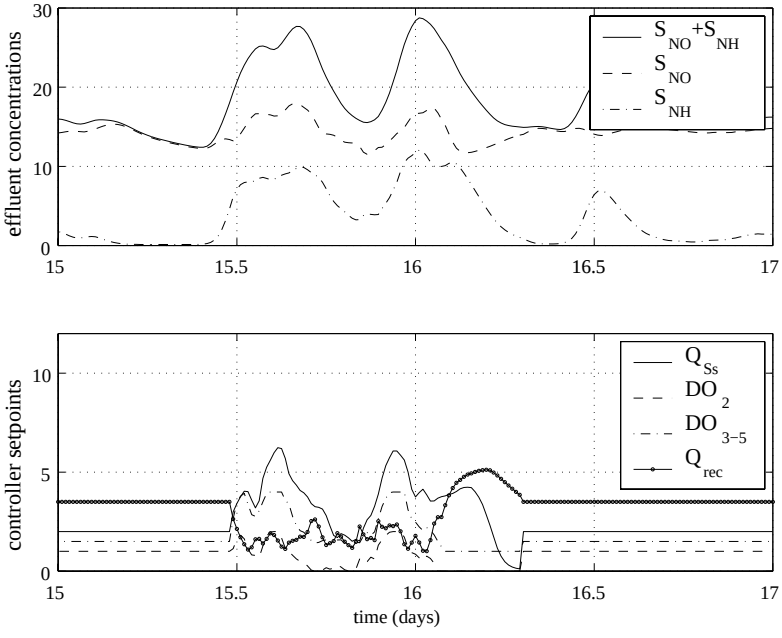

Figure F.9: Set-points determined by MPC with weights on u.

Variable (cost)	No change	Look-up table	SSC	MPC, $R_1 = \mathbf{0}$	MPC, $R_1 \neq \mathbf{0}$
S_{NO}	1	0.74	0.78	0.77	0.86
S_{NH}	1	0.64	0.69	0.67	0.88
K_La	1	1.36	1.33	1.14	1.06
Q_{rec}	1	1.22	0.90	0.74	0.86
Q_C	1	3.03	2.48	2.50	1.38
$S_{NO} + S_{NH}$	1	0.72	0.76	0.75	0.86

Table F.2: Relative discharges/costs for the different setpoint determination methods.

the control set points are considerably lower as a compromise between discharge levels and costs is obtained. The difference between the weighted and the non-weighted MPC clearly indicates how the weighting matrices can be tuned so that a desired balance between cost and effluent discharge can be obtained.

Comparison

It is not obvious from the plots in Figures F.6-F.8 which setpoint determination method yields the best result. By investigating the integral of each variable during a time period, a clearer picture on the efficiency of each method can be obtained. In Table F.2, the numbers show the nitrogen discharge, total air added to the system (in terms of K_La), external carbon addition and internal recirculation during the period from day 15.4 to 17 relative to the reference case (no changes in the setpoints). As expected, the look-up table yields the least effluent discharge of dissolved nitrogen, but at a high cost. A bit surprising is that the relatively simple SSC approach yields almost as good results as the MPC. However, SSC does not have the advantage of being able to find a compromise between effluent discharge and control costs. Even though the total amount of air as well as external carbon added to the system is lower for the MPC than SSC, the total ammonia (and nitrate) discharge is lower. This indicates that the MPC coordinates the control actions more efficiently than SSC.

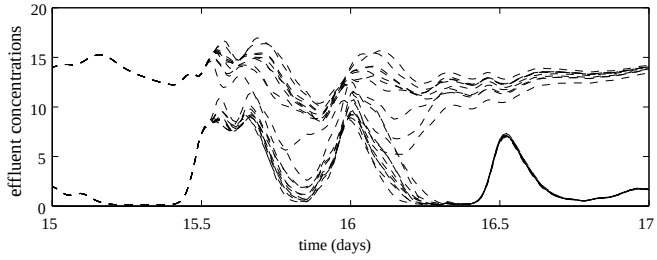

Figure F.10: Effluent nitrogen for a number of simulations where the model parameters are afflicted with random errors.

Parameter sensitivity

The reduced system model is based on a number of parameters that need to be identified. This is seldom an easy task and the margin for errors must be considered to be relatively high. To get an indication on how sensitive the model is to parameter errors, a number of simulations are carried out where the parameters are randomly afflicted with errors. All parameter values are independently given a random normally distributed additive error with a standard deviation of 0.15 times the parameter value. The result can be seen in Figure F.10 where MPC was used. It is clear that when the model is used for the extreme event control discussed here, it is relatively insensitive to parameter errors. The reason is that during extreme events it is not crucial to obtain exactly the right results. Instead, it is important that the control gradient, that is the direction of a certain control contribution, points in the desired direction, since the model is continuously updated. Similar, or slightly better, results as depicted in Figure F.10 are obtained if the sensitivity of SSC is investigated.

General remarks

Employing a different controller setpoint vector during the high influent ammonia event, considerably shortens the time during which the system is affected by the disturbance (compare Figures F.6 and F.7-F.9). To do this automatically requires a detection and classification approach that can detect and discern different disturbances so that the appropriate setpoints can be determined.

Whether this determination is based on empirical knowledge (such as a look-up table) or models is a compromise between simplicity and flexibility. Also, it is important that the determination is robust. The look-up table is simple, but not flexible and it is sensitive to an incorrect classification. The model-based approach, however, is flexible but more complex.

In this work the MPC is only implemented in extreme situations. However, since the reduced system model is there and it needs to be continuously updated, why not use it all the time? This is of course possible if the model covers all possible operational conditions. Then the monitoring and classification algorithms are used to determine appropriate weighting matrices, i.e. Q, R_1 and R_2. For instance, during normal operational conditions, the aim of the control may be to minimise the cost provided the requirements on the effluent water quality are fulfilled. In an extreme situation, the operational cost is of less importance and therefore R_1 and R_2 are reduced or set to zero.

Conclusions and future work

It has been shown that the framework based on operational state classification combined with model based setpoint determination shows promising results for supervisory control of wastewater treatment operation during extreme events. Relatively coarse model reductions, including just a few states and parameters, can be employed to determine appropriate setpoints. This is possible since the disturbances are identified and models designed to suit the specific disturbances are assigned to calculate new controller set points.

Only fast disturbances have been considered here. This is because these are the ones most crucial to detect and control as fast as possible. Slower disturbances can also be approached in the same way. However, if there are several temporal layers in the supervisory control framework, coordination of the local controller becomes ever so important. Multi-time scale coordination of local controllers is a challenging problem and a topic for future work. Another interesting area for further studies is automatic model reduction. Automatic temporal model reduction would fit well into the framework, and work is currently carried out in that direction.

Appendix

Reduced model for nitrogen control

The reaction rates of the reduced model used in the case study presented in this paper are expressed by:

$$\begin{aligned}
\frac{dS_S}{dt} &= \left(-r_H \frac{S_S}{K_S + S_S} \frac{S_O}{K_O + S_O} \right. \\
&\quad \left. -r_H \frac{S_S}{K_S + S_S} \frac{K_O}{K_O + S_O} \frac{S_{NO}}{K_{NO} + S_{NO}} + \alpha \right) \eta_{BH} TSS \qquad \text{(F.3)} \\
\frac{dS_{NO}}{dt} &= -n_H r_H \frac{S_S}{K_S + S_S} \frac{K_O}{K_O + S_O} \frac{S_{NO}}{K_{NO} + S_{NO}} \eta_{BH} TSS + \\
&\quad r_A \frac{S_{NH}}{K_{NH} + S_{NH}} \frac{S_O}{K_O + S_O} \eta_{BA} TSS \qquad \text{(F.4)} \\
\frac{dS_{NH}}{dt} &= -r_A \frac{S_{NH}}{K_{NH} + S_{NH}} \frac{S_O}{K_O + S_O} \eta_{BA} TSS \qquad \text{(F.5)}
\end{aligned}$$

The parameters of the model are assumed to be known from identification or literature. An explanation of all model parameters as well as their values is given in Table F.3.

Parameter	Explanation	Value
r_H	Reaction rate factor for heterotrophs (d^{-1})	5.97
r_A	Reaction rate factor for autotrophs (d^{-1})	2.08
n_H	Factor for anoxic growth of heterotrophs	0.115
K_S	Half-saturation constant for heterotrophs (mg COD d-1)	10.0
K_O	Oxygen half-saturation constant (mg COD d^{-1})	0.4
K_{NO}	Nitrate half-saturation constant for denitrifying heterotrophs (mg NO-N d^{-1})	0.5
K_{NH}	Ammonia half-saturation constant for autotrophs (mg NO-N d^{-1})	1.0
η_{BH}	Fractionation factor for heterotrophs in TSS	0.78
η_{BA}	Fractionation factor for autotrophs in TSS	0.046
α	Conversion rate factor from heterotrophs to readily biodegradable substrate (d^{-1})	0.225

Table F.3: Explanation and values for the parameters of the reduced system model.

State and influent concentration estimates

The estimations of non-measurable states and influent concentrations are based on simplified mass-balance calculations and correlations between variables:

$$\hat{S}_{s,1} = \frac{Q_{in}\hat{S}_{S,in}/V + Q_C S_{S,C}/V + 0.1TSS}{(Q_{in} + Q_{ret} + Q_{rec})/V + 0.15TSS} \quad \text{(F.6)}$$

$$\hat{S}_{S,in} = 2S_{NH,in} \quad \text{(F.7)}$$

$$\hat{S}_{S,2} = 2 \quad \text{(F.8)}$$

$$\hat{S_{S,3}} = 1 \quad \text{(F.9)}$$

$$\hat{S}_{NH,1} = \frac{Q_{in}S_{NH,in} + (Q_{ret} + Q_{rec})\, S_{NH,5}}{Q_{in} + Q_{ret} + Q_{rec}} \quad \text{(F.10)}$$

$$\hat{S}_{NH,2} = \hat{S}_{NH,1} - 2S_{O,2} \quad \text{(F.11)}$$

$$\hat{S}_{NO,2} = S_{NO,1} + 2S_{O,2} - 1.5 \quad \text{(F.12)}$$

Paper F

Addendum

Parameter sensitivity

It was shown in the paper that the supervisory controller was relatively insensitive to parameter errors. The fact that the control objective is to return to normal process operation is the main reason for this. The SSC is even less sensitive, which is an important feature of SSC. In Figure F.11, the result of the same manipulation of the parameter values as is done for the MPC in the paper, is shown using SSC.

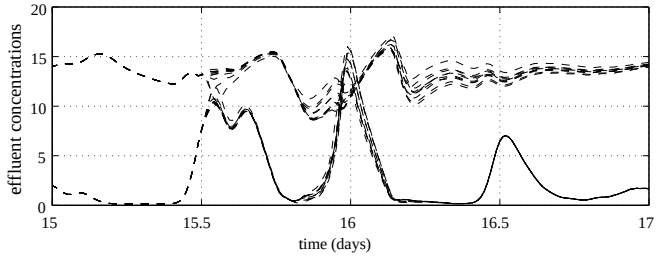

Figure F.11: Effluent nitrogen concentration for a number of simulations using SSC where the model parameters are afflicted with random errors (compare with Figure F.10)

MPC implementation

It is not obvious from the paper how the MPC is implemented. A full description of the implementation would be extensive and is outside the scope of this work. However, there are standard ways to implement linear MPC and algorithms are included in many application and research software packages. For the implementation in Paper F, MATLAB/SIMULINK and its function `constr.m` (or `fmincon.m`) are used.

The procedure for MPC can be summarised in a few steps. 1) Before implementing MPC, a process model is needed. Various model structures, describing the relations between controller outputs and process outputs, may be used. 2) A cost (or loss) function that relates deviation from the reference value, use of control, violation of constraints, etc., is determined. 3) At each updating instance, an optimisation is carried out online, using the current state as initial condition. MPC is often implemented with a control horizon as well as a prediction horizon. Up till the control horizon, each control vector may be varied. However, between the control horizon and the prediction horizon the control vectors are kept constant. The control horizon is often limited to just a few updating instances, whereas the prediction horizon is chosen similar to the settling time of the dominant dynamics. It is evident that a long control horizon makes the optimisation task computationally demanding. 4) The optimal (in terms of minimising the cost function) choice of control is applied to the process. At the next updating instance, steps 3 and 4 are repeated.

MPC is today a standard tool for control engineers and many text books are available on the subject (e.g. Camacho and Bordons (1999)). Overviews of MPC can also be found in, e.g. Morari and Lee (1999) and Mayne et al. (2000).

Supervisory controller

The supervisory control structure is not emphasised in the paper. This is instead done in Paper E. If Paper F is read independently of Paper E, it is somewhat unclear how the monitoring/classification part is coupled to the setpoint determination part. As is described in Paper E, the output of the setpoint determination is weighted using the membership function obtained from the mon-

itoring/classification algorithm. In this paper, crisp classification was used to facilitate comparison between the different setpoint determination strategies, i.e. look-up tables, SSC and MPC. In a real application, a fuzzy membership function is probably the most appropriate choice, since it provides seamless transition between operational states.

Applicability

It is not clear that the dynamic setpoint determination scheme presented in this paper improves the performance sufficiently to compensate for the large increase in complexity. It is the author's opinion that it may be one step too far, unless there are already models available for other reasons (e.g. simulation models for process analysis). Moreover, it is probably easier to get acceptance among operators, if a look-up table based on their own experiences is used.

Paper G

A chemometric approach to supervisory control of wastewater treatment operation

C. Rosen and U. Jeppsson

Based on Rosen and Jeppsson (2001b)
J. Chemometr. (submitted).

Abstract: *In this paper, a supervisory controller for wastewater treatment processes is presented. The controller is a steady-state controller with a PCA model representing the process. The objective is to control the effluent nitrogen concentration in terms of ammonia and nitrate from a five reactor activated sludge pre-denitrification plant. A PCA model is identified from data on a set of manipulated variables, process variables and one or several output/target variables. New data are projected onto the model and the difference in the principal component space between the desired location and the current location is computed. The control law is expressed in terms of changes in manipulated variables by mapping the difference in PC-space onto the measurement space. The resulting controller is a multivariate controller with integral action only. To compensate for model errors due to changes in the controlled process, identification in open loop, nonlinearities and controller saturations, a compensation term is applied. This term expresses the the difference between the setpoint and the current location and is implemented in a PI fashion. The fact that the controller only describes the steady-state relationship between the variables may appear disturbing. From a water recipient point of view, the mean load over longer time periods is more important and, hence, less consideration should be taken to fast disturbances. However, since grab sample strategies sometimes are used to verify that the output quality requirements are*

met, a feed-forward term can be included in the controller, with a significant decrease in the output quality variation as a result. Simulation studies show that the controller can be used to control the process to desired output setpoints as well as to reduce the quality variation significantly.

Keywords: PCA, process control, supervisory control, wastewater treatment.

Nomenclature

A	system matrix in state-space representation
B	input matrix in state-space representation
C	output matrix in state-space representation
K_Δ	controller gain
K_I	controller gain, integral part of compensation term
K_P	controller gain, proportional part of compensation term
$\mathbf{P}$	loading matrix
$\mathbf{P}_m$	part of loading matrix associated with $\mathbf{Z}_m$
$\mathbf{P}_X$	part of loading matrix associated with $\mathbf{Z}_x$
$\mathbf{P}_Y$	part of loading matrix associated with $\mathbf{Y}$
Q_{rec}	nitrate recirculation flow rate
S_{NH}	ammonia concentration
S_{NO}	nitrate concentration
S_S	substrate concentration
S_O	dissolved oxygen concentration
$\mathbf{T}$	score matrix
$\mathbf{t}$	score vector
$\mathbf{t}_{sp}$	score setpoint
u	local control signal
u_s	supervisory control signal
$\mathbf{X}$	process data matrix
$\mathbf{x}$	process data vector
$\mathbf{Y}$	process output data matrix
$\mathbf{y}$	process output data vector
$\mathbf{y}_{sp}$	process output setpoint
$\delta\mathbf{y}$	process output control error
$\mathbf{Z}_m$	manipulated process data matrix
$\mathbf{Z}_X$	non-manipulated process data matrix
$\mathbf{z}_m$	manipulated process data vector
$\mathbf{z}_{m,sp}$	manipulated process data setpoint

Introduction

Wastewater treatment plant (WWTP) operation is subject to a number of effluent water quality standards, which are becoming increasingly strict. From a control point of view, not only the absolute effluent concentrations are of interest. The way the requirements are policed will have great impact on the control strategy. In Sweden and some other countries, average values over weeks and months are used to evaluate the compliance with the requirements. However, in other countries, for example Germany, grab samples are used. Common to both strategies is that the plant pays a discharge fee for excessive discharge. In a few countries (e.g. Denmark) the plant pays for every discharged unit. It is obvious that the control strategies used are dependent on what type of legislation is used. In Sweden, the operation aims at meeting the requirements in average and less emphasis has to be put on daily or weekly quality variations. In Germany it is important to keep the effluent quality within the requirements at all times and in Denmark there must be a trade-off between control and discharge costs.

Whatever legislation used, the need to control the plant in a coordinated way is significant. This involves determining appropriate setpoints for local controllers to achieve target or quality requirements put on the overall process. This task is often referred to as supervisory control. Compared to local control, supervisory control distinguishes itself as high level control and is often carried out by operators. To help the operators, monitoring and surveillance systems are used to monitor the process and to detect and isolate faults and disturbances. During normal conditions, the local controllers (typically PID controllers) will often, with appropriate setpoints, yield a result, which meets the requirements on the process without significant changes in the controller setpoints. However, there are situations when this is not true. Obviously, during abnormal conditions or extreme events, local controllers cannot compensate for the disturbances (this is normally how abnormal conditions are defined) and the setpoints need to be corrected and/or new actuators have to be initiated to meet the requirements or to minimise the effect of a disturbance. Extreme event supervisory control in wastewater treatment has been addressed in other publications (Rosen and Yuan; 2000; Rosen et al.; 2001). However, often the process is subjected to slow changing disturbances (e.g. ambient temperature and seasonal effects), which must be considered as normal, and in these cases corrections must be made

to the setpoints. Another example is when the overall process requirement is changed. When the number of controllers is high and the process is complex, for example with recirculation streams, it is not obvious how to correct the set points to obtain the desired output.

Chemometric methods for process control have been addressed before in the literature (Kaspar and Ray; 1992, 1993; Piovoso and Kosanovich; 1994; Chen and McAvoy; 1996; Chen et al.; 1998). It is a natural step to counteract deviations within the process using the information from the monitoring algorithm. The are, however, some difficulties that need to be considered. Using information from the monitoring algorithm to control the process will introduce a new situation and the process system can no longer be considered as an open-loop system, since actions based on the information closes the loop. The closed-loop system will have different characteristics compared to the open-loop system, such as a change in the covariance structure of the process, process gain as well as system dynamics (Chen et al.; 1998; Pasadyn et al.; 1999). If a model is identified in open loop, these changes will yield a model error that must be compensated for in order to control the process to a certain process output setpoint. Another difficulty has to do with applying a linear model to nonlinear systems. The nonlinearities may be inherent in the process, but may also be introduced by controller saturation or physical limitations. Process changes due to varying operational conditions as well as measurement disturbances are also difficulties that must be addressed.

In this paper a chemometric approach to supervisory control of an activated sludge WWTP is presented. The difficulties mentioned above are addressed by introducing a compensation term, which compensates for model errors and controller saturation. The approach is based on a multivariate feedback control law, using a principal component analysis (PCA) model as a steady-state representation of the process. The principal component space (PCS) supervisory controller is implemented on top of the local PID control systems. Simulations using the international benchmark system developed within the European COST action 624 collaboration(Pons et al.; 1999; COST624; 2001), is used to investigate the applicability of such a supervisory controller. The aim of the supervisory controller is to control the effluent nitrogen concentration in terms of ammonia and nitrate, both in terms of daily average effluent concentration, but also in terms of variation and maximum values.

Multivariate controller in PC-space

The approach presented below is similar to that of Piovoso and Kosanovich (1994). However, instead of using PCA on which an output Y is regressed (i.e. PCR), we include the output variable/variables in the PCA. This yields a more direct algorithm, but the difference in performance is small. Another difference is that we include a model error compensation term, which will be discussed later.

Basic PCS controller

Assume that during different operational conditions, a number of process variables, $\mathbf{X}$, can be measured in the process. In $\mathbf{X}$, let $\mathbf{Z}_m$ be manipulated process variables, $\mathbf{Z}_x$ be the non-manipulated process variables and $\mathbf{Y}$ be the process output variables. $\mathbf{X}$ can the be partitioned as:

$$\mathbf{X} = [\mathbf{Z}_m|\mathbf{Z}_X|\mathbf{Y}] \tag{G.1}$$

A PCA on X then gives:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T \tag{G.2}$$

This equation is written as:

$$\mathbf{X} = \mathbf{T}\,[\mathbf{P}_m|\mathbf{P}_X|\mathbf{P}_Y] \tag{G.3}$$

where

$$\mathbf{P}^T = [\mathbf{P}_m|\mathbf{P}_X|\mathbf{P}_Y]$$

If $\mathbf{t}_{sp}$ is the desired operating point in the PC-space, the corresponding point in the measurement space is then:

$$\mathbf{x}_{sp} = \mathbf{t}_{sp}\mathbf{P}^T \tag{G.4}$$

The difference between the current location and the desired position in the PC-space at time k is denoted $\Delta\mathbf{t}_k$. Then the difference can be transformed to the measurement space as:

$$\begin{aligned} \Delta\mathbf{x}_k &= \Delta\mathbf{t}_k\mathbf{P}^T \\ &= (\mathbf{t}_{sp} - \mathbf{t}_k)\,\mathbf{P}^T \end{aligned} \tag{G.5}$$

The difference in the PC-space can be expressed as a difference in $\mathbf{z}_m$:

$$\Delta\mathbf{z}_{m,k} = \Delta\mathbf{t}_k\mathbf{P}_m \tag{G.6}$$

It is now possible to design a control law based on the desire to drive the process from its current location in the PC-space to a desired location. The control law is incremental so that the change is added to the previous value at each control step:

$$\begin{aligned}\mathbf{z}_{m,k+1} &= \mathbf{z}_{m,k} + \Delta\mathbf{z}_{m,k} \\ &= \mathbf{z}_{m,k} + (\mathbf{t}_{sp} - \mathbf{t}_k)\,\mathbf{P}_m\end{aligned} \tag{G.7}$$

This control law inherently has integral action and, consequently, the controller error (in the PC-space) will be forced to zero in steady state. One possible strategy is to set the desired location in the PC-space to the origin, i.e. the control strives to keep the process as 'normal' as possible. Another way, albeit cumbersome when $\mathbf{P}$ spans more than just a few dimensions, is to locate the process in desired regions using a score plot. However, it is desirable to control the location in the PC-space to different locations depending on the current situation. Since most control is carried out in order to comply with output requirements, the output is used as target. Let $\mathbf{y}_{sp}$ denote the desired output. A corresponding location in the PC-space is found by solving:

$$\mathbf{t}_{sp}\mathbf{P}_Y = \mathbf{y}_{sp} \tag{G.8}$$

This is an under-determined system (assumed that there are a higher number of scores than output variables) and the solution will be a hyperplane. Using the pseudo inverse of $\mathbf{P}_y$ yields a solution with the smallest Euclidean norm. This is a reasonable choice since it results in a solution that gives the desired output as well as minimises the distance from the origin to the process location in PC-space. The setpoint in PC-space becomes:

$$\mathbf{t}_{sp} = \mathbf{y}_{sp}\mathbf{P}_Y^{\dagger} \tag{G.9}$$

where $\mathbf{P}_Y^{\dagger}$ is the pseudo inverse of $\mathbf{P}_y$. The relation between the setpoint and the manipulated variables then becomes:

$$\mathbf{z}_{m,k+1} = \mathbf{z}_{m,k} + K_\Delta\left(\mathbf{y}_{sp}\mathbf{P}_Y^{\dagger} - \mathbf{t}_k\right)\mathbf{P}_m \tag{G.10}$$

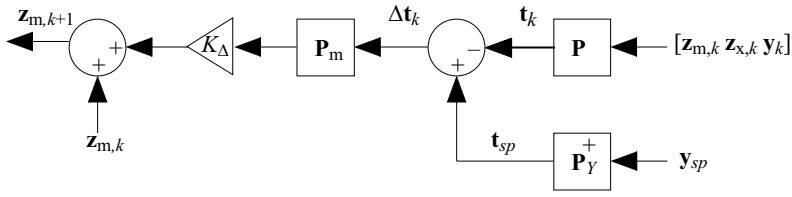

Figure G.1: Basic PCS controller structure.

where K_Δ is the gain constant that determines the fraction of the implemented change. It should be noted that the discussion above assumes that the process output can be considered a function of process variables and, consequently, the choice of variables is crucial.

Because the controller described above is a steady-state controller, the controller moves may be too aggressive. By implementing only a fraction of the computed change this can be overcome. Piovoso and Kosanovich (1994) suggest that if the system dynamics are known, the appropriate gain can be determined by use of pole placement techniques. The basic controller structure is illustrated in Figure G.1.

PCS controller with compensation

The basic PCS controller does not yield an output that equals the set point. There are several reasons for this. Firstly, there will certainly be a discrepancy between the PCA model and the system, either due to identification difficulties or due to unknown variations in process parameters. Secondly, as mentioned before, closing the loop will change the behaviour of the system. This model error cannot be eliminated by the integral action since the control law is expressed in the model domain (PC-space). A solution to this is to introduce a compensation term in Equation G.10:

$$\begin{aligned} \mathbf{z}_{m,k+1} &= \mathbf{z}_{m,k} + K_\Delta \Big[\Big(\mathbf{y}_{sp}\mathbf{P}_Y^\dagger - \mathbf{t}_k \Big) \mathbf{P}_m + \\ &\quad \delta\mathbf{y}_k \mathbf{P}_Y^\dagger \mathbf{P}_m \Big] \end{aligned} \tag{G.11}$$

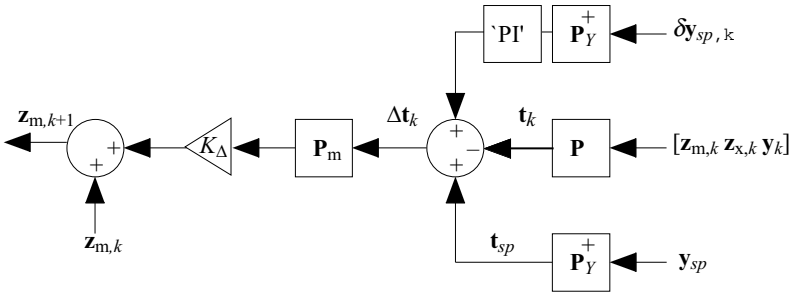

Figure G.2: PCS controller structure with a compensation term.

where

$$\delta \mathbf{y}_k = \mathbf{y}_{sp} - \mathbf{y}_k$$

Equation G.11 implies that the setpoint $\mathbf{y}_{sp}$ will be moved slightly to compensate for the model error. To eliminate small errors the term can be computed in a proportional/integral (PI) fashion:

$$\begin{aligned} \mathbf{z}_{m,k+1} = \mathbf{z}_{m,k} + K_\Delta \Big[\Big(\mathbf{y}_{sp} \mathbf{P}_Y^\dagger - \mathbf{t}_k \Big) \mathbf{P}_m + \\ K_P \delta \mathbf{y}_k \mathbf{P}_Y^\dagger \mathbf{P}_m + K_I \sum \delta \mathbf{y}_k \mathbf{P}_Y^\dagger \mathbf{P}_m \Big] \end{aligned} \quad \text{(G.12)}$$

K_P is the proportional gain of the compensation term and K_I is the discrete integral gain. When tuning the controller parameters one needs to pay attention to the fact that the controller contains two integrators if a PI configuration is used. The structure of the controller with a compensation term is shown in Figure G.2.

Example

To demonstrate the properties of the compensation term, the basic controller is exemplified using an oscillatory but stable linear system. Consider the 5:th order linear MISO system:

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx \end{aligned}$$

where

$$A = \begin{bmatrix} -0.9505 & 0.6142 & -0.9289 & -1.5075 & 1.7737 \\ -1.0977 & -1.2835 & 1.0512 & 0.9968 & -0.2439 \\ 1.1866 & -0.4332 & -1.0984 & 0.8531 & 0.1399 \\ 1.7525 & -0.8986 & 0.3685 & -1.2803 & -0.0615 \\ -0.9873 & 1.4109 & 0.4892 & -0.1558 & -1.0668 \end{bmatrix}$$

$$B = \begin{bmatrix} 0.0412 & -0.9812 & -0.5062 & 0 \\ -0.7562 & -0.6885 & 1.6197 & 1.9375 \\ -0.0891 & 1.3395 & 0 & 0 \\ -2.0089 & -0.9092 & -1.0811 & -1.2559 \\ 1.0839 & -0.4129 & -1.1245 & -0.2135 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Let u be the manipulated variables ($\mathbf{z}_m$) in Equation G.3 and y be the process output ($\mathbf{y}$). A PCA model can be identified by exciting the system through u in open loop. The model, $\mathbf{P}$, is decomposed into $\mathbf{P}_m$ and $\mathbf{P}_Y$ and the controller is implemented according to Equation G.12 (non-manipulated variables ($\mathbf{z}_x$) are left out). To mimic a small change in the process, a disturbance in the system matrix A is introduced:

$$\delta A = \begin{bmatrix} 0.0858 & -0.0400 & 0.0669 & -0.1604 & 0.0529 \\ 0.1254 & 0.0690 & 0.1191 & 0.0257 & 0.0219 \\ -0.1594 & 0.0816 & -0.1202 & -0.1056 & -0.0922 \\ -0.1441 & 0.0712 & -0.0020 & 0.1415 & -0.2171 \\ 0.0571 & 0.1290 & -0.0157 & -0.0805 & -0.0059 \end{bmatrix}$$

so that the new system matrix is $A + dA$. The disturbed system displays qualitatively the same behaviour as the undisturbed system. In Figure G.3, a step change in the setpoint for y is shown.

As can be seen, the controller will not yield the desired output. Now, the compensation term is implemented as the integrated error of the setpoint, i.e. as an I-controller ($K_\Delta = 0.9$, $K_P = 0$ and $K_I = 0.5$). The figure shows that the desired output is obtained. The over-shoot is due to the integral action in the compensation term, but can be reduced at the cost of transient speed. Note that the implemented controller is purposely delayed to avoid algebraic loops. Consequently, the example should not be interpreted analytically.

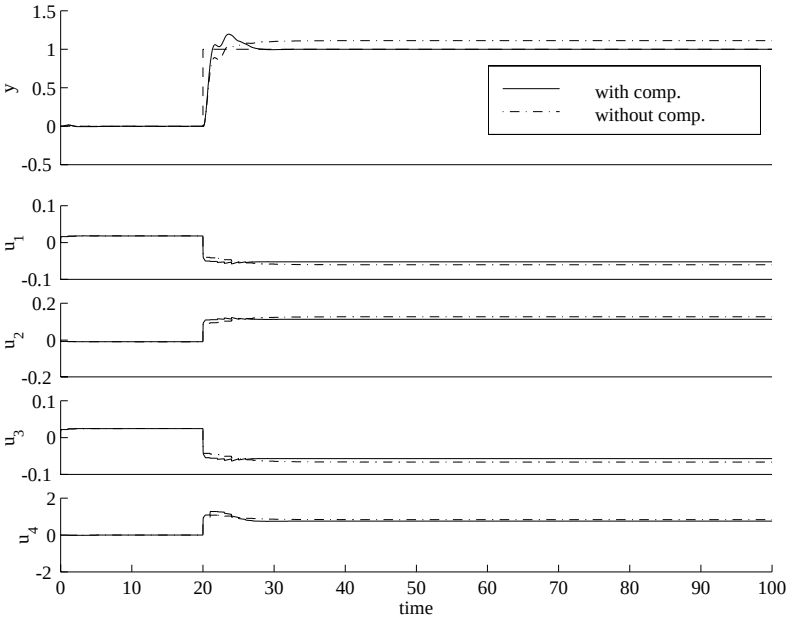

Figure G.3: Output response for a step change in reference value (top) and the corresponding controller outputs (bottom). Without compensation (dash-dotted) and with compensation (solid).

Supervisory controller

To use the above-discussed controller for supervisory control, only minor modifications must be made to the algorithm. The supervisory controller is used to derive appropriate setpoints for local controllers. If the local control is perfect and the controller response can be considered instantaneous compared to the supervisory controller sampling time, the only modification to the algorithm is to replace $\mathbf{z}_m$ with $\mathbf{z}_{m,sp}$ in Equation G.12. However, this is rarely applicable since there may be controller saturation due to limitations imposed on the controllers. This implies that $\mathbf{z}_m \neq \mathbf{z}_{m,sp}$ and it may therefore be wise to implement Equation G.12 so that the old setpoint of the local controllers is

replaced by the actual value of the manipulated variables:

$$\begin{aligned}\mathbf{z}_{m,sp,k+1} &= \mathbf{z}_{m,k} + K_\Delta \left[\left(\mathbf{y}_{sp}\mathbf{P}_Y^\dagger - \mathbf{t}_k \right) \mathbf{P}_m + \right. \\ &\quad \left. K_P \delta\mathbf{y}_k \mathbf{P}_Y^\dagger \mathbf{P}_m + K_I \sum \delta\mathbf{y}_k \mathbf{P}_Y^\dagger \mathbf{P}_m \right] \end{aligned} \tag{G.13}$$

Obtaining a model $\mathbf{P}$

To obtain a model that describes the plant in an appropriate way, the need for experimental planning and design increases as the number of manipulated variables increases. It is important to keep in mind that the models identified using the multivariate techniques discussed in this paper are based on correlation between variables and not on cause-effect relationships. The structure of the identification excitations is therefore extremely important. Identification of complex processes is a large topic of its own and will not be discussed in here.

Dynamics

The supervisory control approach described above can be regarded as a multivariate feedback controller, based on the inverse steady-state relation between the local controller outputs, process measurements and the process outputs. This implies that the controller has some limitations to what can be achieved in a dynamic situation. With a feedback configuration based on a steady-state representation of the process, the closed system cannot be made arbitrarily fast. The implication of this is that disturbances slower than the dominant time constants of the system can be reduced or eliminated by the controller. However, faster disturbances cannot be attenuated, since the controller response is not fast enough.

A way to circumvent this limitation is to introduce a feed-forward term in Equation G.13. A feed-forward term can easily be incorporated in the expression for $\delta\mathbf{y}_k$. Then

$$\delta\mathbf{y}_k = \mathbf{y}_{sp} - \hat{\mathbf{y}}_k \tag{G.14}$$

where

$$\hat{\mathbf{y}}_k = f\left(\mathbf{y}_k, \mathbf{x}_k, \dots\right)$$

If the estimation of $\hat{\mathbf{y}}_{sp}$ is done using upstream measurements, the model error is estimated beforehand resulting in a faster controller response.

Model validity

The validity of the controller model can be monitored (Chen and McAvoy; 1996) using normal monitoring procedures such as the sum of the squared prediction error (SPE) (Kresta et al.; 1991). The SPE is useful, since it indicates if the current operational point is close to the model plane. Thus, monitoring the SPE may provide information whether the model is likely to produce appropriate control actions and is recommended to avoid inappropriate controller setpoints. When used in conjunction with control limits, the SPE can also be used to invoke predefined controller setpoints from look-up tables to handle unknown or uncertain situations. However, the distortion of the model due to closing of the loop, controller saturations, etc. makes it difficult to use the confidence limit calculated from the identification phase. If limits are to be used, these generally need to be calculated using data from the closed loop system.

Simulation of wastewater treatment processes

In this section, the simulation model used to test and to exemplify the applicability of the proposed supervisory controller is briefly described. A number of test scenarios, which represent frequently occurring phenomena, are defined and a short description of the implemented supervisory controller is given.

Simulation model

The simulation model used is a benchmark model developed within the collaboration of COST Action 624 (Pons et al.; 1999). The COST model is based on the Activated Sludge Model No 1 (ASM1) (Henze et al.; 1987) and a ten layer Takacs (Takács et al.; 1991) settler model. The nature of wastewater treatment processes is nonlinear, which is reflected in the simulation model. The complete model comprises more than 200 states, with time constants in the

range of minutes to months. The complete configuration, including simulation model parameter values, can be found at the COST 624 web site (COST624; 2001).

Plant configuration

The simulated plant comprises five completely mixed activated sludge reactors and a settler. Thus, only the biological stage of a WWTP is considered. The plant configuration is shown in Figure G.4. The measured and manipulated variables are listed in Table G.1. Note that in addition to constraints within the local controllers, limits are imposed on the calculated setpoints to assure in practice achievable values. Also, a dead band is used for the dissolved oxygen controllers, since aeration below 0.5 mg/l is seldom used in practice.

Simulation scenarios

To test the applicability of the supervisory controller, a number of scenarios are simulated and evaluated. The scenarios involve a number of different types of disturbances that may be encountered and that have to be accepted within normal operational conditions.

I. *Varying effluent nitrogen setpoint.* To test the controller and evaluate its ability to track setpoint changes, a sequence of different controller setpoints are applied. The minimum and maximum setpoint, respectively, are considered to be on the fringe of normal effluent discharge.

II. *Inhibition of nitrification.* A very important parameter for the operation of WWTPs is the nitrification rate (μ_A), that is, how fast ammonia is converted into nitrate. Inhibition occurs more or less frequently in many plants (due to pH changes or toxicity etc.). In the model, a nitrification rate of $\mu_A = 0.5$ is used as the nominal value. In the scenario, the nitrification rate is changed at day 3 from its nominal value to $\mu_A = 0.35$ and then linearly increasing to its nominal value over a period of ten days.

III. *Measurement disturbance.* To evaluate the robustness of the controller, an error is imposed on the substrate ($S_{S,1}$) measurement in the first reactor.

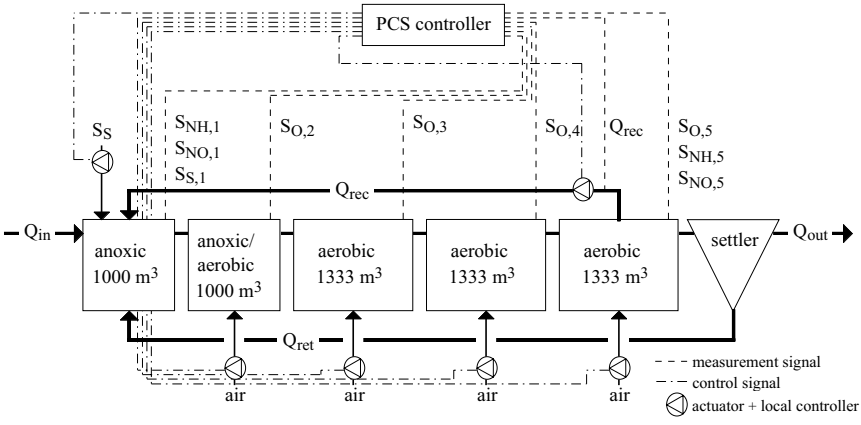

Figure G.4: Simulated plant configuration with the measurement and control signals indicated.

Variable	Notation	Meas.	Manip.	Range	Control
Ammonia reactor 1 [mg N/l]	$S_{NH,1}$	yes	no	-	-
Nitrate, reactor 1 [mg N/l]	$S_{NO,1}$	yes	no	-	-
Substrate, reactor 1 [mg COD/l]	$S_{S,1}$	yes	yes	0-30	u_1
Dissolved oxygen, reactor 2 [mg O_2/l]	$S_{O,2}$	yes	yes	0, 0.5-5	u_2
Dissolved oxygen, reactor 3 [mg O_2/l]	$S_{O,3}$	yes	yes	0, 0.5-5	u_3
Dissolved oxygen, reactor 4 [mg O_2/l]	$S_{O,4}$	yes	yes	0, 0.5-5	u_4
Dissolved oxygen, reactor 5 [mg O_2/l]	$S_{O,5}$	yes	yes	0, 0.5-5	u_5
Ammonia reactor 5 (effluent) [mg N/l]	$S_{NH,5}$	yes	no	-	-
Nitrate, reactor 5 (effluent) [mg N/l]	$S_{NO,5}$	yes	no	-	-
Nitrate recirculation rate [m^3/d]	Q_{rec}	yes	yes	0-1.2 E5	u_6

Table G.1: Measured and manipulated variables.

The nature of the disturbance imitates a drift in the sensor, with a linearly increasing offset starting at 0 mg COD/l at day 3 and reaching 2 mg COD/l at day 7. From day 7 to day 11 the offset is constant. At 2 mg COD/l, the disturbance constitutes approximately 45 % of the normal concentration in reactor 1 and, consequently, the disturbance must be considered as severe.

IV. *Control handle loss*. Sometimes a control handle is lost due to malfunction of equipment. At day 15 the aeration in reactor 4 is lost, that is, no air is added. The challenge for the controller is to compensate for this.

V. *Varying influent characteristics*. The most dominant disturbance in wastewater treatment is the variation in the influent wastewater characteristics (as a matter of fact, this is normally considered as the state of things rather than a disturbance). One of the test files in the COST benchmark, dry weather data, is used to create the diurnal and weekly patterns. The variations are faster than the hydraulic retention time of the plant and, thus, the dynamic behaviour of the controller can be evaluated.

VI. *Combination of scenarios II, III, IV and V*. As the last scenario, varying influent conditions is combined with the disturbances discussed above to evaluate the ability of the controller to handle multiple disturbances.

Supervisory controller

To identify a controller model, a data training set is used. Training data comprise operation during different local controller setpoints and with constant influent conditions. The setpoints of the training set are chosen so that the desired mean of the effluent nitrogen concentration is obtained (10 mg N/l). This is done to ensure that the desired operating point will be close to the origin of the PC-space. A PCS controller is developed according to the algorithm discussed above. Five principal components are retained and together they capture approximately 90 % of the variations in training data. The variables included are listed in Table G.1. However, nitrate and ammonia in reactor 5 are added up and constitute the controlled variable $\mathbf{Y}$ in Equation G.1. Since the controller will be applied to varying influent conditions, a simple feed-forward term is included in the controller. The effluent nitrogen concentration is estimated as

proportional to the nitrogen concentration in reactor 1. To avoid unnecessary variation in the local controller setpoints, the output of the controller is filtered with a linear first order filter using a time constant of 0.01 days. The parameters of the supervisory control are chosen so that a balance between controller speed and control signal variance is obtained.

Results and discussion

In this section, the result of the simulation study is presented. For scenarios I to IV, that is the scenarios with constant influent wastewater characteristics, no feed-forward term is included in the controller. For scenarios V and VI, the feed-forward term is included to deal with the dynamic disturbances. In all scenarios, the PCS controller is implemented with a compensation term according to Equation G.13.

Scenario results

In Figure G.5, the result of the first scenario is shown. The PCS controller can track the setpoint changes without large deviations. This is achieved by alterations in all the local controller setpoints. It is interesting to note the different behaviour of the transient at the setpoint changes. This is a clear indication of the nonlinear nature of the system. For comparison, the result of a PCS controller without compensation term is included. Here, it is obvious that the controller has problems in tracking the setpoint. Note that the controller error is not a constant offset and that it changes signs. There are a number of reasons why. Firstly, there will be a model error due to reasons discussed before. However, more important in this case are the controller output limits. Since some regions in the controller space are not usable, the output will deviate significantly from the setpoint.

The result of the inhibition of the nitrification rate in scenario II can be seen in Figure G.6. The impact on the effluent quality is reduced significantly using the PCS controller (mainly by coordination of the substrate concentration in reacor 1 and the nitrate recirculation). After the initial decrease in nitrification rate at day 3, the controller drives the output back to its setpoint within a day whereas

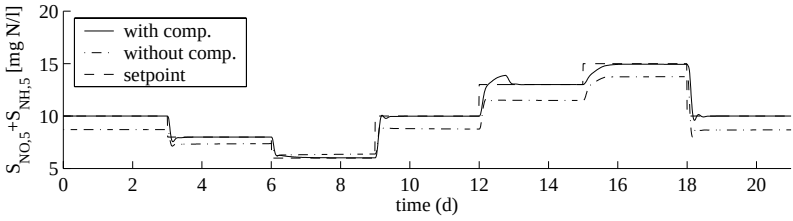

Figure G.5: *Scenario I.* Sequence of changing reference values for the effluent nitrogen concentration. Results with and without compensation term.

for the case without supervisory control the effluent concentration remains well above the setpoint for more than a week.

In scenario III, the robustness of the controller is challenged. From days 3 to 7, the increasing offset in the $S_{S,1}$ measurement results in a hardly visible deviation from the setpoint (Fig. G.7). However, when the offset is constant between days 7 and 11, the integral part of the controller is capable of meeting the setpoint without difficulties. When the offset is corrected, there is a transient of less than a day before the effluent concentration is back at its setpoint. Most local controller setpoints are varied significantly to attain the overall setpoint. In the case with no supervisory control, the measurement disturbance has great effect on the effluent nitrogen concentration, since the local controller does not provide the process with sufficient carbon (remember that the measurement disturbance results in an overestimation of the substrate concentration).

At day 15, scenario IV is applied to the system (Fig. G.7). The loss of aeration

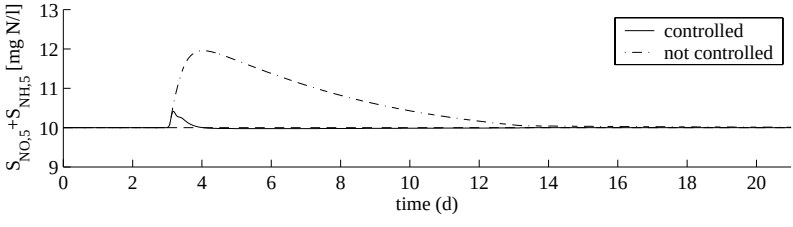

Figure G.6: *Scenario II.* Inhibition of nitrification with adaptation from day 3. Effluent nitrogen with supervisory controller (—) and without supervisory controller (--).

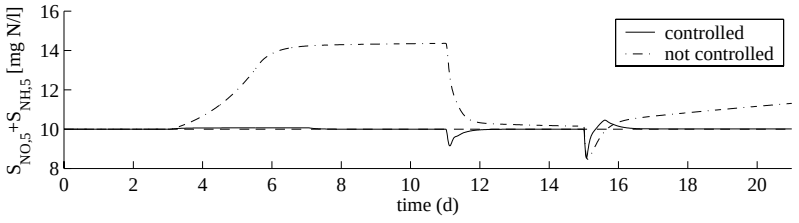

Figure G.7: *Scenarios III and IV*. Output concentration with $S_{S,1}$ measurement disturbance at days 3-11 and loss of control handle at day 15.

in reactor 4 leads (after a slight improvement due to system dynamics) to a deterioration of the effluent quality if no supervisory control is applied. In the controlled case the effluent concentration is restored in a day. The reason this is possible is that the dissolved oxygen concentration in reactor 4 ($S_{S,4}$) is closely correlated with the dissolved oxygen concentrations in other reactors and the controller compensates for the loss by increasing the aeration in reactors 3 and 5. Thus, the controlled system is redundant in this particular PC-direction. However, if the lost control handle is very dominant in a certain dimension in the PC-space, the controller would have great difficulties to compensate for this within the controller restrictions (if there are no controller restrictions or severe nonlinearities the controller would in theory be able to compensate for such a loss).

From the initial results, it is seen that the PCS controller is capable of handling different types of disturbances. However, as was mentioned before, the most severe disturbance is normally constituted by the influent characteristics. In most WWTPs, the control authority of the control system is limited in comparison to the disturbances imposed on the system through the influent and to completely attenuate the variation in the effluent quality would be extremely costly. However, by using supervisory control, the variation can be reduced considerably. In Figure G.8, the effluent nitrogen is shown when scenario V is applied to the system. To illustrate the importance of a feed-forward term, the resulting effluent concentrations of both a feedback (fb) and a feedback + feed-forward configuration are displayed (top panel). It is clearly seen that the variation is reduced when a feed-forward term is included. The feedback configuration marginally reduces the variation compared to the case when no supervisory control is applied (not shown for clarity reasons). An interesting

observation is that both configurations do not handle mid day influent peaks acceptably. The reason for this is discovered when the manipulated variables are studied. The increase in effluent nitrogen is substantial due to an increase of effluent ammonia. However, the dissolved oxygen concentration is lowered in both reactors 3 and 5. The intuitive reaction is that this is not a correct strategy, as conversion speed of ammonia to nitrate is roughly positively proportional to the dissolved oxygen concentration. A look at the SPE for the same period gives the answer (Fig G.9). During peak loads, the model displays poor fit and is, consequently, not appropriate for calculating controller setpoints. The explanation for this is that during peak loads, the operational state of the plant deviates substantially from that of the remaining time. Thus, the range of operating conditions in this scenario is too wide to be sufficiently described by the model. Introducing a second controller model or predefined setpoints for this operational state may decrease the peaks. The switching between controller models can be implemented with, for instance, a combination of PCA and fuzzy c-means clustering (FCM) as discussed in Rosen and Yuan (2000).

The result from the last scenario shows that the PCS controller can handle multiple disturbances. In Figure G.10, it is shown that the effect of the measurement disturbance in $S_{S,1}$ measurements between days 3 and 11 is hardly discernible. The loss of a control handle from day 15 manifests itself as an increase in the output concentration variation. The mean and maximum values, however, are not affected.

The controller is able to control the process so that the sequence of setpoints of scenario I is met during varying influent conditions (not shown). The only exception is during the very lowest setpoint between days 6 and 9. At these conditions, the discrepancy between the model and the process is simply too large, which manifests itself as high SPE values. Even though the setpoint was reachable during constant input, the varying conditions drive the process out of control. This is not surprising since the low setpoint is a rather extreme setpoint for the plant configuration used.

Controller performance

From an operational point of view, it is interesting to investigate the PCS controller performance according to a number of different criteria. However, it is

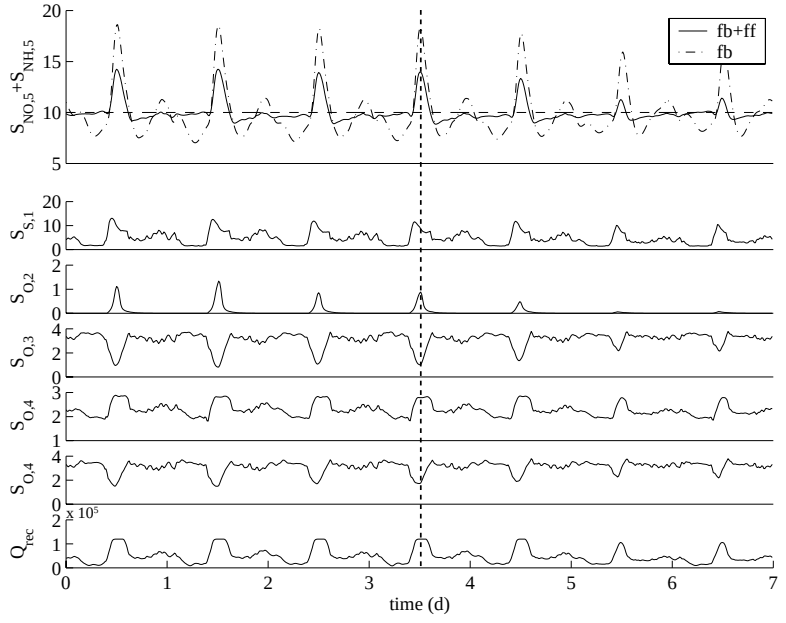

Figure G.8: *Scenario V*. Varying influent conditions during one week. Output concentration with feedback + feed-forward (fb+ff) configuration and only feedback (fb) configuration (top). Effluent total nitrogen setpoint 10 mg N/l. Manipulated variables in the fb+ff configuration case (below).

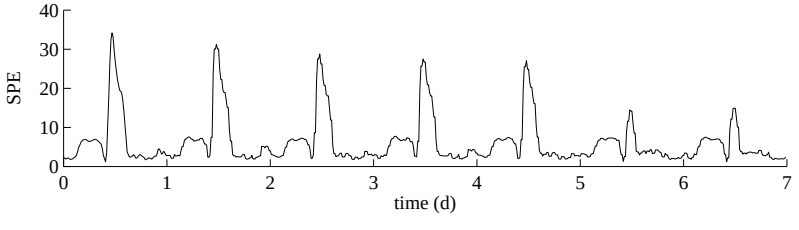

Figure G.9: *Scenario V*. SPE for feedback + feed-forward configuration. Due to the distortion of the model, the confidence limit calculated from the identification data cannot be used.

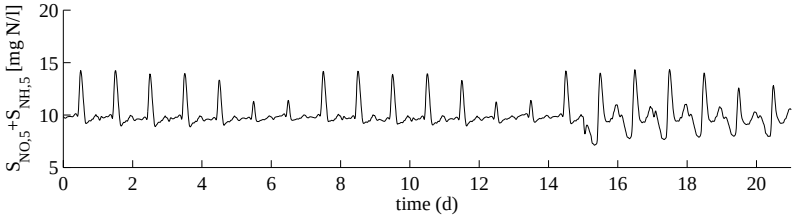

Figure G.10: *Scenario VI*. Output concentration during period of multiple disturbances.

important to point out that no consideration has been taken to evaluate how efficiently the various control handles are used. This is outside the scope of this work and is only briefly discussed below.

The PCS controller with and without feed-forward term is evaluated and compared with constant control (i.e. constant local control setpoints). In finding the local setpoints for the constant control case, the mean of the setpoints of the two PCS controllers is used and modified slightly to obtain the same output average. It must be pointed out that this choice may by all means not be optimal in terms of control costs. The constant control values should only be used for qualitative comparisons.

The controller configurations are evaluated in terms of effluent mean, max, min, 95-percentile and standard deviation (σ). The 95-percentile gives an indication on how far from a maximum limit the setpoint can be set. From Table G.2 it can be seen that the PCS controller with feedback only yields the desired mean and that it decreases the effluent quality variation. However, it does not perform especially well in terms of maximum and minimum values and the 95 percentile is located at the same value as in the non-controlled case. Thus, the only advantage with this configuration is that the mean can be controlled to a certain setpoint. The relative cost is not significantly higher, except for the aeration in reactor 2 (there is no aeration in reactor 2 when no supervisory control is used).

When a feed-forward term is used, the desired mean is achieved at the same time as the variation is reduced significantly. It can also be seen that the maximum value and especially the 95-percentile are far lower than if no supervisory control

Controller	Process output properties				
	Mean	Max	Min	95-perc.	σ
Constant	10.04	17.33	4.23	16.03	3.37
PCS(fb)	10.00	18.57	7.07	16.14	2.49
PCS(fb+ff)	10.00	14.21	8.84	12.79	1.01
	Relative control costs				
	u_1	u_2	u_3	u_4	u_5	u_6
Constant	1	0	1	1	1	1
PCS(fb)	1.06	-	1.06	0.97	1.03	1.06
PCS(fb+ff)	1.05	-	1.03	0.99	0.96	1.15

Table G.2: Comparison of controller performance for one week of operation during varying influent conditions.

Controller	Local control signals					
	$\sigma_{u,1}$	$\sigma_{u,2}$	$\sigma_{u,3}$	$\sigma_{u,4}$	$\sigma_{u,5}$	$\sigma_{u,6}$
Constant	2.14	0	43.8	48.3	63.1	0
PCS(fb)	3.27	35.9	38.5	51.9	70.8	20065
PCS(fb+ff)	3.29	55.2	44.5	57.0	57.8	29682
	Supervisory control signals (local control setpoints)					
	$\sigma_{us,1}$	$\sigma_{us,2}$	$\sigma_{us,3}$	$\sigma_{us,4}$	$\sigma_{us,5}$	$\sigma_{us,6}$
Constant	0	0	0	0	0	0
PCS(fb)	2.15	0.113	0.200	0.260	0.291	20065
PCS(fb+ff)	4.02	0.183	0.573	0.252	0.394	29682

Table G.3: Comparison of controller performance for one week of operation during varying influent conditions.

is used. This is important, since if a grab sample policing strategy is used, the setpoint can be set close to the maximum limit with lower control cost as a result. The relative costs are not affected significantly.

In Table G.3, the control signal standard deviation for each controller is listed. The local control signal standard deviation (σ_u) is not affected notably, which is slightly surprising since the variation in the supervisory controller control output is considerable. Note that the local controller signals and the supervisory controller signals have different units.

Practical aspects

The approach discussed above involves a number of practical difficulties of which the identification of the controller model (**P**) perhaps is the most problematic. To obtain a model, which describes the plant in an appropriate way, the need for experimental planning increases as the number of controlled variables rises. In this work, identification is carried out with constant influent characteristics. This is seldom possible in real wastewater operation. Therefore, it is important that the excitation of the process is sufficient so that the effects of the controller changes do not "drown" in the variation caused by the diurnal, weekly and seasonal patterns. Furthermore, due to seasonal effects, the need for adaptive models is significant (Rosen and Jeppsson; 2001a). Work is currently carried out to solve these two identification problems.

Another important objection to the approach above is more of a technical nature. In this work, the carbon addition control in the first reactor utilises a substrate-measuring device. These devices are not yet available, at least for direct measurement. However, indirect measurements through respirometry (Spanjers et al.; 1994; Vanrolleghem et al.; 1994) are available, though with a response delay. A way around this problem could be to use another approach for carbon addition control (Vanrolleghem et al.; 1993; Lindberg and Carlsson; 1996; Yuan et al.; 1996) or to estimate the substrate concentration based on other measurement since it has been shown that the controller is relatively insensitive to measurement disturbances.

In this work, no consideration has been taken to the control costs and how efficient the control handles are used. In a real application, a cost-benefit analysis must be carried out in conjunction with the controller design to avoid excessive control costs. A possible way to achieve this would be to include the cost according to a cost function as a controlled variable in the controller.

Conclusions

An approach to supervisory control of wastewater treatment operation by means of principal component space (PCS) control is presented. The supervisory controller models the steady state relationship between manipulated variables and

the process output using principal component analysis. The controller is implemented on top of the local control systems of the manipulated variables. The main objective of the supervisory controller is to control the average effluent quality to certain setpoints. However, a secondary objective of the controller is to minimise the effluent quality variation during varying influent wastewater quality conditions. The controller is based on an earlier presented method, but a compensation term has been introduced to compensate for model errors due to identification problems or changing conditions (disturbances). The compensation term also makes it possible to deal with loss of control handles and local controller saturation.

The supervisory controller is demonstrated using the COST 624 wastewater treatment benchmark simulation model. The results show that the controller is able to meet setpoints imposed on the effluent nitrogen concentration, both under constant and varying influent concentrations. Moreover, the variation in the effluent concentration is reduced significantly by the introduction of a feed-forward term in the controller. It is also shown that the controller can compensate for controller saturation or actuator loss if the loss or saturation occurs in a PC direction covered by other actuators and that it is relatively insensitive to measurement disturbances.

The result of this study indicates that the PCS controller can successfully be used in systems like wastewater treatment systems. However, further analysis of the controller is needed to validate the general applicability of the controller.

Paper G

Addendum

Choice of pseudo inverse

It is stated in the paper that the pseudo inverse (the term pseudo inverse is somewhat carelessly used synonymously to the Moore-Penrose-pseudo inverse) is a sensible choice when solving Equation G.8:

$$\mathbf{t}_{sp}\mathbf{P}_Y \quad = \quad \mathbf{y}_{sp}$$

However, it is not the only choice and a few words on a more general choice of pseudo inverse are appropriate.

Any matrix $\mathbf{P}_Y^{\dagger}$, which fulfils $\mathbf{P}_Y^{\dagger}\mathbf{P}_Y\mathbf{P}_Y^{\dagger} = \mathbf{P}_Y^{\dagger}$ and $\mathbf{P}_Y\mathbf{P}_Y^{\dagger}\mathbf{P}_Y = \mathbf{P}_Y$, is a pseudo inverse to $\mathbf{P}_Y$. Using the Moore-Penrose-pseudo inverse corresponds to minimising

$$J \quad = \quad \mathbf{t}\mathbf{t}^T$$

while $\mathbf{t}_{sp}\mathbf{P}_Y = \mathbf{y}_{sp}$ is fulfilled. However, from a control perspective, it may be interesting to find the solution that minimises a cost function expressed in the manipulated variables so that actual control costs can be considered. Thus, minimising

$$J \quad = \quad f\left(\mathbf{z}_{m,sp}\right)$$

subject to $\mathbf{z}_{m,sp}$ would yield a different solution to the one of the Moore-Penrose-pseudo inverse. The actual design of such a cost function will not be discussed, but some important limitations need to be mentioned. Remember

that an assumption of the PCS controller is that data are mean centred. This implies that the pseudo inverse will be independent of the value of $\mathbf{y}_{sp}$. Now, assume the opposite and that a $\mathbf{y}_{sp}$ close to zero is chosen. Then, the elements of the pseudo inverse will be large, and controller stability can no longer be assumed. Thus, if the data of the general cost function above are not mean centred, one must proceed with caution. This does not restrict us to use a cost function based on $\mathbf{z}_m$. A basic cost function that will yield the same pseudo inverse for all $\mathbf{y}_{sp}$ is:

$$\begin{aligned} J &= \mathbf{z}_m \mathbf{Q} \mathbf{z}_m^T \\ &= \mathbf{t}_{sp} \mathbf{P}_m \mathbf{Q} \left(\mathbf{t}_{sp} \mathbf{P}_m \right)^T \end{aligned}$$

where $\mathbf{Q}$ is a diagonal weighting matrix, allowing for separate weighting of the individual control signals. A limitation with this cost function is that negative control signals will be as heavily penalised as positive.

When the score vector $\mathbf{t}$ has been obtained we need the pseudo inverse for the compensation term in Equations G.11-G.13. This can be obtained as (assumed that $\mathbf{y}_{sp} \neq 0$)

$$\mathbf{P}_Y^{\dagger} = \mathbf{y}_{sp}^{\dagger} \mathbf{t}_{sp}$$

where $\mathbf{y}_{sp}^{\dagger}$ is the Moore-Penrose-pseudo inverse of $\mathbf{y}_{sp}$. In the single output case (as in Paper G) the pseudo inverse is $\mathbf{t}_{sp}/y_{sp}$.

In Figure G.11, the result of a different choice of pseudo inverse is shown for constant influent conditions and a setpoint for effluent nitrogen concentration of 15 mg N/l from days 1 to 3 and 12 mg N/l from days 3 to 7. Separate weights were put on each control setpoint, using the $\mathbf{Q}$-matrix. What can be seen is a decrease in the substrate concentration (high penalty) of 5% and an increase in internal recirculation (low penalty) of 20-40% compared to Moore-Penrose-pseudo inverse. It is clear that the dynamic or transient behaviour is similar. However, the qualitative behaviour of the dissolved oxygen levels in reactors 2, 3 and 5 is different from the Moore-Penrose solution and this further stresses the fact that the choice of pseudo inverse is not obvious. More distinct differences can be obtained by more sophisticated cost functions and no exhaustive analysis has been carried out. However, the example shows that there is a potential improvement of the PCS controller by investigating the choice of pseudo inverse.

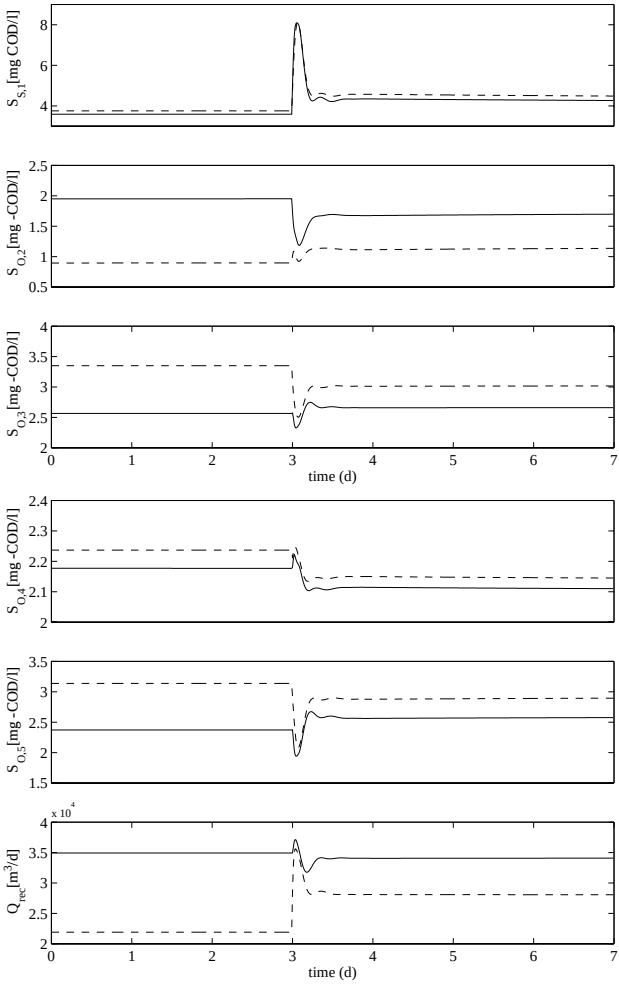

Figure G.11: Difference in the local controller setpoints generated by the PCS controller using different choices of pseudo inverses. (—) pseudo inverse obtained by optimising over $\mathbf{z}_m$, (--) Moore-Penrose-pseudo inverse. Constant influent conditions and effluent nitrogen setpoints of 15 mg N/l (days 1-3) and 12 mg N/l (days 3-7).

Sensitivity to identification

The identification part must be considered as the bottleneck of the methodology. As was mentioned in the paper, it is not realistic to assume that a model can be identified during constant influent conditions. However, the compensation term makes the identification less sensitive as the term corrects minor discrepancies between the identified model and the real process. What is important, is that the model is qualitatively correct, that is the model describes the major directions of the process in a correct manner. This can be shown by manipulating the loading matrix $\mathbf{P}$ (here denoted $\mathbf{P}_{orig}$). Two manipulations are carried out. In the first, a disturbance matrix $\delta\mathbf{P}$ is added to $\mathbf{P}_{orig}$ to form $\mathbf{P}_{dist}$. The disturbance matrix is a random, normally distributed matrix and the matrix element values are chosen so that the norm of the matrix is 0.25 (the norm of $\mathbf{P}_{orig}$ is 1). The second manipulation consists of using an 'indicator' matrix, $\mathbf{P}_{ind}$, representing $\mathbf{P}_{orig}$. To let the elements be 1 or 0 is a too crude approximation. Instead, a four level indicator matrix is used. The indicator matrix is simply calculated by multiplying $\mathbf{P}_{orig}$ with 3, rounding each element to the closest integer and then dividing the matrix with 3. The test matrices, together with the original matrix, are shown below to demonstrate how relatively severe the manipulations are.

$$\mathbf{P}_{orig} = \begin{bmatrix} -0.3461 & 0.4648 & -0.0166 & 0.1207 & -0.0154 \\ -0.2474 & -0.3500 & -0.1320 & 0.0412 & -0.0146 \\ 0.0212 & -0.0898 & -0.3643 & 0.8851 & 0.2143 \\ 0.0345 & 0.0543 & -0.6230 & -0.4200 & 0.6474 \\ -0.1007 & 0.1434 & -0.6350 & -0.1103 & -0.7075 \\ -0.4452 & -0.3751 & 0.1460 & -0.0633 & 0.0541 \\ 0.2280 & -0.5793 & -0.0665 & -0.0799 & -0.1466 \\ 0.5010 & 0.3339 & 0.1396 & 0.0312 & 0.0088 \\ 0.5532 & -0.2051 & -0.1137 & 0.0198 & -0.0976 \end{bmatrix}$$

$$\mathbf{P}_{dist} = \begin{bmatrix} -0.4758 & 0.4634 & 0.0052 & 0.1175 & 0.0700 \\ -0.3594 & -0.3315 & -0.1769 & 0.0665 & -0.0478 \\ -0.0170 & -0.0192 & -0.3743 & 0.8631 & 0.1397 \\ 0.0221 & 0.0957 & -0.7863 & -0.4534 & 0.7012 \\ -0.1001 & 0.0266 & -0.6035 & -0.1127 & -0.7047 \\ -0.3894 & -0.3287 & 0.1538 & -0.0749 & 0.0036 \\ 0.1798 & -0.5252 & -0.1059 & -0.1437 & -0.1525 \\ 0.4529 & 0.3763 & 0.0960 & 0.1174 & -0.1251 \\ 0.5398 & -0.1178 & -0.1857 & 0.0492 & -0.0253 \end{bmatrix}$$

$$\mathbf{P}_{ind} = \begin{bmatrix} -0.3333 & 0.3333 & 0 & 0 & 0 \\ -0.3333 & -0.3333 & 0 & 0 & 0 \\ 0 & 0 & -0.3333 & 1.0000 & 0.3333 \\ 0 & 0 & -0.6667 & -0.3333 & 0.6667 \\ 0 & 0 & -0.6667 & 0 & -0.6667 \\ -0.3333 & -0.3333 & 0 & 0 & 0 \\ 0.3333 & -0.6667 & 0 & 0 & 0 \\ 0.6667 & 0.3333 & 0 & 0 & 0 \\ 0.6667 & -0.3333 & 0 & 0 & 0 \end{bmatrix}$$

The results of a simulation run with numerous setpoint changes are displayed in Figure G.12 together with results using the original $\mathbf{P}_{orig}$. It is evident that the $\mathbf{P}_{orig}$ is relatively insensitive to the disturbances. It can be seen that it takes longer time for the controller to reach new setpoints and the variation is larger in the cases where $\mathbf{P}_{dist}$ and $\mathbf{P}_{ind}$ are used, but it is clear that the controller still delivers acceptable results. A few conclusions can be drawn from this: 1) The compensation term makes the controller less sensitive to identification errors. 2) It may be possible to infer the controller model from simulations and process knowledge in the form of an indicator matrix. 3) The problem of non-stationary data becomes simplified since no new correlation structure needs to be identified. Instead, only adaptive scaling parameters should suffice. It should also be noted that variations in the influent wastewater characteristics files for the COST 624 benchmark is relatively high. If the variations are less dominant, it may very well be possible to separate the effects of the changes in the controller setpoints from those of the varying influent conditions.

Different operational states

From Figures G.8 and G.9 it is evident that the controller does not handle the influent peaks especially well. A possible explanation is related to the region in which the model is valid. This region is determined by the data used for identification and how well these can be approximated by a linear model. Consequently, there will be regions not covered by the model. The peak loads falls into such a region. During low and medium load periods, there is an excess of nitrification capacity. In this region, the model yields results that lower the DO setpoints in some reactors to obtain a decrease in the effluent nitrogen concentration. However, during high loads this is not appropriate. Instead, the

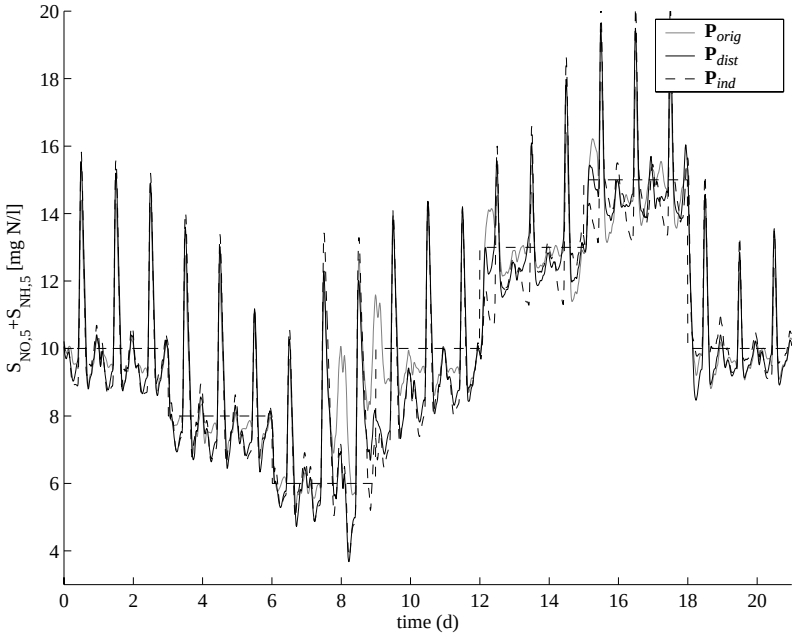

Figure G.12: The effluent nitrogen from the plant during various effluent setpoints.

correct action would be to increase the DO setpoints. There is, consequently, a nonlinear relation between the effluent nitrogen concentration and the DO concentrations in the reactors. Since the model is identified during constant influent conditions, corresponding to the mean values of the dry weather data (excess of nitrification capacity), the model does not reflect this behaviour. This conclusion is supported by results from other simulation results on the same simulation model (Vanrolleghem and Gillot; 2001). The fact that the model lowers the DO setpoints in some reactors to obtain a decrease effluent nitrogen concentration, probably puts a limit to how low the effluent nitrogen concentration can be pushed, since extremely low effluent nitrogen concentration also falls outside the region, in which the model is valid. Thus, the experimental design to produce the identification data is crucial to the performance of the controller.

By implementing the PCS controller in combination with the supervisory control framework discussed in Papers E and F, this problem can be resolved. Also, by introducing a second model that better describes the operational state of insufficient nitrification capacity could provide another solution. Adaptive models may also be applicable, but the task of making the model adaptive is not as straightforward as in the monitoring case.

More control handles

In the paper, only six control handles are used. A logical extension of the supervisory control system would be to include sludge return flow rate and wastage flow rate. The sludge return flow rate are then used in shorter time scales as been discussed in Paper E and in Yuan et al. (2001a), and the wastage flow rate then works in the longer time scales to control the total amount of sludge in the system. Another possible control handle is step feed where the influent stream can be directed to more than one tank. Preferably, this should be done in variable fashion, so that a varying percentage of the flow can be directed to each tank.

Part III

Bibliography

Bibliography

Aarnio, P. and Minkkinen, P. (1986). Application of partial least-squares modelling in the optimization of a waste-water treatment plant, *Anal. Chim. Acta* **191**: 457–460.

Alsberg, B. K. (1999). Multiscale cluster analysis, *Anal. Chem.* **71**: 3092–3100.

Alsberg, B. K., Woodward, A. M. and Kell, D. B. (1997). An introduction to wavelet transforms for chemometricians: a time-frequency approach, *Chemometrics Intell. Lab. Syst.* **37**: 215–239.

Åström, K. J. and Wittenmark, B. (1989). *Adaptive control*, Addison-Wesley, Reading, Massachusetts, USA.

Åström, K. J. and Wittenmark, B. (1997). *Computer controlled systems, theory and design*, 3rd edn, Prentice Hall, Inc., Englewood Cliffs, New Jersey, USA.

Baffi, G., Martin, E. B. and Morris, A. J. (1999). Non-linear projection to latent structures revisited: the quadratic PLS algorithm, *Comput. Chem. Eng.* **23**: 395–411.

Baffi, G., Martin, E. B. and Morris, A. J. (2000). Non-linear dynamic projection to latent structures modelling, *Chemometrics Intell. Lab. Syst.* **52**: 5–22.

Bakshi, B. R. (1998). Multiscale PCA with application to multivariate statistical process monitoring, *AIChE J.* **44**(7): 1596–1610.

Bakshi, B. R. (1999). Multiscale analysis and modeling using wavelets, *J. Chemometr.* **13**: 415–434.

Bendwell, N. (2000). Monitoring of a wastewater treatment plant with a multivariate model, *Control Systems 2000*, Victoria, BC, Canada, pp. 185–187.

Bergh, S.-G. (1996). *Diagnosis problems in wastewater settling,* Lic. thesis, Dept. of Industrial Electrical Engineering and Automation, Lund University, Lund, Sweden.

Berglund, A. and Wold, S. (1997). INLR, implicit non-linear latent variable regression, *J. Chemometr.* **11**: 141–156.

Bissel, D. (1994). *Statistical methods for SPC and TQM*, Chapman and Hall, London, UK.

Box, G. and Luceno, A. (1997). *Statistical control by monitoring and feedback adjustment*, John Wiley & Sons, Inc., New York, New York, USA.

Bro, R., Smilde, A. K. and de Jong, S. (2001). On the difference between low-rank and subspace approximation: improved model for multi-linear PLS regression, *Chemometrics Intell. Lab. Syst.* **58**: 3–13.

Camacho, E. F. and Bordons, C. (1999). *Model predictive control - advanced textbooks in control and signal processing*, Springer Verlag, Berlin, Germany.

Carstensen, J. (1994). *Identification of wastewater processes*, PhD thesis, Inst. of Mathematical Statistics and Operations Research (IMSOR), Technical University of Denmark, Lyngby, Denmark.

Champely, S. and Doledec, S. (1997). How to separate long-term trends from periodic variation in water quality monitoring, *Wat. Res.* **31**(11): 2849–2857.

Chapman, D. T. (1998). Statistics for treatment plant operation, *in* J. F. Andrews (ed.), *Dynamics and control of the activated sludge, Water Quality Library*, Vol. 6, Technomic Publishing Company, Inc., Lancaster, Pennsylvania, USA.

Chen, G. and McAvoy, T. J. (1996). Process control utilizing data based multivariate statistical models, *Can. J. Chem. Eng.* **74**: 1010–1024.

Chen, G., McAvoy, T. J. and Piovoso, M. J. (1998). A multivariate statistical controller for on-line quality improvement, *J. Process Control* **8**(2): 139–149.

Copp, J. B. (2000). *The COST simulation benchmark - description and simulator manual*, COST (European Cooperation in the field of Scientific and Technical Research), Brussels, Belgium.

COST624 (2001). COST Action 624 website, http://www.ensic.inpl-nancy.fr/COSTWWTP.

Dahl, K. S., Piovoso, M. J. and Kosanovich, K. A. (1999). Translating third-order data analysis methods to chemical batch processes, *Chemometrics Intell. Lab. Syst.* **46**: 161–180.

Davis, J. F., Bakshi, B. R., Kosanovich, K. A. and Piovoso, M. J. (1996). Process monitoring, data analysis and data interpretation, *AIChE Symposium Series* **92**: 1–11.

Dayal, B. S. and MacGregor, J. F. (1997a). Improved PLS algorithms, *J. Chemometr.* **11**: 73–85.

Dayal, B. S. and MacGregor, J. F. (1997b). Recursive exponentially weighted PLS and its application to adaptive control and prediction, *J. Process Control* **7**(3): 169–179.

Di Rusco, D. (1998). The partial least squares algorithm: a truncated Cayley-Hamilton series approximation used to solve the regression problem, *Model. Identif. Control* **19**(3): 117–140.

Dohan, K. and Whitfield, P. H. (1997). Identification and characterization of water quality transient using wavelet analysis. I. Wavelet analysis methodology, *Wat. Sci. Tech.* **36**(5): 325–335.

Dold, P. L., Ekama, G. A. and Marais, G. (1980). A general model for the activated sludge process, *Prog. Wat. Tech.* **12**: 47–77.

Dong, D. and McAvoy, T. J. (1996a). Batch tracking via nonlinear principal component analysis, *AIChE J.* **42**(8): 2199–2208.

Dong, D. and McAvoy, T. J. (1996b). Nonlinear principal component analysis - based on principal curves and neural networks, *Comput. Chem. Eng.* **20**(1): 65–78.

Ekama, G. A. and Marais, G. (1979). Dynamic behaviour of the activated sludge process, *J. Water Pollution Control Fed.* **51**: 534–556.

Ekama, G. A., Barnard, J. L., Günthert, F. W., Krebs, P., McCorquodale, J. A., Parker, D. S. and Wahlberg, E. J. (1997). Secondary settling tanks: theory, modelling, design and operation, *Scientific and Technical Report No. 6*, IAWQ, London, UK.

Eriksson, L., Johansson, E., Kettaneh-Wold, N. and Wold, S. (2001). *Multi- and megavariate data analysis - principles and applications*, Umetrics AB, Umeå, Sweden.

Esbensen, K. and Geladi, P. (1990). The start and early history of chemometrics: selected interviews. Part 2, *J. Chemometr.* **4**: 389–412.

Fisher, R. and MacKenzie, W. (1923). Studies in crop variations. II. The manurial response of different potato varieties, *J. Agr. Sci.* **13**: 311–329.

Flehmig, F., v Watzdorf, R. and Marquardt, W. (1998). Identification of trends on process measurement using the wavelet transform, *Comput. Chem. Eng.* **22**(Suppl.): S491–S496.

Fourie, S. H. and de Vaal, P. (2000). Advanced process monitoring using an online non-linear multiscale principal component analysis methodology, *Comput. Chem. Eng.* **24**: 755–760.

Gallagher, N. B. and Wise, B. M. (1997). Development and benchmarking of multivariate statistical process control tools for a semiconductor etch process: improving robustness through model updating, *ADCHEM 97*, Banff, Canada, pp. 78–83.

Garcia, C. E., Prett, D. M. and Morari, M. (1989). Model predictive control: theory and practice - a survey, *Automatica* **25**(3): 335–348.

Geladi, P. (1988). Notes on history and nature of partial least squares (PLS) modelling, *J. Chemometr.* **2**: 231–246.

Geladi, P. and Esbensen, K. (1990). The start and early history of chemometrics: selected interviews. Part 1, *J. Chemometr.* **4**: 337–354.

Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial, *Anal. Chim. Acta* **185**: 1–17.

Glad, T. and Ljung, L. (2000). *Control theory - multivariable and nonlinear methods*, Taylor & Francis, London, UK.

Grung, B. and Manne, R. (1998). Missing values in principal component analysis, *Chemometrics Intell. Lab. Syst.* **42**: 125–139.

Gujer, W., Henze, M., Mino, T. and van Loosdrecht, M. C. M. (1999). Activated sludge model no. 3, *Wat. Sci. Tech.* **39**(1): 183–193.

Hammer, M. J. (1986). *Water and wastewater technology*, Prentice Hall, Inc., Englewood Cliffs, New Jersey, USA.

Helland, K., Berntsen, H. E., Borgen, O. S. and Martens, H. (1991). Recursive algorithm for partial least squares regression, *Chemometrics Intell. Lab. Syst.* **14**: 129–137.

Henze, M., Grady Jr, C. P. L., Gujer, W., Marais, G. and Matsuo, T. (1987). Activated sludge model no. 1, *Scientific and Technical Report No. 1*, IAWQ, London, UK.

Henze, M., Gujer, W., Mino, T. and van Loosdrecht, M. C. M. (2000). Activated sludge model ASM1, ASM2, ASM2d and ASM3, *Scientific and Technical Report No. 9*, IWA Publishing, London, UK.

Henze, M., Gujer, W., Mino, T., Matsuo, T., Wentzel, M. C. and Marais, G. (1995). Activated sludge model no. 2, *Scientific and Technical Report No. 3*, IAWQ, London, UK.

Henze, M., Gujer, W., Mino, T., Matsuo, T., Wentzel, M. C., Marais, G. and van Loosdrecht, M. C. M. (1999). Activated sludge model no. 2d, *Wat. Sci. Tech.* **39**(1): 165–182.

Himes, D. M., Storer, R. H. and Georgakis, C. (1994). Determination of the number of principal components for disturbance detection and isolation, *Amer. Control Conf. 1994*, Baltimore, USA, pp. 1279–1283.

Höskuldsson, A. (1988). PLS regression methods, *J. Chemometr.* **2**: 211–228.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *J. Edu. Psych.* **24**: 417–441.

Hwang, D. H. and Han, C. (1999). Real-time monitoring for a process with multiple operating modes, *Control Eng. Practice* **7**: 891–902.

ICA (2001). Conference preprints, *1st IWA Conference on Instrumentation, Control and Automation (ICA2001),* Malmö, Sweden, Vol. 1-2, IEA, Lund Univ., Lund, Sweden.

Jackson, J. E. (1980). Principal components and factor analysis: part I - principal components, *J. Qual. Tech.* **12**(4): 201–213.

Jackson, J. E. (1981). Principal components and factor analysis: part II - additional topics related to principal components, *J. Qual. Tech.* **13**(1): 46–58.

Jackson, J. E. and Mudholkar, G. S. (1979). Control procedures for residual associated with principal component analysis, *Technometrics* **21**(3): 341–349.

Jaeckle, C. and MacGregor, J. F. (1996). Product design through multivariate statistical analysis of process data, *Comput. Chem. Eng.* **20**(Suppl.): S1047–S1052.

Jaeckle, C. and MacGregor, J. F. (2000). Industrial application of product design through the inversion of latent variable models, *Chemometrics Intell. Lab. Syst.* **50**: 199–210.

Jeppsson, U. (1996). *Modelling aspects of wastewater treatment processes*, PhD thesis, Dept. of Industrial Electrical Engineering and Automation, Lund University, Lund, Sweden.

Jeppsson, U., Alex, J., Pons, M. N., Spanjers, H. and Vanrolleghem, P. A. (2001). Status and future trends of ICA in wastewater treatment - a European perspective, *Wat. Sci. Tech.* (accepted).

Jia, F., Martin, E. B. and Morris, A. J. (1998). Non-linear principal component analysis for process fault detection, *Comput. Chem. Eng.* **22**(Suppl.): S851–S854.

Joliffe, J. (1986). *Principal component analysis*, Springer Verlag, Berlin, Germany.

Kano, M., Hasebe, S., Hashimoto, I. and Ohno, H. (2001). A new multivariate statistical process monitoring method using principal component anlaysis, *Comput. Chem. Eng.* **25**: 1103–1113.

Kano, M., Nagao, K., Hasebe, S., Hashimoto, I., Ohno, H., Strauss, R. and Bakshi, B. R. (2000a). Comparison of statistical process monitoring methods: application to the Eastman challenge problem, *Comput. Chem. Eng.* **24**: 175–181.

Kano, M., Nagao, K., Ohno, H., Hasebe, S. and Hashimoto, I. (2000b). Dissimilarity of process data for statistical process monitoring, *ADCHEM 00*, Piza, Italy, pp. 231–236.

Kaspar, M. H. and Ray, W. H. (1992). Chemometric methods for process monitoring and high-performance controller design, *AIChE J.* **38**(10): 1593–1608.

Kaspar, M. H. and Ray, W. H. (1993). Dynamic PLS modelling for process control, *Chem. Eng. Sci.* **48**(20): 3447–3461.

Katebi, M. R., Johnson, M. A., Wilkie, J. and McCluskey, G. (1998). Control and management of wastewater treatment plants, *UKACC International Conference on Control '98* pp. 433–438.

Kennel, M. B. (1997). Statistical test for dynamical nonstationarity in observed time-series data, *Phys. Rev. E* **56**(1): 316–321.

Kim Oanh, N. T. and Bengtsson, B. E. (1995). Development of a wastewater monitoring program incorporated into process control for mitigation of chemical and fiber loss from the Bai Bang Paper Company (BAPACO), a bleached kraft pulp and paper mill in Vietnam, *Resour. Conserv. Recycl.* **14**: 53–66.

Kosanovich, K. A. and Piovoso, M. J. (1997). PCA of wavelet transformed process data for monitoring, *Intell. Data Anal.* **1**(2): http://www–east.elsevier.com/ida/.

Kourti, T. and MacGregor, J. F. (1994). Multivariate SPC methods for monitoring and diagnosing of process performance, *PSE'94*, Kyongju, Korea, pp. 739–746.

Kourti, T. and MacGregor, J. F. (1995). Tutorial: Process analysis, monitoring and diagnosis, using multivariate projection methods, *Chemometrics Intell. Lab. Syst.* **28**: 3–21.

Kourti, T., Lee, J. and MacGregor, J. F. (1996). Experiences with industrial applications of projection methods for multivariate statistical process control, *Comput. Chem. Eng.* **20**: 745–750.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks, *AIChE J.* **37**(2): 233–243.

Kramer, M. A. and Mah, R. S. H. (1994). Model-based monitoring, *in* D. Rippin, J. Hale and J. Davis (eds), *Second Int. Conf. on Foundations of computer aided process operations*, pp. 45–68.

Kresta, J. V., MacGregor, J. F. and Marlin, T. E. (1991). Multivariate statistical monitoring of process operating performance, *Can. J. Chem. Eng.* **69**: 35–47.

Krofta, M., Herath, B., Burgess, D. and Lampman, L. (1995). An attempt to understand dissolved air flotation using multivariate data analysis, *Wat. Sci. Tech.* **31**(3-4): 191–201.

Ku, W., Storer, R. H. and Georgakis, C. (1995). Disturance detection and isolation by dynamic pricipal component analysis, *Chemometrics Intell. Lab. Syst.* **30**: 179–196.

Kynch, G. J. (1952). A theory of sedimentation, *Trans. Faraday Soc.* **48**: 166–176.

Lang, M., Guo, H., Odegard, J. E., Burrus, C. S. and Wells Jr, R. O. (1996). Noise reduction using an undecimated discrete wavelet transform, *IEEE Signal Process. Lett.* **3**(1): 10–12.

Larsson, J.-E. (1994). Diagnosis based on explicit mean-end models, *Artif. Intell.* **80**: 29–93.

Larsson, M., Hill, D. J. and Olsson, G. (2000). Emergency voltage control using search and predictive control, *Int. J. Electr. Power Energy Syst.* (accepted).

Lennox, J. A. (2001). *Multivariate subspaces for fault detection and isolation: with applications to the wastewater treatment process*, PhD thesis, Dept. of Chemical Engineering, The University of Queensland, Brisbane, Australia.

Lennox, J. A. and Rosen, C. (2001). Adaptive multiscale principal component analysis for online monitoring of wastewater treatment, *Wat. Sci. Tech.* (accepted).

Li, W. and Qin, S. J. (2001). Consistent dynamic PCA based on errors-in-variables subspace identification, *J. Process Control* **11**: 661–678.

Li, W., Yue, H. H., Valle-Cervantes, S. and Qin, S. J. (2000). Recursive PCA for adaptive process monitoring, *J. Process Control* **10**: 471–486.

Lin, W., Qian, Y. and Li, X. (2000). Non-linear dynamic principal component analysis for the on-line process monitoring and diagnosis, *Comput. Chem. Eng.* **24**: 423–429.

Lindberg, C. F. (1997). *Control and estimation strategies applied to the activated sludge process*, PhD thesis, System and Control Group, Uppsala University, Uppsala, Sweden.

Lindberg, C. F. and Carlsson, B. (1996). Adaptive control of external carbon flow rate in an activated sludge process, *Wat. Sci. Tech.* **34**(3-4): 173–180.

Lindgren, F., Geladi, P. and Wold, S. (1993). The kernel algorithm for PLS, *J. Chemometr.* **7**: 45–59.

Louwerse, D. J. and Smilde, A. K. (2000). Multivariate statistical process control of batch processes based on three-way models, *Chem. Eng. Sci.* **55**: 1225–1235.

Luo, R., Misra, M. and Himmelblau, D. M. (1999). Sensor fault detection via multiscale analysis and dynamic PCA, *Ind. Eng. Chem. Res.* **38**: 1489–1495.

MacGregor, J. F. (1997). Using on-line process data to improve quality: challenges for statisticians, *Int. Stat. Rev.* **65**(3): 309–323.

MacGregor, J. F. and Kourti, T. (1995). Statistical process control of multivariate processes, *Control Eng. Practice* **3**(3): 403–414.

MacGregor, J. F., Jaeckle, C., Kiparissides, C. and Koutoudi, M. (1994). Process monitoring and diagnosis by multiblock PLS methods, *AIChE J.* **40**(5): 826–838.

Malthouse, E. C., Mah, R. S. H. and Tamhane, A. C. (1995). Some theoretical results on nonlinear principal component analysis, *Amer. Control Conf. 1995*, Seattle, USA, pp. 744–748.

Malthouse, E. C., Tamhane, A. C. and Mah, R. S. H. (1997). Nonlinear partial least squares, *Comput. Chem. Eng.* **21**(8): 875–890.

Marsili-Libelli, S. and Müller, A. (1996). Adaptive fuzzy pattern recognition in the anaerobic digestion process, *Pattern Recognit. Lett.* **17**: 651–659.

Mayne, D. Q., Rawlings, J. B., Rao, C. V. and Scokaert, P. O. M. (2000). Constrained model predictive control: stability and optimality, *Automatica* **36**: 789–814.

Messick, N. J., Kalivas, J. H. and Lang, P. M. (1997). Selecting factors for partial least squares, *Microchem. J.* **55**: 200–207.

Misiti, M., Misiti, Y., Oppenheim, G. and Poggi, J. M. (1996). *Wavelet TOOLBOX, For use with MATLAB*, The MathWorks, Inc., Natick, Massachusetts, USA.

Morari, M. and Lee, J. H. (1999). Model predictive control: past, present and future, *Comput. Chem. Eng.* **23**: 667–682.

Mujunen, S.-P., Minkkinen, P., Teppola, P. and Wirkkala, R.-S. (1998). Modeling of activated sludge plants treatment efficiency with PLSR: a process analytical case study, *Chemometrics Intell. Lab. Syst.* **41**: 83–94.

Negiz, A. and Çinar, A. (1997). Statistical monitoring of multivariable dynamic processes with state-space models, *AIChE J.* **43**(8): 2002–2020.

Nelson, P. R. C., Taylor, P. A. and MacGregor, J. F. (1996). Missing data methods in PCA and PLS: score calculations with incomplete observations, *Chemometrics Intell. Lab. Syst.* **35**: 45–65.

Nounou, M. N. and Bakshi, B. R. (1999). On-line multiscale filtering of random and gross errors without process models, *AIChE J.* **40**(5): 1041–1058.

Olsson, G. (1989). Practical experiences of identification and modeling from experiments, *in* G. G. Patry and D. Chapman (eds), *Dynamic modeling and expert systems in wastewater engineering*, Lewis Publishers, Inc., Chelsea, Michigan, USA.

Olsson, G. and Jeppsson, U. (1994). Establishing cause-effect relationships in acivated sludge plants - what can be controlled, *8^{th} Forum Applied Biotechnology (FAB)*, Gent, Belgium, pp. 2057–2070.

Olsson, G. and Newell, B. (1999). *Wastewater treatment systems – modelling, diagnosis and control*, IWA Publishing, London, UK.

Olsson, G. and Piani, G. (1992). *Computer systems for automation and control*, Prentice Hall International (UK) Ltd, Hemel Hampstead, UK.

Orhon, D. and Artan, N. (1994). *Modelling of activated sludge systems*, Technomic Publishing Company, Inc., Lancaster, Pennsylvania, USA.

Ouyang, S., Bao, Z. and Liao, G.-S. (2000). Robust recursive least squares learning algorithm for principal component analysis, *IEEE Trans. Neural Netw.* **10**(1): 215–221.

Pasadyn, A. J., Qin, S. J. and Valle-Cervantes, S. (1999). Closed-loop and open-loop identification of an industrial wastewater reactor, *Amer. Control Conf. 1999*, San Diego, USA, pp. 3965–3969.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *Philos. Mag.* **2**: 559–572.

Percival, D. B. and Mofjeld, H. O. (1997). Analysis of subtidal coastal level fluctuations using wavelets, *J. Am. Stat. Assoc.* **92**(439): 868–880.

Phatak, A. and de Jong, S. (1997). The geometry of partial least squares, *J. Chemometr.* **11**: 311–338.

Piovoso, M. J. and Kosanovich, K. A. (1994). Applications of multivariate statistical methods to process monitoring and controller design, *Int. J. Control* **59**(3): 743–765.

Piovoso, M. J., Kosanovich, K. A. and Pearson, R. K. (1992). Monitoring process performance in real-time, *Amer. Control Conf. 1992*, Chicago, USA, pp. 2359–2363.

Pons, M. N., Spanjers, H. and Jeppsson, U. (1999). Towards a benchmark for evaluating control strategies in wastewater treatment plants by simulation , *Comput. Chem. Eng.* **23**(Suppl.): S403–S406.

Qin, S. J. (1998). Recursive PLS algorithm for adaptive data modeling, *Comput. Chem. Eng.* **22**(4-5): 503–514.

Qin, S. J. and Dunia, R. (2000). Determining the number of principal components for best reconstruction, *J. Process Control* **10**(2-3): 245–250.

Qin, S. J. and McAvoy, T. J. (1992). Nonlinear PLS modeling using neural networks, *Comput. Chem. Eng.* **16**(4): 379–391.

Rännar, S., MacGregor, J. F. and Wold, S. (1998). Adaptive batch monitoring using hierarchical PCA, *Chemometrics Intell. Lab. Syst.* **41**: 73–81.

Ricker, N. L. (1988). The use of biased least squares estimators for parameters in discrete-time pulse-response models, *Ind. Eng. Chem. Res.* **27**: 343–350.

Rioul, O. (1993). A discrete-time multiresolution theory, *IEEE Trans. Signal Process.* **41**(8): 2591–2606.

Roda, I. R., Sànchez-Marrè, M., Comas, J., Baeza, J., Colprim, J., Lafuente, J., Cortés, U. and Poch, M. (2001). A hybrid supervisory system to support wastewater treatment plant operation, *Wat. Sci. Tech.* (accepted).

Rosen, C. (1998a). *Monitoring wastewater treatment systems,* Lic. thesis, Dept. of Industrial Electrical Engineering and Automation, Lund University, Lund, Sweden.

Rosen, C. (1998b). Time delays and fault propagation in multilevel flow models, *Technical Report TEIE-7132*, Dept. of Industrial Electrical Engineering and Automation, Lund University, Lund, Sweden.

Rosen, C. and Jeppsson, U. (2001a). A chemometric approach to supervisory control of wastewater treatment operation, *J. Chemometr.* (submitted).

Rosen, C. and Jeppsson, U. (2001b). Supervisory control of wastewater treatment operation by PC-space control, *7th Scandinavian Symposium on Chemometrics (SSC7)* p. A77.

Rosen, C. and Lennox, J. A. (2001). Multivariate and multiscale monitoring of wastewater treatment operation, *Wat. Res.* **35**(14): 3402–3410.

Rosen, C. and Olsson, G. (1997a). Analysis of on-line measurements of Pt Loma wastewater treatment plant, San Diego, *Technical Report TEIE-7131*, Dept. of Industrial Electrical Engineering and Automation, Lund University, Lund, Sweden.

Rosen, C. and Olsson, G. (1997b). From data to information - analysis of operational data from wastewater treatment plants, *Ny teknik inom avloppsvattenrening 1997*, Lund, Sweden.

Rosen, C. and Olsson, G. (1998). Disturbance detection in wastewater treatment plants, *Wat. Sci. Tech.* **37**(12): 197–205.

Rosen, C. and Yuan, Z. (2000). Supervisory control of wastewater treatment plants by combining principal component analysis and fuzzy c-means clustering, *Wat. Sci. Tech.* **43**(7): 147–156.

Rosen, C., Larsson, M., Jeppsson, U. and Yuan, Z. (2001). A framework for extreme-event control in wastewater treatment, *Wat. Sci. Tech.* (accepted).

Röttorp, J. and Jansson, Å. (2001). Discussion on multivariate real-time systems for process monitoring and control, *J. Chemometr.* (submitted).

Sànchez, M., Cortés, U., Bejar, J., De Gracia, J., Lafuente, J. and Poch, M. (1997). Concept formation in WWTP by means of classification techniques: a compared study, *Appl. Intell.* **7**: 147–165.

Sànchez, M., Cortés, U., Lafuente, J., Roda, I. R. and Poch, M. (1996). DAIDEPUR: an integrated and distributed architecture for wastewater treatment supervision, *Artif. Intell. Eng.* **10**(3): 275–285.

Schreiber, T. (1997). Detecting and analysing nonstationarity in a time series using nonlinear cross predictions, *Phys. Rev. Lett.* **78**(5): 843–846.

Shao, R., Jia, F., Martin, E. B. and Morris, A. J. (1999). Wavelets and non-linear principal component analysis for process monitoring, *Control Eng. Practice* **7**: 865–879.

Shensa, M. J. (1992). The discrete wavelet transform: wedding the à trous and mallat algorithms, *IEEE Trans. Signal Process.* **40**(10): 2464–2482.

Spanjers, H., Olsson, G. and Klapwijk, A. (1994). Determining short-term biochemical oxygen-demand and respiration rate in an aeration tank by using respirometry and estimation, *Wat. Res.* **28**(7): 1571–1583.

Spanjers, H., Vanrolleghem, P. A., Nguyen, K., Vanhooren, H. and Patry, G. G. (1998a). Towards a simulation-benchmark for evaluating respirometry-based control strategies, *Wat. Sci. Tech.* **37**(12): 219–226.

Spanjers, H., Vanrolleghem, P. A., Olsson, G. and Dold, P. L. (1998b). Respirometry in control of the of the activated sludge process: principles, *Scientific and Technical Report No. 7*, IAWQ, London, UK.

Stephanopoulos, G. and Ng, C. (2000). Perspectives on the synthesis of plant-wide control structures, *J. Process Control* **10**(2-3): 97–111.

Stephanopoulos, G., Dyer, M. and Karsligil, O. (1997). Multi-scale modeling, estimation and control of process systems, *Comput. Chem. Eng.* **21**(Suppl.): S7979–S803.

Stork, C. L. and Kowalski, B. R. (1999). Weighting schemes for updating regression models—a theoretical approach, *Chemometrics Intell. Lab. Syst.* **48**: 151–166.

Strang, G. and Nguyen, T. (1996). *Wavelets and filter banks*, Wellesley-Cambridge Press, Wellesley, Massachusetts, USA.

Takács, I., Patry, G. G. and Nolasco, D. (1991). A dynamic model of the clarification-thickening process, *Wat. Res.* **25**(10): 1263–1271.

Teppola, P. and Minkkinen, P. (1999). Possibilistic and fuzzy c-means clustering for process monitoring in an activated sludge waste-water treatment plant, *J. Chemometr.* **13**: 445–459.

Teppola, P., Mujunen, S.-P. and Minkkinen, P. (1997). Partial least squares modeling of an activated sludge plant: a case study, *Chemometrics Intell. Lab. Syst.* **38**: 197–208.

Teppola, P., Mujunen, S.-P. and Minkkinen, P. (1998a). A combined approach of partial least squares and fuzzy c-means clustering for the monitoring of an activated-sludge waste-water treatment plant, *Chemometrics Intell. Lab. Syst.* **41**: 95–103.

Teppola, P., Mujunen, S.-P. and Minkkinen, P. (1998b). Kalman filter for updating the coefficients of regression models. A case study from an activated sludge waste-water treatment plant, *Chemometrics Intell. Lab. Syst.* **45**: 371–384.

Teppola, P., Mujunen, S.-P. and Minkkinen, P. (1999). Adaptive fuzzy c-means clustering in process monitoring, *Chemometrics Intell. Lab. Syst.* **45**: 23–38.

Teppola, P., Mujunen, S.-P., Minkkinen, P., Puijola, T. and Pursiheimo, P. (1998c). Principal component analysis, contribution plots and feature weights in the monitoring of sequential process data from a paper machine wet end, *Chemometrics Intell. Lab. Syst.* **44**: 307–317.

Thompson, J. R. and Koronacki, J. (1993). *Statistical process control for quality improvement*, Chapman and Hall, New York, New York, USA.

Torrence, C. and Compo, G. P. (1998). A practical guide to wavelet analysis, *Bull. Amer. Meteorol. Soc.* **79**(1): 61–78.

Trygg, J. and Wold, S. (1998). PLS regression on wavelet compressed NIR spectra, *Chemometrics Intell. Lab. Syst.* **42**: 209–220.

Trygg, J., Kettaneh-Wold, N. and Wallbäcks, L. (2001). 2D wavelet analysis and compression of on-line industrial process data, *J. Chemometr.* **15**(4): 299–319.

Tsung, F. (2000). Statistical monitoring and diagnosis of automatic controlled process using dynamic PCA, *Int. J. Prod. Res.* **38**(3): 625–637.

Van Haandel, A. C., Ekama, G. A. and Marais, G. (1981). The activated sludge process: Part 3 - Single sludge denitrification, *Wat. Res.* **15**: 1135–1152.

Vanhooren, H. and Nguyen, K. (1996). Development of a simulation protocol for evaluation of respirometry-based control strategies, *Technical report*, BIOMATH, Univ. Gent, Gent, Belgium.

Vanrolleghem, P. A. (1994). *On-line modelling of activated sludge processes: development of an adaptive sensor*, PhD thesis, Laboratory of Microbial Ecology, Univ. Gent, Gent, Belgium.

Vanrolleghem, P. A. and Gillot, S. (2001). Robustness and economic measures as control benchmark criteria, *Wat. Sci. Tech.* (accepted).

Vanrolleghem, P. A., Kong, Z., Rombouts, G. and Verstraete, W. (1994). An online respirographic biosensor for the characterization of load and toxicity of wastewaters, *J. Chem. Technol. & Biotechnol.* **59**(4): 321–333.

Vanrolleghem, P. A., Vermeersch, L. K., Dochain, D. and Vansteenkiste, G. C. (1993). Modelling of a nonlinear distributed parameter bioreactor: optimisation of nutrient removal process, *in* A. Pavé (ed.), *Modeling and Simulation*, SCS, San Diego, USA, pp. 563–567.

Vetterli, M. and Herley, C. (1992). Wavelets and filter banks: theory and design, *IEEE Trans. Signal Process.* **40**(9): 2207–2232.

Wakeling, I. A. and Morris, J. J. (1993). A test of significance for partial least squares regression, *J. Chemometr.* **7**: 291–304.

Walczak, B. and Massart, D. L. (2001a). Dealing with missing data. Part I, *Chemometrics Intell. Lab. Syst.* **58**: 15–27.

Walczak, B. and Massart, D. L. (2001b). Dealing with missing data. Part II, *Chemometrics Intell. Lab. Syst.* **58**: 29–42.

Wangen, L. E. and Kowalski, B. R. (1988). A multiblock partial least squares algorithm for investigating complex chemical systems, *J. Chemometr.* **3**: 3–20.

WATERMATEX (2000). Symposium preprints, *5^th International Symposium on System Analysis and Computing in Water Quality Management (WATERMATEX 2000)*, BIOMATH, Univ. Gent, Gent, Belgium.

Weijers, S. (2000). *Modelling, identification and control of activated sludge plants for nitrogen removal*, PhD thesis, Techische Universiteit Eindhoven, The Netherlands.

Weiss, B., Ferry, M., Pons, M. N., Roche, N., Cecile, J. L. and Prost, C. (1998). Detection of pollution by toxics in wastewater treatment plants, *7^th International Conference on Computer Applications in Biotechnology*, Osaka, Japan, pp. 535–540.

Westerhuis, J. A., Kourti, T. and MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models, *J. Chemometr.* **12**(5): 301–321.

Whitfield, P. H. and Dohan, K. (1997). Identification and characterization of water quality transient using wavelet analysis. II. Application to electronic water quality data, *Wat. Sci. Tech.* **36**(5): 337–348.

Wikström, C., Albano, C., Eriksson, L., Fridén, H., Johansson, E., Kettaneh-Wold, N. and Wold, S. (1998). Multivariate process and quality monitoring applied to an electrolysis process part I. Process supervision with multivariate control charts, *Chemometrics Intell. Lab. Syst.* **42**: 221–231.

Wise, B. M. (1991). *Adapting multivariate analysis for monitoring and modeling of dynamic systems*, PhD thesis, Dept. of Chemical Engineering, University of Washington, Seattle, Washington, USA.

Wise, B. M. and Gallagher, N. B. (1996a). *PLS_Toolbox for use with MATLAB*, Eigenvector Research, Inc., Manson, Washington, USA.

Wise, B. M. and Gallagher, N. B. (1996b). The process chemometrics approach to process monitoring and fault detection, *J. Process Control* **6**(6): 329–348.

Wise, B. M. and Ricker, N. L. (1993). Identification of finite impulse response models with continuum regression, *J. Chemometr.* **7**: 1–14.

Wise, B. M., Veltkamp, D. F. and Kowalski, B. R. (1990). A theoretical basis for the use of principal component models for monitoring multivariate processes, *Process Control Qual.* **1**: 41–51.

Wold, H. (1966). Nonlinear estimation by iterative least squares procedures, *in* F. David (ed.), *Research papers in statistics*, John Wiley & Sons, Inc., New York, New York, USA, pp. 411–444.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models, *Technometrics* **20**(4): 397–405.

Wold, S. (1991). Chemometrics, why, what and where to next, *J. Pharm. Biomed. Anal.* **9**(8): 589–596.

Wold, S. (1994). Exponentially weighted moving principal components analysis and projection to latent structures, *Chemometrics Intell. Lab. Syst.* **23**: 149–161.

Wold, S., Esbensen, K. and Geladi, P. (1987a). Principal component analysis, *Chemometrics Intell. Lab. Syst.* **2**: 37–52.

Wold, S., Geladi, P., Esbensen, K. and Öhman, J. (1987b). Multi-way principal components- and PLS-analysis, *J. Chemometr.* **1**: 41–56.

Wold, S., Kettaneh-Wold, N. and Skagerberg, B. (1989). Nonlinear PLS modeling, *Chemometrics Intell. Lab. Syst.* **7**: 53–65.

Wold, S., Kettaneh-Wold, N. and Tjessem, K. (1996). Hierarchical multiblock PLS and PC models for easier interpretation and as an alternative to variable selection, *J. Chemometr.* **10**: 463–482.

Yoo, C., Choi, S. W. and Lee, I. (2001). Disturbance detection and isolation in the activated sludge process, *Wat. Sci. Tech.* (accepted).

Yoon, S. and MacGregor, J. F. (2001). Fault diagnosis with multivariate statistical models part I: using steady state fault signatures, *J. Process Control* **11**: 387–400.

Yuan, Z., Bogaert, H., Rosen, C. and Verstraete, W. (2001a). Sludge blanket height control in secondary clarifiers, *1st IWA Conference on Instrumentation, Control and Automation (ICA2001)*, Vol. 1, Malmö, Sweden, pp. 81–88.

Yuan, Z., Bogaert, H., Vanrolleghem, P. A., Thoye, C., Vansteenkiste, G. C. and Verstraete, W. (1996). Carbon dosage control for predenitrification processes, *10th Forum Applied Biotechnology (FAB)*, Gent, Belgium, pp. 1733–1743.

Yuan, Z., Keller, J. and Lant, P. (2001b). Optimization and control of biological nitrogen removal activated sludge processes: a review of recent developments, *in* S. N. Agathos and W. Reineke (eds), *Focus on Biotechnology*, Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Zhang, J., Martin, E. B. and Morris, A. J. (1997). Process monitoring using non-linear statistical techniques, *Chem. Eng. J.* **67**: 181–189.

Zullo, L. (1996). Validation and verification of continuous plants operating modes using multivariate statistical methods, *Comput. Chem. Eng.* **20**(Suppl.): S683–S688.