

**Faculty of Science and Engineering  
Department of Chemical Engineering**

**Development of an Intelligent Dynamic Modelling System for the  
Diagnosis of Wastewater Treatment Processes**

**Muhammad Imran Khalid**

**This thesis is presented for the Degree of  
Master of Philosophy (Chemical Engineering)  
of  
Curtin University of Technology**

**March 2010**

## **Declaration**

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: Imran Khalid.....

Date: 11/10/2010

# Abstract

In the 21<sup>st</sup> Century, water is already a limited and valuable resource, in particular the limited availability of fresh water sources. The projected increase in global population from 6 billion people in 2010 to 9 billion in 2050 will only increase the need for additional water sources to be identified and used. This situation is common in many countries and is frequently exacerbated by drought conditions. Water management planning requires both the efficient use of water sources and, increasingly, the re-use of domestic and industrial wastewaters. A large body of published research spanning several decades is available, and this research study looks specifically at ways of improving the operation of wastewater treatment processes.

Process fault diagnosis is a major challenge for the chemical and process industries, and is also important for wastewater treatment processes. Significant economic and environmental losses can be attributed to inappropriate Abnormal Event Management (AEM) in a chemical/processing operation, and this has been the focus of many researchers. Many researchers are now focusing on the application of several fault diagnosis techniques simultaneously in order to improve and overcome the limitations experienced by the individual techniques. This approach requires resolution of the conflicts ascribed to the individual methods, and incurs additional costs and resources when employing more than one technique. The research study presented in this thesis details a new method of using the available techniques. The proposal is to use different techniques in different roles within the diagnostic approach based upon their inherent individual strengths. The techniques that are excellent for the detection of a fault should be employed in the fault detection, and those best applied to diagnosis are used in the diagnosis section of a diagnostic system.

Two different techniques are used here, namely a mathematical model and data mining are used for detection and diagnosis respectively. A mathematical model is used which is based upon the principal of analytical redundancy in order to establish the presence of a fault in a process (the fault detection), and data mining is

used to produce production rules derived from the historical data for the diagnosis. A dataset from an industrial wastewater treatment facility is used in this study.

A diagnostic algorithm has been developed that employs the techniques identified above. An application in Java was constructed which allows the algorithm to be applied, eventually producing an intelligent modelling agent. Thus the focus of this research work was to develop an intelligent dynamic modelling system (using components such as mathematical model, data mining, diagnostic algorithm, and the dataset) for simulation of, and diagnosis of faults in, a wastewater treatment process where different techniques will be assigned different roles in the diagnostic system.

Results presented in Chapter 5 (section 5.5) show that the application of this combined technique yields better results for detection and diagnosis of faults in a process. Furthermore, the dynamic update of the set value for any process variable (presented in Chapter 5, section 5.2.1) makes possible the detection of any process disturbance for the algorithm, thereby mitigating the issue of false alarms. The successful embedding of both a detection and a diagnostic technique in a single algorithm is a key achievement of this work, thus reducing the time taken to detect and diagnose a fault. In addition, the implementation of the algorithm in the purpose-built software platform proved its practical application and potential to be used in the chemical and processing industries.

# Table of contents

Declaration .....	I
Abstract .....	II
Table of contents.....	IV
List of Figures.....	VII
List of Tables.....	VIII
Acknowledgements.....	IX
<b>1 Research Overview.....</b>	<b>1</b>
1.1 Background .....	1
1.2 Thesis Objectives .....	2
1.3 Thesis Outline .....	3
<b>2 Literature Review.....</b>	<b>5</b>
2.1 Introduction.....	5
2.2 Process Diagnosis.....	5
2.2.1 Process Model Based Methods .....	7
2.2.2 Process History Based Methods.....	9
2.3 Hybrid Diagnosis Systems .....	9
2.4 Software Platform.....	10
2.5 Wastewater Treatment.....	12
2.6 Concluding Remarks .....	13
<b>3 Process Model-Based Methods for Wastewater Treatment Processes... 15</b>	
3.1 Introduction.....	15
3.2 Wastewater.....	15
3.3 Wastewater Treatment.....	16
3.4 Preliminary Treatment Stage .....	18
3.5 Primary Treatment Stage .....	19
3.6 Secondary Treatment Stage .....	19
3.6.1 Biological Treatment .....	21

3.7	Tertiary Treatment Stage .....	22
3.8	PMB Techniques .....	22
3.8.1	Qualitative Model-Based Techniques .....	22
3.8.2	Quantitative Model-Based Techniques.....	24
3.9	Mathematical Modelling .....	25
3.10	Wastewater Treatment Mathematical Model.....	26
3.11	Results from Model Simulations .....	31
3.11.1	Batch Process .....	32
3.11.2	Completely Mixed Reactor .....	35
3.12	Model Validation.....	38
3.13	Summary Comments .....	39
<b>4</b>	<b>Process History-Based Methods .....</b>	<b>40</b>
4.1	Introduction.....	40
4.2	PHB Techniques.....	40
4.3	Data Mining .....	41
4.3.1	Data Mining Process Description.....	42
4.3.2	Applications for Data-Mining .....	44
4.3.3	Data Mining Approaches .....	45
4.3.3.1	Descriptive approaches .....	45
4.3.3.2	Predictive approaches .....	46
4.4	Inductive Data Mining.....	48
4.5	Wastewater Treatment Plant (WWTP) dataset .....	51
4.6	Development of Production Rules from WWTP data.....	54
4.7	Weighting of Production Rules .....	58
4.8	Validation of Production Rules .....	66
4.9	Concluding Remarks .....	66
<b>5</b>	<b>Diagnostic Algorithm and Java Application.....</b>	<b>67</b>
5.1	Introduction.....	67
5.2	Diagnostic Algorithm .....	67
5.2.1	The Algorithm.....	68
5.3	Software Platform.....	77

5.4 Architecture of the Application.....	79
5.5 Results and Discussion .....	82
5.6 Concluding Remarks .....	86
<b>6 Conclusions and Recommendations.....</b>	<b>88</b>
6.1 Conclusions.....	88
6.2 Recommendations for Future Work .....	89
References.....	90
Appendix .....	98

## List of Figures

3.1 Structure of a wastewater treatment plant.....	18
3.2 Simulation results on batch process using excess DO.....	33
3.3 Simulation result from batch process with extended residence time .....	33
3.4 Effect of oxygen concentration on wastewater treatment.....	35
3.5 Simulation dynamics for a continuous process using excess DO.....	36
3.6 Wastewater treatment dynamics for controlled oxygen supply.....	37
4.1 An overview of the knowledge discovery process.....	44
5.1 A schematic of fault detection and diagnosis.....	68
5.2 Integrated view of Fault detection.....	73
5.3 An integrated view of the components in the diagnosis section.....	77
5.4 Graphical User Interface (GUI) of the application.....	79
5.5 Flowsheet screen of the WWTP application.....	80
5.6 Detection and Diagnosis Algorithm command window.....	81
5.7 Screen for the display of the diagnostic results.....	82
5.8 Input of process parameters for simulation.....	83
5.9 Graphical presentation of the simulation results.....	83
5.10 Diagnostic results for high SS value.....	84
5.11 Simulation indicating low SS value.....	85
5.12 Set of diagnostic results for low value of SS.....	85



## List of Tables

3.1 Parameters used for batch process simulation.....	33
3.2 Simulation parameters for continuous process.....	36
3.3 Parameters used to study effect of oxygen.....	37
3.4 Parameters used in modelling and simulation with literature values.....	38
3.5 Results for Model Validation.....	39
4.1 Process variables of the wastewater treatment plant.....	52
4.2 WWTP data in different classes.....	56
4.3 Production rules for SS-S.....	57
4.4 Production rules for low and high class by their weight function.....	60
4.5 Production rules for SS-S arranged by weight function.....	61
4.6 Production rules for VSS-S arranged by weight function.....	62
4.7 Production rules for SED-S arranged by weight function.....	63
4.8 Production rules for DBO-S arranged by weight function.....	64
4.9 Production rules for DQO-S arranged by weight function.....	65
4.10 Validation of Production Rules.....	66
5.1 The fault libraries used in the knowledge base of the diagnosis system.....	75
5.2 Results of the diagnostic methodology.....	86

## **Acknowledgements**

I would like to thank my academic supervisor Dr. Nicoleta Maynard and co-supervisor Dr. Martyn Ray of Curtin University of Technology for their guidance and help in accomplishing this research task.

I would also like acknowledge Curtin University for providing me with the Curtin International Postgraduate Research Scholarship (CIPRS).

Muhammad Imran Khalid

# Chapter 1

## Research Overview

### 1.1 Background

Water sources, and their use, cannot be considered as an unlimited resource. The treatment and reuse of wastewater is now accepted practice worldwide and within Australia. Wastewater treatment is a very active area of research yielding many publications, however there are many challenges and significant aspects requiring improvement within the wastewater industries. In broad terms, there are two types of wastewater, namely municipal and industrial. Although different operations produce different wastewaters, the schematic process applied to water treatment is often very similar.

Wastewater treatment facilities are mostly government or state operated and, unlike other chemical and process industries, there is no revenue directly generated from the end product (the reclaimed water), except for reducing the use and cost of freshwater. Reviewing the treatment process highlights the need to increase the efficiency and effectiveness of operation by using a robust fault detection and diagnostic system, optimization of utilities, work force, assets, etc. Models have been developed to assist the engineers and operators understanding of the process, and hence apply effective fault detection and diagnostics and optimization techniques, e.g. Activated Sludge Model (ASM) series (Gujer et al., 1999) and Anaerobic Digestion Model (Copp et al., 2005).

Process fault diagnosis is one of the major challenges facing the chemical and process industries and significant economic and environmental losses can be attributed to poor Abnormal Event Management (AEM). There are two key approaches to fault diagnosis, namely process model based diagnosis and data driven diagnosis. Process model based approaches assume that a fault will cause changes to certain physical parameters, which in turn will lead to changes in some of the model parameters or states. It is then possible to detect and diagnose these faults by monitoring the estimated model parameters or states. This technique uses a

diagnostic-driven mathematical model of a process (Simani et al., 2003) and generally exhibits good accuracy, since the process models are developed from the underlying fundamental principles. However, comprehensive theoretical models of complex processes are extremely challenging to develop due to their often inherent non-linear nature. Alternatively the use of computer aided systems means that data driven models are relatively easy to develop, and analysis of the data enables fault identification and diagnosis (Li, 2003).

A comprehensive review of the techniques available for fault detection and diagnostics, the strengths and challenges of each technique, and the key attribute required by an ideal diagnostic system, is given by Venkatasubramanian et al. (2003b, 2003c, 2003a). They determined that no single method has all the ideal desirable features for a diagnostic system, but some of these methods can complement one another resulting in an improved overall diagnostic system. The use of a hybrid system is a promising future research direction leading to developments in diagnostic systems. Integrating these complementary features is one way to develop hybrid systems that may overcome the limitations of an individual strategy. The drawbacks of single-method-based diagnostic systems are serious enough to limit their applications to small case studies and render them unsuitable for large-scale industrial situations. This makes the design and development of hybrid systems important (Venkatasubramanian et al., 2003c).

Few papers have been published on wastewater treatment in relation to process monitoring and control and fault diagnostics (Lee et al. 2006). Since 1998, 29 journal papers were published concerning the application of hybrid fault diagnostic systems. There are only two publications for fault detection and mathematical models (in 2003 and 2008).

## **1.2 Thesis Objectives**

The main objective of this research work was to develop an intelligent modelling system to be used for dynamic simulation, and the diagnosis of wastewater processes using a hybrid of model based (mathematical model) and data driven (inductive data mining) to address the limitations exhibited by the individual approaches. Particular

emphasis was placed upon the development of a hybrid diagnostic system for its reliability, flexibility, robustness, relative low-cost of development, and ease of utilization in wastewater treatment processes. The overall objective of this thesis was to develop an effective fault detection and diagnostic platform which employs a hybrid system of a mathematical model and a data driven technique to overcome the limitations of the individual fault detection and diagnostic methodologies. This was achieved by the following approach:

- Develop a software platform using Java for dynamic simulation and fault detection on the wastewater treatment process; this includes the implementation of the model libraries of full-scale wastewater treatment processes.
- Select the most suitable data driven technique for fault detection and then implement on the wastewater treatment data.
- Implement the hybrid algorithm in the Java-based platform; this includes the development of the hybrid algorithm.
- Validate the results against wastewater treatment process data.

### **1.3 Thesis Outline**

The organization of the thesis is as follows:

Chapter 2 discusses in detail the process diagnostics including an overview, the techniques available, and a critical literature review of current developments identifying the limitation of process diagnostics.

Chapter 3 focuses on the wastewater process. It describes briefly the types of operating units in a wastewater process before presenting a mathematical model to be used for the fault detection. It also includes the rationale which supports implementation of the derived model over other available process models.

Chapter 4 provides a comparative study of available data driven techniques leading to the selection of the preferred data driven technique (inductive data mining), the wastewater data is then used to generate results. Discussion on the output format of results and manipulation of the format for use in the diagnostic system is also included.

Chapter 5 contains the development of the software platform in Java and discusses its flexibility, ease of use, and customization. The wastewater model is imported and initial simulation results are discussed. It includes the development of the algorithm for a hybrid approach of the model plus data driven technique. The results and validation of the wastewater data are presented together with the accuracy and the computational efficiency.

Chapter 6 presents the research conclusions, and identifies areas of research recommended for further study.

# Chapter 2

## Literature Review

### 2.1 Introduction

This chapter provides a general overview of process diagnosis and wastewater treatment, in addition to a critical review of selected topics. This chapter includes details of hybrid diagnosis systems and the software platform used in wastewater treatment and process diagnosis.

### 2.2 Process Diagnosis

Diagnosis is defined as “the process of identifying the nature and causes of certain phenomenon”. It has been observed that the word “diagnosis” is always attributed to abnormal phenomenon rather than the normal situation. Similarly, process diagnosis is the study used to identify the nature and causes of an abnormal phenomenon occurring in a process. The abnormal phenomenon in the processing and manufacturing industries is termed a “fault”. Himmelblau (1978) described a fault as a departure from an acceptable range of an observed variable or a calculated parameter associated with a process. Hence, fault is a process abnormality or symptom, such as high temperature in a reactor or low product quality, etc. The underlying causes of this abnormality are called basic events, or root causes. The basic event is also referred to as a malfunction or a failure.

In broad terms, a fault is generally related to one (or more) of three main classes of failures or malfunctions. These are:

*Gross parameter changes in a model:* Parameter changes arise when there is a disturbance entering the process from the environment. An example is the change in the heat transfer coefficient due to fouling of a heat exchanger.

*Structural changes:* Structural changes refer to changes in the model itself. They occur due to the equipment hardware failures. An example is a controller failure

which would imply that the manipulated variable is no longer functionally dependent on the controlled variable.

*Malfunctioning sensors and actuators:* Gross errors usually occur with actuators and sensors. These could be due to a fixed failure, a constant bias (positive or negative) or an out-of-range failure.

Process diagnosis is the study undertaken to identify the nature and causes of a fault in a process. This process usually includes the following four steps (Chiang et al., 2001):

*Fault Detection* - determining if a fault has occurred by observing the values of process variables or by user-defined fault indices.

*Fault Identification* - identifying the variable most relevant to the fault. The task of this step is to focus the attention of the plant operator on the particular subsystem which is most pertinent to the fault.

*Fault Diagnosis* - isolating the cause, type, location, and time of the fault.

*Process Recovery* - removing the fault and bringing the process back to normal conditions, or taking optimal steps to minimize loss of production.

Process diagnosis was traditionally performed by the process operators but this task has become more difficult due to the variety, uncertainty and time delay of malfunctions which can be very complex in their nature,. In the last decade, computer-aided systems have been investigated, implemented and proved to be a successful tool for fault diagnosis. A fault diagnostic system has two main components:

- (i) type of knowledge used, and
- (ii) type of diagnostic search strategy.

Diagnostic search strategy is usually strongly dependent on the knowledge representation scheme, which in turn is largely influenced by the type of prior knowledge available. Hence, the type of prior knowledge used is the most important distinguishing feature in diagnostic systems. The prior domain knowledge may be developed from a fundamental understanding of the process using a first principles approach. Such knowledge is referred to as deep, causal, or model-based knowledge (Simani et al., 2003). However it may also be obtained from past experience of the



process and is then referred to as shallow, compiled, evidential, or process-history-based knowledge.

The process diagnostic techniques can be broadly classified into two categories, i.e. process-model-based methods and process-history-based (data-driven) methods as discussed below.

### **2.2.1 Process Model Based Methods**

These methods use qualitative knowledge and quantitative models extracted from an understanding of the process principles. The models present the interacting relationships between the process variables based upon the assumption that a fault will cause changes to certain physical parameters which in turn lead to changes in some of the model parameters or states. It is then possible to detect and diagnose these faults by monitoring the estimated model parameters or states.

The process model based methods can be further sub-divided into qualitative causal models and quantitative methods (Dash and Venkatasubramanian, 2000, Venkatasubramanian et al., 2003a, 2003b).

#### ***Quantitative model based methods***

Relying on an explicit model, all model based fault detection and isolation (FDI) methods require two steps. The first step is to generate inconsistencies between the actual and expected behavior, these are called residuals and reflect the potential faults of the system. The second step chooses a decision rule for diagnosis.

Some form of redundancy is required to check for the inconsistencies. There are two types of redundancies, namely hardware and analytical redundancy. Hardware redundancy uses redundant sensors and has been utilized in the control of safety-critical systems such as aircraft, space vehicles and nuclear power plants, but its applicability is limited due to the additional costs and space required. Alternatively, analytical redundancy is achieved from the functional dependence among the process variables and is usually provided by a set of algebraic or differential relationships among the states, the inputs, and the outputs of the system (Lou et al., 1986, Michle, 1988).

The essential element of analytical redundancy is to check the actual system behavior against the system model for consistency. Any inconsistency, expressed as residuals, can be used for detection and isolation purposes. The residuals should be close to zero when no fault occurs, but show a significant value when the underlying system changes. To generate the diagnostic residuals requires an explicit mathematical model of the system. The model may be obtained either analytically using first principles or empirically as a black-box model.

Most of the work on model-based diagnostic systems reported to date was mainly in the aerospace, mechanical or electrical engineering literature. There has not been much published research on its application for fault diagnosis in chemical process systems. There are some serious limitations that apply for its application in the chemical industries. One issue is the lack of availability, and the complexity, of models for chemical processes and their inherent non-linear nature. In addition to the modelling challenges, the model based qualitative methods do not include an explanation and descriptive facility. Furthermore, an estimation of classification errors cannot be provided when using these methods. Another disadvantage with these methods is that if a fault is not specifically and appropriately modelled, then there is no guarantee that the residuals will be able to detect it.

### ***Qualitative model based methods***

Qualitative models are usually developed based on the fundamental understanding of the physics and chemistry of the process. Various forms of qualitative models such as causal models and abstraction hierarchies have been developed. The strategy employed in qualitative models is causal-effect reasoning related to the system behavior. The most popular methods are fault-trees and signed digraphs (SDG). Fault trees (Lapp and Powers, 1977) use backward chaining until a primary event is found that presents a possible root cause for the observed process deviation from normal operation. SDG (Iri et al., 1979) is another representation of the causal information in which the process variables are represented as graph nodes and causal relations by directed arcs. Causal model-based methods mimic human reasoning and so generation of explanation is relatively straightforward making them more interactive. There were drawbacks of the early SDG methods mainly because their expressive capability is often limited (Hunag and Wang, 1999a). In an SDG, a node or a branch

can often only take three values. i.e., -, 0, and + representing for example low, normal and high values for a node. This over simplified expression could create ambiguous solutions.

The two main concerns with qualitative model based methods are ambiguities and spurious/inauthentic solutions. Considerable research has been done in relation to the reduction of spurious solutions while reasoning with qualitative models.

### **2.2.2 Process History Based Methods**

In contrast to the model-based approaches where a priori knowledge (either quantitative or qualitative) about the process is needed, in process history based methods, only the availability of a large amount of historical process data is needed. There are different ways in which this data can be transformed and presented as a priori knowledge to a diagnostic system. This is known as feature extraction, and this extraction process can be either qualitative or quantitative in nature. Two of the major methods that extract qualitative history information are expert systems and trend modelling methods. Methods that extract quantitative information can be broadly classified as non-statistical or statistical methods. Neural networks are an important class of non-statistical classifiers. Principal component analysis (PCA)/partial least squares (PLS), data mining and statistical pattern classifiers form a major component of statistical feature extraction methods. The knowledge can be available as rules and formulations.

## **2.3 Hybrid Diagnosis Systems**

Venkatasubramanian and co-workers (see Venkatasubramanian et al., 2003b, 2003c, 2003a) have provided a very comprehensive review of the methods available for process diagnosis. A set of desirable characteristics that a diagnostic system should possess were also identified and listed. Different approaches were evaluated against a common set of requirements or standards. From their evaluation, it was revealed that no single method has all the desirable features stipulated for a diagnostic system. It was postulated that some of these methods can complement one another, resulting in better diagnostic systems. Integrating these complementary features is one way to develop hybrid methods that could overcome the limitations of individual solution strategies. Hence, hybrid approaches are attractive where different methods work in

conjunction to solve parts of the problem. Although all the methods possess limitations, in the sense that they are only as good as the quality of information provided, it was shown that some methods might be better adapted to the knowledge available than others. For example, fault explanation through a causal chain is best done through the use of digraphs, whereas fault isolation might be very difficult using digraphs due to the qualitative ambiguity and then analytical model-based methods might be superior. It is expected that hybrid methods will provide a general, powerful problem-solving platform.

## **2.4 Software Platform**

One of the challenges in the implementation of a diagnostic system is the software architecture (Venkatasubramanian et al., 2003c). The diagnostic task can be performed either off-line or on-line. In an off-line diagnostic task, the process behavior is recorded in the form of data or graphical trends. This data is then analyzed off-line using a suitable diagnostic method. This type of diagnostic is often used as a preventive action to save the process from repeating the same malfunction. Due to the complexity of chemical processes and the significant losses attributed to the poor management of faults, the online diagnostic is increasingly employed in the chemical industries for corrective actions. The online diagnostic system helps operators and engineers to manage an abnormal event (fault) as soon as it is encountered in the process. Using its knowledge base, the online diagnostic system will search for the most likely culprit for a specific fault. This vital information, along with the prior knowledge that the operators and engineers have about the process, will be critical in the isolation and diagnosis of a fault, thus reducing the losses and downtime and increasing process and personnel safety associated with the management of an abnormal event (fault).

From the literature review (Venkatasubramanian et al., 2003c), a software platform intended to be used for online diagnosis should possess the following characteristics.

### **i) Flexibility**

Flexibility refers to the ability of software to accommodate different configurations of plant items and an ability to use different types of equipment for any unit

operation or process. In relation to wastewater treatment where different configurations exist in different facilities due to variations of influents within a region (state or country), the software should have the timely ability to accommodate for such changes in configuration.

ii) Detection algorithm

The software should incorporate an algorithm that is capable of detecting any abnormality in the process. The algorithm will be highly dependent on the type of benchmarking information used to detect faults, i.e. if a mathematical model is used then the algorithm should be able to: (a) obtain values from the mathematical model; (b) obtain information from real processes; and (c) should have the threshold value for noise/bias in real data. After the comparison of values of a process variable from real operations and benchmarking, the algorithm will be able to detect a fault and draw the attention of operators to that variable.

iii) Data/Information import

Based on the detection algorithm requirements, it is essential that the software should be able to obtain information from the control system of a process for use in fault detection. The integration of a diagnostic software platform with the control module in a plant is a major challenge that requires precision skills in software architecture, and the availability of resources (e.g. a distributed control system of a pilot scale plant) for trials to validate the effectiveness, efficiency and robustness to the software platform.

iv) Diagnostic algorithm

After the detection of a fault, the most important aspect of the diagnostic system is to then diagnose the fault. This necessitates a diagnostic algorithm that will diagnose the fault using the knowledge-base used for diagnosis purposes.

v) Results display

An essential ingredient that a diagnostic software platform should possess is the display of diagnostic results in an easily understood format. This may be production rules, fault trees, decision trees, or any other acceptable form of results.

vi) Cost effectiveness

Although management of a fault in a process is essential, and is associated with significant economic benefits, it would be an optimal solution to have a cost effective diagnosis system without compromising the effectiveness of the system. Commercial software such as Gensym G2, which is widely used in the chemical industries for

better management of abnormal process conditions, costs between \$100,000 and \$1 million. Although it may be cost effective, this is a critical investment decision for small-scale and service industries.

## **2.5 Wastewater Treatment**

Water recycling and reuse is now generally accepted and adopted worldwide as an essential water resource. The average composition of municipal wastewater is 99.94% H<sub>2</sub>O and 0.06% dissolved and suspended solids, which indicates the potential for the reuse of waste water.

Wastewater treatment facilities can process wastewater using either an aerobic, anaerobic or anoxic process. Each process has its own advantages and disadvantages. Wastewater treatment facilities are mostly government or state operated and, unlike other chemical and process industries, there is no direct revenue generated from the end product, i.e. the reclaimed water (although it can be used for irrigation or drinking purposes). Hence there is more attention given to reducing the running costs by optimization of utilities, workforce, assets, etc. Models have been developed to assist the engineers and operators understanding of the process, e.g. Activated Sludge Model (ASM) series (Gujer et al., 1999) and Anaerobic Digestion Model (Copp et al., 2005). Wastewater processes have been widely researched in recent years. Lee et al. (2006) addressed fault diagnosis of sensors in wastewater treatment processes but the methodology proposed is unable to identify the faulty sensor which causes process transitions. Kim et al. (2002) calibrated ASM1 using genetic algorithms but the components of ASM1 were not calibrated in detail. Gernaey et al. (2004) used artificial intelligence and white-box based modelling and simulation for wastewater treatment processes, and demonstrated how different methodologies can complement and support the process knowledge included in white-box activated sludge models. Puteh et al. (1999) present a mathematical model of the aeration tank and the secondary settler of a wastewater treatment facility, proposing that the performance of wastewater treatment processes consisting of an incomplete mixing reactor described by tanks-in-series model is better than that of a completely mixed aeration tank. Rigger et al. (2006) proposed a model for the response time of the aeration systems concluding that if more calibrated applications of the aeration system model are available, it should be possible to develop a classification system for design

purposes. Wintgens et al. (2003) presented the modelling of a membrane bioreactor replacing the conventional aeration system in wastewater treatment processes, which is more efficient than the traditional aeration tank in terms of investment cost but the operating cost is higher. Van Hulle and Vanrolleghem (2004) showed that model-based optimization is an efficient and cost-effective way to ensure that an industrial wastewater treatment plant functions well, but a more holistic evaluation is required before the proposed methodology can be applied.

Much research on wastewater is related to the treatment of industrial wastewaters with very specific requirements. Acharya et al. (2009b,2009a) used activated carbon prepared from Tamarind wood with zinc chloride activation for the removal of lead (II) and chromium (VI) from industrial wastewater. Hunag et al. (2009) developed an integrated neural-fuzzy process controller to control aeration in an aerated submerged biofilm wastewater treatment process (ASBWTP) which saved 33% of the operating costs during the time when the controller was used. Pai (2008) employed grey models (GM) to predict the effluent quality of a wastewater and compared the results with the use of artificial neural networks (ANN). The results indicated that GM can be used for effluent prediction while using less data than required in ANN. The amount of research work reported on wastewater is significant, and wastewater is indeed one research area that has attracted major attention worldwide.

## **2.6 Concluding Remarks**

Process diagnosis is a key research area receiving significant attention from both academia and industry. In recent years this has led to significant process improvements in abnormal event management with proposals for numerous new techniques. With the development of these techniques, the focus of research should now be on the development of a hybrid methodology that can address the weaknesses of individual techniques in order to further enhance the effectiveness of process diagnostics. So far, this area has not been studied in depth.

By comparison with other chemical processes, the wastewater treatment process is quite small in terms of the processes involved and the physical size of a facility. Furthermore, as discussed in Section 2.5, a municipal wastewater facility does not

generate any direct revenue from its process, unless it is involved in the treatment of an industrial wastewater. Hence, the optimization of capital and operating costs must be a key focus. A purpose-built, cheap and effective software platform that can be used for process diagnosis in wastewater treatment, or any other small chemical industry, is one application for budget optimization. A significant contribution of this research study is the development of a flexible and low-cost software platform.



# Chapter 3

## Process Model-Based Methods for Wastewater Treatment Processes

### 3.1 Introduction

In this chapter, Sections 3.1 to 3.7 present the necessary background understanding of the typical wastewater treatment process. There is a significant amount of published literature on wastewater treatment processes and this large volume of information can pose a problem for someone new to this field needing an overview of the essential processes involved in water treatment. Therefore Sections 3.1-3.7 present a comprehensive, yet concise, review of wastewater and its treatment.

Process Model-Based (PMB) methods/techniques for Fault Detection and Diagnosis (FDD) are then discussed from Section 3.8. PMB methods are generally classified as quantitative or qualitative. An introduction and brief discussion of the quantitative and qualitative PMB techniques are presented, followed by explanation of how mathematical modelling will be used in this work for FDD. A mathematical model of a municipal wastewater treatment plant is presented in Section 3.10. Results obtained from the process simulations are reported and discussed in Section 3.11.

### 3.2 Wastewater

Wastewater is a general term for any water that has been used for either domestic or industrial purposes, and hence becomes contaminated by various waste materials. It can contain human excreta, food waste, and industrial toxins, along with many other pollutants in the form of dissolved and suspended material, and hence is unfit for human consumption and can damage aquatic systems. The composition of wastewater in terms of the water itself and the contained wastes varies widely

between countries, and within each country, but the average composition of a wastewater is around 99% water and 1% solid.

Water shortages and deterioration in quality are major challenges faced by many countries, especially those pursuing economic and social development. Due to population growth and the associated increased use of water for agriculture, industry and recreation, the human consumption of natural waters has steadily increased over several centuries thus making it an increasingly valuable resource. This situation had lead many researchers to think about whether there will be enough water to accommodate the needs of future generations (Kumar, 2004). Water management strategies and specific techniques are being adopted in many countries for optimum water utilization based on defined water policies. Wastewater treatment compliments water management in two ways: (1) it increases the total available water resource by converting the wastewater into useable water; and (2) when the treated wastewater is discharged into a receiving body (lakes or rivers) after treatment, it does not deteriorate the overall water quality by being free from pollutants that may affect the aquatic system.

The composition of wastewater indicates the great potential for the application of wastewater treatment processes. Wastewater treatment is being increasingly adopted worldwide for optimizing water management systems.

### **3.3 Wastewater Treatment**

Wastewater treatment is defined as the processing of wastewater for the removal or reduction of all undesirable constituents. Three types of processes are involved in the treatment of wastewater:

- a) Mechanical,
- b) Biological, and
- c) Chemical.

Mechanical or physical treatment processes are used for the removal of large objects, heavy inorganic matter, oil, greases and particulate solids from the raw wastewater before it is treated using biological processes.

The biological process then converts the influent from mechanical treatment unit, to be almost free from dissolved and suspended solids (pollutants). This goal is achieved by the action of microorganisms that thrive on the pollutants in wastewater

for their nutrients. Oxygen is required in this treatment stage as it is essential for the survival of living organisms.

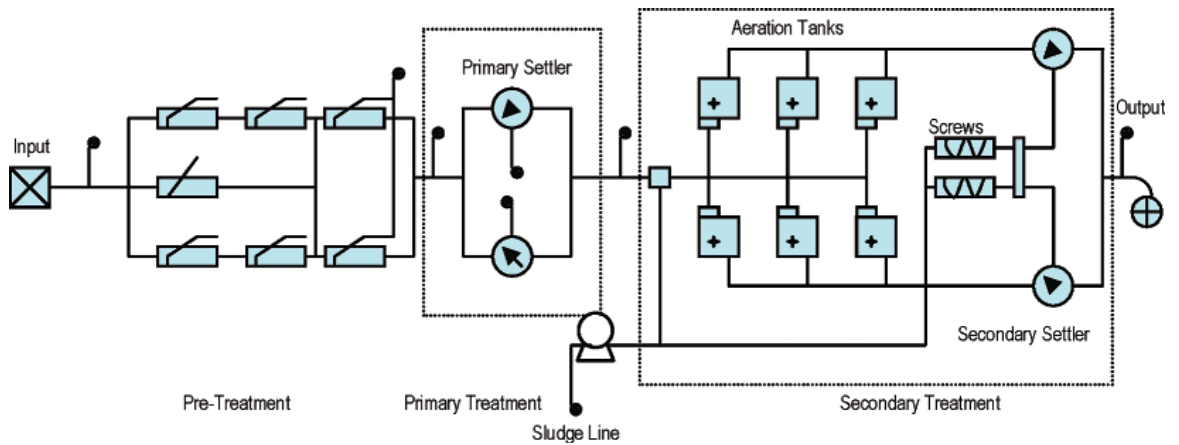
The main purpose of chemical treatment is to achieve the final required water quality (specification) before it can be sent to a receiving body. Chemical agents, such as chlorine gas, are mainly used for disinfection of water in order to kill any microorganisms remaining after the biological treatment stage.

In any type of wastewater treatment plant, there are essentially the following stages:

- a) Preliminary treatment,
- b) Primary treatment,
- c) Secondary treatment,
- d) Tertiary treatment and
- e) Advanced treatment systems.

The last stage, i.e. tertiary treatment, depends on the specified required quality for the effluent water and, unless a high quality of water is required, this stage is seldom used in treatment plants.

A typical preliminary treatment stage consists of screens, comminutors and grits. Only mechanical treatment is carried out in the preliminary stage. In the primary treatment stage, heavy solids and oils or greases are separated from the wastewater using gravity action in a sedimentation tank often called a primary settler/clarifier. Effluent from the primary settler/clarifier is treated in a bio-reactor where microorganisms eliminate, or significantly reduce, the amount of dissolved and suspended matter (pollutants). The effluent is then passed to a secondary settler where the clean water as supernatant is either released into the receiving body or is passed to tertiary treatment for further purification. Settled solids at the bottom of the settler are either recycled back to the reactor or disposed of after further treatment. The bio-reactor and the secondary settler comprise the secondary stage for wastewater treatment. An example of a typical wastewater treatment plant is given in the schematic of Figure 3.1.



**Figure 3.1** Structure of a wastewater treatment plant (Wang et al., 2004)

Wastewater has to pass through these three treatment stages before it can be released back into the environment. The following section provides further details of these stages and the equipment used in each stage.

### 3.4 Preliminary Treatment Stage

The preliminary stage consists of screens, comminutors and grits. A screen is a device with openings of uniform size that is used to retain solids present in the influent wastewater. Its principal role is to remove coarse material from the influent that could damage subsequent process equipment (Tchobanoglous et al., 2003). Screens are followed by Comminutors which are used to reduce the particle size of wastewater solids as large, stringy solids can easily plug pump impellers (Degremont., 1991, Forster, 2003). The last unit in the primary treatment is Grit chamber which is used to remove the heavy inorganic material present in wastewater, such as sand, eggshells, gravel and cinders, which have settling velocities and specific gravities substantially greater than the organic solids in the wastewater. (Tchobanoglous et al., 2003). Only mechanical treatment of wastewater is carried out with the aim to eliminate the large sized materials and heavy inorganic material from the wastewater before it can enter the primary settler for further purification.

### 3.5 Primary Treatment Stage

After the preliminary treatment of raw wastewater, the remaining solids are extracted by gravity in large sedimentation tanks in the primary stage of treatment. Sedimentation tanks further slow the influent flow of wastewater so that organic and inorganic suspended solids can settle to the bottom of the clarifiers. Floatable solids and grease are skimmed off by a rotating arm and deposited as a scum. A clarified supernatant liquid leaves from the top of the sedimentation tank, while concentrated sludge exits from the bottom of the sedimentation tank. The primary clarifiers remove about 60% of the Total Suspended Solids and about 30% of the Biochemical Oxygen Demand in the incoming wastewater. Two types of sedimentation tanks/clarifier are used in wastewater treatment plants, namely rectangular tanks and circular tanks (Tchobanoglous et al., 2003).

### 3.6 Secondary Treatment Stage

The secondary treatment stage of a wastewater treatment process is the most important as the removal of organic material (which represents the pollutants in wastewater) takes place in this stage. It comprises a biological unit that facilitates the growth of microorganisms requiring a sufficient supply of oxygen. The microorganisms use the organic material in the wastewater stream as food, thus reducing the pollutant contents in the wastewater. Treated wastewater from the biological unit is then passed to a secondary clarifier/tank where the solids settle to the tank bottom. The treated effluent stream is either released into the receiving body, or is subjected to tertiary treatment if a high quality of water purity is required. A brief overview of the principal types of reactors used in this stage is given below.

#### a) Batch Reactors

A batch reactor used for wastewater treatment will incorporate the following operational phases (Buchanan and Seabloom, 2004):

- i. Fill:* Raw wastewater that has been through primary treatment is added to the reactor. During this phase, aeration may or may not be supplied in order to provide alternating periods of high or low dissolved oxygen. This mode may occupy 25% of the total cycle time.

*ii. React:* Aeration is provided in an effort to obtain rapid biodegradation of organic compounds. This mode will typically require about 35% of the total cycle time.

*iii. Settle:* Aeration is shut off to allow the wastewater to become anoxic (for denitrification) and to allow for quiescent conditions that allow very effective liquid-solid separation. Clarification will usually take about 20% of the overall cycle time.

*iv. Draw:* Clarified raw water is removed as the supernatant liquid. The decanting is accomplished using adjustable or floating weirs. Periodically the excess biosolids must be removed. Decanting generally takes about 15% of the total cycle time.

An important requirement in the batch reactor process is that a tank is never completely emptied, and a portion of the settled solids are left to seed the next cycle. This allows the establishment of a population of organisms uniquely suited to treating the wastewater. By subjecting the organisms to periods of high and low oxygen levels, and to high and low food availability, the population of organisms becomes very efficient at treating wastewater. A typical hydraulic retention time for a batch reactor varies between 20 to 40 hours (Tchobanoglous et al., 2003).

#### **b) Complete Mix Reactors**

It is assumed that a complete mixing occurs instantaneously and uniformly throughout a complete mix reactor as fluid particles enter the reactor. Fluid particles leave the reactor in proportion to their statistical population. The actual time required for complete mixed conditions depend on the reactor geometry and the power input.

#### **c) Plug-Flow Reactors**

Fluid particles pass through the reactor with a little or no longitudinal mixing and exit from reactor in the same sequence in which they enter. The particles retain their identity and remain in the reactor for a time equal to the theoretical detention time.

#### **d) Packed Bed Reactors**

A packed bed reactor is filled with some type of packing material, such as rock, slag, ceramic or plastic with plastic being the most commonly used. With respect to the flow, a packed bed reactor can be operated in either down or up flow mode. The

input to the reactor can be continuous or intermittent. The packing material can also be continuous or arranged in multiple stages with flow from one stage to another.

**e) Fluidized Bed Reactors**

The fluidized reactor is similar to the packed bed reactor in many aspects, but the packing material is expanded by the upward movement of fluid through the bed.

### **3.6.1 Biological Treatment**

Biological methods of wastewater treatment are based upon induced contact with microorganisms, which feed on the organic materials in the wastewater, thereby reducing the Biological Oxygen Demand (BOD) content of wastewater. BOD in wastewater is used as an indicator of pollutant level, where the greater the BOD, the greater the degree of pollution (Green-Ideas, 2009).

The basic principle behind biological treatment lies in the microorganisms consuming the suspended organic material present in the wastewater as their food source. The organic material is transformed into cellular mass by the metabolic process which is no longer suspended and hard to separate from the water, but can be precipitated by gravity at the bottom of a settling tank. Thus, the water exiting the biological system (biological treatment unit and clarifier) is much clearer than the entering water. Biological treatment based on the metabolic action of the microorganisms can be classified as anaerobic, anoxic and aerobic treatments.

The anaerobic treatment, carried out in the absence of oxygen, utilizes anaerobic bacteria to decompose suspended organic substances. Wastewater or sludge is introduced into a closed tank which is kept under anaerobic conditions and the retention time in the tank is from several days to several weeks. Anaerobic treatment is generally suitable for the treatment of wastes containing high concentrations of organic substances (often used for sludge treatment).

Anoxia is defined as a condition where water is without, or has very low levels of, dissolved oxygen (U.S-EPA, 2006) . Anoxic treatment refers to the growth of microorganisms in anoxic conditions. In wastewater treatment process, this treatment is carried out in anoxic tanks that ultimately reduce the concentration of

nitrate in wastewater. Although this is not the primary metabolic reaction in aerobic treatment, anoxic treatment exists with aerobic treatment under favorable conditions.

Aerobic treatment is a means of oxidizing and decomposing organic substances in wastewater using aerobic microorganisms. Suspended organic substances are oxidized and decomposed by metabolic reactions of microorganisms, which also produces energy. Microorganisms multiply using a portion of this energy and the organic substances present, any excess of microorganisms grown must be separated and disposed of as excess sludge.

### **3.7 Tertiary Treatment Stage**

The removal of nutrients such as phosphorous and nitrogen from the treated water using chemicals is considered as tertiary treatment of wastewater, although recently this has been performed using biological mass. Therefore, the removal of nutrients is performed close to the biological unit. The plant configuration can be such that the nutrient removal is either before or after the biological unit. Recent practice has used only the disinfection of treated water in the tertiary treatment stage (Norweco, 2006). The purpose of disinfection in the treatment of wastewater is to substantially reduce the number of microorganisms in the water that will be discharged back into the environment. Common means of disinfection include ozone, chlorine, or ultraviolet light.

### **3.8 PMB Techniques**

Process Model-Based (PMB) methods/techniques for Fault Detection and Diagnosis (FDD) are discussed in this section. PMB methods are generally classified as quantitative or qualitative, an introduction and brief discussion is presented below. A detailed, but concise, review of these different methodologies is presented below.

#### **3.8.1 Qualitative Model-Based Techniques**

For qualitative model-based techniques, the relationships developed are based upon a fundamental understanding of the physical phenomena controlling the process that are expressed in terms of qualitative functions. The qualitative models can be



developed either as qualitative causal models or abstraction hierarchies. Digraphs, fault trees and qualitative physics are the most popular techniques that belong to the class of casual models. The abstraction hierarchy used for the FDD in a process can be further classified as structural or functional hierarchy. The following is a brief explanation of some of the most commonly used techniques used in qualitative model-based FDD (Venkatasubramanian et al., 2003b).

**a) Digraphs based causal models**

Cause and effect relations, or models, can be represented in the form of signed digraphs (SDG). Digraph is a graph with directed arcs between the nodes, and SDG is a graph in which the directed arcs have a positive or negative sign attached to them. The directed arcs lead from the ‘cause’ nodes to the ‘effect’ nodes. Each node in the SDG corresponds to the deviation from the steady-state value of a variable. SDGs have been the most widely used form of causal knowledge for process fault diagnosis and safety.

**b) Fault Trees**

Fault tree is a logic tree that propagates primary events or faults to the top level event or a hazard. The tree usually has layers of nodes. At each node different logic operations such as AND and OR are performed for propagation. Fault trees have been used in a wide range of risk assessment and reliability analysis studies. Before the construction of the fault tree, the analyst should possess a complete understanding of the system. The fault tree is constructed by asking questions such as: “What could cause a top level event?” In answering this question, one generates other events connected by logic nodes. Fault trees provide a computational means for combining logic in order to analyze system faults. The attraction of using a fault tree stems from the fact that different logic nodes can be used (OR, AND, XOR) instead of the predominantly OR node used in the digraphs. This helps in eliminating spurious solutions and representing the system in a concise manner. The biggest problem with fault trees is that the development is prone to mistakes at different stages. It is of primary importance that the underlying logic of the fault tree construction is correct, otherwise the entire model is faulty from the outset.

### c) **Qualitative Physics**

Qualitative physics or “common sense” reasoning about physical systems has been an area of major interest in the artificial intelligence community. An important approach in qualitative physics is the derivation of qualitative behavior from the ordinary differential equations (ODEs). These qualitative behaviors for different failures can be used as a knowledge source. The goals of these methodologies are to reason from qualitative physical and equation-based descriptions. The advantage of these methods is their ability to yield partial conclusions from incomplete and often uncertain knowledge of the process.

### d) **Abstraction Hierarchy**

Another form of model knowledge is through the development of abstraction hierarchies based on decomposition. There are two-dimensions along which abstraction at different levels is possible, i.e. structural and functional. The structural hierarchy represents the connectivity information of the system and its subsystems. The functional abstraction hierarchy represents the means-end relationships between a system and its subsystems. The majority of the work on fault diagnosis in chemical engineering depends on the development of functional decomposition, and the reason for its popularity is due to the complex functionalities of various units that cannot be expressed in terms of structure.

## **3.8.2 Quantitative Model-Based Techniques**

Relying on an explicit model of the monitored plant, all model-based FDD methods require two steps. The first step generates inconsistencies between the actual and expected behavior known as “residuals”. The second step chooses a decision rule for diagnosis. Some form of redundancy is always required in order to generate residuals to evaluate the inconsistency. There are two types of redundancies, hardware redundancy and analytical redundancy. Hardware redundancy has been utilized in the control of safety-critical systems such as aircraft, space vehicles and nuclear power plants, and requires redundant sensors. Its applicability in the chemical and process industry sector has been limited due to the additional costs and space required for the extra sensors. However, analytical redundancy is achieved from the functional dependence among the process variables and is usually provided by a set of algebraic or differential relationships among the states, and the inputs and the outputs of the

system. The main concept used in the quantitative model-based FDD techniques is analytical redundancy based upon checking the actual system behavior against the system model for consistency. Any inconsistency, expressed as residuals, can be used for detection and isolation purposes. The residuals should be close to zero when no fault occurs, but show 'significant' values when the underlying system changes.

The generation of the residuals requires an explicit mathematical model of the system, either a model derived analytically using first principles or a black-box model obtained empirically. The first principles models are obtained based on a physical understanding of the process. In a chemical engineering process, mass, energy and momentum balances are used in the development of model equations. Historical models developed from first principles were seldom used in process control and fault diagnosis mainly because of their complexity. In addition, chemical engineering processes are often nonlinear which makes the design of fault diagnosis procedures more difficult. However, this is changing due to increased computational power and speed and better understanding of nonlinear controller design and synthesis(Venkatasubramanian et al., 2003a,b,c).

### **3.9 Mathematical Modelling**

A mathematical model describes the fundamental physical phenomena controlling the process expressed in terms of mathematical functional relationships between the inputs and outputs of the system. Most of the work on quantitative model-based approaches has been based on general input-output and state-space models. However, there are a wide variety of quantitative model types that have been considered in fault diagnosis such as first-principles models, frequency-response models, etc. The first-principles models have not been very popular in fault diagnosis studies because of the computational complexity in utilizing these models in real time fault diagnostic systems, and the difficulty in developing accurate models. The most important class of models that have been heavily investigated in fault diagnosis studies are the input-output or state-space models. In this work, the equations describing first principals are used rather than the inputs-output or state-space model of the secondary stage of wastewater treatment. This dynamic model will provide transient values of the process measurements which can be used for the generation of

residuals using the actual values from the process operation for the detection of a fault in the process.

### **3.10 Wastewater Treatment Mathematical Model**

In order to promote development, and facilitate the application of practical models for design and operation of biological wastewater treatment systems, the International Association on Water Quality (IAWQ) formed a task group in 1983. The first goal was to review existing models and the second goal was to reach a consensus concerning the simplest mathematical model having the capability of realistically predicting the performance of single-sludge systems. The final result was presented in 1987 as the IAWQ Activated Sludge Model No. 1 (ASM1). Although the model has been extended since then, for example to incorporate more fractions of COD (i.e. chemical oxygen demand), to describe growth and population dynamics of floc forming and filamentous bacteria and to include new processes for describing enhanced biological phosphorus removal, the original model is probably still the most widely used for describing WWT (wastewater treatment) processes all over the world (Jeppsson, 2003). Since then ASM1 has been the core of numerous models with a number of supplementary details added in almost every case. The model has grown more complex over the years, from ASM1 to ASM2, including biological phosphorus removal processes and to ASM2d including denitrifying PAOs. In 1998 the task group decided to develop a new modelling platform, the ASM3, in order to create a tool for use in the next generation of activated sludge models. The ASM3 is based on recent developments in the understanding of the activated sludge processes, among which are the possibilities of following internal storage compounds, which have an important role in the metabolism of the organisms (Henez et al., 2000).

The use of ASM1 as the core of recent models was the source of inspiration for adopting ASM1 as the basis for the mathematical modelling of wastewater in this research along with the fact that the data available for the validation (dated back to 1991) of the proposed model only covers the basic processes in the water treatment. The concepts of Monod's kinetics, population balance and Activated Sludge Model 1 (ASM1) are used to derive a mathematical model for the activated sludge treatment

of wastewater which can then be used for fault detection. The major biological and chemical processes occurring in the activated sludge system for the treatment of wastewater are:

- 1) Production and decay of microorganisms (under different conditions)
- 2) Utilization of suspended organic material (substrate, i.e. food for microorganisms)
- 3) Oxygen consumption
- 4) Production of volatile suspended solids

### 1) Production and decay of microorganisms

The change in microorganism population due to production is given (Morley, 1979) as:

$$\left(\frac{dX}{dt}\right)_p = \mu X \quad (3.1)$$

where

$X$  = concentration of microorganisms. This is the  $\text{g/m}^3$  of microorganisms present in the activated sludge system for the conversion of wastewater into treated water.

$\mu$  = specific growth rate

The specific growth rate in equation 3.1 can be modelled using Monod's kinetics (Shuler and Kargi, 2002) as given by:

$$\mu = \mu_{\max} \left( \frac{S}{K_S + S} \right) \quad (3.2)$$

where

$\mu_{\max}$  = maximal specific growth rate

$K_S$  = half saturation coefficient

$S$  = substrate concentration (i.e. the food for micro-organisms). Substrate is the suspended solids present in the wastewater. As it is a source of food for the

microorganisms to live on, the suspended solids are called “substrate” in this modelling exercise.

On substitution of equation 3.2 in 3.1, the change in microorganism population due to production is:

$$\left(\frac{dX}{dt}\right)_p = \mu_{\max} \left(\frac{S}{K_s + S}\right) X \quad (3.3)$$

Now, the decay rate of microorganisms due to endogenous metabolism (Morley, 1979):

$$\left(\frac{dX}{dt}\right)_d = -k_d X \quad (3.4)$$

Hence, the net change in microorganism population can be modelled by:

$$\left(\frac{dX}{dt}\right) = \left[ \mu_{\max} \left(\frac{S}{K_s + S}\right) - k_d \right] X \quad (3.5)$$

## 2) Utilization of suspended organic material

This section describes the dynamics of suspended solids (SS) in the wastewater treatment. The rate of substrate utilization (consumption) due to microorganisms (Morley, 1979) is given by:

(3.6)

$$\frac{dS}{dt} = -\frac{\mu X}{Y_{X/S}}$$

where

$S$  = concentration of substrate (suspended organic material in wastewater)

$Y_{X/S}$  = yield factor (indicates how many units of microorganisms are produced per unit of substrate. Similar to  $Y_{X/O_2}$  that is the unit of  $O_2$  consumed by a unit of microorganisms)

On substitution of equation 3.2 in 3.6, the equation will become:

$$\frac{dS}{dt} = -\mu_{\max} \left(\frac{S}{K_s + S}\right) \frac{X}{Y_{X/S}} \quad (3.7)$$

### 3) Oxygen Concentration

The oxygen concentration in the wastewater system is reduced by aerobic growth of microorganisms. Similar to the consumption of suspended organic material (equation 3.7), the rate of oxygen depletion from the process can be derived as:

$$\frac{dS_o}{dt} = - \left[ \mu_{\max} \left( \frac{S}{K_S + S} \right) \left( \frac{S_o}{K_o + S_o} \right) \frac{X}{Y_{X/O_2}} \right] \quad (3.8)$$

Another important consideration is that the model equations are derived assuming the excess oxygen demand in the process. If the oxygen supply is limited or accounted for, then the factor  $(S_o/K_o + S_o)$  is to be incorporated in equations 3.5, and 3.7. The resulting models equations are given below by equations 3.9 and 3.10.

Net microorganism growth (equation 3.5):

$$\left( \frac{dX}{dt} \right) = \left[ \mu_{\max} \left( \frac{S}{K_S + S} \right) \left( \frac{S_o}{K_o + S_o} \right) - k_d \right] X \quad (3.9)$$

and the rate of substrate utilization (equation 3.7):

$$\frac{dS}{dt} = - \mu_{\max} \left( \frac{S}{K_s + S} \right) \left( \frac{S_o}{K_o + S_o} \right) \frac{X}{Y_{X/S}} \quad (3.10)$$

### 4) Production of Volatile Suspended Solids

The decay of microorganisms contributes towards the volatile suspended solids. A portion of the dead microorganisms is recycled as a source of food for other microorganisms, while the particulate part of the dead microorganisms contributes to the concentration of volatile solids. The decay rate can be modelled using death-regeneration hypothesis and the model equation (Jeppsson, 2003) is given as:

$$\frac{dS_{VS}}{dt} = f_p (k_d X) \quad (3.11)$$

$k_d$  = decay rate of microorganisms

$f_p$  = fraction of biomass yielding particulate products

Equation 3.11 above describes the dynamics of Volatile Suspended Solids (VSS).

The derivation of an equation for the production of “Substrate” from equation 3.11 is based upon a fraction of biomass yielding particulate products, and the remainder is utilized as a food source. The equation then derived for the production of substrate is:

$$\frac{dS}{dt} = (1-f_p)(k_d X) \quad (3.12)$$

This equation affects equation 3.7 such that the net rate of substrate utilization is:

$$\frac{dS}{dt} = -\mu_{\max} \left( \frac{S}{K_s + S} \right) \frac{X}{Y_{X/S}} + (1-f_p)(k_d X) \quad (3.13)$$

### **Mass Balance around Activate Sludge System**

If F is the influent flow rate in the activated sludge process, a mass balance will yield:

$$\text{In-Out+Generation-Consumption=Accumulation} \quad (3.14)$$

Substrate (suspended organic solids) mass balance:

$$FS_o - FS + 0 - (\mu X / Y_{X/S})V = V \frac{dS}{dt} \quad (3.15)$$

Simplifying the above equation yields:

$$\frac{F}{V} S_o - \frac{F}{V} S - (\mu X / Y_{X/S}) = \frac{dS}{dt} \quad (3.16)$$

where

V = volume of reactor

S = concentration of substrate in reactor

S<sub>o</sub> = initial concentration of substrate in influent. This is the concentration of the suspended solids in influent wastewater whereas S is the concentration of suspended solids in effluent (treated wastewater).



$F/V = D$  (dilution rate, i.e. inverse of residence time) (Shuler and Kargi, 2002)

Equation 3.16, using  $D$  will become:

$$DS_o - DS - (\mu X / Y_{X/S}) = \frac{dS}{dt} \quad (3.17)$$

Equation 3.17 predicts the dynamic behavior of the concentration of suspended solids in the activated sludge system.

### Cell mass balance

Similarly to the substrate analysis, the microorganisms (cell) balance on the activated sludge process is modelled as:

$$FX_o - FX + \mu XV - k_d XV = V \frac{dX}{dt} \quad (3.18)$$

Simplifying equation 3.18 by using  $D$  for  $F/V$ :

$$DX_o - DX + \mu X - k_d X = \frac{dX}{dt} \quad (3.19)$$

Equation 3.19 is similar to equation 3.17 and predicts the dynamics of cell mass in the activated sludge process.

### Volatile Suspended Solids Balance

Using equation 3.14, a mass balance for the volatile solids is:

$$FS_{VS} - FS_{VS} + f_p (k_d X) V = V \frac{dS_{VS}}{dt} \quad (3.20)$$

Simplifying the above equation and using  $D$  for  $F/V$ :

$$DS_{VS} - DS_{VS} + f_p (k_H X_H + k_A X_A) = V \frac{dS_{VS}}{dt} \quad (3.21)$$

## 3.11 Results from Model Simulations

The following simulation results were obtained by using the model equations for the activated sludge system. The biological treatment of wastewater can be carried out using a batch or continuous process, the later is more widely used.

### 3.11.1 Batch Process

The simulation study on a batch reactor is further divided into three steps in order to obtain a better understanding of the process and the effects of different parameters.

#### 1. Microorganisms and organic material

The first simulation is for aerobic growth of the heterotrophic microorganism, and their effect on the concentration of the organic material in the wastewater. It is assumed that an excess of dissolved oxygen (DO) exists in the process (i.e. no effect of oxygen concentration on microorganisms). The equations used are given below.

##### *Microorganism concentration*

The net change in the microorganism's concentration can be predicted from equation 3.5:

$$\frac{dX}{dt} = \left[ \mu_{\max} \left( \frac{S}{K_S + S} \right) - k_d \right] X$$

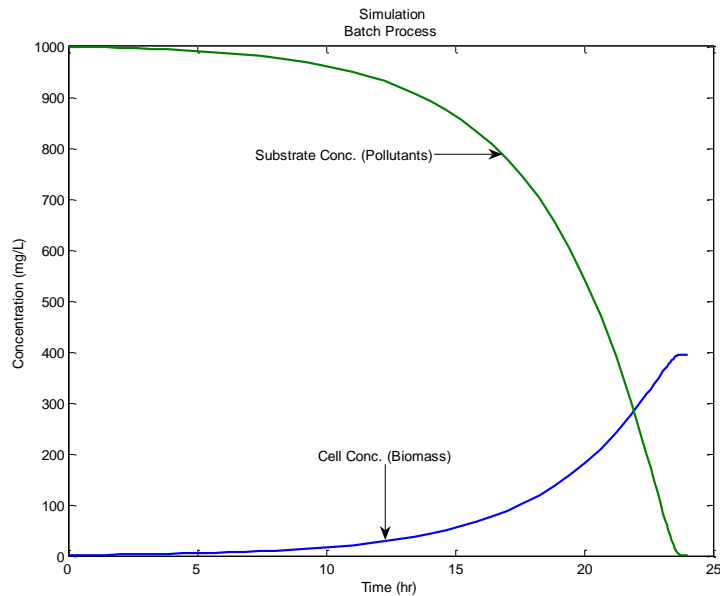
##### *Organic material concentration*

Using Equation 3.13 for the net concentration of the organic material (i.e. substrate) in the wastewater:

$$\frac{dS}{dt} = -\mu_{\max} \left( \frac{S}{K_S + S} \right) \frac{X}{Y_{X/S}} + (1-f_p)k_d X$$

Simulation results from the above equations show how the concentration of microorganisms and organic material will change with time in a batch process.

## Simulation Results and Discussion



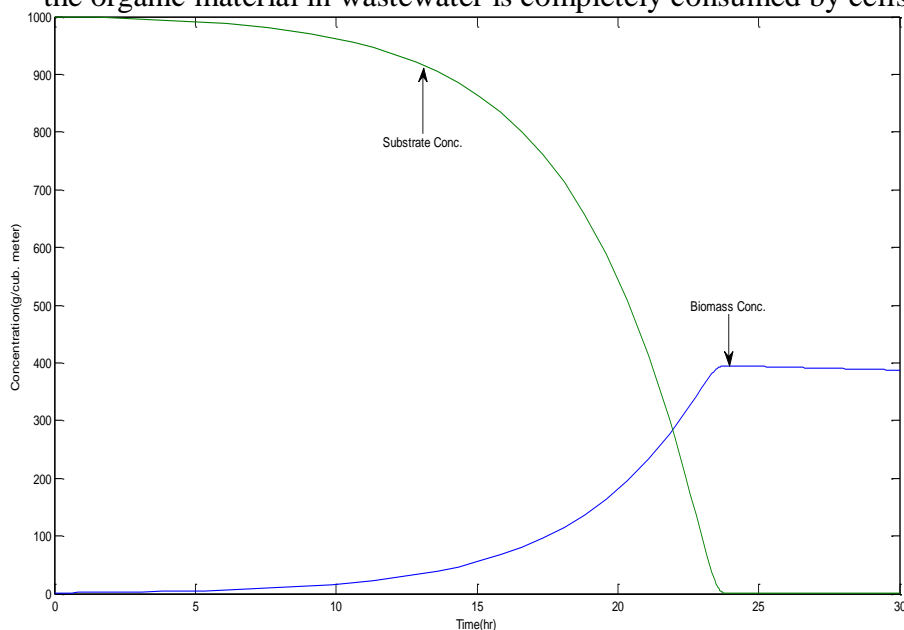
**Table 3.1** Parameters used for the simulation in Figure 3.1

Parameter	Value	Units
Substrate initial concentration	1000	$\text{g}/\text{m}^3$
Cell initial concentration	1.5	$\text{g}/\text{m}^3$
Residence time	24	hr

**Figure 3.2** Simulation result from batch process assuming excess DO

Figure 3.2 illustrates the dynamics of the wastewater reclamation process using activated sludge in batch conditions. The concentration of organic materials (i.e. pollutant or substrate) reduces to near zero at the end of batch operation.

Figure 3.3 is an extension of Figure 3.2 using the parameters in Table 3.1, it shows how the activated sludge (microorganisms or cell) concentration starts decreasing as the organic material in wastewater is completely consumed by cells.



**Figure 3.3** Simulation result from batch process assuming excess DO and a residence time of 30hr

## 2. Microorganisms, organic material and dissolved oxygen

For this simulation, the earlier assumption of excess DO is rejected. Now the microorganism growth (concentration) depends on the oxygen concentration at any time in the process. If enough DO is provided, the concentration of microorganism will increase resulting in the decrease of organic material concentration. The rate of growth will be reduced if DO is less than the oxygen demand of the microorganisms. The equations used are given below.

### *Microorganism concentration*

Using equation 3.9 for the net growth of microorganism including oxygen concentration is:

$$\frac{dX}{dt} = \left[ \mu_{\max} \left( \frac{S}{K_S + S} \right) \left( \frac{S_o}{K_{OH} + S_o} \right) - k_d \right] X$$

### *Organic material concentration*

The model equation for substrate utilization is then derived from equation 3.13 to include the effect of oxygen as:

$$\frac{dS}{dt} = - \left[ \mu_{\max} \left( \frac{S}{K_S + S} \right) \left( \frac{S_o}{K_{OH} + S_o} \right) \frac{X}{Y_{X/S}} \right] + (1 - f_p) k_d X \quad (3.22)$$

Although oxygen is consumed in the batch process, it is provided continuously using turbines. Oxygen supply will fulfill the oxygen demand of microorganisms and also produce turbulence in the reactor that will help to keep the solution of organic material and microorganism suspended in the reactor, thus improving the contact between both phases and resulting in efficient removal of pollutants from wastewater. The amount of DO depends upon the Biological Oxygen Demand (BOD) of the wastewater and, in general, a minimum residual of 1mg of DO per liter of wastewater must be maintained (Buchanan and Seabloom, 2004).

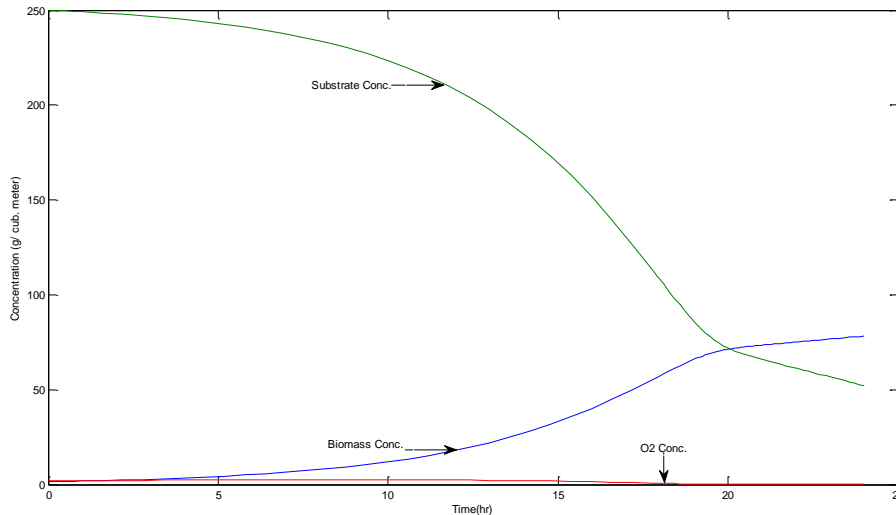
The equation for oxygen consumption derived from equation 3.8 is:

$$\frac{dS_o}{dt} = DS_{o_{in}} - \left[ \mu_{max} \left( \frac{S}{K_S + S} \right) \left( \frac{S_o}{K_{OH} + S_o} \right) \frac{X}{Y_{X/O_2}} \right] \quad (3.23)$$

Simulation from the model equations above will predict the microorganisms and organic material concentration under the influence of the oxygen supplied.

### *Simulation Results and Discussion*

Oxygen is always used in excess in the activated sludge treatment of wastewater treatment. For illustration purposes, Figure 3.4 shows the dynamics of activated sludge system using 2g/m<sup>3</sup> of air.



**Figure 3.4** Dynamics of activated sludge system using oxygen concentration as 2g/m<sup>3</sup>

### **3.11.2 Completely Mixed Reactor**

Assume that complete mixing occurs instantaneously and uniformly throughout the reactor as the fluid-particles enter.

#### 1) Microorganisms and organic material

For the activated sludge process, the model equations assuming an excess of dissolved oxygen (DO) are:

Microorganism concentration predicted using equations 3.5 and 3.14:

$$\frac{dX}{dt} = DX_{in} - DX_{out} + \mu_{max} \left( \frac{S}{K_S + S} \right) X - k_d X \quad (3.24)$$

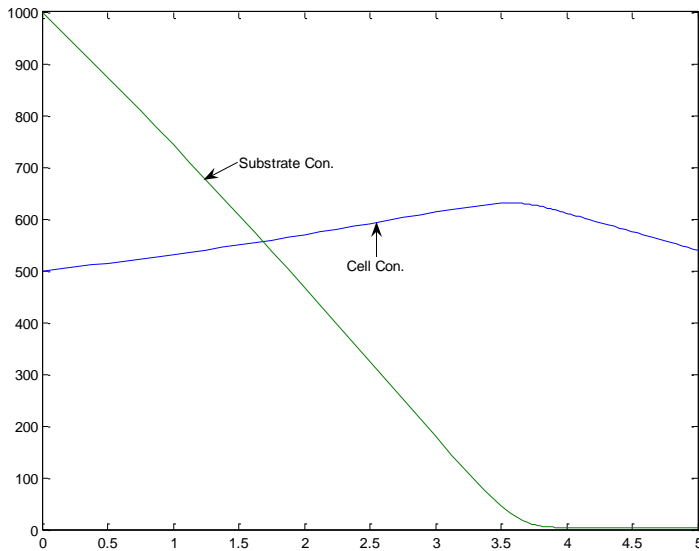
Similarly, the organic material concentration from equations 3.13 and 3.15:

$$\frac{dS}{dt} = DS_{in} - DS_{out} - \mu_{max} \left( \frac{S}{K_S + S} \right) \frac{X}{Y_{X/S}} + (1 - f_p) k_d X \quad (3.25)$$

Simulation results from the above equations will predict how the concentration of microorganisms and organic material will change with time in the activated sludge process.

### Simulation Results and Discussion

It is desirable to study the effect of continuous operation on the wastewater process. This dynamic behavior is predicted by simulations using the parameters in Table 3.2, as shown in Figure 3.5.



**Table 3.2** Simulation parameters for continuous process

Parameter	Value	Units
Substrate initial concn.	1000	g/m <sup>3</sup>
Cell initial concn.	500	g/m <sup>3</sup>
Residence time	4	hr
Substrate influent	250	g/m <sup>3</sup>
Substrate effluent	10	g/m <sup>3</sup>
Cell influent	0	g/m <sup>3</sup>
Cell effluent	400	g/m <sup>3</sup>

**Figure 3.5** Simulation dynamics for a continuous process using excess DO

## 2) Microorganisms, organic material and dissolved oxygen

The model equations for microorganisms and organic material considering the DO concentration are:

Microorganism concentration using equations 3.9 and 3.14:

$$\frac{dX}{dt} = DX_{in} - DX_{out} + \mu_{max} \left( \frac{S}{K_S + S} \right) \left( \frac{S_o}{K_{OH} + S_o} \right) X - k_d X \quad (3.26)$$

Similarly for the organic material:

$$\frac{dS}{dt} = DS_{in} - DS_{out} - \mu_{max} \left( \frac{S}{K_S + S} \right) \left( \frac{S_o}{K_{OH} + S_o} \right) \frac{X}{Y_{X/S}} + (1-f_p)k_d X \quad (3.27)$$

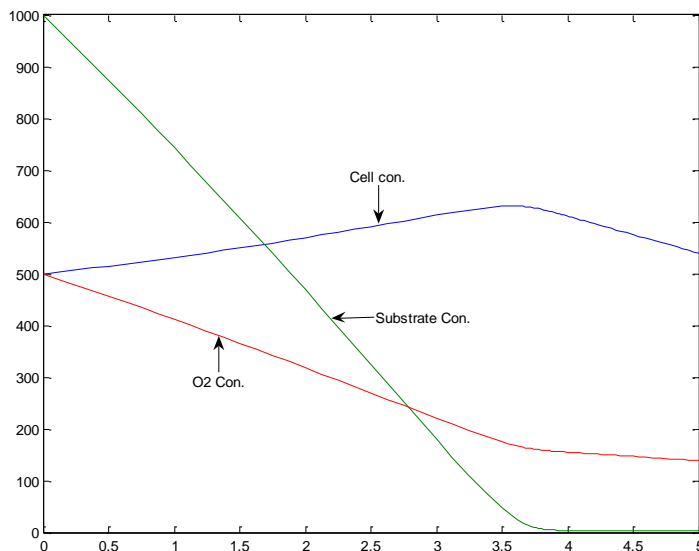
Oxygen concentration is predicted using equation 3.23:

$$\frac{dS_o}{dt} = DS_o - \left[ \mu_{max} \left( \frac{S}{K_S + S} \right) \left( \frac{S_o}{K_{OH} + S_o} \right) \frac{X}{Y_{X/O_2}} \right]$$

Simulations obtained using these model equations predict the microorganisms and organic material concentrations under the influence of the oxygen supplied.

### *Simulation, results and discussion*

Simulation results used to study the dynamic behavior of the processes discussed above are shown in Figure 3.6, and using the simulation parameters given in Table 3.3.



**Table 3.3** Parameters used to study the effect of O<sub>2</sub> consumption

Parameter	Value	Units
Substrate initial concn.	1000	g/ m <sup>3</sup>
Cell initial concn.	500	g/ m <sup>3</sup>
Residence time	4	hr
Substrate influent	250	g/ m <sup>3</sup>
Substrate effluent	10	g/ m <sup>3</sup>
Cell influent	0	g/ m <sup>3</sup>
Cell effluent	400	g/ m <sup>3</sup>
DO	500	g/ m <sup>3</sup>

**Figure 3.6** Wastewater treatment dynamics for controlled oxygen supply

Table 3.4 shows all parameters that are used in the simulations discussed above. The parameters are adopted from the ASM1 model described by Jeppsson (2003).

**Table 3.4** Parameters used in modelling and simulation with literature values

Notation	Explanation	Value
$\mu_{max}$	Maximum specific growth rate	0.25 hr <sup>-1</sup>
$Y_{X/S}$	Yield factor (unit of microorganisms produced per unit of substrate)	0.4
$Y_{X/O_2}$	Unit of microorganisms per unit of O <sub>2</sub>	0.9 - 1.4
$Y_{X/NO_2}$	Unit of microorganisms per unit of NO	0.24
$K_S$	Half saturation coefficient	23 m <sup>-3</sup>
$K_{OH}$	Oxygen half saturation constant	0.2 m <sup>-3</sup>
$K_{NO}$	Nitrate half saturation constant	0.5 m <sup>-3</sup>
$\eta_g$	Correction factor for anoxic growth	0.8
$f_p$	Fraction of biomass yielding VSS	0.15
$k_d$	Decay rate of microorganisms	0.005 hr <sup>-1</sup>
$F$	Influent flow rate	936 hr <sup>-1</sup> m <sup>3</sup>
$V$	Volume of reactor	4021 m <sup>3</sup>
$D$	F/V	0.23 hr <sup>-1</sup>

### 3.12 Model Validation

The model developed is then validated before it can be used for the diagnostic purposes. It is essential to establish that the model does actually predict the behaviour of activated sludge process for the treatment of wastewater. Though it is evident from the simulation results (Figure 3.2 to 3.6) that the model does follow the expected behaviour of an activated sludge system, it is still to be established that how accurate the proposed model is. The wastewater treatment plant (WWTP) data (discussed in detail in section 4.5) is used for the validation of the model.

The value of an input parameter from the WWTP data (Appendix A) is used as the initial value for the simulations study of the model to obtain the output value of the parameter. This output value is then compared to the actual output given in the



data set to validate the proposed model. Table 3.5 summarizes the results obtained during the validation of proposed model.

**Table 3.5** Results for model validation

Date	Input Parameter	Input Value (WWTP data set)	Output Parameter	Output Value (WWTP data set)	Output value (simulations)	Difference (%)
11/1/90	SS-E	192	SS-D	100	107	7
1/3/90	SS-E	166	SS-D	94	98	4
3/5/90	SS-D	88	SS-S	49	53	8
29/7/90	SS-D	90	SS-S	34	37	9
28/09/90	SSV-E	57	SSV-D	77	81	5
30/11/90	SSV-E	75	SSV-D	83	89	7
15/1/90	SSV-D	71	SSV-S	76	81	7
8/3/91	SSV-D	64	SSV-S	85	88	4
21/5/91	DBO-E	238	DBO-D	101	105	4
31/7/91	DBO-E	170	DBO-D	101	100	1
6/8/91	DBO-D	90	DBO-S	16	18	12
16/10/91	DBO-D	121	DBO-S	33	36	9

Table 3.5 summarizes the results obtained from the validation of the activated sludge model proposed. The discrepancy in the model and actual values ranges from 1 – 12% with an average of 6.5%.

### 3.13 Summary Comments

A brief review of the processes involved in the wastewater treatment facility is presented, and a mathematical model derived from the ASM1 model. This model is used in this research work for the detection step of the fault detection and diagnosis system. The ideology used for fault detection is the analytical redundancy where the value of an observed parameter in a process is compared to the value obtained from mathematical model simulations, in order to decide whether the parameter under observation is out of the normal operating limit.

# Chapter 4

## Process History Based Method

### 4.1 Introduction

Process History Based (PHB) methods for fault detection and diagnosis are discussed in this chapter. This includes an introduction and explanation of the different techniques available including a detailed review of data-mining, a technique used in this research work. Data mining is then applied to the dataset of a wastewater treatment plant, and the results are presented, analyzed and discussed. Some modifications are suggested and implemented in the selected technique of data mining that further improves its efficiency and effectiveness in order to interpret data for fault detection and diagnosis. An introduction and explanation of the Wastewater Treatment Plant (WWTP) dataset used in this work is included.

### 4.2 PHB Techniques

In contrast to the depth of knowledge required for fault detection and diagnosis when using Process Model Based methods, PHB methods only require access to the historical and/or operational data. This data is then used to extract knowledge for input to a diagnostic system. This process of knowledge extraction is known as feature extraction. The PHB methods are classified as either qualitative or quantitative on the basis of the type of knowledge extracted from the database. Expert System (ES) and Qualitative Trend Analysis (QTA) are the most important and widely applied techniques from the class of qualitative PHB approaches. Quantitative techniques are further divided into statistical and non-statistical approaches. Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Partial Least Squares (PLS) form the majority of the statistical quantitative PHB approaches, while the best known techniques in non-statistical quantitative PHB approach is Artificial Neural Networks (ANN). Each of these techniques has its own strengths and weaknesses. During the diagnosis stage of FDD,

the onsite operator or engineer has to identify the symptoms, analyze the symptomatic information, interpret the various error messages and indications, and decide upon the correct diagnosis for the situation. Due to the inherent complexity of chemical processes, the diagnosis requires extensive technical skills and process experience, in addition to a complete understating of the process and some general concepts of diagnosis, in order to carry out the diagnostic operation when a fault is identified in the process. This requires a very experienced engineer with the deep domain-specific knowledge and the knowledge of the “ins-and-outs” of the system (Sun et al., 2007). Many diagnosis methods have been proposed to help operators and engineers perform diagnostic fault analysis. For example, expert systems, neural networks, and genetic algorithms are the most popular approaches among the PHB techniques. The application of an expert system is limited due to the knowledge acquisition because the knowledge-based systems developed from expert rules are very system specific, their representational ability is quite limited, and they are difficult to update. The complexity of a process makes the diagnosis methods based on ANN and Genetic Algorithm (GA) difficult, in addition the inherent nature of ANN approach means they lack the explanation and adaptability properties of a diagnostic system (Chen and Mo, 2004, Venkatasubramanian et al., 2003c, Sun et al., 2007, Yang et al., 2005).

The most helpful presentation of a domain-specific knowledge for a non-expert is the cause-symptom relationship that enables the quick comprehension of the situation. Production rules are the knowledge formalized into “rules” containing an If part and a Then part that explains the cause-symptom relationship in a process. Production rules are one of the most popular and widely used knowledge representation languages.

### **4.3 Data Mining**

In modern processes, the computer control and data logging systems are able to easily collect large amounts of data. This data can be used for process monitoring and fault diagnosis, as well as in other decision making activities - if properly analyzed. Data mining is a powerful new technique with the potential to help engineers explore and focus on the most important information available from the

analysis of the historical and/or operational data available from a process. Data mining is defined as "The nontrivial process of extracting implicit, previously unknown, and potentially useful, information from data" (Fayyed et al., 1996). It uses machine learning, statistical and visualization techniques to discover and present discovered knowledge in a form which can be easily understood. It allows users to analyze large databases to solve decision problems encountered in an industry or business sector. The primary goal of data mining is the extraction of knowledge from the available data and is often known as Knowledge Discovery and Data Mining (KDDM).

### **4.3.1 Data Mining Process Description**

Data Mining is a complex process which typically involves the following procedures (Fayyed et al., 1996).

***Understanding:*** Developing an understanding of the application domain, the relevant prior knowledge, and the goals of mining.

***Creating a target data set:*** Selecting a data set, or focus on a subset of variables or data samples, on which discovery is to be performed.

***Data pre-processing and cleaning:*** This is frequently time consuming; data pre-processing is needed because most large databases were created for a different purpose from their current applications. Therefore, the data within these databases are not immediately ready to use in knowledge discovery algorithms or other information processing techniques. For example, the data may contain information that is not uniform, the data may be blank or inconsistent, certain data may be continuous while others are categorical, some data may contain sensitive information and require encryption, and finally, some data may contain uncertainties. Hence data pre-processing and cleaning involves basic operations such as the removal of noise or outliers if appropriate, collecting the necessary information to model, deciding on strategies for handling missing data fields, and accounting for noise, time sequence information and known changes.

***Data reduction and projection:*** Finding useful features to represent the data depending on the aim of the task, using dimensionality reduction or transformation, or finding invariant representation from data.

***Choosing the data mining task:*** Depends mainly on the application domain and on the interest of the miner. Decide whether the goal of the KDDM process is summarization, clustering, classification and regression, etc., and identification of several types of data mining tasks for which data mining offers possible answers.

***Choosing the data analysis algorithm(s):*** Selection of the methods to be used to search for patterns in the data. This includes deciding which models and parameters may be appropriate (e.g. models of categorical data are different from models on vectors over the real data) and matching a particular data mining method with the overall criteria of KDDM process.

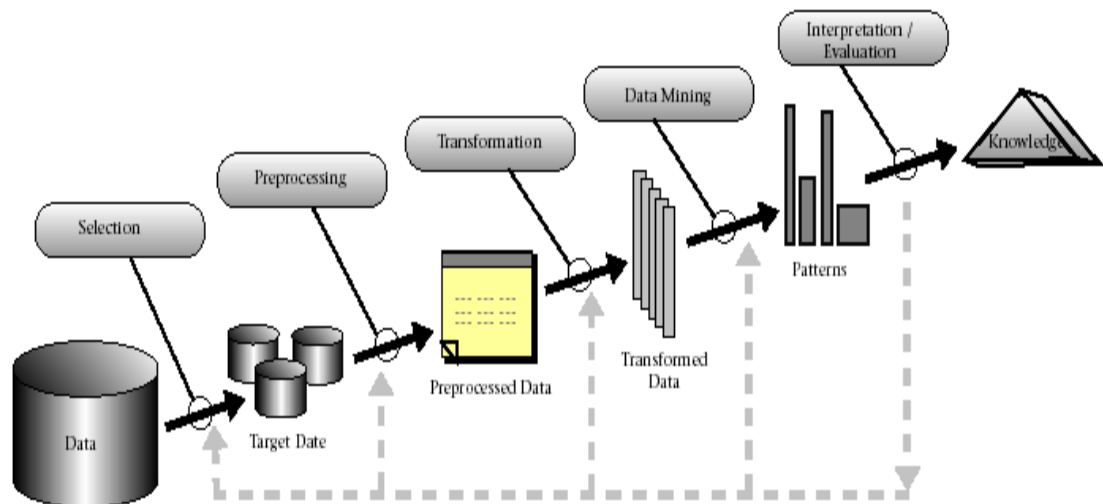
***Data mining:*** searching for patterns of interest in a particular representational form, or a set of such representations, including clustering, dependency modelling, analysis, visualization, etc. The following steps can significantly aid the data mining process.

a) *Interpretation* - interpreting mined patterns, and possible return to any of the previous steps. This step can also involve visualization of the data given the extracted models.

b) *Using discovered knowledge* - this step involves acting directly on discovered knowledge, incorporating the knowledge into another system for further action, or documenting and reporting the knowledge. It also includes checking and resolving potential conflicts with previously believed or extracted knowledge.

c) *Evaluation of KDDM purpose* - newly discovered knowledge is often used to formulate new hypotheses; also new questions may be posed using the enlarged knowledge base. In this step, the KDDM process is evaluated for possible further use in both refinement and expansion.

The overall process, representing the move from data to information, and ultimately knowledge, is shown in Figure 4.1. Sometimes the analysis step itself is referred to as data mining, although this term is better applied to the whole process.



**Figure 4.1** An overview of the knowledge discovery process, showing the move from data to knowledge or information, through various steps. (Buontempo, 2005)

### 4.3.2 Applications for Data-Mining

The technique of data mining has attracted much interest, not only from information technology companies but also from the industrial and business sectors. The following is the list of domains where this technique is being, or potentially can be, applied (Wang, 1999).

***Manufacturing Process Analysis*** - identifying the causes of faults in manufacturing processes.

***Production Design*** - developing a system which will give product designers access to data and information from a range of corporate databases deemed essential to their function, in particular, customer complaints, product material features, R&D testing. Access to this data may point to fundamental design anomalies or inefficiencies which would not have been otherwise apparent.

*Scientific Data Analysis* - cataloguing in surveys, the basic processing needed before high-level scientific analysis can occur, scientific discovery over a large data set, e.g. the SKICAT system from JPL/Caltech was used to automatically identify stars and galaxies in a large-scale sky survey for cataloguing and scientific analysis. In the global climate area, spatio-temporal patterns such as cyclones were predicted from large simulated and observational datasets.

*Experimental Results Analysis* - summarizing the experimental results and the predictive models.

*Marketing and Sales Data Analysis* - identifying potential customers, establishing the effectiveness of a sales campaign.

*Investment Analysis* - predict a portfolio return on investment.

*Intelligent agents and World Wide Web (WWW) navigation* - model user preferences from data, collaborative filtering, advertising, etc.

*Fraud detection* - identify fraudulent transactions.

*Loan approval* - establishing the credit worthiness of a customer requesting a loan.

*Portfolio Trading* - trade a portfolio of financial investments by maximizing returns and minimizing risks.

### **4.3.3 Data Mining Approaches**

Data mining approaches can be broadly categorized as either descriptive or predictive. Descriptive approaches aim to discover patterns that characterize the data, whereas predictive approaches aim to construct models to predict the outcome of a future event by learning from the observed parameters (Charaniya et al., 2008).

#### **4.3.3.1 Descriptive approaches**

The descriptive approaches fall into two categories; (a) identifying interesting patterns in the data; and (b) clustering the data into meaningful groups.

**a) Pattern discovery**

Algorithms for finding patterns in very large datasets are one of the key success stories of data mining research. These methods aim to analyze the parameters of various runs to identify a pattern that is observed in a large number of runs. Patterns discovered from process data can provide insights into the relationship between different parameters, and can also be used to discover association rules. Various algorithms have been developed that can mine process data to discover relationships between the parameters of the different runs that satisfy certain properties. The most efficient approaches for finding these patterns are FPgrowth and LPminer (Han et al., 2004).

**b) Clustering**

Clustering methods can be used to group different process runs into subsets (groups) according to the similarity in the behavior of some parameters. Clustering methods can be differentiated along multiple dimensions, one of them being the top-down (partitional) or bottom-up (agglomerative) nature of the algorithm. Partitional methods commence with all process runs (or object/record) belonging to one cluster and they are divided into designated number of clusters. *K*-means, Partitioning Around Medoids (PAM), Self-Organizing Maps (SOM), and graph-based clustering methods are popular examples of partitional algorithms.

By contrast, agglomerative methods start with each run belonging to a separate cluster and the clusters are merged, based on the similarities of their parameter profiles, until the runs have been grouped into a pre-specified number of clusters. Hierarchical agglomerative clustering is the most commonly used agglomerative method (Jain et al., 1999). Most statistical packages, such as S-Plus and R Project provide a range of clustering methods (R-Foundation, 2008, TIBCO-Software-Inc, 2008)

**4.3.3.2 Predictive approaches**

Predictive approaches can be used to analyze a set of process runs that exhibit different outcomes (e.g. final product concentration) to identify the relationship between process parameters and the outcome. The discovered relationships (called model or classifier) can be used to predict the process outcome and provide key



insights into how the predicted outcome might affect other parameters of the run, thereby allowing for an intelligent outcome-driven refinement of the process parameters. Commonly used predictive methods include regression, Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Decision Trees (DT). These methods have been designed for problems that arise when process runs are divided into classes. Two of the commonly used predictive methods are discussed below.

### **Artificial Neural Networks (ANN)**

ANN models attempt to imitate the signal processing events that occur in the interconnected network of neurons in the brain. An ANN consists of several nodes that are organized into two or more layers. The first layer serves as input for process parameters and the final layer determines the run outcome. Any intermediate layers are referred to as hidden layers. Every node of a hidden layer receives all inputs from the previous layer, performs a weighted average of the inputs, and sends its output to the next layer after a threshold transformation. This process is continued until the final output layer is reached. The weighting factors and threshold parameters are learnt from the training runs in an attempt to minimize the error in classifying the runs (Krogh, 2008).

### **Decision Trees (DT)**

DT-based methods classify runs recursively based on chosen thresholds for one or more parameters. The process parameter that provides most information about the classes is used to split the runs into two or more branches. Splitting thus results in ‘child’ nodes that are most separated from each other in terms of the class. Thus, selecting a parameter and its threshold for the split is a key exercise for DT classifiers. This division is repeated until all the runs at a particular node belong to a single class (terminal node) or one or more stopping rules are satisfied. A top-down interpretation of a decision tree is intuitive and it also allows ranking of process parameters according to their relevance (Quinlan, 1990).

## 4.4 Inductive Data Mining

Inductive data mining refers to the technique used for the generation of the decision tree and production rules from a dataset. It is also an effective approach for automated acquisition of expert knowledge to be built into the knowledge base of an expert system. Classification problem is the current research focus in the area of data mining, and decision tree is one of the most widely used classification methods.

The appeal of decision trees for data analysis and as classifier systems originate primarily from three inherent properties: their ability to model non linear relationships; their ease of interpretability; and their nonmetric nature. Decision trees have been found to be able to handle large-scale problems due to their computational efficiency, to provide interpretable results and, in particular, to identify the most representative attributes for a given task. The traditional approach to inducing decision trees based upon given training data involves recursive partitioning which selects partitioning variables and their value in a greedy (indiscriminate acquisition?) manner to optimize a given measure of purity. A greedy algorithm makes each choice in a locally optimized manner and progresses making one greedy choice after another and reduces the problem to a smaller one this way. This methodology has numerous benefits including classifier interpretability and the capability of modelling non linear relationships.

While capable of modelling nonlinear relationships, decision trees retain a high level of interpretability. The typical structure of a decision tree consists of a root node linked to two or more child nodes which may or may not link to further child nodes. Each nominal node within the tree represents a point of decision or data splitting based upon the data (DeLisle and Dixon, 2004)

Most inductive data mining methods for decision tree generation use supervised learning, i.e. learning from a set of pre-classified cases. Many algorithms have been proposed for decision tree generation, e.g. ID3, C4.5, See5.0, CART, SLIQ, SPRINT and BOAT. The best known algorithms used are CART and See5.0 (with earlier versions as ID3 and C4.5). The decision tree created by CART is a binary tree in which each split generates exactly two branches. In the decision tree created by See5.0, each split can generate more than two branches; also See5.0 can solve the

classification problem with continuous-valued attributes. It was developed by Quinlan (1993, 1986, 1990, 1996), and is similar to most other decision tree algorithms in that See5.0 consists of two phases, a building (growing) phase followed by a pruning phase.

### a) Building Phase

In the building phase, a subset of the training set called the window is chosen at random and a decision tree is formed from it; this tree correctly classifies all objects in the window. All other objects in the training set are then classified using the tree. If the tree gives the correct answer for all these objects, then it is correct for the entire training set and the process terminates. If not, then a selection of incorrectly classified objects is added to the window and the process continues. A decision tree is then constructed in a top-down fashion by iteratively selecting the most informative attribute at the current node in the tree. The most informative attribute for the current node is determined by the splitting criterion, i.e. Gain Ratio (Quinlan, 1993). The Gain Ratio is calculated in the following manner.

**Step 1:** Calculate  $Info(S)$  to identify the class in the training set  $S$ .

$$Info(S) = - \sum_{i=1}^n \left[ \left\{ \frac{freq(C_i, S)}{|S|} \right\} \log_2 \left\{ \frac{freq(C_i, S)}{|S|} \right\} \right] \quad (4.1)$$

where,

$n$  = number of classes

$C_i$  = a class

$|S|$  = total number of cases in the training set  $S$

$freq(C_i, S) = |S_i|$  = number of cases in  $S$  belonging to the class  $C_i$

**Step 2:** Calculate the expected information value,  $Info_x(S)$  for test X to partition S:

$$Info_x(S) = - \sum_{i=1}^m \left[ \left( \frac{|S_i|}{|S|} \right) Info(S_i) \right] \quad (4.2)$$

when,

m = number of outputs from test X

**Step 3:** Calculate the information gain after partition according to test X:

$$Gain(X) = Info(S) - Info_x(S) \quad (4.3)$$

**Step 4:** Calculate the partition information value  $SplitInfo(X)$  acquiring for S partitioned into m subsets:

$$SplitInfo = - \frac{1}{2} \sum_{i=1}^m \left[ \left( \frac{|S_i|}{|S|} \right) \log_2 \left( \frac{|S_i|}{|S|} \right) + \left\{ 1 - \left( \frac{|S_i|}{|S|} \right) \right\} \log_2 \left\{ 1 - \left( \frac{|S_i|}{|S|} \right) \right\} \right] \quad (4.4)$$

**Step 5:** Calculate the Gain Ratio of  $Gain(X)$  over  $SplitInfo(X)$ :

$$GainRatio(X) = \frac{Gain(X)}{SplitInfo(X)} \quad (4.5)$$

The Gain Ratio (X) compensates for the weak point of  $Gain(X)$  which represents the quantity of information provided by X in the training set. Therefore, an attribute with the highest Gain Ratio (X) is taken as the root of the decision tree.

## b) Pruning phase

A large decision tree constructed from a training set usually does not retain its accuracy over the whole sample space for over-training or over-fitting. Therefore, a fully grown decision tree needs to be pruned by removing the less reliable branches to obtain better classification performance over the whole instance space, even though it may have a higher error over the training set. A number of empirical methods have been proposed for pruning a decision tree and they can be divided into two types: construction-time pruning (or pre-pruning) and pruning after building a fully grown tree (or post pruning). Pre-pruning methods (e.g. threshold method and  $\chi^2$  test method) are used to decide when to stop expanding a decision tree. A serious limitation in the pre-pruning method is that the criterion to stop a tree is often based

on local information. In contrast, the post-pruning methods (e.g. cost-complexity, critical value and reduced error) use global information. The See5.0 algorithm applies an error-based post-pruning strategy to deal with the over-training problem, which is a pessimistic error pruning method. In practice for each classification node, See5.0 calculates a predicted error rate based on the total aggregate of misclassifications at that particular node (See5.0, 2008).

## **4.5 Wastewater Treatment Plant (WWTP) dataset**

A wastewater treatment plant database containing 527 cases representing 527 days of operation is used in this study for the generation of production rules (to be used as the fault libraries for process fault diagnosis). It was collected by Poch and made publicly available by Bejar and Corts of the University of Catalonia, Spain (Sanchez et al., 1997). Each data case is represented by 38 attributes, i.e. process parameters/variables. Out of the 38 attributes, 7 are the output variables, 9 are related to process operational performance, and the rest are the variables related to the influent into the biological unit of a wastewater treatment plant. The 38 attributes are listed in Table 4.1. All attributes are numeric and have continuous values. The units for all parameters are  $\text{g}/\text{m}^3$  except the input flow to plant (Q-E) which has the unit  $\text{m}^3/\text{day}$ . The WWTP used in this study consists of Sequential Batch Reactor (SBR).

**Table 4.1** Process variables of the wastewater treatment plant (Wastewaterdatabase, 2006)

---

No.	Attribute	
<b><i>Pre Treatment</i></b>		
1	Q-E	(input flow to plant)
2	ZN-E	(input zinc to plant)
3	PH-E	(input pH to plant)
4	DBO-E	(input biological demand of oxygen to plant)
5	DQO-E	(input chemical demand of oxygen to plant)
6	SS-E	(input suspended solids to plant)
7	SSV-E	(input volatile suspended solids to plant)
8	SED-E	(input sediments to plant)
9	COND-E	(input conductivity to plant)
<b><i>Primary Treatment</i></b>		
10	PH-P	(input pH to primary settler)
11	DBO-P	(input biological demand of oxygen to primary settler)
12	SS-P	(input suspended solids to primary settler)
13	SSV-P	(input volatile suspended solids to primary settler)
14	SED-P	(input sediments to primary settler)
15	COND-P	(input conductivity to primary settler)
<b><i>Secondary Treatment</i></b>		
16	PH-D	(input pH to secondary settler)
17	DBO-D	(input biological demand of oxygen to secondary settler)
18	DQO-D	(input chemical demand of oxygen to secondary settler)
19	SS-D	(input suspended solids to secondary settler)
20	SSV-D	(input volatile suspended solids to secondary settler)
21	SED-D	(input sediments to secondary settler)
22	COND-D	(input conductivity to secondary settler)
<b><i>Performance Inputs</i></b>		
23	RD-DBO-P	(performance input biological demand of oxygen in primary settler)
24	RD-SS-P	(performance input suspended solids to primary settler)
25	RD-SED-P	(performance input sediments to primary settler)
26	RD-DBO-S	(performance input biological demand of oxygen to secondary settler)
27	RD-DQO-S	(performance input chemical demand of oxygen to secondary settler)
28	RD-DBO-G	(global performance input biological demand of oxygen)
29	RD-DQO-G	(global performance input chemical demand of oxygen)
30	RD-SS-G	(global performance input suspended solids)
31	RD-SED-G	(global performance input sediments)
<b><i>Outputs</i></b>		
32	PH-S	(output pH)
33	DBO-S	(output biological demand of oxygen)
34	DQO-S	(output chemical demand of oxygen)
35	SS-S	(output suspended solids)
36	SSV-S	(output volatile suspended solids)
37	SED-S	(output sediments)
38	COND-S	(output conductivity)

---

## The Process Details

The plant is an activated sludge process located in Manresa, a town near Barcelona (Catalonia, Spain) population of 100,000 inhabitants. It treats a daily flow of approx. 35,000 m<sup>3</sup> comprising mainly domestic wastewater, although other wastewaters from industries located near the town are also received in the plant. The plant consists of three main treatment sections (Albazzaz et al., 2005):

- (i) Pre-treatment,
- (ii) Primary treatment and
- (iii) Secondary treatment by means of activated sludge.

The database has been used for studies in classification by Sanchez et al., (1997) where two methods, the K-means clustering method and Linneo+ methodology, a knowledge acquisition tool with unsupervised learning strategy, were investigated. (Sanguesa and Cortes (1997) used the data to study a possibilistic network. Hunag and Wang (1999) used the data in developing fuzzy casual networks. Wang et al. (2004) used this data set to present an approach for multidimensional visualization of multiple principal coordinates using a technique called parallel coordinates for the purpose of process monitoring. This data was also used for the historical data analysis and an empirical comparison was made between multidimensional visualization using parallel coordinates, PCA based multivariable statistical process control charts, the T2 and SPE charts, and a clustering approach (Albazzaz et al., 2005). Ma and Wang (2009) used the data for a new approach to data mining using Genetic Programming. Dellana and West (2009) used the data in their work for the predictive modelling of wastewater. West and Mangiameli (2000) employed the WWTP data for the identification of process conditions.

There is an adequate amount of research work in literature where people have used different dataset for the fault detection and diagnosis systems but the data set they used, is normally not available publically. Researchers usually use dataset from industry with which they have research ties. Furthermore, all the researchers who have used this WWTP data-set which is non-linear in nature, only rely on it and

didn't use any other data-set for any validation etc. There is only one publication in the writer's knowledge that use another dataset for validation but the dataset used there was from their own source and that dataset is not available.

Out of the 38 attributes, the database has some missing values for about 144 days out of the 527. Missing data is a common problem for data mining, because in many of these situations, the missing data cannot be re-collected or reproduced. Albazzaz et al. (2005) used eight different methods available in the commercial statistics software system SPSS (ver. 11.5) to deal with the missing values in the database. Only three approaches, i.e., linear trend at point, series mean, and series median were able to give estimations for all the missing values. It was found that the estimations for the missing values using these three methods were reasonably close. A further comparison was made between series mean and linear trend. Both approaches were used to fill in all the missing values, and then calculated such statistics as minimum, maximum, mean, median, and standard deviation for all the 38 attributes in the wastewater treatment plant dataset. It was found that the differences in these statistics between the two approaches were negligible. Eventually the series mean method was used to fill in the missing values. The cleaned data (data after the missing values were filled) is used in this work.

## **4.6 Development of Production Rules from WWTP data**

The data discussed above is used for the development of production rules via decision trees that are intended to be used in the diagnostic section of the hybrid system.

### ***Step 1***

To classify the data into normal, high or low operating conditions, either a descriptive approach of data mining (i.e. pattern recognition or clustering) or a prior knowledge about the process parameter under observation can be used. In this work, the normal value of a parameter using prior knowledge is used rather than using pattern recognition and clustering. The rationale is that it is mandatory to provide the normal value (value that can be used to classify the data) of a parameter under study



for Inductive Data Mining using See5.0. Furthermore, this information is readily available from the process. Initially the parameter SS-S, i.e. Suspended Solids out from Secondary stage of wastewater is used for the study and the normal value is 20mg/L.

### ***Step 2***

The final product value (concentration, flow rate, etc.) in the chemical or process industries is not always fixed to a single numeric value. There exists a range (or limits) for a parameter such that if the output value for the parameter falls between the limits, it is assumed to be normal; otherwise it is high or low if the value is above or below the range for normal operation. From the diagnostic point of view, this is known as the Tolerance Limit of a parameter. As an example, tolerance limit of 15% is used below. This 15% tolerance limit means that if the value of the parameter under observation changes more than 15% of its set value, it indicates the presence of a fault.

### ***Step 3***

By applying the tolerance limit on the normal value, the range for normal operation of SS-S is calculated to be 17-23mg/L. The entire data is then split into normal, high and low classes using this limit with each class representing the corresponding operating condition. Although the low value of output product (concentration of pollutant in this case) in a wastewater is desired, the data that corresponds to a low value of SS-S is classified as at low operating condition.

### ***Step 4***

This data is then segregated into training and test data. Training data is used by the algorithm for its learning process, whereas the test data is unseen (by the algorithm) data which it uses for the validation of the model and its results. In this exercise, 75% of the total data is used for training and 25% for test.

## Results from WWTP Dataset

After the processing of data as described above, it is then used in See5.0 to obtain production rules via decision tree. As we intend to use these results in fault diagnosis, the set of results that explains the cause-consequence relationship for process parameters while the process is in normal operating condition is rejected. This means that the production rules belonging to the normal class are not used any further, although this information may be important and helpful for other decision-making activities. The number of cases occurring in different classes for the WWTP data is reported in Table 4.2.

**Table 4.2** WWTP data in different classes

<b>Class</b>	<b>Total Data</b>	<b>Training</b>	<b>Test</b>
<i>Normal</i>	199	149	50
<i>Low</i>	193	143	50
<i>High</i>	135	101	34

The production rules explaining only low and high classes are listed in Table 4.3.

**Table 4.3** Production rules for SS-S

---

**CLASS HIGH**

---

**RULE NO 1:**

IF RD-SS-G  $\leq$  86.9  
& DBO-E  $>$  89.0  
& SS-E  $>$  168.0  
(53)

**RULE NO 3:**

IF RD-SS-G  $\leq$  83.3  
& DBO-E  $>$  89.0  
& SS-E  $\leq$  168.0  
(20)

**RULE NO 2:**

IF  $86.9 < \text{RD-SS-G} \leq 90.2$   
& SS-E  $>$  214.0  
& DBO-E  $\leq$  283.0  
(17)

**RULE NO 4:**

IF RD-SS-G  $>$  90.2  
& SED-S  $>$  0.03  
& SSV-E  $\leq$  71.7  
& Q-E  $\leq$  33999  
& SS-E  $>$  238  
(7)

---

**CLASS LOW**

---

**RULE NO 1:**

IF  $86.9 < \text{RD-SS-G} \leq 90.2$   
&  $144.0 < \text{SS-E} \leq 214.0$   
& DQO-E  $\leq$  340.0  
(2)

(61)

**RULE NO 2:**

IF  $90.2 < \text{RD-SS-G} \leq 90.8$   
& SED-S  $\leq$  0.03  
& SS-E  $\leq$  176.0  
(15/2)

**RULE NO 6:**

IF RD-SS-G  $>$  92.7  
& SED-S  $\leq$  0.03  
&  $194.0 < \text{SS-E} \leq 262.0$   
& RD-DQO-S  $>$  66.7  
(23/1)

**RULE NO 3:**

IF RD-SS-G  $>$  92.7  
& SED-S  $\leq$  0.03  
& SS-E  $>$  194.0  
& RD-DQO-S  $\leq$  66.7  
(27)

**RULE NO 7:**

IF RD-SS-G  $>$  92.7  
& SED-S  $\leq$  0.03  
& SS-E  $>$  262.0  
& DQO-S  $\leq$  17.0  
& RD-DQO-S  $>$  81.8  
& PH-D  $\leq$  7.9  
(7)

**RULE NO 4:**

IF RD-SS-G  $>$  90.2  
& SED-S  $>$  0.03  
& SSV-E  $>$  71.7  
(3)

**RULE NO 8:**

IF RD-SS-G  $>$  92.7  
& SED-S  $\leq$  0.03  
& SS-E  $>$  262.0  
& DQO-S  $\leq$  17.0  
& RD-DQO-S  $>$  81.8  
& PH-D  $>$  7.9  
& Q-E  $\leq$  41073  
(2)

**RULE NO 5:**

IF RD-SS-G  $>$  90.8  
& SED-S  $\leq$  0.03  
& SS-E  $\leq$  194.0

---

These derived rules are then used in fault libraries for the diagnosis of an abnormal event. Fault libraries are a knowledge-base that the diagnostic algorithm uses to determine the reason for, and solution to, an abnormal event in a process. Rule No.1 from the high class corresponding to high operating condition of the process is used for illustration and explanation as follows.

***Rule No 1:***

If  $RD-SS-G \leq 86.9$

&  $DBO-E > 89.0$

&  $SS-E > 168.0$

**(53)**

Production Rule No. 1 implies that if at any stage of operation the value of global performance input for suspended solids  $RD-SS-G$  starts decreasing, and at that instance the biological demand of oxygen and suspended solids inputs to the plants are greater than 89.0 and 168.0 respectively, there is a probability of at least 53% that the output value for suspended solids from the plant will be higher than 23 mg/l (i.e. the maximum allowable limit for  $SS-S$ ). The number (53) is the number of cases that have been classified to be at high operating condition (w.r.t.  $SS-S$ ) by the information obtained from Rule No. 1. This probability can then determine if the total numbers of cases that belong to the high operating condition are known (101 in this case, see Table 4.2).

## **4.7 Weighting of Production Rules**

A modification is suggested here for the enhancement of the production rules to be used in a diagnosis system. A methodology for weighting the production rules is presented and the production rules that correspond to the same class (for example high or low class) are arranged according to their weighting to further simplify and enhance the process of diagnosis.

There are a total of four production rules that describe the reason for a shift of the process to high operating condition. If a fault has been detected in the process

indicating high operating condition and the plant operator/engineer is provided with the four possible cause-symptom relationships in the form of production rules for diagnosis, then the first essential question to be answered is: “Which of the four production rules is most important and requires attention first?” As a timely response is very critical at that stage, information that can provide an operator/engineer with guidance in decision making regarding the most promising production rules is very helpful. As an example, there are 8 production rules that give comprehensive information about the reason why the process is in low operating condition. The number of production rules normally increases with an increase in the amount of data, therefore a large number of production rules are possible if there is much historical/operational data available from a process. In this case, the number of production rules can itself pose a challenge in the diagnosis of abnormal event. A simple concept of “weighting function” is introduced here that will help prioritize the knowledge obtained from production rules in the diagnostic process. Weighting function can be obtained from the relationship described in equation 4.6 below:

$$wtF_i = \frac{W_i}{\sum_{i=1}^n W_i} \quad (4.6)$$

where

$wtF_i$  = weight function of production rule “i”

$\sum_{i=1}^n W_i$  = total weight for all production rules (belonging to one class)

$W_i$  = weight of production rule “i” (calculated from equation 4.7) given as:

$$W_i = \left( \frac{C_c - C_{mc}}{T_c} \right) / n \quad (4.7)$$

when

$C_c$  = number of cases correctly classified by the production rule “i”

$C_{mc}$  = number of cases misclassified by the production rule “i”

$T_c$  = total number of cases that belong to a class “i” (e.g. high or low)

$n$  = number of leaf nodes/variables in production rules

If the production rules are arranged according to their weight function then this will help a plant operator/engineer to prioritize the knowledge during the decision-making process. The weight function for the production rules is reported in Table 4.4. The production rules reported in Table 4.3 are then rearranged using this weighting concept. This concept will help develop a hierarchy of production rules on the basis of their ability to diagnose a fault more accurately. The production having the highest weight function becomes the first candidate out of all those possible in the event of a fault diagnosis. The revised production rules are reported in Table 4.5.

**Table 4.4** Production rules for low and high class by their weight function (ascending order)

<b>Class Low</b>		<b>Class High</b>	
<i>Production Rule No.</i>	<i>Weight Function</i>	<i>Production Rule No.</i>	<i>Weight Function</i>
3	0.143	1	0.173
5	0.048	3	0.067
6	0.038	2	0.056
4	0.030	4	0.014
7	0.010		
2	0.006		
1	0.003		
8	0.002		

**Table 4.5** Production rules for SS-S arranged by weight function

---

**CLASS HIGH**

---

**RULE NO 1(RANK UNCHANGED):**  
IF RD-SS-G  $\leq$  86.9  
& DBO-E  $>$  89.0  
& SS-E  $>$  168.0  
(53)

**RULE NO 3(FORMALY RULE NO 2):**  
IF  $86.9 <$  RD-SS-G  $\leq$  90.2  
& SS-E  $>$  214.0  
& DBO-E  $\leq$  283.0  
(17)

**RULE NO 2(FORMALY RULE NO 3):**  
IF RD-SS-G  $\leq$  83.3  
& DBO-E  $>$  89.0  
& SS-E  $\leq$  168.0  
(20)

**RULE NO 4(FORMALY RULE NO 4):**  
IF RD-SS-G  $>$  90.2  
& SED-S  $>$  0.03  
& SSV-E  $\leq$  71.7  
& Q-E  $\leq$  33999  
& SS-E  $>$  238  
(7)

---

**CLASS LOW**

---

**RULE NO 1(FORMALY RULE NO 3):**  
IF RD-SS-G  $>$  92.7  
& SED-S  $\leq$  0.03  
& SS-E  $>$  194.0  
& RD-DQO-S  $\leq$  66.7  
(27)

**RULE NO 5(FORMALY RULE NO 7):**  
IF RD-SS-G  $>$  92.7  
& SED-S  $\leq$  0.03  
& SS-E  $>$  262.0  
& DQO-S  $\leq$  17.0  
& RD-DQO-S  $>$  81.8  
& PH-D  $\leq$  7.9  
(7)

**RULE NO 2(FORMALY RULE NO 5):**  
IF RD-SS-G  $>$  90.8  
& SED-S  $\leq$  0.03  
& SS-E  $\leq$  194.0  
(61)

**RULE NO 6(FORMALY RULE NO 2):**  
IF  $90.2 <$  RD-SS-G  $\leq$  90.8  
& SED-S  $\leq$  0.03  
& SS-E  $\leq$  176.0  
(15/2)

**RULE NO 3(FORMALY RULE NO 6):**  
IF RD-SS-G  $>$  92.7  
& SED-S  $\leq$  0.03  
&  $194.0 <$  SS-E  $\leq$  262.0  
& RD-DQO-S  $>$  66.7  
(23/1)

**RULE NO 7(FORMALY RULE NO 1):**  
IF  $86.9 <$  RD-SS-G  $\leq$  90.2  
&  $144.0 <$  SS-E  $\leq$  214.0  
& DQO-E  $\leq$  340.0  
(2)

**RULE NO 4(FORMALY RULE NO 4):**  
IF RD-SS-G  $>$  90.2  
& SED-S  $>$  0.03  
& SSV-E  $>$  71.7  
(3)

**RULE NO 8(FORMALY RULE NO 8):**  
IF RD-SS-G  $>$  92.7  
& SED-S  $\leq$  0.03  
& SS-E  $>$  262.0  
& DQO-S  $\leq$  17.0  
& RD-DQO-S  $>$  81.8  
& PH-D  $>$  7.9  
& Q-E  $\leq$  41073  
(2)

---

Similar to the development of production rules for the one output variable discussed above (i.e. Suspended Solids, SS-S), the production rules for other output variables namely Volatile Suspended Solids(VSS-S), Sediments (SED-S), Biological Oxygen Demand (DBO-S), and Chemical Demand of Oxygen (DQO-S) are developed and presented in Tables 4.6 to 4.9 respectively.

**Table 4.6** Production rules for VSS-S arranged by weight function

<b>CLASS HIGH</b>	
<p><b>RULE NO 1(FORMALY RULE NO 2):</b>            IF COND – P &gt; 921            &amp; COND – D &gt; 51            &amp; SS-D &gt; 2550            (6)</p>	<p><b>RULE NO 2(FORMALY RULE NO 1):</b>            IF COND – P &gt; 921            &amp; COND – D &gt; 51            &amp; DBO – D ≤ 78.9            &amp; PH – D ≤ 104            &amp; SS – D &gt; 1846            &amp; SED – D &gt; 11            &amp; ZN – E ≤ 1.8            &amp; RD – DBO – P ≤ 8.0            &amp; 235 &lt; RD – SED – P ≤ 274            (3)</p>
<b>CLASS LOW</b>	
<p><b>RULE NO 1(RANK UNCHANGED):</b>            IF COND – P ≤ 921            &amp; SS – D ≤ 863            (5)</p> <p><b>RULE NO 2(FORMALY RULE NO 4):</b>            IF COND – P ≤ 921            &amp; SS – D &gt; 863            &amp; COND – E &gt; 829            &amp; RD-DBO-G &gt; 90.4            (3)</p>	<p><b>RULE NO 3(FORMALY RULE NO 2):</b>            IF COND – P &gt; 921            &amp; COND – D &gt; 51            &amp; DBO – D ≤ 78.9            &amp; SS – D ≤ 921            (3)</p> <p><b>RULE NO 4(FORMALY RULE NO 3):</b>            IF COND – P ≤ 921            &amp; SS – D &gt; 863            &amp; COND – D &gt; 829            &amp; RD – DBO – G ≤ 90.4            &amp; SED – D ≤ 12            (3)</p>



**Table 4.7** Production rules for SED-S arranged by weight function

---

***CLASS HIGH***

---

**RULE NO 1(RANK UNCHANGED):**

IF RD-SS – G > 79.5  
& RD – DQO –G > 81.7  
& COND – D ≤ 39  
& RD – SED – P > 204  
(3)

**RULE NO 2(RANK UNCHANGED):**

IF RD-SS – G > 79.5  
& SSV – P > 63.6  
& PH – E > 7.6  
& 39 < COND – D ≤ 52  
& RD – DQO –G > 86.8  
(4)

---

***CLASS LOW***

---

**RULE NO 1(FORMALY RULE NO 2):**

IF RD – DQO – G ≤ 55.6  
& COND – P > 1165  
(3)

**RULE NO 3(FORMALY RULE NO 1):**

IF RD – DQO – G ≤ 55.6  
& COND – P ≤ 1165  
& RD – DQO – S ≤ 74.4  
(3)

**RULE NO 2(FORMALY RULE NO 3):**

IF COND – D > 103  
& 55.6 < RD – DQO – G ≤ 66  
& RD – SED – P ≤ 220.0  
(3)

**RULE NO 4(FORMALY RULE NO 4):**

IF COND – D > 103  
& 55.6 < RD – DQO – G ≤ 66  
& RD – SED – P > 220.0  
& DBQ – E > 488  
(3)

---

**Table 4.8** Production rules for DBO-S arranged by weight function

---

**CLASS HIGH**

---

**RULE NO 1(FORMALY RULE NO 2):**

IF COND – P > 2340

& SS –D > 2550

(6)

**RULE NO 2(FORMALY RULE NO 1):**

IF COND – P > 2340

& 921 < SS –D ≤ 2550

& RD – SED – P ≤ 295

(2)

---

**CLASS LOW**

---

**RULE NO 1(RANK UNCHANGED):**

IF SS-D ≤ 921

& DQO-D ≤ 0.1

(6)

**RULE NO 2(FORMALY RULE NO 3):**

IF SS-D ≤ 921

& SS-P ≤ 278

& DQO-D > 0.2

(5)

**RULE NO 3(FORMALY RULE NO 2):**

IF SS-D ≤ 921

& SS-P ≤ 278

& 0.1 < DQO-D ≤ 0.2

& COND-D > 64

(3)

---

**Table 4.9** Production rules for DQO-S arranged by weight function

---

**CLASS HIGH**

---

**RULE NO 1(FORMALY RULE NO 2):**

IF SED-E  $\leq$  7.0  
& RD-DBO-G  $>$  88.7  
& RD-SS-P  $\leq$  64  
& DBO-P  $>$  134  
**(4)**

**RULE NO 2(FORMALY RULE NO 1):**

IF SED-E  $\leq$  7.0  
& RD-DBO-G  $>$  88.7  
& RD-SS-P  $>$  64  
& RD-SED-G  $\leq$  99.7  
& DBO-D  $>$  64.5  
& DBQ-E  $>$  297  
& DBO-P  $>$  296  
& ZN-E  $>$  0.4  
**(2)**

---

**CLASS LOW**

---

**RULE NO 1(FORMALY RULE NO 4):**

IF SED-E  $\leq$  7.0  
& RD-DBO-G  $\leq$  88.7  
& DBO-P  $\leq$  145  
& RD-SED-P  $\leq$  317  
& RD-DBO-P  $\leq$  7.8  
& RD-DBO-S  $>$  16  
& DBQ-E  $>$  319  
**(6)**

**RULE NO 2(FORMALY RULE NO 3):**

IF SED-E  $\leq$  7.0  
& RD-SS-P  $>$  64  
& RD-SED-G  $>$  99.7  
& RD-SS-G  $>$  90.7  
& DBO-P  $\leq$  146  
& RD-DBO-G  $>$  92.7  
**(3)**

**RULE NO 3(FORMALY RULE NO 2):**

IF SED-E  $\leq$  7.0  
& DBQ-E  $>$  297  
& DBO-P  $>$  145  
& DBO-D  $>$  64.5  
&  $84.1 < \text{RD-DBO-G} \leq 88.7$   
& COND-D  $>$  100  
&  $159 < \text{SS-P} \leq 228$   
& RD-DBO-P  $\leq$  7.8  
& RD-DBO-S  $\leq$  31  
**(4)**

---

## 4.8 Validation of Production Rules

Validation of a model is a very critical step in model development (either a mathematical model or model in the form of production rules). A model may initially produce very encouraging results but its accuracy on the new process or data set actually decides if the model is fit for the intended purposes. In this case study, the WWTP data set was divided into two portions i.e. 75% and 25%. The major portion of the data was used as training data-set for the learning of See 5.0 and the development of production rules. The remaining 25% is then used as a test data-set to validate the model (to measure the accuracy of the production rules See 5.0 developed on the training data-set). The following table summarizes the results obtained on the output variables used in this study.

**Table 4.10** Validation results of Production Rules developed by See 5.0

	<b>See 5.0 Results (% accuracy)</b>	
<b>Variable</b>	<b>Training data-set</b>	<b>Test data-set</b>
SS-S	97.5	96.9
VSS-S	97.3	95.5
SED-S	98.1	96.6
DBO-S	99.7	94.7
DQO-S	96.4	91.7

## 4.9 Concluding Remarks

A review of different data driven techniques is presented. The underlying principles of Inductive Data Mining, a technique selected for the research work, are discussed. Production Rules are developed using Inductive Data Mining on the Spanish data set of a wastewater treatment plant in order to build the knowledge base (fault libraries) of the diagnostic system. A new concept of Weight Factor is introduced and implemented on the production rules generated in the chapter. This concept helps arrange the production rules for easy retrieval.

# Chapter 5

## Diagnostic Algorithm and Java Application

### 5.1 Introduction

An algorithm is presented that uses the mathematical model from Chapter 3 for the detection of faults (as discussed in chapter 3) and the fault libraries that are made up of the production rules (derived from the WWTP dataset in chapter 4) for the diagnosis purposes. These component parts provide a complete fault detection and diagnosis system. The ideology behind development of this diagnostic system, together with the assumptions made, is discussed below, and is followed by the development of a software platform (a Java application). This Java application is used for simulation of the process using the mathematical model for the detection of faults. The implementation of the diagnostic algorithm enables the Java application to be used for the diagnosis of faults in the process. A discussion of the development and architecture of the application is presented, followed by the results obtained.

### 5.2 Diagnostic Algorithm

As discussed in Section 2.4, it has been shown by Venkatasubramanian and co-workers (2003b, 2003c, 2003a) that a diagnostic system is a valuable area of research in fault detection and diagnosis. This section presents a new diagnostic system for the detection and diagnosis of fault in a process.

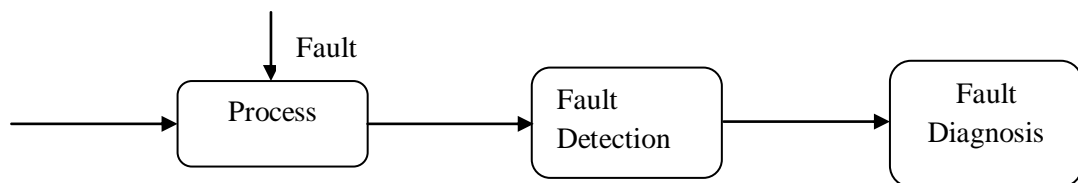
Analogous to the medical profession, where graduates have the opportunity to apply their expertise in many fields of medicine as a physician or a surgeon, it is proposed here that instead of using available techniques to perform simultaneously detection and diagnosis of faults, these techniques should be classified into Detection and Diagnosis techniques on the basis of their strengths. Furthermore, any future research should be focused on improving the specialty of a technique (either detection or diagnosis of faults). Based on this idea, a diagnostic algorithm is

presented here that uses analytical redundancy based upon the mathematical model for the detection of a fault and production rules for the diagnosis.

While developing the algorithm, the following assumptions were made:

- a) Due to the complexity, multiple faults are not considered in this study.
- b) For the detection of faults, the data is considered to be noise free.
- c) The term: “multi faults” used in this work (later in Table 5.2) refers to the two faults that occur simultaneously with distinct characteristics and no interconnecting relationship exists between the two.

The structure of a typical fault detection and diagnosis system is shown in Figure 5.1.



**Figure 5.1** A schematic of fault detection and diagnosis

**Fault detection** is the first step in a diagnosis system. Fault detection is determining whether a fault has occurred in the process under observation. It helps engineers/operators identify if a fault exists.

**Fault diagnosis** is the determination of the cause of the observed out-of-control status of a process variable. In other words, fault diagnosis is determining the type, location, magnitude and time of the fault. This step is essential in counteraction or elimination of a fault.

Some researchers split diagnosis further into the fault isolation and its identification - which when combined have the same meaning as that defined above. In this work, rather than fault isolation and identification, the term fault diagnosis is used. Another term, “diagnosis system”, when used in this work refers to both the fault detection and the diagnosis system rather than only the diagnosis system.

### 5.2.1 The Algorithm

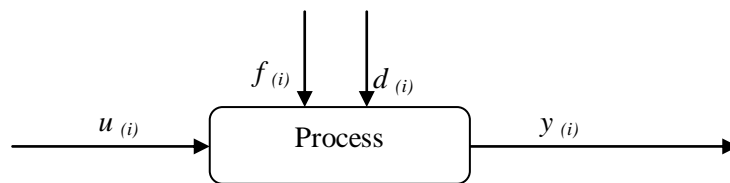
A comprehensive survey of available and most commonly used techniques for the detection and diagnosis of faults is given by Venkatasubramanian et al. (2003b, 2003c, 2003a). In their extensive review, it was shown that despite all the research for improvement in detection and diagnosis of process faults, it is a widely accepted

by most researchers that none of the techniques introduced so far have the potential to meet the key requirements of a practical diagnosis system (as defined by the ten key characteristics of a diagnostic system in Venkatasubramanian et al. (2003a)). One possibility for enhancing the effectiveness of a diagnosis system is to adopt the hybrid methodology. This is based on the assumption that these methods can complement each other's limitations in different areas, thus resulting in an overall improved diagnostic system. As an example, fault explanation through a causal chain is best done through the use of digraphs, whereas fault isolation might be difficult using digraphs due to the qualitative ambiguity (Venkatasubramanian et al., 2003a).

Building upon the detailed review by Venkatasubramanian et al. (2003b, 2003c, 2003a), this work presents a new approach for fault detection and diagnosis employing both mathematical models and historical data from a wastewater treatment plant (for detection and diagnosis respectively). The knowledge base of the diagnostic system is built up of the historical data using inductive data mining and application of the commercial software See5.0. The mathematical model is used for the detection of faults in the process. A step by step overview of the system is given below.

**a) *Fault Detection***

Assume the system has inputs  $u_{(i)}$  and outputs  $y_{(i)}$  ( $i = 1$  to  $n$ ). Under the fault free operational mode, the output  $y_{(i)}$  will conform to the input  $u_{(i)}$ . Now consider the system with a fault  $f_{(i)}$  and/or a process disturbance  $d_{(i)}$  as shown below:

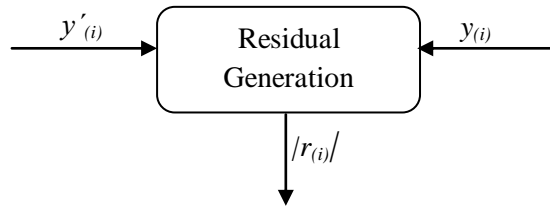


**Step 1: Residual Generation**

The first step in the system is the residual generation where residual ( $r$ ) is defined as an error in a result. In the context of fault detection and diagnosis, it is the difference between the actual and the desired value of any process parameter. The residual in the output parameter can be calculated from the following equation:

$$r_{(i)} = y_{(i)} - y'_{(i)}$$

where  $y_{(i)}$  is the output variable and  $y'_{(i)}$  is the desired (set) value for parameter  $y_{(i)}$ . For the purpose of its implementation in computer code, although the value of  $r_{(i)}$  is calculated above but  $|r_{(i)}|$  will be used in next steps (the Java application need only positive value to decide if the process is performing above or under the control limits).



The desired value  $y'_{(i)}$  can be obtained in a number of ways. It includes historical data, prior knowledge and principal models of the process. A statistical method can be applied on historical data to determine a value of the output that is considered to be in the normal operating state. Prior knowledge from plant engineers and operators can be useful in deciding the normal operating value for any output variable. Another way to obtain the normal operating value for a process variable as used in this work is the use of mathematical models. A mathematical model can be obtained by carrying out a simple mass and energy balance on the process under consideration. For a given input  $\mathbf{u}$ , it is possible to find out the expected normal value of the output via mathematical models by the aid of computer simulations. In this study, the wastewater mathematical model (chapter 3) is used for the simulation. And before it can be employed for the residual generation in this step, it was validated against the wastewater data-set.

## Step 2: Residual Evaluation

Once the residual is generated in Step 1, the next critical step is to evaluate this residual to establish if any inconsistency exists in the system. This is known as residual evaluation. The inputs to this step are the outputs from residual generation, i.e.  $|r_{(i)}|$ , and  $\tau'_{(i)}$  which is the tolerance for each residual generated. Tolerance is similar to the desired value and can be obtained in different ways. Tolerance helps to identify if the process, although deviated from the expected value, is still within the

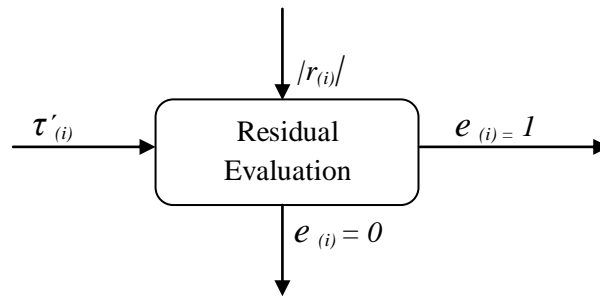


normal operating state. If not, then determine the direction of the drift, i.e. +ve (operating state above the allowable maxima; i.e. high) or -ve (operating state below the allowable minima; i.e. low). The final calculation in this step yields the value of error  $e$  (which potentially indicates a fault), and it can be calculated from the following relationships:

$$e_{(i)} = 0, |r_{(i)}| \leq \tau'_{(i)}$$

*and*

$$e_{(i)} = 1, |r_{(i)}| > \tau'_{(i)}$$

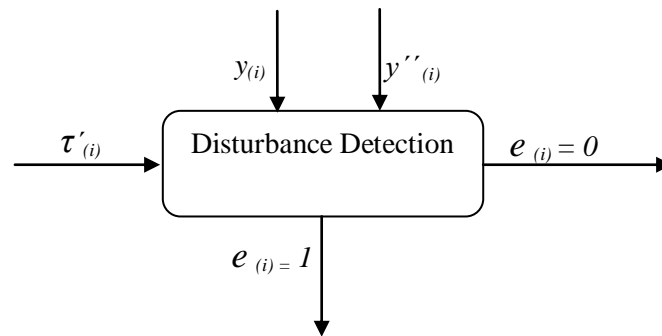


A value of  $e_{(i)} = 0$  indicates that the system is in normal operating state, where a value of  $e_{(i)}=1$  indicates the presence of a potential fault. Further analysis is carried out once the value of  $e_{(i)}$  has been assigned.

### Step 3: Disturbance Detection

The decoupling of a disturbance from a fault is a key research area in fault detection and diagnosis. Recursive use of mathematical models is employed in this diagnosis system which can isolate a disturbance from a process fault. A mathematical model is employed using process input  $u_{(i)}$  to calculate the expected value of the output variable, i.e.  $y'_{(i)}$ . With the new expected value of the output variable and output value from the process  $y_{(i)}$ , then Step 2 and Step 3 are repeated for the detection of disturbance in this step. The value of  $e$  now determines whether this unwanted incident in the process is an uncontrollable input (process disturbance) or a fault;

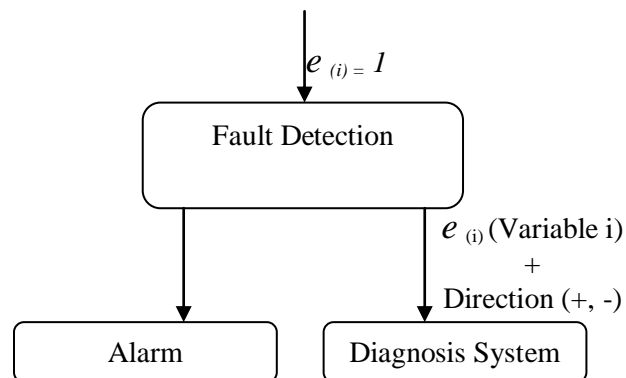
where  $e_{(i)} = 0$  indicates a disturbance in input  $u_{(i)}$ ; and  $e_{(i)} = 1$  indicates a fault in the process.



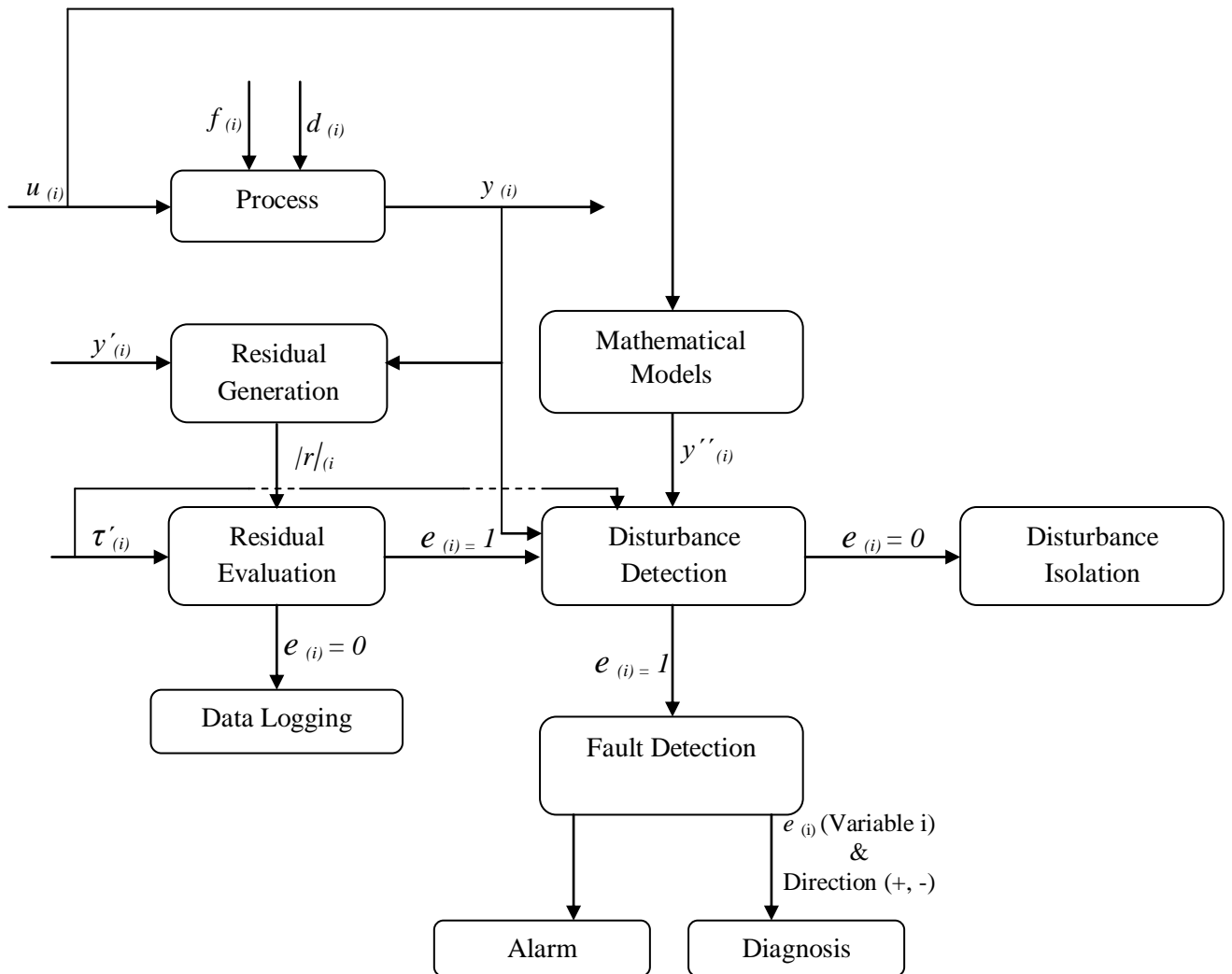
Depending on the nature of the disturbance, i.e. step, ramp or continuous (reconfiguration of input), the new calculated value of  $y''_{(i)}$  can be passed onto Step 2 for either temporary or permanent use as the set value of output  $y_{(i)}$ .

#### Step 4: Fault Detection

If Step 3 produces  $e_{(i)} = 1$ , then it indicates the presence of a fault in the process which effects the value of the output variable  $y_{(i)}$ . The results are then passed to the diagnosis section of the system along with the information about the variable, i.e.  $y_{(i)}$ , and the direction of fault, i.e. (+ve or -ve drift).



An integrated overview of the fault detection section for this diagnostic methodology is given in Figure 5.2.



**Figure 5.2** Integrated view of Fault detection

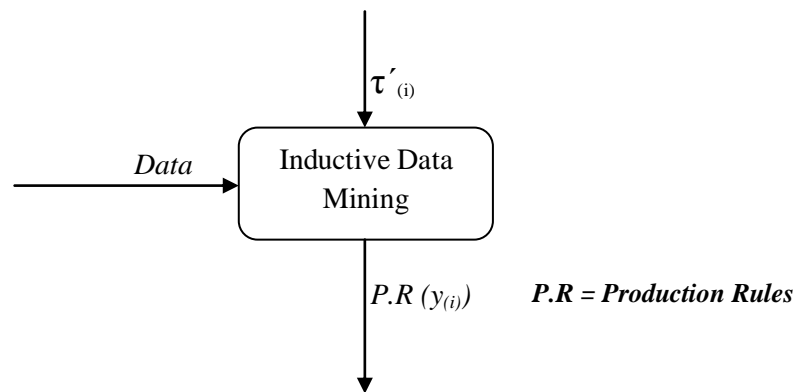
**b) Fault Diagnosis**

After the detection of a fault, the output variable away from the normal operating condition and its direction (+ve drift or -ve drift, where +ve refers to the process performing above the control limit and -ve refers to its performance under the control limit) is then passed onto the diagnosis section of this diagnostic system. For diagnosis of a fault, the major requirement is a knowledge base that can be used to compare symptoms and produce a corrective action for the fault under observation. In this diagnostic methodology, the knowledge base is produced from the historical data. Valuable information and hidden interactions between the process variables can

be obtained by exploring the data. The commercial software See5.0 is used in this work to conduct inductive data mining on the historical data available from a wastewater plant. The results can be obtained in the form of a decision tree or production rules.

### Step 1: Inductive data mining

Two inputs are required in this step in order to produce a meaningful relationship between the measured variables of the process, i.e. data, and the classification basis for the data. The data is classified using the numeric values from the tolerance  $\tau'_{(i)}$  used in the detection section of this work. It should be noted that classification of data needs to be carried out on every measurable output  $y'_{(i)}$  using its respective tolerance  $\tau'_{(i)}$ . The production rules for each of the variables used are obtained here. These production rules explain the reason for the deviation of a variable from its normal operating state. There will be essentially two sets of production rules, one that explains the +ve drift of the variable and other for -ve drift.



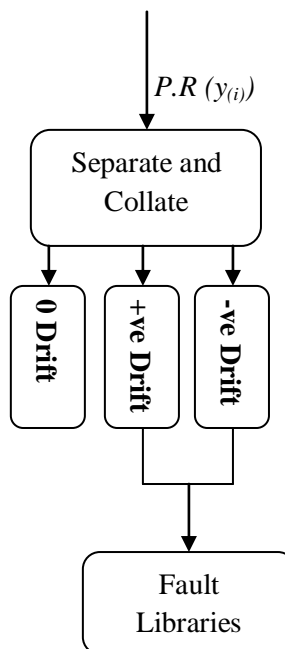
### Step 2: Fault Libraries

From the production rules obtained in the above step, fault libraries are built in order to compare the fault symptoms and hence correctly diagnose the fault. The structure of a fault library used in this work is as follows;

**Table 5.1** The fault libraries used in the knowledge base of the diagnosis system

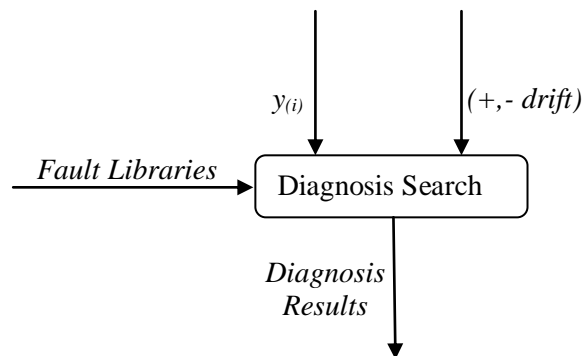
Fault Library								
Parameter	$y_{(1)}$		$y_{(2)}$		$y_{(3)}$		$y_{(n)}$	
Drift	+	-	+	-	+	-	+	-
Production Rules	If....	If....	If....	If....	If....	If....	If....	If....
	and...	and...	and...	and...	and...	and...	and...	and...
	and...	and...	and...	and...	and...	and...	and...	and...
	and...	and...	and...	and...	and...	and...	and...	and...
	and...	and...	and...	and...	and...	and...	and...	and...
	and...	and...	and...	and...	and...	and...	and...	and...
	and...	and...	and...	and...	and...	and...	and...	and...
	then....	then....	then....	then....	then....	then....	then....	then....

The production rules belonging to each parameter need to be organized in the fault library for access by the system when required.

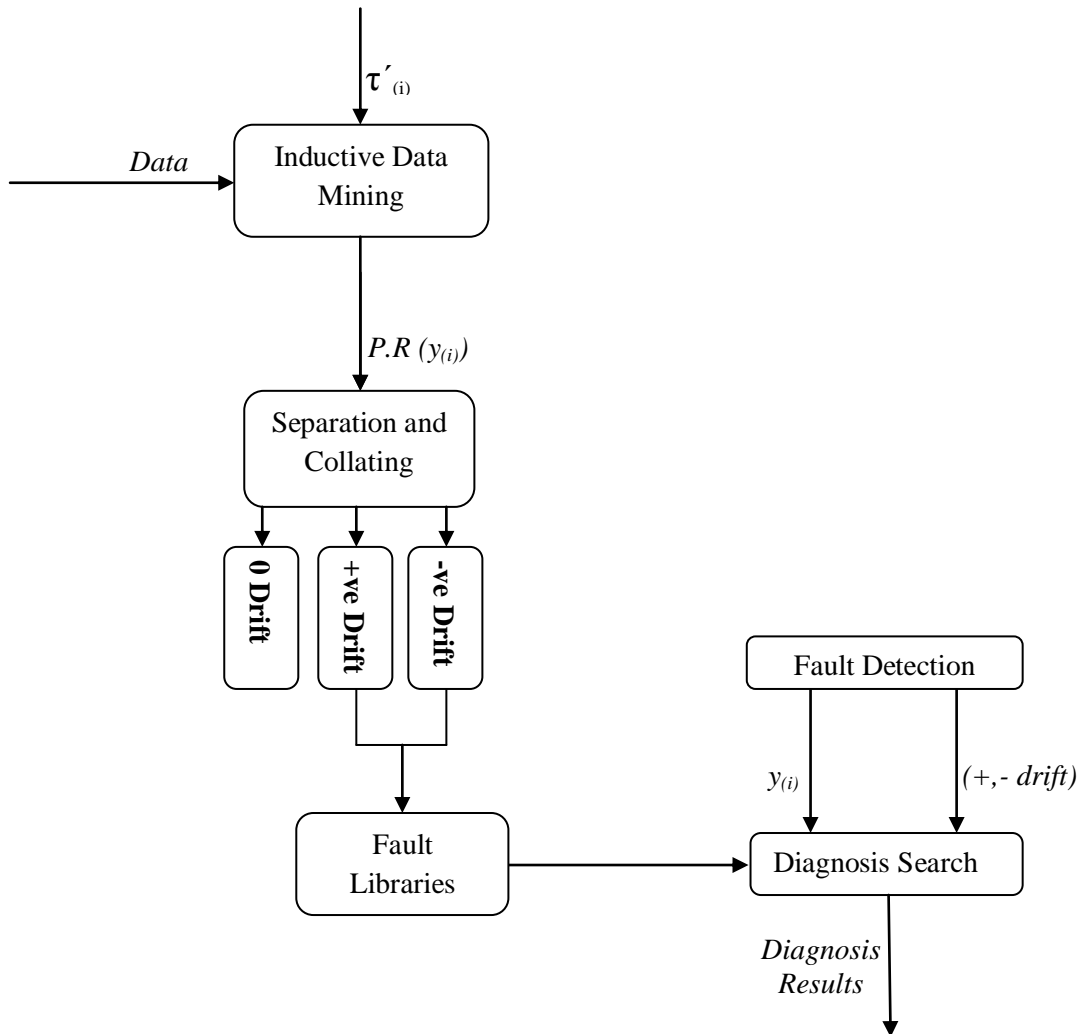


### Step 3: Diagnosis Search

This is the final step for the fault diagnosis system before it can present results for the mitigation of a fault, and hence recovery of the process back to its normal operating state. Using the information obtained from the fault detection system, i.e.  $y(i)$  and (+,- drift), the diagnosis system searches for the best candidate out of the fault library. The initial step of the search methodology is to match the variable to from those available in fault library. When the variable is identified; the search method will display the result (production rules) from the available drift (+ve or -ve).



An integrated overview of the complete diagnosis section for the proposed diagnostic system is presented in Figure 5.3.



**Figure 5.3** An integrated view of the components in the diagnosis section

### 5.3 Software Platform

After the successful development of a new diagnostic algorithm, efforts were made to develop a software platform that can be used for the simulation (to be used in the detection of faults using mathematical models), and diagnosis of faults using production rules from the fault libraries. Java Development Kit (**JDK** version 6) and NetBeans IDE (Integrated Development Environment version 6.1) is used for the development of the application. A brief introduction to Java and NetBeans is given below before presenting the development of the Java application.

## Java

Java is the most influential and widely used programming languages of the 21<sup>st</sup> century (TIBCO-Software-Inc, 2008). The language derives much of its syntax from languages C and C++ (once the most widely used languages and now second in application). Java applications can run on any Java Virtual Machine (JVM) regardless of computer architecture. Java is general-purpose, object-oriented and is specifically designed to have as few implementation dependencies as possible. Java is used for application software through to web applications. It was originally designed for use on digital mobile devices, such as cell phones. However, when Java 1.0 was released to the public in 1996, its main focus had shifted to use on the Internet. Since 1996 it has evolved as a successful language for use both Internet and other uses. A decade later, it is still an extremely popular language used by over 6.5million developers worldwide (Palmer, 2003).

The following reasons lead to the choice of Java rather than other languages in this work, relating to a few key principles from the original Java design (Leahy, 2010):

- **Easy to Use:** The fundamentals of Java came from a programming language called C++. Although a powerful language, it was felt to be too complex in its syntax, and inadequate for all of Java's requirements. Java is built using improved ideas, to provide a programming language that is powerful and simple to use.
- **Reliability:** Java needed to reduce the likelihood of fatal errors from programmer mistakes. With this in mind, object-oriented programming was introduced. Once data and its manipulation were packaged together in one place, it increased Java's robustness.
- **Security:** As Java was originally targeting mobile devices that would be exchanging data over networks, it was built to include a high level of security. Java is probably the most secure programming language to date.
- **Platform Independent:** Programs needed to work regardless of the machine upon which they were being executed. Java was written to be a portable language that does not depend upon a particular operating system, or the computer hardware.



In addition to these points, Java is available under the GNU General Public License (GPL), thus making it free software.

## NetBeans

The NetBeans IDE is an open-source integrated development environment. NetBeans IDE supports development of all Java application types. It provides the "plumbing" for the Graphical User Interface (GUI) in any Java application that conventionally every developer had to write themselves otherwise. NetBeans provides these entire straight "out of the box" thus saving a developer a significant amount of time and work (NetBeans, 2010).

NetBeans is available as open source free software. It can be run on most operating systems including Windows, Linux, Mac OS X and Solaris (Fears, 2008).

## 5.4 Architecture of the Application

The GUI of the application "WWTP-Simulation and Diagnosis" is presented in Figure 5.4. The application consists of the following five screens and a graphical applet.

1. U.D. Inputs
2. Flowsheet
3. Results
4. Online Data
5. Faults

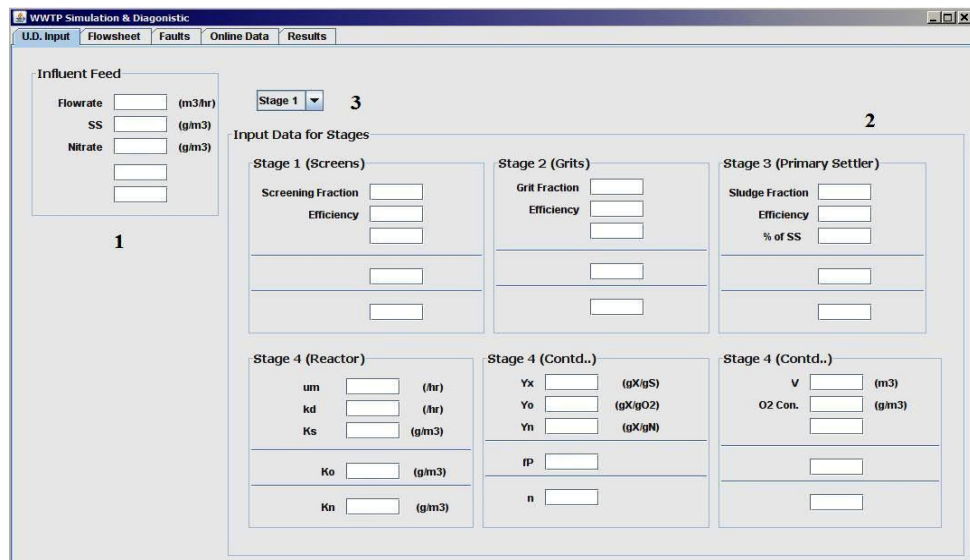


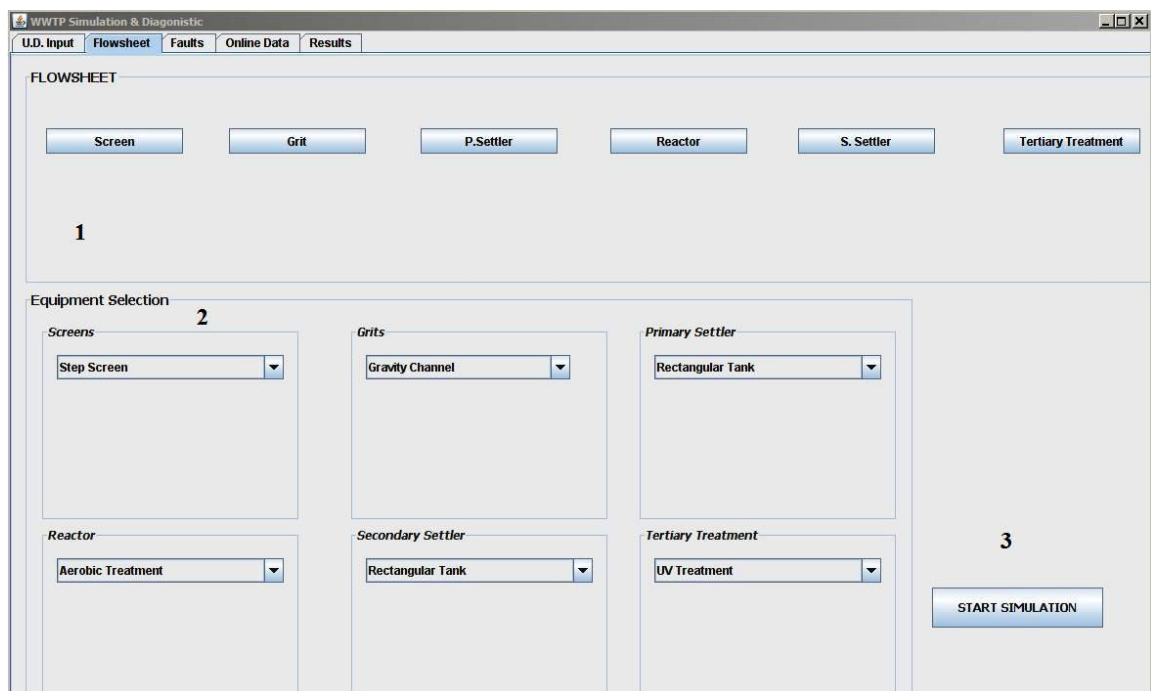
Figure 5.4 The Graphical User Interface (GUI) of the application

Figure 5.4 shows the main screen of the “WWTP-Simulation and Diagnostic” application. The first screen to be discussed is the “U.D. Inputs (used defined inputs)”.

The inputs to the WWTP are declared in the field marked as “1” in figure 5.4.

Field “2” points to the process parameters used in the different stages of the WWTP. Most of the process parameters are discussed in Chapter 3.

This application can be used for the simulation of different stages of the wastewater in a sequential and cumulative manner by selecting the number of stages (section “3” on figure 5.4). A point of emphasis here is that if Stage 2 is selected, then this will give the simulation of stages 1 and 2; whereas selecting stage 3 will simulate stage 1, 2 and 3.



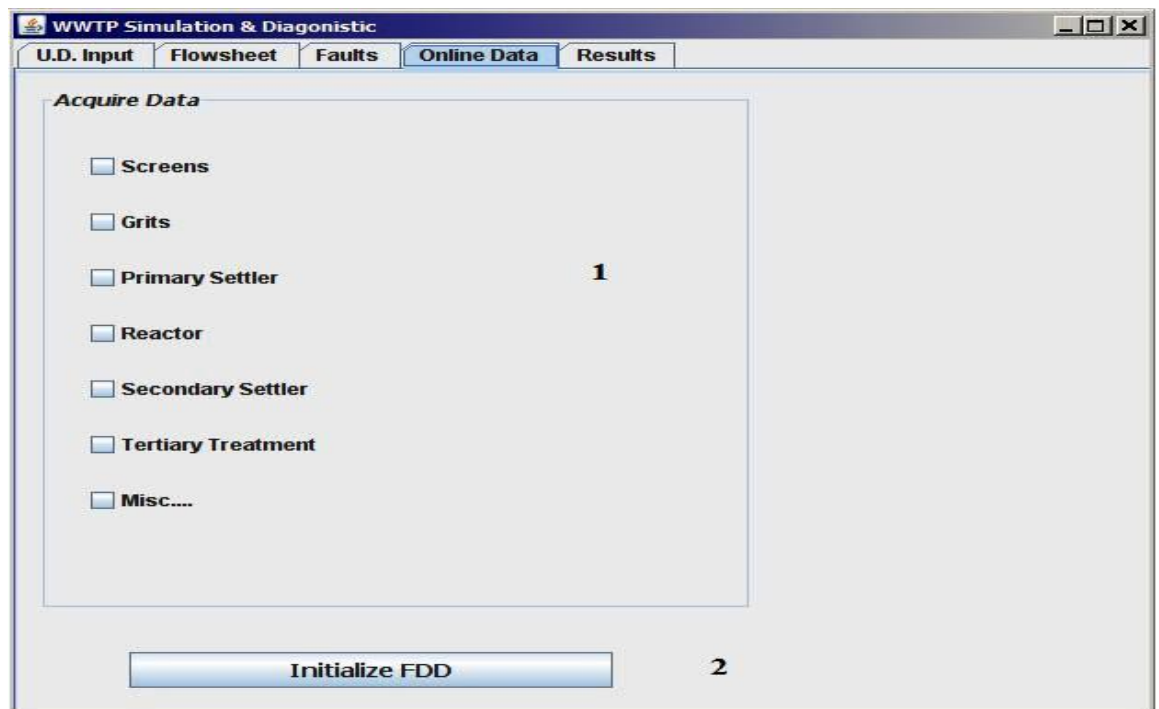
**Figure 5.5** Flowsheet screen of the WWTP application

Equipment selection, simulation command, and selection of an individual stage is developed in the “Flowsheet” tab of WWTP application. In figure 5.5, “3” indicates the simulation command button of the application that will initiate the simulation of required stages (as discussed above under figure 5.4) using the inputs on the “U.D. Inputs” screen and a file “WWTP.java” that includes the model

equations. The file “WWTP.java” is solved using the Runge-Kutta 4 (RK4) method (also coded in Java).

The next screen next to Flowsheet is the “Results” of this application. The numeric results obtained from the simulation of the model are displayed in the screen on the execution of the software. The “Results” screen is used mainly to display the results from the simulation. Although graphical presentation is used to display the simulation results, it was also considered useful to display the results in numeric form in the WWTP application. The reasons were the use of numeric data for the detection of faults, and the assumption that the simulated data may be used in future research work. For example, the data may be required for the validation of some experimental results.

The detection of faults occurred in the screen labeled as “Online Data”.

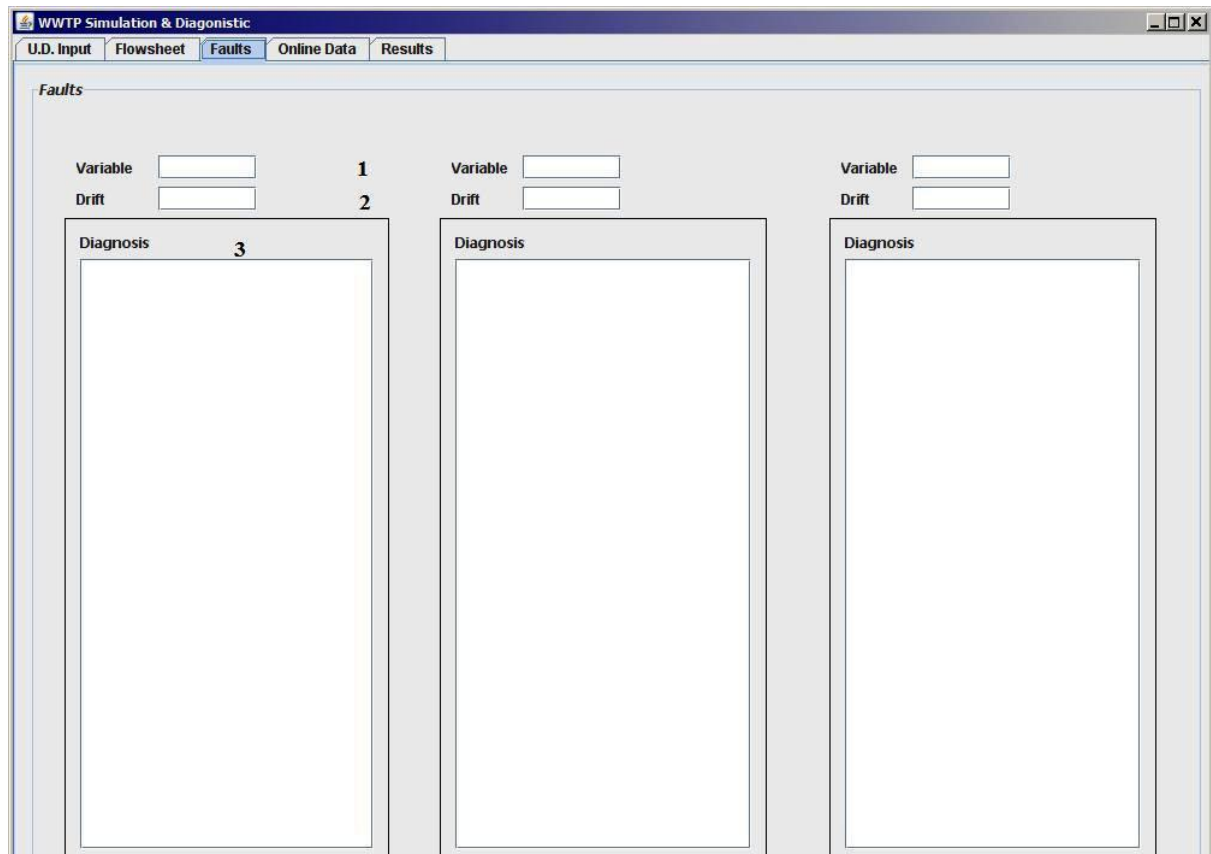


**Figure 5.6** Detection and Diagnosis Algorithm command window

The check boxes (No. “1” in figure 5.6) were designed so that that once selected, it will import the online data when coupled to the control system of a wastewater treatment facility. However, it was not possible to obtain technical support from any local wastewater treatment facility (preventing activation and use of this screen). For the WWTP application, the data for the detection of a fault is entered manually into the file. After the acquisition of data (in this application,

manual entry of data in the file), the detection and diagnosis of a fault can be initiated using the command box labeled “Initialize FDD” (labeled as “2” in figure 5.6).

The diagnostic results are the displayed in the last tab labeled as “Faults”. A screenshot of this screen is given in figure 5.7 below.



**Figure 5.7** Screen for the display of the diagnostic results

The variable suspected of faulty behavior (for example, SS, VSS, etc.) is displayed in the field labeled as “1” in figure 5.7, whereas the drift (i.e. high or low) is displayed in the field below labeled as “2”. The production rules (developed in Chapter 4) are displayed in the field labeled “3” with the title “Diagnosis”.

## 5.5 Results and Discussion

The diagnostic algorithm proposed earlier in this chapter, and the WWTP application developed, along with the WWTP dataset were used to generate results. From the WWTP dataset, input values were selected and used in the WWTP Java application, the known output values (from WWTP dataset) were fed to the software for detection purposes. The production rules for the diagnostic suggestion were used for the development of the fault libraries as the knowledge base of the FDD platform. When

the suspended solids reduction was inhibited due to the poor production of microorganisms, the algorithm was able to detect and eventually provide suggestions for the diagnostics. Figure 5.8 below shows the inputs used for the fault scenario when the suspended solids concentration in effluent eventually increased. Figures 5.9 and 5.10 present the graphical results from the simulation and the diagnostic results after the detection of fault, respectively.

The screenshot shows the 'Input Data for Stages' section of the WWTP Simulation & Diagnostic software. The interface includes several input fields for different stages:

- Influent Feed:** Flowrate (924 m<sup>3</sup>/hr), SS (312 g/m<sup>3</sup>), Nitrate (1 g/m<sup>3</sup>).
- Stage 1 (Screens):** Screening Fraction (0.01), Efficiency (0.99).
- Stage 2 (Grits):** Grit Fraction (0.02), Efficiency (0.99).
- Stage 3 (Primary Settler):** Sludge Fraction (0.05), Efficiency (0.99), % of SS (25).
- Stage 4 (Reactor):**  $\mu_m$  (0.13 hr<sup>-1</sup>),  $k_d$  (0.005 hr<sup>-1</sup>),  $K_s$  (23 g/m<sup>3</sup>),  $K_o$  (0.2 g/m<sup>3</sup>),  $K_n$  (0.5 g/m<sup>3</sup>).
- Stage 4 (Contd.):**  $Y_x$  (0.4 gX/gS),  $Y_o$  (1.42 gX/gO<sub>2</sub>),  $Y_n$  (0.24 gX/gN),  $f_P$  (0.15),  $n$  (0.8).
- Stage 4 (Contd.):**  $V$  (4021 m<sup>3</sup>), O<sub>2</sub> Con. (0.5 g/m<sup>3</sup>).

Figure 5.8 Input of process parameters for simulation

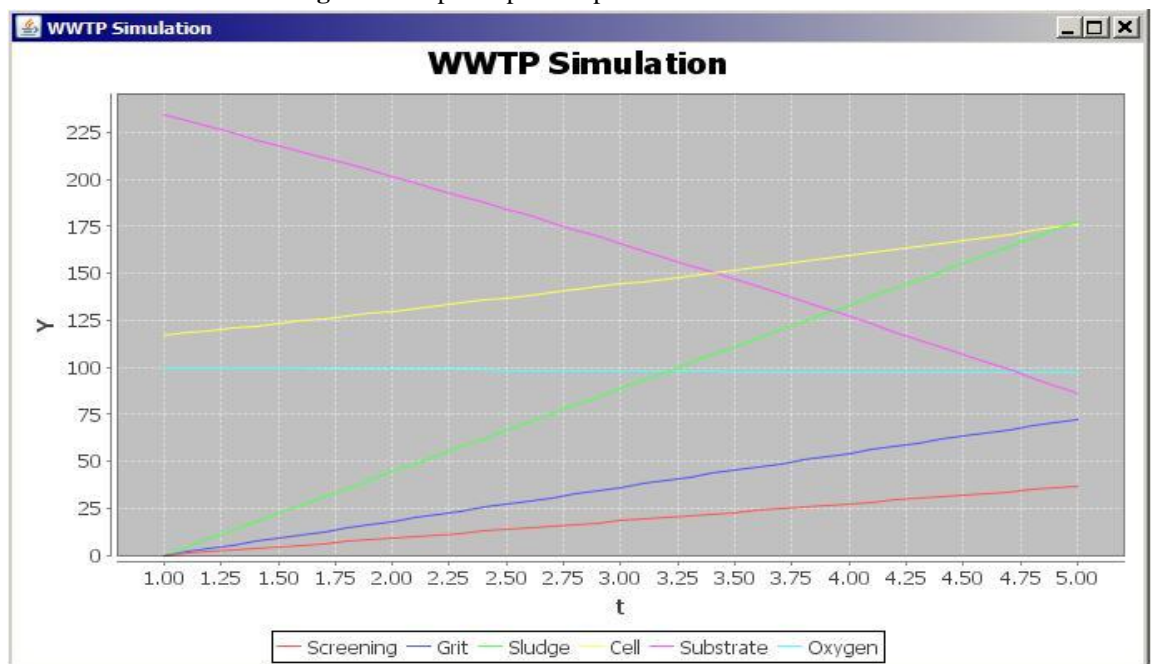
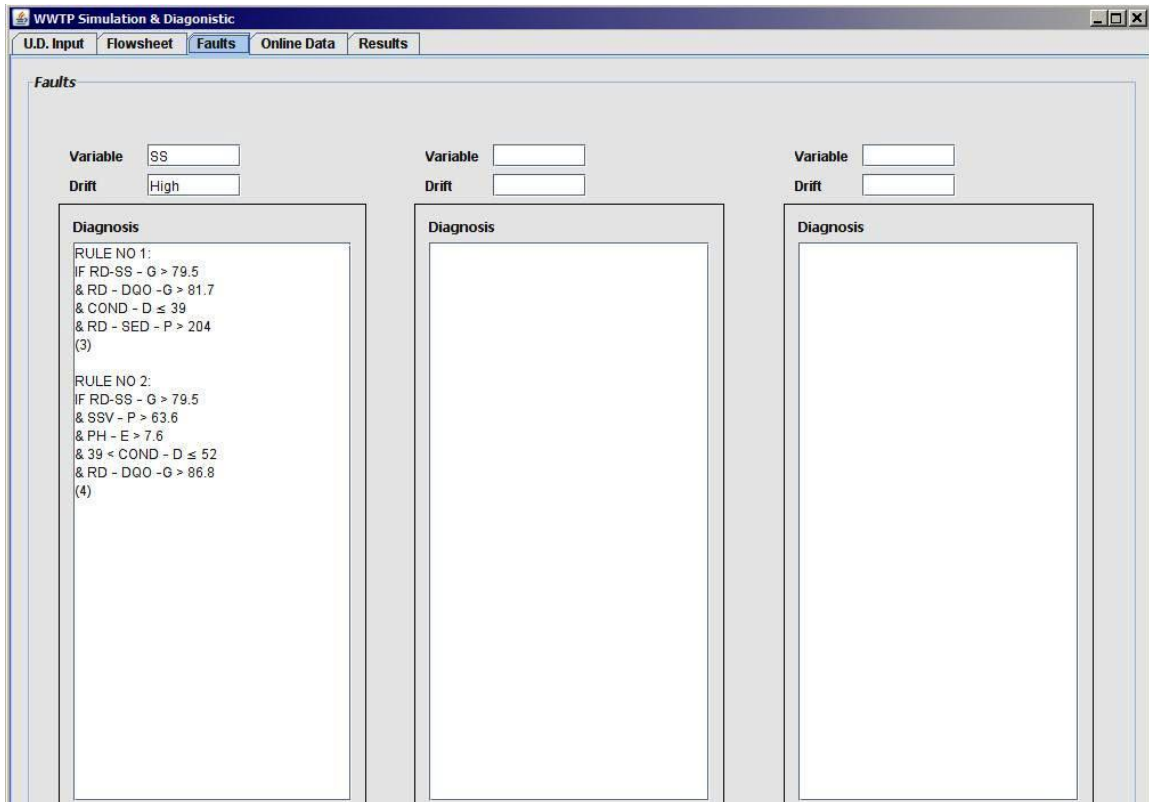


Figure 5.9 Graphical presentations of the simulation results



**Figure 5.10** Diagnostic results for high SS value

Similarly when a low concentration of suspended solids in effluent was detected (although in practice, the low concentration of suspended solids is not a fault, but is considered as a fault in this study), the WWTP software was able to detect and diagnose this fault.

The graphical presentation of the results and the diagnostic suggestions from the fault scenario discussed above are presented in figures 5.11 and 5.12 respectively.

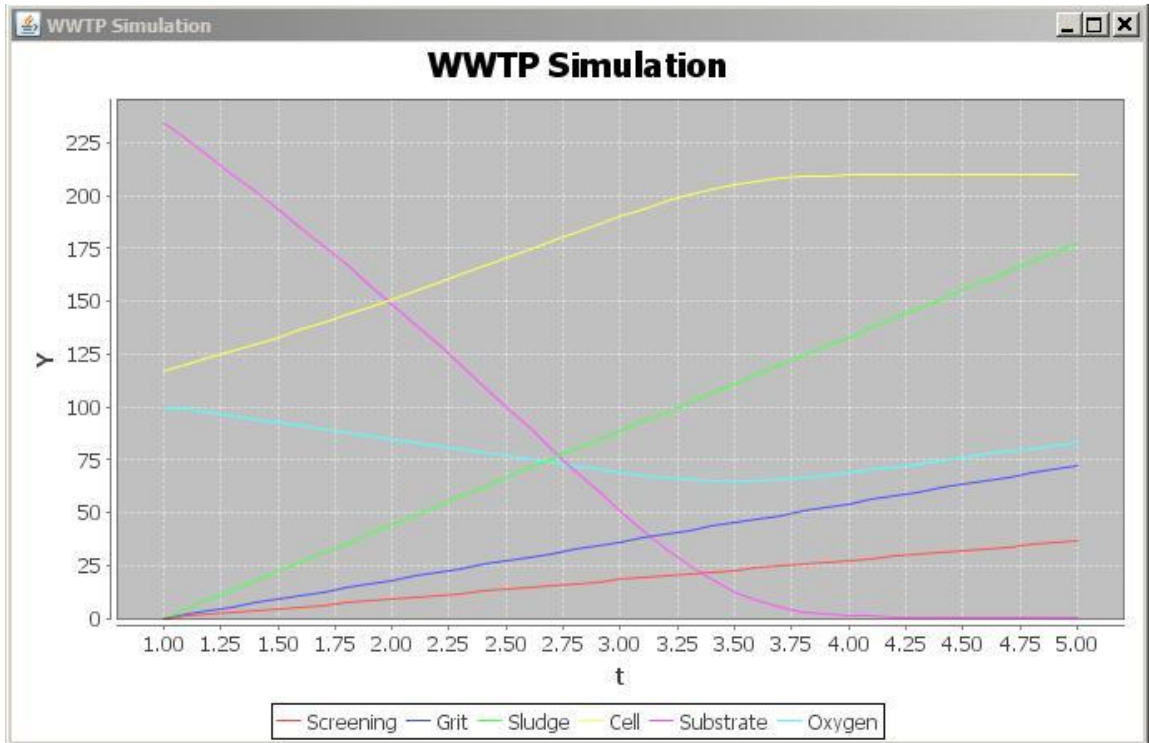


Figure 5.11 Simulation indicating low SS value

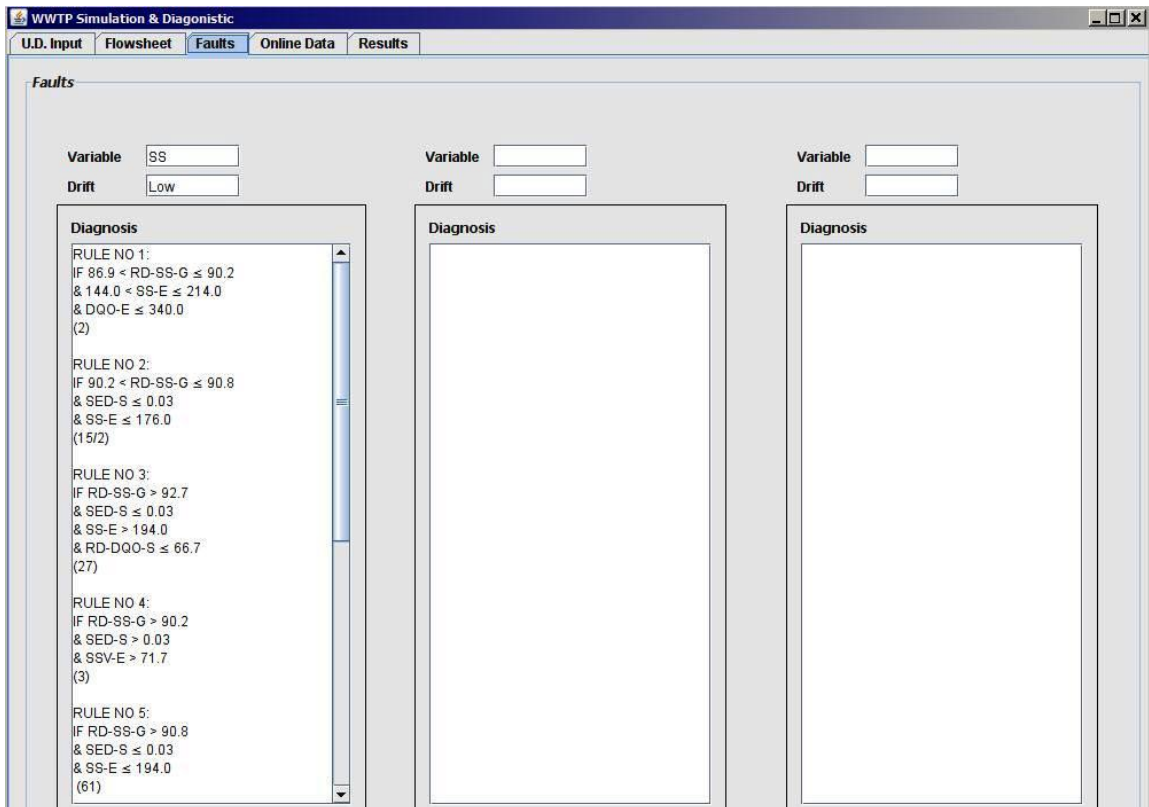


Figure 5.12 Set of diagnostic results for low value of SS

Further scenarios using other output variables were studied and results are reported in Table 5.2.

**Table 5.2** Results of the diagnostic methodology

	<b>Variable</b>	<b>Direction</b>	<b>Detection</b>	<b>Diagnosis</b>	<b>Time(Sec)</b>
<b>Single Fault</b>	SS	+	✓	✓	1.5
	SS	-	✓	✓	1.5
	SSV	+	✓	✓	1.5
	SSV	-	✓	✓	1.5
<b>Disturbance</b>	SS	+	✓	✓	1
	SS	-	✓	✓	1
	SSV	+	✓	✓	1
	SSV	-	✓	✓	1
<b>Multi Fault</b>	SS & SSV	+ & +	✓	No	2
	SS & SSV	- & -	✓	No	2
	SS & SSV	+ & -	✓	No	2
	SS & SSV	- & +	✓	No	2

The results show that the most troublesome scenario for the proposed methodology is if two faults occur simultaneously. It is determined that there is a need to expand the knowledge base obtained from data mining in order to accommodate multiple faults. Recommendations to enhance this methodology are discussed in Chapter 6.

## 5.6 Concluding Remarks

An algorithm based on a new approach is presented in this chapter, and the results obtained confirmed its suitability. A software application is also developed to implement the algorithm and yielded reliable simulation results. Although there remain some challenges to be overcome with the proposed software application (see Chapter 6: Conclusions and Future Work) and also with the proposed algorithm (also



see Chapter 6), this study serves its objective to develop an initial platform for future researchers intending to use different techniques to perform different roles in process diagnostics.

# Chapter 6

## Conclusions and Recommendations

### 6.1 Conclusions

An algorithm is proposed which has been applied to a data set from a wastewater treatment plant. A mathematical model was used for detection, whereas the production rules were employed for the diagnosis, of process faults. The accuracy of detection of faults using the mathematical model was calculated to be 93% (Table No. 3.5) whereas, the production rules exhibited 95% accuracy on the validation step (Table No. 4.10). It was not possible to compare the response time for the detection and diagnosis of faults using this technique with other systems because that data is not available in the literature. A Java application was designed for the implementation of the algorithm and it performed successfully, although some aspects still require improvement.

A new technique of ranking the production rules is proposed and was applied in this research to reduce the large number of production rules and improve the efficiency of the predictions. In this study, the production rules are used to build the knowledge base of the algorithm, and are subsequently used in the diagnosis. The production rules were obtained by using a data-set of WWTP, and an inductive data mining technique using See 5.0 was applied to yield the production rules.

A mathematical model was derived, mainly based upon the ASM1 model, to be used for the detection of faults. The model was validated against the WWTP data-set. An accuracy of 93% on the validation was considered acceptable in this work, but it is expected that calibration of the proposed model would improve the effectiveness of this technique.

One major issue encountered in this research work was the lack of technical input and data sets from any local wastewater treatment facilities. Therefore, a data-set from 1991 was used, which mainly covers the dynamics of suspended solids removal from an activated sludge system.

Despite the lack of extensive technical data, the average response time of 1.5 sec for the proposed software (for the detection of fault and/or disturbance, search of the best possible diagnostic strategy from the fault libraries, and the display of results on screen), 93% model accuracy, 95% production rules accuracy demonstrate the validity and represent the key achievements from this work.

## **6.2 Recommendations for Future Work**

The following areas are identified for continuation of this research work:

- a. Modeling errors are a common problem especially when a model is to be used for process fault detection and diagnosis. It is suggested that if the mathematical model and operational data for a process is available, then genetic algorithm or programming (a technique well established and used for optimization) can be used to fit that model to the operational data thus improving the output (predictions) obtained from the model.
- b. It is expected that the “WWTP-Simulation and Diagnostic” has the potential to be used in the chemical and processing industries. Thus it is recommended to apply this technique on a more complex process where: (a) an elaborated and exhaustive mathematical model of the process is available; and (b) where abundant historical and process data covering almost all aspects and variables of the process are available, in order to test the robustness of this technique.
- c. Detection of multiple faults in a process is a challenging task. It is proposed that if sufficient operational or historical data containing multiple faults scenarios is available, then data mining should be employed for determination of the root causes and the process behavior when multiple faults are present.

## References

“Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged”.

Acharya, J., Sahu, J. N., Mohanty, C. R. & Meikap, B. C. (2009a) Removal of lead(II) from wastewater by activated carbon developed from Tamarind wood by zinc chloride activation. *Chemical Engineering Journal*, vol. 149, no. 1-3, pp. 249-262.

Acharya, J., Sahu, J. N., Sahoo, B. K., Mohanty, C. R. & Meikap, B. C. (2009b) Removal of chromium(VI) from wastewater by activated carbon developed from Tamarind wood activated with zinc chloride. *Chemical Engineering Journal*, vol. 150, no. 1, pp. 25-39.

Aguado, D., Ferrer, A., Ferrer, J. & Seco, A. (2007) Multivariate SPC of a sequencing batch reactor for wastewater treatment. *Chemometrics and Intelligent Laboratory Systems*, vol. 85, no. 1, pp. 82-93.

Albazzaz, H., Wang, X. Z. & Marhoon, F. (2005) Multidimensional visualisation for process historical data analysis: a comparative study with multivariate statistical process control. *Journal of Process Control*, vol. 15, no., pp. 285-294.

Buchanan, J. R. & Seabloom, R. W. (2004) *Aerobic Treatment of Wastewater and Aerobic Treatment Units*. The University of Washington.

Buontempo, F. V. (2005) *Rapid toxicity prediction of organic chemicals using data mining techniques and QSAR based on genetic programming for decision tree generation* Department of Chemical Engineering, University of Leeds, Leeds, PhD.

Charaniya, S., Hu, W. S. & Karypis, G. (2008) Mining bioprocess data: opportunities and challenges. *Trends in Biotechnology*, vol. 26, no. 12, pp. 690-699.

Chen, C. Z. & Mo, C. T. (2004) A method for intelligent fault diagnosis of rotating machinery. *Digital Signal Processing*, vol. 14, no. 3, pp. 203-217.

Chiang, L. H., Russell, E. & Braatz, R. D. (2001) *Fault detection and diagnosis in industrial systems*, Springer, London.

Copp, J. B., Belia, E., Snowling, S. & Schraa, O. (2005) Anaerobic digestion: a new model for plant-wide wastewater treatment process modelling. *Water Science and Technology*, vol. 52, no. 10-11, pp. 1-11.

Dash, S. & Venkatasubramanian, V. (2000) Challenges in the industrial applications of fault diagnostic systems. *Computers and Chemical Engineering*, vol. 24, no. 2-7, pp. 785-791.

Degremont, G. (1991) *Water treatment handbook*, Lavoisier Publishing, Paris.

DeLisle, R. K. & Dixon, S. L. (2004) Induction of decision trees via evolutionary programming. *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 3, pp. 862-870.

Dellana, S. A. & West, D. (2009) Predictive modeling for wastewater applications: Linear and nonlinear approaches. *Environmental Modelling & Software*, vol. 24, no. 1, pp. 96-106.

Fayyed, U. M., Shapiro, G., Smith, P. & Uthursuamy, R. (1996) *Advances in Knowledge Discovery and Data Mining*, MIT Press.

Fears, N. (2008) What Is Netbeans and What Can It Do For You?  
<http://www.brighthub.com/internet/web-development/articles/8472.aspx#ixzz0fHVxjHBo>

Forster, C. F. (2003) *Wastewater treatment and technology*, Thomas Telford, London.

Gernaey, K. V., van Loosdrecht, M. C. M., Henze, M., Lind, M. & Jorgensen, S. B. (2004) Activated sludge wastewater treatment plant modelling and simulation: state of the art. *Environmental Modelling & Software*, vol. 19, no. 9, pp. 763-783.

Green-Ideas (2009) Green Ideas, Inc. <http://www.egreenideas.com/glossary.php?group=b>

Gujer, W., Henze, M., Mino, T. & van Loosdrecht, M. (1999) Activated Sludge Model No. 3. *Water Science and Technology*, vol. 39, no. 1, pp. 183-193.

Han, J. W., Pei, J., Yin, Y. W. & Mao, R. Y. (2004) Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, vol. 8, no. 1, pp. 53-87.

Himmelblau, D. M. (1978) *Fault detection and diagnosis in chemical and petrochemical processes*, Elsevier Scientific Publishing , Amsterdam.

Henze, M., Gujer, W., Mino, T. & van Loosdrecht, M. (2000) *Activated sludge models : ASM1, ASM2, ASM2d and ASM3*, IWA Publishing, London.

Huang, M. Z., Wan, J. Q., Ma, Y. W., Wang, Y., Li, W. J. & Sun, X. F. (2009) Control rules of aeration in a submerged biofilm wastewater treatment process using fuzzy neural networks. *Expert Systems with Applications*, vol. 36, no. 7, pp. 10428-10437.

Hunag, Y. C. & Wang, X. Z. (1999) Application of fuzzy casual networks to wastewater treatment plants. *Chemical Engineering Science*, vol. 54, pp. 2731-2738.

Iri, M., Aoki, K., O'Shima & Matsuyama, H. (1979) An algorithm for diagnosis of system failures in the chemical processes. *Computers and Chemical Engineering*, vol. 3, no. 1-4, pp. 489-493.

Jain, A. K., Murty, M. N. & Flynn, P. J. (1999) Data clustering: A review. *Association for Computing Machinery Computing Surveys*, vol. 31, no. 3, pp. 264-323.

Jeppsson, U. (2003) A General Description of the IAWQ Activated Sludge Model No. 1. Lund Institute of Technology, Lund, Sweden.

Khalid, M. I. (2005) *Automatic Generation of Decision Trees from Process Historical Data using Genetic Programming* Institute of Particle Science and Engineering, The University of Leeds, Leeds, MSc (Eng).

Kim, S., Lee, H., Kim, J., Kim, C., Ko, J., Woo, H. & Kim, S. (2002) Genetic algorithms for the application of Activated Sludge Model No. 1. *Water Science and Technology*, vol. 45, no. 4-5, pp. 405-411.

Krogh, A. (2008) What are artificial neural networks? *Natural Biotechnology*, vol. 26, no. 2, pp. 195-197.

Kumar, A. (2004) State-of-the-art report on Water Management, Documentation and Research training centre, Indian Statistical Institute, Bangalore.

Lapp, S. A. & Powers, G. A. (1977) Computer-aided synthesis of fault trees. *IEEE Transactions and Reliability*, vol. 37, no., pp. 2-13.

Leahy, P. (2010) What is Java? <http://java.about.com/od/gettingstarted/a/whatisjava.htm>

Lee, C., Choi, S. W. & Lee, I. B. (2006) Sensor fault diagnosis in a wastewater treatment process. *Water Science and Technology*, vol. 53, no. 1, pp. 251-257.

Lee, D. S., Vanrolleghem, P. A. & Park, J. M. (2005) Parallel hybrid modeling methods for a full-scale cokes wastewater treatment plant. *Journal of Biotechnology*, vol. 115, no. 3, pp. 317-328.

Li, R. F. (2003) *Advanced process monitoring & control using principal & independent component analysis* Department of Chemical Engineering, University of Leeds, Leeds, PhD.

Lou, X. C., Willsky, A. S. & Verghese, G. C. (1986) Optimally robust redundancy relations for failure detection in uncertain systems. *Automatica*, vol. 22, no. 3, pp. 333-344.

Ma, C. Y. & Wang, X. Z. (2009) Inductive data mining based on genetic programming: Automatic generation of decision trees from data for process historical data analysis. *Computers & Chemical Engineering*, vol. 33, no. 10, pp. 1602-1616.

Michle, B. (1988) Detecting changes in signals and systems—a survey. *Automatica*, vol. 24, no. 3, pp. 309-326.

Morley, D. A. (1979) *Mathematical modelling in water and wastewater treatment*, Applied Science, London.

NetBeans (2010) NetBeans IDE - Get Tomorrow Today <http://netbeans.org/features/index.html>

Norweco (2006) Disinfection of Water and Wastewater  
<http://www.norweco.com/html/lab/Disinfection.htm>

Pai, T. Y. (2008) Gray and neural network prediction of effluent from the wastewater treatment plant of industrial park using influent quality. *Environmental Engineering Science*, vol. 25, no. 5, pp. 757-766.

Palmer, G. (2003) *Technical Java : developing scientific and engineering applications*, Prentice Hall, Upper Saddle River, NJ.

Puteh, M., Minekawa, K., Hashimoto, N. & Kawase, Y. (1999) Modeling of activated sludge wastewater treatment processes. *Bioprocess Engineering*, vol. 21, no. 3, pp. 249-254.

Quinlan, J. R. (1986) Induction of decision trees. *Machine Learning*, vol. 1, pp. 81-106.

Quinlan, J. R. (1990) Decision trees and decision-making. *IEEE transactions on systems, man, and cybernetics*, vol. 20, no. 2, pp. 339.



Quinlan, J. R. (1993) *C4.5 : programs for machine learning*, Morgan Kaufmann, Calif.

Quinlan, J. R. (1996) Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, vol. 4, no., pp. 77-90.

R-Foundation (2008) The R Project for Statistical Computing <http://www.r-project.org/>

Radcliffe, J. C. (2004) *Water Recycling in Australia*, The Australian Academy of Technological Sciences and Engineering, Melbourne.

Rieger, L., Alex, J., Gujer, W. & Siegrist, H. (2006) Modelling of aeration systems at wastewater treatment plants. *Water Science and Technology*, vol. 53, no. 4-5, pp. 439-447.

Sanchez , M., Cortes, U., Bejar, J., Gracia, J. D., Lafuente , J. & Poch, M. (1997) Concept formation in WWTP by means of classification techniques:A compared study. *Applied Intelligence*, vol. 7, no., pp. 147-165.

Sanguesa, R. & Cortes, U. (1997) Learning causal networks from data:a survey and a new algorithm for recovering possibilistic causal networks. *A.I. Communications*.

See5.0 (2008) See 5.0. 2.06 ed. RuleQuest Research Pty Ltd, NSW, Australia.

Shuler, M. L. & Kargi, F. (2002) *Bioprocess engineering*, Prentice Hall, NJ.

Simani, S., Fantuzzi, C. & Patten, R. J. (2003) *Model based fault diagnosis in dynamic systems using identification techniques*, Springer, NY.

Sun, W. X., Chen, J. & Li, J. Q. (2007) Decision tree and PCA-based fault diagnosis of rotating machinery. *Mechanical Systems and Signal Processing*, vol. 21, no. 3, pp. 1300-1317.

Tchobanoglous, G., Burton, F. L., Stensel, H. D. & Metcalf & Eddy. (2003) *Wastewater engineering : treatment and reuse*, McGraw-Hill, Boston.

TIBCO-Software-Inc (2008) TIBCO:The Power of Now\_ <http://www.insightful.com/>

U.S-EPA (2006) Lake Erie <http://www.epa.gov/lakeerie/glossary.html>

Van Hulle, S. W. H. & Vanrolleghem, P. A. (2004) Modelling and optimisation of a chemical industry wastewater treatment plant subjected to varying production schedules. *Journal of Chemical Technology and Biotechnology*, vol. 79, no. 10, pp. 1084-1091.

Venkatasubramanian, V., Rengaswamy, R. & Kavuri, S. N. (2003a) A review of process fault detection and diagnosis: Part I: Quantitative model based methods. *Computers and Chemical Engineering*, vol. 27, no. 3, pp. 293-311.

Venkatasubramanian, V., Rengaswamy, R. & Kavuri, S. N. (2003b) A review of process fault detection and diagnosis: Part II: Qualitative models and search strategies. *Computers and Chemical Engineering*, vol. 27, no. 3, pp. 313-326.

Venkatasubramanian, V., Rengaswamy, R. & Kavuri, S. N. (2003c) A review of process fault detection and diagnosis: Part III: Process history based methods. *Computers and Chemical Engineering*, vol. 27, no. 3, pp. 327-346.

Wang, X. Z. (1999) *Data mining and knowledge discovery for process monitoring and control*, Springer-Verlag, London.

Wang, X. Z., Medasani, S., Marhoon, F. & Albazzaz, H. (2004) Multidimensional visualization of principal component scores for process historical data analysis. *Industrial & Engineering Chemistry Research*, vol. 43, no., pp. 7036-7048.

West, D. & Mangiameli, P. (2000) Identifying process conditions in an urban wastewater treatment plant. *International Journal of Operations & Production Management*, vol. 20, no. 5-6, pp. 573-590.

Wimberger, D. & Verde, C. (2008) Fault diagnosticability for an aerobic batch wastewater treatment process. *Control Engineering Practice*, vol. 16, no. 11, pp. 1344-1353.

Wintgens, T., Rosen, J., Melin, T., Brepols, C., Drensla, K. & Engelhardt, N. (2003) Modelling of a membrane bioreactor system for municipal wastewater treatment. *Journal of Membrane Science*, vol. 216, no. 1-2, pp. 55-65.

Yang, B. S., Lim, D. S. & Tan, A. C. C. (2005) VIBEX: an expert system for vibration fault diagnosis of rotating machinery using decision tree and decision table. *Expert Systems with Applications*, vol. 28, no. 4, pp. 735-742.

## Appendix

### Data Statistics:-

No.	Attrib.	Min	Max	Mean	St-dev
1	Q-E	10000	60081	37226.56	6571.46
2	ZN-E	0.1	33.5	2.36	2.74
3	PH-E	6.9	8.7	7.81	0.24
4	DBO-E	31	438	188.71	60.69
5	DQO-E	81	941	406.89	119.67
6	SS-E	98	2008	227.44	135.81
7	SSV-E	13.2	85.0	61.39	12.28
8	SED-E	0.4	36	4.59	2.67
9	COND-E	651	3230	1478.62	394.89
10	PH-P	7.3	8.5	7.83	0.22
11	DBO-P	32	517	206.20	71.92
12	SS-P	104	1692	253.95	147.45
13	SSV-P	7.1	93.5	60.37	12.26
14	SED-P	1.0	46.0	5.03	3.27
15	COND-P	646	3170	1496.03	402.58
16	PH-D	7.1	8.4	7.81	0.19
17	DBO-D	26	285	122.34	36.02
18	DQO-D	80	511	274.04	73.48
19	SS-D	49	244	94.22	23.94
20	SSV-D	20.2	100	72.96	10.34
21	SED-D	0.0	3.5	0.41	0.37
22	COND-D	85	3690	1490.56	399.99
23	PH-S	7.0	9.7	7.70	0.18
24	DBO-S	3	320	19.98	17.20
25	DQO-S	9	350	87.29	38.35
26	SS-S	6	238	22.23	16.25
27	SSV-S	29.2	100	80.15	9.00
28	SED-S	0.0	3.5	0.03	0.19
29	COND-S	683	3950	1494.81	387.53
30	RD-DBO-P	0.6	79.1	39.08	13.89
31	RD-SS-P	5.3	96.1	58.51	12.75
32	RD-SED-P	7.7	100	90.55	8.71
33	RD-DBO-S	8.2	94.7	83.44	8.4
34	RD-DQO-S	1.4	96.8	67.67	11.61
35	RD-DBO-G	19.6	97	89.01	6.78
36	RD-DQO-G	19.2	98.1	77.85	8.67
37	RD-SS-G	10.3	99.4	88.96	8.15
38	RD-SED-G	36.4	100	99.08	4.32

### WWTP Data:-

The WWTP data used in this research work is included here.





















D-29/8/91	32190	0.3	7.3	200	545	258	65	4	1260	7.4	191	226	67	3.5	1198	7.5	115	244	77	77	0.1	1351	7.7	21	71	27	71	0	1326	40	66	97	82	71	90	87	90	100
D-30/8/91	30488	0.2	7.5	152	300	132	70	4.5	1073	7.4	150	210	60	4.5	1081	7.4	93	233	64	84	0.3	1188	7.3	17	55	18	80	0	1224	40	70	93	82	76	90	82	86	100