



Data-driven soft-sensors for monitoring and fault diagnosis in wastewater treatment plants

Pezhman Kazemi

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.



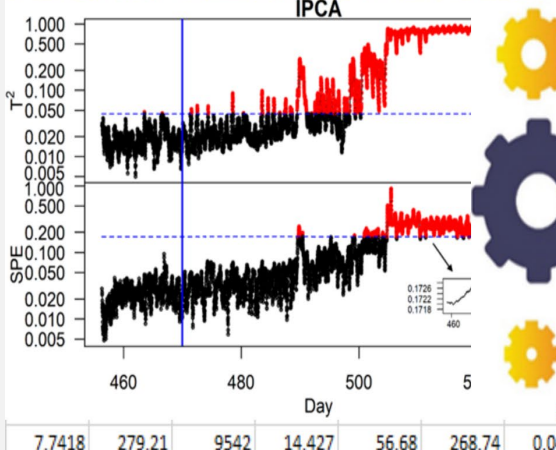
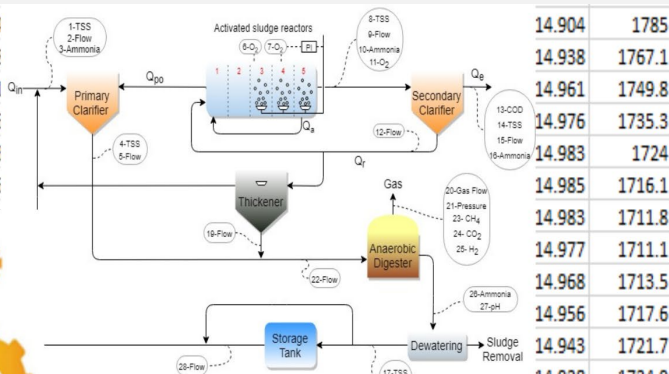
UNIVERSITAT
 ROVIRA I VIRGILI

Data-driven soft-sensors for monitoring and fault diagnosis in wastewater treatment plants

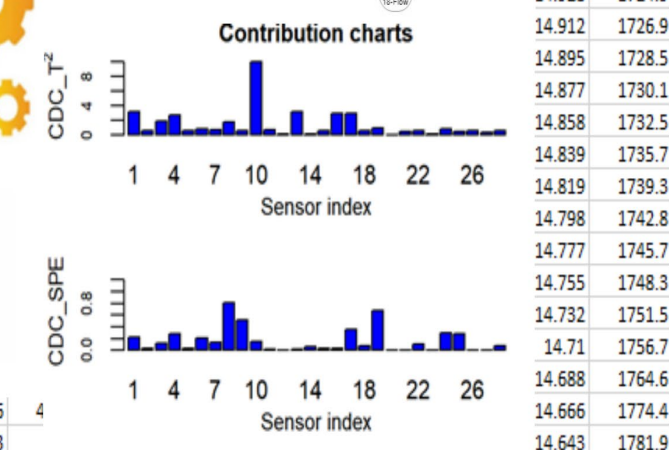
PEZHMAN KAZEMI



233	7.7008	200.06	15031	14.904
294	7.7063	201.68	14457	14.938
324	7.7113	202.44	13795	14.961
352	7.716	202.61	13147	14.976
392	7.7204	202.32	12550	14.983
451	7.7246	201.61	12026	14.985



325	7.7578	152.95	8594.6	14.666						
7.7418	279.21	9542	14.427	56.68	268.74	0.0452	7.7569	150.07	9475.2	14.643



14.904	1785
14.938	1767.1
14.961	1749.8
14.976	1735.3
14.983	1724
14.985	1716.1
14.983	1711.8
14.977	1711.1
14.968	1713.5
14.956	1717.6
14.943	1721.7
14.928	1724.9
14.912	1726.9
14.895	1728.5
14.877	1730.1
14.858	1732.5
14.839	1735.7
14.819	1739.3
14.798	1742.8
14.777	1745.7
14.755	1748.3
14.732	1751.5
14.71	1756.7
14.688	1764.6
14.666	1774.4
14.643	1781.9

DOCTORAL THESIS
 2020

Data-driven soft-sensors for monitoring and fault diagnosis in wastewater treatment plants

DOCTORAL THESIS

Author:

Pezhman Kazemi

Advisors:

Prof. Jaume Giralt i Marcé

Dr. Jean-Philippe Steyer

Departament d'Enginyeria Química (DEQ)



UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2020



UNIVERSITAT ROVIRA I VIRGILI

Departament d'Enginyeria Química (DEQ)
Av. Paisos Catalans, 26
43007 Tarragona, Spain

We STATE that the present study, entitled "Data-driven soft-sensors for monitoring and fault diagnosis in wastewater treatment plants", presented by Pezhman Kazemi, for the award of the degree of Doctor, has been carried out under our supervision at the Departament d'Enginyeria Química (DEQ).

Tarragona, 1st October 2020.

Doctoral Thesis Supervisors,

A handwritten signature in blue ink, appearing to be 'JG', written over a horizontal line.

Prof. Jaume Giralt i Marcé

A handwritten signature in blue ink, appearing to be 'J. Steyer', written in a cursive style.

Dr. Jean-Philippe Steyer

UNIVERSITAT ROVIRA I VIRGILI

Data-driven soft-sensors for monitoring and fault diagnosis in wastewater treatment plants

Pezhman Kazemi

تقدیم بہ پدر بزرگوار و مادر عزیزم

Dedicated to My kind father and dear mother

Abstract

Failing to reach the specific effluent properties in wastewater treatment plants can adversely affect human health and environmental. Due to this, there are significant pressures on authorities for efficient design and operation of wastewater treatment plants (WWTPs). Therefore, to achieve regulatory standards for wastewater effluent in a cost-efficient way, the development of an advanced information framework for the control and supervision of the WWTPs is mandatory. For the implementation of this framework, the real-time measurements of crucial parameters (e.g., concentrations of nitrate and total nitrogen, phosphate and total phosphorus, suspended solids, biochemical oxygen demand (BOD) and chemical oxygen demand (COD), total volatile fatty acids (VFA)) are necessary. Measurement of such parameters is often associated with capital and maintenance costs, as well as the time delay.

The focus of this thesis was to design soft-sensors that can be used besides conventional instrumentation to improve the process operation and safety. Due to the availability of the massive amount of process data in most modern WWTPs, data-driven methods have attracted significant attention. Therefore, in this thesis, we developed different data-driven soft-sensors for online prediction of a crucial parameter (for instance, VFA) and fault detection (FD) and diagnosis in WWTPs.

Firstly, we propose different data-driven soft-sensor for estimating total VFA concentration in the anaerobic digester. We evaluated random forest (RF), artificial neural network (ANN), extreme learning machine (ELM), support vector machine (SVM) and genetic programming (GP) based on synthetic data obtained from the International Water Association (IWA) Benchmark

Simulation Model No. 2 (BSM2). In addition, the model robustness was assessed to determine the performance of each soft-sensor under different process states.

Second, to prevent failures and serious consequences during the running of the anaerobic digestion (AD) plant, the VFA soft-sensors using different advanced techniques such as SVM, ELM and ensemble of neural network (ENN) are tested and compared in terms of accuracy and robustness for detecting process and instrument faults. To compare the proposed approaches with the traditional FD method, a principal component analysis (PCA) model was also developed. By applying soft-sensors, the residual signal, i.e., the difference between estimated and measured VFA values, can be generated. This residual signal was used in combination with univariate statistical control charts to detect the faults.

Third, we propose a complete adaptive process monitoring framework based on incremental principal component analysis (IPCA). This framework updates the eigenspace by incrementing new data to the PCA at a low computational cost. The contribution of variables is also recursively provided using a complete decomposition contribution (CDC). For the imputation of missing values, the empirical best linear unbiased prediction (EBLUP) method is incorporated into this framework.

Overall, this thesis presents the application of different data-driven soft-sensors for online prediction and FD in WWTP; it is also shown that they have strong potential for providing support to the operation of water treatment facilities.

Keywords: BSM2, Bootstrapping, Anaerobic digestion, Soft-Sensor, Neural network, CUSUM chart, Incremental PCA, BSM2, EBLUP, Fault detection, Fault isolation, Time-varying processes; data-driven, genetic programming.

UNIVERSITAT ROVIRA I VIRGILI

Data-driven soft-sensors for monitoring and fault diagnosis in wastewater treatment plants

Pezhman Kazemi

Acknowledgments

I would like to express my gratitude to my supervisors Prof. Jaume Giralt i Marcé and Dr. Jean-Philippe Steyer, for their useful guidance, insightful comments, and considerable encouragement to complete this thesis. They have guided me to pursue important problems that will have a practical impact and were always available to guide me whenever I approached them. This work would not have been completed without their encouragement and patience.

I would like to express my gratitude to Dr. Esther Torrens Serrahima and Ms. Núria Juanpere Mitjana for all the administrative support that they have done for me.

I had the chance to go to the Laboratory of Environmental Biotechnology (LBE), Narbonne, France, for four months. I would like to thank all the people at the LBE for giving me this opportunity to stay with them.

My family has always supported me, and for this I would like to thank my mother, father, and sister.

I also thank my friends Moein, Hossein, Sepehr, and Nasibeh, whom I can always count on, to share my woes and my happiness.

Last but not least, for the financial support, I would like to acknowledge Universitat Rovira I Virgili Marti Franques scholarship 2016PMF-PIPF-28.

UNIVERSITAT ROVIRA I VIRGILI

Data-driven soft-sensors for monitoring and fault diagnosis in wastewater treatment plants

Pezhman Kazemi

Contents

Abstract	i
Acknowledgments	iv
List of Figures.....	xi
List of Tables.....	xiv
Nomenclature	xvi
1. Introduction	1
1.1. Background.....	2
1.2. Objectives and scope of the research.....	4
1.3. Scientific dissemination	5
2. Wastewater treatment plant	7
2.1. Introduction to wastewater treatment	8
2.2. Wastewater characterization	9
2.3. Unit operations in WWTP.....	11
2.3.1. Pre-treatments	11
2.3.2. Primary treatments	12
2.3.2.1. Equalization tank.....	13
2.3.2.2. Primary settling	13
2.3.3. Secondary treatments	13
2.3.3.1. Activated Sludge Process	14
2.3.3.2. Secondary settler	15
2.3.4. Tertiary treatments.....	16
2.3.4.1. Filtration.....	16
2.3.4.2. Disinfection.....	17
2.3.5. Sludge treatment	17

2.3.5.1. Thickening	17
2.3.5.2. Anaerobic digester.....	18
2.3.5.3. Dewatering	18
2.3.5.4. Disposal.....	19
2.4. Benchmark simulation No.2 (BSM2).....	19
2.4.1. Plant layout.....	20
2.4.1. Control strategy.....	23
3. Data-driven soft-sensors	25
3.1. Introduction to soft-sensors	26
3.2. Designing data-derived soft-sensors.....	29
3.2.1. Data acquisition	29
3.2.2. Data pre-processing	30
3.2.2.1. Variable selection	30
3.2.2.2. Outlier detection.....	32
3.2.2.3. Model selection, training and validation.....	33
3.2.2.3. Soft-sensor maintenance.....	35
3.3. Applications of soft-sensors	36
3.3.1. Online prediction	36
3.3.1.1. Transfer function models.....	36
3.3.1.2. Multiple regression.....	37
3.3.1.3. Artificial neural networks (ANNs).....	38
3.3.1.4. Support vector machines (SVMs).....	40
3.3.2. Monitoring and fault detection	41
3.3.2.1. Control charts	42
3.3.2.2. Principal component analysis (PCA)	44
3.3.2.3. Partial least squares (PLS)	45
3.3.2.4. Artificial neural networks (ANNs).....	46
3.3.2.5. Support vector machines (SVMs).....	48

4. Data-Driven Soft Sensors for Online Monitoring of Volatile Fatty Acids in Anaerobic Digestion Processes.....	49
4.1. Introduction	50
4.2. Materials and Methods.....	54
4.2.1. Data Collection	54
4.2.2. Pre-Processing of the Data	54
4.3. Data-Driven Methods.....	56
4.3.1. Artificial Neural Network (ANN)	56
4.3.2. Extreme Learning Machine (ELM)	58
4.3.3. Random Forest (RF)	60
4.3.4. Support Vector Machine (SVM)	60
4.3.5. Genetic Programming (GP).....	61
4.3.6. Feature Ranking.....	63
4.4. Results and Discussion.....	64
4.4.1. Studying the Relationship between Input and Output Data	64
4.4.2. Choosing the Most Influential Variables Using the Feature Ranking Method	67
4.4.3. Soft Sensor Design.....	69
4.4.4. Evaluation of the Robustness of Soft Sensors	73
4.5. Conclusions	75
5. Data-driven techniques for fault detection in anaerobic digestion process	77
5.1. Introduction	78
5.2. Materials and methods	83
5.2.1. Data collection and pre-processing.....	83
5.2.2. Soft-sensor assessment.....	83
5.2.3. Fault detection assessment	84
5.3. Data-driven techniques	84

5.3.1. Artificial neural network (ANN).....	84
5.3.2. Support vector machine (SVM).....	85
5.3.3. Principal component analysis (PCA).....	85
5.3.4. Feature selection	86
5.4. Statistical process control.....	86
5.4.1. Control charts.....	86
5.4.2. Bootstrap confidence limits.....	87
5.5. Proposed approach.....	88
5.6. Developing soft-sensors for FD	90
5.7. Fault detection and discussion.....	93
5.7.1. Fault diagnosis without missing values	96
5.7.2. Fault diagnosis with missing values	100
5.8. Conclusion	101
6. Fault detection and diagnosis in water resource recovery facilities using incremental PCA	103
6.1. Introduction	104
6.2. Conventional PCA	107
6.3. Incremental PCA.....	108
6.3.1. Estimating the number of PCs.....	110
6.3.2. Process monitoring statistics thresholds	111
6.3.3 Fault Isolation	111
6.3.4. Missing data imputation.....	112
6.4. Adaptive Fault Detection and Isolation	113
6.5. Simulation results	115
6.5.1. Description of the Water resource recovery facility	115
6.5.2. Fault detection and isolation.....	115
6.5.2.1. Normal operation of the WWTP	117
6.5.2.2. Drift fault in the dissolved oxygen sensor.....	118

6.5.2.3. Step change in inorganic nitrogen in the anaerobic digester.....	120
6.5.2.4. Step change in the settling velocity of the secondary clarifier....	121
6.5.2.5. Step change in the bioreactor parameters	122
6.5.2.6. Storm events.....	123
6.6. Fault detection with missing values.....	124
6.7. Conclusion	126
7. Concluding remarks	129
7.1. Summary of the results	130
7.1.1. Online prediction	130
7.1.2. Fault detection and isolation.....	131
7.2. Future research lines	132
References	136

List of Figures

Figure 2. 1 Location of different unit operations in a conventional wastewater treatment plant diagram.	12
Figure 2. 2 A generalized schematic diagram of an activated sludge process (“Activated sludge - Wikiwand”).....	15
Figure 2. 3 A simplified flow diagram of the biogas process (Angelidaki et al., 2003).....	19
Figure 2. 4 Plant layout for BSM2 (Nopens et al., 2010).	20
Figure 2. 5 Definition of the response time (Rieger et al., 2003).	23
Figure 2. 6 Schematic of the control strategy for aerobic reactors in BSM2.....	24
Figure 3. 1 Model vs. data-driven soft sensors (Fortuna et al., 2007).	26
Figure 3. 2 Schematic of simple soft-sensor.	27
Figure 3. 3 Application fields of soft-sensors (Kadlec, 2009).....	28
Figure 3. 4 Overview of steps needed for developing data-driven soft-sensors (Haimi et al., 2013; Kadlec, 2009).....	29
Figure 4. 1 Simple scheme of a neuron.....	57
Figure 4. 2 Scheme of the extreme learning machine (ELM) model.....	58
Figure 4. 3 Computation flowchart of genetic programming (GP).	63
Figure 4. 4 Importance of variables on a scale from 0 to 100 obtained with the fscaret method.	68
Figure 4. 5 Trend of input and output variables used for developing soft sensors.	69
Figure 4. 6 Prediction results of different soft sensors; black is the actual values and blue is the predicted volatile fatty acids (VFA) values. GP: genetic programming; SVM: support vector	

machine; ANN: artificial neural network; ELM: extreme learning machine; RF: random forest.	72
Figure 4. 7 Results of the robustness evaluation for each model: (a) for $k_{m,ac}$; (b) for $k_{hyd,ch}$; (c) for $k_{I,NH3}$. NRMSE: normalized root-mean-squared error; avNRMSE: average normalized root-mean-squared error.	74
Figure 5. 1 General diagram of the proposed FD framework.	90
Figure 5. 2 Trend of input and output variables used for developing soft-sensors. Left side of the red line used for training and the right side used for testing.....	92
Figure 5. 3 Density estimation and probability plots of soft-sensor residual for normal operation.	95
Figure 5. 4 FD performance for variation of $k_{I,NH3}$ for the complete data set.....	97
Figure 5. 5 FD performance for pH sensor faults among different charts for complete data	99
Figure 5. 6 FD performance for variation of $k_{I,NH3}$ among different charts for uncompleted data.	101
Figure 6. 1 Adaptive IPCA monitoring diagram.....	114
Figure 6. 2 Schematic diagram of BSM2 and locations of the measured variables.	116
Figure 6. 3 Fault detection results of PCA and IPCA for normal operation (dashed blue line: 99% confidence limit; solid vertical blue line: onset of the fault). The small plot in the IPCA chart shows the adaptive threshold for SPE.....	117
Figure 6. 4 Fault detection and diagnosis results of IPCA for drift fault in the dissolved oxygen sensor (dashed blue line: 99% confidence limit; solid vertical blue line: onset of the fault). The small plot inside the IPCA chart shows the adaptive threshold for SPE.	119

Figure 6. 5 Fault detection and diagnosis results of IPCA for a step change in AD inorganic nitrogen (dashed blue dashed: 99% confidence limit; solid vertical blue line: onset of the fault). The small plot inside the IPCA chart shows the adaptive threshold for SPE.	120
Figure 6. 6 Fault detection and diagnosis results of IPCA for a step-change in the settling velocity of the secondary clarifier (dashed blue line: 99% confidence limit; solid vertical blue line: onset of the fault). The small plot inside the IPCA chart shows the adaptive threshold for SPE.	122
Figure 6. 7 Fault detection and diagnosis results of IPCA for a step change in the specific growth rate of the autotrophs (dashed blue line: 99% confidence limit; solid vertical blue line: onset of the fault). The small plot inside the IPCA chart shows the adaptive threshold for SPE.	123
Figure 6. 8 Sensors' contribution to T^2 and SPE statistics at different days of operation under storm condition.	124
Figure 6. 9 Pattern of missing values during the normal running of the WWTP (from day 457 to day 488).	125

List of Tables

Table 2. 1 Main components of domestic wastewater (Henze et al., 2008).....	9
Table 2. 2 Typical composition of raw municipal wastewater with minor contributions of industrial wastewater (Henze et al., 2008).....	10
Table 2. 3 Definition of variables included in the input file of BSM2.	21
Table 2. 4 Sensor classes (Rieger et al., 2003).....	22
Table 3. 1 Abnormal patterns in univariate WWTP data that could indicate a fault and potential causes of the fault pattern (Capizzi and Masarotto, 2017).....	42
Table 4. 1 Obtained variables from Benchmark Simulation Model No.2 (BSM2).	54
Table 4. 2 Performance of the linear regression (LR) models based on the default influent file and different input vectors.	65
Table 4. 3 Summary of default and the modified values of inorganic nitrogen (S_{in}), composite (X_c) and carbohydrate (X_{ch}).....	66
Table 4. 4 Performance of the LR models based on the modified anaerobic digestion (AD) variables and different input vectors.	66
Table 4. 5 Result of different subsets trained by support vector machine (SVM).....	69
Table 4. 6 Final tuning parameters of soft sensors obtained with the grid search.	70
Table 4. 7 Results of soft sensors for the training and validation sets.	71
Table 4. 8 Coefficient table for the equation obtained by GP.....	72
Table 5. 1 Results of soft-sensors on the training and validation set.	93
Table 5. 2 Faults in BSM2 simulation.....	93

Table 5. 3 Robustness of SVM, ELM and ENN in the prediction of VFA affected by $k_{I,NH3}$ changes.....	95
Table 5. 4 Performance of FD methods for different faults (the meaning of the fault number is presented in Table 5.2).....	96
Table 5. 5 Detection delays (in samples) for different charts (the meaning of the fault number is presented in Table 2).....	98
Table 6. 1 Description of the simulated faults	117
Table 6. 2 Fault detection performance for complete and uncompleted data set.....	126

Nomenclatures

AD	Anaerobic Digestion
ADM1	Anaerobic Digestion Model
AIC	Akaike's Information Criterion
ALR	Alkalinity Loading Rate
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Network
ARIMA	Integrated Moving Average
ASM1	Activated Sludge Model No. 1
BOD	Biochemical Oxygen Demand
BSM1	Benchmark Simulation Model No. 1
BSM1_LT	Benchmark Simulation Model No. 1 Long-Term
BSM2	Benchmark Simulation Model No. 2
CDC	Complete Decomposition Contribution
CEPT	Chemically Enhanced Primary Treatment
COD	Chemical Oxygen Demand
CP	Mallow's Coefficient
CPV	Cumulative Percent Variance
CUSUM	Cumulative Sum Control Chart
D	Granule Size
EBLUP	Best Linear Unbiased Prediction
ELM	Extreme Learning Machine
ENN	Ensemble of Neural Network
EWMA	Exponentially Weighted Moving Average
EWPLS	Exponentially Weighted PLS
FAR%	False Alarm Rate

FD	Fault Detection
FN	False Negatives
FOP	First-Order Perturbation Analysis
ForceCA	Forecastable Component Analysis
GA	Genetic Algorithm
GP	Genetic Programming
GS	Grid Search
HLR	Hydraulic Loading Rate
IPCA	Incremental Principal Component Analysis
IWA	International Water Association
KDE	Kernel Density Estimation
KPLS	Kernel Partial Least Squares
L1 And L2	Regularization Term
LCL	Lower Control Limits
LLE	Linear Embedding
LR	Linear Regression
MAPE	Mean Absolute Percentage Error
MBPLS	Multiblock Partial Least Squares
MBR	Membrane Bioreactor
MDR%	Missed Detection Rate
MLP	Multi-Layer Perceptron
MLR	Multiple Linear Regression
MLSS	Mixed Liquor Suspended Solids
MLVSS	Mixed Liquor Volatile Suspended Solids
MSE	Mean Squared Error
MSPC	Multivariate Statistical Process Controls
MWPCA	Moving Window PCA
NFS	Neuro-Fuzzy System

NNPL	Neural Network PLS
NRMSE	Normalized Root-Mean-Squared Error
OLR	Organic Loading Rate
PCA	Principal Component Analysis
PCR	Principal Component Regression
PCs	Principal Components
PI	Proportional-Integral
PLS	Partial Least Squares
R ²	Coefficient of Determination
RAM	Random-Access Memory
RBF	Radial Basis Function
RBFN	Radial Basis Function Network
RF	Random Forest
RMSE	Root Mean Squared Error
RNNs	Recursive Neural Networks
RPCA	Recursive PCA
RPLS	Recursive PLS
SGD	Stochastic Gradient Descent
SLFN	Single Hidden Layer Feed-Forward Neural Network
SOM	Self-Organizing Map
SPC	Statistical Process Control
SPE	Squared Prediction Error
SVI ₃₀	Sludge Volume Index At 30 Min
SVI ₅	Sludge Volume Index At 5 Min
SVM	Support Vector Machine
SVR	Support Vector Regression
T ₂	Hotelling's T-Squared
TF	False Negatives

TOC	Total Organic Carbon
TP	Total Phosphorus
TP	True Positives
TSS	Total Suspended Solid
UCL	Upper Control Limits
UV	Ultraviolet
VE	Variable Extraction
VFA	Volatile Fatty Acid
VMP	Volumetric Methane Production Rate
VS	Variable Selection
WMA	Weighted Moving Average
WWQI	Wastewater Quality Index
WWTPs	Wastewater Treatment Plants

CHAPTER 1

Introduction

This first chapter highlights the details about the challenges in the control and monitoring of the wastewater treatment plants and explains the significant objectives of this thesis. Also, it presents some of the leading publications concluded from the thesis.

1.1. Background

The recent strict regulations established by governments to improve the effluent quality of the wastewater treatment plants (WWTPs) caused significant pressures on companies for efficient design and operation of these plants. There are two options available to meet the government's requirements: either building new high-performance facilities or enhancing the efficiency of existing plants by incorporating advanced monitoring and control techniques. The first option is likely impossible due to the high capital investment needed for the construction of new WWTPs. Moreover, the required land for the construction of new wastewater treatment sites may not be readily available because of environmental restrictions (Olsson, 2012; Olsson et al., 2005). Therefore, if the second option is adequately implemented, it can increase the effluent water quality, reduce the chemical and energy consumption, and decrease the operational cost. To solve the current WWTPs problems, the development of an advanced information system for the control and supervision of the process is mandatory. The efficient implementation of advanced monitoring and control systems in WWTP is highly reliant on the availability of different sensors and high-performance controllable actuators. The standard monitoring practice in WWTPs relies on the online and offline analysis of the primary variables such as concentrations of ammonia, nitrate and total nitrogen, phosphate and total phosphorus, suspended solids, Biochemical Oxygen Demand (BOD) and Chemical Oxygen Demand (COD). Measuring these variables is often associated with the capital and maintenance costs; moreover, due to their time-delayed responses, they are not suitable for real-time monitoring.

Furthermore, the harsh environment of WWTPs (solids deposition, biofilm formation and precipitates) changes the accuracy and reliability of the measurements and increases the need for maintenance of the instrumentation. Due to these reasons, very few wastewater treatment plants

were only equipped with some simple sensors and control loops, which were mainly used for flow measurement and primary plant performance monitoring. However, with the widespread advent of intelligent sensors that are capable of self-calibration, self-cleaning and self-reconfiguration, this situation changed rapidly over the last years. However, despite the remarkable advances in the field of intelligent sensors, the accurate process measurement for some primary variables is still challenging. Currently, for new and upgraded WWTPs, the trend is toward the utilization of advanced control and monitoring systems; for this reason, many online easy-to-measure process variables such as temperature, pressure, flow rate, level measurements, conductivity, turbidity, pH, and perhaps, dissolved oxygen and ammonia concentration are frequently measured. Thus, one solution to the real-time measurement challenges is utilizing these historical process data to estimate the primary variables that are useful for monitoring both processes and instruments. Primary variables are always dependent on some of the secondary variables; therefore, the availability of the secondary variables allows developing models that can reconstruct the relationship between these two categories of variables. These models are often called soft-sensor; they are computer programs that are capable of predicting the primary variables that are crucial for efficient process operation where their hardware measurements are not sufficiently reliable, and in some plants, they do not exist at all. Another successful application of soft-sensors is process monitoring and process fault detection (FD) by providing information to plant operators about abnormal events or changing process states (Kadlec, 2009).

1.2. Objectives and scope of the research

The main objective of this thesis was to investigate the utilization of data-driven techniques for developing different soft-sensors which can be used in modeling and monitoring of WWTPs.

In order to achieve this objective, the following detailed objectives were addressed:

1. Developing data-driven soft-sensors to predict the total volatile fatty acids concentration (VFA) in AD process using simple measurements such as ammonia concentration, pressure, CO₂ mole fraction, and pH. Using VFA soft-sensor can overcome the hardware sensor problems (high capital and maintenance cost, time-delayed responses) and achieve lower costs and greater simplicity that is highly attractive for the biogas sector. The detailed procedure for developing VFA soft-sensors using a variety of artificial intelligence methods is presented in Chapter 4.
2. Due to the complexity of the AD process, recovering from failures is time-consuming and expensive; for this reason, a reliable early FD procedure is needed. This problem was addressed using the developed VFA soft-sensor in Chapter 5. The secondary objective was to find out which artificial intelligence techniques perform better in terms of FD. The other secondary objective was to improve the developed VFA soft-sensors accuracy for early FD in AD by integrating it with the statistical process control (SPC) approaches mainly the cumulative sum control chart (CUSUM).
3. The focus of previous objectives was mainly on the AD process. In this regard, VFA concentration in the AD was chosen as a quality parameter and predicted by applying soft-sensor. Applying the soft-sensors provides the residual signal, i.e., the difference between estimated and measured VFA values. This residual can be used for FD. However, specifying

the quality parameter for monitoring all parts of WWTP is not easy, as many parameters can be considered for this task. Moreover, due to a large number of sensors measurements in the WWTP compared to the AD process, supervised methods are not the best option. Therefore, to solve this problem, an unsupervised method (incremental principal component analysis (IPCA)) is proposed. Due to the fluctuation in the flow rate and composition of the feed stream, the WWTP exhibits very nonstationary behaviors. The main objective was to cope with this behavior by adapting the IPCA for FD and isolation instead of normal PCA. The secondary objective was to analyze the accuracy of the IPCA during different faults in WWTP. The other secondary objective was to check the ability of the proposed framework in accurate isolation of the faults. In Chapter 6, these objectives were addressed.

1.3. Scientific dissemination

- 1- Fault detection and diagnosis in water resource recovery facilities using incremental PCA, **Pezhman Kazemi**, Jaume Giralt, Christophe Bengoa, Armin masoumian, Jean-Philippe Steyer. Water Sci Technol, (Accepted) (<https://doi.org/10.2166/wst.2020.368>).
- 2- Robust data-driven soft-sensors for online monitoring of volatile fatty acids in anaerobic digestion processes. **Pezhman Kazemi**, Christophe Bengoa, Jean-Philippe Steyer, Josep Font, Jaume Giralt. Processes, 2020, 8(1), 67 (<https://doi.org/10.3390/pr8010067>).
- 3- Data-driven fault detection methods for detecting small-magnitude faults in anaerobic digestion process. **Pezhman Kazemi**, Jaume Giralt, Christophe Bengoa, Jean-Philippe Steyer. Water Sci Technol (2020) 81 (8): 1740–1748 (<https://doi.org/10.2166/wst.2020.026>).

4. Data-driven techniques for fault detection in anaerobic digestion process. **Pezhman Kazemi**, Christophe Bengoa, Jean-Philippe Steyer, Jaume Giralt. Process Safety and Environmental Protection (Accepted).

CHAPTER 2

Wastewater treatment plant

In this chapter, the brief description of the wastewater treatment plant and the main phenomena that take place in this process will be discussed. This description provides an overview of the unit operations and commonly available measurements of modern WWTPs, which designed for removal of total nitrogen.

2.1. Introduction to wastewater treatment

The propose of the wastewater treatment plants is to eliminate the toxic compounds from incoming wastewater by concentrating them into the sludge using a combination of mechanical, physical, chemical, and biological processes. The effluent of the wastewater treatment plants can be either returned to surface water with minimal environmental impacts or reused in the industrial process that generated it. The sludge stream, which is the by-product of the process, contains high levels of various toxic contaminants; thus needs to be adequately treated before disposal to the environment.

The term "wastewater" includes broad ranges of polluted water coming from different sources; however, it is often summarized in two main categories of municipal and industrial wastewater. The primary source of municipal wastewater is typically originated from household activities and sometimes contains liquid waste from small industries and stormwater or urban runoff. Industrial wastewater may contain different pollutions according to its source. Due to its contaminants that are toxic to microbial growth (which is the base of traditional wastewater treatment plants), this kind of wastewater can not be treated using traditional wastewater treatment plants. Therefore, in order to discharge the industrial wastewater to the sewer system, several preliminary treatments need to be done to eliminate the incompatible pollutants. For most of the industrial processes that are not fulfilled the regulation for discharging the wastewaters to the sewer system, special treatment must be performed in situ in the production place.

2.2. Wastewater characterization

The composition of wastewater is affected by many factors. For instance, not all human wastewater has the same composition; it is very dependent on behavior, lifestyle, and standard of living of inhabitants. Generally, the components of wastewater can be divided into different main categories, according to Table 2.1.

Table 2. 1 Main components of domestic wastewater (Henze et al., 2008).

Component	Source	Effect
Microorganisms	Pathogenic bacteria, virus, antibiotics, worm eggs	Risk when bathing and eating shellfish Fish
Biodegradable organic materials	Oxygen depletion in rivers, and lakes	Fish death, odors
Other organic materials	Detergents, pesticides, fat, oil and grease, coloring, solvents, phenols, cyanide	Toxic effect, aesthetic inconvenience, bioaccumulation in the food chain
Nutrients	Nitrogen, phosphorous, ammonium	Eutrophication, oxygen depletion, toxic effect
Metals	Nitrogen, phosphorous, ammonium	Toxic effect, bioaccumulation
Other inorganic materials	Acids (typically hydrogen sulfide) bases	Corrosion, toxic effect Changing
Thermal effect	Hot water	Changing living conditions for flora and fauna
Odor (and taste)	Hydrogen sulfide	Aesthetic inconvenience, toxic effect
Radioactivity		Toxic effect, accumulation

The main source of pollutions in wastewater is organic matter, which their contents traditionally measured in terms of COD and BOD. Measuring the first one is quick, while the second one requires a slow procedure. The COD measures the organic contents of wastewater indirectly and can be estimated by the amount of oxygen needed to chemically oxidize the organic

matter. BOD measures the organic pollutions indirectly and representing the amount of oxygen consumes per liter by aerobic bacteria to oxidize the organic contents at 20 °C during a specific amount of time. The standard BOD analysis takes five days (BOD₅); however, sometimes, the other alternative may be used.

The composition of municipal wastewater varies by location and by time in the same location. The variation of water consumption by households and infiltration and exfiltration during transport to the sewage system are the primary reasons for this composition variation. Table 2.2 shows the typical municipal wastewater composition. The high values represent the wastewater with low water consumption and/or infiltration. On the other hand, low values represent high water consumption and/or infiltration. Due to the high load of water during the storm events, the concentration of components in most stormwater is lower compared to the very diluted wastewater.

Table 2. 2 Typical composition of raw municipal wastewater with minor contributions of industrial wastewater (Henze et al., 2008).

Parameters	High	Medium	Low
COD total	1200	750	500
COD soluble	480	300	200
COD suspended	720	450	300
BOD	560	350	230
VFA (as acetate)	80	30	10
N total	100	60	30
Ammonia-N	75	45	20
P total	25	15	6
Ortho-P	15	10	4
TSS	600	400	250
VSS	480	320	200

A diurnal pattern of the water consumption related to human activities is another important factor that has to take into account for the design of the WWTP and its control systems. The diurnal pattern is typically showing lesser flow in the early morning while it peaks around lunch and dinner time. The level of these fluctuations directly depends on the size of the communities, with smaller communities, the difference between the lower and the upper discharge peak is higher.

2.3. Unit operations in WWTP

Wastewater treatment plants consist of a sequence of unit operations in which physical, chemical, and biological processes or their combinations are used to remove pollutants from wastewater (Hreiz et al., 2015). The streams in WWTPs can be divided into two main production lines: the water line and the sludge line. The typical sequence of treatment on the water line in an activated sludge plant includes preliminary (removal of the heaviest solid materials and oil separation), secondary (removal of suspended organic and pollutant compounds by biological processes) and tertiary (polish the biological treatment before delivering the water to the river body) treatments. The sludge line is devoted to treating the sludge, which is produced during the process, and its main goals are reducing the water content, the volume, and the microbial content in this kind of waste. Figure 2.1 shows a diagram of unit operations for a common wastewater treatment plant.

2.3.1. Pre-treatments

The first operation in WWTPs is Pre-treatments. Municipal wastewater always contains a

wide verity of large solids matter such as branches, dead leaves, rocks, trash, rugs and etc. For this reason, pre-treatment is essential in order to protect the downstream operations (pumps, pipelines, and other equipment) against blockage and abrasion by removing these solid matters from the influent stream. Pre-treatment operations include screening, shredding, grit removal, pre- aeration, and chemical addition (Metcalf & Eddy et al., 2014).

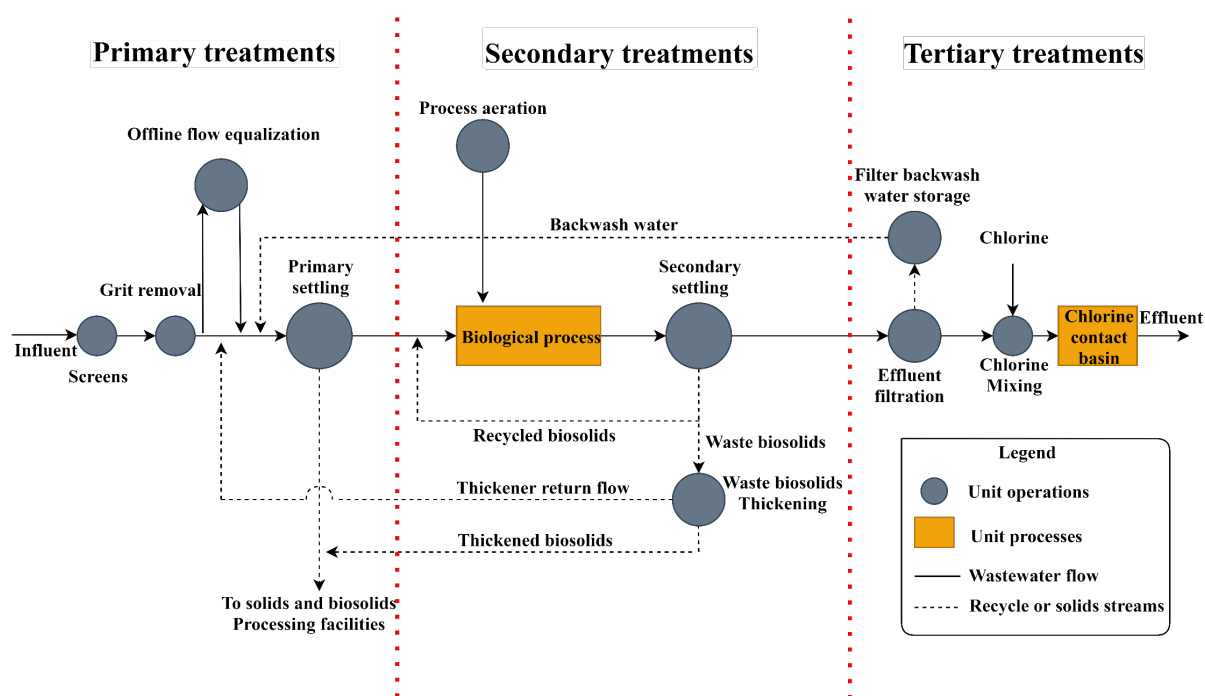


Figure 2. 1 Location of different unit operations in a conventional wastewater treatment plant diagram.

2.3.2. Primary treatments

The purpose of primary treatments is to reduce the fluctuations in the flow and the organic load of the influent stream and simultaneously, removing settleable solids and floatable materials from wastewater. Typically, during primary treatments, the BOD will be reduced by around 30% to 40% (Gray, 2004; Metcalf & Eddy et al., 2014).

2.3.2.1. Equalization tank

The biological processes are very sensitive to the variation of flow, and they perform better under uniform flow conditions. For this reason, the equalization tank is placed at the entrance of the influent stream to dampen the diurnal flow rate fluctuations, which are beyond the capacity of the treatment plant. Due to the high hydraulic retention time, the sedimentation process may occur partially in this tank (Pons and Corriou, 2001).

2.3.2.2. Primary settling

In primary settling operation, the settleable solids and floatable materials are separated from the liquid suspension via gravity separation in a large basin. To reach the optimum gravity separation, the dispersion due to turbulence needs to be reduced; for this reason, the basin should be designed with shallow height (not less than 1.80 m) and hydraulic retention time around 0.5-2 hours. Solids that are heavier than water are collected at the bottom of the basin and drained through the sludge line for further treatment. While solids that are lighter than water, such as oil and grease, are skimmed from the top of the basin. The primary settling also can act as an equalization tank due to its high retention time (Metcalf & Eddy et al., 2014; Spellman, 2013).

2.3.3. Secondary treatments

Secondary treatment is an important part of WWTPs, which involve the biological process to reduce suspended, colloidal, and dissolved organic matter in the effluent from primary treatment. Different species of microorganisms such as heterotrophic, autotrophic, yeasts, algae, fungi,

filamentous bacteria, and protozoa are presented in this process (Gray, 2004). These microorganisms consume organic pollutants in the wastewater as the food and energy sources. They can either grow suspended in the liquid phase or on the surface of the media as a biofilm (Metcalf & Eddy et al., 2014; Spellman, 2013). In the biofilm system, the biological growth occurs on some form of media like a trickling filter. Wastewater flows around the media, where the organisms remove and oxidize the organic compounds. On the other hand, in the suspended system, the biological growth is mixed with the wastewater. Usually, the suspended growth systems are more efficient and compact compared to the trickling filter; for this reason, only more details about this system will be presented in the next section.

2.3.3.1. Activated Sludge Process

One of the most commonly used treatment processes in the wastewater industry is the activated sludge process, which mainly consists of biological reactors and settlers. Four important phenomena, including oxidation of carbonaceous wastes, oxidation of nitrogenous wastes, removal of fine solids, and removal of heavy metals, occur during this process. The transformation of biodegradable materials into new biomass, carbon dioxide, water, and residual organic matter has occurred through the growth of diverse bacteria. Normally after biological treatment, there is a settler that is utilized to separate the suspended solids and biomass from the aerated sewage and thicken the sludge before it is recycled to the reactor (Metcalf & Eddy et al., 2014; Spellman, 2013).

As mentioned earlier, activated sludge processes is a process in which microorganisms oxidize the organic matter. Thus, keeping the high concentration of the mixed culture of

organisms, in the aerated reactors is essential for this process. The microorganisms reproduce in the aerated tank and are kept suspended either by blowing air into the tank or by using agitators. The mixture of microorganisms and water, known as the mixed liquor suspended solids (MLSS), enter the secondary settler where the suspended solid is separated from the liquid phase. To maintain high concentration of microorganism in the aerated tank, part of the separated solids (sludge) in the settler is returned to the aeration tank. The rest of the sludge is transported to the sludge treatment section. Figure 2.2 shows the basic schematic of the biological process.

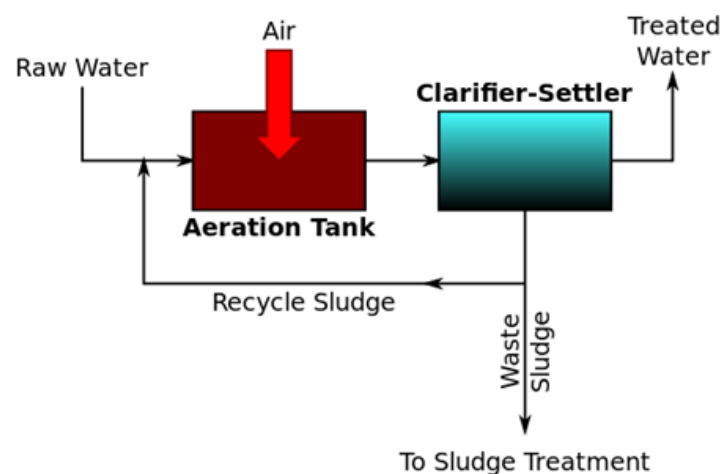


Figure 2. 2 A generalized schematic diagram of an activated sludge process (“Activated sludge - Wikiwand”).

The growth rate of biomass depends on many variables, such as the amount of biomass, the substrate, temperature, pH, and the presence of toxins.

2.3.3.2. Secondary settler

The secondary settler is the main component of the activated sludge process. The objective of this process is to thicken the biomass and separate it from water. It also removes the floating foam

and scum, which is produced in the aeration tank (Spellman, 2013). In the operation of secondary settlers, three adverse phenomena, namely bulking, rising, and dispersed sludge may frequently have occurred. The bulking sludge will happen when the sludge has poor settling characteristics and poor compatibility. The reason for sludge bulking is the growth of filamentous bacteria, which leads to sludge washout into the treated water. Rising sludge is caused by denitrification in the secondary settler. Due to the denitrification process, nitrogen gas may be trapped in the sludge layer, causing the sludge to rise. Another operational problem present in the absence of filamentous organisms is dispersed sludge, which thickens easily but gives an effluent with a high concentration of fine suspended solids (Schütze et al., 2002). Hence the excessive rise of sludge blanket, which may lead to low effluent quality, should be prevented.

2.3.4. Tertiary treatments

Tertiary treatment of effluent involves advanced treatment steps after secondary treatment to further reduce BOD, solids, and nutrients.

2.3.4.1. Filtration

The objective of filtration is to remove the possible solids which are still suspended in the effluent of secondary settler. In order to remove these compounds and reduce the turbidity of the effluent, the water can be filtered using textiles that would retain the small particles.

2.3.4.2. Disinfection

The aim of this treatment is to reduce the number of pathogens from discharge water, especially when it is used for household consumption. Usually, to disinfect water, chlorine in gas or liquid form or as sodium hypochlorite can be used. Other common disinfection methods involve the use of ozone or ultraviolet (UV) light. Generally, the efficiency of disinfection processes is directly related to the dosage of disinfectant and the contact time.

2.3.5. Sludge treatment

The main objective of this unit is to stabilize the by-product of the water treatment process (sludge), which contains a high concentration of pollutants and organic material. The other objective of this unit is to minimize the sludge disposal costs. During this process, the extracted water from sludge will be recycled back for further treatment. The treatment processes can be either chemical, biological, or physical/thermal.

2.3.5.1. Thickening

The produced sludge by secondary settler has a high content of water, thus to improve the handling of sludge, its overall volume needs to be reduced by the thickening process. The gravity thickener is a vertical tank where the phase separation is achieved by gravity and the interparticle forces inside it. During this process, the sludge loses 70-75% of its water content. The thickened sludge deposited on the bottom is then collected and sent to anaerobic AD for further process.

2.3.5.2. Anaerobic digester

AD is a multi-step biological process in which the organic solids content of the sludge decomposed into biogas, which is a mixture of methane, carbon dioxide, and trace gases such as hydrogen sulfide and hydrogen (Angelidaki et al., 2003). By applying this process, in addition to the reduction of the total mass of solids, any present pathogens in the sludge will be destroyed. Four important steps in AD have occurred are hydrolysis, fermentation, acetogenesis, and methanogenesis, where hydrolysis is subject to the fermentation process, while acetogenesis and methanogenesis are linked. In the hydrolysis step, the hydrolytic and fermentative bacteria excrete enzymes to break the complex organic materials into smaller units. The hydrolyzed products are then utilized by fermentative bacteria. The fermentative step product such as acetate, hydrogen, and carbon dioxide can directly be used by methanogenic microorganisms producing methane and carbon dioxide. Other more reduced products such as alcohols and higher volatile fatty acids are further oxidized by acetogenic bacteria in syntrophic with the methanogens. Figure 2.3 illustrates the simplified flow diagram of the biogas process.

2.3.5.3. Dewatering

After the sludge has been used in the AD process, it still has a significant amount of water, as much as 70 percent. Therefore, it is important to dry and dewater the sludge before its disposal. Dewatering the sludge can be done either by mechanical or natural treatment. Due to the slow nature of natural dewatering, mechanical processes such as centrifugation are slowly becoming one of the most preferred methods for the dewatering process. Other alternatives are the rotary drum vacuum filter and the belt filter press.

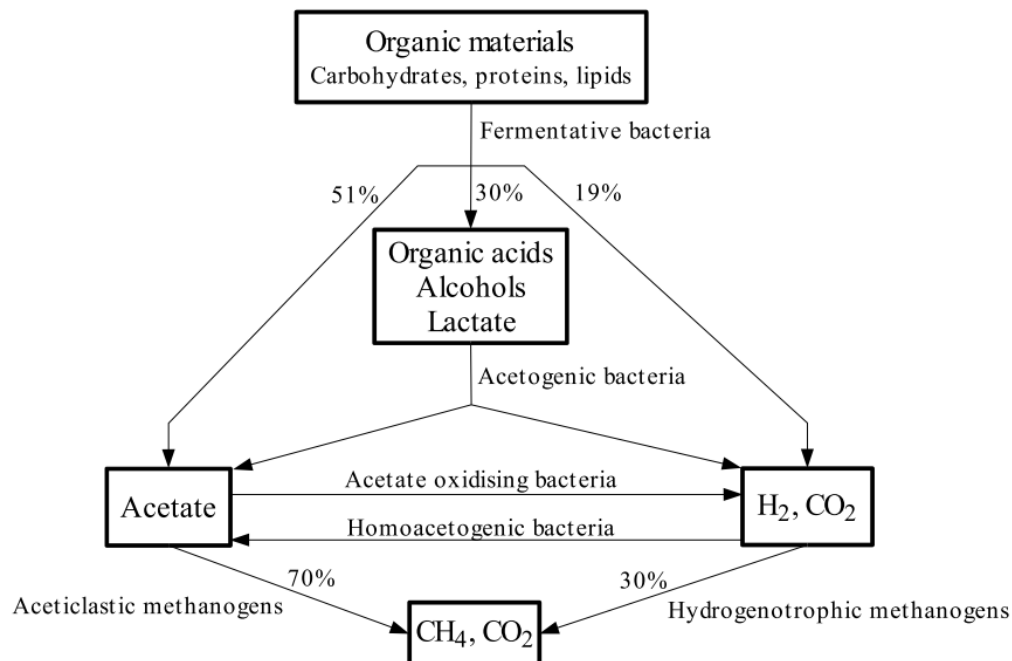


Figure 2. 3 A simplified flow diagram of the biogas process (Angelidaki et al., 2003).

2.3.5.4. Disposal

Once the sludge has been dewatered effectively, it can be either disposed into the sanitary landfill or, based on its chemical compositions, can be used as a fertilizer.

2.4. Benchmark simulation No.2 (BSM2)

This section explains BSM2, which is an extended version of the benchmark simulation model No. 1 (BSM1)(Jeppsson et al., 2006; Nopens et al., 2010). BSM2 is a well-known and powerful dynamic mathematical simulation model, able to simulate the physicochemical and biological phenomena in WWTP. It contains plant layout, a simulation model, the procedures for carrying

out the tests, the criteria for evaluating the results, and a default control strategy. The on-going research and development of BSM2 are being performed within the framework of the IWA Task Group on Benchmarking of Control Strategies for WWTPs, established in 2005 (see www.benchmarkwwtp.org). BSM2 has been under development for several years, with the preliminary concepts first introduced to a general audience at IWA's Watermatex 2004 symposium (Jeppsson et al., 2006; Vanrolleghem et al., 2010). All data for training and testing the performance of proposed soft-sensors in this thesis are obtained by using BSM2.

2.4.1. Plant layout

The BSM2 is composed of different unit operations such as primary clarifier, activated sludge biological reactor, secondary clarifier, thickener, anaerobic digester, dewatering unit, and storage tank. The schematic representation of BSM.2 is illustrated in Figure 2.4.

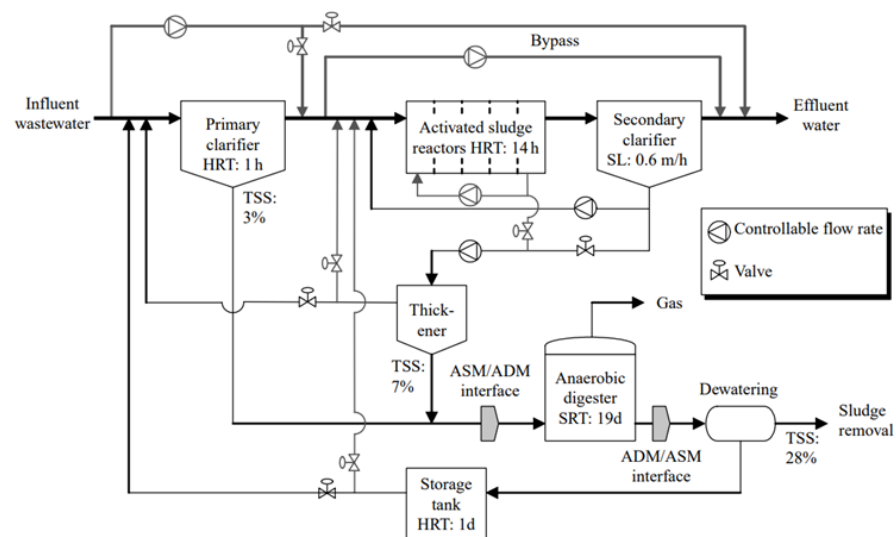


Figure 2. 4 Plant layout for BSM2 (Nopens et al., 2010).

Five reactors are included in the activated sludge process: two anoxic reactors with a total volume of 3000 m³ and three aerobic reactors with a total volume of 9000 m³, which are used for nitrification and predenitrification, respectively. The plant capacity is 20648 m³ d⁻¹ of average influent dry weather flow rate with 592 mg L⁻¹ of average biodegradable COD. The Activated Sludge Model No. 1 (ASM1) and the Anaerobic Digestion Model (ADM1) were used to describe the biological phenomena that take place in the activated sludge and AD reactor, respectively.

Moreover, to model the settling process double-exponential settling velocity function, which has the ability to model both the flocculants and the hindered settling conditions, was used (Takács et al., 1991). The influent characteristics consist of a 609 days dynamic influent data file (sampling frequency equal to a data point every 15 min) that includes rainfall and seasonal temperature variations over the year (Nopens et al., 2010; Jeppsson et al., 2006). The first 245 days serve for stabilization under dynamic conditions. The input file provided different variables. The meaning of these variables are explained in Table 2.3.

Table 2. 3 Definition of variables included in the input file of BSM2.

Definition	Notation
Soluble inert organic matter	S_I
Readily biodegradable substrate	S_S
Particulate inert organic matter	X_I
Slowly biodegradable substrate	X_S
Active heterotrophic biomass	$X_{B,H}$
Active autotrophic biomass	$X_{B,A}$
Particulate products arising from biomass decay	X_P

Oxygen	S_O
Nitrate and nitrite nitrogen	S_{NO}
$NH_4^+ + NH_3$ nitrogen	S_{NH}
Soluble, biodegradable organic nitrogen	S_{ND}
Particulate biodegradable organic nitrogen	X_{ND}
Alkalinity	S_{ALK}
Total suspended solids	TSS
Flow	Q
Temperature	T

It is recommended to run the benchmark simulation in steady-state mode first and then use that as a starting point for the dynamical simulations, in order to limit the influence of the starting condition on the dynamical process.

One important feature of BSM2 is the possibility to implement sensors and actuators that allow the testing of the custom control strategies designed (Rieger et al., 2003). Different classes of sensors that can be used in the simulation are shown in Table 2.4. The classification contains both continuous (A, B₀, C₀) and time-discrete (B₁, C₁, D) sensor models. The main parameter that describes the dynamic behavior of sensors during a step change is shown in Figure 2.5.

Table 2. 4 Sensor classes (Rieger et al., 2003)

Sensor class	T_{90}^*	Examples
A	1 min	Ion-sensitive, optical without filtration
B ₀	10 min	Gas-sensitive + fast filtration
B ₁	10 min	Photometric + fast filtration
C ₀	20 min	Gas-sensitive + slow filtration

C_1	20 min	Photometric+ slow filtration, sedimentation
D	30 min	Photometric od titrimetric for total components

* Response time

The response time is defined as the sum of the delay time T_{10} , defined as the time to reach 10% of the final value of step response and the rise (or fall) time. Together they represent the time to reach and not leave a band between 90% and 110% of the final step response:

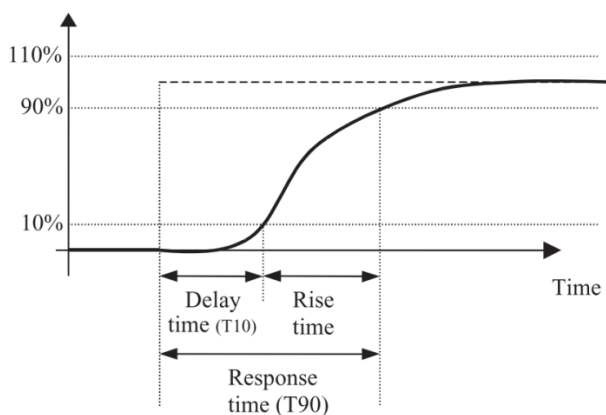


Figure 2. 5 Definition of the response time (Rieger et al., 2003).

2.4.1. Control strategy

In BSM2 to control the amount of oxygen in aerobic reactors, the proportional-integral (PI) controller is proposed (Jeppsson et al., 2007). The PI controller controls the value of dissolved oxygen (S_o) in the fourth reactor ($S_{o,4}$) at 2 mg/l by changing the value of K_{La} (The oxygen transfer coefficient) in the third reactor (K_{La3}), K_{La} in the fourth reactor (K_{La4}) and K_{La5} . It should be noted that in this configuration, the value of K_{La5} is half of the value of K_{La3} and K_{La4} . The schematic diagram of the explained controller configuration is shown in Figure 2.6.

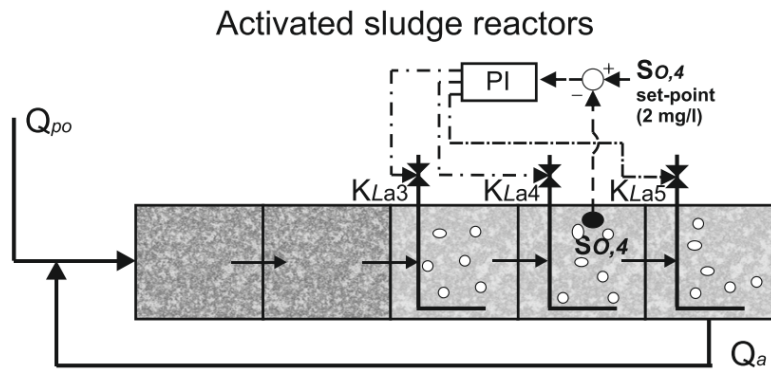


Figure 2. 6 Schematic of the control strategy for aerobic reactors in BSM2.

CHAPTER 3

Data-driven soft-sensors

Soft-sensors are a computer program that can be used in a similar way as their hardware counterparts. The core of a soft-sensor is a model that processes information produced typically by hardware instruments. Based on their core model, soft-sensors can be classified into two different classes, namely model-driven and data-driven. Model-driven soft-sensors are made based on the mathematical models of phenomena that happen in the processes, while data-driven soft-sensors are made based on the collected historical data of processes.

3.1. Introduction to soft-sensors

Nowadays, industrial processes are heavily equipped with various types of sensors. Previously, the collected data from these sensors could only be used for process monitoring and control. However, with the emerging “big data” technology, researchers started using collected data to build predictive models. In terms of the process industry, these predictive models are called *soft-sensors* (Kadlec et al., 2009). Soft-sensors can be divided into two main categories of the model-driven and data-driven type (see Figure 3.1).

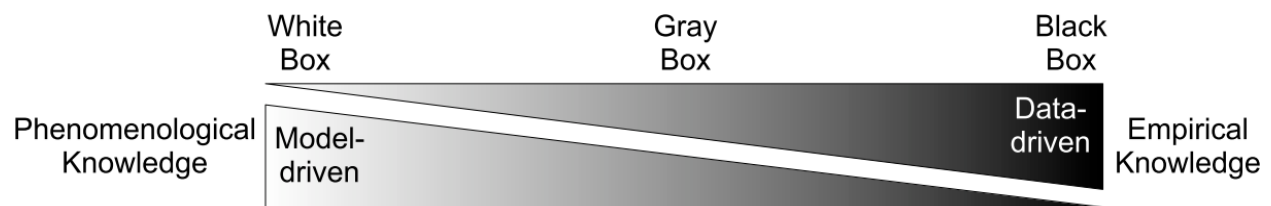


Figure 3. 1 Model vs. data-driven soft sensors (Fortuna et al., 2007).

Model-driven soft-sensors are developed based on the physical and chemical phenomena of the process and mainly used for the planning and design of the processing plants. As these kinds of models are developed under the ideal states of the processes, their practical usage as soft-sensors may be associated with the problem. Due to this reason, data-driven soft-sensors increasingly gained in popularity.

A data-driven soft-sensors is an input-output model in which the model inputs consist of variables that can be easily measured at a reasonable cost. The output of the model consists of information about the variables that their measurement is not easy and associated with incurring high costs. Figure 3.2 shows the schematic of a soft-sensor with six inputs and one output.

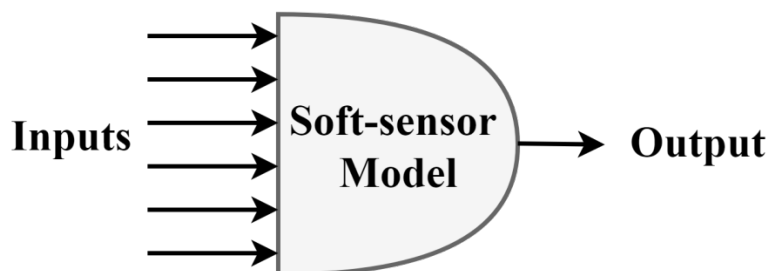


Figure 3. 2 Schematic of simple soft-sensor.

Different methods such as PCA, Principal Component Regression (PCR), Partial Least Squares (PLS), ANN, Neuro-Fuzzy System (NFS) and SVM have been used for designing data-driven soft sensors (Bishop, 1995; Cortes and Vapnik, 1995; Jang et al., 2005; Jolliffe, 2011; Wold et al., 2001).

Soft-sensors can be used for different applications. They are mainly applied for the *prediction of variables* that their estimation is associated with high sampling time or through off-line analysis only. From the process viewpoint, estimation of these variables is essential because they are often shown the process quality or other critical aspects of the process. Therefore, it would be very beneficial to estimate them with lower sampling time and lower financial costs. The supervised methods can be used for this task, and their application field is further referred to as *online prediction*. The other important application of soft-sensors is *monitoring and fault detection*. By using soft-sensors, the state of the process and the possible deviation from the normal operation can be detected. Traditionally, the state of processes was monitored by the control room operators using the univariate process control charts, and based on their experiences, the decision about process condition could be made. By using soft-sensors, the process state can be predicted, and it allows the operators to make faster, better, and more objective decisions. Two of the main methods

commonly used in developing process monitoring soft-sensors are PCA and self-organized maps (SOMs) (Kohonen, 2001).

Another application of soft-sensors is sensor failure detection. Generally, this application field can be described as *sensor fault detection and reconstruction*. Once the fault in the sensors network is detected and identified, it can be reconstructed, or it can be replaced by the soft-sensor, which is already trained as a backup of a hardware sensor. Figure 3.3 shows the application field of data-driven soft-sensors and the potential techniques that can be applied for each application.

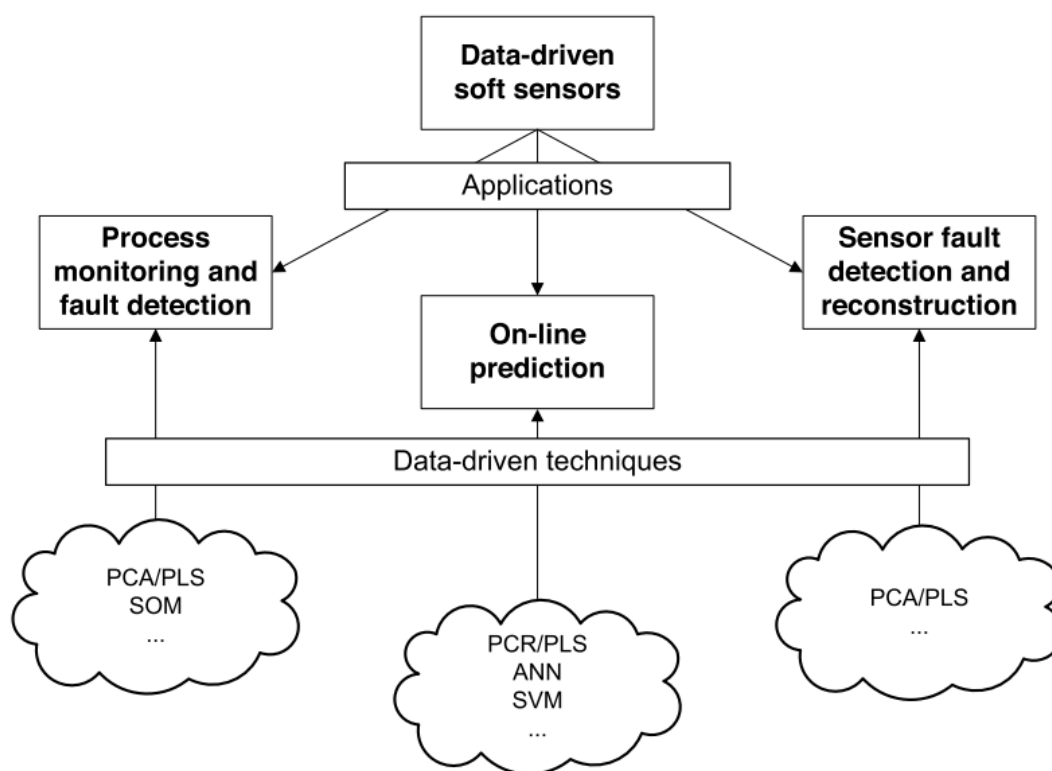


Figure 3. 3 Application fields of soft-sensors (Kadlec, 2009)

3.2. Designing data-derived soft-sensors

In this section, the soft-sensor design framework, and the common methods which are used in each step of framework development will be briefly discussed. An overview of the soft-sensors design procedure is shown in Figure 3.4. The designing procedure consists of different stages, including data acquisition, data pre-processing, model design, and model maintenance (Haimi et al., 2013; Kadlec et al., 2009).

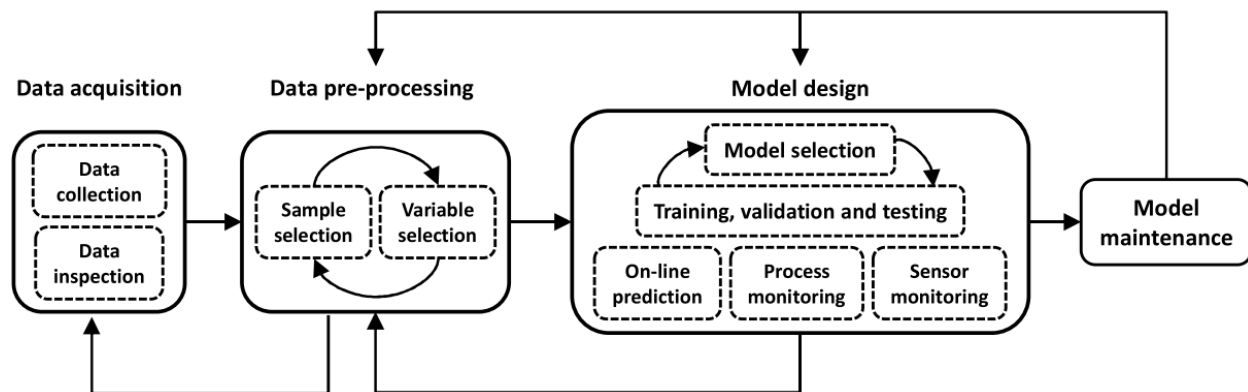


Figure 3. 4 Overview of steps needed for developing data-driven soft-sensors (Haimi et al., 2013; Kadlec, 2009)

3.2.1. Data acquisition

Nowadays, in the most modern process plants, the historical data are routinely stored in the data acquisition system. The first step in designing a data-driven soft-sensor is data collection and, subsequently, data inspection. During this step, the initial inspection of data is performed in order to understand the data structure and to recognize the possible problems within the data (e.g., locked measurements, missing and drifting data, and measurements outside the operating range of the

instruments). The choice of model complexity is also assessed during this step. As the data inspection is commonly performed manually, much time and expertise regarding the process under study are needed.

3.2.2. Data pre-processing

The data from process plants are usually associated with noise, outliers, and missing values. Apart from this, the high dimensionality of data can be considered as one of the limitations for developing the soft-sensors. Thus it is mandatory to prepare the data prior to any model development. To pre-processing the data, prior knowledge about the process is needed. Such knowledge can be used in combination with statistical methods for *variable selection* and *outlier detection* to improve the pre-processing step (Kadlec et al., 2009; Slišković et al., 2011).

3.2.2.1. Variable selection

One of the crucial steps that need to be done prior to data-driven soft-sensors development is input variable selection, usually accomplished by using both the plant expert's knowledge and suitable mathematical strategies. The aim of variable selection step is to select the most informative input variables for the soft-sensor. This allows to obtain a reliable estimation of the output with a low number of independent variables. It also reduces the model complexity, improves model accuracy, and decreases the measuring costs by lowering the number of hardware sensors needed to acquire the input variables (Curreri et al., 2020; Xibilia et al., 2017). The methods for variable selection can be classified into two main categories Variable Extraction (VE) and Variable Selection (VS).

By applying VE methods, the new variables are created based on transformations or combinations of the original input variables. The transformed variables don't have physical meaning and can not be interpreted. The common method which is widely used for this task is PCA (Xibilia et al., 2017).

By applying VS methods, the best subset from the original variables set is selected without performing any further transformation. The methods for this task are mainly supervised methods; therefore, they require processing both input and output data. The model structure may or may not be involved in the investigation. VS methods can be classified as:

- **Filters:** in this method, a subset of input variables is selected by evaluating the relation between input and output of the considered system. The subset of input variables is obtained independently of the machine learning method that is used to build the model. The main advantage of filter approach is the low computational complexity ensuring speed to the model. One of the famous filter methods is Pearson's correlation coefficient, which is a measure of linear correlation. Filter methods are mainly used as the first step in hybrid approaches (Cateni et al., 2013; Curreri et al., 2020).

- **Wrappers:** Wrapper approaches considered the machine learning methods as a black-box in order to select the best subsets of input variables based on the prediction performance of the given machine learning method. Different performance criteria such as the Mean Squared Error (MSE), Akaike's Information Criterion (AIC), or CP (Mallow's Coefficient) statistics can be used (Curreri et al., 2020). The wrapper approach is more computationally intensive and slower compared to the filter approach.

- **Embedded (model-based):** Unlike previous methods, embedded approach performs the variable selection in the machine learning method. A specific characteristic of the model or of its learning process is used to define the criterion. These methods are slower than filters and give low-performance results when not enough data is available (Cateni et al., 2013).
- **Hybrid:** Combining different methods often improves the results of the input selection procedure. More information regarding hybrid methods can be found in (Souza et al., 2016).

3.2.2.2. Outlier detection

During the analysis of process plant data, it is common to face observations that are not consistent with the majority. Such observations are often called outliers. The presence of outliers in the data set is resulting from hardware failure, incorrect readings from instrumentation, transmission problems, 'strange' process working conditions, etc. There are two types of outliers, namely obvious outliers and non-obvious outliers (Kadlec et al., 2009). Obvious outliers are those measurements that fall outside physical or technological limitations. For example, the absolute pressure can not be negative, or the flow sensor may not deliver values that exceed the technological limitations of the sensor.

In contrast, non-obvious outliers are the observations that do not violate any limitations but still fall outside of the typical ranges and do not reflect the correct variable states (Haimi et al., 2013; Kadlec et al., 2009). Outlier detection is a part data pre-processing step, and it is essential for designing soft-sensors because undetected outliers may decrease the performance of the models. One of the famous outlier detection methods is 3σ , which is based on the observations of the variable distributions. In this method, the sample data that fall outside of the $\mu(x) \pm 3\sigma(x)$,

where $\mu(x)$ is the mean value and $\sigma(x)$ the standard deviation of the variable x , are considered as outliers (Lin et al., 2007). Multivariate approaches are another method that uses a combination of variables to detect the outliers. An example of these approaches is a combination of PCA and Hotelling's T^2 measure. In the first step, PCA is applied in order to identify the main disturbances. Then each identified disturbance is analyzed by Scheffé's Test in order to detect data outliers using statistical criteria (Gonzalez et al., 2003).

3.2.2.3. Model selection, training and validation

As the model is a major part of soft-sensors, searching for the best type of model which leads to higher performance of soft-sensors is necessary. Despite the significant progress in the field of soft-sensor development still, there is no straight forward approach for finding the model type and its hyperparameters. This is because the choice of model depends on its application, and it is often subjected to the developer's experience and personal preference (Haimi et al., 2013; Kadlec et al., 2011). The best practice is to use simpler models (e.g. linear regression model) in the initial stage of soft-sensor development and gradually increase model complexity as long as a significant improvement in the model's performance can be observed. During this task, the performance of the model should be assessed by independent data (Hastie et al., 2009). Most of the data-driven models have several basic parameters and several meta-parameters that need to be tuned in order to increase the generalization performance of the model. The regression coefficients of linear regression methods and the connection weights of the neural network are two examples of basic parameters.

On the other hand, the number of neurons and layers in the neural network or number of retained principal components in the PCA method are considered as meta-parameters (Haimi et al., 2013). Usually, the basic parameters of the models are tuned during the training stage. In contrast, the meta-parameters needs to be tuned by the model developers. Different methods can be used for meta-parameters tuning; however, *grid search* and *random search* are two widely used methods for this task.

- **Grid search:** this approach is simply an extensive search through a manually specified subset of the meta-parameters space of a learning algorithm. In order to find out the best subset of meta-parameters, the grid search must be guided by some metric, typically measured by cross-validation on the training set or evaluation on a held-out validation set (Chicco, 2017). In K-fold cross-validation, a part of data is used to calibrate the model, and the other part of data is used for validation. First, the data is split to K roughly equal-size parts, and then the k th part should be set aside, and the model is calibrated to the other $K-1$ parts and, finally, the models' accuracy over the k th part is calculated. After repeating the procedure for all K parts, the overall accuracy can be obtained by averaging all K parts for the specific set of meta-parameters. The model with the highest overall accuracy is finally trained on the whole data and assessed against a testing data set. Although cross-validation is a common approach for meta-parameters optimization, alternative methods such as bagging (Breiman, 1996) and boosting (Freund and Schapire, 1997) can be used for this task.

- **Random search:** random search is a technique where the random combinations of meta-parameters are used to find the best solution for model development. It is very similar to grid search, but it has a better performance compared to grid search. As the combination of meta-

parameters is chosen randomly, the chance of finding the optimal parameters is comparatively higher in a random search (Bergstra et al., 2012).

After finding the optimal meta-parameters of the model and training the model, the trained model must be evaluated on independent data once again. There are several metrics for measuring the performance of models, in case of a regression problem MSE, which measures the average square distance between the predicted and the correct value can be used. Other approaches, such as the visual representation of the predictions, are also used for this task. By using this approach, useful information about the relation between the predictions and the correct values, together with the analysis of the prediction residuals, can be obtained (Fortuna et al., 2007).

3.2.2.3. Soft-sensor maintenance

One of the main difficulties in the application of soft-sensors is model performance degradation. The predictive performance of soft-sensors tends to decrease slowly after sometimes due to several reasons, including changes in the state of the process and instrumental characteristics or operating conditions (Kaneko and Funatsu, 2014). In wastewater treatment applications, the reason for this may be, e.g., variations in influent wastewater composition, temperature and flow rate, instrument recalibrations, or operational changes inside the plant. To solve this issue, the soft-sensors should be maintained and updated regularly. The soft-sensors update can be performed automatically or by its developer. To automatically update the soft-sensors variety of approaches have been proposed. The majority of these methods are based on adaptive versions of the multivariate statistical methods like PCA and PLS (Haimi et al., 2013; Kadlec et al., 2011; Kaneko and Funatsu, 2014).

3.3. Applications of soft-sensors

In this section, the main applications of soft-sensors will be discussed. Generally, in WWTPs, data-driven soft-sensors can be used for online prediction of the primary process variables or monitoring and FD. The following sections also list the recent examples of these two most common application types of soft-sensors across the wastewater treatment process.

3.3.1. Online prediction

The most common application of soft-sensors in WWTPs is the prediction of variables that can not be measured directly, or their measurement is associated with high cost due to maintenance and sensors price. Soft-sensors are often used for the prediction of critical variables, which represents the final product quality. Soft-sensors are very useful when the dynamic of the process under study can not be easily defined using the first principle approach, and there is no possibility to collect the necessary information online. For this task, different supervised data-driven methods can be used.

3.3.1.1. Transfer function models

In the transfer function model, the relationship between input and output variables of the linear system is described using a mathematical function. The transfer function will be the right choice for the dynamic systems when the number of output parameters is just one or two (Box et al., 2008). One of the most used methods within this modeling approach is univariate autoregressive integrated moving average (ARIMA) models. This model is a particular case of transfer function

models that do not depend on the input variables and commonly used for linear time series forecasting.

In WWTPs, the ARIMA model is mainly used to predict the effluent parameters. Park and Koo (2015) used the transfer function ARIMA model for the prediction of turbidity on sedimentation reservoir outflow. They used the turbidity of raw water, pH, alkalinity, flow rate, and coagulant as inputs of their model. By using these inputs, the coefficients of determination of the predicted model were obtained higher than 0.95 (Park and Koo, 2015). West et al. (2002) developed a transfer function model that can monitor output biochemical oxygen demand from a wastewater treatment process (West et al., 2002). As the WWTPs' behavior is highly nonstationary, the performance of the ARIMA model may be deteriorated after long-term usage.

3.3.1.2. Multiple regression

Multiple linear regression is the most common form of linear regression analysis in which a dependent variable is modeled as a function of several independent variables. Ordinary least squares are commonly used to estimate the model parameters (Sheather, 2006).

Ebrahimi et al. (2017) used multiple regression models to predict different quality parameters such as BOD, Total Phosphorus (TP), and Wastewater Quality Index (WWQI) for real WWTP. Their model showed high levels of statistical significance in addition to admissible accuracy in terms of fitting with the training data parameters, with 81.8% average accuracy, and validating with the testing dataset, with an average relative prediction error of 2.9% (Ebrahimi et al., 2017). In other work, Al Bazedí and Abdel-Fatah (2020) applied multiple regression models to chemically enhanced primary treatment (CEPT). The modeling was undertaken using the simulation of the

data obtained from pilot plant experimental studies using different types of coagulants (FeCl₃, alum, lime, and Magna-floc155). The obtained results showed that the empirical model could predict removal efficiencies with $R^2 = 0.973$, and 0.978 for COD and TSS (Al Bazed and Abdel-Fatah, 2020).

PLS regression, which is a combination of PLS and multiple regression is commonly used for developing industrial soft-sensors to predict water quality variables, such as COD, TSS, nitrate, and oil and grease concentrations. Lourenço et al. developed and published a soft-sensor based on PLS. The soft-sensor uses UV spectra of water samples collected at the outlet of a fuel park WWTP as inputs to estimate total organic carbon (TOC). The root mean squared error (RMSE) of cross-validation values for the developed PLS model based on the row sample was 2.3 mg Cl^{-1} , and the RMSE of validation was 1.8 mg Cl^{-1} (Lourenço et al., 2008). The use of non-linear PLS (kernel PLS or KPLS) shows significant improvement of this approach in simulating the behavior of non-linear systems such as an anaerobic filter and conventional activated sludge (D. S. Lee et al., 2006; Woo et al., 2009).

3.3.1.3. Artificial neural networks (ANNs)

ANNs are developed based on the operation of biological neurons, which are the primary information processing units in nervous systems. Both the biological and artificial neurons get the information as inputs and process them as outputs. A general discussion about the theory of ANN, learning algorithms, application areas, etc. is given in Bishop (Bishop, 1995). Among different variations of ANNs, the feed-forward networks, like the multi-layer perceptron (MLP) (Bishop,

1995) and the radial basis function network (RBFN) (Poggio and Girosi, 1990) are widely used for developing soft-sensors for industrial processes.

One of the variations of ANNs is recursive neural networks (RNNs). RNNs are very similar to feed-forward networks; the only difference is the feedback connection, which enhanced the network capability to extract and learn temporal sequences from the data (Mandic and Chambers, 2001).

Another variation of ANNs is the self-organizing map (SOM) or Kohonen map, which can deal with unsupervised problems and thus is an excellent choice for process monitoring tasks. It provides a topology preserving mapping from the high dimensional space to map units. Map units, or neurons, usually form a two-dimensional lattice, and thus the mapping is a mapping from high dimensional space onto a plane (Kohonen, 2001).

The performance of two ANN variants, namely the RBFN and MLP, are compared to a Support Vector Regression (SVR) model in Najafzadeh and Zeinolabedini (Najafzadeh and Zeinolabedini, 2019). The data set for the comparison is based on the real WWTP located in Kerman, Iran. It is clearly shown that the performance of the SVR soft-sensor is superior in comparison to the other two methods. In Bekkari and Zeddouri (2019), the ANN is applied to real WWTP in Touggourt, Algeria, to predict effluent COD (Bekkari and Zeddouri, 2019). The influent variables such as pH temperature, suspended solids, Kjeldahl Nitrogen, BOD, and COD were used as input variables of ANN. The results showed that the ANN model could predict the experimental results with a high correlation coefficient of 0.89, 0.96, and 0.87 for learning, validation, and testing phases, respectively. In other work, Pisa et al. (2018) used RNN to predict effluent total nitrogen and the ammonium concentrations in order to reduce possible violations (Pisa et al., 2018). The adopted input and output data are generated through the usage of the BSM2 model.

The inputs for the RNN model were as follows: ammonium concentration of the primary clarifier, overflow rate of the primary clarifier, multiplication of ammonium concentration of the primary clarifier and overflow rate of the primary clarifier, recycle flow rate and temperature. The mean absolute percentage error (MAPE) for the prediction of ammonia and total nitrogen in the effluent were around 8% and 5%, respectively. Another soft-sensor for rapid prediction of effluent BOD has been developed by Rustum et al. (2008) using the SOM method (Rustum et al., 2008). They trained three different SOM models, each SOM model having different input variables. The inputs to the models were the variables inflow, COD, suspended solids, ammonia nitrogen, pH, and temperature. The authors conclude that SOM is capable of achieving a satisfactory performance for the prediction of BOD.

3.3.1.4. Support vector machines (SVMs)

Recently SVMs gained considerable attention for application in the process industry due to their theoretical background in statistical learning. The SVMs theories can be found in (Cortes and Vapnik, 1995). SVMs have been demonstrated to work very well for a broad spectrum of applications, so it is not surprising that they have also been successfully applied as soft-sensors. While SVMs used widely in designing various soft-sensors, there is still much work needed when dealing with huge data sets for which the computational complexity of the SVM training process can be prohibitive (Kadlec et al., 2011).

In a detailed study by Yasmin et al. (2019), SVM and MLP models were used for estimation of COD, total nitrogen, ammonia nitrogen, and total phosphorus in an aerobic granular sludge (Yasmin et al., 2019). The simulation was done using the experimental data obtained from the

sequencing batch reactor under a hot temperature of 50°C. The results indicated that the performance of the SVM approach was superior for the considered task compared with the MLP model.

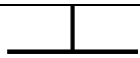



More recently, Zaghoul et al. (2020) investigated the capabilities of two artificial intelligence algorithms for the development of predictive models for aerobic granular sludge-based on influent characteristics and operational conditions using adaptive neuro-fuzzy inference system (ANFIS), and SVM (Zaghoul et al., 2020). The model presented in this work is a two-stage prediction process in which the MLSS, mixed liquor volatile suspended solids (MLVSS), sludge volume index at 5 min (SVI₅), sludge volume index at 30 min (SVI₃₀) and granule size (D) are predicted in the first stage of the model, then the effluent COD, ammonia nitrogen, and phosphates are predicted in the second stage. The inputs to the first sub-model were influent ammonia and phosphates, influent pH, organic loading rate, hydraulic retention time, superficial air velocity, water temperature in the reactor, and settling time. It was shown that ANFIS needs a more computational resource due to the large rule base, while SVR was able to achieve high accuracies due to the penalty placed on the prediction errors. SVR proved to have higher prediction accuracy than ANFIS, but ANFIS had a better generalization ability.

3.3.2. Monitoring and fault detection

Another application of soft-sensor in WWTPs is process monitoring. The faults are the events that cause irregularities in the performance of WWTPs. Due to the nature of the different types of faults, it is essential to consider the versatility of an analytical approach (i.e., which types of faults can be detected) when designing a FD framework. For instance, if the fault occurs in a sensor that

is included in a control loop, many variables could be affected. In contrast, the fault in a sensor that is not included in the control loop may only affect the measured sensor variable (Haimi et al., 2013; Kadlec et al., 2011). Typical univariable faults that may occur in WWTPs are illustrated in Table 3.1. For FD and monitoring, soft-sensors can be designed using either unsupervised or supervised methods. Conventional methods such as PCA, PLS, and SPC widely used in WWTPs to detect faults. The SPC is a method that uses statistical approaches to monitor and control processes.

Table 3. 1 Abnormal patterns in univariate WWTP data that could indicate a fault and potential causes of the fault pattern (Capizzi and Masarotto, 2017).

Pattern	Cause
Isolated 	Power spike, air bubble on the sensor, spike of the contaminant in influent
Sustained 	Change in operational status, mechanical performance variation, sensor recalibration
Transient 	State change, sensor malfunction, cycle fluctuation
Drift 	Sensor or mechanical device degradation, biological shift, fluid flow restriction

3.3.2.1. Control charts

One of the essential tools used in statistical process control (SPC) is control charts. Control charts are useful to determine if a manufacturing process is in a state of control. The Shewhart control chart is a widely used control chart, which is introduced by Walter Shewhart of Bell Labs. This control chart uses upper and lower control limits (UCL and LCL) for a process variable or statistic by adding or subtracting k standard deviations from the variable's mean. The fault is detected once an observation pass either lower or upper control limits. The Shewhart control chart

can only be applied to variables that have a normal distribution and whose observations are independent and stationary (Montgomery, Douglas, 2009; Shewhart, 1926). The exponentially weighted moving average (EWMA) is a control chart that can handle nonstationary data by updating the UCL and LCL (Wold, 1994). The EWMA chart gives less weight to previous observations and more weight to recent observations; for this reason, it is frequently used for data smoothing (Mina and Verde, 2006). Another control chart that is widely used in the industry is the CUSUM; it shows the accumulation of information on current and previous observations. For this reason, the CUSUM chart is useful in the detection of small and moderate shifts in the mean of process observations (Abujiya et al., 2015).

Wąsik et al. (2017) studied the possibility of the use of Shewhart control charts to monitor changes in the forms of nitrogen for the wastewater treatment plant in Krosno. By using statistical analysis, it was shown that the highest number of cases of elevated nitrate-nitrogen and/or ammonium nitrogen have occurred when the temperature of treated wastewater was below 8-9°C (Wąsik et al., 2017).

The performance of five univariate FD methods (Shewhart, EWMA, and residuals of EWMA) applied to the simulated results of the Benchmark Simulation Model No. 1 long-term (BSM1_LT) were compared by Corominas et al. (Corominas et al., 2011). In order to check the ability of each method in detecting faults, different faults in dissolved oxygen (DO) sensor, including shift, drift, fixedvalue, complete failure, wrong gain, and calibration, have been simulated. The results clearly showed the better performance of adaptive methods (residuals of EWMA) in detecting a sensor measurement shift and when monitoring the actuator signals in a control loop (e.g., airflow). Spindler and Vanrolleghem applied CUSUM charts to continuous mass balancing to detect off-balance periods. They showed that CUSUM is a suitable and more reliable approach than mass

balances based on long term averages of data. Continuous mass balancing following this method requires individual balance equations that describe the redundancy of the measured data (Spindler, 2014; Spindler and Vanrolleghem, 2012).

3.3.2.2. Principal component analysis (PCA)

PCA is one of the widely used statistical methods for FD and monitoring. In this method, the number of variables reduced by building linear combinations (principal components or PCs) of them in such a way that these combinations cover the highest possible variance in the input space. For FD, PCA should be trained (calculating the PCs) by the training data set that represents the normal operation of the process, then the new data or testing data are transformed into the model subspace (defined by the PC) (Haimi et al., 2013; Kadlec et al., 2011). By calculating squared prediction error (SPE) and Hotelling's T^2 statistics, the overall distance of new observation to the PCA-model will be estimated. If the overall distance was higher than the desired threshold, the observation is considered as a possible fault. Various modifications of PCA, including adaptive (moving window and recursive), dynamic and kernel PCA, have been applied widely for FD in dynamic and nonstationary processes such as WWTPs.

Baggiani and Marsili-Libelli (2009) described the development of real-time FD and isolation system based on an adaptive PCA algorithm. They used the data sampled at 1 min intervals from three nitrogen sensors in a conventional pre-denitrifying wastewater treatment plant with a capacity of 88,600 PE, during a period of nine months. The performance of developed adaptive PCA was assessed by organizing the sequential data in two differing moving windows: a short-horizon window to test the response to single malfunctions and a longer time-horizon to simulate multiple unrepaired failures. In both cases, the algorithm performance was very satisfactory, with

a 100% failure detection in the short window case, which decreased to 84% in the long window setting (Baggiani and Marsili-Libelli, 2009). Kazor et al. (2019) published an extensive discussion about the comparison of monitoring performances of static, dynamic, adaptive, and adaptive–dynamic versions of PCA, KPCA, and locally linear embedding (LLE). These methods are applied to real data collected from a membrane bioreactor (MBR) located at the Mines Park Water Reclamation Test Site in Golden, CO (Kazor et al., 2016). They also evaluated a nonparametric estimation of thresholds for monitoring statistics and compared results with the standard parametric approaches. The monitoring of the MBR was conducted based on 28 quantitative variables that are consistently measured. From their simulation study, it is clear that the adaptive–dynamic versions of all three methods improved results when applied to the autocorrelated and nonstationary MBR process. They also concluded that the false alarm rates would be reduced once the nonparametric thresholds are applied. Haimi et al. (2016) investigated a PCA anomaly detection system for a large-scale municipal WWTP located in Helsinki, Finland. They employed two methodologies, including moving-window PCA extensions with adaptive and fixed window-lengths. The experimental results showed that the monitoring systems with the adequate sets of parameters could detect, drifts, and peaks in measurements, as well as process anomalies. Moreover, the correct isolation of the abnormal variables is demonstrated (Haimi et al., 2016).

3.3.2.3. Partial least squares (PLS)

Similar to PCA, PLS estimates the independent linear combinations of the measured variables. Unlike PCA in PLS, the inputs are maximally linearly related to the output variable. PLS is considered as supervised techniques, and it just monitors the output variables that are

affected by the input variables. In PLS, the fault can be detected if an abnormal observation has relativity to the output variables (Chen et al., 2016). PLS is a prevalent technique in chemical engineering; several modifications of PLS such as multi-way PLS, Neural Network PLS (NNPLS), recursive PLS (RPLS), multiblock partial least squares (MBPLS) and exponentially weighted PLS (EWPLS) have been used for FD (Bro, 1996; Dayal and MacGregor, 1997; Qin, 1998; Qin and McAvoy, 1992).

Chen et al. (2016) applied PCA and PLS techniques for FD in WWTP (Chen et al., 2016). The performance of proposed FD methods is estimated using simulated results of the BSM1. Both methods could detect the simulated leakage fault, although with some delays.

Choi and Lee (2005) investigated the application of the MBPLS method to analyze and diagnose of a WWTP in a steel mill plant. The measured process variables from WWTP are partition into several blocks for multiblock analysis. They divided all variables into three blocks, each of which is associated with influents, an equilibrium tank and recycling, and aeration tanks and settlers. Moreover, the effluent COD is modeled based on the collected data and predicted online. Monitoring of the process is performed based on four kinds of statistics (T^2 and Q for block, and T^2 and Q for the whole process). The contribution to the four monitoring statistics is also estimated. It is shown that the proposed method based on the newly defined variable and block contribution could can and isolate the faults precisely (Choi and Lee, 2005).

3.3.2.4. Artificial neural networks (ANNs)

ANNs can also be used in the context of FD and monitoring. This can be done either by supervised or unsupervised training of ANN. Supervised training can be done by labeled data in

such a way that inputs and outputs are defined. Whereas, unsupervised training of the ANN uses data that are unlabelled. Autoencoder ANN is one of the types of neural networks that can be used as an unsupervised method for FD and monitoring. It can be trained to model a process by estimating the values of inputs and comparing the estimation to the actual values. The other type of unsupervised ANN, which is discussed in the previous section, is SOM (Newhart et al., 2019; Xiao et al., 2017).

Miron et al. (2018) developed a feed-forward neural network classifier for FD in WWTP. The inputs to their model were 90 variables; including nine variables (dilution rate, aeration rate, recirculating rate, influent substrate concentration, the rate of the sludge in excess, biomass concentration, substrate concentration, dissolved oxygen concentration, and recirculated biomass concentration) and the corresponding lags equal to 10 for each variable. The output of the ANN classifier is consists of seven classes (one class represents the normal operation state and six classes correspond to the six types of faults). The architecture of the developed ANN is straightforward, including one hidden layer with ten neurons. By using this model, they obtained the overall correct recognition rate of 97.2% and the rate of false classification (2.8%) (Miron et al., 2018). Chi and Guo (2019) proposed a sensor fault diagnosis method based on interval prediction, which using radial basis function (RBF) neural network with set membership estimation. The influent BOD, COD, and TSS are selected as the input of the prediction model, and the effluent COD and TSS are used as the output. They developed two interval fault diagnosis models in order to further determine that the sensor has failed. When one of the detection models detects that the corresponding effluent water quality exceeds the predicted interval, it proves that the detection sensor has failed, when both signals are exceeded, it is proved that the system has failed. The simulation results demonstrate that the proposed sensor fault diagnosis method is

practical and useful (Chi and Guo, 2019). Xiao et al. (2017) investigated the application of auto-associative ANN with the shallow and deep structure for fault diagnosis in WWTP. The developed model is applied to BSM1 and real WWTP. The proposed methodology provides a recursive minimization strategy to deal with missing values; it also offers kernel density estimation (KDE) to calculate the confidence interval. In order to detect the faults in the early stage of their occurrence, the multi-step ARMA model has been used to predict the sum of squared residuals (SPE) over a long horizon. The results showed that the proposed methodology is capable of detecting sensor faults and process faults with reasonable accuracy (Xiao et al., 2017).

3.3.2.5. Support vector machines (SVMs)

Like ANN, SVM also can be trained either supervised or unsupervised for FD. Cheng et al. present (2019) a FD framework based on SVM classification. In their proposed framework, first, the dimension of data is reduced using forecastable component analysis (ForeCA) then the quadratic Grid Search (GS) algorithm is utilized to optimize the meta-parameters of the SVM. The performance of the proposed method is validated by BSM1 and simulated oxygen sensor fault (Cheng et al., 2019). In other work, Zeng et al. (2006) applied SVM classifier for FD in real WWTP located in China. Due to the unbalanced distribution of classes in the training data set, the risk function with the weight coefficient based on leave-one-out errors is used to improve SVM classification performance. Furthermore, the Genetic Algorithm (GA) is applied to optimize the risk function globally. The improved SVM indicates high classification accuracy (Zeng et al., 2006).

CHAPTER 4

Data-Driven Soft Sensors for Online Monitoring of Volatile Fatty Acids in Anaerobic Digestion Processes

The concentration of VFAs is one of the most important measurements for evaluating the performance of AD processes. In real-time applications, VFAs can be measured by dedicated sensors, which are still currently expensive and very sensitive to harsh environmental conditions. Moreover, sensors usually have a delay that is undesirable for real-time monitoring. Due to these problems, data-driven soft sensors are very attractive alternatives. This chapter proposes different data-driven methods for estimating reliable VFA values. We evaluated RF, ANN, ELM, SVM and GP based on synthetic data obtained from the IWA BSM2.

4.1. Introduction

AD is a well-established process for stabilizing municipal sewage sludge and treating organic waste products and wastewaters from different industries, households and farms. In this process, the organic matter biodegrades in an oxygen-free environment. This leads to decomposition and bioconversion of organic matter into biogas, which mainly consist of CH_4 (50–60%) and carbon dioxide (30%–40%), with some other trace gases such as hydrogen sulfide and water vapors, etc. (Abu Qdais et al., 2010). The methane gas produced can be used as an energy source for generating electricity and heating the AD reactor. The benefits of the AD process are high organic load treatment, low sludge production, energy is recovered when the biogas produced is used and operating costs are reduced due to the oxygen-free operation (Yordanova et al., 2005). Many operational parameters affect the performance and effluent quality of the AD process; therefore, frequent monitoring of these parameters is crucial to ensure a stable performance. Among these parameters, pH, partial alkalinity and VFAs are the most important measurements for monitoring the stability and performance of the digesters with low buffering capacity. However, in the highly buffered digester, although the process is extremely under-stressed, the pH may vary very little; in this case, VFAs are the only reliable measurement for process monitoring. VFAs are key intermediate products for the reactions that produce CH_4 , while their accumulation inside the reactor inhibits the bacteria and causes a lower methane production rate so that the process fails (Franke-Whittle et al., 2014). The common VFA monitoring approach in a wastewater treatment plant is online and/or offline analysis. However, the measuring procedure is very costly and characterized by time-delayed responses that are often undesirable for real-time monitoring (Haimi et al., 2013). Moreover, due to the complex media and severe operating conditions in AD processes, the online sensors are not always reliable. This is because solid deposition, slime build-

up and precipitation, among others, mean that sensors require regular maintenance and calibration. Therefore, it is imperative to develop cost-effective measuring techniques to provide the necessary information based on easily measured available variables and without installing new instruments (Dürrenmatt and Gujer, 2012). Thanks to recent progress in measurement and instrumentation technologies, many easy to measure parameters, such as pH, temperature, flow rates, pressure and gas composition, can be measured online (Corona et al., 2013; Jimenez et al., 2015). One alternative for dealing with these issues is using software sensors. Soft sensors are models that estimate a hard-to-measure property by using relatively easy measurements. Soft sensors have been successfully applied in monitoring and controlling wastewater treatment plants (Haimi et al., 2013) and can be classified into three main categories: mechanistic, data-driven and hybrid models, depending on their underlying methods (James et al., 2000). Hybrid modelling approaches are often preferred over more complex mechanistic approaches, such as the anaerobic digestion model (ADM1). In hybrid modelling, the known linear/non-linear behaviors of the system can be described by a mechanistic approach and the unknown relationships among variables can be defined by data-driven models. Hybrid models can be more precise than mechanistic models because they integrate two approaches (Dürrenmatt and Gujer, 2012). Unlike hybrid models, data-driven models depend solely on a priori knowledge. The algorithm determines connections between input and output variables, and therefore it is a very attractive replacement of mechanistic models when they are not valid or available (Gernaey et al., 2004). Different techniques such as multivariate statistical methods, multiple linear regression (MLR), PCR, PLS, ANNs and fuzzy systems are used to design different soft sensors for wastewater treatment processes (Corominas et al.; Haimi et al., 2013; Newhart et al., 2019).

Tay et al. (Tay and Zhang, 2000) applied a neuro-fuzzy model to predict the response of high-rate AD based on different system disturbances. Their model inputs were organic loading rate (OLR), hydraulic loading rate (HLR), alkalinity loading rate (ALR), volumetric methane production rate (VMP), TOC and VFA. Their model outputs were VMP, TOC and VFA prediction one hour ahead. They showed that their model can predict the response of anaerobic wastewater for treatment systems in the presence of OLR, HLR and alkalinity loading shocks.

Mullai et al. (Mullai et al., 2011) studied the performance of an anaerobic hybrid reactor for treating penicillin-G wastewater. The experimental data were modelled by an ANFIS. Time and influent COD were considered as inputs and effluent COD was the only output of the model. Prior to modelling, the fuzzy clustering method was used to separate the different operation phases of the digester. The R^2 correlation for prediction versus actual values was found to be 0.9718, 0.9268 and 0.9796 for different phases of the digester operation. The results show that their model had a good prediction ability for COD removal efficiency. In another work, Güçlü and co-authors (Güçlü et al., 2011) implemented back-propagation ANN models for predicting effluent volatile solid concentration and methane yield. Effluent volatile solid and methane yields were predicted with the ANN using pH, temperature, flowrate, VFA, alkalinity, dry matter and organic matter as model inputs. The gradient descent with an adaptive learning rate algorithm was used. The R^2 correlations were 0.89 and 0.71 for volatile solid and methane yield respectively. The authors stated that they only used conventional parameters as model inputs, which is inappropriate because VFA and alkalinity are generally not measured frequently in most real AD plants due to the above issues. Rangasamy et al. (Rangasamy et al., 2007) studied modelling of an anaerobic tapered fluidized bed reactor for starch wastewater treatment using a multilayer perceptron neural network. ANN with two hidden layers was trained by using the back-propagation algorithm to predict different

process responses, including effluent COD, biogas production, VFA, alkalinity and effluent pH. The OLR and Influent pH were considered as model inputs. Briefly, most of the studies in the past were focused on predicting VFA for a laboratory-scale anaerobic digester by using available input parameters without considering the difficulty of measuring them. Furthermore, most of the developed models were trained based on the very limited operational conditions, thus the generalization ability and performance of the models in different situations is ambiguous (Güçlü et al., 2011; Rangasamy et al., 2007).

In this chapter, we study the ability of different data-driven modelling techniques, such as ANNs, ELM, SVM, RF and GP, to develop robust VFA monitoring software sensors from exclusively online easy-to-measure variables. The wrapper feature ranking method combining different models was used to select the most influential process variables for developing a data-driven soft sensor to increase the accuracy and reduce computation time. The procedure was applied to the wastewater treatment BSM2 running in the Matlab Simulink environment. The developed software sensors were compared in terms of accuracy, robustness and transparency. Transparency is important because transparent models are very beneficial for controlling and gaining insight into the modelling procedures of AD processes. Although the Genetic Algorithm, which is very similar to the GP, has been applied for modelling AD processes, there is no soft sensor designed with the GP technique in this context (Beltramo et al., 2019; Huang et al., 2016). Therefore, in this chapter, we also discuss using the GP model for designing more transparent and interpretable soft sensors.

4.2. Materials and Methods

4.2.1. Data Collection

The first step in designing a data-driven soft sensor is obtaining the process data. Therefore, we used synthetic data produced by BSM2. It should be noted that this chapter describes a preliminary study for designing soft sensors for AD processes by applying different techniques to synthetic data; therefore, for applying the approaches discussed in this chapter to real systems, real process data is necessary. Thirteen process variables obtained from the simulation are listed in Table 4.1. These variables are measured from the influent, effluent and gas line of AD every 15 min.

Table 4. 1 Obtained variables from Benchmark Simulation Model No.2 (BSM2).

Parameters	Unit	Parameters	Unit
Effluent COD	gm ⁻³	CH ₄ mol_fraction	-
Effluent alkalinity	Molm ⁻³	CO ₂ mol_fraction	-
Influent TSS	gm ⁻³	H ₂ mol_fraction	-
Effluent TSS	gm ⁻³	Pressure	bar
Effluent pH	-	Effluent ammonia	g m ⁻³
Effluent BOD	gm ⁻³	Influent Flow	m ³ d ⁻¹
Gas flow	m ³ d ⁻¹		

COD: chemical oxygen demand; TSS: total soluble solid; BOD: biological oxygen demand.

4.2.2. Pre-Processing of the Data

To achieve successful soft sensor development, the data set should be pre-processed to eliminate the missing and redundant values, outliers and signal noise. In this chapter, as the data

set is obtained synthetically from a simulator, there is no need to perform any further steps for removing the outliers and processing the missing values. However, the signal noise which is incorporated in BSM2 to obtain more comparable and realistic benchmark simulation results needs to be considered. Therefore, the signal noise should be handled appropriately for soft sensors in order to estimate VFA values accurately in different conditions. Prior to model construction and prediction, the weighted moving average (WMA) method is adopted to reduce signal noise, as it is fast to compute and easy to use, compared to the other methods. In the WMA method each data point in the sample window is multiplied by a different weight based on its position (Hota et al., 2017) as given in Equation 4.1:

$$F_t = \frac{\sum_{i=1}^n w_i A_{t-i}}{\sum_{i=1}^n w_i} \quad (4.1)$$

where F_t is the smoothed signal occurrence at time t , W_i is the weight to be given to the actual occurrence for the time $t-i$, A_i is the actual occurrence for the time $t-i$ and n is the total number of window lengths in the prediction. A window length of 100 sampling times is chosen for all model constructions.

As the measured variables have different units, they need to be normalized before the different models are developed. The variable values were normalized according to their mean and standard deviation. To assess model performances, normalized root-mean-squared error (NRMSE) and the coefficient of determination (R^2) were calculated according to Equations 4.2 and 4.3 respectively.

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^n (y_{prd,i} - y_{act,i})^2}{n}}}{\max(y_{act}) - \min(y_{act})} \quad (4.2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{prd,i} - y_{act,i})}{\sum_{i=1}^n (y_{prd,i} - y_m)} \quad (4.3)$$

where y_{act} and y_{prd} are the actual and predicted values, respectively, i is the data record number, y_m is average of the experimental value, and n is the total number of records.

In addition, to examine the generalization ability of the models, the overall data set is split into a training set, used to fit the model, and a validation set, used to calculate the error. The obtained simulated data from day 245 to 450 and day 451 to 609 were used as training and validation sets respectively. It should be noted that due to the high number of data points (58,465), the training set was randomly sampled and finally 1252 data points were uniformly selected to reduce the computation time during model training.

4.3. Data-Driven Methods

4.3.1. Artificial Neural Network (ANN)

An ANN is a non-linear model that has at least three main layers, called input, hidden and output layers. All layers are connected by components called neurons (Figure 4.1)—in which, three main operations are carried out. First n -elements of the input vector (z_1, z_2, \dots, z_n) are multiplied by weights ($w_{1,1}, w_{1,2}, \dots, w_{1,n}$). Second, the weighted inputs are added together with bias signal b to obtain a value (Gil et al., 2018):

$$a = z_1 \cdot w_{1,1} + z_2 \cdot w_{1,2} + \dots + z_n \cdot w_{1,n} + b \quad (4.4)$$

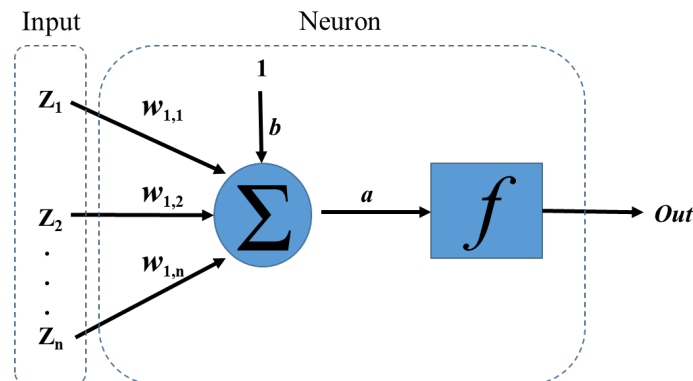


Figure 4. 1 Simple scheme of a neuron.

Finally, the output signal is a function of a , the weighted sum of the inputs.

The purpose of an activation function is to ensure that the input space is mapped to a different space in the output. Linear, sigmoid and hyperbolic transfer functions are generally used in ANNs (Eskandarian et al., 2017; Gil et al., 2018).

The structure of neurons in each layer, which are grouped and connected, is called the topology of the network. There are many different topologies; however, the MLP is the most commonly used. The MLP topology can be characterized by the same number of network inputs and outputs, equal to the number of input and output variables of the system to be modelled (Gil et al., 2018). There are no specific rules for finding the best topology of the network, and therefore different methods such as trial and error or evolutionary algorithms (e.g., genetic algorithm) can be used for this purpose (Stanley and Miikkulainen, 2002). Once the topology is obtained, the network should be trained. We used a back-propagation algorithm using Stochastic Gradient Descent (SGD) with an adaptive learning rate algorithm (Gil et al., 2018). There are many parameters that need to be tuned for ANN, the most important ones include the number of layers, number of neurons in each layer, type of transfer function and regularization terms (L1 and L2), which prevent overfitting and improve generalization, were considered for a grid search. The

number of epochs was considered to be high (2000) because the early stopping method was used. The ANN models were trained in R (R, 2017) using the H2O package (Candel et al., 2018). To determine the best ANN structure, the number of neurons in the hidden layer was varied from 5 to 160.

4.3.2. Extreme Learning Machine (ELM)

The ELM is a single hidden layer feed-forward neural network (SLFN) that was first introduced by Huang et al. (Guang-Bin Huang et al., 2004). In traditional ANN, all the parameters (number of neurons and hidden layers) have to be tuned, which leads to dependency between different parameters (weights and biases) in each layer. However, in ELM, the hidden layer does not need to be tuned (Abdullah et al., 2015). In ELM, the input layer weights and biases are randomly initialized, then fixed without any tuning iteration. The output layer weights are calculated analytically. As there is no iterative procedure for the tuning phase, ELM has a faster learning speed than traditional ANN and has a better generalization performance. Figure 4.2 shows the structure of the proposed ELM model.

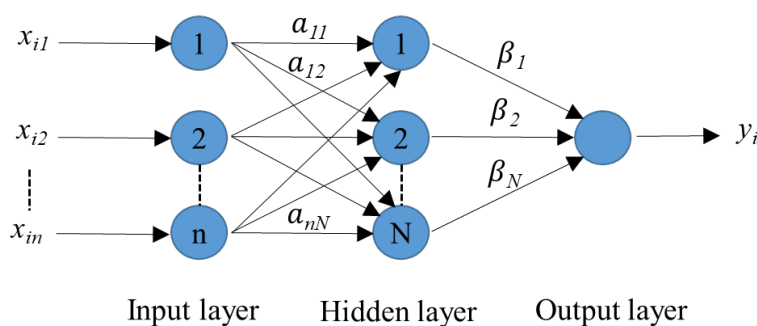


Figure 4. 2 Scheme of the extreme learning machine (ELM) model.

Considering M arbitrary distinct samples (x_i, y_i) , where $x_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in R^n$ is the input vector and $y_i = [y_{i1}, y_{i2}, \dots, y_{im}]^T \in R^m$ the output vector, then the output of SLFN with N hidden neurons can be calculated as:

$$y_i = \sum_{j=1}^N \beta_j \cdot f(a_j x_i + b_j) \quad i = 1, \dots, M \text{ and } j = 1, \dots, N \quad (4.5)$$

where $\beta_j = [\beta_{j1}, \beta_{j2}, \dots, \beta_{jm}]$ are the weights of the output layer, which need to be estimated; $f(\cdot)$ is the transfer function; a_j and b_j are the input weights and biases, respectively. Equation 4.5 can be written in matrix form (Zhang et al., 2017):

$$y = H\beta \quad (4.6)$$

where, $y = (y_1, y_2, \dots, y_M)^T$, $\beta = (\beta_1, \beta_2, \dots, \beta_N)^T$ and H given by:

$$H = \begin{pmatrix} f(a_1 x_1 + b_1) & \cdots & f(a_N x_1 + b_N) \\ \vdots & \ddots & \vdots \\ f(a_1 x_M + b_1) & \cdots & f(a_N x_M + b_N) \end{pmatrix} \quad (4.7)$$

The β is determined via Moore-Penrose's generalized inverse:

$$\beta = H^+ y \quad (4.8)$$

where

$$H^+ = (H^T H)^{-1} H^T \quad (4.9)$$

The value of the input weights (a_j) and biases (b_j) can be randomly initialized; however, the weights of the output layer (β_j) need to be determined with experimental data. Usually, the number of neurons in the hidden layer is higher than in the input layer ($N > n$). For the grid search, the number of neurons in the hidden layer was varied from 20 to 130.

4.3.3. Random Forest (RF)

The RF model was first developed by Breiman (Breiman, 2001). In this technique, a model is built based on a set of unpruned single regression trees. This is equal to combining different nonlinear relationships to form a more accurate non-linear model. In this method, trees are generated based on bootstrap sampling from the original training data set. In bootstrapping, random subsets of data are chosen from the original training data set to ensure the diversity among the ensemble of trees and enhance the prediction ability. The best node splitting feature for each node is selected from a set of m features that are randomly chosen from the total M features ($m < M$). By choosing m random features for node splitting, the correlation between different trees and thus the average response of multiple regression trees is expected to have lower variance compared to single regression trees (Breiman, 2001; Eskandarian et al., 2017). RF has three main tuning parameters: the number of trees in the forest ($ntrees$), the number of features randomly sampled as candidates at each node split ($mtry$), and the maximum number of nodes in the trees ($maxnode$). The parameters of the RF model for the grid search were set at 2:7 (with step size 1) for $mtry$, and 1000 to 2000 (with step size 100) for $ntrees$ and 5 to 30 (with step size 5) for $maxnode$.

4.3.4. Support Vector Machine (SVM)

SVM is a relatively new type of machine learning method that was introduced by Cortes and Vapnik for regression and classification problems (Cortes and Vapnik, 1995; Smola and Schölkopf, 2004). The objective of SVM is to map the input vectors X onto a very high-

dimensional feature space via a kernel function and then to make a linear regression in this space.

The regression function can be obtained as follows (Najafzadeh et al., 2016):

$$y(x) = \sum_{i=1}^l w_i \cdot K(x, x_i) + b \quad (4.10)$$

where $y(x)$ represents predicted values, $K(x, x_i)$ is a kernel function for input features, and w_i and b are coefficients. The most famous kernel functions are the polynomial kernel, the radial basis, the exponential radial basis, and the multilayer perceptron kernel function (Liu and Lei, 2018; Smola and Schölkopf, 2004). Due to the high prediction ability, the radial basis kernel function was used. The commonly used radial basis kernel has the form:

$$K(x, x_i) = \exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right) \quad (4.11)$$

where x and x_i are support vectors satisfying the equations of kernel function $K(x, x_i)$ and σ is the width of the Gaussian kernel function. More information on SVM can be found in references (Liu and Lei, 2018; Najafzadeh et al., 2016). To find the precise model, the tuning parameters of SVM, mainly the regularization parameter C and the inverse kernel width σ used by the radial basis kernel function, should be determined. The SVM model parameters for the grid search were set as 0.001, 0.01, 0.02, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 for σ , and 200 to 800 (with step size 100) for C .

4.3.5. Genetic Programming (GP)

GP is a powerful method for developing a mathematical expression. It was first introduced by Koza (Koza, 1994). In this method, the mathematical expressions are generated using input-output data by a biologically inspired algorithm, where the population continuously evolves towards the best-fitted model (Bahrami et al., 2016). GP and the GA have some similarities, while the most

important difference is the output format. The output of GA is a value, whereas the output of GP is a computer program. The complicated structures of computer programs, mathematical expressions, and process system models in GP are represented with trees (Bahrami et al., 2016; Sonolikar et al., 2017).

The GP tree structure consists of different nodes, classified into internal or external based on their position. The internal nodes can be chosen through the operators $\{+, -, \times, /, \sin, \cos, \log, \text{abs}\}$, mathematical functions, conditional statements or even the user-defined operators. The external nodes include the constants and the model variables. After introducing data into the algorithm, the population is randomly generated. This stage is crucial to increase diversity among the models. In the next stage, each model is evaluated by a fitness function to determine how well the developed models fit the observed data. The new generation is built based on the models, having a lower fitness error. The next generation of models is reproduced by using genetic operators, such as the reproduction, crossover and mutation operators (Bahrami et al., 2016; Koza, 1994). In the crossover operator, two models exchange the sub-trees to generate two new models, while in the mutation operator, the new model is formed from the previously generated model by substituting a randomly selected sub-tree with a newly generated one. The mutation operator is used to increase the genetic diversity of the population. The reproduction operator copies models without any change to the next generation. The fitness evaluation step is repeated for the newly generated population, and the whole procedure continues until an acceptable fitness value is achieved or the algorithm reaches its generation limit. By using this iterative procedure, the accuracy of each model improves in each iteration and finally the best one is considered as the output of the GP algorithm (Bahrami et al., 2016; Koza, 1994). Figure 4.3 shows a flow chart of the GP computation procedure:

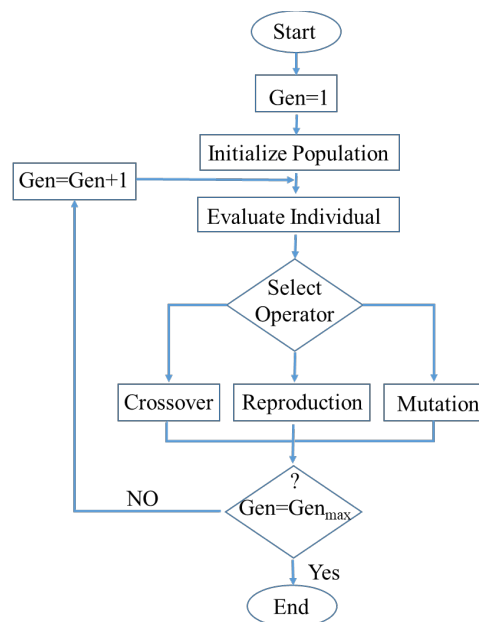


Figure 4. 3 Computation flowchart of genetic programming (GP).

4.3.6. Feature Ranking

Using many features for model building can cause many problems and directly affect the model accuracy. Hence, the redundant and unnecessary features should be eliminated as they may add noise and have no impact on the dependent variable. Using feature ranking helps to determine the importance of features and reduce the dimension of the data set (James et al., 2013). The other benefits of the feature ranking method are shorter training time, ease of interpreting models, overfitting reduction and lower cost in data collection. The wrapper method obtains different variable importance results depending on the feature ranking model applied. We used the fscaret package of the R environment to avoid this problem (R, 2017; Szlek and Aleksander, 2015). Briefly, variable ranking is performed in three steps: model training, variable ranking extraction, and variable ranking scaling according to the generalization error. The final variable ranking is

obtained by multiplying the raw variable importance with the fraction of minimal error obtained from models by the model's actual error according to the equation below (Szłęk et al., 2016):

$$weightedImp_i = \frac{rawImp_i}{\sum_{i=1}^i rawImp_i} \times \frac{minError}{Error_i} \times 100 \quad (4.12)$$

where $weightedImp_i$ is the weighted average of the individual model; $rawImp_i$ is the raw importance of the variable obtained from model i ; $minError$ is the minimal error obtained (RMSE or MSE) for all models; and $Error_i$ is error for model i .

4.4. Results and Discussion

4.4.1. Studying the Relationship between Input and Output Data

Before fitting non-linear models to the calibration data set, it is necessary to check whether a linear model can describe the relationships between parameters. If a linear model describes the relationships, then there is no need to use the more complicated models. Therefore, a simple linear regression (LR) was performed to predict VFA. We ran the BSM2 simulation based on the default influent file and the data signals recorded according to Table 4.1. As mentioned earlier, in most of the previously published studies, the authors used hard to measure parameters, such as VFA, COD, alkalinity, etc., to develop different soft sensors for the wastewater treatment processes. Therefore, the COD, alkalinity and BOD were initially eliminated from the model's input candidate list. Due to the simplicity of the LR model, it is not necessary to perform feature ranking prior to modelling, thus the rest of the parameters according to Table 4.1 were used for the LR model. Next, to determine the effect of other hard to measure parameters, such as TSS and Ammonia, on the

outcome, they were removed one by one from the input data of the LR model and the error was calculated. The results of the LR models with different input vectors are shown in Table 4.2.

Table 4. 2 Performance of the linear regression (LR) models based on the default influent file and different input vectors.

	All_Inputs	All_Input_Except _TSS	All_Input_Except _Ammonia	All_Input_Except _TSS & Ammonia
Training_NRMSE	0.028	0.029	0.028	0.029
Test_NRMSE	0.039	0.039	0.038	0.038
Training_R ²	0.981	0.978	0.979	0.976
Test_R ²	0.967	0.966	0.967	0.968

NRMSE: normalized root-mean-squared error; R²: coefficient of determination.

The results clearly show that the relationship between parameters is highly linear. The LR model can still predict VFA with high accuracy even when TSS and Ammonia are eliminated from the input vector. This is contrary to the real behavior of the AD process, which is a very non-linear system. The most logical reasons for this linearity are the default influent data file and the design of the AD reactor itself. The default influent file is designed in such a way that the variation in the organic load to the AD reactor is not very intense, and thus, inhibition phenomena, which are the main source of non-linearity in the AD process, will not occur. In addition, the AD reactor is quite over-designed compared to its feed load, and small disturbances from the sludge recovery unit do not have a great impact on its operation. Therefore, it is possible that the process is pushed towards very narrow linear operational ranges. Thus, to increase the non-linearity and make the simulated AD process more realistic and challenging, we manipulated the feed load to the reactor. Furthermore, by manipulating the feed load, the AD's behavior is similar to the co-digestion, which is more interesting than the mono-digestion process. We randomly changed the

concentration of the inorganic nitrogen (S_{in}), the composite (X_c), and the carbohydrate (X_{ch}) as well as the feed flow rate of the reactor. The summary of default and modified values of the parameters are shown in Table 4.3.

Table 4. 3 Summary of default and the modified values of inorganic nitrogen (S_{in}), composite (X_c) and carbohydrate (X_{ch}).

	Default Values				Modified Values			
	S_{in}	X_c	X_{ch}	Flow	S_{in}	X_c	X_{ch}	Flow
	(kmolm^{-3})	(kgm^{-3})	(kgm^{-3})	(m^3d^{-1})	(kmolm^{-3})	(kgm^{-3})	(kgm^{-3})	(m^3d^{-1})
Min.	0.0006	0	0.000	56.55	0.0006	0.00	0.000	1.993
1st Qu.	0.0015	0	2.941	137.07	0.0019	0.00	3.293	82.257
Median	0.0019	0	3.952	175.81	0.1156	22.54	4.897	138.704
Mean	0.0020	0	3.830	183.57	0.0957	16.41	7.256	140.602
3rd Qu.	0.0022	0	4.833	217.90	0.1584	28.71	9.684	193.301
Max.	0.0325	0	8.607	479.96	0.2718	39.52	40.464	477.957

After manipulating the AD feed signal, a new LR model with the same approach was implemented with the new data set to determine the impact of changes on the VFA prediction.

The prediction results are shown in Table 4.4.

Table 4. 4 Performance of the LR models based on the modified anaerobic digestion (AD) variables and different input vectors.

	All_Inputs	All_Input_Except _TSS	All_Input_Except _Ammonia	All_Input_Except _TSS & Ammonia
Training_NRMSE	0.103	0.133	0.143	0.189
Test_NRMSE	0.197	0.213	0.122	0.304
Training_R ²	0.865	0.794	0.748	0.586
Test_R ²	0.654	0.511	0.841	0.663

It can be seen that by manipulating the AD feed carbon, inorganic nitrogen load and feed flow, the process is pushed towards the non-linear operational ranges. The best Test_NRMSE is obtained by using all inputs except ammonia concentration. Although the accuracy of this model is better than the other LR models, its performance is far from an acceptable range. Thus, there is a demand for accurate prediction of VFA by non-linear models. In the next sections of this chapter, non-linear approaches are applied to the newly obtained data set to develop more accurate soft sensors.

4.4.2. Choosing the Most Influential Variables Using the Feature

Ranking Method

As mentioned earlier, using redundant and unnecessary features may add noise to the model and increase the risk of over-fitting. Therefore, a combination of influential sensor measurements should be chosen as model inputs. The measurements must be the most influential and, at the same time, they have to be easy to measure to satisfy the soft sensor definition. The best combination of parameters is generally chosen with feature ranking techniques. In the present chapter, we used the fscaret feature ranking technique. Before carrying out the feature ranking method, hard to measure parameters, including COD, alkalinity and BOD, were removed from the data set. The gas flow and CH₄ mole fraction were also removed due to their direct correlation with pressure and CO₂ mole fraction respectively. It should be noted that the same results could be obtained by using gas flow and CH₄ mole fraction instead of using pressure and CO₂ mole fraction; therefore, the decision to eliminate correlated parameters can be made based on the simplicity and availability of measurements. The remaining parameters, listed in Table 4.1, were used as an input

vector for the fscaret method. Figure 4.4 shows the importance of the variables on a scale from 0 to 100 obtained with the fscaret method for VFA prediction.

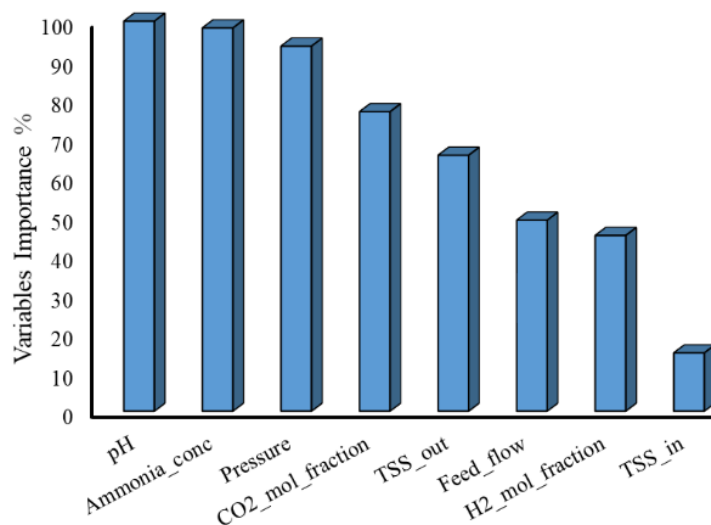


Figure 4. 4 Importance of variables on a scale from 0 to 100 obtained with the fscaret method.

In Figure 4.4, pH, Ammonia concentration and pressure have the most influence on VFA. To determine the best subset of inputs, we used the SVM method as a non-linear technique. SVM was chosen because it is a fast and accurate method for modelling a non-linear system. Therefore, the most important variables (pH, Ammonia concentration and pressure) were chosen as the core subset and the other variables were added one by one to the model based on their importance values. The validation error is estimated for each subset after the models are trained. Table 4.5 shows the NRMSE of each subset based on the trained models. It can be seen that the second subset has the best NRMSE; therefore, this subset was selected for further development of VFA soft sensors based on ANN, ELM, SVM and RF. It should be noted that, as the GP model selects influential features inherently, there is no need to perform feature selection before its training. It is interesting that, based on the obtained data set, the flow does not have a significant effect on the VFA compared to the other variables and it is not included in the best subset. Figure 4.5 shows

the trend of four input and output variables over the total experiment period. The input and output trends show that the behavior of the AD process is very complex, which results in absence of direct correlations between parameters.

Table 4. 5 Result of different subsets trained by support vector machine (SVM).

Inputs	R ²	NRMSE
pH + Ammonia_conc + pressure	0.813	0.182
pH + Ammonia_conc + pressure + CO ₂ _mol fraction	0.990	0.033
pH + Ammonia_conc + pressure + CO ₂ _mol_fraction + TSS_out	0.972	0.058
pH + Ammonia_conc + pressure + CO ₂ _mol_fraction + TSS_out+Flow	0.977	0.044
pH + Ammonia_conc + pressure + CO ₂ _mol_fraction + TSS_out + Flow + H ₂ _mol_fraction	0.971	0.049

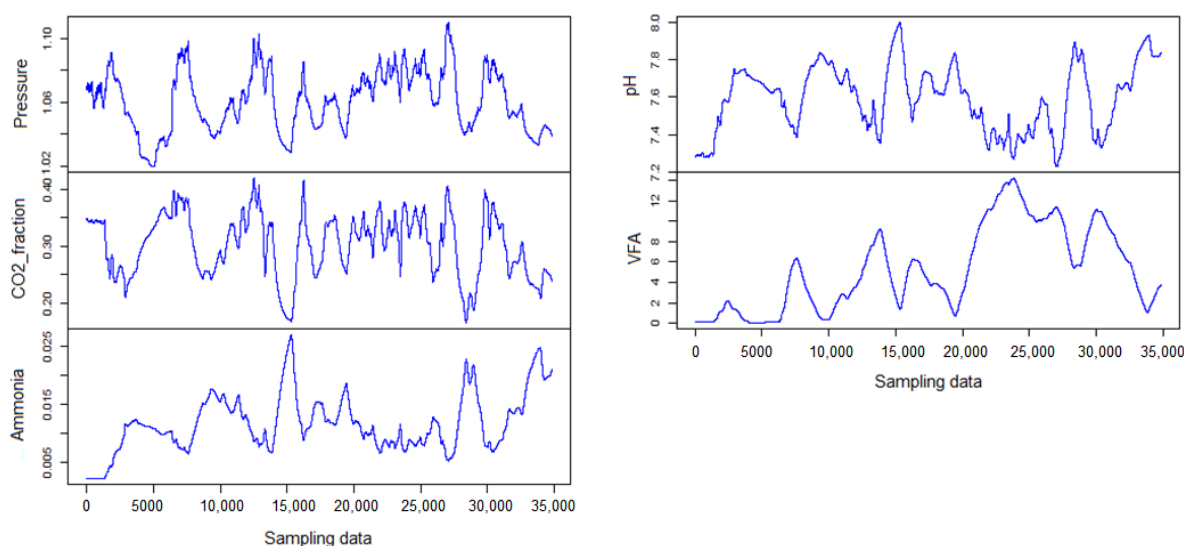


Figure 4. 5 Trend of input and output variables used for developing soft sensors.

4.4.3. Soft Sensor Design

Before the soft sensors were designed, the data set was split into training and validation partitions. The uniformly sampled data from day 245 to 450 were used for training and the rest were used for validation of the final models. All calculations were performed using Amazon

Elastic Compute Cloud with 36 cores and 60 GB of random-access memory (RAM) running on the *openSUSE* operating system (SUSE, Nürnberg, Germany). All modelling procedures were performed in the *R* environment, except GP, which was developed using the *Eureka Formulize* software package (Schmidt and Hod, 2014; Schmidt and Lipson, 2009). This package has been optimized to find simple models, which are expected to have good generalization ability. The tuning parameters of each model were estimated using an extensive grid search. In this method for searching algorithm parameters, a tuning grid must be specified manually. In the grid, each algorithm tuning parameter can be specified as a vector of possible values. These vectors are combined to define all the possible combinations to try. As previously mentioned, different models such as RF, ELM, ANN and SVM have been used to design VFA soft sensors. Each of these models has its own parameters that must be tuned to generate an accurate model. Table 4.6 shows the final tuning parameters obtained with the grid search. No major parameters need to be tuned for the GP model implemented in the *Eureka Formulize* software package.

Table 4. 6 Final tuning parameters of soft sensors obtained with the grid search.

Algorithm	Tuning parameters				
ANN	Neuron Size	Transfer Function	Number of Hidden Layers	L1	L2
	108	Tanh	1	1×10^{-5}	1×10^{-5}
RF	mtry	Number of trees	Maximum nodes		
	4	1600	20		
ELM	Neuron size	Transfer function			
	126	Sigmoid			
SVM	Sigma	C			
	0.2	500			

ANN: artificial neural network; RF: random forest; ELM: extreme learning machine; SVM: support vector machine.

The NRMSE and R2 are estimated based on the training and validation sets for each model. The results obtained using different techniques are shown in Table 4.7.

Table 4. 7 Results of soft sensors for the training and validation sets.

Algorithm	NRMSE Training	R ² Training	NRMSE Validation	R ² Validation
ANN	0.0089	0.9992	0.0192	0.9969
RF	0.1432	0.7533	0.3419	0.5784
ELM	0.0003	0.9999	0.0169	0.9977
SVM	0.0165	0.9966	0.0390	0.9941
GP	0.0025	0.9999	0.0037	0.9998

GP: genetic programming.

The lowest validation error was achieved by the model developed with the GP algorithm; thus, it should be considered as the final, ready-to-use model. The ANN, ELM and SVM algorithms gave a slightly higher error, but they were still better than the RF model. It can be seen that the RF predictions failed with a NRMSE of 34%. The high error of RF is because it is a rule-based method, in which the data is categorized into different classes. Therefore, if applied to temporal data, weak results will be obtained due to the high number of classes. Although ANN, ELM and SVM are potential non-linear function approximation methods with a broad application, they are still “black box” models whose structure and parameters do not provide any insight into the phenomena underlying the process being modelled. In contrast, the GP model is transparent and can generate explicit equations that are very convenient for direct online implementation in the existing process information and control systems. The GP optimal soft sensor model generated by the Eureka package is given as a coefficient in Table 4.8.

Table 4. 8 Coefficient table for the equation obtained by GP.

Term *	Coef	Term *	Coef
constant	1132.95	P	-11.28
$[NH_3]$	1469.52	$[CO_2]$	-33.72
$\sqrt{[NH_3]}$	3354.73	pH	-295.38
pH^2	19.45	$pH \times \sqrt{[NH_3]}$	-441.87
$[CO_2] \times [NH_3] \times e^{(1.60 \times [CO_2])}$	20271.09	$[CO_2] \times pH \times [NH_3] \times e^{(1.60 \times [CO_2])}$	-2670.07

* $[NH_3]$, pH , $[CO_2]$, and P correspond to the ammonia concentration, pH , CO_2 fraction and pressure, respectively.

The GP was trained based on the full input vector; however, Table 4.8 shows that the final model contains four variables. This is in accordance with the result obtained by the *fscaret* method. In addition, to check the prediction ability of each model, the final models were tested using the whole data set without sampling. Figure 4.6 shows the prediction result of each model. It can be seen that the performance of all models is satisfactory except for the RF model.

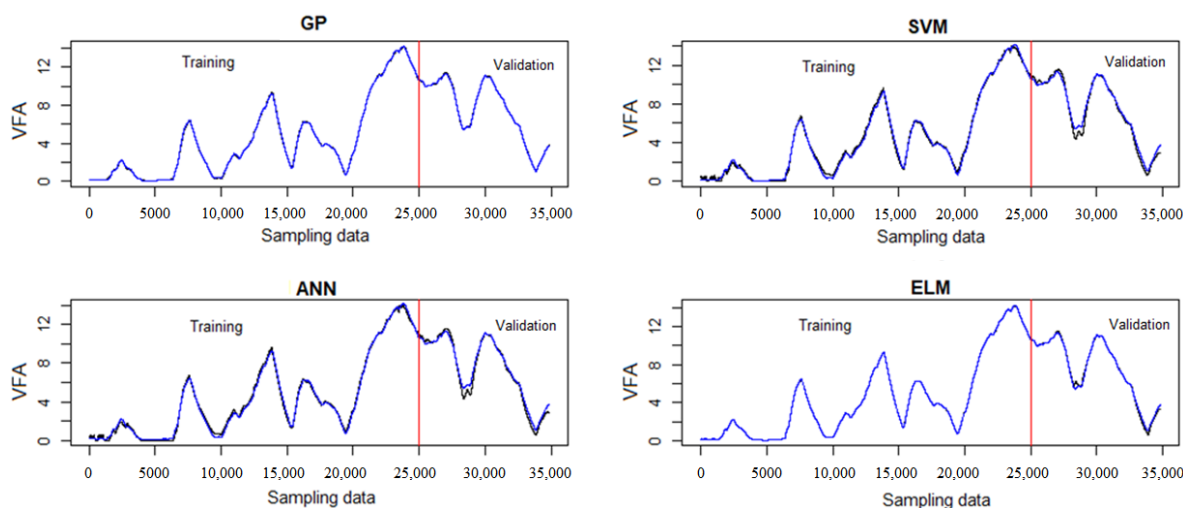
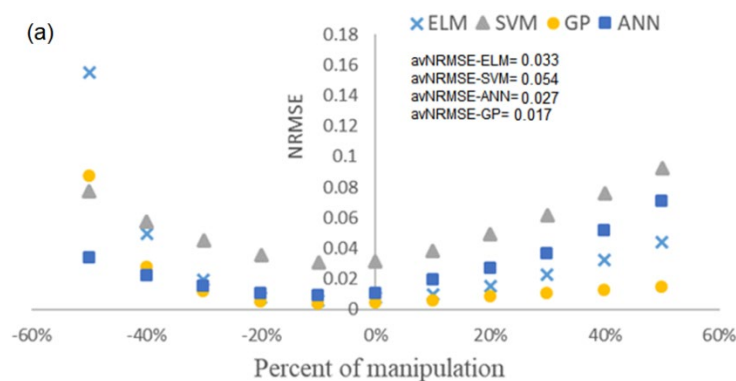


Figure 4. 6 Prediction results of different soft sensors; black is the actual values and blue is the predicted volatile fatty acids (VFA) values. GP: genetic programming; SVM: support vector machine; ANN: artificial neural network; ELM: extreme learning machine; RF: random forest.

4.4.4. Evaluation of the Robustness of Soft Sensors

The prediction accuracy of soft sensors tends to drop after a period of their online operation due to changing process states. This change in soft sensor accuracy may cause some issues in the process operation, including increasing the maintenance cost and decreasing the quality of final products. Therefore, it is very beneficial to examine how this accuracy degradation affects the final prediction. To do this, we selected three important biochemical parameters of anaerobic digestion: the hydrolysis rate of carbohydrates ($k_{\text{hyd,ch}}$), the maximum uptake rate of acetate ($k_{\text{m,ac}}$), and the ammonia inhibition constant ($k_{\text{I,NH}_3}$). We then varied them by $\pm 50\%$ around their default values. The default values of $k_{\text{hyd,ch}}$, $k_{\text{m,ac}}$ and $k_{\text{I,NH}_3}$ were 10 d^{-1} , 8 d^{-1} and $0.0018 \text{ kmol.m}^{-3}$ respectively. To evaluate the robustness of each model, new data sets were generated by the BSM2 simulation with modified biochemical parameters. Then, each trained model was tested based on the newly obtained data set. Figure 4.7 shows the results of the robustness evaluation for each model. As the RF soft sensor failed, it is not considered for the robustness evaluation.



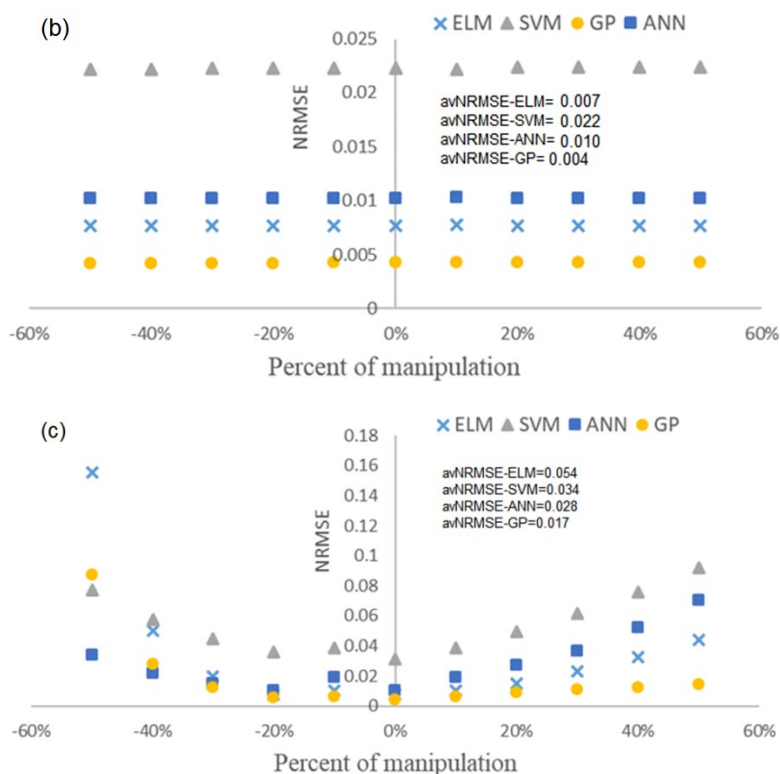


Figure 4. 7 Results of the robustness evaluation for each model: (a) for $k_{m,ac}$; (b) for $k_{hyd,ch}$; (c) for $k_{I,NH3}$. NRMSE: normalized root-mean-squared error; avNRMSE: average normalized root-mean-squared error.

From the variation of $k_{m,ac}$, it can be concluded that GP is the most robust approach, although NRMSE increases at -50% . For $k_{hyd,ch}$, the NRMSE is quite constant which shows that all techniques are insensitive to variations of this parameter; nonetheless, GP still has a lower error compared to the other model. For $k_{I,NH3}$, again GP is more robust as the error does not change significantly. The least robust model is SVM, which has the highest error compared to the other models. To make comparison easier the average NRMSE for variation of each parameter is also shown in Figure 4.7.

Comparing avNRMSE for all the soft sensors shows that the GP has lower prediction error during changes in AD state parameters. Overall, the performance of other soft sensor models is also acceptable due to low avNRMSE.

4.5. Conclusions

In the present chapter, different data-driven soft sensors are proposed for predicting the effluent VFA of the AD process. The performance of these soft sensors has been successfully demonstrated with a case study based on synthetic data obtained from BSM2. Analyzing the simulated data with the default BSM2 influent file, the LR models showed that the relationship between input and output data is highly linear. The BSM2 model was modified to introduce non-linearity in the simulated data. Moreover, by applying this modification the behavior of the AD was similar to the anaerobic co-digestion process. The best subset of input variables including pH, ammonia concentration, pressure and CO₂ mole fraction was obtained by using the *fscaret* method along with SVM. After training the models, we obtained the prediction and generalization performances of each model based on a specific validation data set. The results show that all models except RF predict the effluent VFA precisely; however, GP performed slightly better than the other models. The RF model totally failed to predict VFA. This suggests that tree based models are not a very appropriate choice for developing models with an extrapolation capability similar to soft sensors. Assessing the robustness of soft sensors shows that the GP model is more robust and less sensitive to the state changes of the AD process. Last but not least, the other benefit of adopting the GP soft sensor, apart from accuracy and robustness, is its transparency, which makes it easy to integrate into process control systems without any further modifications.

UNIVERSITAT ROVIRA I VIRGILI

Data-driven soft-sensors for monitoring and fault diagnosis in wastewater treatment plants

Pezhman Kazemi

CHAPTER 5

Data-driven techniques for fault detection in anaerobic digestion process

Monitoring AD is a crucial task to ensure optimized operation and to prevent failures and serious consequences during the running of the plant. To fulfill this task, a useful data-driven framework is proposed and validated on a simulated data set obtained using the BSM2 from the IWA. The proposed framework is based on data-driven soft-sensors predicting total VFA, mainly acetate, propionate, valerate and butyrate concentrations inside the digester. The VFA concentration is considered because it does not only reflect the current process health, but it is also sensitive to the incoming feeding imbalances. VFA soft-sensors using different advanced techniques such as SVM, ELM and ENN are tested and compared in terms of accuracy FD robustness. To compare the proposed approaches with the traditional FD method, a PCA model was also developed.

5.1. Introduction

AD is indeed a complex process in which four consecutive biological steps are involved in the degradation of the organic matter. When operating such complex processes, various human health and environmental risk can be identified, including explosion, fire, and biological and ecological pollution risk. For instance, if an AD plant has totally failed, because of biological community problems, the reactor should be drained and filled up again with a new inoculum. Due to the high amount of organic content and various pathogenic bacteria (e.g. *Salmonella*, *Enterobacter*, *Clostridia*, *Listeria*), parasites (e.g. *Ascaris*, *Trichostrongylidae*, *Coccidae*), fungi, viruses (Ritari et al., 2012) which could be considered as environmental biohazard, the sludge inside the AD cannot be easily drained to the environment. Moreover, draining and filling procedures may take several months, during which the plant cannot be operated in nominal conditions. Therefore, to implement process safety and prevent environmental risk, early detection of abnormal conditions that may cause failures and more catastrophic consequences in the future is imperative. Many parameters influence the AD process performance; therefore, the monitoring of these parameters is mandatory to ensure a stable and safe operation. These process parameters can be classified into three groups: (i) parameters that characterize the process, (ii) process operating variables and (iii) early indicators of process imbalance. The first group (feed quality and quantity, biogas production, temperature, pH, ammonia concentration and total solids/dry matter) shows the overall AD plant health; therefore, it is mandatory to monitor them regularly to identify the possible changes from normal operation of the process. However, these parameters are not suitable for the early detection of abnormal conditions. For instance, if the gas production or pH decreases, it means that the instabilities already occurred in the process. The variables involved in the second group are the organic loading rate and the hydraulic retention time. Their

values depend on the plant operator's decision. Usually, these parameters are changed when there is an alteration in feed composition or due to the process instability. The third group gathers early indicators of process imbalance such as gas composition measurement (CH_4 , CO_2 , H_2), redox potential, alkalinity and VFA concentration. Although these parameters can indicate the process imbalance beforehand, they do not give direct information regarding the exact cause of process imbalance (Boe et al., 2010; Dixon et al., 2007; Weiland, 2008).

The main goal of every FD system is to identify any abnormal events. Abnormal events can be defined as the situations when the process deviates significantly from its normal operation. FD methods are generally classified into three categories: model-based, knowledge-based and data-driven methods (Sánchez-Fernández et al., 2018). In model-based methods, a mathematical model is developed based on the knowledge of the process dynamics. This mathematical model is capable of describing the process dynamics and revealing the physical meaning, which is very important in practical applications. Several studies on the supervision and diagnosis of AD processes by applying model-based techniques have been reported. For example, González et al. (Alcaraz-González et al., 2012) have designed an interval observer that can detect and isolate faults in sensors and design input hypotheses in the presence of unmeasured input disturbances. Their approach was applied to anaerobic digestion process carried out in a continuous fixed bed reactor to treat industrial wine distillery wastewater. To design the FD framework, five observers were designed by using different combinations of six available online measurements, including COD, VFA, total inorganic carbon, strong ions, CO_2 flow and CO_2 pressure. These observers were used for building residuals by comparing the estimated and the measured variables that allow, in a simple but highly efficient way, the detection of faults in the AD process. However, the complexity of the development and the amount of a priori knowledge, which has to be available for the model

development, is a major drawback of this approach. Knowledge-based methods are performed based on sets of rules that are extracted from the past experience of human experts. The extracted information can be the locations of input and output process variables, patterns of abnormal process conditions, fault symptoms, operational constraints, and performance criteria (Carrasco et al., 2004; Genovesi et al., 2000; Nan et al., 2008; Steyer et al., 1997). One of the well-known knowledge-based methods in the FD system for AD is fuzzy logic (Genovesi et al., 1999). Steyer et al. (Steyer et al., 1997) developed a hybrid approach that used both fuzzy logic and artificial neural networks for online detection and diagnosis of faults such as foaming, sudden changes in the effluent, pipe clogging, or bad temperature regulation in a fluidized bed reactor for the treatment of wine distillery wastewater. On-line measurements including pH, temperature, recirculation flow rate, input flow rate and gas flow rate were pre-processed using fuzzy logic to build a features vector. Then, according to the discrimination fuzzy rules, these feature vectors were classified into predefined categories that show the state of the process. The role of neural network was to classify the process states and to identify the faulty or dangerous ones. In another study, Carrasco et al. (Carrasco et al., 2004) developed a fuzzy-logic-based diagnosis system for the determination of acidification states of an anaerobic wastewater treatment plant. In order to develop the diagnosis system, Takagi–Sugeno–Kang method of fuzzy inference was used. The membership for the on-line measurements variables were then determined using expert knowledge. As outputs of the system, seven possible results were considered as fuzzy sets with their corresponding membership functions. Each possible result corresponds to a situation: Organic Overload (meaning high acidification by organic overload), Medium Acidification by Organic Overload, Low Acidification by Organic Overload, Normal condition, Low Acidification by Hydraulic Overload, Medium Acidification by Hydraulic Overload and Hydraulic Overload or

high acidification by hydraulic overload. The success of this approach heavily depends on the operator and engineer's knowledge which is implemented into this method. Moreover, learning and extracting such information from process history is always a difficult operation and sometimes impossible, especially when the process is very complex and has non-linear behaviours (Nan et al., 2008; Sánchez-Fernández et al., 2018). In contrast, the data-driven methods are purely designed based on historical records and online data of the process independent of any mathematical model or intervention of human knowledge. This method could be very beneficial in complex industrial processes, where the mathematical models and human knowledge are not easy to obtain in practice. Recently, this method has become more attractive due to the availability of large amounts of data collected by the utilization of distributed control systems. The main disadvantage of data-driven method is ambiguity in detection of faults and disturbances before the design stage due to the lack of real process data availability. This method has already been applied to wastewater treatment plants (Sánchez-Fernández et al., 2018), but to the best of our knowledge, it has not been used yet for FD in AD processes. Therefore, the objective of the present chapter is to compare several data-driven approaches capable of detecting random faults in the process state as well as different sensor faults. Moreover, due to the importance of VFA (mainly consisting of acetate, propionate, valerate and butyrate) as a high potential state indicator of AD processes, VFA soft-sensors based on different data-driven techniques such as SVM, ELM and ENN have been tested and compared. The training of these soft-sensors has been performed by a data set obtained from the IWA BSM2 (Jeppsson et al., 2006; Nopens et al., 2010). Implementing data-driven approaches allows the generation of error or residual signal, resulting from the differences between measured VFA and prediction by the soft-sensor. This signal can be used directly or as an input for univariate SPC charts such as CUSUM charts to detect the faults when there is a shift in the

residual signal (Bin Shams et al., 2011). Finally, the performance of data-driven soft-sensors greatly relies on data of good quality (Xiao et al., 2017). In practice, due to maintenance or failures, some sensor's signals may become unavailable. Therefore, to examine the sensitivity of proposed FD approach during signal failures, different cases with and without missing values were simulated and studied. Another point that should be considered during the design of an FD system is the assumption that data follow Gaussian distributions. This assumption may not be fulfilled, leading to either too narrow or too wide fault control limits. To avoid this drawback, all the control limits used in this chapter were developed using the non-parametric bootstrapping method (Phaladiganon et al., 2011). For the sake of comparison between the proposed FD framework and the traditional method, PCA algorithm is also used. In this case, the same inputs vector as the soft-sensors plus VFA was used as input for the PCA method. The major contribution of the current work is to develop a data-driven FD framework which is very robust to different magnitudes of random faults occurring in AD processes.

The current chapter is organized as follows: First, VFA soft-sensors using different data-driven techniques based on the simulated data from BSM2 are developed. Then, the FD framework is described in parallel with the application of generated residual signals along with univariate statistical charts to detect faults. Finally, the approach is validated by implementing some artificial faults in BSM2. In this regard, two different (with and without missing values) data sets were generated and the results are discussed.

5.2. Materials and methods

5.2.1. Data collection and pre-processing

To design a data-driven soft-sensor, fourteen possible process variables are collected from BSM2. A list of these variables is presented in Table 4.1. The sampling time for recording these variables is 15 min and they are measured from different streams (influent, effluent and gas line) of the AD process. Before model construction to reduce the signal noise, the WMA method is adopted.

All model construction is performed by window length of 100 sampling data. Due to the different scale of measured variables, it is also crucial to normalize them before developing the soft-sensors. It should be noted that for the training of the soft-sensors, there were no missing values in the data set, the missing values being induced and imputed only during the FD procedure. Thus, the explanation of the method for missing values imputation will be discussed in the next section.

5.2.2. Soft-sensor assessment

To assess soft-sensor performances, NRMSE and the R^2 were used. Additionally, to examine the generalization ability of the soft-sensors, the overall data set is split into a training set, used to fit the model and a validation set, used to calculate the error. The obtained simulated data from day 245 to 453 and day 453 to 474 were used as training and validation set for developing soft-sensor, respectively. Due to the high number of collected data (20000), 2000 data points were

randomly sampled from the training set to reduce the computation time and resource during model training.

5.2.3. Fault detection assessment

To assess the performance of the proposed FD framework, precision, recall, and F1 score were calculated by using Equations 5.1 to 5.3. True positives (TP), False positives (FP) and false negatives (FN) are data points correctly labeled as faults, normal data points which incorrectly labeled as faults and faulty data points incorrectly labeled as normal, respectively.

$$Precision = \frac{TP}{TP+FP} \quad (5.1)$$

$$Recall = \frac{TP}{TP+FN} \quad (5.2)$$

$$F1 \text{ score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5.3)$$

The rate of false detection and miss detection is captured by precision and recall, respectively. The F1-score is the harmonic mean of precision and recall. The higher values of these indicators show the higher performance of the FD method under evaluation (Sokolova and Lapalme, 2009).

5.3. Data-driven techniques

5.3.1. Artificial neural network (ANN)

For developing VFA soft-sensors, two different types of ANN, including ENN and ELM have been used. In ENN, a bootstrap aggregation (bagging) method was used. Adopting this technique allows greater model stability and more efficient use of training data relative to early stopping

methods (methods that stop when the validation error is minimal instead of continuing optimization until errors on the training data are minimized). Different data samples generated by using bagging are used to train several neural network models. The final prediction of VFA is an average of predicted VFA by each model (Cannon, 2019; Cannon and Whitfield, 2002). Because of no iterative procedure for the tuning phase, learning speed of ELM is faster than traditional ANN and has a better generalization performance (Abdullah et al., 2015).

5.3.2. Support vector machine (SVM)

Discussion about the SVM method is already presented in Chapter 4.

5.3.3. Principal component analysis (PCA)

Nowadays, in most modern industrial processes massive amounts of data are generated due to many sensors measurements. PCA can handle this high dimensionality by projecting the correlated data onto space of uncorrelated PCs, which linearly uncorrelated and contains most of the variance of the original data. To detect faults using PCA method, two statistics namely Hotelling's (T^2) and the SPE should be estimated (Sánchez-Fernández et al., 2018). T^2 index represents the squared Mahalanobis distance of the retained PCs and measures the variability of the mean and covariance within these PCs. SPE statistic is the measure of the lack of fit for the PCA model. They can be estimated as follows (Jackson, 1991):

$$T^2 = x^T P \Lambda_k^{-1} P^T x \quad (5.4)$$

$$SPE = x^T \tilde{P}^T x \quad (5.5)$$

where x is a new observation, Λ_k contains in its diagonal the k most significant eigenvalues of the covariance matrix of original data matrix $X(n \times m)$, which is collected during normal operation of the process, in decreasing order. Their associated eigenvectors are contained in P . The residual eigenvectors and eigenvalues $(m-k)$ can be found in \tilde{P} .

5.3.4. Feature selection

Using many features for developing machine learning models may increase the noise and have a near-zero effect on the dependent variable. Therefore, before developing a high accuracy VFA, soft-sensor feature selection should be applied. The same procedure, as written in Chapter 4, has been used for feature selection.

5.4. Statistical process control

5.4.1. Control charts

One of the popular industry-standard methodologies to monitor processes and detect abnormal behaviors is SPC. During the operational state of the process, the real-time process measurement can be obtained. These measurement data are then used to calculate specific statistics plotted over time and for the normal operation of the process, these statistics must not pass the estimated threshold. In this chapter, univariate control charts (CUSUM and SPE) using the residual data obtained from VFA soft-sensors are depicted and used to determine the abnormal events.

- SPE chart: this chart illustrates the squared residual error obtained by comparing the real values of VFA and the predicted ones obtained by different soft-sensors. For PCA method, the SPE is obtained by Equation 5.5.
- CUSUM chart: here, the method represents the cumulative addition of deviations in every observation. This chart is more sensitive to small magnitude faults and reacts faster than the other chart. CUSUM chart can be obtained by using the following equations (Khusna et al., 2018):

$$C_i^+ = \max[0, C_{i-1}^+ + x_i - (\mu_{i,c} + k)] \quad C_0^+ = C_0^- = 0 \quad (5.6)$$

$$C_i^- = \max[0, C_{i-1}^- + (\mu_{i,c} + k) - x_i] \quad C_i = \max[C_i^+, C_i^-] \quad (5.7)$$

where k , $\mu_{i,c}$, C_i^+ and C_i^- are the slack variable, the mean of the variable under normal operation, and the upper and the lower CUSUM statistics, respectively. To introduced the robustness to the calculate statistics, the slack variable is used. Typically, k is estimated by half of the standard deviation of samples representing normal process operation.

5.4.2. Bootstrap confidence limits

As mentioned earlier, to use most control charts the monitoring statistics should follow a specific probability distribution. However, in many industrial processes, the probability of measurements do not follow a specific distribution. Therefore, considering a normal distribution assumption for systems that follow a non-normal distribution could increase false alarms or miss detection rate. To address these limitations, researchers suggested using different statistics based on a KDE technique (Chou et al., 2001; Phaladiganon et al., 2011; Xiao et al., 2017). The KDE-based control chart estimates the distribution of statistics and determines the control limits without a priori normality assumption. However, to perform a precise density distribution estimation using

the KDE method, several parameters should be determined. These include the types of kernel functions, a smoothing parameter and the number of points accounted for in the range of data. Besides, to calculate the percentile value of the estimated distribution, numerical integration is involved, which adds more complexity to this method. Due to these reasons, the control limit of all charts is calculated by using the bootstrap method. The bootstrap approach is relatively more straightforward than KDE because it does not require any specification of the parameters or a procedure for numerical integration. The interested reader may refer to (Khusna et al., 2018; Phaladiganon et al., 2011) for further information regarding the bootstrap method.

5.5. Proposed approach

In the proposed FD framework, the generated residual by the soft-sensors can be analyzed using a univariate SPC control chart. The goal of this framework is to detect the abnormal events in the AD process as fast as possible. This framework consists of three steps: designing VFA soft-sensors, estimating control limits and FD. The general diagram of the proposed approach is illustrated in Figure 5.1 and the following steps:

Step 1. The first step is soft-sensor training. Different algorithms such as SVM, ELM and ENN were used for developing soft-sensors. This step is composed of sub-steps as follows:

1. The normal operation data is collected from BSM2.
2. Data pre-processing, which consists of scaling the data to the adequate range and implementing WMA method to reduce noise in signals.
3. Feature selection to select the most appropriate variables for developing VFA soft-sensors.
4. Developing soft-sensors, as described in the previous section using different techniques. PCA model is also developed for comparison with the methods mentioned above.

Step 2. Control limit estimation. In this step, the control limits of every chart used in this framework is calculated. The control limits for soft-sensors were obtained by the bootstrap method using the residual values. For PCA method, the bootstrap was performed on the T^2 and SPE statistics.

Step 3. Fault detection. This step is calculated online and consists of different stages, including:

1. Collecting the data from the BSM2. To validate the proposed FD framework, different types of faults are artificially simulated and the data is collected for analysis.
2. The same pre-process step for training is performed on the new data.
3. If any missing values are recognized, they are imputed by the most recent present value before it.
4. Generating the residuals which are the difference between the measurements and the output of soft-sensors constructed in step 1, for VFA.
5. Compute the CUSUM and SPE statistics from the soft-sensor residual. For the PCA model, the SPE and T^2 statistic are estimated separately.
6. Recognizing the faulty events by comparing the defined statistics with their correspondence control limit. If each of the defined control statistics (SPE and T^2) exceeds their respective limits, it can be considered as a faulty event. If the estimated statistics are under defined limits, then there is no fault and the monitoring process is continued.

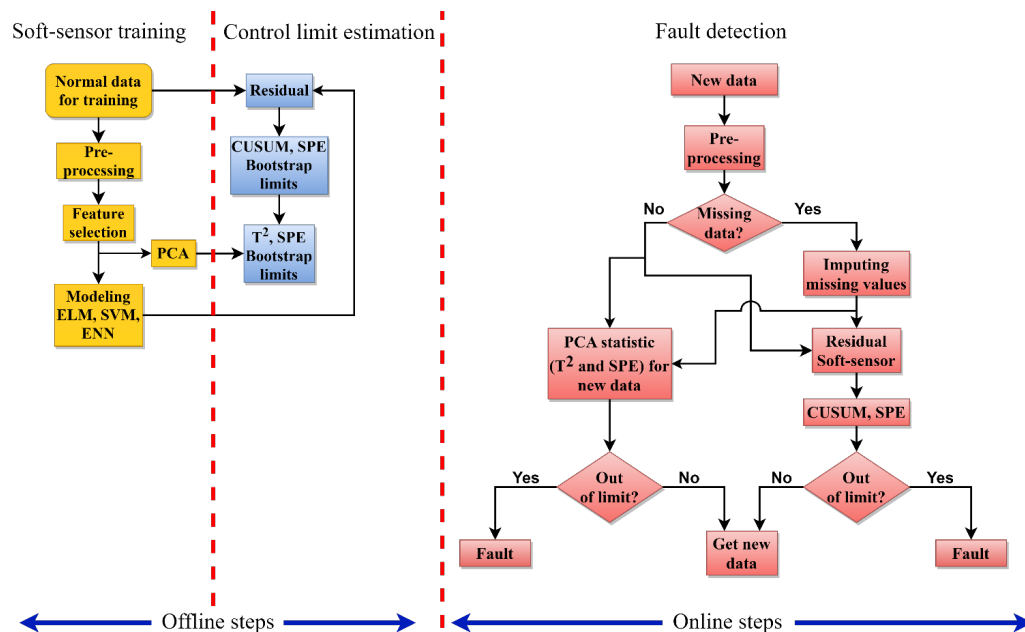


Figure 5. 1 General diagram of the proposed FD framework.

5.6. Developing soft-sensors for FD

The design and evaluation of the soft-sensors have been carried out by collecting data from BSM2 (Jeppsson et al., 2006). Similar to Chapter 4, to increase the nonlinearity and make the objective more challenging, the input vector (feed) to AD, which consists of the concentration of composite (X_c), inorganic nitrogen (S_{in}) and carbohydrate (X_{ch}) and also the feed flow rate, was manipulated.

As mentioned earlier, using redundant and unnecessary features may add noise to the model and increase the risk of over-fitting. Therefore, *fscaret* feature ranking technique was used (Szlęk et al., 2016). Prior to the feature ranking, hard to measure variables such as BOD, COD and alkalinity were removed from the data set. The CH_4 mole fraction and gas flow were also removed due to their direct correlation with CO_2 mole fraction and pressure, respectively. Finally, the pH, ammonia concentration, CO_2 mole fraction and pressure were considered as the most influential

variables for developing VFA soft-sensors. Figure 5.2 shows their trends over the total period of experiments. Before training soft-sensors, the data set was split into training and validation partitions. As can be seen in Figure 5.2, these two splits are divided by a red vertical line; the left side was used for training and the other side was used for validating the soft-sensors. Due to a large number of training data, uniform sampling was performed to enhance the soft-sensor training speed.

The whole training procedure was performed in R environment by using *Caret* package (Kuhn, 2008). All models were tuned using random hyperparameter *Search* which is incorporated in this package. In this method, the tuning parameters of all models are randomly selected from the tuning space which is defined beforehand. The number of randomly sampled values from the tuning space can be explained by “*tuneLength*” parameter of the *Caret* package (Bergstra et al., 2012). The tuning parameters for each model are:

- SVM with a radial basis kernel. For this method, using the random search of the *Caret* package leads to values of 245.88 and 0.0020 for the parameters C and γ , respectively (Cortes and Vapnik, 1995).
- For ELM, the transfer function was sinusoidal and the number of hidden neurons was 114.
- For ENN, one hidden layer with 16 neurons gave the best performance. The number of ensemble network was fixed at 10.

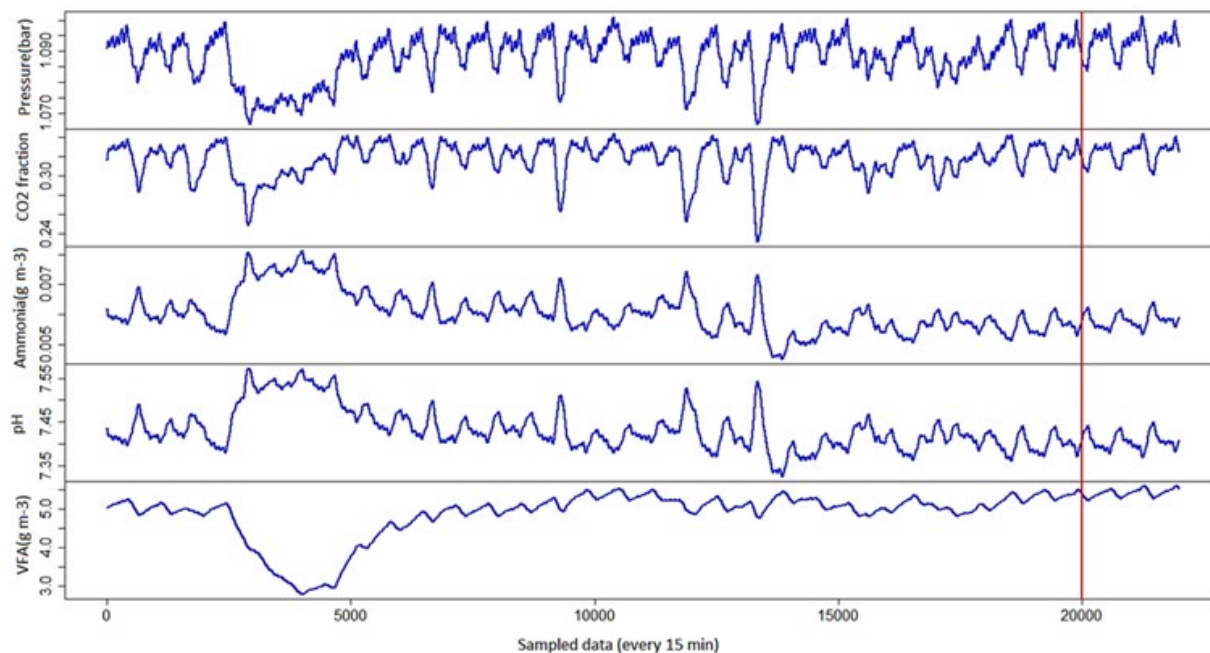


Figure 5. 2 Trend of input and output variables used for developing soft-sensors. Left side of the red line used for training and the right side used for testing.

As previously mentioned, different models such as SVM, ELM and ENN were used to design VFA soft-sensors. It should be noted that the developed soft-sensors predict the current amount of VFA inside the digester. After training all models, the NRMSE and R^2 are estimated based on the training and validation set for each model. The obtained results using different techniques are presented in Table 5.1.

The lowest validation error was achieved for the model developed with ELM and ENN algorithms. The SVM algorithm gave a slightly higher error, but still these were satisfactory results. For training the PCA model, all the input variables for training soft-sensors plus VFA were used.

Table 5. 1 Results of soft-sensors on the training and validation set.

Algorithm	Training		Validation	
	NRMSE	R ²	NRMSE	R ²
SVM	0.014	0.998	0.041	0.983
ELM	0.008	0.999	0.010	0.999
ENN	0.008	0.999	0.010	0.999

5.7. Fault detection and discussion

In order to check the detection accuracy of the proposed methods, two types of faults including faults in the state parameters of AD and sensor faults were studied. The ammonia inhibition constant (k_{I,NH_3}) in BSM2 was varied from $\pm 5\%$ to $\pm 15\%$ around its default value ($0.0018 \text{ kmol.m}^{-3}$) as a simulated artificial fault in the state parameter. The variation of k_{I,NH_3} can be similar to manipulating of ammonia concentration inside the digester. Moreover, three different faults including bias, drift and fixed value were simulated for pH sensor as sensor faults (see Table 5.2). The data from day 453 to 530 were defined as the normal data set and used for estimating the control limits of the entire charts. The fault starts from day 530 and lasts until the end of the simulation (609 days).

Table 5. 2 Faults in BSM2 simulation.

Fault#	Description
1 to 6	Fault in AD state parameter k_{I,NH_3} , from -15% to 15% (5% step)
7	pH sensor drift (0.00006 drift speed)
8	pH sensor Bias (+0.1 bias)
9	pH sensor fixed value

In this chapter, for comparison reasons, three different approaches were used for developing VFA soft-sensor. Therefore, initially, different trained soft-sensors will be compared in terms of

robustness with respect to faulty events and after that, the most robust one will be chosen for further discussion. The robustness can be defined as the soft-sensor ability to predict the value of VFA in abnormal events when there is a fault in the process. In the abnormal events, a robust soft-sensor predicts the value of VFA as if the process was in normal operation; therefore, the difference between the measured and the predicted VFA (residual) can be easily considered as a fault (Baraldi et al., 2015). The robustness with respect to a fault i can be estimated by the following equation:

$$S_i = \frac{\sum_{k=1}^N (VFA_{pred}^{fa} - VFA_{real}^{no})^2}{N} \quad (5.8)$$

where S_i , VFA_{pred}^{fa} , VFA_{real}^{no} and N are the robustness with respect to fault i , VFA predicted by soft-sensors during the faulty event, the real VFA in normal operation of the process and number of all faulty samples, respectively. A global robustness measure over all faults (m) can be constructed as follow:

$$S = \sum_{i=1}^m S_i \quad (5.9)$$

Therefore, a low value of S means high robustness.

As variations of $k_{I,NH3}$ affect both predicted VFA by soft-sensors and the real output of the process, the manipulation of $k_{I,NH3}$ is considered as different faulty events for soft-sensors robustness calculation. Table 5.3 shows the results in terms of the soft-sensors robustness for different magnitude of faults and the global robustness with respect to all faults.

Table 5. 3 Robustness of SVM, ELM and ENN in the prediction of VFA affected by k_{L,NH_3} changes.

Fault magnitude	Soft-Sensors Robustness		
	SVM	ELM	ENN
k_{L,NH_3} (+5%)	0.088	0.105	0.105
k_{L,NH_3} (-5%)	0.083	0.115	0.105
k_{L,NH_3} (+10%)	0.342	0.413	0.409
k_{L,NH_3} (-10%)	0.336	0.453	0.437
k_{L,NH_3} (+15%)	0.758	0.903	0.906
k_{L,NH_3} (-15%)	0.758	0.950	1.021
Global robustness	2.365	2.939	2.983

From Table 5.3, ELM and ENN soft-sensors are the worst-performing, whereas SVM is robust with respect to k_{L,NH_3} variations. The obtained results confirm that ELM and ENN have excellent robustness in terms of predicting VFA; however, they are not as robust as SVM toward detection of faults. Therefore, SVM was chosen for further discussion and FD.

Figure 5.3 illustrates the density estimate and probability plots of the residual obtained by

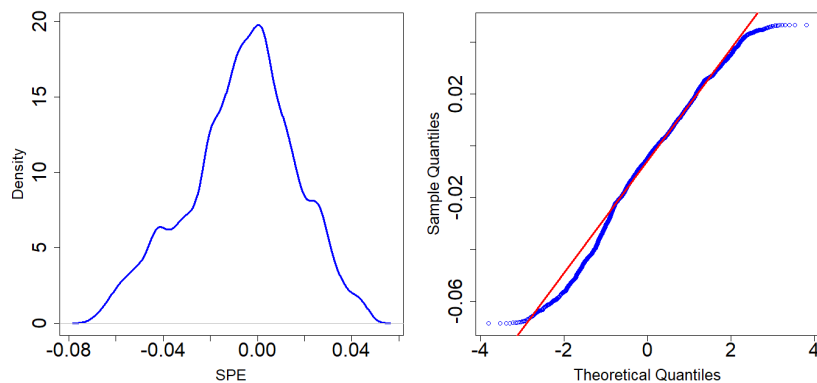


Figure 5. 3 Density estimation and probability plots of soft-sensor residual for normal operation.

applying SVM soft-sensor to the normal operating data collected during normal operation of BSM2. It is obvious that the residual does not follow normal distribution. Therefore, applying the normality assumption for estimating the control limits may cause a false result when used for

monitoring. Due to this reason, the control limits for all statistics were determined using the bootstrap method.

5.7.1. Fault diagnosis without missing values

In this section, step three of the proposed approach is discussed. The FD performance has been compared with PCA method. In the PCA model, four PCs were selected to explain 90% of variance and its T^2 and SPE control limit obtained theoretically using the bootstrap method with confidence level α equal to 0.99. The FD performance for different faults is presented in Table 5.4. The chart that has the highest F1 score considered as the most precise one. Comparing average F1 scores shows that CUSUM chart based on VFA residual has the best performance among the other charts (Table 5.4). Although PCA- T^2 chart has the highest precision, it achieved the worst F1 score due to the low average value of recall. In comparison, SPE statistic obtained from VFA soft-sensor performs better than the SPE obtained by PCA method. All models showed a high value of precision which means that all charts have a low false detection alarm rate.

Table 5. 4 Performance of FD methods for different faults (the meaning of the fault number is presented in Table 5.2).

Fault#	CUSUM statistic (VFA residual)			SPE (VFA residual)			PCA- T^2			PCA-SPE		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	0.979	0.941	0.960	0.987	0.904	0.944	1	0.789	0.882	0.989	0.740	0.847
2	0.978	0.906	0.941	0.990	0.691	0.814	1	0.268	0.423	0.983	0.480	0.646
3	0.972	0.682	0.802	0.977	0.306	0.466	1	0.020	0.040	0.969	0.254	0.403
4	0.979	0.945	0.962	0.947	0.124	0.219	1	0.019	0.037	0.738	0.022	0.043
5	0.980	0.962	0.971	0.986	0.523	0.683	1	0.057	0.108	0.857	0.047	0.089
6	0.980	0.970	0.975	0.992	0.885	0.935	1	0.253	0.404	0.941	0.126	0.222
7	0.971	0.674	0.796	0.990	0.739	0.846	1	0.012	0.024	0.959	0.186	0.311
8	0.972	1	0.985	0.980	1	0.989	1	0.038	0.074	0.979	1	0.989

9	0.976	1	0.988	0.888	0.973	0.928	1	0.003	0.007	0.982	0.931	0.956
Av.	0.976	0.897	0.931	0.970	0.682	0.758	1	0.162	0.222	0.933	0.420	0.500

P: Precision, R: Recall, AV.: Average

Figure 5.4 shows the FD performance for positive variations of k_{I,NH_3} among different charts for the complete data set (similar results were obtained for negative variations). The horizontal and vertical blue lines show the obtained confidence limits and the fault onsets, respectively. The confidence limits of each plot are obtained by the bootstrap method as discussed previously.

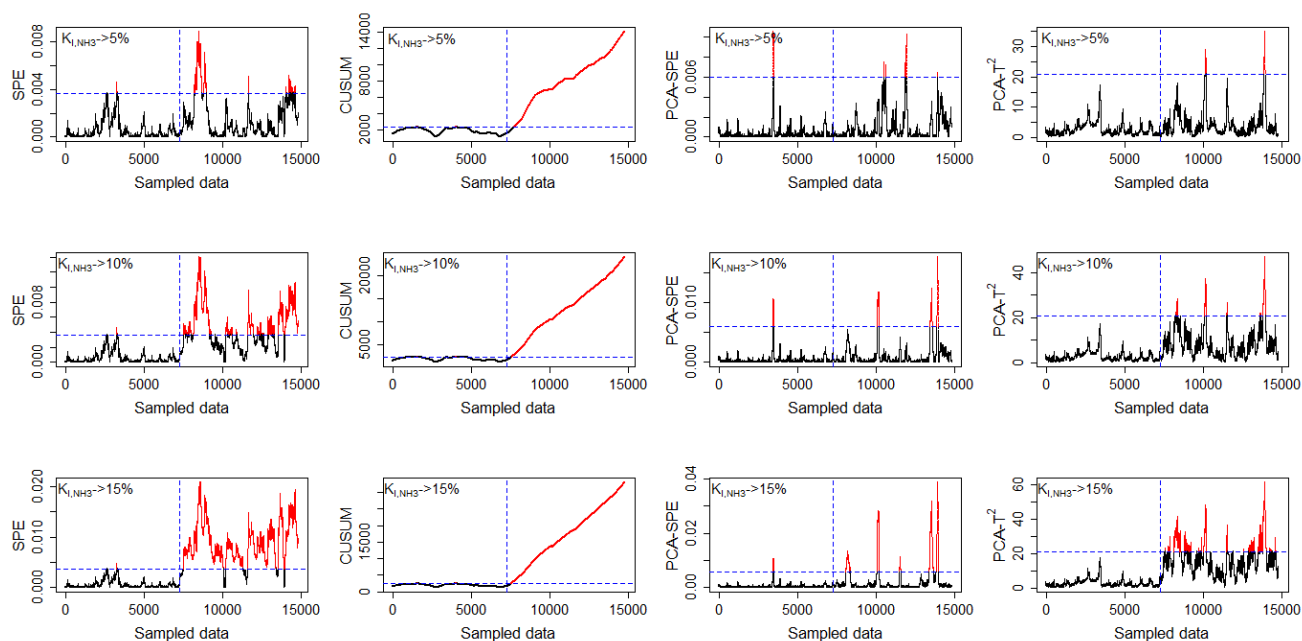


Figure 5. 4 FD performance for variation of k_{I,NH_3} for the complete data set

As it can be seen, the proposed approach based on VFA soft-sensor has superior performance compared to the PCA method. By increasing the magnitude of the fault, the performance of the charts also improved. From Figure 5.4, it can be concluded that SPE chart performs better on larger magnitude faults, whereas CUSUM chart can perform better on smaller magnitude faults.

The detection delay obtained by each method is shown in Table 5.5. A comparison of charts delay showed that due to the small magnitude of faults, all charts roughly have a high delay. This is due to the fact that the magnitude of the faults is very small; therefore, their impact on the whole process and the measured variables is not significant. The delay of SPE chart is the lowest one; however, due to its fluctuation around the control limit, it is not clear whether it is a real detection or a false alarm (Figure 5.4). The delay of CUSUM chart is due to its cumulative behavior; therefore, it takes some time for the signal to pass the control limit, but as it has an upward trend, it can definitely be considered as a fault. The performance of the proposed control charts for faults in pH sensor is outlined in Figure 5.5.

Table 5. 5 Detection delays (in samples) for different charts (the meaning of the fault number is presented in Table 2).

Fault#	SPE (VFA residual)	Cusum statistic (VFA residual)	PCA-T ²	PCA-SPE
1	1	439	20	40
2	324	704	1379	1224
3	1840	2375	2814	1819
4	861	407	2821	3168
5	187	280	954	2744
6	165	224	759	278
7	1841	2439	2828	3070
8	1	1	2758	1
9	1	1	2817	1
Average	580.11	763.33	1905.55	1371.66

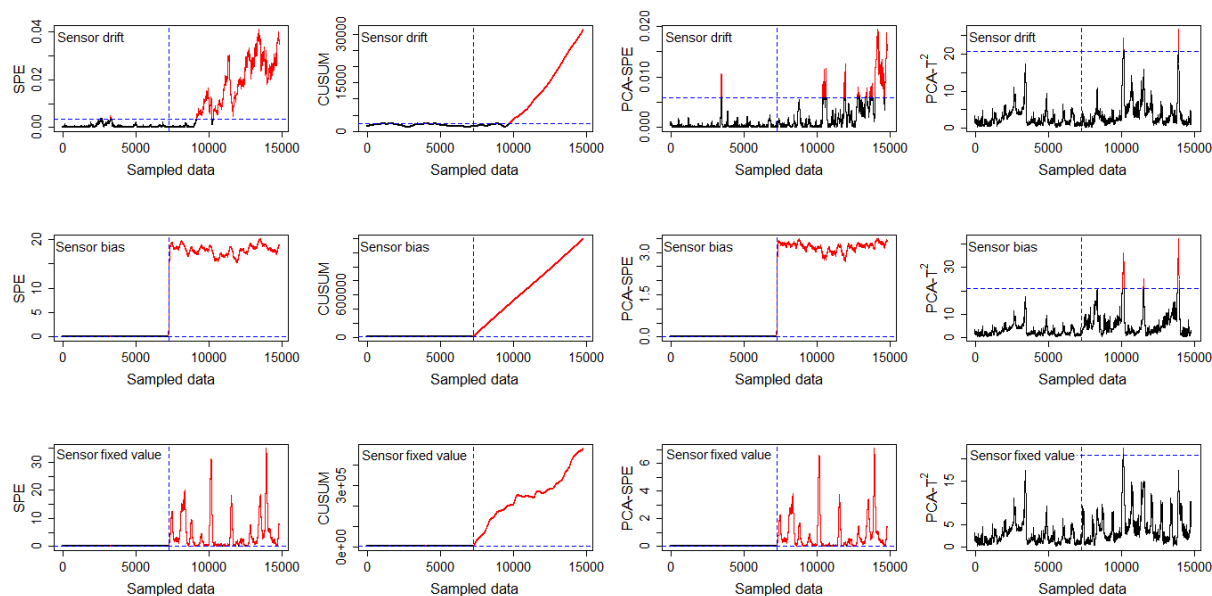


Figure 5.5 FD performance for pH sensor faults among different charts for complete data

As it can be seen, SPE, CUSUM and PCA-SPE charts perform well on the different sensors faults, while none of the faults can be captured by PCA-T² which reveals that the faults occurred in the residual space rather than PCs space. This is because the variations of variables, such as sensor faults, tend to be captured by PCA-SPE, whereas the variations of process condition, such as faults in state parameters of processes, are usually captured by PCA-T² chart (Yoo et al., 2004). Due to the small drift speed (0.00006 per 15 minutes), it takes some times for the charts to detect the drift fault. However, from Table 5.5, SPE chart has less delay and performs better than the other ones. While the CUSUM chart shows better performance on small magnitude faults, SPE chart is more appropriate for larger faults magnitude due to inherent CUSUM delay. Therefore, to enjoy the advantages of both methods, it would be optimal to use both charts simultaneously.

5.7.2. Fault diagnosis with missing values

Most monitoring systems fail in case of missing data. Therefore, to have a robust and accurate FD system, good sensor data is needed. However, missing values due to sensor failure and maintenance activities are inevitable. Thus, it is mandatory to examine the sensitivity of FD approaches during missing sensors signal events. To do this, 15% of the signal is randomly eliminated under k_{I,NH_3} faults. To impute missing values, the last value is used, simulating a constant measurement. Figure 5.6 shows the performance of different charts during k_{I,NH_3} faults with imputed missing values. As can be seen, SPE chart is very sensitive compared to the other ones. PCA method is less sensitive; however, the detection performance of this method is very low. On the contrary, CUSUM chart outperformed compared to the other charts in detection and sensitivity to the imputed missing values. The better performance of CUSUM chart in case of missing data is due to its cumulative behavior that makes the CUSUM statistics for each point depending on the previous points. Therefore, even with some missing points in the data set, the performance has not deteriorated.

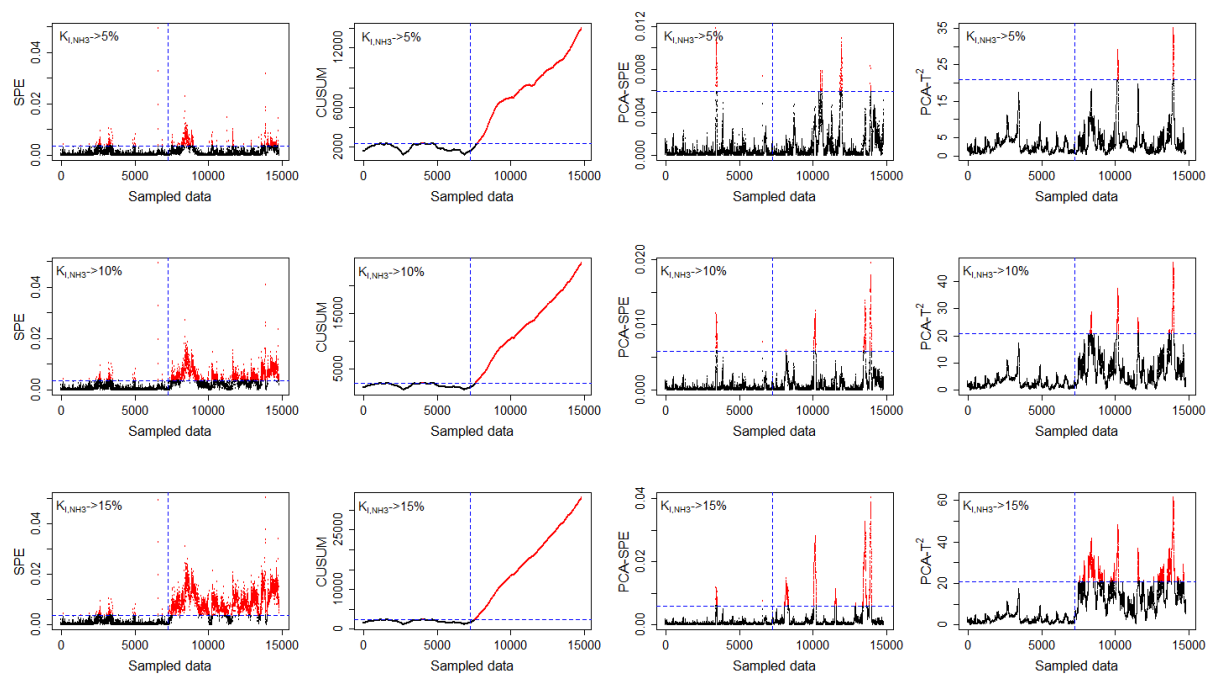


Figure 5. 6 FD performance for variation of k_{I,NH_3} among different charts for uncompleted data.

5.8. Conclusion

In this chapter, a framework for detecting random faults in AD processes was successfully developed based on VFA soft-sensors which allow early detection of faults. In this regard, different data-driven soft-sensors such as ELM, SVM and ENN were trained and compared in terms of accuracy and FD robustness by applying a simulated data set obtained from BSM2. Prior to soft-sensor design, the most appropriate subset of input variables was found by using feature selection method. Ammonia concentration, pH, pressure and CO_2 mole fraction were selected as the best subset of input variables. Each soft-sensor was examined in terms of robustness to the fault and it was found that although ELM and ENN methods have shown high accuracy in predicting VFA during normal operating conditions, they could not be effectively used for FD purposes because they are not robust enough. On the other hand, SVM is satisfactory from a

robustness point of view. The SVM soft-sensor was applied to the data obtained by BSM2 to generate a residual signal. Examining the residual signal suggests that it follows a non-normal distribution. Due to this reason, non-parametric bootstrap method was used for estimating the control limit for the entire charts. This residual signal was applied to univariate charts such as SPE and CUSUM charts to detect the faults. The simulation results using changes in the k_{I,NH_3} parameter showed that residual signals applied to the CUSUM and SPE charts have better detection performances than PCA.

Moreover, CUSUM chart performs better for small magnitude faults, while SPE is superior in higher magnitude faults due to the inherent CUSUM delay. For pH sensor faults generally, all charts performed well; nonetheless, for small drift faults, SPE chart detection is faster among the other charts. A study on the data set with missing values suggests that CUSUM chart is very robust in FD during missing signals events. Finally, it can be concluded that due to the distinct performance of CUSUM and SPE charts in different faulty events, they can be used simultaneously within the same FD framework.

CHAPTER 6

Fault detection and diagnosis in water resource recovery facilities using incremental PCA

Because of the static nature of conventional PCA, natural process variations may be interpreted as faults when it is applied to processes with time-varying behavior. In this chapter, therefore, we propose a complete adaptive process monitoring framework based on IPCA. This framework updates the eigenspace by incrementing new data to the PCA at a low computational cost. Moreover, the contribution of variables is recursively provided using CDC. To impute missing values, the EBLUP method is incorporated into this framework. The effectiveness of this framework is evaluated using BSM2.

6.1. Introduction

Because of the complexity of modern industrial processes, the demand has increased to implement FD and a diagnosis framework for those processes. WWTPs, are no exception (Sánchez-Fernández et al., 2018). Abnormal or faulty events are those that occur when the process indicates a deviation from its normal behavior. In WWTP, abnormal events include changes in influent quality (e.g. rainfall, industrial discharge), outbreaks of microorganisms (e.g. filamentous bacteria, algae) that impact treatment quality, irregularities or damage to treatment units (e.g. membranes, clarifiers), mechanical failures (e.g. pumps, air blowers) and sensor failure (e.g. drift, bias, electrical interference) (Newhart et al., 2019). All these faults can undermine process performance, so they must be detected and diagnosed as soon as they occur. FD techniques are generally divided into three main groups: model-based methods, knowledge-based methods, and data-driven methods. To implement techniques in the first two groups, an in-depth knowledge of the system's behavior is required. However, obtaining in-depth knowledge of reactions or pathways phenomena (especially for WWTPs, which are very complicated processes), is both time-consuming and challenging.

Data-driven methods, on the other hand, rely only on historical and online data and do not require in-depth prior knowledge of the process (Kazemi et al., 2020). Thanks to the advanced process control framework, WWTPs collect large amounts of data using measurements from numerous online sensors (Wang and Shi, 2010). Data-driven techniques have therefore received significant attention in process monitoring over the last few years. Multivariate statistical process controls (MSPCs) are a subset of data-driven methods used to analyze the performance of the processes and identify the parameters that govern them. One of the most widely used techniques

for MSPC is PCA (Newhart et al., 2019). Various types of PCA methods, such as conventional PCA (Garcia-Alvarez et al., 2009; Sanchez-Fernández et al., 2015; Xiao et al., 2017), kPCA (Jun et al., 2006; Lee et al., 2004; Xiao et al., 2017) and dynamic PCA (C. Lee et al., 2006; Mina and Verde, 2006) have been successfully used to monitor processes and detect faults in time-invariant processes. However, because of the complexity of the biological reactions and non-stationary plant influent, WWTPs have time-varying characteristics. Using conventional PCA to monitor such processes may therefore lead to excessive rates of false alarms and missing detections. An adaptive process monitoring framework is therefore needed. Although adaptive PCAs are more suitable for non-stationary processes, due to high computational costs not every adaptive approach is applicable for realtime FD. Several adaptive approaches include recursive PCA (RPCA), moving window PCA and EWMA have recently been used for FD (Choi et al., 2006; Elshenawy et al., 2010; Li et al., 2000; Rosen and Lennox, 2001; Shang et al., 2015). Haimi et al. implemented a moving window PCA technique to detect and isolate the faults in WWTP (Haimi et al., 2016). These authors took into account the variable length of the historical data in the model's construction to prevent sub-optimal monitoring performance. Li et al. applied an RPCA monitoring framework that recursively updates the correlation matrix. These authors used two approaches, namely rank-one modification and Lanczos tridiagonalization, to compute the eigenvalues of the updated correlation matrix (Li et al., 2000). Recently, Elshenawy et al. proposed RPCA based on first-order perturbation analysis (FOP), which is a rank-one update of the eigenvalues and their corresponding eigenvectors from a sample covariance matrix (Elshenawy et al., 2010; Elshenawy and Mahmoud, 2018). These authors stated that the computational cost of their approach is lower than that of previously used methods such as RPCA using Lanczos tridiagonalization and moving window PCA (MWPCA).

In this chapter we propose a new low-computational-cost adaptive FD framework for WWTP that uses IPCA. This approach is based on the IPCA proposed by Hall et al., which was motivated by developments in the field of computer vision (Arora et al., 2012; Brand, 2002; Hall et al., 1998). The main benefit of IPCA over other adaptive PCA approaches is that it does not require all eigenvalues to be computed since only the largest ones are needed. This considerably accelerates computation time. Cardot et al. (Cardot and Degras, 2018) compared the computation time of adaptive PCA methods such as perturbation and stochastic approximation and concluded that IPCA outperforms other methods, particularly when the data have many features (sensor measurements). These authors also studied the accuracy of adaptive methods in determining the eigenspace. Their results revealed that IPCA is more accurate than methods such as RPCA, which is based on FOP theory. The IPCA algorithms offer an excellent compromise between statistical accuracy and computational speed. Another major drawback of PCA approaches are missing data during realtime monitoring. In practice, due to sudden mechanical breakdown, maintenance, hardware sensor failure or malfunction of the data acquisition system, etc., some sensor signals may become unavailable (Zhang and Dong, 2014). To meet the requirements of realtime FD and diagnosis, therefore, the problem of missing data needs to be considered. To handle missing data, the EBLUP approach is incorporated into the proposed IPCA method (Cardot and Degras, 2018).

Once a fault is detected, it is essential to find out its primary source. The most common fault isolation method are contribution charts (Nomikos and MacGregor, 1995). With this method, the variables that contribute to the T^2 (Hotelling's T-squared) and SPE statistics are calculated and the variable with the highest contribution is considered the primary cause of the fault. Because of the recursive nature of IPCA, the contribution plots of T^2 and SPE are updated in accordance with the adaptive eigendecomposition. To test the proposed framework, several types of process

parameters and sensor faults were simulated using BSM2 (Jeppsson et al., 2006). This chapter is organized as follows: in section 6.2 we provide a brief review of the theoretical background behind the conventional PCA method; in sections 6.3 and 6.4 we discuss the IPCA algorithm and its realtime application framework for FD and diagnosis; and in sections 6.5 and 6.6 we analyze the performance of the monitoring framework on the complete and incomplete (i.e. with missing values) data sets using BSM2.

6.2. Conventional PCA

PCA is a well-known method designed to convert a data set with possibly correlated variables into a set of values of linearly uncorrelated variables called PCs (Xiao et al., 2017). Let a data matrix $X \in \mathcal{R}^{n \times m}$ be composed of n samples and m sensors. The X matrix is scaled to zero mean and unit variance (Z-score normalization) in order to avoid the scaling problem. The PCA model can be defined as:

$$X = TP^T + E \quad (6.1)$$

where $T \in \mathcal{R}^{n \times m}$, $P \in \mathcal{R}^{m \times m}$, and $E \in \mathcal{R}^{n \times m}$ are score (PCs), loading, and residual matrices, respectively. The scores are generated by projecting X onto to the loading matrix. To solve the PCA model the covariance matrix, $S \in \mathcal{R}^{m \times m}$, of X should be decomposed as follows:

$$S = \frac{1}{n-1} X^T X = P \Lambda P \quad (6.2)$$

The columns of P are eigenvectors of S , and $\Lambda \in \mathcal{R}^{m \times m}$ is a diagonal matrix containing the eigenvalues $\Lambda = \text{diag}(\lambda_1, \lambda_2 \dots \lambda_m)$ arranged in descending order. Usually, the number of (β) principal components, which sufficiently explain the variability of the process, are selected.

Therefore, $P \in \mathbb{R}^{m \times \beta}$ and its corresponding eigenvalues are $\Lambda \in \mathbb{R}^{\beta \times \beta}$. For FD, two statistics (mainly T^2 and SPE) are normally used. Variations in the mean and covariance in PCs are measured using T^2 while the variation in the residual subspace is measured using SPE (Elshenawy and Awad, 2012). These are given by:

$$T_k^2 = x_k^T P_{k(\beta_k)} \Lambda_{k(\beta_k)}^{-1} x_k P_{k(\beta_k)}^T \quad k = 1 \dots n \quad (6.3)$$

$$SPE_k = x_k (1 - P_{k(\beta_k)} P_{k(\beta_k)}^T)^2 x_k^T \quad (6.4)$$

Here, as mentioned earlier, β_k is the number of PCs to retain. To perform FD, suitable thresholds for these two statistical indices must be obtained. This enables small faults to be detected with a minimum rate of false alarms. Process operation is normal if these statistical indices remain below these thresholds. The procedure for estimating the thresholds of T^2 and SPE will be discussed later.

6.3. Incremental PCA

Detecting and monitoring faults using conventional PCA is suitable for time-invariant processes. If it were used for time-variant processes, it would be difficult to monitor the typical time-varying characteristics caused by external disturbances and changes in operating conditions. To handle time-varying issues, an adaptive FD framework must be developed (Hu et al., 2012). The key idea behind IPCA is to update the eigenvector by incrementing the new data to the PCA. To update process monitoring, both the mean $m_k \in \mathbb{R}^{1 \times m}$ and the standard deviation $\delta_k \in \mathbb{R}^{1 \times m}$ need to be re-estimated whenever a new sample $x_k \in \mathbb{R}^{1 \times m}$ data vector becomes available (Artač et al., 2002; Cardot and Degras, 2018; Hall et al., 1998). These are given by

$$m_k = (1 - f)m_{k-1} + fx_k \quad (6.5)$$

$$\delta_k = \sqrt{(1 - f)\delta_{k-1}^2 + f(x_k - m_k)^2} \quad (6.6)$$

where $0 \leq f < 1$ is a forgetting factor. Let us assume that eigenvector P_{k-1} has already been derived using the data from X . Also, Λ_{k-1} is a diagonal matrix of the corresponding eigenvalues λ , and m_k and δ_k are the mean and variance vector of a new sample x_k , respectively. Each sample must be standardized according to:

$$x'_k = \frac{x_k - m_k}{\delta_k} \quad (6.7)$$

where x'_k is the standardized sample. Next, the projection of the new sample a_k is computed using the current eigenvector P_{k-1} :

$$a_k = P_{k-1}^T x'_k \quad (6.8)$$

Finally, residual vector h_k is estimated using the feature vector a_k . The residual vector is orthogonal to the eigenvector.

$$h_k = P_{k-1} a_k - x'_k \quad (6.9)$$

To update the eigenvector, h_k needs to be normalized according to the condition outlined in the equation below:

$$\hat{h}_k = \begin{cases} \frac{h_k}{\|h_k\|_2}, & \text{if } \|h_k\|_2 \neq 0 \\ 0, & \text{otherwise} \end{cases}, \quad \|h_k\|_2 = \sqrt{h_k^T \cdot h_k} \quad (6.10)$$

where $\|h_k\|_2$ is the Euclidean norm of matrix h_k .

The new eigenvector P_k can be estimated by adding \hat{h}_k to the current eigenvector P_{k-1} and rotating the result using rotation matrix R :

$$P_k = [P_{k-1}, \hat{h}_k] R_k \quad (6.11)$$

R is calculated by solving the eigenvector decomposition problem as shown below:

$$D_k R_k = R_k \Lambda_k \quad (6.12)$$

where matrix D_k is composed as follow:

$$D_k = (1 - f) \begin{bmatrix} \Lambda_{k-1} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix} + f(1 - f) \begin{bmatrix} a_k a_k^T & \gamma a_k \\ \gamma a_k^T & \gamma^2 \end{bmatrix} \quad (6.13)$$

where $\mathbf{0}$ is a vector of zero and $\gamma = \hat{h}_k x'_k$. By updating the eigenvector P_k and the eigenvalue matrix Λ_k for each new sample, the process monitoring statistics must also be updated according to Equation 6.3 and Equation 6.4.

6.3.1. Estimating the number of PCs

To fully implement IPCA, the number of PCs must be updated after a new sample becomes available. Numerous methods are available for calculating the number of PCs (Li et al., 2000). In this chapter we used cumulative percent variance (CPV), which can be obtained from:

$$CPV(\beta_k) = \frac{\sum_{j=1}^{\beta_k} \lambda_j}{\sum_{j=1}^m \lambda_j} 100\% \quad (6.14)$$

where β_k is a number of selected PCs. The number of PCs is chosen when CPV reaches a predetermined limit, e.g. 95%.

6.3.2. Process monitoring statistics thresholds

Because of the non-stationary behavior of water resource recovery facilities, the thresholds for the detection statistics (T^2 and SPE) change over time. For realtime monitoring, therefore, it is necessary to adapt these limits (Li et al., 2000). The T^2 threshold σ_k is approximated using the chi-distribution χ^2 with β_k degrees of freedom and a confidence limit α :

$$\sigma_k = \chi_{\alpha, \beta_k}^2 \quad (6.15)$$

and the threshold γ_k for the SPE statistic is given by:

$$\gamma_k = \theta_1 \left[\frac{\eta_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{h_0}} \quad (6.16)$$

where η_α is the normal deviate corresponding to the $(1-\alpha)$ percentile, and

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2} \quad (6.17)$$

$$\theta_i = \sum_{j=\beta_k+1}^m \lambda_j^i; \quad i = 1, 2, 3 \quad (6.18)$$

The fault is detected once the T^2 and SPE statistics exceed their respective adaptive thresholds, σ_k and γ_k .

6.3.3 Fault Isolation

Once a fault has been detected, the variables that contribute to the deviation must be identified in order to determine the location of primary causes. The idea behind the contribution chart is that variables with the largest contributions to the FD statistics are probably the faulty variables (Alcala

and Qin, 2009). Many methods are available for calculating the contributions of variables. We used CDC, which is widely used in industry. The CDC for the T^2 and SPE statistics is defined as:

$$CDC_k^{T^2} = (x_k P_{k(\beta_k)} \Lambda_{k(\beta_k)}^{-1/2} P_{k(\beta_k)}^T)^2 \quad (6.19)$$

$$CDC_k^{SPE} = (x_k - (x_k P_{k(\beta_k)} P_{k(\beta_k)}^T))^2 \quad (6.20)$$

where $CDC_k^{T^2}$ and CDC_k^{SPE} are the vectors whose elements are the variables that contribute to the T^2 and SPE statistics for each column (sensor) of a sample, respectively.

6.3.4. Missing data imputation

Methods such as mean, regression, hot-deck, maximum likelihood, multiple imputations, etc. are suggested for imputing missing values in the realtime application of PCA (Cardot and Degras, 2018). In this chapter we used the EBLUP method described by Brand (2002) to impute missing values in the new observation vector x_k . With this method, the missing observations in vector x_k are approximated using the mean value m_{k-1} and the eigenvector decomposition (P_{k-1} and Λ_{k-1}) of the current sample. The x_k can be split into two sub-vectors: x_k^o (observed values) and x_k^m (missing values), respectively. Similarly, m_{k-1} and P_{k-1} are split into m_{k-1}^o , m_{k-1}^m and p_{k-1}^o , p_{k-1}^m respectively. Let $\Lambda_{k-1}^{-1/2}$ be the diagonal matrix containing the square roots of the diagonal eigenvalues arranged in descending order (Brand, 2002; Cardot and Degras, 2018). By applying the conditional expectation for multivariate normal distribution, the EBLUP can be obtained as follows:

$$\hat{x}_k^m = \mathbb{E}(x_k^m | x_k^o, m_{k-1}, P_{k-1}, \Lambda_{k-1}) = m_{k-1}^m + (p_{k-1}^m \Lambda_{k-1}^{-1/2}) (p_{k-1}^o \Lambda_{k-1}^{-1/2}) (x_k^o - m_{k-1}^o) \quad (6.21)$$

where \hat{x}_k^m is the imputed missing value.

6.4. Adaptive Fault Detection and Isolation

In this section we discuss the complete structure of the adaptive FD and isolation scheme. The proposed adaptive FD and isolation framework can be divided into two stages: (i) offline training and (ii) online monitoring. The following steps summarize the overall procedure.

Offline Training

- (1) Collect the training data matrix $X \in \mathcal{R}^{n \times m}$ and normalize it to zero mean m_{k-1} and unit variance δ_{k-1} .
- (2) Use Equation 6.2 to compute S and its corresponding eigenpairs P_{k-1} , Λ_{k-1} .
- (3) Calculate the number of retaining PCs (β_{k-1}) by applying Equation 6.14.
- (4) Calculate the thresholds of the FD statistics σ_{k-1} and γ_{k-1} using Equations 6.15 and 6.16.

Online Monitoring

- (1) Collect a new data vector x_k and normalize it to zero mean and unit variance using Equations 6.5, 6.6, and 6.7. If there are any missing values, compute them using Equation 6.21.
- (2) Calculate the monitoring statistics T_k^2 and SPE_k using Equations 6.3 and 6.4.
- (3) If the values of the monitoring statistics are below their corresponding thresholds, the process status is normal. The updating procedure continues as follows:
 - (i.) Update m_k and δ_k using Equations 6.5 and 6.6.

- (ii.) Calculate the updated eigenpairs P_k and Λ_k using Equations 6.8-6.13.
 - (iii.) Calculate the number of retained PCs, β_k , using Equation 6.14.
 - (iv.) Update the monitoring statistics thresholds σ_k and γ_k using Equations 6.15 and 6.16.
 - (v.) Return to step one.
- (4) If the values of the monitoring statistics exceed their corresponding thresholds, the process status is faulty and the updating procedure should be stopped. The fault isolation procedure (CDC) needs to be run to identify the process variables responsible for the detected fault.
- (i.) The contribution statistics for the faulty points need to be calculated according to Equations 6.19 and 6.20.
 - (ii.) The process variables with the largest contributions are responsible for the fault.

The complete diagram for the proposed adaptive IPCA monitoring procedure is shown in Figure 6.1.

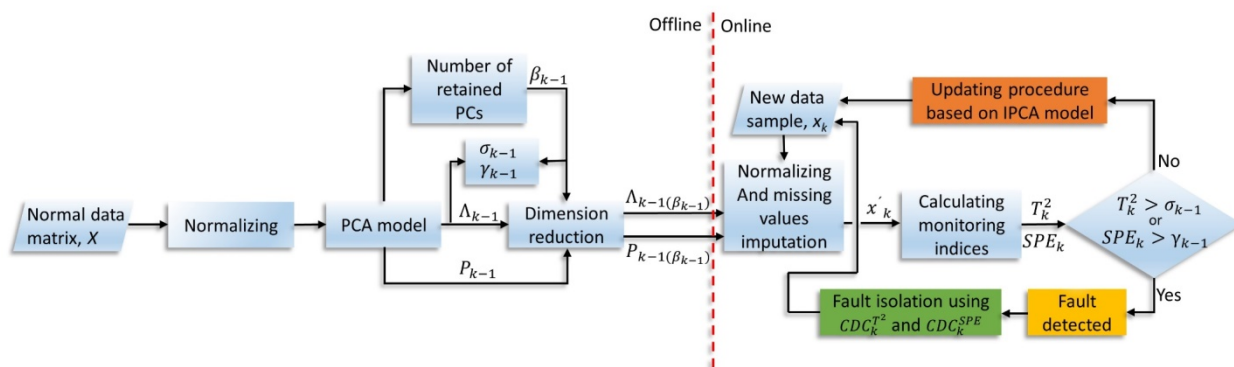


Figure 6. 1 Adaptive IPCA monitoring diagram.

6.5. Simulation results

6.5.1. Description of the Water resource recovery facility

In this section the proposed adaptive IPCA framework is applied to BSM2 (Jeppsson et al., 2006; Nopens et al., 2010).

6.5.2. Fault detection and isolation

BSM2 simulation was modified to obtain sensor data from various parts of the process by simulating different types of fault. By performing this modification, 320 data measurements (16 state variables \times 20 measurement points) can be obtained. However, measuring all these variables is not common practice in a real WWTP. In this chapter, therefore, we considered only realistic and commonly available sensor measurements. Twenty-eight process measurements obtained from various parts of BSM2 simulation (Figure 6.2) were recorded every 15 min as the inputs for the proposed monitoring framework. All these variables were corrupted with white noise to simulate real sensor measurements.

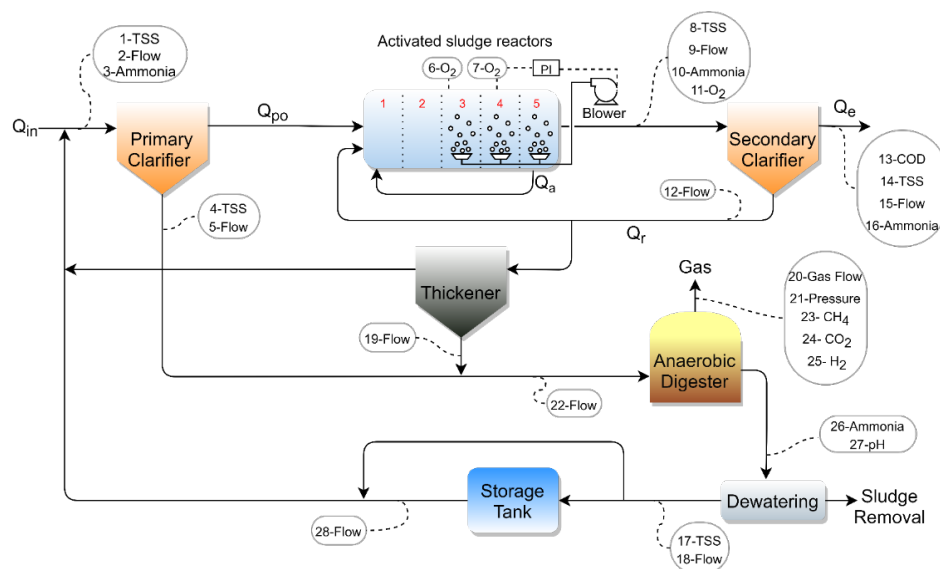


Figure 6. 2 Schematic diagram of BSM2 and locations of the measured variables.

To build the offline adaptive IPCA process monitoring framework, 21 days of data measurements (day 435 to day 456) were recorded under normal operating conditions. We chose this period because there were no rain or storm events on those days. The measurements from day 457 to 530 were used to test the performance of the IPCA during time-varying characteristics and abnormal behavior. On those days, there were three storm events, which we will discuss in the next section. The main sources of time-variant behavior in this simulation are drift and abrupt changes in the total suspended solid (TSS) concentration of the reject flow of the dewatering process. These changes in the TSS concentration of the dewatering unit are due to seasonal and diurnal patterns. The number of retained PCs, β_k , was estimated recursively using the CPV method in such a way that the retained PCs explained almost 99% of total variance. Several varieties of faults in the sensors, process variables and process parameters were simulated according to Table 6.1. Each fault began on day 470 and continued until day 530. The thresholds confidence limit for all charts was considered equal to 99%.

Table 6. 1 Description of the simulated faults

Fault#	Description		Fault type
1	Sensor fault	Sensor no. 7	Drift fault ($0.1 \text{ g.m}^{-3}.\text{d}^{-1}$)
2	Change in inorganic nitrogen of anaerobic digester	S_{in}	Added 0.2 kmol.m^{-3} to its normal value
3	Secondary clarifier parameter	v_s	Decreased by 50%
4	Step change in bioreactor parameters	μ_A	Decreased from 0.5 to 0.1 d^{-1}

6.5.2.1. Normal operation of the WWTP

First we simulated the normal operation of the process (i.e. with no fault added) to show that conventional PCA is unable to deal with the time-variant characteristics of WWTP. The monitoring statistics obtained by conventional PCA and IPCA are shown in Figure 6.3.

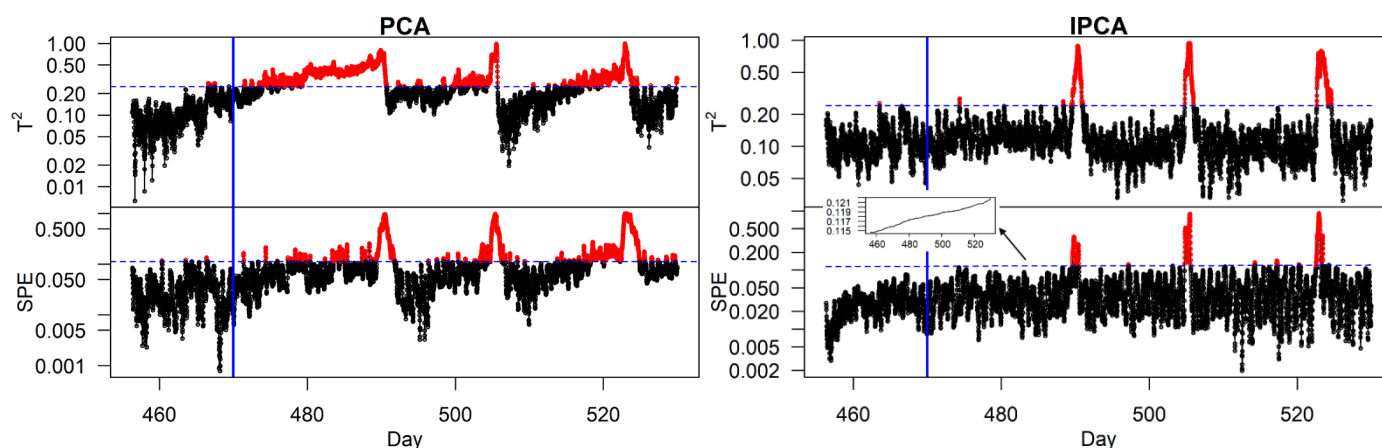


Figure 6. 3 Fault detection results of PCA and IPCA for normal operation (dashed blue line: 99% confidence limit; solid vertical blue line: onset of the fault). The small plot in the IPCA chart shows the adaptive threshold for SPE.

Before we discuss the monitoring results, note that the three severe peaks in this figure for both PCA and IPCA represents storm events (around days 490, 505 and 523) during normal operation of the WWTP. These storm events, which exist in all the FDs studied in this chapter, are

classified as faults. As Figure 6.3 shows, the statistics for conventional PCA exceeded their thresholds for several normal operational points in addition to these storm events. In other words, these results indicate the excessive rate of false alarms, which illustrates a major limitation of using conventional PCA for time-varying processes such as WWTP. What conventional PCA captures in this case is only the dynamic and statistical characteristics of the process supplied by the training data set.

However, in time-varying processes, these dynamic or statistical characteristics change over time and must therefore be updated rather than considered constant. Unlike PCA (which produces an excessive rate of false alarms), IPCA produces a very low rate of false alarms. Apart from storm events, which are correctly labeled as faults, data points mislabeled by the T^2 and SPE statistics of IPCA are very few (shown in red). Also, since the correlation for T^2 did not change, the threshold for T^2 remained constant. On the other hand, due to adaptation, the threshold for SPE varied more although its range remained quite low, i.e. between 0.115 and 0.121 (see Figure 6.3).

6.5.2.2. Drift fault in the dissolved oxygen sensor

The first fault, shown in Table 6.1, is simulated by applying a drift in the oxygen sensor of the fourth reactor at day 470. This sensor is used in combination with a proportional-integral (PI) controller to manipulate the concentration of dissolved oxygen in the third, fourth, and fifth reactors. The size of the drift was $0.1 \text{ g.m}^{-3}.\text{d}^{-1}$. This was intentionally small in order to evaluate the adaptive behavior of IPCA during such a fault. The monitoring statistics and results of diagnosis for IPCA are shown in Figure 6.4. As we can see, both monitoring statistics detected the fault, though there were some delays. These delays in detection were due to the fact that, as the

drift starts, the PI controller tries to compensate for them and brings the dissolved oxygen concentration back to the setpoint by decreasing the blower speed. The fault therefore cannot be detected directly from the faulty sensor at the initial stage of its occurrence. The impact of the PI controller action is more significant on dissolved oxygen in the third and fifth reactors at the initial stage of the fault occurrence.

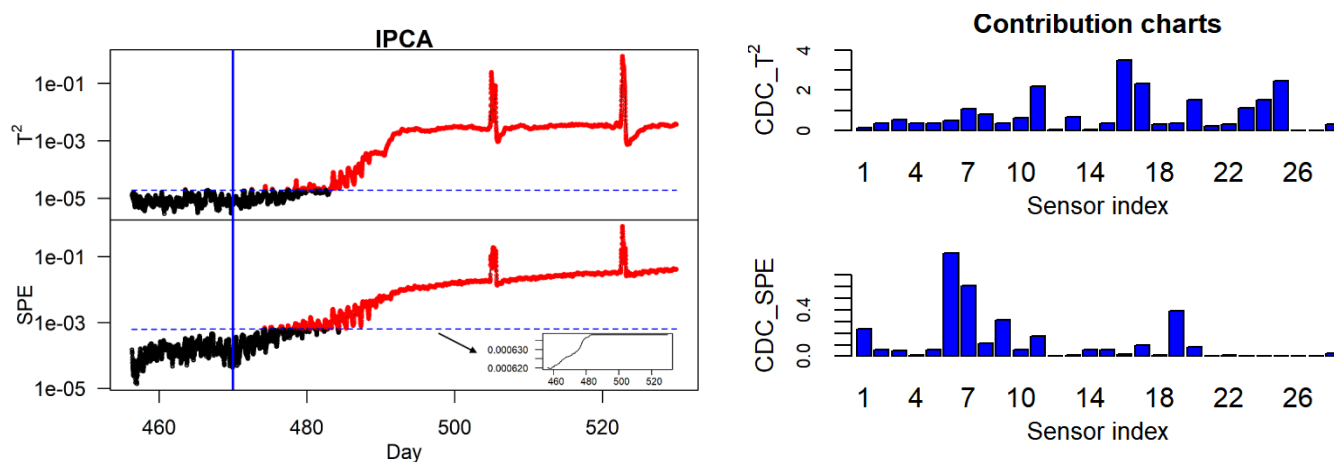


Figure 6. 4 Fault detection and diagnosis results of IPCA for drift fault in the dissolved oxygen sensor (dashed blue line: 99% confidence limit; solid vertical blue line: onset of the fault). The small plot inside the IPCA chart shows the adaptive threshold for SPE.

The SPE statistic therefore begins to violate the threshold around day 475. This is due to the change in the dissolved oxygen concentrations in the third and fifth reactors. However, it is difficult to allocate the drift because it is very small. The T^2 statistic begins to violate the threshold significantly around day 483, which indicates that the PI controller reduces the blower speed to its minimum. At this point, most of the variables begin to deviate and the fault becomes more visible and easier to detect. The contribution chart for T^2 shows that the ammonia, H_2 , TSS, and O_2 sensors (numbers 16, 25, 17, and 11, respectively; see Figure 6.2 for the type and location of sensors) contribute the most. Of these sensors, sensor 11, which represents the dissolved oxygen in the fifth reactor, is isolated correctly. However, the other sensors, such as the concentration of ammonia in

the effluent stream (number 16), may provide indirect clues as to the root of the fault. As the concentration of dissolved oxygen decreases due to the action of the PI controller, the nitrification rate decreases, so the concentration of ammonia increases. The most contributing sensors for the SPE statistic are numbers 6 and 7, which represent the concentration of dissolved oxygen in the third and fourth reactors, respectively. Sensor 6 contributes more than faulty sensor number 7, which shows that the PI controller has more impact on the concentration of dissolved oxygen in the third reactor at the initial stage of the fault.

6.5.2.3. Step change in inorganic nitrogen in the anaerobic digester

The nitrogen level in the AD process is critical because of its inhibitory impact on microbial activity and needs to be monitored carefully. Fault number 2 (Table 6.1) is simulated by inducing a step change at day 470 equal to $+0.2 \text{ kmol.m}^{-3}$ in the AD inorganic nitrogen level. The monitoring statistics contribution charts are shown in Figure 6.5.

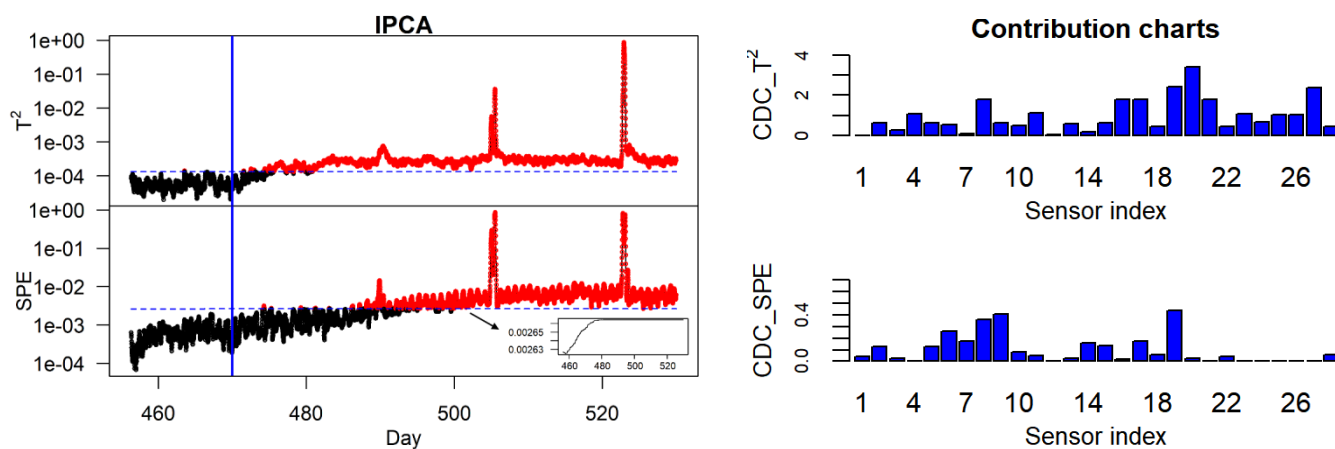


Figure 6. 5 Fault detection and diagnosis results of IPCA for a step change in AD inorganic nitrogen (dashed blue dashed: 99% confidence limit; solid vertical blue line: onset of the fault). The small plot inside the IPCA chart shows the adaptive threshold for SPE.

As we can see, both monitoring statistics detect the fault, though with some delays. The delay is shorter in T^2 than in SPE. These delays are due to the small size of the faults and their lesser impact at the initial stage of their occurrence. Figure 6.5 shows that sensor 20 has the largest contribution to the T^2 statistic. Sensor 20 is the gas flow of AD, which is correctly isolated. Due to the inhibition phenomena caused by the increase in inorganic nitrogen, the activity of microbial communities decreases, which accounts for the reduction in gas flow. The contribution chart for SPE could not isolate the fault correctly.

6.5.2.4. Step change in the settling velocity of the secondary clarifier

To simulate fault number 3 (Table 6.1), the double exponential settling velocity function (v_s) of the second clarifier was reduced by fifty percent. This fault can be classified as a change in the process parameters. Figure 6.6 shows the detection performance and contribution charts for the T^2 and SPE statistics. Figure 6.6 shows that both statistics can detect the fault instantly after its occurrence. However, SPE has a stronger detection than T^2 . Clearly, both statistics provide accurate isolation. Sensors 13 (COD) and 14 (TSS) contribute the most to T^2 and SPE, respectively. Since there is no sensor to measure settling velocity, COD and TSS are considered the most relevant sensors (these variables are directly related to the settling velocity). Note that the smaller magnitude of this fault could still be detected, though the detection rate may be lower due to its lesser impact on the process.

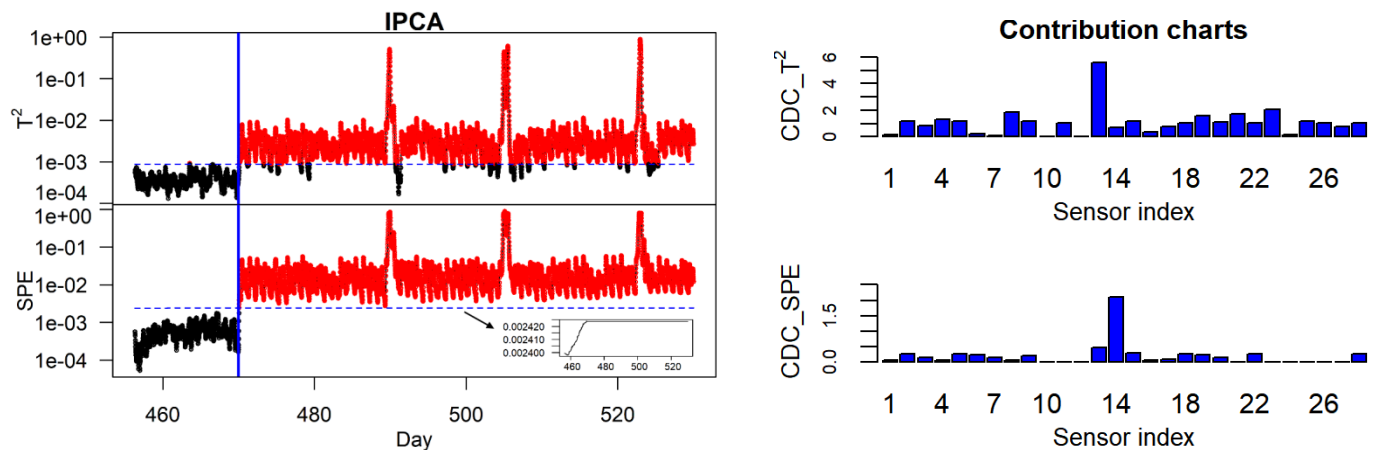


Figure 6. 6 Fault detection and diagnosis results of IPCA for a step-change in the settling velocity of the secondary clarifier (dashed blue line: 99% confidence limit; solid vertical blue line: onset of the fault). The small plot inside the IPCA chart shows the adaptive threshold for SPE.

6.5.2.5. Step change in the bioreactor parameters

One of the most crucial parameters in WWTP is the specific growth rate for the autotrophs (μ_A), which determines the speed of ammonia conversion into nitrite (nitrification rate). If the inhibition (due to toxicity or changes in pH, etc.) occurs in the activated sludge reactors, the nitrification rate is altered due to the lower microbial activity. To simulate fault number 4 (Table 6.1), the value of μ_A was changed from 0.5 to 0.1 d^{-1} . The monitoring statistics and contribution charts are shown in Figure 6.7. This figure shows that the T^2 statistic begins to violate the threshold earlier than the SPE statistic. Although T^2 begins violating the threshold in the initial stage of the fault occurrence, it took almost 30 days to be completely above the threshold. As the nature of this fault is very similar to the drift fault, the detection delay for both methods is very high due to the lesser impact of the faults at the initial stage of their occurrence. The variable contribution chart shows that sensor 10 (the concentration of ammonia in reactor 5) has clearly contributed the most to the deviation in the T^2 statistic. As the nitrification rate decreases, the

concentration of ammonia increases. The contribution chart of the SPE statistic provides no information about the root cause of the fault.

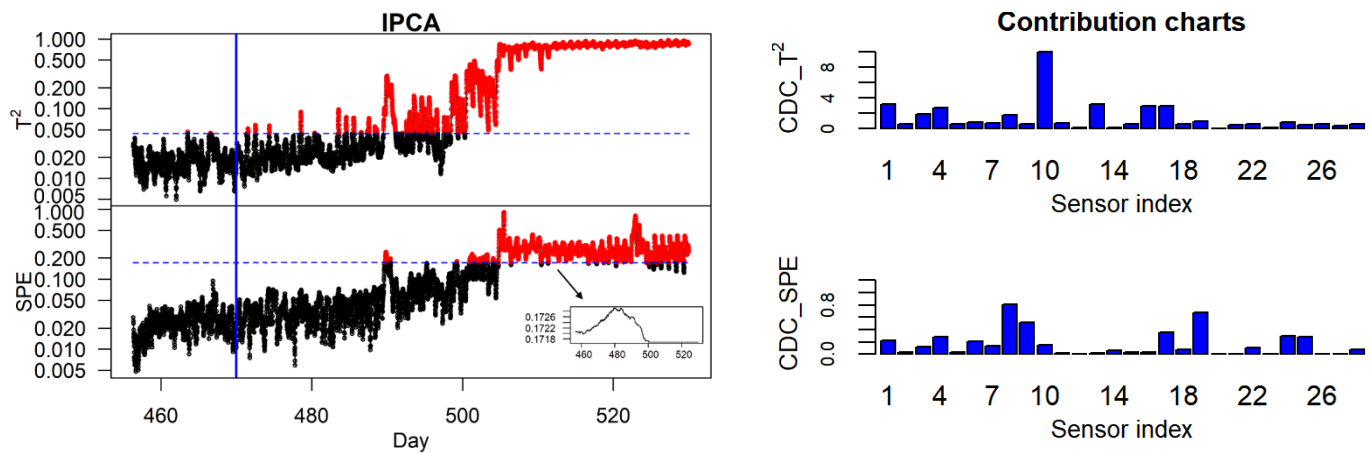


Figure 6. 7 Fault detection and diagnosis results of IPCA for a step change in the specific growth rate of the autotrophs (dashed blue line: 99% confidence limit; solid vertical blue line: onset of the fault). The small plot inside the IPCA chart shows the adaptive threshold for SPE.

6.5.2.6. Storm events

As we mentioned earlier, there are three storm events in the simulated data set. These events are visible in Figure 6.3 around days 490, 505, and 523. Figure 6.3 also shows that both the T^2 and SPE statistics were able to detect these events accurately. Moreover, as the storm events end, the statistical indices immediately go back to their normal values, which indicates the robustness of the method during such a fault. Figure 6.8 shows the variables responsible for the deviation in the T^2 and SPE statistics during the storm events. The sensors' contributions to these three storm events are quite similar. As well as the increasing influent flowrate, many other variables are influenced by storm events; therefore, it is difficult to isolate the fault just by looking at the contribution charts. The trends in all correlated variables therefore need to be examined by experienced process operators. The contribution chart shows that sensors 13 and 19 (effluent COD

and thickener underflow flow rate) contribute the most to the deviation in T^2 . During the storm events, the COD also decreases due to the dilution effect caused by the increase in plant influent flow.

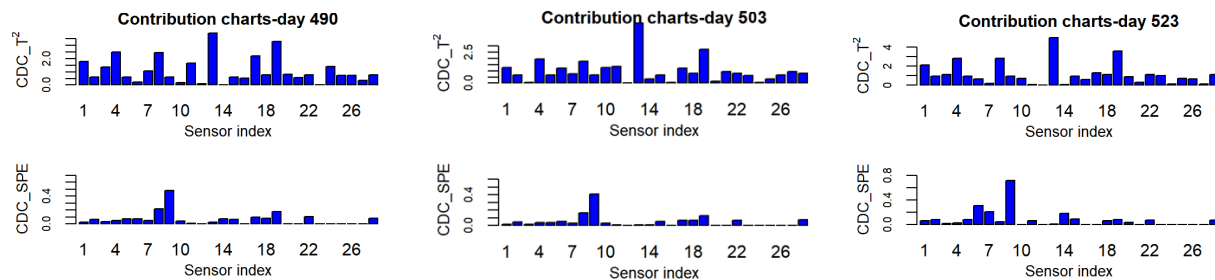


Figure 6. 8 Sensors' contribution to T^2 and SPE statistics at different days of operation under storm condition.

The thickener underflow flow rate also increases as a result of the increase in plant influent flow. Sensor 9 (inlet flow rate of the second clarifier) contributes the most to SPE deviation, which is directly related to the plant influent flow rate. Therefore, as the plant influent flow rate increases due to the storm, the thickener underflow flowrate also increases.

6.6. Fault detection with missing values

Missing values in the input data vector of the monitoring framework is a complex and challenging issue. However, in practice, missing values due to sensor failure or maintenance activities are common. A robust monitoring framework must therefore be designed to handle these missing values. In this study, missing values are imputed using EBLUP. The advantage of this method is that these values can be introduced as they appear during the FD procedure, which is

useful for realtime applications. To assess the accuracy of our proposed method in the presence of missing sensor signals, several measurements were deleted randomly from the data set. Figure 6.9 shows the pattern of missing values from day 457 to day 488 during normal operation of the WWTP. This period was chosen to avoid the storm events during the normal running of the plant. Sampled data are shown in a continuous grey/black color scheme (the darker the color, the higher the value of the sensor), while missing values are shown in red. As we can see, the pattern of the missing values is extremely intricate since many sensor signals can become unavailable at the same time, which is challenging for any FD and monitoring framework.

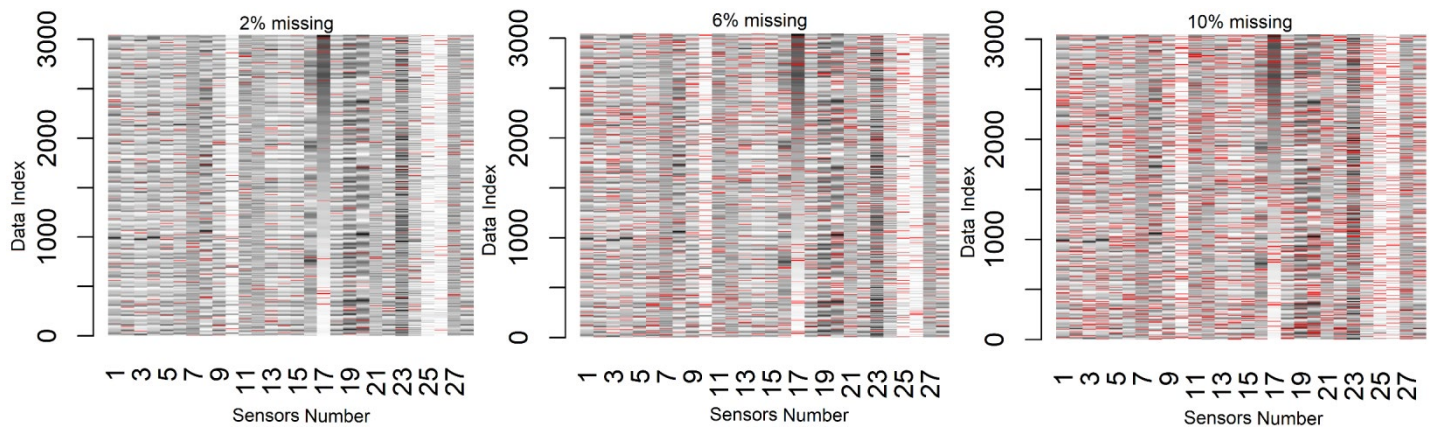


Figure 6. 9 Pattern of missing values during the normal running of the WWTP (from day 457 to day 488).

Table 6.2 shows the false alarm rate (FAR%) and missed detection rate (MDR%) of the monitoring framework for the complete and incomplete data sets during normal and faulty operation of the WWTP. These rates can be estimated as follows:

$$FAR\% = \frac{\text{Number of normal samples incorrectly classified as faulty}}{\text{Total number of normal samples}} \quad (6.22)$$

$$MDR\% = \frac{\text{Number of faulty samples incorrectly classified as normal}}{\text{Total number of faulty samples}} \quad (6.23)$$

We can see that the FAR% for the T^2 statistic is not affected by the missing values. On the other hand, the FAR% for the SPE statistic increases significantly in the presence of missing values. The most significant changes to the FAR% occurred with the largest level of missingness (10%). This may be because the imputation of missing values has a more significant impact on the residual subspace compared to the mean and covariance in PCs. Since the variation in the residual subspace is measured by SPE, it is more sensitive to the amount of imputed missing values.

Table 6. 2 Fault detection performance for complete and uncompleted data set.

	Fault#	0%*		2%*		6%*		10%*	
		T ²	SPE	T ²	SPE	T ²	SPE	T ²	SPE
FAR %	normal**	0.45	0.26	0.59	3.87	0.49	14.05	0.52	26.49
MDR%		-	-	-	-	-	-	-	-
FAR%	1	0.07	0	0.22	3.20	0.15	15.70	0.15	30.10
MDR%		18.53	14.70	18.4	12.99	18.5	10.33	18.58	8.80
FAR%	2	0.07	0	0.22	3.20	0.15	15.70	0.15	30.10
MDR%		8.56	34.26	8.57	29.95	8.44	24.07	8.68	19.86
FAR%	3	0.07	0	0.22	3.20	0.15	15.7	0.15	30.10
MDR%		7.91	0.06	8.45	0.08	9.72	0.08	11.41	0.06
FAR%	4	0.07	0	0.22	3.20	0.15	15.70	0.15	30.10
MDR%		37.86	55.99	37.84	55.40	38.10	54.16	38.34	53.90

* number of missing values; ** from day 457 to day 488 during normal operation; MDR%, missed detection rate; FAR%, false alarm rate.

6.7. Conclusion

We have proposed a novel FD and isolating framework based on IPCA and CDC. The advantage of IPCA over other adaptive PCA approaches is that it does not require all eigenvalues to be computed. This characteristic speeds up computation time. The proposed adaptive IPCA monitoring framework consists of (i) adaptive estimation of the mean and variance of data, (ii) adaptive estimation of selected PCs and threshold, (iii) estimation of eigenvalue decomposition by incrementing the new data to the PCA, and (iv) adaptive estimation of the contribution charts

for T^2 and SPE. This framework was applied to BSM2. Our results, demonstrated with simulated faulty scenarios, show that IPCA is able to adapt time-varying process behavior while detecting and isolating faults. Although there are some delays in detecting the faults, the performance is acceptable since the faults under study were small at the initial stage of their occurrence. The contribution charts showed that this method was able to isolate the faults correctly. Sometimes, however, due to a lack of sensor measurements, it was not possible to isolate the sensor directly and only the impact of the fault on the other measurements could be detected. Our proposed framework can also handle in realtime highly intricate patterns of missing values (e.g. more than one sensor signal failure at the same time) by applying the EBLUP method. Our study of FD performance with missing values shows that SPE is more sensitive to the imputation of missing values than T^2 . This is due to the significant impact of the imputed values on the residual subspace.

UNIVERSITAT ROVIRA I VIRGILI

Data-driven soft-sensors for monitoring and fault diagnosis in wastewater treatment plants

Pezhman Kazemi

CHAPTER 7

Concluding remarks

This final chapter presents the most important contributions and main conclusions of this dissertation, emphasizing their significance. This chapter also includes approaches for future work.

7.1. Summary of the results

The thesis has been presented the application of different soft-sensors for online prediction of crucial parameters and FD and isolation in WWTP by employing advance data-driven techniques. The primary purpose of this thesis was to design the soft-sensors that work beside conventional instrumentation to enable a more efficient and safer operation of WWTPs. Recently most WWTPs are very well-instrumented; as a result, a massive amount of data is stored in their data acquisition systems. The stored data contain useful information about the process operation and provide valuable materials for designing soft-sensors. For this reason, in the present thesis, the soft-sensors were developed using data-driven techniques. The current thesis presented two different applications of data-driven soft-sensors in the wastewater treatment industry.

7.1.1. Online prediction

The VFA measurement in the AD process needs dedicated sensors, which are still currently expensive and very sensitive to harsh environmental conditions. Moreover, their measurement is always associated with a delay that is undesirable for real-time monitoring. Therefore, to overcome this issue, we investigated the application of data-driven techniques for online prediction of VFA. The prediction and generalization performances of different data-driven methods were estimated on a specific validation data set obtained from BSM2. Based on our results, it was found that not all data-driven methods are suitable for developing soft-sensors. For instance, the performance of the RF method for the prediction of VFA was very low. The low performance of RF is because it is a rule-based method, in which the data is categorized into different classes. Therefore, if applied

to temporal data, weak results will be obtained due to the high number of classes. In contrast, the other methods such as ANN, ELM, SVM, and GP showed higher performance in the prediction of VFA. Among these methods, GP soft-sensor has a high potential for implementation in the control system due to its robustness and transparency compared to the other methods used in this thesis.

7.1.2. Fault detection and isolation

The AD is considered as a complex process due to the biological steps that occur in it. Thus, precise monitoring of this complex process is mandatory to make the biogas production process more efficient, reliable, and profitable. The developed soft-sensors for online prediction of VFA were used for FD in AD by applying a simulated data set obtained from BSM2. By comparing different soft-sensors, it was found that although some soft-sensors have shown high accuracy in predicting VFA during normal operating conditions, they could not be effectively used for FD purposes because they are not robust enough. Combining the data-driven soft-sensors with SPC charts such as CUSUM chart showed significant improvement in the detection performance of small magnitude faults. A study on the data set with missing values suggests that CUSUM chart is very robust in FD during missing signals events.

The behavior of wastewater treatment processes is usually very non-stationary. It is often challenging to distinguish the normal operation of the process from those caused by the varying influent conditions. In addition to the mentioned difficulties, the dynamic and nonlinear aspects involved in these processes must be considered. In this thesis, MSPC is proposed as a remedy for these difficulties. PCA is one of the MSPC methods used which used for FD. PCA accounts for

collective effects, as it allows for the simultaneous analysis of all included variables. This would be very beneficial when there are many measurements available similar to the situation in WWTPs. A PCA model is trained using data from normal operation of the process and then, it is used to detect deviations from normal behavior. However, due to changing conditions, for instance, diurnal variations, seasonal changes and long term trends, the monitoring model must be updated. For this reason, we proposed the novel FD framework based on the IPCA. Our results, demonstrated with simulated faulty scenarios, show that IPCA is able to adapt time-varying process behavior while detecting and isolating faults. There are some delays in detecting the faults, which is acceptable since the simulated faults were small at the initial stage of their occurrence. IPCA can correctly isolate the fault, although in some cases, it was not possible to isolate the sensor directly and only the impact of the fault on the other measurements could be detected. Based on the proposed framework, very complex missing values pattern (e.g., more than one sensor signal failure at the same time) can be imputed in a real-time framework.

7.2. Future research lines

The work explained in this thesis and the achieved results provide several insights for further development and future research. Some of them are studies that were omitted from the thesis due to time restrictions, and others are new predictions that originated from issues that emerged during the outcome of the investigation.

Although in BSM2 for each process, compromises were pursued to combine plainness with realism, still the collected data could not be the same as real process data. More practical

experience of soft-sensors in WWTPs would be needed. Therefore, the proposed methods in this thesis need to be validated on real process data from WWTP.

Due to the time-varying behavior of WWTPs, model maintenance is crucial for successful soft-sensor implementation. More study on the soft-sensor maintenance, for instance, by model adaptation would be beneficial.

One of the essential properties of soft-sensors, which is vital for industrial application, is transparency. Industrial practitioners need to understand how the models estimate the predictions. This aspect could be useful for the troubleshooting of the soft-sensors in case of failure. Although we studied the application of the GP technique as a transparent model, yet further studies are needed in this area to achieve higher acceptance of soft-sensors in the process industry.

In order to use the developed soft-sensors in the process industry, professional implementation is necessary. At this moment, the developed frameworks in this thesis are just implemented in the R programming environment. For a real-life application, the developed codes need to be modified according to the programming environment of the industrial process. Further studies are needed to develop a general framework that can be easily integrated into the monitoring and control system of industrial processes.

To find out the primary source of the fault in this thesis, we combine the traditional CDC method with IPCA. Although CDC is a standard method for fault diagnostic, sometimes it has a lower performance compared to the other methods. Therefore, for the future study, it would be an excellent idea to compare the performance of the CDC with the other potential diagnostic methods such as reconstruction methods, structured residuals based on multivariate statistical methods, angle-based methods and reconstruction-based contribution based.

In addition to non-stationary behaviors, WWTPs exhibit non-linear and autocorrelated behaviors as well. In order to overcome these problems, i.e., to take into account the serial correlation in the data or to capture the process dynamic, the potential of dynamic IPCA (DIPCA), which uses an augmenting matrix with time-lagged variables need to be studied. It should be noted that by adding time-lagged variables, the number of variables significantly increases, which can cause an increase in computational cost.

UNIVERSITAT ROVIRA I VIRGILI

Data-driven soft-sensors for monitoring and fault diagnosis in wastewater treatment plants

Pezhman Kazemi

References

- Abdullah, S.S., Malek, M.A., Abdullah, N.S., Kisi, O., Yap, K.S., 2015. Extreme Learning Machines: A new approach for prediction of reference evapotranspiration. *J. Hydrol.* 527, 184–195. <https://doi.org/10.1016/J.JHYDROL.2015.04.073>
- Abu Qdais, H., Bani Hani, K., Shatnawi, N., 2010. Modeling and optimization of biogas production from a waste digester using artificial neural network and genetic algorithm. *Resour. Conserv. Recycl.* 54, 359–363. <https://doi.org/10.1016/J.RESCONREC.2009.08.012>
- Abujiya, M.R., Riaz, M., Lee, M.H., 2015. Enhanced Cumulative Sum Charts for Monitoring Process Dispersion. *PLoS One* 10, e0124520. <https://doi.org/10.1371/journal.pone.0124520>
- Activated sludge - Wikiwand [WWW Document], URL https://www.wikiwand.com/en/Activated_sludge (accessed 7.9.20).
- Al Bazed, G.A., Abdel-Fatah, M.A., 2020. Correlation between operating parameters and removal efficiency for chemically enhanced primary treatment system of wastewater. *Bull. Natl. Res. Cent.* 44, 1–6. <https://doi.org/10.1186/s42269-020-00368-y>
- Alcala, C.F., Qin, S.J., 2009. Reconstruction-based contribution for process monitoring. *Automatica* 45, 1593–1600. <https://doi.org/10.1016/j.automatica.2009.02.027>
- Alcaraz-González, V., López-Bañuelos, R.H., Steyer, J.-P., Méndez-Acosta, H.O., González-Álvarez, V., Pelayo-Ortiz, C., 2012. Interval-Based Diagnosis of Biological Systems - a Powerful Tool for Highly Uncertain Anaerobic Digestion Processes. *CLEAN - Soil, Air, Water* 40, 941–949. <https://doi.org/10.1002/clen.201100721>
- Angelidaki, I., Ellegaard, L., Ahring, B.K., 2003. Applications of the anaerobic digestion process.

Adv. Biochem. Eng. Biotechnol. https://doi.org/10.1007/3-540-45838-7_1

Arora, R., Cotter, A., Livescu, K., Srebro, N., 2012. Stochastic optimization for PCA and PLS.

2012 50th Annu. Allert. Conf. Commun. Control. Comput. Allert. 2012 861–868.

<https://doi.org/10.1109/Allerton.2012.6483308>

Artač, M., Jogan, M., Leonardis, A., 2002. Incremental PCA for on-line visual learning and

recognition, in: Proceedings - International Conference on Pattern Recognition. pp. 781–784.

<https://doi.org/10.1109/icpr.2002.1048133>

Baggiani, F., Marsili-Libelli, S., 2009. Real-time fault detection and isolation in biological

wastewater treatment plants. Water Sci. Technol. 60, 2949–2961.

<https://doi.org/10.2166/wst.2009.723>

Bahrani, P., Kazemi, P., Mahdavi, S., Ghobadi, H., 2016. A novel approach for modeling and

optimization of surfactant/polymer flooding based on Genetic Programming evolutionary algorithm. Fuel 179, 289–298.

Baraldi, P., Di Maio, F., Genini, D., Zio, E., 2015. Comparison of Data-Driven Reconstruction

Methods For Fault Detection. IEEE Trans. Reliab. 64, 852–860.

<https://doi.org/10.1109/TR.2015.2436384>

Bekkari, N., Zeddouri, A., 2019. Using artificial neural network for predicting and controlling the

effluent chemical oxygen demand in wastewater treatment plant. Manag. Environ. Qual. An

Int. J. 30, 593–608. <https://doi.org/10.1108/MEQ-04-2018-0084>

Beltramo, T., Klocke, M., Hitzmann, B., 2019. Prediction of the biogas production using GA and

ACO input features selection method for ANN model. Inf. Process. Agric. 6, 349–356.

<https://doi.org/10.1016/J.INPA.2019.01.002>

Bergstra, J., Ca, J.B., Ca, Y.B., 2012. Random Search for Hyper-Parameter Optimization Yoshua Bengio, *Journal of Machine Learning Research*.

Bin Shams, M.A., Budman, H.M., Duever, T.A., 2011. Fault detection, identification and diagnosis using CUSUM based PCA. *Chem. Eng. Sci.* 66, 4488–4498.
<https://doi.org/10.1016/J.CES.2011.05.028>

Bishop, C.M., 1995. *Neural Networks for Pattern Recognition*. Clarendon press, Oxford.

Boe, K., Batstone, D.J., Steyer, J.-P., Angelidaki, I., 2010. State indicators for monitoring the anaerobic digestion process. *Water Res.* 44, 5973–5980.
<https://doi.org/10.1016/J.WATRES.2010.07.043>

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 2008. *Time Series Analysis, Time Series Analysis: Forecasting and Control: Fourth Edition*, Wiley Series in Probability and Statistics. Wiley.
<https://doi.org/10.1002/9781118619193>

Brand, M., 2002. Incremental singular value decomposition of uncertain data with missing values. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 2350, 707–720. https://doi.org/10.1007/3-540-47969-4_47

Breiman, L., 2001. Random Forests. *Mach. Learn.* 45, 5–32.
<https://doi.org/10.1023/A:1010933404324>

Breiman, L., 1996. Bagging predictors. *Mach. Learn.* 24, 123–140.
<https://doi.org/10.1007/bf00058655>

Bro, R., 1996. Multiway calibration. *Multilinear PLS. J. Chemom.* 10, 47–61.

[https://doi.org/10.1002/\(SICI\)1099-128X\(199601\)10:1<47::AID-CEM400>3.0.CO;2-C](https://doi.org/10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C)

Candel, A., Ledell, E., Bartz, A., 2018. Deep Learning with H2O.

Cannon, A.J., 2019. monmlp: Monotone Multi-Layer Perceptron Neural Network.

Cannon, A.J., Whitfield, P.H., 2002. Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models. *J. Hydrol.* 259, 136–151.

[https://doi.org/10.1016/S0022-1694\(01\)00581-9](https://doi.org/10.1016/S0022-1694(01)00581-9)

Capizzi, G., Masarotto, G., 2017. Phase I Distribution-Free Analysis of Multivariate Data. *Technometrics* 59, 484–495. <https://doi.org/10.1080/00401706.2016.1272494>

Cardot, H., Degras, D., 2018. Online Principal Component Analysis in High Dimension: Which Algorithm to Choose? *Int. Stat. Rev.* 86, 29–50. <https://doi.org/10.1111/insr.12220>

Carrasco, E.F., Rodriguez, J., Puñal, A., Roca, E., Lema, J.M., 2004. Diagnosis of acidification states in an anaerobic wastewater treatment plant using a fuzzy-based expert system. *Control Eng. Pract.* 12, 59–64. [https://doi.org/10.1016/S0967-0661\(02\)00304-0](https://doi.org/10.1016/S0967-0661(02)00304-0)

Cateni, S., Vannucci, M., Vannocci, M., Coll, V., 2013. Variable Selection and Feature Extraction Through Artificial Intelligence Techniques, in: *Multivariate Analysis in Management, Engineering and the Sciences*. InTech. <https://doi.org/10.5772/53862>

Chen, A., Zhou, H., An, Y., Sun, W., 2016. PCA and PLS monitoring approaches for fault detection of wastewater treatment process, in: *IEEE International Symposium on Industrial Electronics*. Institute of Electrical and Electronics Engineers Inc., pp. 1022–1027. <https://doi.org/10.1109/ISIE.2016.7745032>

Cheng, H., Liu, Y., Huang, D., Liu, B., 2019. Optimized Forecast Components-SVM-Based Fault

- Diagnosis with Applications for Wastewater Treatment. *IEEE Access* 7, 128534–128543.
<https://doi.org/10.1109/ACCESS.2019.2939289>
- Chi, B., Guo, L., 2019. Wastewater treatment sensor fault detection using RBF neural network with set membership estimation, in: *Proceedings of the 31st Chinese Control and Decision Conference, CCDC 2019*. Institute of Electrical and Electronics Engineers Inc., pp. 2685–2690. <https://doi.org/10.1109/CCDC.2019.8832519>
- Chicco, D., 2017. Ten quick tips for machine learning in computational biology. *BioData Min.*
<https://doi.org/10.1186/s13040-017-0155-3>
- Choi, S.W., Lee, I.B., 2005. Multiblock PLS-based localized process diagnosis. *J. Process Control* 15, 295–306. <https://doi.org/10.1016/j.jprocont.2004.06.010>
- Choi, S.W., Martin, E.B., Morris, A.J., Lee, I.B., 2006. Adaptive multivariate statistical process control for monitoring time-varying processes. *Ind. Eng. Chem. Res.* 45, 3108–3118.
<https://doi.org/10.1021/ie050391w>
- Chou, Y.-M., Mason, R.L., Young, J.C., 2001. The Control Chart For Individual Observations From A Multivariate Non-Normal Distribution. *Commun. Stat. - Theory Methods* 30, 1937–1949. <https://doi.org/10.1081/STA-100105706>
- Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cort Es, U., Poch, M., Cortés, U., Poch, M.,. Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques 106, 89–103. <https://doi.org/10.1016/j.envsoft.2017.11.023>
- Corominas, L., Villez, K., Aguado, D., Rieger, L., Rosén, C., Vanrolleghem, P.A., 2011. Performance evaluation of fault detection methods for wastewater treatment processes.

Biotechnol. Bioeng. 108, 333–344. <https://doi.org/10.1002/bit.22953>

Corona, F., Mulas, M., Haimi, H., Sundell, L., Heinonen, M., Vahala, R., 2013. Monitoring nitrate concentrations in the denitrifying post-filtration unit of a municipal wastewater treatment plant. *J. Process Control* 23, 158–170. <https://doi.org/10.1016/j.jprocont.2012.09.011>

Cortes, C., Vapnik, V., 1995. Support-Vector Networks. *Mach. Learn.* 20, 273–297. <https://doi.org/10.1023/A:1022627411411>

Curreri, F., Graziani, S., Xibilia, M.G., 2020. Input selection methods for data-driven Soft sensors design: Application to an industrial process. *Inf. Sci. (Ny)*. 537, 1–17. <https://doi.org/10.1016/j.ins.2020.05.028>

Dayal, B.S., MacGregor, J.F., 1997. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. *J. Process Control* 7, 169–179. [https://doi.org/10.1016/S0959-1524\(97\)80001-7](https://doi.org/10.1016/S0959-1524(97)80001-7)

Dixon, M., Gallop, J.R., Lambert, S.C., Lardon, L., Healy, J. V, Steyer, J.-P., 2007. Data mining to support anaerobic WWTP monitoring. *Control Eng. Pract.* 15, 987–999. <https://doi.org/10.1016/j.conengprac.2006.11.010>

Dürrenmatt, D.J., Gujer, W., 2012. Data-driven modeling approaches to support wastewater treatment plant operation. *Environ. Model. Softw.* 30, 47–56. <https://doi.org/10.1016/j.envsoft.2011.11.007>

Ebrahimi, M., Gerber, E.L., Rockaway, T.D., 2017. Temporal performance assessment of wastewater treatment plants by using multivariate statistical analysis. *J. Environ. Manage.* 193, 234–246. <https://doi.org/10.1016/j.jenvman.2017.02.027>

- Elshenawy, L.M., Awad, H.A., 2012. Recursive fault detection and isolation approaches of time-varying processes. *Ind. Eng. Chem. Res.* 51, 9812–9824. <https://doi.org/10.1021/ie300072q>
- Elshenawy, L.M., Mahmoud, T.A., 2018. Fault diagnosis of time-varying processes using modified reconstruction-based contributions. *J. Process Control* 70, 12–23. <https://doi.org/10.1016/j.jprocont.2018.07.017>
- Elshenawy, L.M., Yin, S., Naik, A.S., Ding, S.X., 2010. Efficient recursive principal component analysis algorithms for process monitoring. *Ind. Eng. Chem. Res.* 49, 252–259. <https://doi.org/10.1021/ie900720w>
- Eskandarian, S., Bahrami, P., Kazemi, P., 2017. A comprehensive data mining approach to estimate the rate of penetration: Application of neural network, rule based models and feature ranking. *J. Pet. Sci. Eng.* 156, 605–615. <https://doi.org/10.1016/J.PETROL.2017.06.039>
- Fortuna, L., Graziani, S., Rizzo, A., G. Xibilia, M., 2007. *Soft Sensors for Monitoring and Control of Industrial Processes*, *Soft Sensors for Monitoring and Control of Industrial Processes*. Springer London, London, UK. <https://doi.org/10.1007/978-1-84628-480-9>
- Franke-Whittle, I.H., Walter, A., Ebner, C., Insam, H., 2014. Investigation into the effect of high concentrations of volatile fatty acids in anaerobic digestion on methanogenic communities. *Waste Manag.* 34, 2080–2089.
- Freund, Y., Schapire, R.E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* 55, 119–139. <https://doi.org/10.1006/jcss.1997.1504>
- Garcia-Alvarez, D., Fuente, M.J., Vega, P., Sainz, G., 2009. Fault Detection and Diagnosis using

- Multivariate Statistical Techniques in a Wastewater Treatment Plant. IFAC Proc. Vol. 42, 952–957. <https://doi.org/10.3182/20090712-4-tr-2008.00156>
- Genovesi, A., Harmand, J., Steyer, J.-P., 2000. Integrated Fault Detection and Isolation: Application to a Winery's Wastewater Treatment Plant. *Appl. Intell.* 13, 59–76. <https://doi.org/10.1023/A:1008379329794>
- Genovesi, A., Harmand, J., Steyer, J.-P., 1999. A fuzzy logic based diagnosis system for the on-line supervision of an anaerobic digester pilot-plant. *Biochem. Eng. J.* 3, 171–183. [https://doi.org/10.1016/S1369-703X\(99\)00015-7](https://doi.org/10.1016/S1369-703X(99)00015-7)
- Gernaey, K. V, Van Loosdrecht, M.C.M., Henze, M., Lind, M., Jørgensen, S.B., 2004. Activated sludge wastewater treatment plant modelling and simulation: state of the art. *Environ. Model. Softw.* 19, 763–783. <https://doi.org/10.1016/j.envsoft.2003.03.005>
- Gil, J.D., Ruiz-Aguirre, A., Roca, L., Zaragoza, G., Berenguel, M., 2018. Prediction models to analyse the performance of a commercial-scale membrane distillation unit for desalting brines from RO plants. *Desalination* 445, 15–28. <https://doi.org/10.1016/j.desal.2018.07.022>
- Gonzalez, G.D., Orchard, M., Cerda, J.L., Casali, A., Vallebuona, G., 2003. Local models for soft-sensors in a rougher flotation bank. *Miner. Eng.* 16, 441–453. [https://doi.org/10.1016/S0892-6875\(03\)00021-9](https://doi.org/10.1016/S0892-6875(03)00021-9)
- Gray, N.F., 2004. *Biology of Wastewater Treatment*, Series on Environmental Science and Management. Published by imperial college press and distributed by world scientific publishing co. <https://doi.org/10.1142/p266>
- Guang-Bin Huang, Qin-Yu Zhu, Chee-Kheong Siew, 2004. Extreme learning machine: a new

- learning scheme of feedforward neural networks, in: IEEE International Conference on Neural Networks - Conference Proceedings. IEEE, pp. 985–990.
<https://doi.org/10.1109/IJCNN.2004.1380068>
- Güçlü, D., Yılmaz, N., Ozkan-Yucel, U.G., 2011. Application of neural network prediction model to full-scale anaerobic sludge digestion. *J. Chem. Technol. Biotechnol.* 86, 691–698.
<https://doi.org/10.1002/jctb.2569>
- Haimi, H., Mulas, M., Corona, F., Marsili-Libelli, S., Lindell, P., Heinonen, M., Vahala, R., 2016. Adaptive data-derived anomaly detection in the activated sludge process of a large-scale wastewater treatment plant. *Eng. Appl. Artif. Intell.* 52, 65–80.
<https://doi.org/10.1016/j.engappai.2016.02.003>
- Haimi, H., Mulas, M., Corona, F., Vahala, R., 2013. Data-derived soft-sensors for biological wastewater treatment plants: An overview. *Environ. Model. Softw.* 47, 88–107.
<https://doi.org/10.1016/J.ENVSOF.2013.05.009>
- Hall, P.M., Marshall, D., Martin, R.R., 1998. Incremental Eigenanalysis for Classification, in: *Proceedings of the British Machine Vision Conference 1998*. British Machine Vision Association, pp. 29.1-29.10. <https://doi.org/10.5244/C.12.29>
- Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition*, Springer series in statistics. <https://doi.org/10.1007/978-0-387-84858-7>
- Henze, M., Loosdrecht, M.C.M. van, Ekama, G.A., Brdjanovic, D., 2008. *Biological wastewater treatment: Principles, modelling and design*. IWA Publishing.

- Hota, H.S., Handa, R., Shrivastava, A.K., 2017. Time Series Data Prediction Using Sliding Window Based RBF Neural Network, *International Journal of Computational Intelligence Research*.
- Hreiz, R., Latifi, M.A., Roche, N., 2015. Optimal design and operation of activated sludge processes: State-of-the-art. <https://doi.org/10.1016/j.cej.2015.06.125>
- Hu, Z., Chen, Z., Hua, C., Gui, W., Yang, C., Ding, S.X., 2012. A simplified recursive dynamic pca based monitoring scheme for imperial smelting process. *Int. J. Innov. Comput. Inf. Control* 8, 2551–2561.
- Huang, M., Han, W., Wan, J., Ma, Y., Chen, X., 2016. Multi-objective optimisation for design and operation of anaerobic digestion using GA-ANN and NSGA-II. *J. Chem. Technol. Biotechnol.* 91, 226–233. <https://doi.org/10.1002/jctb.4568>
- Jackson, J.E., 1991. *A user's guide to principal components*. Wiley.
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. *An Introduction to Statistical Learning*, Springer Texts in Statistics. Springer New York, New York, NY. <https://doi.org/10.1007/978-1-4614-7138-7>
- James, S.C., Legge, R.L., Budman, H., 2000. ON-LINE ESTIMATION IN BIOREACTORS: A REVIEW. *Rev. Chem. Eng.* 16, 311–340. <https://doi.org/10.1515/REVCE.2000.16.4.311>
- Jang, J.S.R., Sun, C.T., Mizutani, E., 2005. Neuro-Fuzzy and Soft Computing-A Computational Approach to Learning and Machine Intelligence [Book Review]. *IEEE Trans. Automat. Contr.* 42, 1482–1484. <https://doi.org/10.1109/tac.1997.633847>
- Jeppsson, U., Pons, M.N., Nopens, I., Alex, J., Copp, J.B., Gernaey, K. V., Rosen, C., Steyer, J.P., Vanrolleghem, P.A., 2007. Benchmark simulation model no 2: General protocol and

exploratory case studies. *Water Sci. Technol.* 56, 67–78.

<https://doi.org/10.2166/wst.2007.604>

Jeppsson, U., Rosen, C., Alex, J., Copp, J., Gernaey, K.V., Pons, M.-N., Vanrolleghem, P.A., 2006. Towards a benchmark simulation model for plant-wide control strategy performance evaluation of WWTPs. *Water Sci. Technol.* 53, 287–295.

<https://doi.org/10.2166/wst.2006.031>

Jimenez, J., Latrille, E., Harmand, J., Robles, A., Ferrer, J., Gaida, D., Wolf, C., Mairet, F., Bernard, O., Alcaraz-Gonzalez, V., Mendez-Acosta, H., Zitomer, D., Totzke, D., Spanjers, H., Jacobi, F., Guwy, A., Dinsdale, R., Premier, G., Mazhegrane, S., Ruiz-Filippi, G., Seco, A., Ribeiro, T., Pauss, A., Steyer, J.-P., 2015. Instrumentation and control of anaerobic digestion processes: a review and some research challenges. *Rev. Environ. Sci. Bio/Technology* 14, 615–648. <https://doi.org/10.1007/s11157-015-9382-6>

Jolliffe, I., 2011. Principal Component Analysis, in: *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1094–1096. https://doi.org/10.1007/978-3-642-04898-2_455

Jun, B.H., Park, J.H., Lee, S.I., Chun, M.G., 2006. Kernel PCA based faults diagnosis for wastewater treatment system, in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, pp. 426–431. https://doi.org/10.1007/11760191_63

Kadlec, P., 2009. On robust and adaptive soft sensors. *Computing*.

Kadlec, P., Gabrys, B., Strandt, S., 2009. Data-driven Soft Sensors in the process industry. *Comput. Chem. Eng.* 33, 795–814. <https://doi.org/10.1016/j.compchemeng.2008.12.012>

- Kadlec, P., Grbić, R., Gabrys, B., 2011. Review of adaptation mechanisms for data-driven soft sensors. *Comput. Chem. Eng.* 35, 1–24. <https://doi.org/10.1016/J.COMPCHEMENG.2010.07.034>
- Kaneko, H., Funatsu, K., 2014. Database monitoring index for adaptive soft sensors and the application to industrial process. *AIChE J.* 60, 160–169. <https://doi.org/10.1002/aic.14260>
- Kazemi, P., Steyer, J.-P.P., Bengoa, C., Font, J., Giralt, J., 2020. Robust data-driven soft sensors for online monitoring of volatile fatty acids in anaerobic digestion processes. *Processes* 8, 67. <https://doi.org/10.3390/pr8010067>
- Kazor, K., Holloway, R.W., Cath, T.Y., Hering, A.S., 2016. Comparison of linear and nonlinear dimension reduction techniques for automated process monitoring of a decentralized wastewater treatment facility. *Stoch. Environ. Res. Risk Assess.* 30, 1527–1544. <https://doi.org/10.1007/s00477-016-1246-2>
- Khusna, H., Mashuri, M., Ahsan, M., Suhartono, S., Prastyo, D.D., 2018. Bootstrap-based maximum multivariate CUSUM control chart. *Qual. Technol. Quant. Manag.* 1–23. <https://doi.org/10.1080/16843703.2018.1535765>
- Kohonen, T., 2001. *Self-Organizing Maps*, Springer Series in Information Sciences. Springer Berlin Heidelberg, Berlin, Heidelberg. <https://doi.org/10.1007/978-3-642-56927-2>
- Koza, J., 1994. Genetic programming as a means for programming computers by natural selection. *Stat. Comput.* 4. <https://doi.org/10.1007/BF00175355>
- Kuhn, M., 2008. Building Predictive Models in *R* Using the **caret** Package. *J. Stat. Softw.* 28, 1–26. <https://doi.org/10.18637/jss.v028.i05>

- Lee, C., Choi, S.W., Lee, I.B., 2006. Sensor fault diagnosis in a wastewater treatment process. *Water Sci. Technol.* 53, 251–257. <https://doi.org/10.2166/wst.2006.027>
- Lee, D.S., Lee, M.W., Woo, S.H., Kim, Y.J., Park, J.M., 2006. Nonlinear dynamic partial least squares modeling of a full-scale biological wastewater treatment plant. *Process Biochem.* 41, 2050–2057. <https://doi.org/10.1016/j.procbio.2006.05.006>
- Lee, J.M., Yoo, C.K., Choi, S.W., Vanrolleghem, P.A., Lee, I.B., 2004. Nonlinear process monitoring using kernel principal component analysis. *Chem. Eng. Sci.* 59, 223–234. <https://doi.org/10.1016/j.ces.2003.09.012>
- Li, W., Yue, H.H., Valle-Cervantes, S., Qin, S.J., 2000. Recursive PCA for adaptive process monitoring. *J. Process Control* 10, 471–486. [https://doi.org/10.1016/S0959-1524\(00\)00022-6](https://doi.org/10.1016/S0959-1524(00)00022-6)
- Lin, B., Recke, B., Knudsen, J.K.H., Jørgensen, S.B., 2007. A systematic approach for soft sensor development. *Comput. Chem. Eng.* 31, 419–425. <https://doi.org/10.1016/j.compchemeng.2006.05.030>
- Liu, L., Lei, Y., 2018. An accurate ecological footprint analysis and prediction for Beijing based on SVM model. *Ecol. Inform.* 44, 33–42. <https://doi.org/10.1016/J.ECOINF.2018.01.003>
- Lourenço, N.D., Menezes, J.C., Pinheiro, H.M., Diniz, D., 2008. Development of PLS calibration models from UV-Vis spectra for TOC estimation at the outlet of a fuel park wastewater treatment plant. *Environ. Technol.* 29, 891–898. <https://doi.org/10.1080/09593330802015581>
- Mandic, D.P., Chambers, J.A., 2001. Recurrent Neural Networks for Prediction, Recurrent Neural

Networks for Prediction. John Wiley & Sons, Ltd. <https://doi.org/10.1002/047084535x>

Metcalf & Eddy, I., Tchobanoglous, G., Stensel, H.D., Tsuchihashi, R., Burton, F., 2014.

Wastewater Engineering - Treatment and Resource Recovery.

Mina, J., Verde, C., 2006. Fault detection for large scale systems using Dynamic Principal

Components Analysis with adaptation, in: IFAC Proceedings Volumes (IFAC-PapersOnline). pp. 220–225. <https://doi.org/10.15837/ijccc.2007.2.2351>

Miron, M., Frangu, L., Caraman, S., Luca, L., 2018. Artificial neural network approach for fault

recognition in a wastewater treatment process, in: 2018 22nd International Conference on System Theory, Control and Computing, ICSTCC 2018 - Proceedings. Institute of Electrical and Electronics Engineers Inc., pp. 634–639. <https://doi.org/10.1109/ICSTCC.2018.8540694>

Montgomery, Douglas, C., 2009. Introduction To Statistical Quality Control., Sixth. ed. John

Wiley & Sons, Inc., New York, United States.

Mullai, P., Arulselvi, S., Ngo, H.-H., Sabarathinam, P.L., 2011. Experiments and ANFIS

modelling for the biodegradation of penicillin-G wastewater using anaerobic hybrid reactor. Bioresour. Technol. 102, 5492–5497. <https://doi.org/10.1016/j.biortech.2011.01.085>

Najafzadeh, M., Etemad-Shahidi, A., Lim, S.Y., 2016. Scour prediction in long contractions using

ANFIS and SVM. Ocean Eng. 111, 128–135.

<https://doi.org/10.1016/J.OCEANENG.2015.10.053>

Najafzadeh, M., Zeinolabedini, M., 2019. Prognostication of waste water treatment plant

performance using efficient soft computing models: An environmental evaluation. <https://doi.org/10.1016/j.measurement.2019.02.014>

- Nan, C., Khan, F., Tariq Iqbal, M., Iqbal, M.T., 2008. Real-time fault diagnosis using knowledge-based expert system. *Process Saf. Environ. Prot.* 86, 55–71.
<https://doi.org/10.1016/j.psep.2007.10.014>
- Newhart, K.B., Holloway, R.W., Hering, A.S., Cath, T.Y., 2019. Data-driven performance analyses of wastewater treatment plants: A review. *Water Res.* 157, 498–513.
<https://doi.org/10.1016/j.watres.2019.03.030>
- Nomikos, P., MacGregor, J.F., 1995. Multivariate SPC charts for monitoring batch processes. *Technometrics* 37, 41–59. <https://doi.org/10.1080/00401706.1995.10485888>
- Nopens, I., Benedetti, L., Jeppsson, U., Pons, M.-N., Alex, J., Copp, J.B., Gernaey, K. V., Rosen, C., Steyer, J.-P., Vanrolleghem, P.A., 2010. Benchmark Simulation Model No 2: finalisation of plant layout and default control strategy. *Water Sci. Technol.* 62, 1967–1974.
<https://doi.org/10.2166/wst.2010.044>
- Olsson, G., 2012. ICA and me - A subjective review. *Water Res.*
<https://doi.org/10.1016/j.watres.2011.12.054>
- Olsson, G., Nielsen, M., Yuan, Z., Lynggaard-Jensen, A., Steyer, J.-P., 2005. *Instrumentation, Control and Automation in Wastewater Systems*. IWA Publishing.
<https://doi.org/10.2166/9781780402680>
- Park, S.-H., Koo, J., 2015. Application of Transfer Function ARIMA Modeling for the Sedimentation Process on Water Treatment Plant. *Int. J. Control Autom.* 8, 135–144.
<https://doi.org/10.14257/ijca.2015.8.10.13>
- Phaladiganon, P., Kim, S.B., Chen, V.C.P., Baek, J.-G., Park, S.-K., 2011. Bootstrap-Based T^2

- Multivariate Control Charts. *Commun. Stat. - Simul. Comput.* 40, 645–662.
<https://doi.org/10.1080/03610918.2010.549989>
- Pisa, I., Santín, I., López Vicario, J., Morell, A., Vilanova, R., 2018. A recurrent neural network for wastewater treatment plant effluents' prediction, in: *Actas de Las XXXIX Jornadas de Automática*. pp. 621–628. <https://doi.org/10.17979/spudc.9788497497565.0621>
- Poggio, T., Girosi, F., 1990. Regularization algorithms for learning that are equivalent to multilayer networks. *Science* (80). 247, 978–982.
<https://doi.org/10.1126/science.247.4945.978>
- Pons, M.-N., Corriou, J.-P., 2001. Implementation of an Equalisation Tank on the Cost 624 Wastewater Treatment Plant Benchmark. *IFAC Proc. Vol.* 34, 433–437.
[https://doi.org/10.1016/s1474-6670\(17\)34258-1](https://doi.org/10.1016/s1474-6670(17)34258-1)
- Qin, S.J., 1998. Recursive PLS algorithms for adaptive data modeling. *Comput. Chem. Eng.* 22, 503–514. [https://doi.org/10.1016/s0098-1354\(97\)00262-7](https://doi.org/10.1016/s0098-1354(97)00262-7)
- Qin, S.J., McAvoy, T.J., 1992. Nonlinear PLS modeling using neural networks. *Comput. Chem. Eng.* 16, 379–391. [https://doi.org/10.1016/0098-1354\(92\)80055-E](https://doi.org/10.1016/0098-1354(92)80055-E)
- R, 2017. R Core Team. R: A language and environment for statistical computing.
- Rangasamy, P., Pvr, I., Ganesan, S., 2007. Anaerobic tapered fluidized bed reactor for starch wastewater treatment and modeling using multilayer perceptron neural network. *J. Environ. Sci.* 19, 1416–1423. [https://doi.org/10.1016/S1001-0742\(07\)60231-9](https://doi.org/10.1016/S1001-0742(07)60231-9)
- Rieger, L., Alex, J., Winkler, S., Boehler, M., Thomann, M., Siegrist, H., 2003. Progress in sensor technology-progress in process control? Part I: Sensor property investigation and

classification.

- Ritari, J., Koskinen, K., Hultman, J., Kurola, J.M., Kymäläinen, M., Romantschuk, M., Paulin, L., Auvinen, P., 2012. Molecular analysis of meso- and thermophilic microbiota associated with anaerobic biowaste degradation. *BMC Microbiol.* 12. <https://doi.org/10.1186/1471-2180-12-121>
- Rosen, C., Lennox, J.A., 2001. Multivariate and multiscale monitoring of wastewater treatment operation. *Water Res.* 35, 3402–3410. [https://doi.org/10.1016/S0043-1354\(01\)00069-0](https://doi.org/10.1016/S0043-1354(01)00069-0)
- Rustum, R., Adeloye, A.J., Scholz, M., 2008. Applying Kohonen Self-Organizing Map as a Software Sensor to Predict Biochemical Oxygen Demand. *Water Environ. Res.* 80, 32–40. <https://doi.org/10.2175/106143007x184500>
- Sánchez-Fernández, A., Baldán, F.J.J., Sainz-Palmero, G.I.I., Benítez, J.M.M., Fuente, M.J.J., 2018. Fault detection based on time series modeling and multivariate statistical process control. *Chemom. Intell. Lab. Syst.* 182, 57–69. <https://doi.org/10.1016/j.chemolab.2018.08.003>
- Sanchez-Fernández, A., Fuente, M.J., Sainz-Palmero, G.I., 2015. Fault detection in wastewater treatment plants using distributed PCA methods. *IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA 2015-Octob.* <https://doi.org/10.1109/ETFA.2015.7301504>
- Schmidt, M., Hod, L., 2014. Schmidt, M., Lipson, H. (2013) Eureka (Version 0.98 beta) [Software]. Available from <http://www.eureka.com/>.
- Schmidt, M., Lipson, H., 2009. Distilling Free-Form Natural Laws from Experimental Data. *Science (80-.)*. 324, 81–85. <https://doi.org/10.1126/science.1165893>

- Schütze, M.R., Butler, D., Beck, M.B., 2002. Modelling, Simulation and Control of Urban Wastewater Systems, Modelling, Simulation and Control of Urban Wastewater Systems. Springer London. <https://doi.org/10.1007/978-1-4471-0157-4>
- Shang, L., Liu, J., Turksoy, K., Min Shao, Q., Cinar, A., 2015. Stable recursive canonical variate state space modeling for time-varying processes. *Control Eng. Pract.* 36, 113–119. <https://doi.org/10.1016/j.conengprac.2014.12.006>
- Sheather, S.J., 2006. A Modern Approach to Regression with R, Design. Springer New York LLC, New York. <https://doi.org/10.1016/j.peva.2007.06.006>
- Shewhart, W.A., 1926. Quality Control Charts ¹. *Bell Syst. Tech. J.* 5, 593–603. <https://doi.org/10.1002/j.1538-7305.1926.tb00125.x>
- Slišković, D., Grbić, R., Hocenski, Ž., 2011. Online data preprocessing in the adaptive process model building based on plant data. *Teh. Vjesn.* 18, 41–50.
- Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. *Stat. Comput.* 14, 199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* 45, 427–437. <https://doi.org/10.1016/J.IPM.2009.03.002>
- Sonolikar, R.R., Patil, M.P., Mankar, R.B., Tambe, S.S., Kulkarni, B.D., 2017. Genetic Programming based Drag Model with Improved Prediction Accuracy for Fluidization Systems. *Int. J. Chem. React. Eng.* 15. <https://doi.org/10.1515/ijcre-2016-0210>
- Souza, F.A.A., Araújo, R., Mendes, J., 2016. Review of soft sensor methods for regression

- applications. *Chemom. Intell. Lab. Syst.* <https://doi.org/10.1016/j.chemolab.2015.12.011>
- Spellman, F.R., 2013. *Handbook of Water and Wastewater Treatment Plant Operations*, Handbook of Water and Wastewater Treatment Plant Operations. CRC Press. <https://doi.org/10.1201/b15579>
- Spindler, A., 2014. Structural redundancy of data from wastewater treatment systems. Determination of individual balance equations. *Water Res.* 57, 193–201. <https://doi.org/10.1016/j.watres.2014.03.042>
- Spindler, A., Vanrolleghem, P.A., 2012. Dynamic mass balancing for wastewater treatment data quality control using CUSUM charts. *Water Sci. Technol.* 65, 2148–2153. <https://doi.org/10.2166/wst.2012.125>
- Stanley, K.O., Miikkulainen, R., 2002. Evolving Neural Networks through Augmenting Topologies. *Evol. Comput.* 10, 99–127. <https://doi.org/10.1162/106365602320169811>
- Steyer, J.-P., Rolland, D., Bouvier, J.-C., Moletta, R., 1997. Hybrid fuzzy neural network for diagnosis - application to the anaerobic treatment of wine distillery wastewater in a fluidized bed reactor. *Water Sci. Technol.* 36, 209–217. [https://doi.org/10.1016/S0273-1223\(97\)00525-8](https://doi.org/10.1016/S0273-1223(97)00525-8)
- Szlek, J., Aleksander, M., 2015. fscaret: Automated Feature Selection from “caret.”
- Szłęk, J., Paclawski, A., Lau, R., Jachowicz, R., Kazemi, P., Mendyk, A., 2016. Empirical search for factors affecting mean particle size of PLGA microspheres containing macromolecular drugs. *Comput. Methods Programs Biomed.* 134, 137–147. <https://doi.org/10.1016/j.cmpb.2016.07.006>

- Takács, I., Patry, G.G., Nolasco, D., 1991. A dynamic model of the clarification-thickening process. *Water Res.* 25, 1263–1271. [https://doi.org/10.1016/0043-1354\(91\)90066-Y](https://doi.org/10.1016/0043-1354(91)90066-Y)
- Tay, J.-H., Zhang, X., 2000. A fast predicting neural fuzzy model for high-rate anaerobic wastewater treatment systems. *Water Res.* 34, 2849–2860. [https://doi.org/10.1016/S0043-1354\(00\)00057-9](https://doi.org/10.1016/S0043-1354(00)00057-9)
- Vanrolleghem, P.A., Corominas, L., Flores-Alsina, X., 2010. Real-Time Control and Effluent Ammonia Violations Induced by Return Liquor Overloads. *Proc. Water Environ. Fed.* 2010, 7101–7108. <https://doi.org/10.2175/193864710798207503>
- Wang, L., Shi, H., 2010. Multivariate statistical process monitoring using an improved independent component analysis. *Chem. Eng. Res. Des.* 88, 403–414. <https://doi.org/10.1016/j.cherd.2009.09.002>
- Wąsik, E., Jurík, L., Chmielowski, K., Operacz, A., Bugajski, P., 2017. STATISTICAL PROCESS CONTROL OF REMOVAL OF NITROGEN COMPOUNDS IN THE WASTEWATER TREATMENT PLANT IN KROSNO. *Infrastruct. Ecol. Rural AREAS* 4, 1699–1711. <https://doi.org/10.14597/infraeco.2017.4.2.128>
- Weiland, P., 2008. Wichtige Messdaten für den Prozessablauf und Stand der Technik in der Praxis BT - Messen, Steuern, Regeln bei der Biogaserzeugung : 15. November 2007, Convention Center, Messe Hannover, in: Gülzower Fachgespräche. Fachagentur Nachwachsende Rohstoffe, Gülzow, pp. 17–31.
- West, D., Dellana, S., Jarrett, J., 2002. Transfer function modeling of processes with dynamic inputs. *J. Qual. Technol.* 34, 315–326. <https://doi.org/10.1080/00224065.2002.11980161>

- Wold, S., 1994. Exponentially weighted moving principal components analysis and projections to latent structures. *Chemom. Intell. Lab. Syst.* 23, 149–161. [https://doi.org/10.1016/0169-7439\(93\)E0075-F](https://doi.org/10.1016/0169-7439(93)E0075-F)
- Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: A basic tool of chemometrics, in: *Chemometrics and Intelligent Laboratory Systems*. Elsevier, pp. 109–130. [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1)
- Woo, S.H., Jeon, C.O., Yun, Y.S., Choi, H., Lee, C.S., Lee, D.S., 2009. On-line estimation of key process variables based on kernel partial least squares in an industrial cokes wastewater treatment plant. *J. Hazard. Mater.* 161, 538–544. <https://doi.org/10.1016/j.jhazmat.2008.04.004>
- Xiao, H., Huang, D., Pan, Y., Liu, Y., Song, K., 2017. Fault diagnosis and prognosis of wastewater processes with incomplete data by the auto-associative neural networks and ARMA model. *Chemom. Intell. Lab. Syst.* 161, 96–107. <https://doi.org/10.1016/J.CHEMOLAB.2016.12.009>
- Xibilia, M.G., Gemelli, N., Consolo, G., 2017. Input variables selection criteria for data-driven Soft Sensors design, in: *Proceedings of the 2017 IEEE 14th International Conference on Networking, Sensing and Control, ICNSC 2017*. Institute of Electrical and Electronics Engineers Inc., pp. 362–367. <https://doi.org/10.1109/ICNSC.2017.8000119>
- Yasmin, N.S.A., Wahab, N.A., Anuar, A.N., Bob, M., 2019. Performance comparison of SVM and ANN for aerobic granular sludge. *Bull. Electr. Eng. Informatics* 8, 1392–1401. <https://doi.org/10.11591/eei.v8i4.1605>
- Yoo, C.K., Lee, D.S., Vanrolleghem, P.A., 2004. Application of multiway ICA for on-line process

monitoring of a sequencing batch reactor. *Water Res.* 38, 1715–1732.

<https://doi.org/10.1016/J.WATRES.2004.01.006>

Yordanova, S., Noikova, N., Petrova, R., Tzvetkov, P., 2005. Neuro-Fuzzy Modelling on Experimental Data in Anaerobic Digestion of Organic Waste in Waters, in: 2005 IEEE Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications. IEEE, pp. 84–88. <https://doi.org/10.1109/IDAACS.2005.282946>

Zaghloul, M.S., Hamza, R.A., Iorhemen, O.T., Tay, J.H., 2020. Comparison of adaptive neuro-fuzzy inference systems (ANFIS) and support vector regression (SVR) for data-driven modelling of aerobic granular sludge reactors. *J. Environ. Chem. Eng.* 8. <https://doi.org/10.1016/j.jece.2020.103742>

Zeng, G.M., Li, X.D., Jiang, R., Li, J.B., Huang, G.H., 2006. Fault diagnosis of WWTP based on improved support vector machine. *Environ. Eng. Sci.* 23, 1044–1054. <https://doi.org/10.1089/ees.2006.23.1044>

Zhang, C., Liu, Q., Wu, Q., Zheng, Y., Zhou, J., Tu, Z., Chan, S.H., 2017. Modelling of solid oxide electrolyser cell using extreme learning machine. *Electrochim. Acta* 251, 137–144. <https://doi.org/10.1016/J.ELECTACTA.2017.08.113>

Zhang, Z., Dong, F., 2014. Fault detection and diagnosis for missing data systems with a three time-slice dynamic Bayesian network approach. *Chemom. Intell. Lab. Syst.* 138, 30–40. <https://doi.org/10.1016/j.chemolab.2014.07.009>

UNIVERSITAT ROVIRA I VIRGILI

Data-driven soft-sensors for monitoring and fault diagnosis in wastewater treatment plants

Pezhman Kazemi



UNIVERSITAT
ROVIRA i VIRGILI