# APPLICATION OF DATA MINING TECHNIQUES FOR INDUSTRIAL PROCESS OPTIMIZATION

Prepared by:

CANMET Energy Technology Centre - Varennes

*– A Literature and Software Review –*

Author names

Radu Platon & Mouloud Amazouz

*We are drowning in information, but starving for knowledge*
(John Naisbitt)

July 2007

# DISCLAMER

This report is distributed for informational purposes and does not necessarily reflect the views of the Government of Canada nor constitute an endorsement of any commercial product or person. Neither Canada nor its ministers, officers, employees or agents makes any warranty in respect to this report or assumes any liability arising out of this report.

# TABLE OF CONTENT

# 1. INTRODUCTION

Data mining can be defined as the science of extracting useful information from large data sets or databases. Data mining is used for building empirical models, which are based not on the underlying theory about the process or mechanism that generated the data. Data mining, as the name suggests it, is data-driven, and it provides a description of the observed data. Its fundamental objective is to provide insight and understanding about the structure of the data and its important features, and to discover and extract patterns contained in the data set. This discipline also referred to as knowledge discovery in databases (KDD), is a process of extracting implicit, previously unknown, and potentially useful information from data (Shapiro et al, 1992). Data mining brings together a multitude of disciplines, such as database systems, statistics, artificial intelligence, data visualization, and others.

The discovered knowledge can be applied to information management, query processing, decision-making, process control, and many other applications. Mining information and knowledge from commercial or industrial data, and applying this information to new and innovative uses has been recognized by companies as an important area for generating revenues and increasing business opportunities.

The availability of high volume data in the industrial sector, have given rise to a new level of interest in the applications of knowledge discovery and data mining for industrial applications.

Data mining products are commercially available and numerous industrial applications are being developed.

The main objective of the report is to provide an overview of data mining methods suitable for industrial process applications, by examining the publications related to data mining applications that are being developed or that are already implemented in the industry. A survey of this published work provides not only current trends, but also a better understanding of the different applications required by the industry, and the data mining methods used for these applications.

In the context of the industrial applications, the scope of this paper covers industrial processes such as pulp & paper, and petrochemical operations, with applications geared mainly towards process monitoring and control, process parameter value inference (soft sensors), detection of abnormal situations (faults) and their diagnostic and a general improvement of the process understanding through discovery of correlations between process parameters.

A general overview of the main tasks performed by data mining and methods used to achieve these tasks is provided.

A summary of the data pre-processing steps performed for a data set containing historical process data of Smurfit Stone's pulp & paper plant at La Tuque (Québec) is presented.

Commercial software packages used for developing and implementing these applications were also examined. This proved to be useful, since most companies provide information about industrial applications of their respective software packages.

# 2. APPLICATION OF DATA MINING IN THE MANUFACTURING PROCESSES

In many industrial plants, large amounts of process data are collected by data historians. Hundreds of variables are continuously measured, and their values are being stored in voluminous databases.

These databases are a rich source of information, but without the use of data mining algorithms, extracting useful information and creating knowledge from a large number of process variables and different operating modes can be a very difficult, if not impossible task

Data mining techniques can be used, for example, to identify patterns in the data, correlations between different process variables, classify or cluster variables in order to ultimately discover new relationships between the variables contained in the database, and thus extract previously unknown process knowledge.

Data mining algorithms are well developed and many industries recognized the benefits of this technology. Data mining is used effectively in different areas, such as retail, marketing, banking, insurance and medicine, amongst others. For example, applying data mining methods to detect patterns of fraudulent credit card usage, identify "loyal" customers, or determine credit card spending by customer groups became common practice for financial institutions.

In the industrial process sector, data mining can be used to unlock useful knowledge from the databases storing large volumes of historical process data; this knowledge can then be used to improve process operation, product quality and energy consumption.

However, there are relatively few applications in the industry, mainly due to the lack of tools, expertise and staff time to properly mine (analyze) the data. This way, the available historical process data, which represents a huge repository of operating plant experience, often obtained at significant expenses, becomes an under-utilized resource.

This indicates that there is a need, in the industrial process sector, to develop a structured framework for implementing data mining applications. This framework will provide users with a simplified, but proven methodology for applying data mining tasks, with less effort and without knowing all the theoretical and statistical details of the algorithms.

Automated tools for analyzing the data, generating results and graphically displaying these results for a fast and more convenient interpretation can assist operators and engineers in decision making regarding plant operation, ultimately leading to process improvements.

# 3. DATA PRE-PROCESSING

The successful application of data mining is highly dependent on the quality of the data, since for quality mining results, quality data is needed. Data mining techniques are very susceptible to the adage "garbage-in/garbage-out".

Therefore, the steps preceding the data mining are essential to ensure that useful knowledge is derived from the data. These steps are referred as data pre-processing, and their main objective is to increase the data quality before the data mining algorithms can be successfully applied.

Typically, up to 90% of the time and effort in a data mining project is spent on data understanding and pre-processing1.

An overview of the knowledge discovery from data steps is illustrated below:

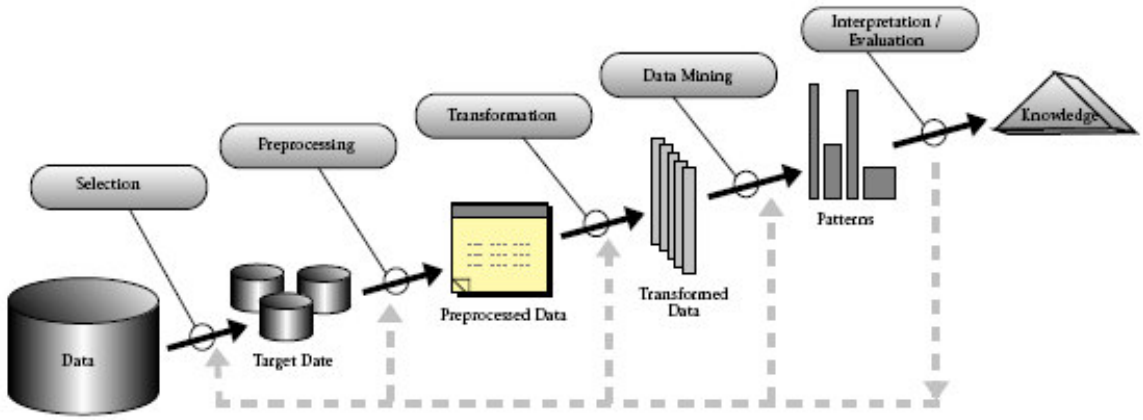


**Fig 1. An overview of the steps that compose the KDD process 2**

In the industrial process context, data quality is a key issue, since historical process data include numerous invalid data, such as:

- incomplete data: missing process variables, missing variable values
- inconsistent data: impossible or out-of-range values
- noisy data: errors, outliers, inaccurate values

[1] CRoss Industry Standard Process for Data Mining, http://www.crisp-dm.org
[2] Fayyad U., Piatetsky-Shapiro G. and Smith P., *From Data Mining to Knowledge Discovery in Databases*, 1996 AI Magazine, vol. 17, pp. 37-54

This can be caused by various process or equipment related factors, such as:

- drift and malfunction of measuring instruments
- starts and stops of key unit operations
- aberrant process behaviour
- relative infrequent laboratory analyses or product quality sampling
- dubious periods of operation, such as shutdowns, low production periods, equipment cleaning

The major data pre-processing steps can be classified as follows:

- Data cleaning, dealing mainly with filling in missing values, smoothing noisy data, identifying/removing outliers and resolving inconsistencies
- Data integration, dealing with combining multiple data sources
- Data transformation, dealing with converting data into forms suitable for data mining
- Data reduction, dealing with obtaining a reduced data set, without a significant loss of information from the raw data set
- Data reconciliation, dealing with the integrity or accuracy of individual records of data

A more detailed presentation of data pre-processing steps, as well as a description of data pre-processing performed for the Smurfit Stone project can be found in the *Data pre-processing techniques for multivariate data analysis* report.

# 4. DATA MINING CONCEPTS AND APPROACHES: A GENERAL OVERVIEW

Data mining techniques can be classified according to different criteria: based on the type of database to be mined, the type of knowledge to be discovered, and the types of methods to be used.

In industrial process application, data mining is used in areas such as process control, process-related decision-making support, process parameter inference, fault detection and diagnosis, amongst others. These applications will be review further on in this report. Each application requires a different technique and different types of databases, or a combination of various techniques in order to achieve the desired objective.

A classification scheme based on the type of knowledge to be discovered presents an accurate indication of the different requirements and techniques necessary (Chen et al, 1996). It also matches the requirements of the variety of industrial applications dealing with different kinds of information that needs to be extracted and understood. Therefore, the following overview of data mining techniques will be organized based on the type of knowledge to be extracted.

From this perspective, the two main tasks that data mining performs are descriptive modeling and predictive modeling. A descriptive model represents, in a convenient way, the main features of the data; it summarizes the information contained in the data set and describes its main characteristics. A predictive model is a model of future behavior, allowing the value of some variables to be predicted from known values of other variables.

## 4.1 Descriptive models

As mentioned previously, a descriptive model represents the main features of the data. It provides insights and understanding about the structure of the data, describing its important features and relationships between variables. A descriptive model can also allow the discovery of certain objects possessing similar properties, and the discovery of pattern or rules that apply only to certain subsets of the data, and not to the whole data space.

In the context of data mining, the terms model and pattern are used to describe the data modeling and pattern extraction performed by specific data mining techniques. Since model and pattern are two different concepts that use different methods, it is important to distinguish between them. A model is a global concept, providing a full description of the data that applies to all points in the measurement space (database), while a pattern is a local description of some subset of the data, holding for some variables, but not for all of them.

Pattern extraction allows the discovery of unusual structures within the data, revealing persistent deviations from the general run of the data. Patterns are very useful for both descriptive and predictive reasons, enabling, for example, the description of variables having unusual characteristics, the detection of complicated relationships between variables, or the prediction of future variables having unusual properties.

Different approaches for constructing global descriptive models are available:

- methods based on probability distributions and densities, which describe data in terms of their underlying distribution or density function
- cluster analysis methods, which decompose the data set into groups (clusters) such that the points in one group are similar to each other but different from points belonging to other clusters
- statistical techniques, indicating correlations between input variables

For finding useful patterns and rules from large data sets, some of the methods for descriptive pattern modeling are:

- association rules, for finding all rules of a certain type that fulfill some frequency and accuracy criteria; for example, given a frequency and accuracy threshold, all patterns satisfying this threshold, as well as their frequency, are identified
- selective discovery of patterns and rules, where a criteria for *interestingness* of a rule is introduced, and thus identifying only the *interesting* patterns
- outlier detection, where abnormal data behavior (data deviating from the natural variability of the data set) is identified. The cut-off value or threshold which divides anomalous and non-anomalous data numerically is often the basis for important decisions.

The techniques used for constructive descriptive models include:

- neural networks – such as the Kohonen self-organizing map for clustering, for example
- statistical factor analysis – such as Principal Component Analysis (PCA) for system modeling, outlier detection
- fuzzy logic – for discovering association rules, for clustering

## 4.2 Predictive models

As mentioned previously, a predictive model has the specific objective of predicting the value or characteristic of some variables on the basis of observed values or characteristics of other variables in the database. A set of available data, called the training data set, is used to construct the model. The training set contains both the variables to be predicted – the model output – and the variables that will be used for prediction – the model input. By learning the relationships between the input and output values, the model estimates a mapping, or a function, that can predict the output value given the input vector. When presented to new, previously unseen input data, this model can forecast the future output values corresponding to this input.

The predictive models can be further classified in two distinct categories, depending on whether the variable to be predicted is categorical or real-valued.

### *4.2.1 Classification models*

Categorical (qualitative variables with no numerical significance) variables are estimated by classification models, which predict the likely class (category) membership of a new variable, based on a series of other known variables. This method is also called supervised classification, since the classes in which the output has to be correctly placed are pre-defined in the training data set.

For example, such a model can use different available process measurements and provide a diagnostic for a process fault by predicting the correct class in which the fault belongs.

Some of the main classification techniques used are presented below:

- linear discriminates, linear classifiers based on the concept of searching for the linear combination of the variables that best separates the classes
- nearest neighbour method, which searches the objects in the training data that are most similar to the object, in terms of input variables, and then classifies the new object into the most heavily represented class among these similar object
- tree models, which partition, in a recursive manner, the input space such that the majority of points in a partition should belong to the same class
- neural networks, such as feed-forward and radial basis function neural networks

### 4.2.2 Predictive models for regression

Predictive models for regression estimate real-valued variables. In these models, the variable to be predicted has a numerical significance. The variables used for prediction are called predictor variables, and they do not need to be numerical. The variable to be predicted is called the response variable. This type of model is essentially similar to the classification model, the only difference being the numerical instead of categorical nature of the model output (the response variable).

Some of the main methods used for predictive regression models are presented below:

- linear models and least squares fitting, where the output is predicted as a linear combination of the inputs
- neural networks, mainly feed-forward neural networks
- Partial Least Squares (PLS, also known as Projection to Latent Structures)

# 5. DATA MINING APPROACHES FOR INDUSTRIAL PROCESS APPLICATIONS

A multitude of methods are available for achieving different data mining tasks. However, in the industrial process context that this report overviews, the main data mining applications are in the areas of process monitoring and control, soft sensors, expert systems for decision-making support and fault detection and diagnosis systems. Some examples of data mining techniques used for these applications are as follows:

- regression techniques used for the development of soft sensors, where a process parameter can be inferred from other available (measured) variables
- classification techniques used to predict whether the final product will be within a certain specification, or to classify faults into the correct classes
- descriptive modelling techniques, such as PCA, used to identify correlations between variables, to further process understanding and for statistical process control purposes

Certain data mining techniques are extensively used for developing these industrial applications:

- factor based statistical models, such as principal component analysis (PCA) and projection to latent structures (PLS)
- neural network models
- fuzzy logic concepts

As it will be shown in the literature review section of this report, these applications along with the techniques used, are the most covered areas in the published literature.

Some techniques can perform more than one task: for example, feed-forward neural networks can be used for both classification and regression purposes. Some tasks can be performed by more than one approach: for example, regressions can be done either with neural networks or with PLS. Some applications are using a combination of different techniques, in order to take advantage of each method's strengths. The literary survey presented in this paper will try to determine the strengths and drawbacks of these methods with respect to the tasks that they can perform.

Before starting with the literary review, the above mentioned techniques will be briefly presented.

## 5.1 Factor Based Statistical Models (PCA/PLS)

Factor based statistical models which are used for data mining applications are principal components analysis (PCA) and partial least squares also known as projection to latent structures (PLS). They are used to transform a number of related process variables to a smaller set of uncorrelated variable. These techniques are often called data reduction methodologies since they identify factors that have a much lower dimension than the original data set and still can properly describe the major trends in the original data set. Often 100 input variables can be reduced to 5 or 6 new variables explaining between 40 to 80% of the process variability of industrial data sets.

Principal component analysis models typically provide a better understanding of the process, indicating correlations between input variables, and input and output variables respectively.

PCA/PLS models are able to cope with the following issues of industrial data:

- Noisy data sets
- Missing data in the data sets
- Correlated variables within the data sets
- Large data sets, both observations and variables
- Data sets with many variables and a small number of observations
- Data sets with many observations and a small number of variables

Typical industrial data mining applications include:

- multivariate statistical process control charts to detect abnormal situations and help locate the cause of the problem
- classifying products
- predicting process parameter values, product characteristics or specifications

## 5.2 Neural Networks

Black box models are typically based on neural net and/or fuzzy logic techniques.

Artificial Neural Networks (ANN) is a type of Artificial Intelligence technique that mimics the behaviour of the human brain; it attempts to describe a nonlinear relationship between the input and output of a complex system using historic process data. Similar to a human neural system, an artificial neural network is an information processing structure that consists of a number of input units and output units connected in a systematic fashion. Between the input and output units, there may be one or more hidden layers, each consisting of a number of units called neurons, nodes or cells. The connections between units lying on different layers are assigned with varying weights. Input signals (or data sets) are fed in from the input layer, and they follow all possible connection paths to reach the next layer.

Along each connection link, the signal suffers a transformation, eventually reaching the output layer. Through this mechanism an ANN learns to identify patterns in the data set and predict variables. Neural net models prefer data sets whose inputs are independent.

They are able to cope with the following issues of industrial data:

- Noisy data sets
- Missing data in the data sets
- Large data sets, both observations and variables
- Data sets with many observations and a small number of variables

Typical industrial data mining applications include:

- detecting abnormal situations and locating the cause of the problem
- classifying products
- predicting process parameter values product characteristics and specifications

## 5.3 Fuzzy logic

Fuzzy logic is an Artificial Intelligence technique that mimics the human reasoning by using gradients of true and false. A fuzzy system can convert a set of user-supplied human language rules into mathematical equivalents. Fuzzy logic algorithms can discover rules and associations in a data set using both linguistic (qualitative) and numerical data. Fuzzy logic modelling is a probability based modeling, and it allows numerical data to be converted into corresponding qualitative attributes. It is mainly used for rule extraction and clustering, and it can generate rules of association can be either linguistic or numerical. It can also use historic process to create a set of rules that are use to estimate the unknown variables. Given an input-output data set of historical values, fuzzy logic concepts can be used to group the data into clusters of similar behaviour, and then evaluate the behaviour of the process from these clusters. Each cluster is represented by a fuzzy if – then rule which characterizes the system behaviour in the vicinity of that cluster.

The fuzzy logic techniques are able to cope with same issues of industrial data as the neural networks.

# 6. DATA MINING APPLICATIONS FOR INDUSTRIAL PROCESS: A LITERARY REVIEW

The publications reviewed and analyzed in this report described data mining industrial applications for industrial processes such as pulp and paper, and petrochemical sectors. Roughly 25 papers were reviewed. They were published in recent issues of several well-established publications, industrial journals and presentations from different international conferences such as:

- Computers & Chemical Energy
- Control Engineering Practice
- Control Engineering
- Process Control
- Energy
- Pulp and Paper Canada
- Various international conferences on data mining related topics

Material published from universities – such as PhD dissertations – was also reviewed and analyzed in this survey.

The number of reviewed publications was considered to be sufficient, since the material covered in these publications is quite representative of the main applications of data mining in the industrial process sector.

## 6.1 Literature review

In *A review of process fault detection and diagnosis, Part III: Process history based methods* (Venkatasubramanian et al, 2003), a systematic and comparative study of various data-driven process diagnostic methods is presented. The main methods used for extraction information from historical process data are stated to be the following:

- non-statistical: neural networks are an important class of non-statistical classifiers
- statistical: principal component analysis (PCA)/partial least squares (PLS), and statistical pattern classifiers

The application of statistical methods, such as PCA, and artificial intelligence methods, such as neural networks, was examined in this paper. The advantages and drawbacks of each method were analysed:

| Advantages & disadvantages of process history based fault detection & diagnosis methods | | |
|---|---|---|
| **Technique** | **Advantages** | **Drawbacks** |
| **PCA** | - good at revealing anomalies<br>- do not need an explicit system model<br>- handle large data sets, with a large number of dimensions | - time-invariant (a recursively updated PCA better suited to deal with time variance)<br>- lack the ability to classify anomalies<br>- better suited for linear processes |
| **Neural Networks** | - robust to noise and non-linearities and noise<br>- novelty identifiability (for NN that generate bounded decision regions)<br>- classification ability<br>- easy to implement, requiring little process a priori knowledge | - lack of explanation and adaptability abilities<br>- poor generalization capability outside the training data<br>- difficulty with multiple faults |

Important information presented in this publication concerns the industrial application of history-based methods:

- the most fault diagnostic applications in process industries are based on these approaches. This is due to the fact that process history based approaches are easy to implement, requiring very little modeling effort and a priori knowledge

- further, even for processes for which models are available, the models are usually steady-state models. It would require considerable effort to develop dynamic models specialized towards fault diagnosis applications.

- statistical approaches that are easy to build and which do very well on fast detection of abnormal situations have been successful in industrial applications

These advantages of process history based methods for constructing a fault and diagnosis system can be generalized for other data mining applications, since the objective is the same, knowledge discovery from process data, and the data mining techniques for extracting this knowledge are the same. The acceptance and implementation of these methods in the process industry confirms that data-mining approaches are likely to be successfully adopted by the industry.

The authors recommend that one of the main directions for future work be in the field of integrating different data mining techniques in the same system, since these methods can complement each other, resulting in better data mining applications.

In *The application of neural networks to the paper-making industry* (Edwards et al, 1999), the application of neural network (NN) and PCA techniques for the prediction of paper curl is presented. Paper curl is defined as the tendency of paper to deviate from a flat form. This parameter is measured off-line, after the production of a complete roll is completed. An out-of-specification curl level results in re-pulping that paper roll, thus increasing production costs and time. Prediction the paper curl during the production leads to significant savings.

Feed-forward NNs were used to predict the level of the curl – a regression task – and to predict if the curl will be within a required specification – "in-specification", or "out-specification", a classification task. Various process parameters measured during the manufacture of a reel of paper were used to train the networks. Prior to NN training, they were processed using a variation of the PCA technique, the Karhunen-Loève transform, for reducing its dimensionality and removing any correlation between parameters. Some of the principal components obtained by applying this transformation were selected and used as network inputs. The neural networks presented good prediction and classification results for the paper curl level.

The work presented in this paper also shows that a combination of different data mining techniques – NNs and modified PCA – was beneficial, mainly by greatly simplifying the NN training.

In *Application of feedforward neural networks and partial least squares regression for modelling kappa number in a continuous Kamyr digester* (Dayal et al, 1994), the application of feedforward NNs and PLS techniques in order to develop a predictive model for the Kappa number and to gain a better understanding of the correlations between the Kappa number and other digester process variables is presented. In kraft pulping, the Kappa number measures pulp quality, relating to the lignin content remaining in the pulp. Because of long delay between off-line Kappa number measurement and potential corrective actions, an inferential on-line model predicting the Kappa number will enable earlier execution of corrective control actions.

Digester data was collected and used to train feed-forward NNs and to build the PLS model. It was observed that both those techniques performed well, and their results were similar. However, while the NN model offered no process insights, the PLS analysis improved process understanding by high-lightening interactions between variables. This can be very useful for reducing the variations in the output variable- the Kappa number, in this case.

*Optimal selection of soft sensor inputs for batch distillation columns using principal component analysis* presents a comparative study of NN, PLS and PCA combinations for predicting purposes, and their performance. The challenges facing proper process variables to be used for developing the model are detailed, and the predictive performances of models having their inputs determined by a PCA analysis versus conventional inputs (not selected with PCA).

The analysis is quite comprehensive, covering different data sets – ranging from traditional inputs to different variations found by the PCA analysis – and different inferring methodologies.

The results of the study prove that proper selection of process parameters, according to their impact on the variability of the output, as determined by the PCA analysis has a significant beneficial impact on increasing model performance.

In *Identification of neural dynamic models for fault detection and isolation: the case of a real sugar evaporation process* (Patan et al, 2005), dynamic neural networks are used to develop a fault and diagnostic system. The dynamic neuron model can easily take into account the time delays between inputs and outputs that characterize many industrial processes. Historical process variables from a sugar factory's evaporation station and four classes of behavior (normal and three abnormal behaviors – faults) are used to train the network. Using a process model representing the nominal operating conditions, this model was applied for detecting the faults and identifying them. The study shows that this kind of model performs effectively if the fault classes were included in the training data set. Unknown faults can be detected, but classified only as "unknown", since their respective classes were not present in the training data.

In *Data driven process monitoring based on neural networks and classification trees* (Zhou, 2004), a process monitoring, fault detection and classification scheme based on NN model is developed for a batch polymerization reactor for the production of polymethylmethacrylate.

In order to reduce the dimensionality of the proposed NN model, the data is processed prior to the NN training. The process measurements go through a polynomial regression step and the polynomial coefficients, which are usually of far lower dimensionality than the original data, are used to build the NN model, thus resulting in a significant reduction in model construction time. A feed-forward NN was used to model the process (feature extraction), and a radial basis function (RBF) neural network was used to perform the classification task. This approach performed very well. A comparison between the NN accuracy when the input data goes through the polynomial-based preprocessing step versus unprocessed data shows a significant reduction in training time, at a cost of a very slight decrease in model performance.

In *Online monitoring of steel casting processes using multivariate statistical technologies: From continuous to transitional operations* (Zamprogna et al, 2005), a model based on the PCA and PLS method was developed to monitor operations and to detect process abnormalities of a steel caster. This system was commissioned at Dofasco's No. 2 caster. It was found that the PCA/PLS scheme is able to systematically discover correlations present in large amounts of industrial data. Dominant patterns are identified in the data, representing the normal operating conditions. Deviations from these patterns represent abnormal behaviors of the process and they can be readily detected. The model provided a deeper understanding of f the interactions between process parameters, and its on-line detection of faults and process variables most likely correlated to the faults help operators ensure a normal operation.

In *Clustering algorithms in process monitoring and control application to continuous digesters* (Ahvenlamp et al, 2005), a combination of NN and fuzzy logic is used to predict the Kappa number and to monitor changes in process variable for fault detection and classification purposes. The system used is a combination of the Kohonen self-organized map (SOM) neural network and the Takagi-Sugeno clustering method. The SOM is a type of unsupervised classification (clustering) NN. The fuzzy clustering mentioned groups the data into clusters of similar behavior, and then fuzzy rules explaining the behavior of the data in the vicinity of clusters as polynomial functions of the inputs are formulated.

The modeling data was normal operation data from an industrial continuous digester. The SOM is trained with the data to be able to predict the Kappa number value and to detect and classify changes in process parameters affecting the Kappa number (process faults). The SOM codebook matrix is used as an input for the fuzzy clustering model, which performs the final Kappa number prediction. It is reported that this configuration achieved good prediction performance, and was able to detect abnormal behaviors, even when the deviations were small.

In *Intelligent lime kiln control system*, (Jarvensivu et al, 2001) a feedfoward neural network model was developed to predict the estimate residual CaCO3 content of the burned lime. The predicted value is used in the control scheme of the kiln, as a part of the fan speed feedforward control and feedback adjustment. The model performance was very good, and it was implemented at a pulp & paper mill in Finland. The main recorded benefits include improved lime kiln efficiency (increased throughput), decreased variations in lime quality, and decreased energy consumption.

In *Application of feedforward neural networks for soft sensors in the sugar industry* (Devogelaere et al, 2002), a feedforward neural network is trained on historical data to predict the electrical conductivity of the mixture of crystals and molasses during the in the crystallization process so that it can replace the lab-test procedure. The network was trained using was a standard back propagation algorithm, and its performance was deemed satisfactory. However, the authors recommend that a recurrent NN model would be more appropriate than the back propagation method, since it will perform better in the case of time dependent sets. The publication also mentions industrial implementations of neural network models in the cane sugar industry, mainly for the evaporation and crystallisation processes.

In *An adaptive neuro-fuzzy inference system as a soft sensor for viscosity in rubber mixing process*, (Merikoski et al), the development of an adaptive neuro-fuzzy inference (ANFIS) system for predicting rubber viscosity using three measured process variables in rubber mixing process. The ANFIS model is a fuzzy inference system formulated as a feed-forward neural network: the fuzzy model extracts rules from the data, and the coefficients of these rules are further tuned with a neural network model. The performance was the system was good, and several improvements have been suggested, such as removal of outliers from the training data.

In *Multivariate process monitoring and early fault detection (MSPC) using PCA and PLS* (Yoon et al, 2003), PCA and PLS techniques were applied to a continuous process at Honeywell, Geismar, LA to detect process fault and predict fuel consumption. Based on normal operating data, a PCA model was developed, and process faults were identified as deviations from normal operation. A PLS model was developed to predict natural gas flowrate. The model performs quite well, detecting the faults a few hours in advance, allowing plant personnel to make the required corrective actions and accurately predicting natural gas consumption.

## 6.2 Other published material

The main data mining applications for industrial processes are reflected very well in the papers mentioned in the previous section: increased process understanding, process monitoring, fault detection and classification and prediction of process variables.

Other publications were reviewed, since they present similar applications to the ones already presented – the industrial process to which they are applied varies, but the data mining techniques used, the models and the task they perform are very close to those presented in the previous section.

Summarizing their content as it was done in the previous section would have become quite redundant, so they are listed, followed by a brief description, above:

- *Use of artificial neural networks process analyzers: a case study* (Duwaish et al, 2002), presents the development of feedforward NN model for predicting O2 contents in the boiler flue gas from other measured variables

- *Bayesan neural networks on the inference of distillation product quality*, (Hall Barbosa et al, 2002) presents the development of different types of NNs for predicting quality of distillation products, for the REPAR refinery in Brazil

- S*oft sensors for product quality monitoring in debutanizer distillation columns*, (Fortuna et al, 2005) presents the development of a multi-layer perceptron NN for predicting the gasoline and butane concentration of a debutanizer column; this application is mimplemented in a refinery in Italy

- *A self-validating inferential sensor for emission monitoring* (Qin et al, 1997) presents a PCA/NN based application for predicting boiler emissions

- *Adaptive neural networks for intelligent operation of the activated sludge process* presents different types of NN models used for data monitoring, filtering, and fault detection and classification in the treatment of wastewater.

*A Good Practice Guide* (number 215, 1996) prepared for UK's Department of the Environment presents the benefits of some data mining techniques applied to industrial processes. It targets the chemical, petrochemical, food and drink, paper and board and metals sectors, as well utility companies. Soft sensors, expert systems and process models developed using data mining techniques are presented are ways of optimizing the process and reducing production and energy costs. Numerous successful industrial implementations of these techniques are mentioned in the guide.

## 6.3 Overview of artificial intelligent systems implemented in the surveyed industrial sectors

A report on the *Application of Artificial Intelligence technology to Increase Productivity, Quality and Energy Efficiency in Heavy Industry*, prepared for CETC-Ottawa in 1995 was also examined, in order to gain a better understanding of the use of data mining techniques in the manufacturing industry.

Even though the report was prepared 10 years ago, it does provide a representative trend concerning the application of different AI tools in the industry. Since these AI tools are essentially neural network and fuzzy logic based, the information presented in this study can also be used as an indication of industry acceptance to these data mining techniques.

The objectives of the study were to:

- asses current uses of AI in heavy industries & related industries
- identify where & how AI may increase productivity, quality and energy efficiency in industry
- identify key barriers to the development & implementation of AI in the industry
- identify application targets in 5 sectors
- iron and steel
- cement
- mining and metallurgy
- oil and gas
- pulp and paper

The study identified 177 intelligent systems prototyped and/or applied in heavy industry:

- process control, including soft sensors (inference mechanisms), fault detection and diagnosis
- process monitoring, including online decision support, model decision processes instead of equipment or operations
- scheduling and planning
- fault diagnosis and maintenance, including reduction of equipment downtime, systematic analysis of information about equipment failure
- design (generation of alternative designs)

The study does not include on university prototypes, research prototypes, or other systems not implemented or tested in an industrial environment. If multiple sites are counted, then the number goes up to 250.

The table below presents an overview of the use of AI tools:

| Use of AI tools overview | | | | | | |
|---|---|---|---|---|---|---|
| | | Iron & Steel | Cement | Mining & Metallurgy | Oil & Gas | Pulp & Paper |
| Application | Control | 39 % | 60 % | 39 % | - | 4 % |
| | Monitoring | 14 % | 14 % | 18 % | 11 % | 45 % |
| | Diagnosis | 15 % | - | 15 % | 11 % | 32 % |
| | Design | 3 % | - | 8 % | 22 % | 5 % |
| | Scheduling | 16 % | - | 3 % | 11 % | 8 % |
| | Planning | 10 % | 20 % | 16 % | 22 % | 3 % |
| | Maintenance | 3 % | 6 % | 1 % | 23 % | 3 % |
| | Total number of applications | 73 | 15 with multiple sites: 80 | 52 (with field tested prototypes) | 9 | 23 with multiple sites: > 50 with university prototypes: 70 |
| Technology | Expert Systems | 82 % | 35 % | 98 % | 90 % | 89 % |
| | Neural Networks | 9 % | - | 2 % | - | 11 % |
| | Fuzzy Logic | 8 % | 65 % | - | 10 % | - |
| | Other | 1 % | - | | - | - |
| Location | North America | 13 % | 20 % | 57 % | 78 % | 48 % |
| | Europe | 31 % | 67 % | 26 % | 10 % | 40 % |
| | Asia | 50 % | 8 % | 6 % | 12 % | 8 % |
| | Other | 6 % | 5 % | 11 % | - | 4 % |

The reported key benefits following the implementation of these AI tools, as reported by plants using them, are increased:

- product quality – control, monitoring and diagnosis
- scheduling – productivity
- energy efficiency.

These factors are considered to be interdependent and are treated as the key components in heavy industry's strategy to achieve higher competitiveness.

# 7. DISCUSSION OF REVIEWED LTERATURE

This review of published work regarding data mining applications in the industrial process sector provided information about current trends of the techniques and methods used to develop these data mining tools. Current applications were identified, along with suitable data mining methods to perform the required task. It was seen that different methods are used for different tasks: for example, classification can be done either with a neural network based model or with a PLS method.

## 7.1 Neural network based methods

This literature assessment indicates that numerous data mining applications for prediction and classification use neural network based models.

Feedforward models using a backpropagation algorithm are widely used for regression and supervised classification, as reported in some papers: Duwaish et al (2002), Fortuna et al (2005), Jarvensivu et al (2001), Dayal et al (1994), Edwards et al (1999), Zamprogna et al (2005), Zhou (2004).

For supervised classification, radial basis function (RBF) neural networks are very appropriate; this type of network was used by Zhou (2004). Kohonen self-organized maps are used for clustering (unsupervised classification); Ahvenlamp et al (2005) presented the development and application of this type of NN.

Devogelare et al (2002) suggest that a recurrent NN model would be more appropriate than the back propagation method when dealing with time dependent data sets. Hall Barbaso et al (2002) present a Bayesan NN that has better prediction accuracy than a standard backpropagation multi-layer perceptron.

Many of these neural network based data mining applications are implemented and perform well on-line.

The main neural network techniques found in this literature review, as well as their application areas are summarized in the table below:

| Application | NN method |
|---|---|
| Numerical value prediction | Feedforward (backprogation); Bayesan |
| Qualitative value prediction | Feedforward (backprogation); Bayesan |
| Supervised classification | Radial Based Function (RBF) |
| Unsupervised classification | Kohonen self-organizing maps |

## 7.2 Factor based statistical methods (PCA/PLS)

This literature assessment indicates that numerous data mining applications for prediction and classification use PCA and PLS models. PCA models are used for descriptive modeling, providing a better understanding of the process by showing correlations between process variables.

PCA analyses were used for studying and identifying interactions between variables, identifying the variables that have the greatest contribution on the process variability, as presented in some papers: Zamprogna et al (2005), Yoon et al (2003), Zhang et al (2006), Edwards et al (1999).

PLS models are used for predictive purposes, either regressive or classification prediction, as shown in the following papers: Dayal et al (1994), Yoon et al (2003).

## 7.3 Fuzzy logic based methods

This literature assessment indicates that data mining applications for modeling and prediction can be developed using fuzzy logic techniques.

Fuzzy logic based clustering is used for extracting association rules from the data set and building a model that can be used for prediction purposes; such models were presented by Merikoski et al, and Ahvenlampi et al (2005).

## 7.4 Comparison of methods

This literature review showed that all the methods described previously – NN, PCA/PLS and fuzzy logic – are used for developing the main data mining applications for industrial processes:

- process modeling and monitoring
- soft sensors, for regression or classification purposes
- fault detection and classification

The predictive models developed using NN, PLS and fuzzy logic performed well, displaying a good accuracy of prediction. It was also shown that linear PLS models can approximate very well a non-linear process. This it is not unusual, since the nominal operating window of a nonlinear process can be almost linear. Dayal et al (1994) performed a comparative between a linear PLS model and a neural network model for predicting a Kappa number in a continuous Kamyr digester, and the results showed that there was no major performance difference between both methods.

The literature review revealed that most of the predictive models are developed using neural networks. These models perform very well in nonlinear conditions, and they are suitable for either regressive or classification tasks. However, no insight into the process could be obtained from the neural network models.

PCA/PLS models have the ability to systematically examine the data in order to detect correlations between variables and explain the variability of the data set.

A PCA analysis summarizes the information in a data set containing many variables into a few new summary variables – the principal components – which provide an overview of the trends and patterns in the data set; it shows the correlations between the variables that form the principal components, by revealing their magnitude (large or small correlation) and the manner (positive or negative correlation).

Predictive accuracy is a critical aspect of models, but it is not the only aspect. Besides having good prediction capabilities, predictive PLS models can also shed insight into which of the predictor variables are most important and into how predictor variables interact, in the sense that the effect that one has on the response variable depends on the values taken by others.

Neural network and fuzzy logic models have very good predictive capabilities, but they do not have the descriptive characteristics of PCA/PLS models. When no insight into the process is required, NN and fuzzy logic are the preferred tools for prediction – they are extensively used in soft sensor development, for example. When a process needs to be better understood, to identify the parameters affecting its variability, PCA/PLS methods are much more suitable.

## 7.5 Hybrid models

This literature assessment indicates a growing trend of combining statistical methods, such as principal component analysis (PCA) with artificial intelligence methods, such as neural networks. When dealing with a process containing a large number of variables, PCA can be used in order to reduce them, by grouping them into a few new summary variables – the principal components – that approximate well the variability of the process. This allows for a reduction in the number of variables without losing significant information. These principal components can then be used as inputs for a neural network based predictive model. This combination presents good prediction accuracy and the reduced dimensionality of the input space results in decreased neural network complexity and training time, as reported in some papers: Edwards et al (1999), Qin et al (1997). Zamprogna et al (2005) used a PCA analysis in order to determine which variables have the greatest effect of the variable to be estimated, and then used them as optimal inputs to a neural network. This proper input identification leads to an increase of the predictive performance of a neural network.

The same data pre-processing principle prior to network training was also applied by Zhou (2004): using a polynomial regression step, instead of PCA, principal components were extracted from the data and then use as inputs for the NN. A comparison between the NN accuracy when the data goes first through this transformation step versus unprocessed data shows a significant reduction in training time, at a cost of a very slight decrease in model performance.

Neuro-fuzzy combinations are also used for developing data mining applications: Merikoski et al presented the development of a soft sensor using a fuzzy model that has its coefficients optimized by a neural network.

In general, these hybrid configurations, integrating more than one methods in an application, perform performs better than configurations based solely on only one method.

# 8. COMMERCIAL SOFTWARE TOOLS

This section reviews the key features of some selected commercial software used to design and/or implement soft sensor applications.

There are a multitude of commercial tools offering modelling and predictive capabilities based on statistical analysis or on artificial intelligence approaches. Some of them have a considerable market penetration in the industrial manufacturing sector.

While these tools are quite powerful of designing and implementing on-line either statistical or Artificial Intelligence based descriptive or predictive models, it is rare that they offer the possibility of combining both methods into a single hybrid system. This is a major drawback, since such a hybrid system is more powerful than a system based on only one method, as it was revealed in the literature examined for this assessment study.

Some of the commercial available statistical and AI-based tools are presented below. The list presented below is not exhaustive, but it can be considered representative of the commercial tools available for developing and implementing data mining applications in the process industry.

**Product:** **Facnet**

Vendor: Pacific Simulation

Web Site: http://www.pacsim.com/FN/

FactNEt is a statistical analysis system based on the factor analysis technique; it analyses data using the assumption that the data are interrelated and it quantifies the relationships between the variables. It offers data analysis capabilities, and it creates predictive models.

Industrial implementation:

- on-line applications in the pulp and paper as well as the power industry

Remarks:

- The *Automated Engineering Intelligence Using Pattern Recognition Tools* success story available on their site describes the FactNet implementation at an electric utility generation station. FactNet possessed considerable market penetration in data analysis and process control for the pulp and paper industry.However, the company states that the reason why Neural Networks were not selected was the lack of the availability of an “off the shelf" neural networks system available at that time.

**Product:** **NeurOn-Line and G2**

Vendor: Gensym

Web Site: http://www.gensym.com

NeurOn-line is object-oriented software for building neural networks for real-time applications. It can used as a stand-alone application, or it can be used as a module for the Gensym G2 Expert System. G2 software applies real-time rule technology for decisions that optimize operations and that detect, diagnose, and resolve process faults and/or equipment failures – such as sensor failure, for example.

Major features:

- the Neur-on-Line can be used in the wizard mode, where a network is automatically designed based on the user's needs – prediction, classification – or in a user-defined design mode, where the user chooses the network configuration – type, number of layers & neurons, functions, etc.
- G2 offers a wide range applications: it can be used as an expert system, supervisory control, fault detection and diagnostics. Its rule database – it also incorporates a fuzzy logic rule and inference mechanisms – allows it to take real-time decisions, based on the data provided by hard and soft sensors.

Industrial implementation:

- on-line applications in a variety of manufacturing industries, such as chemical, cement and automotive

Remarks:

- while the interface is user-friendly, making the software relatively easy to use, the selection of neural networks types and of the activation functions is very limited: only backprogation and radial-based networks, with either sigmoid or linear functions are available.
- the user does not have much control over the network configuration, even in the user-designed mode. Access to the network configuration parameters is very limited.

**Product:** **AspenIQ**

Vendor: Aspen Tech

Web Site: http://www.aspentech.com

AspenIQ is an inferential technology package (value prediction); models based on neural networks, fuzzy logic, linear and non-linear PLS.

Major features:

- contains modules for implementing inferential sensors on-line
- there are also modules for data collection, lab data validation, and inference model updating
- variable selection can be done using a cascaded genetic algorithm that determines the most important variables

Remarks:

- - it generates linearized inferential models.

**Product:** **Simca-P**

Vendor: Umetrics

Web Site: http://www.umetrics.com

SIMCA-P uses multivariate data analysis tools such as principal components analysis and partial least squares, with a graphically oriented, Windows based visualization interface in order to create predictive models that can be used as soft sensors.

Major features:

- user-friendly, easy to use interface
- accepts various data formats, such as Excel spreadsheets, for example
- can be used plant or process data analysis and trouble shooting, process monitoring, fault detection and classification, batch data analysis, structure-property relationships, or spectroscopic calibration, SIMCA-P and P+ provide the necessary tools in a simple and easy to use format.

Industrial implementation:

- chemical, pharmaceutical, pulp and paper production

**Product: Property Predictor**

Vendor: Pavillion Technologies

Web site: http://www.pavtech.com

Property Predictor uses multivariate data analysis tools – and neural networks, maybe not really clear from their website what are the predictive methods used in the software – to build predictive models acting for paper machine sheet properties. It also can be used as a sensor validation tool to identify defective sensors. It can be integrated with ProcessPerfecter®, Pavilion's multivariable predictive control and optimization software, to provide a complete advanced process control solution.

Industrial implementation:

- pulp and paper, cement

**Product: Pegasus OS2003**

Vendor: Pegasus Technologies

Web site: http://www.pavtech.com

Pegasus OS2003 uses neural networks models to optimize boiler operation, by performing on-line combustion operation. It acts as a soft sensor analyzing emissions and using this information to optimize combustion, by computing optimal air and fuel flows.

Industrial implementation:

- power industry

**Product:** **Matlab Neural Networks, Fuzzy Logic and Statistical toolboxes**

Vendor: Mathworks

Web site: http://www.mathworks.com

These are toolboxes for the Matlab programming language. They offer a variety of functions that are used for neural network design and statistical analysis purposes.

Major features:

- not that user-friendly, it requires a good knowledge of the Matlab programming language in order to use them
- it offers quite an exhaustive collection of code (functions), for both neural network development and statistical analysis
- very suitable for designing and testing soft sensor programs, given its powerful functions already available
- not so suitable for deploying them, given its relative high level of Matlab programming required to write a complete program, to incorporate the available functions into the program and to test different configurations in order to find the optimal one

Industrial implementation:

- widely used in universities, R&D centers and other institutions for the development (programming) of neural network, fuzzy logic and statistical based models

**Product:** **PEPITo**

Vendor: PEPITe

Web site: http://www.pepite.be

PEPITe uses a variety of data mining tools to analyze industrial data. It allows users to code complex and recurring data processing and modeling tasks or even to test new data mining algorithms. It has data analyses capabilities, such as PCA, and it provides the user with a large selection of modeling and knowledge discovery methods, which include: decision and regression trees, fuzzy logic and different NN models.

Industrial implementation:

- pulp and paper , metals, glass, power system management in different industries

Other tools to design neural networks – such as NeuroSolutions – or to perform statistical analysis on data sets were reviewed, but since they are conceived only for programming purposes and not for an on-line implementation, they are considered to be beyond the scope of this assessment.

# 9. CONCLUSION

The review of literature provided information about current trends of data mining applications developed in the industrial process sector. Current applications were identified, along with suitable methods to perform the required assignment. A brief explanation of different data mining tasks, as well as different methods was provided.

The main data mining applications in the industrial process sector are:

- process monitoring
- soft sensors
- fault detection and classification
- improved process understanding

The literature review showed that different methods can be used to achieve the same objective. For prediction, most of the published work reviewed concerning predictive models dealt with neural network-based approaches, and, in a lesser degree, with fuzzy logic and PCA/PLS methods. Linear PLS can also be successfully applied for non-linear processes.

Neural networks and fuzzy logic model have very good predictive capabilities, but they do not offer any process insight, as PCA/PLS do.

PCA/PLS techniques are very powerful tools for creating descriptive models, since they reveal the correlations between variables, providing a better understanding of the process.

Some data mining techniques can be combined into a hybrid model, resulting in applications with superior performances. Integrating the complementary features of different methods into a hybrid system could overcome the limitations of individual solution strategies. The most encountered hybrid system in the literature review was PCA/NN, and, to a lesser extent, NN/fuzzy logic.

The PCA/NN hybrid systems yielded very good results, since the input variables were properly selected by a PCA analysis. The PCA analysis also allowed for a reduction in the number of variables without losing significant information, resulting in a decreased neural network complexity and training time.

In the industrial process sector, a structured framework for implementing data mining applications will provide users with a simplified, but efficient methodology for applying data mining tasks; automated tools for analyzing the data and generating the results can assist operators and engineers in decision making regarding an optimal plant operation.

A brief survey of the commercial data mining packages used in the industrial process sector was also presented. It can be seen that most of them do not allow the development of hybrid systems.

# Annex A
# Bibliographical list

*Principles of data mining*, David Hand, Heikki Mannila, Padhraic Smyth

Frawley W., Shapiro G. and Matheus C., *Knowledge Discovery in Databases: An Overview*, Fall 1992 AI Magazine, pp. 213-228.

Chen M., Han J. and Yu P., *Data mining: an overview from a data base perspective*

Venkatasubramanian V., Rengaswamy R., Yin K. and Navuri S., *A review of process fault detection and diagnosis*, Part III: *Process history based methods*, 2003 Computers and Chemical Engineering, vol. 27, pp. 327-346

Edwards, P.J., Murray, A.F., Wallace, A.R. and Barnard, J., *The application of neural networks to the paper-making industr*y, 1999 European Symposium on Artificial Neural Networks proceedings, pp. 69-74

DayalB.S., MacGregor J.F., Taylor P.A., Kildaw R. and Marcikic S., *Application of feedforward neural networks and partial least squares regression for modelling kappa number in a continuous Kamyr digester*, 1994, Pulp & Paper Canada 95:1, pp.26-32

Zamprogna E., Barolo M. and Seboprg D., *Optimal selection of soft sensor inputs for batch distillation columns using principal component analysis*, 2005 Journal of Process Control, vol. 15, pp. 39-52

Patan K. and Parisi T., *Identification of neural dynamic models for fault detection and isolation: the case of a real sugar evaporation process*, 2005 Journal of Process Control, vol. 15, pp. 67-69

Zhou, Y., *Data driven process monitoring based on neural networks and classification trees*, 2004, PhD dissertation, Texas A&M University

Zhang, Y., Dudzic, M.S., *Online monitoring of steel casting processes using multivariate statistical technologies: From continuous to transitional operations*, 2006 Journal of Process Control 16, pp. 819-829

Ahvenlampi T., and Kortela U., *Clustering algorithms in process monitoring and control application to continuous digesters*, 2005, Informatica 29, pp. 101-109

Jarvensivu, M., Saari K. and Jamsa-Jounela S., *Intelligent lime kiln control system*, 2001, Control Engineering Practice, vol. 9, pp. 589-606

Devogelaere, D., Rijckaert M., Osvaldo G. and Lemus G. , *Application of feedforward neural networks for soft sensors in the sugar industry* , 2002, IEEE September issue

Merikoski, S., Laurikkala, M. and Koivisto, H., *An adaptive neuro-fuzzy inference system as a soft sensor for viscosity in rubber mixing process*, Automation and Control Institute, Tampere University of Technology

Yoon, S., Kettaneh, N., Wold, S., Landry, J. and Pepe, W., *Multivariate process monitoring and early fault detection (MSPC) using PCA and PLS*, 2003, Plant Automation and Decision Support Conference

Duwaish A., Halawani L. and Mohandes M., *Use of artificial neural networks process analyzers: a case study*, 2002, European Symposium on Artificial Neural Networks 2002, proceedings pp. 465-470

Hall Barbosa C., Vellasco M., Melo B., Pacheco M. and Vasconcellos L., *Bayesan neural networks on the inference of distillation product quality*, 2002, VII Brazilian symposium on neural networks

Fortuna L., Graziani S. and Xibilia M., *Soft sensors for product quality monitoring in debutanizer distillation columns*, 2005 Control Engineering Practice, vol. 13, pp. 499-508

Qin S., Yue H. and Dunia R., *A self-validating inferential sensor for emission monitoring*, 1997 AACC vol. 9

Barnett W., *Adaptive neural networks for intelligent operation of the activated sludge process*, 1997 Water Environment Federation Specialty Conference: Computer Technologies for the Competitive Utility, proceedings

*Reducing energy costs in industry with advanced computing and control*, Good Practice Guide 215