INPUT VARIABLE SELECTION FOR MULTIVARIATE STATISTICAL PROCESS MONITORING

MD. MUSFIQUR RAHMAN

Input Variable Selection for Multivariate Statistical Process Monitoring

by

©Md. Musfiqur Rahman

A Thesis
submitted to the School of Graduate Studies
in partial fulfillment of the requirements for the degree of

Master of Engineering

Faculty of Engineering and Applied Science

Memorial University of Newfoundland

May 2013

St. John's Newfoundland

Abstract

Multivariate statistical methods are widely used in process operations for predicting unmeasured quality, and detection and diagnosis of faults. Performance of these monitoring tools greatly depends on selecting the right set of variables as input to the tools. In a typical chemical process on average 1500 variables are logged. Selection of appropriate input variables from these large set of variables is a daunting task. This thesis investigates the application of retrospective Taguchi method in selecting input variables for multivariate statistical monitoring tools. Taguchi's design of experiment (DoE) approach has been widely used in industrial process design, primarily in manufacturing industries for optimizing process parameters. Instead of relying on an arbitrary selection of levels, experiments are conducted following an orthogonal array as determined by the Taguchi method. In the current research, the method is adapted for selecting important input variables for process monitoring tools, namely, support vector regression (SVR) and principal component analysis (PCA). Taguchi's DoE assumes that variables are uncorrelated which is contrary to process data. Process variables are highly correlated and show dynamic variations due to the frequent changes made in the set points causing difficulty to select data to match the orthogonal array of the Taguchi method. These implementation difficulties were addressed in the proposed methodology. Retrospective Taguchi method was adapted for dealing with process data. Additional data preprocessing and correlation analysis steps

were proposed to condition process data for Taguchi method. Detailed methodologies to apply Taguchi method to select input variables for SVR and PCA are described in the thesis. The methodologies were demonstrated using industrial data from a petrochemical process and a hydrometallurgy process respectively. The performance of the proposed Taguchi based method was compared with variable importance in projection (VIP) method. The industrial case studies clearly show that the proposed methodology can minimize the computational efforts in variable selection and it can improve the performance of the monitoring tools.

Acknowledgements

First and foremost, I would like to express profound gratitude to my supervisors Dr. Kelly Hawboldt and Dr. Syed Ahmad Imtiaz, for their invaluable support, motivation, encouragement, supervision and useful suggestions throughout this research work. Their moral support and continuous guidance enabled me to complete my work successfully.

I would like to thank John Halfyard of Vale for his continuous support in providing me necessary information and guidance on the data. I thank my fellow lab mates in process control engineering group, Mohammad Aminul Islam Khan and Md. Raihan Mallick, for stimulating discussions, valuable ideas and encouragements in every aspects and for all the fun we have had in the last two years. Also, I thank my home friends in Memorial University, Md. Kabirul Islam and Sujan Dutta, for supporting and encouraging me in many ways.

I would also like to thank my family members, especially my wife Dr. Most Hosne Ara Khatun and my son Md. Muktadir Rahman, for supporting and encouraging me to pursue this degree. I thank them for their love, support, and confidence throughout the whole career.

Finally, the financial support from School of Graduate Studies (SGS), Memorial University of Newfoundland and Faculty of Engineering and Applied Science is much appreciated.

Table of Contents

List of Tables

List of Figures

List of Abbreviation

CCD = Counter Current Decantation

4-CBA = 4-Carboxy Benzaldehyde

FDD = Fault Detection and Diagnosis

LRT = Leach Residue Thickener

MKNN = Modified K-Nearest Neighbor

PCA = Principal Component Analysis

PTA = Purified Terephthalic Acid

PLS =Partial Least Square

RMSECV = Root Mean Square Error of Cross Validation

RMSEP = Root Mean Square Error of Predication

S/N = Signal to Noise

SVR = Support Vector Regression

VIP = Variable Importance in Projection

Chapter 1

Introduction

1.1 Importance of variable selection

Multivariate statistical methods are widely used in process industries for monitoring purposes. On an average, about 1500 variables are logged at anytime in a process. Due to the sheer number of variables, it is difficult to select the most relevant variables for any application. Process knowledge is typically used to select input variables. However, often limited process knowledge is available to the application developer and has to rely on statistical packages to complement process knowledge.

It is always a challenge to decide which variables should be included in a model to achieve the best performance [Abrahamsson et al., 2003]. Today, variable selection procedures are an integral part of virtually all widely used statistics packages [George, 2000]. The main objective of variable selection methods for monitoring tools is to have a concise model that can improve the prediction & fault detection performance of the monitoring tool. The major concern with multivariate methods is the high probability of over-fitting, which is aggravated when the number of variables is large. Recently, support vector regression (SVR) is gaining much attention as an in-

ferential predictor due to its ability to capture process non-linearity. However, SVR is a computationally intensive method. A concise model can cut down the development time and make online implementation easier. Principal component analysis (PCA) is a widely used fault detection and diagnosis tool in process industry. The performance of a PCA based fault detection & diagnosis tool depends largely on the selection of right kind of input variables. It is important to select those variables that bear the fault signatures in the process. A concise list of variables also facilitates identification of the root cause of fault. Both SVR and PCA do not have any built-in method to select input variables. Developers mainly rely on process knowledge, and trial and error to select input variables. Therefore, it is important to develop systematic methods to select input variables for these monitoring tools.

1.2 Objectives of the Current Study

This research is aimed at developing an input variable selection methodology for an SVR inferential predictor and PCA.

The following objectives are set for the current study:

- Adapt Taguchi's experimental design method for dealing with correlated process data.
- Develop a comprehensive input variable selection methodology for a SVR inferential predictor.
- Apply SVR along with the variable selection method to build an inferential predictor for predicting the quality variable (4CBA) of a Purified Terephthalic Acid (PTA) process.
- Develop a comprehensive variable selection methodology for PCA.

- Apply PCA along with the proposed variable selection method to build fault detection models for Leach Residue Thickener (LRT), CCD 1 Thickener and CCD 2 Thickener of a nickel hydromet process.

1.3 Thesis Organization

The first chapter of this thesis briefly describes the motivations for this research in the context of the importance of variable selection and the objectives of the study.

Chapter 2 covers an extensive review of literature on existing variable selection methods. A brief introduction of Taguchi's experimental design method, and variable selection in SVR and PCA are also described.

Chapter 3 describes the input variable selection methodology for an inferential predictor using support vector regression (SVR). The methodology is demonstrated through a case study from a petrochemical process. Prediction performance is compared with the partial least square (PLS) and the variable importance in projection (VIP) method.

Chapter 4 is devoted to describing the input variable selection methodology for an inferential predictor from a large set of correlated variables. The methodology is demonstrated through the same case study used in Chapter 3.

Chapter 5 describes the methodology of variable selection for PCA. The methodology is demonstrated by developing fault detection models for the leach residue thickener (LRT), counter current decantation (CCD) 1 thickener and counter current decantation (CCD) 2 thickener of a nickel-hydromet process. The root causes of the faults are discussed through contribution and trend plots. The effectiveness of variable selection is also demonstrated through a comparison study.

Chapter 6 briefly outlines the concluding remarks with a summary of useful

findings. Recommendations for future work are also provided.

Chapter 2

Literature Review

2.1 Classification of variable selection methods

Subsets of variables can be selected in two forms: (a) feature or input variable selection, and (b) feature or input variable extraction. The feature selection method basically selects subsets of original variables. Application of filter-based feature selection methodology in classification problems using Least-Squares Support Vector Machines can be found in [Herrera et al., 2006],[Rossi et al., 2006]. Alternatively, in the feature extraction method, subsets of variables are extracted by linear or nonlinear transformations of the original ones. Principal component analysis is a popular feature extraction method where principal components are extracted as a linear relationship of the original input variables. Subsequently, only the major PCs are used for further analysis. Application of feature extraction using PCA in the context of Artificial Neural Network (ANNs) can be found in [Jalali-Heravi et al., 2007]. Feature extraction can reduce the number of variables by grouping correlated variables.

Common variable selection methods for multivariate data analysis can be grouped into two categories: filter methods and wrapper methods [Pierna et al., 2009]. Filter

methods utilize an indirect estimator to measure the prediction ability of the selected subsets of input variables. Alternatively, wrapper methods directly utilize the multivariate method of interest to measure the prediction performance. In the following subsections, these two variable selection methods will be described in detail.

2.1.1 Wrapper methods

Classical wrapper methods are based on sequential techniques to eliminate variables. The process starts by choosing a subset of variables from the input variables. A regression model is subsequently built using that subset to predict the response variables. Cross validation is performed, and the root mean square error of cross validation (RMSECV) or prediction (RMSEP) is used as a criterion to evaluate the subset. Wrapper methods directly measure the generalization ability of the subset of input variables using the learning algorithm of interest. For example, for the selection of important variables for a PLS model, the model itself can be used as a regression model to quantify the prediction ability of the selected subset. The subset that provides the best prediction performance is selected as the final model. The selection of variables can use a forward selection technique, a backward selection technique or a combination of both.

Forward selection methods start with a single variable, and then variables are added one at a time. After each addition, a model is built to evaluate the performance. The main drawback of forward selection is that it produces weaker subsets at the initial stage. As a result, the importance of a certain variable is not assessed in the context of other variables which are not yet included in the model.

Backward selection is the opposite of forward selection, which starts with all the variables, and, subsequently, variables are removed to see the performance. Backward variable selection method for partial least square (PLS) can be found in [Pierna et al.,

2009]. The first step is to fit a model with all the variables. In the subsequent steps, one variable is dropped at a time and a new model is constructed using a training data set. The new model is applied to a test data set to see the performance in terms of RMSEP. If the RMSEP of the new model is found to be less than the previous one, it indicates that the variable, which is left out, is not significant in terms of prediction, and thereby, is removed. The procedure is repeated n (number of variables) times by successively re-fitting reduced models. The final model is constructed with the variables which give the minimum RMSEP. In the backward elimination method, there might be a situation in which the variable, which is removed at an early stage, may have a significant effect when added to the final reduced models [Pierna et al., 2009].

Iterative PLS (IPLS) is an example of a wrapper-based variable selection method which combines both backward and forward steps. It initially starts with a small number of variables, and, subsequently, new variables are added to the list or removed from the list based on the improvement of the model. The initial variable selection is done randomly and a PLS model is built using the selected variables followed by evaluation using cross validation. In the next step, a variable is added or withdrawn from the model randomly, and a new PLS model is built and evaluated by cross validation. If RMSE of the new cross validation is lower than the original, the new set of variables replaces the original. The algorithm is terminated when every variable is tested atleast once without providing any improvements [Osborne et al., 1997].

In the recent past, extensive research has been done to use genetic algorithm (GA) for variable selection. GA is an optimization method applied to identify a subset of the measured variables that provides the lowest RMSECV for the target regression model. Details of GA algorithm for variable selection in PLS regression can be found in [Wise et al., 2007]. The algorithm has several steps. The first step of the GA

is to generate several subsets of randomly selected variables. The pool of all these subsets is termed as population. Each subset of variables in the population is called an individual. Each variable entered in a subset is converted to a binary number, and this binary structure is termed gene. A regression model is built for each subset (individual) and RMSECV is calculated, which is considered as a fitness value for each individual. The median of fitness values is considered as a threshold for the selection of an individual. In the second step, the individuals with fitness greater than the median fitness values are retained and all others are discarded. In the third step, the GA breeds between the retained individuals to replace the discarded variables. The genes from two random individuals are split at some random points in the gene. The first part of the gene from the first individual is swapped with the first part of the gene from the second individual, thus it produces two new individuals of hybrid variables. All the new subsets created in the breeding stage are added to the population. In the next stage all the subsets' genes are given a chance for random mutation. After all the subsets have been paired and bred, the population returns to its original size and the process returns again to the fitness evaluation step. The GA will terminate after a finite number of iterations or after some percentage of the individuals in the population are using identical variable subsets [Andersen and Bro, 2010].

The major concern with the application of GA is over fitting, which can lead to improper prediction [Leardi et al., 2002]. In situations where the variables, especially the output variables, are very noisy, the number of samples are very small, or the variables to objects ratio is very high, GA may model the noise instead of information [Leardi and Lupianez Gonzalez, 1998]. Another problem with the GA is that it generates very few variables which explore a very small part of the domain. One has to make several runs to extract the final list of variables. The final model is selected using a step wise approach where the variables are selected based on the frequency

of selection of each variable in all the runs [Leardi and Lupianez Gonzalez, 1998]. Application of GA for the feature selection of PLS in spectral data sets is reported by [Leardi, 2000].

The major limitation of the wrapper approach is that it uses the same target model to select the variables. These methods do not consider the different levels of operation explicitly; therefore, they do not give information on the range of operation, and there is a risk that model may be built only using data from a narrow operating range. A model developed using a certain range of operation data may not work when the process is operated at a different range.

2.1.2 Filter methods

Filter methods use an indirect estimator which solely relies on the properties of the data. These methods are usually used at the pre-processing stage to screen the important variables. For example, the correlation coefficient or signal-to-noise-ratio can be used as a ranking criterion to measure the input-output relationship. Based on the criterion, variables are ranked for selection. This essentially improves the prediction ability and reduces the inclusion of redundant variables.

Variable importance in projection (VIP) is an example of a filter-based method used for PLS, where VIP value provides a combined measure of contribution of a variable in X block in describing the dependent variable in Y block. A VIP value smaller than 1 indicates a non-important variable which can be removed. The main advantage of the VIP method is that it is able to select variables that are important not only for predicting Y, but also for describing X [Andersen and Bro, 2010].

[Hoskuldsson, 2001] proposed a filter-based variable selection method for PLS regression based on correlation coefficient and data intervals. In this approach, first, the squared correlation coefficient for each variable is calculated with the response

variable. Based on the results, the variables showing high correlation are selected. Then, the sample intervals, for which these variables show good results, are selected as training data to build the final model. The method thus selects both variables and the data range for the model.

Filter methods have some advantages over wrapper methods. The main advantage is that the target model is not used during the variable selection process, and thereby, has no influence on the variable selection. As a result, they can be used as a completely separate preliminary step, and give additional confirmation on variable selection. Based on the ranking of the variables, the selected important variables can be used to build the final target model.

Many prediction and monitoring algorithms utilize filter-based variable ranking as a principal or auxiliary selection mechanism because of its simplicity, scalability and good empirical success [Guyon and Elisseeff, 2003]. The main advantage of the filter method is that it deals solely with the properties of data that will be used for building the model. Data quality severely affects all data-based methods. If the data contain noise, it deteriorates the prediction performance. As such, it is always a good idea to select variables which have high signal to noise ratio. Taguchi's experimental design method is a systematic quantitative method which can be used to calculate the signal-to-noise-ratio. The main advantage of Taguchi method is that it allows for the analysis of many different parameters without having a high amount of experimentation. Also, it is straight forward and easy to apply to many engineering situation. In the following section, the method will be described in detail.

2.1.2.1 Taguchi experimental design method

The Taguchi experimental design method is a powerful statistical design approach developed by Dr. Genichi Taguchi for improving product quality and process by re-

ducing variability in the process [Antony and Antony, 2001]. The main objective of this method is to determine the optimal values for factors of a process which have high influence on process improvement. The chosen factors are arranged in an orthogonal array for experimental analysis to determine the signal-to-noise-ratio. Orthogonal arrays are designed to compute the main interaction effects of the factors with the use of a minimum number of experimental trials.

Successful application of the Taguchi method in the automotive, plastics, semiconductors, metal fabrication and foundry industries can be found in [Rowlands et al., 2000]. Improvement in process yield in a chemical process using the method is reported by [Antony and Antony, 2001]. Selection of an optimal set of design parameters to achieve fast convergence speed and network accuracy of a neural network model is described in [Khaw et al., 1995]. Application of Taguchi method in tuning parameters (K_p, K_I, K_D) of a PID controller based on performance index (Integral-Squarred-Error) is reported in [Vlachogiannis and Roy, 2005].

Taguchi experiments are usually conducted on the actual process facilities, rather than in a laboratory situation, which may lead to plant shut-down, and are often uneconomical or impractical specially for a large chemical process [Sukthomya and Tannock, 2005]. However, this situation can be avoided using the retrospective Taguchi method which is based on historical process data. [Sukthomya and Tannock, 2005] used the retrospective Taguchi approach to determine important process parameters of the superplastic forming (SPF) process which creates a design-specified fan blade for an aircraft engine. The maximum aerofoil thickness of the finished blade was considered as the target output of the process. The application of the retrospective Taguchi method was carried out in two different ways. The first approach used a matched data-set for each experiment in the orthogonal array from the historical database. The corresponding aerofoil thickness for that data-set was considered as

the output of the experiment. In the second approach, a neural network model was trained using the retrospective data of the parameters. The network was then used for Taguchi experimentation to emulate the response of the SPF process using the same experimental design set in the array. Finally, based on the signal-to-noise ratio, important parameters for the SPF process that affect the quality of the blade in terms of thickness were selected. For the first method, it is often difficult to fill the orthogonal array using historical data; while the second method can overcome this difficulty as it requires additional work of building a NN model, and the performance of the method will depend largely on the prediction ability of NN model.

2.1.3 Variable selection in SVR

[Rakotomamonjy, 2003] utilizes an idea similar to wrapper based backward selection to rank variables in classification problems. The algorithm starts with all features and repeatedly removes a feature until r features are left or all variables have been ranked. The ranking criteria is derived from support vector machines, and are based on weight vector $||w||^2$ or generalization error bounds sensitivity with respect to a variable. After the removal of a feature, if it minimizes the generalization error, the feature is treated as non-significant, and is thereby removed from the model. In machine learning algorithm like support vector regression (SVR), although variable selection using the wrapper method can be a good alternative to evaluate the selected subset of input variables, it requires a high computational cost [Hand et al., 2000].

For SVR, variable selection can help in reducing dimension. Lowering dimension is important for cases where training data is small. It also improves generalization errors, as irrelevant features cause the performance to deteriorate. Finally, the computational cost, which is a critical factor for online application, will be reduced [Weston et al., 2001]. Feature selection for support vector machines using backward or forward

selection methods are expensive to compute and are time consuming [Weston et al., 2001].

In the case of data having redundant variables, the wrapper method may create different subsets of variables with identical predictive power [Guyon and Elisseeff, 2003]. Therefore, there is a need for the development of new variable selection methods that can deal with correlated data, and give a clear decision on input variables without being computationally too expensive.

2.1.4 Variable selection in PCA

Principal component analysis (PCA) is a dimensionality reduction technique widely used in process monitoring. Even though PCA can handle large data sets, it also suffers from the curse of over fitting and other application difficulties discussed earlier. The need to develop a concise model using PCA is stressed in [Imtiaz et al., 2007]. For a paper mill application, they mainly used process knowledge & signal quality to select variables.

PCA treats the data symmetrically, and it does not divide the data matrix into input and output blocks. For a fault detection model using PCA, it is necessary to select those variables which bear fault signatures. In cases where there is no output, one still needs to select significant variables with respect to a defined criterion. This is called unsupervised variable selection [Guyon and Elisseeff, 2003]. Therefore, the challenge for variable selection in PCA is two-fold. First, a suitable output for the PCA model that can be used for ranking the variables is needed, and, second, a consistent methodology to calculate the contribution of each variable on the output should be developed.

Chapter 3

Variable selection for inferential predictor using retrospective Taguchi method

3.1 Introduction

Inferential predictors or soft sensors are a valuable tool for inferring difficult-to-measure product qualities from real-time process measurements. A key issue in inferential predictor design is the selection of input variables that have the greatest influence on the prediction. This minimizes the complexity of the model and, according to the principle of parsimony, the simplest model that can explain the data well is preferred. Also, from a practical point of view, a concise model is desirable because having a large number of variables in the predictor will increase the probability of bad values in the input variables, which may adversely affect the model's prediction ability. The thesis proposes a new approach for variable selection based on Taguchi's experimental design method. Taguchi (1986) introduced a simplified design of exper-

iment (DoE) approach using an orthogonal array. The method has been widely used in industrial process design, primarily in the development of trials to generate enough process information to establish the optimal conditions for a particular process while keeping the number of experiments to a minimum [Cobb and Clarkson, 1994]. Instead of relying on an arbitrary selection of levels, experiments are conducted following an orthogonal array, as suggested by Taguchi. The method works best at the initial design stage or when experiments can be carried out without upsetting the process. In many systems, conducting experiments is either too costly or simply not possible. An alternative to experimentation is to carry out the analysis using historical data, which is known as the retrospective Taguchi method [Khoei et al., 2002]. The retrospective Taguchi method has been used primarily for selecting important optimizing parameters in the manufacturing industry. In the current research, the method is adapted for selecting important input variables for inferential predictors in the process plant.

The core of the retrospective Taguchi method is to fill in the orthogonal array using historical data. It is often difficult to fill in the orthogonal array using historical process data for several reasons. First, chemical processes are dynamic systems; operators make frequent adjustments to set points whose effects are felt in the process for an extended period of time. Second, processes are typically operated at a narrow range, as such data is not available at all levels. Third, a high degree of correlation between variables is observed in historical process data because of the correlated moves made by the operators. For example, if the feed rate in a reactor goes up, the operator will typically adjust the reactor level in order to maintain the same residence time. Therefore, it will be difficult to find data with a high feed flow rate and low reactor level. The proposed methodology seeks to overcome these challenges.

The modified retrospective Taguchi variable selection method has been used in combination with support vector regression (SVR) for developing an inferential

predictor. SVR is a nonlinear regression method that maps the nonlinear data to a high-dimensional feature space where linear regression is performed. The performance of the proposed retrospective Taguchi variable selection method is compared with the variable importance in projection (VIP) method, which is a variable selection method based on Partial Least Squares (PLS), introduced by Eriksson et al. [Eriksson et al., 2001].

The chapter is divided into the following sections. A brief review of the theoretical framework of the SVR method is given in Section 3.2. Section 3.3 describes in detail the modified retrospective Taguchi method, while Section 3.4 explains the methodology through an industrial case study. The effectiveness of the proposed method in variable selection for SVR is presented in Section 3.5. Finally, Section 3.6 identifies key conclusions.

3.2 Support Vector Regression

Support vector machine is a supervised learning method that can be used for nonlinear regression. The main feature of the SVR methodology is that it possesses good generalization ability of the regression function, robustness of the solution, sparseness of the regression, and an automatic control of the solution complexity [Desai et al., 2006]. The original SVM algorithm was invented by Vladimir N. Vapnik and the current standard soft margin was proposed by Cortes and Vapnik [Cortes and Vapnik, 1995]. Later, a version of SVM for regression known as ε-support vector regression (ε-SVR) was proposed by Drucker et al. [Drucker et al., 1997]. In SVR the lower dimensional input space (x) is transformed into a high-dimensional feature space, F, via a nonlinear mapping and provides an output which is a linear function of the weights and the kernels. The process is carried out by mapping the input data into higher di-

mensional feature space using the kernel trick; a linear regression is then performed in this feature space. The basic characteristic of SVR, which makes it unique from other methods, is that it follows the structural risk minimization (SRM) technique, when other conventional methods follow empirical risk minimization (ERM). ERM does not guarantee a good generalization performance with the resultant model, as it only minimizes error on the training data, while SRM minimizes an upper bound on the expected risk. The SRM feature generalizes the input-output relationship during its training phase and produces an optimized model in such a way that both the prediction error and model complexity are minimized simultaneously. In the following section the mechanism of ε-SVR will be described in detail.

3.2.1 Algorithm of ε-Support Vector Regression

Consider a set of training data points, $[(x_1, y_1), \cdots (x_i, y_i) \cdots, (x_l, y_l)]$, where $x_i \in \mathbb{R}^{l \times n}$ is a feature vector and $y_i \in \mathbb{R}^l$ is the target output. The objective is to find a function, $f(x)$ (Figure 3.1), that has maximum ε deviation from the targets y_i for all training data and at the same time remains as flat as possible. That means, as long as the error is within ε, which is termed an ε-insensitive zone, there is no effort to fit these variations by function, because this ε-insensitive zone should ideally contain the noise in the data. The SVR methodology considers the following estimation function

$$f(x) = \boldsymbol{\omega}^{\boldsymbol{T}}.\boldsymbol{\phi}(x) + \ b \tag{3.1}$$

where $\boldsymbol{\omega}$ denotes the weight vector, b is a constant; $\boldsymbol{\phi}$(x) denotes a function termed feature, and $\boldsymbol{\omega}.\boldsymbol{\phi}(x)$ denotes the dot product in the feature space, F, such that $\boldsymbol{\phi} : x \rightarrow F$, $\boldsymbol{\omega} \in F$. The flatness can be achieved by seeking a small $\boldsymbol{\omega}$.

Under given parameters of the cost function (measuring empirical risk), $C > 0$ and $\varepsilon > 0$, the standard error function of SVR is

$$min\{\boldsymbol{\omega}, b, \xi, \xi^*\} \Longrightarrow \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + (C\sum_{i=1}^{l}\xi_i + C\sum_{i=1}^{l}\xi_i^*) \tag{3.2}$$

subject to

$$\boldsymbol{\omega}^T\boldsymbol{\phi}(x_i) + b - y_i \leq \varepsilon + \xi_i \tag{3.3a}$$

$$y_i - \boldsymbol{\omega}^T\boldsymbol{\phi}(x_i) - b \leq \varepsilon + \xi_i^* \tag{3.3b}$$

$$\xi_i, \xi_i^* \geq 0, \ i = 1, \cdots\cdots, l. \tag{3.3c}$$

To avoid over-fitting and thereby improving generalization capability, Eqn. 3.2 involves summation of empirical risk, and a complexity term in terms of $\boldsymbol{\omega}^2$. By minimizing the objective function in Eqn. 3.2, the SVR optimizes the position of the ε-tube around the data which is shown in Figure 3.1. Eqn. 3.2 penalizes those data points which lie more than ε distance away from the fitted function, $f(x)$. The stated excess positive and negative deviations beyond the ε distance are defined in terms of the slack variables ξ and ξ^* respectively, as shown in Figure 3.1. The slack variable is introduced to allow some flexibility to function $f(x)$ when it is not possible to approximate an input pair with ε precision. The ε-insensitive loss function in Figure 3.1 can be defined by Eqn. 3.4 :

$$L_\varepsilon(y) = \begin{cases} 0 & \text{for } |f(x) - y| \leq \varepsilon \\ |f(x) - y| - \varepsilon & \text{otherwise.} \end{cases} \tag{3.4}$$

During the fitting of the prediction function to the training data, the SVR mini-

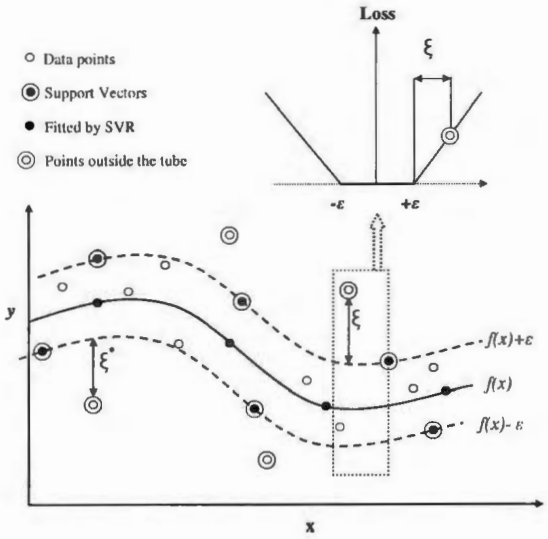

Figure 3.1: A schematic presentation of SVR, along with its ε-insensitive loss function, in which the slope is determined by the value of C [Lahiri and Ghanta, 2008]

mizes the training set error by minimizing both ξ, ξ^* and $||\omega^2||$ in order to increase the flatness of the function or to penalize over-complexity. The prediction accuracy and generalization performance are controlled by two free parameters, C and ε. The cost function parameter C determines the trade-off between the flatness and the tolerance amount of the prediction errors beyond the magnitudes of ε. The model produced by SVR depends on a subset of the training data called support vectors, because any training data that is within the width parameter ε is ignored by the cost function C used in building the model. The tube width parameter ε determines the number of support vectors. As ε decreases, the number of support vectors (SV) increases and thus enhances the risk of model over-fitting and poor generalization. Again, a large ε value produces relatively better generalization performance but provides a high training set error.

According to Vapnik [Vapnik, 1995] [Vapnik, 1998], the solution to the optimization

problem described in Eqn. 3.2 and Eqn. 3.3 is:

$$f(x, \alpha, \alpha^*) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)(\boldsymbol{\phi}(x_i).\boldsymbol{\phi}(x)) + \ b \tag{3.5}$$

where, α_i, $\alpha_i^* \geq 0$, are Lagrange multipliers pertaining to the input data vector, x_i and satisfying $\alpha_i \ \alpha_i^* = 0$, for $i = 1, 2, \cdots\cdots, l$; and the weight vector can be calculated by the following equation:

$$\boldsymbol{\omega} = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\boldsymbol{\phi}(x_i). \tag{3.6}$$

The solution to the optimization problem in Eqn. 3.5 require the computation of a dot product in a feature space, F. This cumbersome computation can be avoided using kernel trick. According to Mercer's theorem, any positive, semi-definite, and symmetric kernel function, K, can be expressed as a dot product in the high-dimensional space. The kernel function is defined in terms of the dot product of the mapping function $\boldsymbol{\phi}$ as shown in Eqn. 3.7:

$$K(x_i, x_j) = \boldsymbol{\phi}(x_i).\boldsymbol{\phi}(x_j) \tag{3.7}$$

The main advantage of this formulation (Eqn. 3.7) is that for many choices of the set $\boldsymbol{\phi_i}(x)$, the form of K is analytically known [Lahiri and Ghanta, 2008]. The kernel function performs all the computations in the input space instead of the feature space. Various kernel functions are available. The most widely used kernel function is the radial basis function (RBF), defined as:

$$K(x_i, x_j) = exp\frac{-||x_i - x_j||^2}{2\sigma^2} \tag{3.8}$$

where σ denotes the width of the RBF. Substituting the dot product in Eqn. 3.5 with

a kernel function, the general form of the SVR-based regression function will be:

$$f(x, \boldsymbol{\omega}) = f(x, \alpha, \alpha^*) = \sum_{i=1}^{l} (\alpha - \alpha^*) K(x, x_i)) + b. \tag{3.9}$$

Here the weight vector $\boldsymbol{\omega}$ is expressed in terms of the Lagrange multipliers, α and α^*. The values of these multipliers are obtained by maximizing the following convex QP problem:

$$\begin{aligned} R(\alpha, \alpha^*) = & -\frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \\ & \times K(x_i, x_j) - \varepsilon \sum_{i=1}^{l} (\alpha_i^* + \alpha_i) \\ & + \sum_{i=1}^{l} y_i (\alpha_i^* - \alpha_i) \end{aligned} \tag{3.10}$$

subject to constraints: $0 \leq \alpha_i \leq C$, $0 \leq \alpha_i^* \leq C$, $\forall_i$, and $\sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0$.
Now a set **S** of support vectors x_s can be found by putting values in the indices 'i' where $0< \alpha<C$ and $\xi_i=0$ (or $\xi_i^*=0$). Using the support vectors, the bias parameter, b in Eqn. 3.9 can be computed as:

$$b = \frac{1}{N_s} \sum_{s \in S} \left[y_s - \varepsilon - \sum_{m \in = S}^{l} (\alpha_m - \alpha_m^*) \boldsymbol{\phi}(x_m) . \boldsymbol{\phi}(x_s) \right]. \tag{3.11}$$

Finally, after getting all the parameters, the approximate function is as shown in Eqn. 3.12:

$$y' = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(x_i, x') + b. \tag{3.12}$$

3.3 Methodology

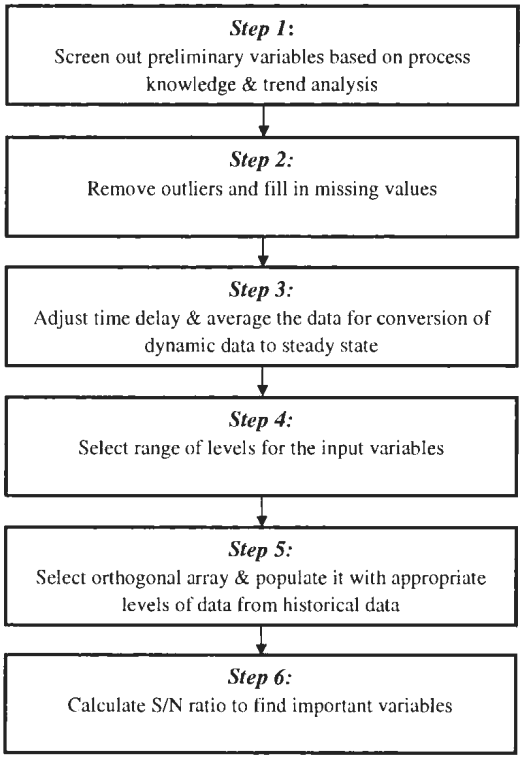

Figure 3.2: Proposed variable selection flow chart

The proposed methodology exploits the merits of Taguchi method to identify important input variables for an inferential predictor. The method initially screens out preliminary variables based on process knowledge and trend analysis. Utilizing process knowledge, an initial list of variables are selected which can have significant contribution to determine the quality variable. Again, trend analysis is useful to select variables based on having significant variation with the quality variable. The methodology has additional preprocessing steps to remove the dynamic effects from the data. Also, the methodology relaxes some of the assumptions of the Taguchi method in order to have sufficient data to fill the orthogonal array. Finally, it applies the Taguchi method to eliminate variables with small S/N ratios. The steps of the

proposed method are shown using a flow chart in Figure 3.2.

Step 1 is the screening of important variables based on prior process knowledge and trend analysis. Through trend analysis, the variation of each variable is observed. Variables which consistently do not show any movement when the quality variable is changing are omitted from the list. This provides a concise list which is used for further analysis.

Step 2 is data preprocessing for outlier removal and filling in missing values. Preprocessing is a crucial step in data analysis, especially for industrial data, as it may contain bad values as a result of process upsets. Also the sensors may contain bias error or variance error. Any outlier or bad data in the data set should be removed since these values can bias the results towards the outliers. This can be done either by visual inspection or using simple rules. Missing values due to outlier removal or slicing of bad data should be filled using appropriate missing data treatment method. For example, use mean of the variable or interpolated values to fill in the missing values.

Step 3 is time delay adjustment and data averaging. Time delay arises mainly from the residence time in vessels and transportation time in pipes. Adjusting the time delay will allow a better capture of correlation in the predictor. After time delay adjustment, data is averaged in order to remove dynamic effects from data. The window for averaging will depend on the dynamics of the system, as well as on the frequency at which the quality variable is available.

Step 4 is the selection of range of each level for all input variables. Two levels are commonly used for analysis [Sukthomya and Tannock, 2005]. The Taguchi method uses a constant value for each level. This is too restrictive for process data. Often it is not possible to match the values from the historical data repository. Instead, a range is assigned for each level. The exact size of the range for the level depends on the range of variation of the variable. Usually, for normal distribution, levels can be

assigned based on the deviation from the mean of the data items. For non-normal distributions, data can be coded by proportion [Sukthomya and Tannock, 2005]. For example, 20%-40% of the range of data can be considered as low level and 60%-80% of the range as high level.

Step 5 is the selection of the appropriate orthogonal array based on the levels and number of variables. Taguchi method utilizes a special design of orthogonal arrays to study the entire process parameter space with a small number of experiments only, and selection of an appropriate orthogonal array depends on the number of levels and parameters used in the analysis [Lin and Lin, 2002]. Orthogonal array is a systematically designed array where each row corresponds to a particular experiment and variables are arranged in columns. Once the array has been selected, it is populated with appropriate values from the data set that fit well with each experimental condition in the array. For highly correlated process variables, this is a challenging task, as the combination of different levels in the orthogonal array may not be available. To overcome this problem, three closely matched measurements are selected from the data-set for each experiment in the array. The target is to keep the average of these three data points within the range of the levels of that particular experiment. Three data points are considered as three trials for each experiment, and the corresponding quality values are considered as the results obtained from the trials.

Step 6 is to calculate the signal-to-noise (S/N) ratio for each experiment utilizing the three trials' results. The S/N ratio is used as a criterion for variable selection. Taguchi has proposed the following three definitions of S/N ratio [Ghani et al., 2004]:

Nominal is the best characteristic: S/N=10 $log\frac{\bar{y}}{s_y^2}$

Smaller the better characteristics: S/N=-10 $\log\frac{1}{n}(\sum y^2)$

Larger the better characteristics: S/N=-10 $\log\frac{1}{n}(\sum \frac{1}{y^2})$

where $\bar{y}$ is the average of observed output data, s_y^2 the variance of y, n the number of observations or trials for each experiment, and y the observed data. Selection of a particular equation depends on the characteristic of the quality variable. For example, to measure the 'liter weight' quality variable of clinker in a clinkerisation process used for determining under-burnt or over-burnt characteristic of produced clinker, 'nominal is the best characteristic' is appropriate. For each type of characteristic described above, the higher the S/N ratio, the more influence the variable has on the experiment. In the current analysis, output is product quality that is a measure of impurity in the product. Therefore, the "smaller the better characteristic" has been used to calculate the S/N ratio.

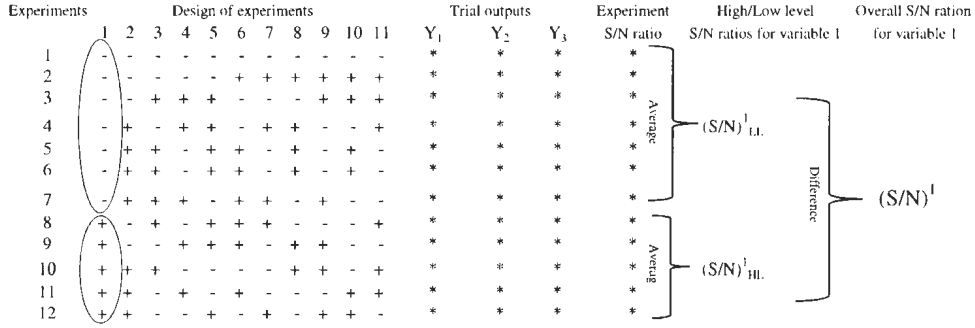

Figure 3.3: Graphical representation of the calculation of S/N ratio

For calculating S/N ration of a particular input variable, the average low and high level S/N ratios are calculated from the corresponding low and high level experiments of that particular variable. The difference of the average low and high level S/N ratios give the final S/N ration of a variable. The calculation of S/N ratio for "Variable 1" is shown using a graphical representation in Figure 3.3. In this case eleven input variables with two levels have been considered. According to Taguchi's design of experiments, a L11 design is appropriate for this case. Design of experiments is shown in the second column, where high level of a variable is denoted by (+) sign and low

level is denoted by (-) sign. Experiments 1 to 7 contain the low level experiments for variable 1 and experiments 8 to 12 contain high level experiments for variable 1. The corresponding trial outputs were used to calculate the high level and low level S/N ratios of variable 1. The difference of the high and low level averages gave the overall S/N ration for variable 1. This process should be repeated for each input variable.

3.4 Industrial Case Study

The proposed variable selection method is validated using data from a petrochemicals plant. An inferential predictor is developed using SVR to predict the product quality. The input variables for the SVR model are selected using the proposed methodology. Because of the proprietary nature of the process, the process details and actual values are withheld in the description.

3.4.1 Data Description

Figure 3.4 shows a process flow diagram of the plant, indicating the sensor locations of the important process variables. The fresh feed, solvent, catalyst, and promoter are fed into the mixing tank. The solvent-to-feed ratio is maintained using a ratio controller. The mixed stream is pumped to the reactor, and air is blown in using a compressor for oxidation. A consecutive oxidation reaction of the form given in Eqn. 3.13 takes place in the reactor.

$$A \longrightarrow B \longrightarrow C \tag{3.13}$$

This is an exothermic reaction. The reactor is operated at constant pressure. The heat of the reaction is removed using a condenser, and the condensed water is recycled back into the system. Part of the condensate is withdrawn from the system. The water

withdrawal rate controls the concentration of reactants inside the reactor. Thus, water withdrawal is an important manipulated variable in the system. About 95% of the achievable conversion takes place in the reactor. Subsequently, reactor effluent is pumped into a series of three crystallizers for secondary reaction and crystallization. In addition to residual reaction, effluent is also depressurized and cooled to the filtering condition in the crystallizer. Air is fed to the first crystallizer for additional reaction of un-reacted feed. Product coming out of the crystallizer undergoes filtering and drying, which complete the process.

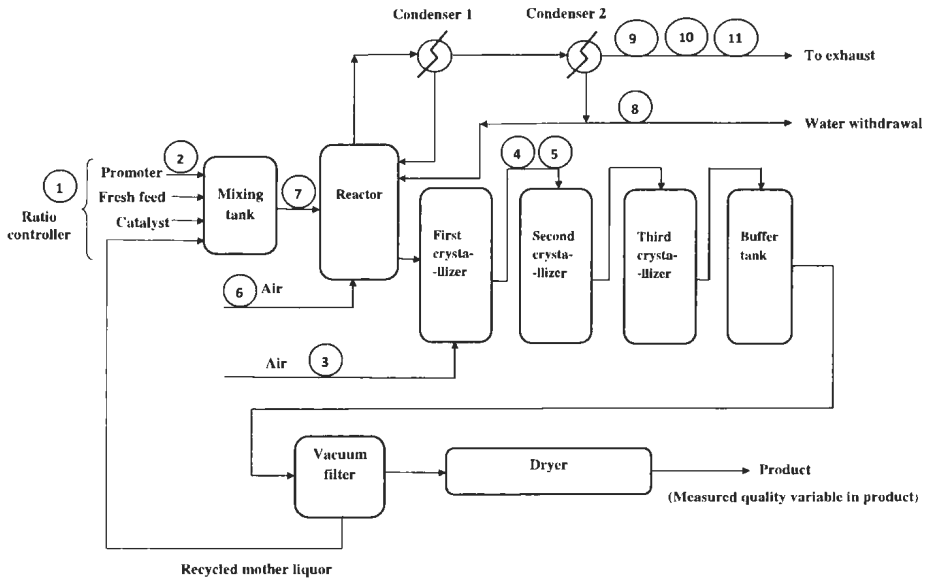

Figure 3.4: Schematic diagram of the process layout

Typically, crude product is analyzed in the lab once or twice a day to ascertain the product quality. The intermediate product B is considered an impurity to the product C. Thus, the measured concentration of B in the product is the quality indicator for the product. The product specification requires the concentration of B to remain below the allowable limit. Keeping the concentration too low consumes more solvent, which is also not desirable from an economic point of view. Thus, the control

objective is to maintain the product quality at the upper allowable limit. However, the measurements of the quality variable were available only twice a day through lab analysis. Therefore an online predictor was required in order to use it as a controlled variable in a model predictive controller (MPC). Data for a period of four months were collected for building the predictor. The process was operated in two different feed levels during the data collection period. Lab analysis of the quality variable was done at four-hour intervals in order to gather sufficient data to build the predictor.

3.4.2 Adjustment for time delay and process dynamic

The quality variable was measured in four-hour intervals while the process variables from the data historian were available at 15-minute intervals. The delay time for the selected process variables with respect to the quality variable varies between 2.5 and 3.5 hours. Therefore, for each lab measurement, the corresponding process data are the process measurements taken 2.5 to 3.5 hours prior to the quality measurement. To account for this time delay, the lab data is lagged by 3 hours. The 4-hour sampling time for the lab data is not suitable to build a dynamic predictor, and the Taguchi method cannot handle the dynamic data; therefore, hourly averages of the process data are taken, which effectively removes the dynamic information from the data. Finally, because product quality is measured at a low frequency, the data matrix has a multivariate structure. Only the complete rows are kept, and all rows which do not contain a quality value are discarded. This is called complete data analysis, which is explained in Table 3.1. For example, a lab measurement is available at 16:00 hrs. The corresponding process data are the averages of process measurements between 12:30 and 13:30 hrs. In this way, a new data matrix is created where only the average values of the process measurements and corresponding quality measurements are retained. Finally, a total of 120 rows are available for building the predictor.

Table 3.1: Complete data analysis

	Variable No.											
Time	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**QV**
12:00	1.28	0.020	2937.7	3.54	5.55	85.26	111.70	31.19	3.33	1.19	0.36	N.M.
12:15	1.28	0.019	2854.9	3.56	5.56	85.27	111.61	32.63	3.43	1.20	0.36	N.M.
12:30	**1.28**	**0.019**	**2805.0**	**3.56**	**5.573**	**85.72**	**111.28**	**32.50**	**3.45**	**1.21**	**0.36**	N.M.
12:45	**1.28**	**0.019**	**2743.9**	**3.48**	**5.58**	**85.83**	**111.77**	**33.30**	**3.38**	**1.22**	**0.37**	N.M.
13:00	**1.28**	**0.019**	**2759.1**	**3.37**	**5.59**	**86.05**	**111.73**	**32.18**	**3.45**	**1.21**	**0.36**	N.M.
13:15	**1.30**	**0.019**	**2810.2**	**3.31**	**5.59**	**85.43**	**111.52**	**30.64**	**3.46**	**1.18**	**0.35**	N.M.
13:30	**1.30**	**0.018**	**2822.3**	**3.23**	**5.60**	**85.73**	**111.75**	**31.1**	**3.44**	**1.19**	**0.35**	N.M.
Average	**1.28**	**0.019**	**2788.1**	**3.39**	**5.59**	**85.75**	**111.61**	**31.95**	**3.44**	**1.20**	**0.36**	
13:45	1.30	0.017	2853.0	3.21	5.60	85.78	111.68	31.55	3.33	1.20	0.36	N.M.
16:00	**1.30**	**0.018**	**2971.5**	**3.31**	**5.59**	**85.39**	**111.38**	**31.19**	**3.40**	**1.20**	**0.36**	**2179**
16:15	1.30	0.018	2968.2	3.33	5.60	85.77	111.73	32.44	3.50	1.19	0.36	N.M.
16:30	**1.32**	**0.018**	**2927.5**	**3.37**	**5.60**	**85.49**	**111.61**	**31.66**	**3.32**	**1.17**	**0.36**	N.M.
16:45	**1.35**	**0.018**	**2954.2**	**3.42**	**5.60**	**87.18**	**111.70**	**31.46**	**3.38**	**1.20**	**0.36**	N.M.
17:00	**1.30**	**0.018**	**2956.1**	**3.46**	**5.60**	**85.42**	**111.54**	**31.21**	**3.34**	**1.21**	**0.37**	N.M.
17:15	**1.26**	**0.018**	**2919.5**	**3.50**	**5.60**	**85.96**	**111.55**	**31.72**	**3.38**	**1.22**	**0.36**	N.M.
17:30	**1.24**	**0.018**	**2914.2**	**3.51**	**5.60**	**85.74**	**111.77**	**34.10**	**3.33**	**1.17**	**0.35**	N.M.
Average	**1.29**	**0.018**	**2934.3**	**3.45**	**5.60**	**85.96**	**111.63**	**32.03**	**3.35**	**1.19**	**0.36**	
19:45	1.30	0.018	2939.9	3.33	5.61	85.59	111.76	30.87	3.33	1.20	0.36	N.M.
20:00	**1.30**	**0.018**	**2943.4**	**3.34**	**5.61**	**86.52**	**112.14**	**34.38**	**3.29**	**1.16**	**0.36**	**2103**

(a) Original data matrix for averaging data

QV=Quality Variable N.M.=Not Measured

Data matrix after complete data analyis

16:00	**1.28**	**0.019**	**2788.1**	**3.39**	**5.59**	**85.75**	**111.61**	**31.95**	**3.44**	**1.20**	**0.36**	**2179**
20:00	**1.29**	**0.018**	**2934.3**	**3.45**	**5.60**	**85.96**	**111.63**	**32.03**	**3.35**	**1.19**	**0.36**	**2103**

(b) New data matrix

Table 3.2: List of variables used in the Taguchi analysis

No.	Variable name	Description
1	Ratio controller	Control of [(solvent+catalyst+promoter)/feed] ratio
2	Promoter flow	Promoter supplied to feed preparation drum
3	Air (Crystallizer)	Air supplied to first crystallizer
4	O_2% (Crystallizer)	O_2% measured in crystallizer-1 outlet gas stream
5	CO_2% (Crystallizer)	CO_2% measured in Crystallizer-1 outlet gas
6	Air (Reactors)	Air supplied to the reactor
7	Feed rate	Feed supplied to the reactor through pump
8	Condensate flow	Condensate withdrawal from condenser-2 bottom
9	O_2% (Condenser)	O_2% measured in condenser-2 outlet gas stream
10	CO_2% (Condenser)	CO_2% measured in condenser-2 outlet gas stream
11	CO% (Condenser)	CO% measured in condenser-2 outlet gas stream

3.4.3 Variable selection by retrospective Taguchi method

Table 3.3: Selection of levels of data for the Taguchi orthogonal experiment using historical data

No.	Variable	Min	Max	Mean	Range	Limit	LL	HL
1	Ratio controller	1.2	2.05	1.56	0.85	0.34	1.54	1.71
2	Promoter flow	0.002	0.028	0.017	0.026	0.010	0.013	0.018
3	Air (Crystallizer)	1514	4024	72940	2510	1004	2518	3020
4	O_2% (Crystallizer)	1.90	3.66	3.28	1.75	0.70	2.61	2.96
5	CO_2% (Crystallizer)	5.23	6.13	5.71	0.89	0.35	5.59	5.77
6	Air (Reactors)	111.2	178.9	165.1	67.6	27.1	138.3	151.8
7	Feed rate	145.8	234.7	216.7	88.9	35.5	181.4	199.2
8	Condensate flow	19.4	48.7	39.8	29.3	11.7	31.1	36.9
9	O_2% (Condenser)	6.39	7.96	7.03	1.57	0.62	7.02	7.34
10	CO_2% (Condenser)	2.23	3.47	2.64	1.24	0.49	2.73	2.98
11	CO% (Condenser)	0.665	0.971	0.749	0.306	0.122	0.787	0.849

LL=Lower level limit ($\leq$), HL=Higher level limit($\geq$), Limit=40% of range value

Table 3.2 shows the list of the variables used in the analysis. The next step is to select the levels for the orthogonal array. A range, rather than a specific value, is assigned for each level. From the available data, for each variable, the lowest forty percent values are considered as low level and the highest forty percent values are

considered as high level. Table 3.3 reports ranges of the levels of different variables. For example, the "recycle ratio" has a range of 0.85; 40% of this range gives the limit for the variable which is 0.34. Adding this limit value (0.34) to its minimum value (1.2) gives a lower level limit (1.54), and subtracting it from the maximum value (2.05) gives higher level limit (1.71). Therefore, values falling in the range of 1.2–1.54 are considered low level and values between 1.71 and 2.05 are categorized as high level.

Table 3.4: Taguchi orthogonal array with low-level and high-level values of eleven variables to design the experiment

	Variable number										
Exp	1	2	3	4	5	6	7	8	9	10	11
1	**1.54**	**0.013**	**2518**	**2.61**	**5.59**	**138.3**	**181.4**	**31.1**	**7.02**	**2.73**	**0.787**
2	**1.54**	**0.013**	**2518**	**2.61**	**5.59**	151.8	199.2	36.9	7.34	2.98	0.849
3	**1.54**	**0.013**	3020	2.96	5.77	**138.3**	**181.4**	**31.1**	7.34	2.98	0.849
4	**1.54**	0.018	**2518**	2.96	5.77	**138.3**	199.2	36.9	**7.02**	**2.73**	0.849
5	**1.54**	0.018	3020	**2.61**	5.77	151.8	**181.4**	36.9	**7.02**	2.98	**0.787**
6	**1.54**	0.018	3020	**2.61**	5.77	151.8	**181.4**	36.9	**7.02**	2.98	**0.787**
7	**1.54**	0.018	3020	2.96	**5.59**	151.8	199.2	**31.1**	7.34	**2.73**	**0.787**
8	1.71	**0.013**	3020	**2.61**	5.77	151.8	199.2	**31.1**	**7.02**	**2.73**	0.849
9	1.71	**0.013**	**2518**	2.96	5.77	151.8	**181.4**	36.9	7.34	**2.73**	**0.787**
10	1.71	0.018	3020	**2.61**	**5.59**	**138.3**	**181.4**	36.9	7.34	**2.73**	0.849
11	1.71	0.018	**2518**	2.96	**5.59**	151.8	**181.4**	**31.1**	**7.02**	2.98	0.849
12	1.71	0.018	**2518**	**2.61**	5.77	**138.3**	199.2	**31.1**	7.34	2.98	**0.787**

The next step is to select the appropriate orthogonal array. An orthogonal array is a matrix which represents the condition of factors in a series of experiments. For eleven variables and two levels, the suggested orthogonal array is L11. Table 3.4 shows the orthogonal array where each row corresponds to an experiment. The eleven variables considered as eleven factors are arranged in column direction. The array is designed for two levels. The low level data appear in bold face and the rest are the high level data. The next step is to find the appropriate combination of these twelve experiments from the data set. This is a challenging task because of the correlation

among the process variables. On the other hand, the Taguchi method assumes variables are uncorrelated, so the array has a standard combination of different levels. To overcome this problem, three values are selected from the data set for each experiment in the array. The measurements are selected in such a way that the average falls within the range of the levels. In the event the value is still outside the limit, an additional 10% margin for the limit is allowed. Using the above methodology, from a total of thirty-six rows of measurements, the twelve experiments for the array are constructed.

Based on the trial results, the signal-to-noise ratio is calculated for each experiment. The quality variable for this particular process is an intermediate product which in turn determines the product quality. The minimum of this value is desired; therefore, "the smaller the better" criterion is appropriate in this case. Table 3.5 shows the calculated signal-to-noise ratios based on the trial results for all experiments. Table 3.6 reports the calculation steps of S/N ratio of each variable with explanation. Variables with larger S/N ratios are considered more important variables.

3.5 Results and Discussion

3.5.1 PLS model

To validate the importance of the eleven initially-screened variables, first a linear PLS model is developed using all eleven variables. Figure 3.5 shows the prediction of both the training and test data sets. The model captured 74.14% variance with three latent variables. The figure clearly indicates that the eleven variables which are selected based on process knowledge are significant in predicting the quality output.

Table 3.5: Calculation of S/N ratio of each experiment in the array from the output quality variable value obtained from each trial data

	Y_i[a]= Output Q.V.				
Exp	Y_1	Y_2	Y_3	$\sum_{i=1}^{3} Y_i^2$	S/N Ratio [b]
1	2006.72	1860.24	2607.82	14288163.98	-66.78
2	1907.87	2002.57	1135.32	8939187.47	-64.74
3	2604.77	2367.88	1393.72	14334113.80	-66.79
4	2386.46	1532.71	2008.05	12076650.55	-66.05
5	1879.88	2003.74	2125.69	12067493.53	-66.04
6	2672.39	2072.05	2118.76	15924205.10	-67.25
7	2296.64	2595.03	2049.69	16210027.66	-67.33
8	2002.57	2365.35	2216.91	14519878.58	-66.85
9	2078.00	1960.33	2142.21	12750022.69	-66.28
10	2141.45	2079.73	2149.94	13533348.86	-66.54
11	2010.96	977.43	2134.29	9554510.00	-65.03
12	1900.15	2089.13	2071.14	12264675.05	-66.12

[a]Output quality variable obtained from three trials data for each experiment

[b]S/N ratio=-10 $\log \frac{1}{N}(\sum_{i=1}^{3} Y_i^2)$, where N=3

Table 3.6: Calculation of S/N ratio for each variable

Variable No.	Variable	LLC[a]	HLC [b]	S/N ratio [c]
1	Ratio controller	-66.43	-66.16	0.26
2	Promoter flow	-66.29	-66.34	0.05
3	Air (Crystallizer)	-65.83	-66.80	0.97
4	O_2% (Crystallizer)	-66.33	-66.30	0.04
5	CO_2% (Crystallizer)	-66.08	-66.48	0.40
6	Air (Reactors)	-66.46	-66.22	0.24
7	Feed rate	-66.39	-66.22	0.17
8	Condensate flow	-66.48	-66.15	0.33
9	O_2% (Condenser)	-66.33	-66.30	0.03
10	CO_2% (Condenser)	-66.64	-66.00	0.64
11	CO% (Condenser)	-66.63	-66.00	0.63

[a]'Low level contribution' for each variable is calculated as average of S/N ratios of those experiments in the orthogonal array where the variable is contributing as low level

[b]'High level contribution' for each variable is calculated as average of S/N ratios of those experiments in the orthogonal array where the variable is contributing as high level

[c]Absolute difference between HLC and LLC

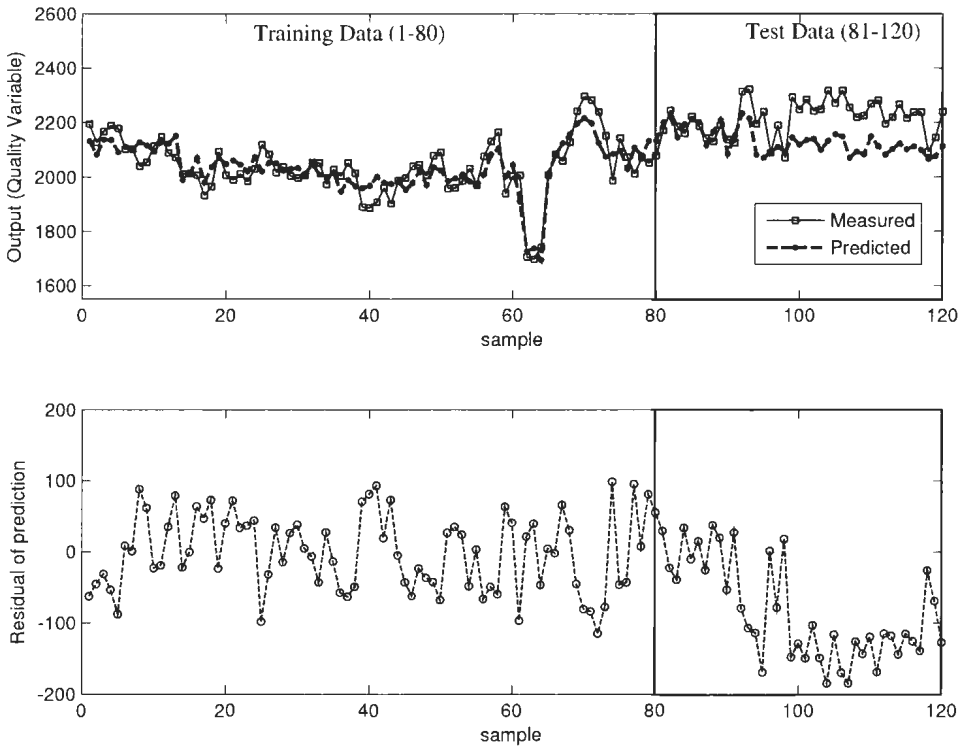

Figure 3.5: Data prediction by PLS model using all eleven variables

3.5.2 VIP method

The proposed retrospective Taguchi variable selection method is compared with the variable importance in projection (VIP) method in the context of SVR model. The VIP score of a predictor variable is a summary of the importance of the projections to find the latent variables. The VIP value is a weighted sum of squares of the PLS weights and therefore, it explains the variance of each PLS dimension. VIP scores demonstrate the importance of each variable, so they are often used for variable selection. Usually the average of squared VIP scores equals one. A score value greater than one is used as variable selection criteria. Table 3.7 shows the rank of the variables based on the VIP scores.

Table 3.7: Comparison of rank of variables

Variables arranged in order of importance (based on process knowledge)	Calculated S/N ratio	Rank (Taguchi)	Calculated VIP value	Rank (VIP)
Ratio controller	0.26	**6**	1.30	**2**
CO_2% (Condenser)	0.64	**2**	1.02	**7**
Air (Reactors)	0.24	**7**	1.08	**5**
CO% (Condenser)	0.63	**3**	1.26	**3**
Air (Crystallizer)	0.97	**1**	1.89	**1**
CO_2% (Crystallizer)	0.40	**4**	0.16	11
Condensate flow	0.33	**5**	1.15	**4**
Promoter flow	0.05	9	0.68	8
O_2% (Crystallizer)	0.04	10	0.59	9
Feed rate	0.17	8	1.07	**6**
O_2% (Condenser)	0.03	11	0.44	10

3.5.3 Selected variables by Taguchi method

In Table 3.7, variables are arranged in order of importance based on process knowledge. The ranks of the variables based on S/N ratio calculated using Taguchi method are given in column 3. From the ranks, it is evident that the seven important variables selected by the Taguchi method are also important based on process knowledge.

The most important variable based on the analysis is "air supplied to first crystallizer." Air is added to the crystallizer for additional conversion of intermediate product to final product. Therefore, this variable has a direct link to the product quality and will make the most significant contribution. The second and third key variables are "CO_2 and CO in the gas stream going out from condenser-2." Both CO_2 and CO are by-products produced during the reaction and directly reflect reaction extent. The next important variable is "CO_2 measured in first crystallizer outlet," which directly measures the additional reaction taking place in the crystallizer. It is less important than the air supplied to the crystallizer because only a small portion of the reaction (<5%) takes place in the crystallizer. The next important variable is "condensate

flow withdrawn from condenser-2." This is an important variable as it controls the rate of condensate flow re-fluxed to the reactor, which in turn controls the water concentration in the reactor. The rate of reaction is proportional to concentration of reactant. The next selected variable is "ratio of solvent, catalyst, and promoter with respect to fresh feed." The ratio has a significant effect, as it controls the catalyst concentration within the reactor. Finally, the variable "air supplied to the reactor" directly effects the oxidation reaction.

3.5.4 Comparison of predictions between Taguchi-SVR and VIP-SVR

Based on the analysis, seven variables are selected using VIP and another set of seven variables are selected using Taguchi approach. The selected variables are used to develop two separate models using the ε-support vector regression method. The SVR algorithm uses a radial basis function (RBF) as kernel. SVR model is developed using PLS Toolbox software [Wise et al., 2007].

Figures 3.5, 3.6, and 3.7 show the predictions by the PLS, Taguchi-SVR, and VIP-SVR approaches respectively. A comparison of Figures 3.6 and 3.7 with Figure 3.5 clearly demonstrates that SVR gives better prediction than PLS. Again, if Figure 3.6 is compared with Figure 3.7, it can be stated that variables selected by Taguchi method has shown good prediction performance compared to the VIP method. The residuals are also plotted in each figure to show the dynamic variation and bias in prediction. It is clearly observed that the Taguchi method captures the dynamic variation well and shows less bias than the VIP method. The model performance is also measured quantitatively by the root-mean-square-error (RMSE). Figure 3.8 shows the RMSE value of the training, validation, and test sets for each case. It clearly shows that the Taguchi-SVR method has less prediction error than the VIP-SVR method.

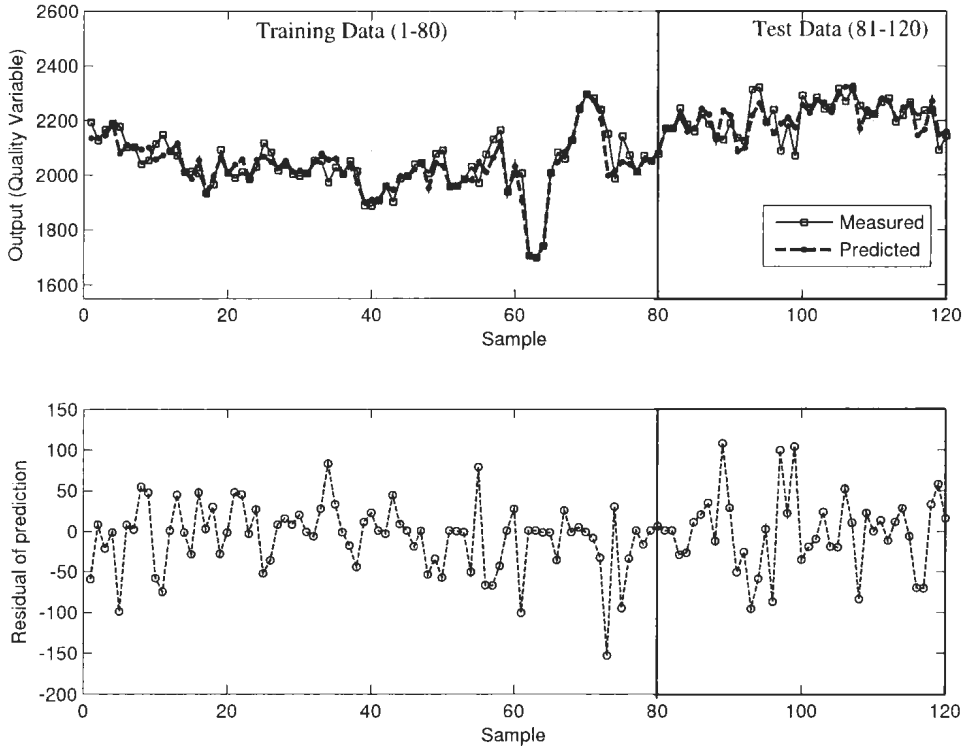

Figure 3.6: Data prediction by SVR model using variables selected from Taguchi approach

3.6 Conclusions

In a typical chemical process, on average 1500 process variables are measured at any given time. Of these, it is difficult to select the most important variables for building an inferential predictor. The proposed method offers a systematic quantitative approach to selecting the most important variables. The method is based on Taguchi's experimental design. However, rather than performing new experiments, the developed method employs historical process data to select important variables for an inferential predictor. The technique can be carried out relatively cheaply and without any process disruption, as it uses only historical data.

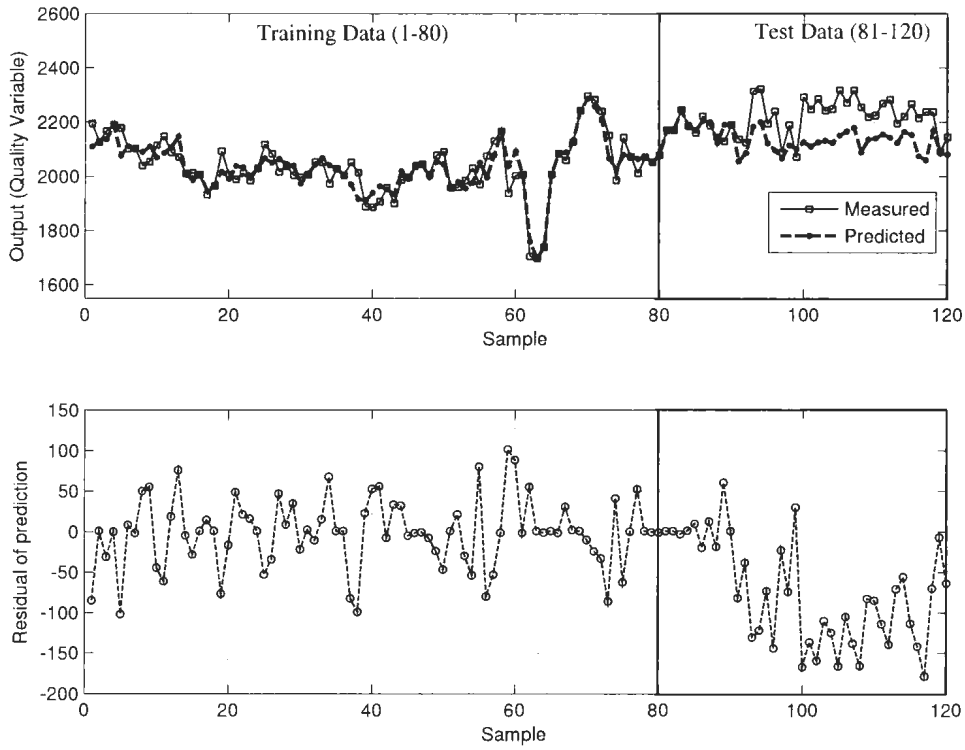

Figure 3.7: Data prediction by SVR model using variables selected from VIP approach

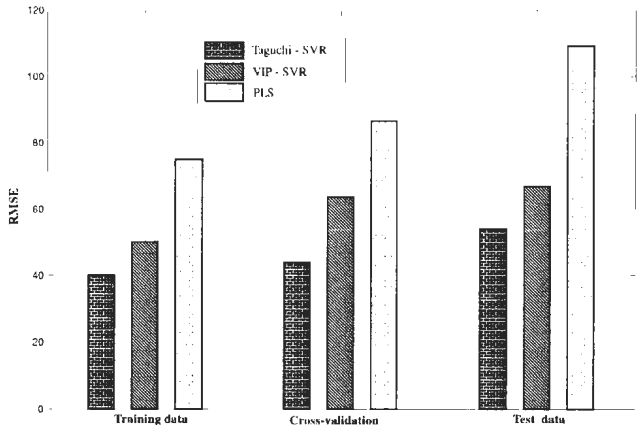

Figure 3.8: Comparison of RMSE value for training, cross-validation and test data

The proposed Taguchi-based methodology is implemented in combination with SVR for building an inferential predictor for a petrochemical process. The method

is compared with VIP method. The Taguchi-SVR method shows significantly better performance in predicting the product quality than the VIP-SVR method.

Implementation difficulties arising from the dynamic changes in the process variables have been addressed. The most significant implementation difficulty is to find orthogonal experiments between the correlated variables. It makes it difficult to fill the orthogonal array in general. This issue is further investigated in Chapter 4. Despite the limitation, the case study illustrates that the technique can be effectively used for variable selection utilizing historical data. It has the ability to identify important variables which can improve the prediction ability of a soft sensor.

Chapter 4

Selection of input variables for inferential predictor from a large set of correlated variables

4.1 Introduction

In Chapter 3, a systematic quantitative approach of variable selection for an inferential predictor using retrospective Taguchi method is explained. It is observed that in the presence of correlated variables, it becomes difficult to fill in the orthogonal array using historical data. Process variables are usually highly correlated, making it difficult to fill the designed orthogonal array of the Taguchi method using historical data. The orthogonal array is designed with different combinations of low and high levels of the factors. In a situation, where two variables are positively correlated, it would be difficult to get values having one at low level and another at high level. For example, if the feed rate in a reactor increases, the reactor level is also raised to keep the residence time the same. As such, in the historical data a high reactor level and low feed rate

combination is typically not available. In the present chapter, this issue is further investigated and a methodology is developed to resolve this issue in a systematic way. The present chapter aims to develop a methodology to select important variables for an inferential predictor from a large set of correlated variables. The proposed method is a combination of filter and wrapper approach to deal with correlated variables. Grouping correlated variables based on their correlation coefficients is at the core of the method. The proposed variable selection method has been used in combination with support vector regression (SVR) for developing an inferential predictor. The performance of the proposed method is compared with the variable importance in projection (VIP) method, which is a variable selection method based on partial least squares (PLS).

The chapter is divided into the following sections. Section 4.2 describes in detail the methodology of variable selection from a large set of correlated variables. Section 4.3 explains the methodology through an industrial case study. Finally, in Section 4.4 key conclusions are described.

4.2 Methodology

The proposed methodology is a hybrid of the retrospective Taguchi method and wrapper method. It uses a correlation-based classifier to group highly correlated variables into different uncorrelated groups. Then, it applies the Taguchi method to eliminate groups of variables with small S/N ratios. Subsequently, it uses a wrapper method to select variables within the group. The key elements of the methodology are described below:

4.2.1 Classification of variables

Correlation matrix is calculated from the available training data. The correlation matrix of n variables X $(n \times n)$ is a $n \times n$ symmetric matrix whose i, j entry is $corr(X_i, X_j)$. Following the calculation of the correlation matrix, variables are re-ordered and grouped based on the correlation. A modified k-nearest neighbor algorithm (MKNN) is applied to group the variables. Detailed theory of the MKNN algorithm can be found in [Parvin et al., 2010]. Matlab built-in function 'corrmap.m.' is used to carry out the above calculation. It produces a pseudo-color map which shows the correlation between variables in a data-set after rearranging the correlated variables in groups. This gives a visual way of classifying large data-set of variables into several uncorrelated groups.

4.2.2 Selection of important groups

Following the classification of variables into groups, a representative variable is selected from each group. For example, if the classification algorithm produces n groups, (n) representative variables will be selected one from each group. Then, the retrospective Taguchi method is applied to the selected variables to find important variables. The advantage of using a single variable from each group is that these variables are relatively uncorrelated with each other; therefore, it is easier to fill the orthogonal array considering these variables as factors.

The retrospective Taghchi method for variable selection has three steps. The first step is the selection of levels for all representative variables. Two levels are commonly used for analysis [Sukthomya and Tannock, 2005]. The Taguchi method uses a constant value for each level. This is too restrictive for process data. Often it is not possible to match the values from the historical data repository. Instead, a range is

assigned for each level. The exact size of the range for a level depends on the range of variation of the variable. The second step is the selection of an appropriate orthogonal array based on the levels and number of variables. An orthogonal array is a systematically designed array where each row corresponds to a particular experiment and variables are arranged in columns. Once the array has been selected, it is populated with appropriate values from the data set that matches closely with each experimental condition in the array. As mentioned previously, because these representative variables come from different groups, as such they are relatively uncorrelated and makes it easier to fill the orthogonal array using historical data. Three data points are usually considered as three trials for each experiment, and the corresponding quality values are considered as the results obtained from the trials. In the final step, the three trials' results are utilized to calculate the signal-to-noise (S/N) ratio for each experiment.

In an orthogonal array, a variable is kept systematically in either low or high level in an experiment. The low level or high level contribution of a variable is the S/N ratio of the experiment, where it acts as either a low level or high level, respectively. After calculating the S/N ratio of each experiment, the S/N ratio for each variable is calculated from the difference of its average low level contribution and average high level contribution in all the experiments. Based on the S/N ratio, a decision is made as to whether a particular variable is important or not. Since each variable is representing a group, if a variable is deemed important, that means that all the variables belonging to that particular group are considered important. In the next step, important variables are selected from within the selected groups.

4.2.3 Selection of variables from within groups

The selection of variables from within the group is done using a wrapper based method. SVR models are built with all variables from the selected groups. Model performance

is evaluated by the root mean square error (RMSE) of cross-validation data or prediction data or a combination of both. A systematic backward elimination approach is followed.

First, a model is developed using all variables of the selected groups. Variables are subsequently eliminated one at a time from each group. If the model performance deteriorates, the eliminated variable is considered important and re-introduced in the model. After testing all the variables from a group, the same exercise is carried out on the variables belonging to other groups. The process terminates after all the groups with multiple variables have been tested.

The steps of the proposed method are shown using a flow chart in Figure 4.1. Step 1 is the screening of important variables based on prior process knowledge and trend analysis. This provides a concise list which is used for further analysis. Step 2 is data preprocessing for outlier removal and filling in missing values. Preprocessing is a crucial step in data analysis, especially for industrial data, as it may contain bad values as a result of process upsets. Also the sensors may contain bias error or variance error; therefore, it is important to validate the measurements before further analysis. Step 3 is time delay adjustment and data averaging. Time delay arises mainly from the residence time in vessels and transportation time in pipes. Adjusting the time delay will allow for a better capture of correlation in the predictor. After time delay adjustment, data is averaged in order to remove dynamic effects from data. The window for averaging will depend on the dynamics of the system as well as the frequency at which the quality variable is available. Step 4 is grouping correlated variables based on the correlation color map. Step 5 is selection of important groups from the list of groups created in Step 4. The retrospective Taguchi algorithm is applied to select the groups. Step 6 &7 are to apply a systematic backward elimination approach using SVR to eliminate least contributing variables from within groups. First, an SVR

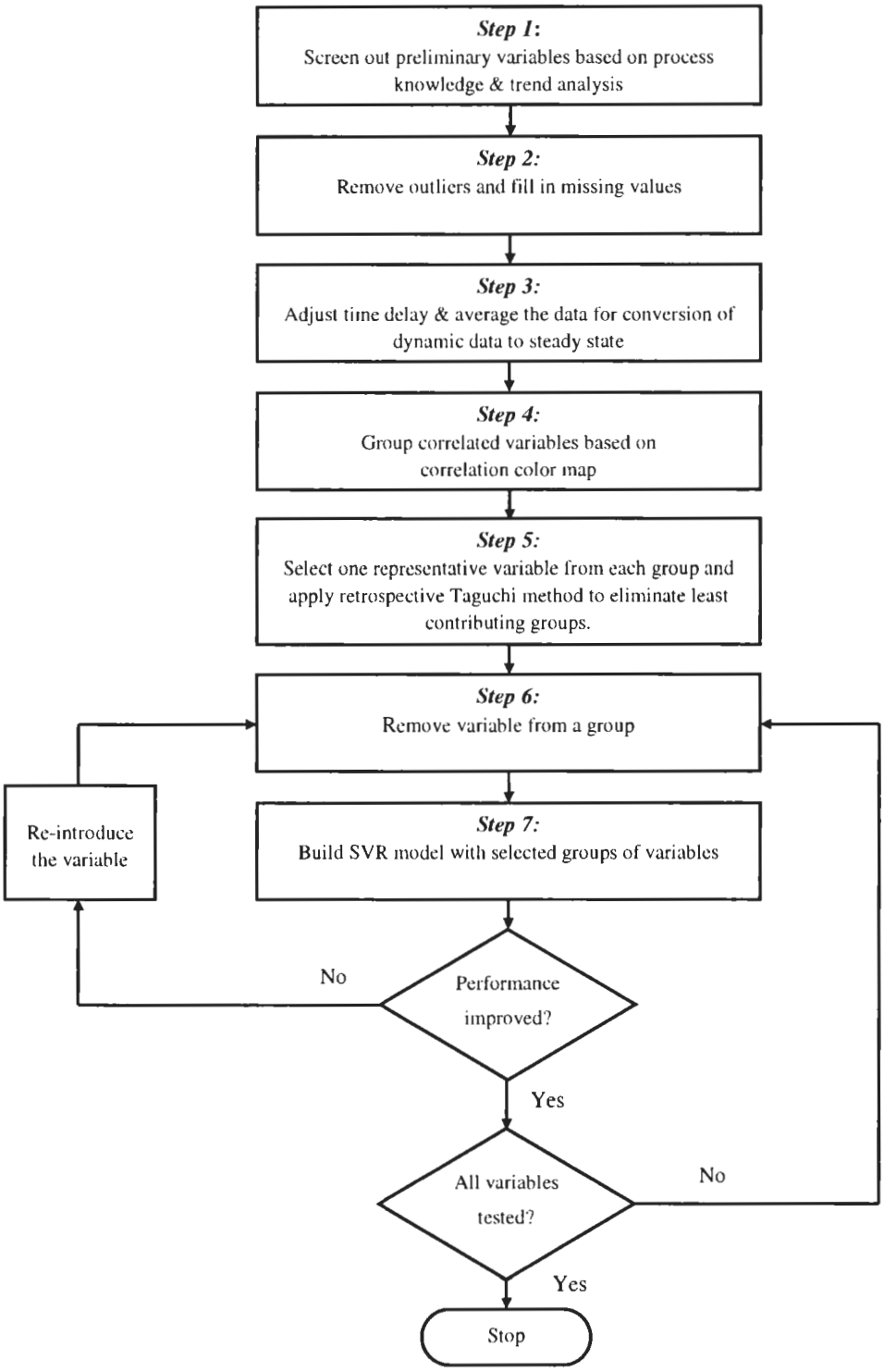

Figure 4.1: Proposed variable selection flow chart for correlated variables

model is built using all the selected groups of variables from Step 5. A variable is removed from a selected group one at a time, and, subsequently, an SVR model is built. If prediction performance is not improved, the variable is considered important and re-introduced in the model. The process terminates after all the variables are tested once.

4.3 Results and Discussion

The list of preliminarily selected variables used as input to the variable selection process are shown in Table 4.1. The classification algorithm was used to classify these variables into different groups.

Table 4.1: List of variables used in the Taguchi analysis

No.	Variable name	Description
1	Ratio controller	Control of [(solvent+catalyst+promoter)/feed] ratio
2	Promoter flow	Promoter supplied to feed preparation drum
3	Air (Crystallizer)	Air supplied to first crystallizer
4	O_2% (Crystallizer)	O_2% measured in crystallizer-1 outlet gas stream
5	CO_2% (Crystallizer)	CO_2% measured in Crystallizer-1 outlet gas
6	Air (Reactors)	Air supplied to the reactor
7	Feed rate	Feed supplied to the reactor through pump
8	Condensate flow	Condensate withdrawal from condenser-2 bottom
9	O_2% (Condenser)	O_2% measured in condenser-2 outlet gas stream
10	CO_2% (Condenser)	CO_2% measured in condenser-2 outlet gas stream
11	CO% (Condenser)	CO% measured in condenser-2 outlet gas stream

4.3.1 Grouping variables using correlation color map

Figure 4.2 shows the correlation color map of eleven variables, where variables are grouped by correlation. The map clearly identifies six groups as reported in Table 4.2. Group 2 has five correlated variables. Air (Reactors) is added to the reactors for oxidation reaction and air (Crystallizer) is added to the crystallizer for the remaining

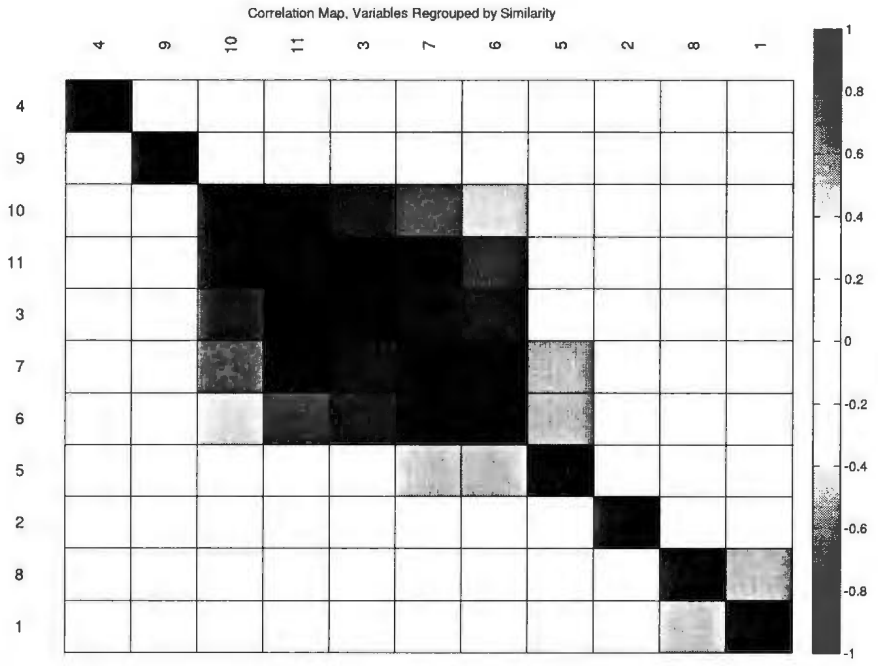

Figure 4.2: Correlation color map for grouping variables based on correlation matrix

conversion. These two variables are therefore correlated with the Feed rate. Again, $CO_2\%$ (Condenser) and $CO\%$ (Condenser) are reaction bi-products which are essentially a measure of the reaction rate. These two variables are inversely correlated with the feed rate and air added to the reactors. An increase in feed rate reduces residence time and affects reaction conversion. This leads to a decrease in reaction bi-products. All these correlation are reflected in Group 2. Group 6 has two correlated variables. Both the ratio controller and condensate flow maintain the desired solvent concentration in the reactor. As such, they are expected to be correlated.

4.3.2 Group selection using the retrospective Taguchi method

In this step, important groups are selected by applying the retrospective Taguchi method. First, a representative variable is selected from each group. Table 4.2 reports the selected variables from each group. Next, ranges of low and high levels are assigned

Table 4.2: Grouping variables based on correlation matrix

Variable No.	Variable name	Group	Selected variable [a]
2	Promoter flow	1	2
3	Air (Crystallizer)	2	3
6	Air (Reactors)		
7	Feed rate		
10	CO_2% (Condenser)		
11	CO% (Condenser)		
9	O_2% (Condenser)	3	9
5	CO_2% (Crystallizer)	4	5
4	O_2% (Crystallizer)	5	4
8	Condensate flow	6	8
1	Ratio controller		

[a]Representative variable selected from each group for using in the variable selection process

for each variable based on the available data. Table 4.3 reports the ranges of the levels of the six representative variables. For each variable, the twenty to forty percent range is considered as low level and the sixty to eighty percent range is considered as high level.

Next step is to select the appropriate orthogonal array for Taguchi experimental design. An orthogonal array is a matrix which gives the levels of factors in a series of experiments. For six variables and two levels, the suggested orthogonal array is L8. Table 4.4 shows the orthogonal array where each row corresponds to an experiment, and the six variables, considered as six factors, are arranged in column direction. The low level data are denoted by bold face and the rest are the high level data. The next step is to find the appropriate combination of these eight experiments from the data set. For each experiment in the array, three closely matched values are selected from the data set that fall within the defined range (as shown in Table 4.4). The measured quality variables of these three data points are considered as the output of the three

Table 4.3: Selection of levels of data of group variables for the Taguchi orthogonal experiment using historical data

No.	Variable	Min	Max	Range	LLR[a]	HLR [b]
2	Promoter flow	0.002	0.028	0.026	0.007-0.013	0.018-0.023
3	Air (Crystallizer)	2207.1	3819.5	1612.3	2529.6-2852.1	3174.5-3497.0
9	O_2% (Condenser)	6.48	7.52	1.04	6.69-6.90	7.10-7.31
5	CO_2% (Crystallizer)	5.39	6.01	0.62	5.51-5.64	5.76-5.88
4	O_2% (Crystallizer)	3.12	3.66	0.54	3.23-3.33	3.44-3.55
8	Condensate flow	23.49	48.40	24.92	28.47-33.45	38.44-43.42

[a]Low level data range is from (min + 20% of range value) to (min + 40 % of range value)
[b]High level data range is from (min + 60% of range value) to (min + 80 % of range value)

trials conducted for each experiment. Based on these three output values, the signal-to-noise ratio for each experiment is calculated. According to Taguchi's experimental design method, calculation of S/N ratio differs based on the nature of the output variables [Sukthomya and Tannock, 2005]. The quality variable for this particular process is an intermediate product which in turn determines the product quality. The minimum of this value is desired; therefore, "the smaller the better" criterion is appropriate in this case. Table 4.5 reports the output values of three trials and the calculation of the S/N ratio for each experiment. Next step is to calculate average S/N ratio for each variable. To calculate this, first the average S/N ratios for the high levels and low levels of each variable were calculated. The difference between these two average S/N ratios gives the final S/N ratio for a particular variable. Table 4.6 reports the calculated values of the S/N ratio of each variable with an explanation. Variables with larger S/N ratios are considered more important variables.

Table 4.4: Taguchi orthogonal array with low-level and high-level range values of six group variables to design the experiment

	Variable number					[a]
Exp.	2	3	9	5	4	8
1	**0.007-0.013**	**2529.6-2852.1**	**6.69-6.90**	**5.51-5.64**	**3.23-3.33**	**28.47-33.45**
2	**0.007-0.013**	**2529.6-2852.1**	**6.69-6.90**	5.76-5.88	3.44-3.55	38.44-43.42
3	**0.007-0.013**	3174.5-3497.0	7.10-7.31	**5.51-5.64**	**3.23-3.33**	38.44-43.42
4	**0.007-0.013**	3174.5-3497.0	7.10-7.31	5.76-5.88	3.44-3.55	**28.47-33.45**
5	0.018-0.023	**2529.6-2852.1**	7.10-7.31	**5.51-5.64**	3.44-3.55	**28.47-33.45**
6	0.018-0.023	**2529.6-2852.1**	7.10-7.31	5.76-5.88	**3.23-3.33**	38.44-43.42
7	0.018-0.023	3174.5-3497.0	**6.69-6.90**	**5.51-5.64**	3.44-3.55	38.44-43.42
8	0.018-0.023	3174.5-3497.0	**6.69-6.90**	5.76-5.88	**3.23-3.33**	**28.47-33.45**

[a]Bold faced-low level data range ; normal-high level data range

Table 4.5: Calculation of S/N ratio of each experiment in the array from the output quality variable value obtained from each trial data

	Y_i[a]= Output Q.V.				
Exp	Y_1	Y_2	Y_3	$\sum_{i=1}^{3} Y_i^2$	S/N Ratio [b]
1	1929.97	1949.96	1945.31	11311354.4	-65.76
2	2052.34	2076.57	2151.66	13153905.0	-66.42
3	2163.03	2171.35	2283.30	14606938.8	-66.87
4	2028.17	2067.56	2052.97	12602974.8	-66.23
5	2082.31	1987.42	2040.44	12449235.0	-66.18
6	2012.23	1984.75	1989.65	11947028.3	-66.00
7	2740.26	2195.76	2225.99	17285379.3	-67.61
8	2138.49	2066.36	2158.97	13504142.2	-66.53

[a]Output quality variable obtained from three trials data for each experiment

[b]S/N ratio=-10 $\log \frac{1}{N}(\sum_{i=1}^{N} Y_i^2)$, where N=3

Table 4.6: Calculation of S/N ratio of each variable for ranking

Group	Variable No.	Variable	LLC[a]	HLC [b]	S/N ratio [c]	Rank
1	2	Promoter flow	-66.32	-66.58	0.257	6
2	3	Air (Crystallizer)	-66.09	-66.81	0.721	1
3	9	O_2% (Condenser)	-66.58	-66.32	0.258	5
4	5	CO_2% (Crystallizer)	-66.61	-66.30	0.309	4
5	4	O_2% (Crystallizer)	-66.29	-66.61	0.316	3
6	8	Condensate flow	-66.18	-66.73	0.547	2

[a]'Low level contribution' for each variable is calculated as the average of S/N ratios of those experiments in the orthogonal array where the variable is contributing as low level

[b]'High level contribution' for each variable is calculated as the average of S/N ratios of those experiments in the orthogonal array where the variable is contributing as high level

[c]Absolute difference between HLC and LLC

4.3.3 Backward elimination of variables from groups using SVR

Table 4.7: Backward elimination of group variables using SVR

Step	Variable used	RMSE			Improve?	Decision
		Cal.	C.V.	Pred.		
1[a]	3, 6, 7, 10, 11, 1, 8, 4, 5	36.19	55.64	55.85		
2[b]	3, 6, 10, 11, 1, 8, 4, 5	36.21	54.05	52.07	Yes	Eliminate variable 7
3[c]	3, 6, 10, 1, 8, 4, 5	38.6	54.15	56.84	No	Keep variable 11
4[d]	3, 6, 10, 11, 1, 4, 5	36.31	43.94	45.61	Yes	Eliminate variable 8

[a]Select all the variables from group 2, 6, 5 and 4
[b]Eliminate variable 7 from group 2
[c]Remove variable 11 from group 2
[d]Reinclude variable 11 and remove variable 8 from group 6

Based on the rank as shown in Table 4.6, The four groups having the highest S/N ratios were selected. The next step was to eliminate variables which are redundant for prediction. Table 4.7 explains the steps involved in the variable elimination process. First, a model was built using all the variables from groups 2, 6, 5 and 4. Variable 7 from group 2 was eliminated and a model was built using the remaining variables. This lead to a decrease in RMSE value; therefore, variable 7 was eliminated. Next, variable 11 from group 2 was removed and a model was built using the remaining variables. The elimination of variable 11 increased RMSE values of both cross-validation and prediction; therefore, it was re-included in the list of variables. Figures 4.3 (a), (b) and (c) show the prediction performance of training and test data of the first three trials. This procedure was repeated until all the variables were tested once.

4.3.4 Final prediction model

The prediction performance of the final SVR model is shown in Figure 4.4. The model was developed using the seven variables emerging from the backward elimination. The

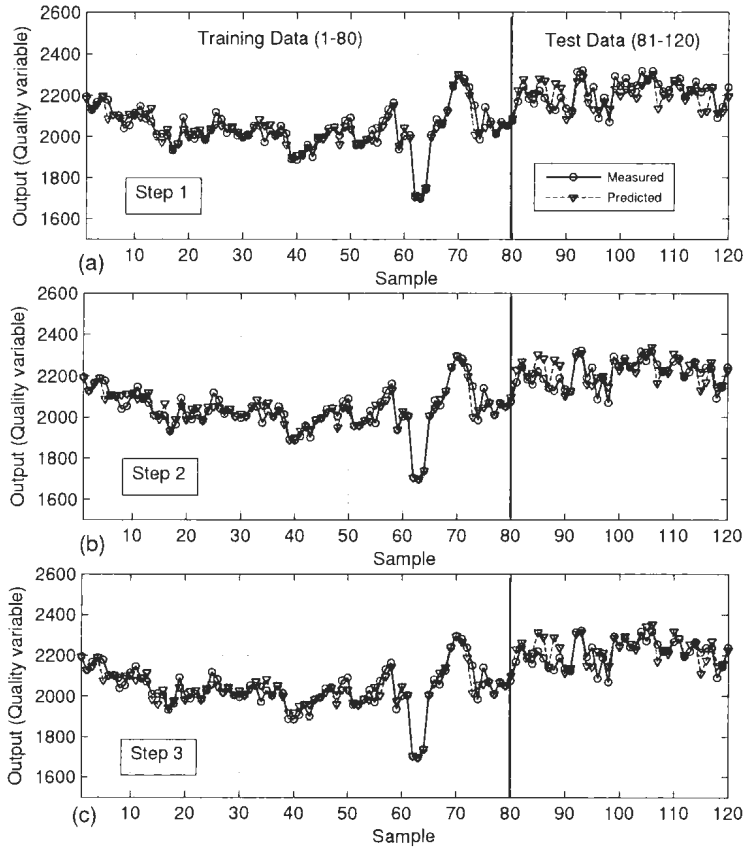

Figure 4.3: Prediction performances at the selection stages of groups variables using SVR: (a) Step 1: Using all the variables from groups 2, 6, 5 and 4 ; (b) Step 2: Eliminating variable 7 from group 2 and (c) Step 3: Eliminating variable 11 from group 2

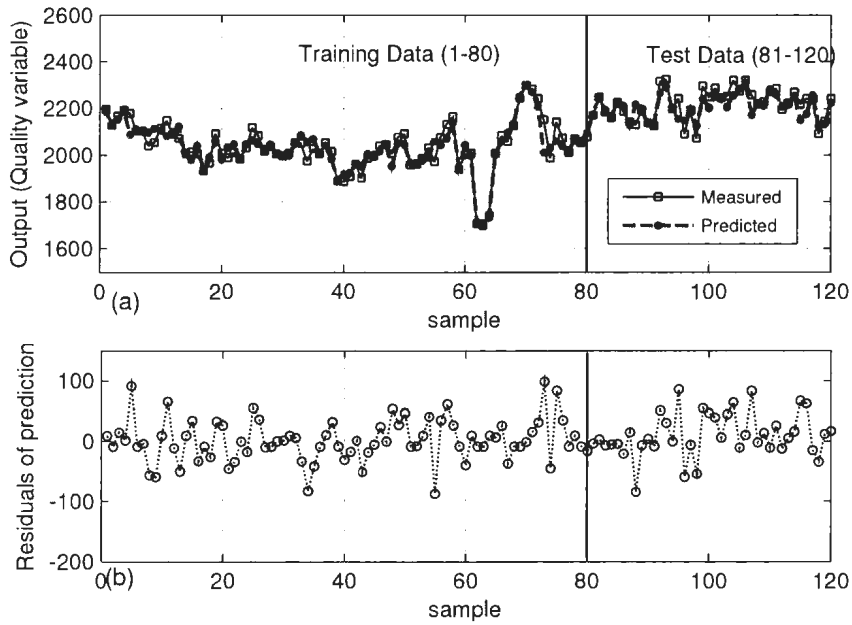

Figure 4.4: Final prediction using seven variables selected from step 4: (a) Predicted and measured output (b) residuals of prediction

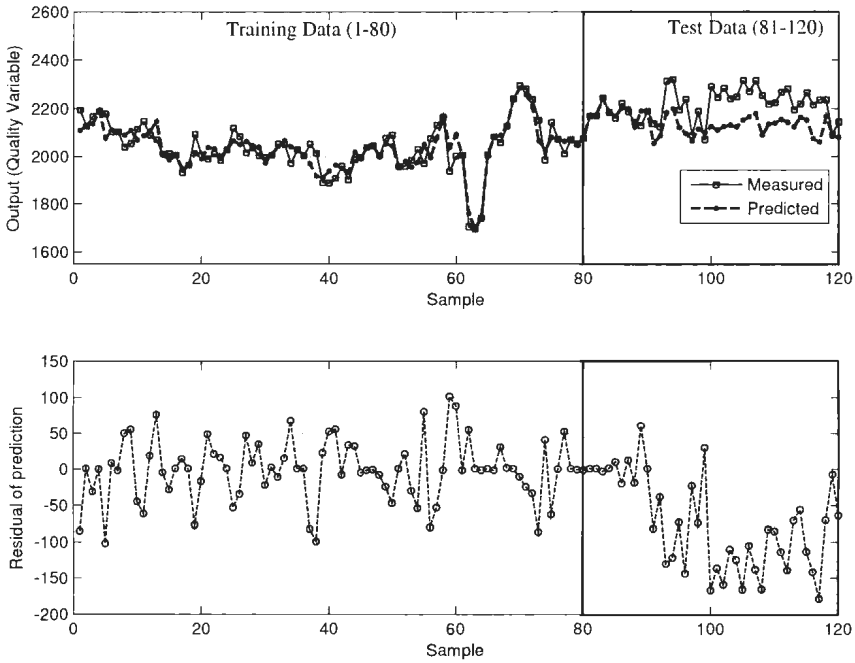

Figure 4.5: Prediction by SVR model using variables selected from VIP approach

ε-SVR algorithm was used with a radial basis function (RBF) as kernel.

The proposed hybrid Taguchi-Wrapper variable selection method is compared with the variable importance in projection (VIP) method in the context of the SVR model. The ranks of the variables based on VIP scores is given in Table 3.7. Seven variables selected from VIP were used to build a separate model using ε-SVR. Figure 4.5 shows the prediction performance of VIP-SVR predictor. A comparison of Figure 4.4 with Figure 4.5, clearly shows that the variables selected by the Taguchi-Wrapper method have superior prediction performance compared to the variables selected by the VIP method. Figure 4.6 shows the RMSE values of the training, validation, and test sets for both methods. It clearly shows that the Taguchi-Wrapper based SVR predictor has less prediction error than the VIP-SVR based predictor.

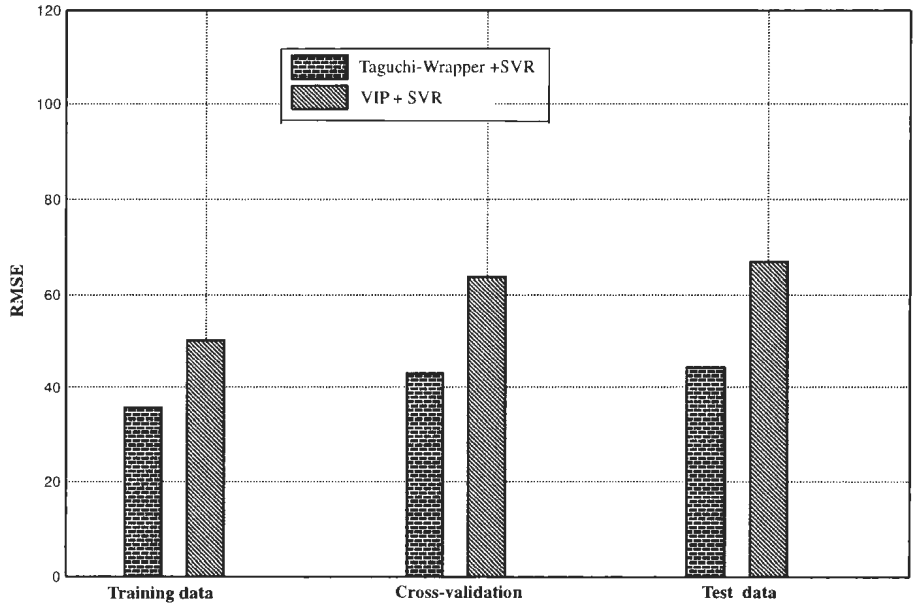

Figure 4.6: Comparison of RMSE values for training, cross-validation and test data

4.4 Conclusions

A systematic quantitative method for selecting the most important input variables for an inferential predictor from a large set of correlated variables is developed. The method is a combination of the retrospective Taguchi method and wrapper approach.

- The method resolves the implementation difficulties of the retrospective Taguchi method arising from correlated nature of process variables. It utilizes a classification algorithm to classify variables into groups of correlated variables. Instead of using all variables, a representative variable from each group is used to fill-in the Taguchi orthogonal array with the appropriate values. Since these variables are uncorrelated, the orthogonal array can be easily filled using historical data.

- The proposed Taguchi-Wrapper-SVR method is applied for building an inferential predictor to predict 4-CBA concentration for a PTA process. The method is compared with the VIP method. The Taguchi-Wrapper-SVR method shows significantly better performance in predicting 4-CBA concentration than the VIP-SVR method. The results clearly indicate that the proposed methodology can improve the prediction ability of SVR predictor.

- Wrapper-based variable selection for SVR requires large computational effort. Taguchi method can effectively eliminate groups of variables and bring the number of variables to a manageable number when wrapper based method can be feasible. A combined retrospective Taguchi and SVR-wrapper approach helps to reduce computational effort substantially.

Chapter 5

Variable selection for PCA model applied to Hydromet process

5.1 Introduction to Fault Detection

Principal component analysis (PCA) has been successfully used as a fault detection & diagnosis (FDD) tool in a wide range of processes [Bakshi, 1998, Kresta et al., 1991, Qin, 2003]. PCA projects data lying on a high dimensional measurement space onto a lower dimension space. Usually, process variables are correlated with each other and this correlation breaks down during any faulty situation. A PCA model identifies this in advance and indicates it as a fault in the process. Early fault detection can provide operators sufficient time to take appropriate action to avoid process shutdown.

A key issue in building a PCA model is the selection of important variables that bear fault signatures. This reduces the complexity of the model, avoids the inclusion of any variable that makes no major contribution in representing the process, and in the event a fault is detected, makes it easier to diagnose the root cause of fault. In building a PCA model, there is no specific guideline about the selection of variables.

It is a trial and error process where variables are selected based on data quality analysis, knowledge of the process and feedback from the plant operators. This Chapter proposes a new systematic approach for the selection of important variables for a PCA model based on the retrospective Taguchi method. A detailed review on the methodology and application of the Taguchi method is discussed in Chapter 3. In this chapter, we describe the monitoring scheme developed to detect and diagnose process faults of leach residue thickener (LRT) and counter current decantation (CCD) circuit of a hydromet process.

This chapter is organized as follows. An overview of the current control practices in mineral processing industries is presented in Section 5.2. A brief review on PCA along with its fault detection criteria is given in Section 5.3. Section 5.4 describes the variable selection methodology for a PCA model using the retrospective Taguchi method. Section 5.5 presents data description along with related operational problems of industrial case studies. Section 5.6 explains the methodology with results. Finally, the chapter ends with concluding remarks in Section 5.7.

5.2 Monitoring practices in mineral processing plants

Mineral processing industry has many regulatory issues and operational challenges which arise due to the solid handling nature of the process. One of the key issues is that, in most of the cases, its dynamic behavior is poorly understood. For example, factors like ore compositions, particle size distribution, ore conditioning etc. influence the process to a greater extent [Jemwa and Aldrich, 2006]. It is difficult to describe these behaviors mathematically. This complex nature of the process creates problems in the automated control system. Essential properties such as mineral texture, libera-

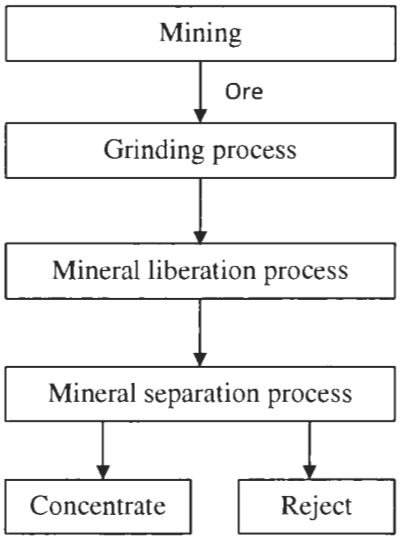

Figure 5.1: Major processing steps involved in a mineral processing plant

tion degree, surface activity, slurry rheology, bubble size distribution and loading are difficult to measure and even to infer from other measurements [Hodouin et al., 2001]. In some cases, the control practice is dependent on visual appearance, which requires highly expert observations by the operators. It requires innovative and effective solutions to ensure smooth operations of these strongly disturbed, poorly modeled and difficult to measure processes [Hodouin, 2011].

To overcome these challenges, mineral processing industries are implementing real time monitoring systems on various process units and streams [Jemwa and Aldrich, 2006]. Various data-based monitoring techniques such as data reconciliation, image processing, pattern recognition, multivariate statistical methods, soft sensors and controller performance monitoring, etc. are gaining popularity in mineral processes. In a mineral processing plant, the main concern is rapid changes in ore grade, mineralogy and grindability. These changes disturb the process, so it is very important to detect this process shifting through a suitable model. PCA has the ability to detect these shifts in operating conditions, and therefore has received much attention

in mineral processing. Application of the PCA approach for fault detection in flotation column is described in [Bergh and Acosta, 2009]. Expert control systems such as neural networks and image processing systems are also gaining popularity. Application of a neural network as an expert control system for determining and tracking the optimal concentrations of zinc and sulfuric acid for the electrolytic process in zinc hydrometallurgy can be found in [Wu et al., 2001]. [Wu et al., 2002] developed an expert control system using a combination of steady-state mathematical models and model switching rules to determine and track the optimal p^H of the neutral and acid leaches coming out of the leaching process in a zinc hydrometallurgy plant. [Hongqiu et al., 2010] developed a model to optimize the cobalt purification process in zinc hydrometallurgy using a combination of fuzzy C-means clustering and fuzzy support vector machine. Application of image processing to extract selected features from the images of a flotation cell and using them to build a PLS model to predict the zinc concentration can be found in [Hatonen et al., 1999].

5.3 Theory of Principal Component Analysis

Principal component analysis is a dimension reduction technique introduced by Pearson (1901), and later developed by Hotelling (1933). It projects correlated set of variables onto a lower dimensional subspace where the transformed data is uncorrelated. The coordinates of this new subspace are called principal components (PC). Thus, it decomposes the data into a few key uncorrelated variables and separates the redundancy which usually arises in process data due to multiple measurements of the same variable or linear relationships between variables. Each PC is a linear combination of original variables. The first few principal components reflect the major trend in the process and can be used in monitoring instead of a large number of variables.

Consider a data matrix, $X \in \mathbf{R}^{N \times m}$ containing N samples and m variables. The objective is to find m linear combinations of the original variables, known as principal components (PCs), which are a set of uncorrelated score variables as shown in Equation 5.1:

$$t_i = Xp_i \ [i = 1, \cdots, m]; \ p_i \in \mathbf{R}^{m \times 1} \tag{5.1}$$

Here, p is known as the loading vector which represents the weights of each variable. The loading vector is calculated in such a way that it can maximize the variance in the score vector. The calculation steps of PCA are described below:

Step 1: Calculate the loading matrix P by applying singular value decomposition (SVD) on the co-variance matrix, $\Sigma = X^T X/(N-1)$ as follows:

$$\Sigma = P \Lambda P^T; \quad P \in \mathbf{R}^{m \times m}; \quad \Lambda \in \mathbf{R}^{m \times m}. \tag{5.2}$$

The loading vectors are the orthonormal column vectors of the matrix P and Λ is the co-variance matrix of the principal components containing non-negative real eigenvalues of decreasing magnitude ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$) along its main diagonals with zero off-diagonal elements. Since the loading vectors are orthogonal to each other, the scores are uncorrelated to each other.

Step 2: Calculate the principal components as $t_i = Xp_i, \quad [\ i = 1, \cdots, m\ ]$.

For m variables, the equal number of PCs are extracted as follows:

$$X = t_1 p_1^T + t_2 p_2^T + \cdots\cdots + t_m p_m^T \tag{5.3}$$

Step 3: Determine the number of PCs, r ($r < m$) required to capture most of the systematic variability in the data. In practice, 2 or 3 PCs are often sufficient to explain most of the systematic variation [Kourti and MacGregor, 1995]. Since these scores

are uncorrelated, they can be monitored individually. These r scores are used to filter noisy data according to Equation 5.4, which is also known as PCA model:

$$\hat{X} = t_1 p_1^T + t_2 p_2^T + \cdots\cdots + t_r p_r^T = X P_r P_r^T \tag{5.4}$$

After building a PCA model, future behavior can be referenced against this 'in control' model. New multivariate observations can be projected onto the plane defined by the PCA loading vectors to obtain their scores as $[t_{i,new} = p_i^T x_{new}]$, and the residuals as $[e_{new} = x_{new} - \hat{x}_{new}]$, where $\hat{x}_{new} = P_r t_{r,new}$; $t_{r,new}$ is the $(r \times 1)$ vector of scores from the model and P_r is the $(m \times r)$ matrix of loadings [Kourti and MacGregor, 1995].

5.3.1 Fault Detection Criteria

Two popular collective test statistics based on PCA known as Hotelling's T^2 and Q-statistics are used for fault detection.

5.3.1.1 Hotelling's T^2-Statistics

The original form of T^2-statistic is defined as:

$$T_i^2 = (x_i - \bar{x})\Sigma^{-1}(x_i - \bar{x})^T \tag{5.5}$$

In PCA Hotelling's T^2-Statistics provide a measure of the variation within the PCA model. T^2 is the sum of normalized squared scores as shown in Equation 5.6:

$$T_i^2 = t_i \Lambda_r^{-1} t_i^T, \tag{5.6}$$

where Λ_r represents a diagonal matrix containing the r largest eigenvalues, t_i refers

to the i-th row of $T_r \in \mathbf{R}^{N \times r}$, the matrix of r score vectors from the PCA model. In calculating the T^2-Statistic, the smaller eigenvalues are not considered, such that it will not be affected by the inaccuracies of the smaller eigenvalues.

Statistical confidence limits for T^2 are directly calculated from the F distribution as shown in Equation 5.7 [Wise and Gallagher, 1996]

$$T^2_{UCL}(\alpha) = \frac{r(N-1)}{(N-r)} F_\alpha(r, N-r). \tag{5.7}$$

Where $F_\alpha(r, N-r)$ is the $100(1-\alpha)\%$ critical point of the F distribution with r and $(N-r)$ degrees of freedom.

Hotelling's T^2 detects abnormal variations in the quality variables in the plane of the first r PCs which are caused by common causes. In case of the occurrence of a totally new type of event which is not present in the reference model, that can be detected by computing the squarred prediction error (SPE) of the residuals of a new observation [Kresta et al., 1991].

5.3.1.2 Squared Prediction Error (SPE) or Q-Statistics

Q-Statistics or Q-Residuals, also known as Rao-statistics, deals with the observation space corresponding to the $(m-r)$ smallest singular values. It represents the squared perpendicular distance of a new multivariate observation from the projection space. The collective test statistics is defined as follows [Kourti and MacGregor, 1995]:

$$Q_i \ \ or \ \ SPE_i = (x_i - \hat{x})(x_i - \hat{x})^T \tag{5.8}$$

The distribution of the Q-statistics as approximated in [Jackson and Mudholkar, 1979] is given below:

$$Q_\alpha \ \ or \ \ SPE_\alpha = \left[\frac{h_o c_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_o (h_o - 1)}{\theta_1^2} \right]^{1/h_o}. \tag{5.9}$$

Where $\theta_i = \sum_{\mathbf{i=\alpha+1}}^{\mathbf{n}} \sigma_j^{2i}$, $h_o = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$, and c_α is the normal deviate corresponding to the $(1-\alpha)$ percentile. The threshold for the SPE is calculated by Equation 5.9 for a given level of significance. This threshold can be used to detect the fault. Geometrically SPE calculates a projected distance from a point in R^m to the hyper-plane defined by the principal components. It measures to which extent the new data is in agreement with the correlation structure identified by the set of PCs. In the case of agreement, SPE only reflects the random variation or measurement noise. SPE will be below the threshold given by Q-statistics indicating that the process is 'in control'. On the contrary, if the correlation of data breaks down, that will be an indication of a faulty operation and will be manifested by violating the control limit by SPE. Usually, PCA has the ability to detect a fault earlier before it is perceived by the operator.

5.4 Methodology for selection of input variables for PCA

The proposed methodology uses the retrospective Taguchi method to select important input variables for PCA model. In the case of an inferential predictor as described in Chapter 3, the measured quality variable is used as a trial result for each experiment in the array. PCA gives a symmetric treatment of data and does not divide variables into input (X) and output (Y) blocks. As such an outcome variable for each experiment has to be defined. In the present analysis, the calculated Hotelling's T^2-statistics is

considered as the outcome variable for each trial. Hotelling's T^2-statistics provides a measurement of process variation within the PCA model. T^2-statistics was used as an outcome variable because it is a determinant of occurrence of common fault in a PCA model. If T^2 value crosses the defined confidence limit, it indicates a fault in the process. The objective of the variable selection method is to select variables which bear the most fault signatures of the unit. Therefore, the variables which will have most deterministic power in indicating the fault will also have the most contribution to the T^2-statistics.

Figure 5.2 shows a flow chart of the steps of the proposed variables selection method. Step 1 is the screening of important variables based on prior process knowledge and trend analysis. Through trend analysis, the variation of each variable is observed. Variables perceived as not important from process point of view are omitted from the list. Step 2 is data preprocessing for outlier removal and filling in missing values. Preprocessing is a crucial step in data analysis, especially for industrial data, as it may contain bad values as a result of process upsets. Also the sensors may contain bias error or variance error; therefore, it is important to validate the measurements before further analysis. Step 3 is the selection of normal operation data from data historian that represents the process well. An ideal PCA model requires the identification of process data that captures the correlation between different variables. Step 4 is building a PCA model to calculate the Hotelling's T^2 values. This T^2 value will be considered as an outcome variable for each experiment fitted in the orthogonal array.

Step 5 is the selection of the ranges of levels for all input variables. For each variable two levels are used for this analysis. The exact size of the range for a level depends on the range of variation of the variable. In Step 6, the appropriate orthogonal array based on the levels and number of variables is selected following Taguchi experimental design. Orthogonal array is a systematically designed array where each row

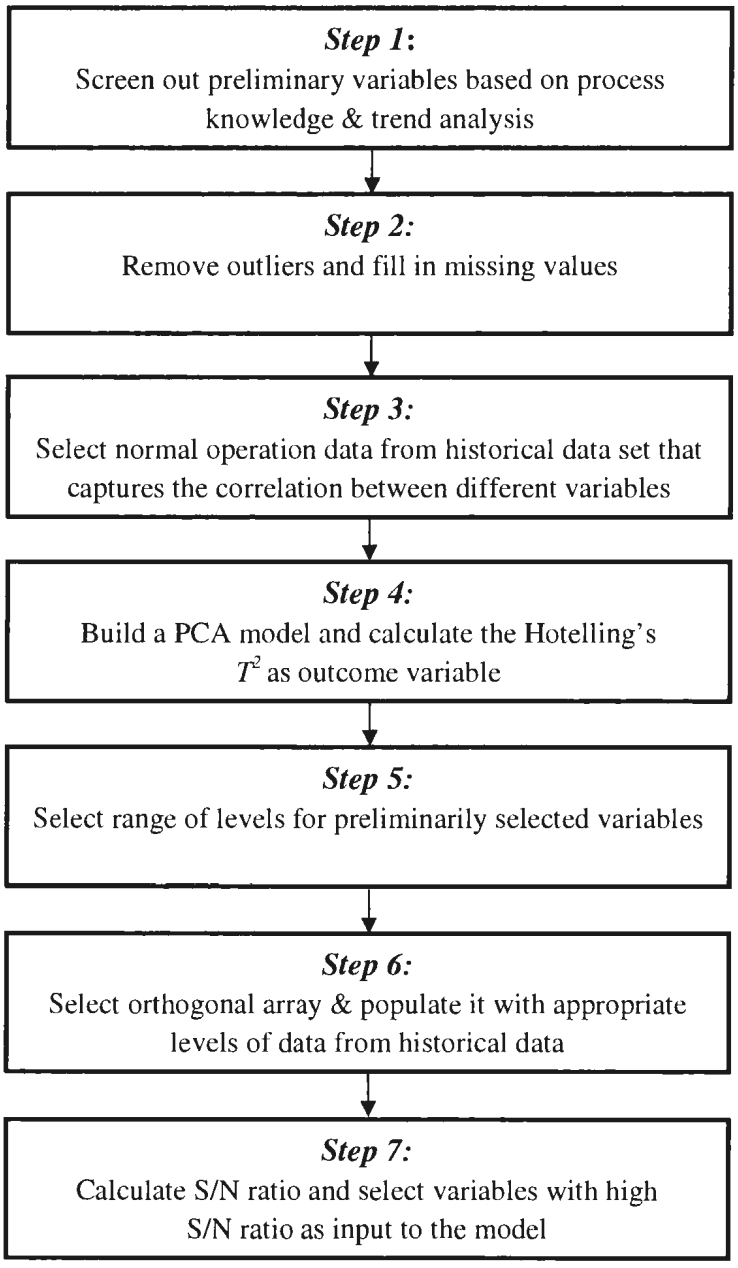

Figure 5.2: Proposed variable selection flow chart for process fault detection

corresponds to a particular experiment and variables are arranged in columns. Once the array has been selected, it is populated with appropriate values from the data set

that fits well with each experimental condition in the array. In the current analysis, nine closely matched measurements are selected from the data-set for each experiment in the array. The target is to keep the average of these nine data points within the range of the levels of that particular experiment. Nine data points are considered as nine trials for each experiment. The corresponding T^2 values are considered as the results obtained from the trials. In a PCA model, Hotelling's T^2-statistics value is a measure of fault in the process. T^2-statistics is expected to be in their nominal range during normal conditions. Therefore, the "nominal is the best characteristic" criteria is used to calculate the S/N ratio for each experiment. Taguchi has proposed the following equation to calculate the S/N ratio for this characteristic [Ghani et al., 2004]:

$$S/N = 10log\frac{\bar{y}^2}{\sigma_y^2} \tag{5.10}$$

where $\bar{y}$ is the average of observed output data and σ_y^2 is the variance of y.
In the final step, the trials' results are utilized to calculate the signal-to-noise (S/N) ratio for each experiment. The contribution of a variable in an experiment is measured from its position as a low level or high level in that experiment. The S/N ratio for a variable is calculated from the difference of its average low level contribution and average high level contribution in all the experiments. Finally, the variables are ranked based on the S/N ratio, and the variables with high S/N ratios are selected as input variables to the PCA model.

5.5 Industrial Case Study : Hydromet Process

PCA-based monitoring schemes are developed for the various thickener units of a hydro-metallurgical plant. The important variables for building the PCA model are selected using the proposed retrospective Taguchi-based method. In the following sec-

tions, we briefly describe the process and various operational problems in the thickener units. Due to the proprietary nature of the process, the tag names and actual values are withheld in the description.

5.5.1 Process description

In a hydromet process, raw ore is concentrated through subsequent metal extraction stages to extract target metals for industrial or commercial uses. Figure 5.1 presents a simplified flow chart showing the major stages in a Nickel hydromet process. Crushing, grinding and size classification are typically used for minerals liberation. Separation processes involve flotation and leaching units followed by thickener units to separate target metals as leach solution from leach slurry. Moreover, there are some peripheral processes such as feeders, conveyors, tailing disposal, effluent treatment, reagent dosage, etc.

The current study focuses on the thickener circuit of a hydro-metallurgical plant. Figure 5.3 shows the different processing units of a typical hydro-metallurgical refining process. Metal concentrate is first finely grounded in the grinding unit and then passed to the pre-leach section. Pre-leached concentrate is then processed in an autoclave where it reacts with oxygen and sulphuric acid at an elevated temperature and pressure to produce an impure sulphate solution of metal called leach slurry. Leach slurry is then thickened in leach residue thickener (LRT) to produce leach solution. The underflow slurry from LRT is further thickened in the CCD circuit to separate the remaining leach solution. Leach residue from the CCD circuit is treated in the effluent treatment section for impoundment. Leach solution coming out of the LRT and CCD units is further processed to remove any impurities from the metal solution and sent to a metal extraction unit for the production of finished metal. The leach residue thickener and counter current decantation (CCD) circuit have a considerable impact

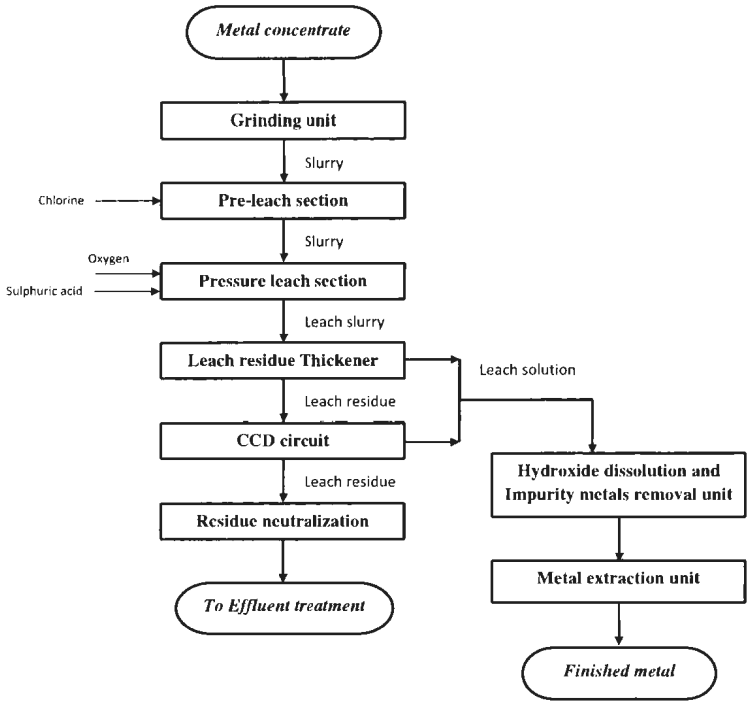

Figure 5.3: Block diagram of different processing units in a hydro-metallurgical plant

on the process in terms of maximizing metal recovery from leach slurry. In the following sections, a brief overview of these units is presented.

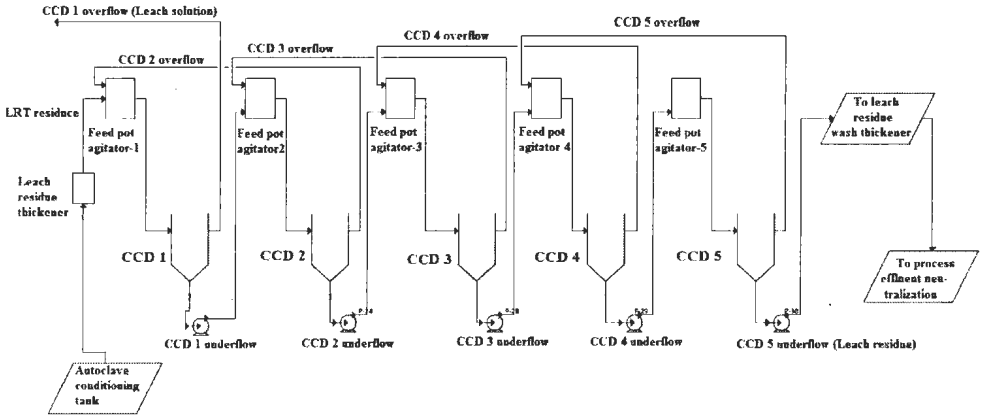

Figure 5.4: Schematic diagram of material flow in a CCD circuit

Leach residue thickener (LRT) performs the primary separation of the target metal from leach slurry. Leach slurry coming out of pressure leach section is thickened in leach residue thickener to separate the target metal as leach solution. Counter current decantation (CCD) thickener circuit is used in the hydro-metallurgical process to recover additional soluble metals from leach residue coming out of LRT thickener. The basis of CCD operation is to obtain a concentrated suspended solid in order to minimize the leach liquor content in underflow slurry. The diluted slurry is then fed to the next thickener. In this way, the suspended solids are concentrated repeatedly in each thickener in order to keep the metal content to a minimum in the underflow slurry. The number of CCD stages required to recover the desired amount of soluble metal depends on the amount of liquor present in the thickener underflow slurry. Figure 5.4 shows the underflow and overflow flow-paths in a five stage CCD circuit. The final product is the leach solution obtained as overflow from CCD 1 thickener. Final leach residue is collected from the underflow of CCD 5 thickener for further processing prior to disposal.

5.5.2 Operational problems in thickener operation

Doughnut-shape formation:

Doughnut-shape formation in the bottom layer of thickener is a major problem in the thickener operation. The reasons for this formation are described below:

- synthetic polymer is commonly used to increase the thickener performance. If its dosage is increased, the underflow viscosity is also increased. As a result, there comes a point when the thickened solids lose their fluidity. The rake arms may no longer cause flow toward the discharge point. Mass tends to travel along in front of the rake arms and starts to accumulate in the rake arms. As a result, it becomes blocked as stationary mass and additional retention time will make them more immovable. The net result is the formation of a fairly solid accumulation that slides along the floor of the thickener and eventually fills the rake truss itself. If allowed to continue long enough, additional solids accumulate in front of the mass contained in the rake and the total accumulation can eventually grow to form a complete ring. This formation is commonly known as doughnut formation. The island effectively blocks settled solids from discharging through the central discharge outlet. Any solids which reach the outlet must pass up and over this island or short circuit directly from the feed inlet. With insufficient detention time in the thickener, the result will be a much lower solids concentration (diluted underflow slurry) [Moss, 1978].

- while operating a thickener, if the density of underflow sludge decreases, the rate of underflow is decreased to control density. This action may work in favor of increased doughnut formation. If the retention time is increased by allowing the solids in the thickener as long as possible, it may then cause doughnut shape formation.

When doughnut forms:

- the viscous drag through the slurry is increased.
- the friction between the doughnut and the bed beneath the rakes increases, which eventually increases rake torque.

The following signs can be indicative of doughnut shape formation:

- an increase in rake torque, at the same time, a decrease in underflow density..

The control strategy during doughnut formation:

- feed rate should be decreased.
- rake should be raised to make the doughnut island slough off and flow or slide into the discharge outlet.

Channeling:

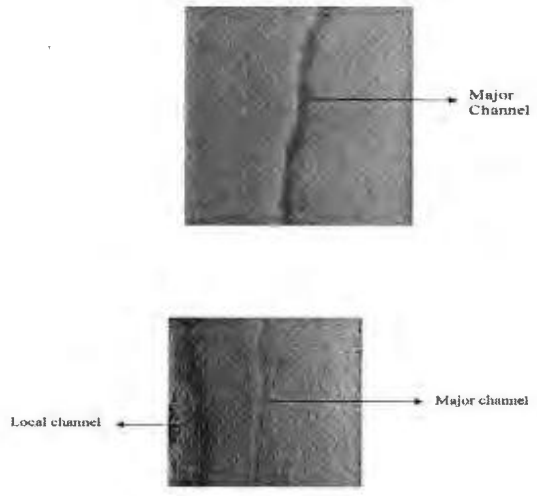

Figure 5.5: Formation of different channels on the settling bed wall [Kurt, 2006]

Channeling is described as the short circuiting of fluid in the bed due to the creation of a higher concentration gradient. Figure 5.5 shows the formation of a major and localized channel on the settling bed wall. The causes of channeling are:

- the high-pressure gradient at the bottom of the bed. The fluid rising from the compression zone causes cracks through the bed, which leads to shear failure of the compression region [Vesilind, 1968].
- another possible cause is the breaking of the solid matrix by impurities and air bubbles that break flow paths through the sludge bed [Glasrud et al., 1993].
- High initial concentration and settling time may also create channeling.
- [DeBoer, 1990] pointed out that the degree of channeling and/or size of the channels are perhaps responsive to changes in the flocculant nature of a given suspension.
- [Dixon, 1979] mentioned the possibility that channeling is the result of wall effects.
- Factors like the size, shape and density of the solid particles also affect the formation of channeling.

The effects of channeling are:

- during compression, it causes the real concentration value in the bed to be higher than the normal value. This leads to overestimation of the thickener capacity.
- it causes other events such as reverse concentration gradients in the settling bed.

Rat holing:

In order to operate a thickener properly, the sludge that has settled on the bottom of the tank has to be moved into the sludge trough, from which it can be pumped to the next unit. If the sludge fails to be moved to the sludge trough, the pump will extract dilute sludge and, as a result, the sludge forms a funnel on top of the sludge trough [Sanin et al., 2011]. This is known as rat holing. Rat holing is considered as fatal to the proper operation of the thickener.

5.6 Results and Discussion

5.6.1 Leach Residue Thickener (LRT)

5.6.1.1 Data Description

Table 5.1: List of preliminarily selected variables for LRT thickener used in proposed variable selection process

No.	Variable name	Description
1	Underflow rate (FFIC)	Measured PV in ratio controller (FFIC)
2	Underflow flow rate (FIC)	Measured PV in flow-controller (FIC) (kg/hr)
3	Underflow slurry flow	Solids mass flow from FQI transmitter (kg/hr)
4	Rake torque	Torque measured from power requirement
5	Bed weight	Measured PV in bed weight controller (kg)
6	Bed pressure	Solids loading of thickener bed(kPa)
7	Rake lift position	Rake's position to measure lifting condition (%)
8	Influent temperature	Measured input feed slurry temperature (°C)
9	Overflow temperature	Measured overflow temperature (°C)
10	Underflow temperature	Measured underflow slurry temperature (°C)
11	Visc loop temperture	Measured visc loop temperature (°C)

From the available data of a total thirty variables, variables that do not show any variation and are perceived as not important from process point of view are eliminated. Table 5.1 shows the list of preliminarily selected variables with description. These eleven variables were used as input to the variable selection process.

5.6.1.2 Variable selection

Using normal operation data, first, a PCA model is built using preliminarily selected eleven variables (shown in Table 5.1) to calculate the Hotteling's T^2 values. The ranges of the levels of eleven variables are also determined from the data as shown in Table 5.2. The standard orthogonal array for eleven variables and two levels is shown in Table 5.3. In the next step, for each experiment in the array, values were searched from the historical data so that the values are within the ranges specified for each

Table 5.2: Selection of levels of data for designing Taguchi orthogonal array using historical data for LRT thickener

No.	Variable	Min	Max	Range	Limit[a]	LL[b]	HL[c]
1	Underflow rate (FFIC)	5.17	66.87	61.70	27.76	32.94	39.11
2	Underflow flow rate (FIC)	1.60	1033.34	1031.74	464.28	465.88	569.06
3	Underflow slurry to CCD flow	0.07	819.86	819.78	368.90	368.98	450.96
4	Rake torque	0.27	13.52	13.25	5.96	6.23	7.55
5	Bed weight	5788.5	9659.7	3871.2	1742.1	7530.5	7917.6
6	Bed pressure	38.57	43.91	5.33	2.40	40.97	41.51
7	Rake lift position	10.84	100.18	89.34	40.20	51.05	59.98
8	Influent temperature	23.41	95.59	72.17	32.48	55.89	63.11
9	Overflow temperature	58.37	84.99	26.62	11.98	70.35	73.01
10	Underflow temperature	29.86	82.44	52.58	23.66	53.52	58.78
11	Visc loop temperture	11.95	39.86	27.90	12.56	24.51	27.30

[a]Limit is 40 percent of range value
[b]Calculated by adding limit value with minimum value; low level data range is from min to LL
[c]Calculated by subtracting limit value from maximum value; high level data range is from HL to max

Table 5.3: Taguchi orthogonal array with low-level and high-level values of eleven variables to design the experiment for LRT thickener

	Variable number										
Exp	1	2	3	4	5	6	7	8	9	10	11
1	**32.94**	**465.88**	**368.98**	**6.23**	**7530.51**	**40.97**	**51.05**	**55.89**	**70.35**	**53.52**	**24.51**
2	**32.94**	**465.88**	**368.98**	**6.23**	**7530.51**	41.51	59.98	63.11	73.01	58.78	27.30
3	**32.94**	**465.88**	450.96	7.55	7917.64	**40.97**	**51.05**	**55.89**	73.01	58.78	27.30
4	**32.94**	569.06	**368.98**	7.55	7917.64	**40.97**	59.98	63.11	**70.35**	**53.52**	27.30
5	**32.94**	569.06	450.96	**6.23**	7917.64	41.51	**51.05**	63.11	**70.35**	58.78	**24.51**
6	**32.94**	569.06	450.96	**6.23**	7917.64	41.51	**51.05**	63.11	**70.35**	58.78	**24.51**
7	**32.94**	569.06	450.96	7.55	**7530.51**	41.51	59.98	**55.89**	73.01	**53.52**	**24.51**
8	39.11	**465.88**	450.96	**6.23**	7917.64	41.51	59.98	**55.89**	**70.35**	**53.52**	27.30
9	39.11	**465.88**	**368.98**	7.55	7917.64	41.51	**51.05**	63.11	73.01	**53.52**	**24.51**
10	39.11	569.06	450.96	**6.23**	**7530.51**	**40.97**	**51.05**	63.11	73.01	**53.52**	27.30
11	39.11	569.06	**368.98**	7.55	**7530.51**	41.51	**51.05**	**55.89**	**70.35**	58.78	27.30
12	39.11	569.06	**368.98**	**6.23**	7917.64	**40.97**	59.98	**55.89**	73.01	58.78	**24.51**

[a]Bold faced-low level values ; normal-high level values

Table 5.4: Calculation of S/N ratio of each experiment in the orthogonal array for LRT thickener

	Y_i= Hotteling T^2 value for each trial														
Exp.	1	2	3	4	5	6	7	8	9	N	Sm_1 [a]	ST_1 [b]	Se_1 [c]	Ve_1 [d]	S/N[e]
1	1.5	5.6	0.3	4.8	1.6	3.3	8.4	1.2	1.4	9	88.5	144.5	56.0	7.0	1.12
2	3.9	6.3	0.7	2.5	10.2	1.4	1.3	0.5	2.1	9	93.4	175.2	81.8	10.2	-0.44
3	2.3	1.8	0.8	6.5	9.6	2.3	1.0	5.8	0.8	9	106.6	183.6	77.1	9.6	0.48
4	1.1	1.7	1.2	6.6	9.8	1.3	1.3	0.9	1.6	9	71.5	151.4	79.9	10.0	-1.65
5	4.3	1.9	2.6	0.8	1.6	1.6	1.3	0.8	2.2	9	32.0	41.1	9.1	1.1	4.78
6	1.6	1.3	1.1	1.3	6.6	1.4	1.7	1.1	0.9	9	31.4	57.0	25.6	3.2	-0.09
7	2.3	0.6	2.3	1.2	1.1	1.5	1.1	0.8	9.8	9	47.4	113.4	66.1	8.3	-2.79
8	1.2	0.8	0.4	1.5	0.6	2.6	1.5	1.1	1.9	9	15.1	19.0	3.9	0.5	5.25
9	0.8	1.5	1.6	1.8	1.4	5.8	0.9	1.6	1.1	9	30.2	48.6	18.4	2.3	1.30
10	1.6	2.1	0.9	1.3	0.8	2.6	1.1	5.9	2.2	9	37.9	57.3	19.4	2.4	2.09
11	4.0	2.3	1.4	0.8	0.3	1.1	1.2	5.4	1.3	9	35.1	58.2	23.1	2.9	0.94
12	1.7	2.3	0.7	2.1	1.7	0.7	0.7	2.3	1.1	9	19.6	23.4	3.8	0.5	6.47

[a] $Sm_1=(\sum_1^9 Y_i)/N$

[b] $ST_1=\sum_1^9 Y_i^2$

[c] $Se_1= ST_1\text{-}Sm_1$

[d] $Ve_1=Se_1/(N\text{-}1)$

[e] $S/N = 10\log\left[\frac{(Sm_1-Ve_1)}{N\times Ve_1}\right]$

Table 5.5: Calculation of S/N ratio of each variable for LRT thickener

No.	Variable	LLC[a]	HLC [b]	S/N ratio [c]	Rank
1	Underflow rate (FFIC)	0.202	3.211	3.01	2
2	Underflow flow rate (FIC)	1.543	1.394	0.15	10
3	Underflow slurry to CCD flow	1.291	1.621	0.33	9
4	Rake torque	2.740	-0.342	3.08	1
5	Bed weight	0.186	2.363	2.18	3
6	Bed pressure	1.704	1.279	0.42	8
7	Rake lift Position	1.518	1.369	0.15	11
8	Influent temperature	1.912	1.000	0.91	5
9	Overflow temperature	1.725	1.187	0.54	7
10	Underflow temperature	0.888	2.024	1.14	4
11	Visc loop temperature	1.797	1.115	0.68	6

[a]'Low level contribution' for each variable is calculated as average of S/N ratios of those experiments in the orthogonal array where the variable is contributing as low level

[b]'High level contribution' for each variable is calculated as average of S/N ratios of those experiments in the orthogonal array where the variable is contributing as high level

[c]Absolute difference between HLC and LLC

experiment. For each experiment, nine closely matched data sets were selected and considered as nine trials. Table 5.4 reports the corresponding Hotteling T^2 values as the outcome variables of nine trials for each experiment. Based on the trials' results, the S/N ratio was calculated for each experiment as given in Figure 5.4. Table 5.5 shows the calculation of S/N ratio for each variable. For example, in case of the underflow rate (FFIC), its low level values are placed in experiments 1 to 7 of the orthogonal array in Table 5.3. Therefore, its low level contribution is the average of S/N ratios of the first seven experiments. Similarly, the high level contribution will be the average of the S/N ratios of experiments 8 to 12. Finally the difference between the low level average S/N ratio and high level average S/N ratio gives the overall S/N ratio for a variable. Based on overall S/N ratios, variables were ranked. The ranks of the variables are shown in Table 5.5. Seven variables which have S/N ratios >0.5 were selected to build the PCA model.

5.6.1.3 Fault detection model

The selected variables were mean-centered and scaled by the reciprocal of the square root of standard deviation. Figure 5.6 (a) shows the eigenvalue plot for seven principal components. Based on the cumulative variance captured as shown in Figure 5.6 (b), one principal component was selected as it is captured 96.11 % of total variance. Figures 5.7 (a) and (b) show the T^2 and Q residuals plots of the model, respectively.

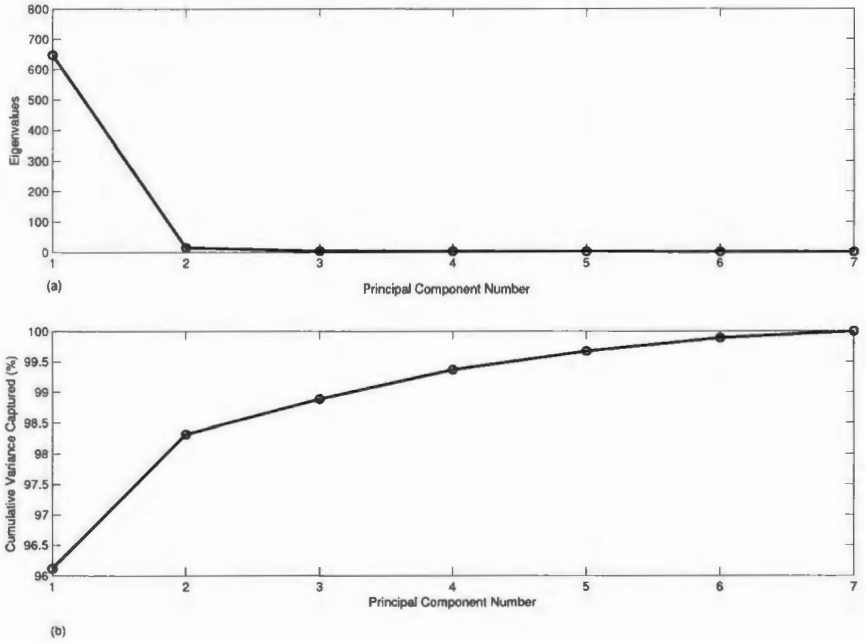

Figure 5.6: (a) Eigenvalue plot and (b) cumulative variance captured (%) plot for LRT thickener

5.6.1.4 Validation

In order to validate the fault detection and diagnosis capability of the model, two faulty data-sets are selected where process was impacted by fault which eventually led to a temporary shutdown of the unit.

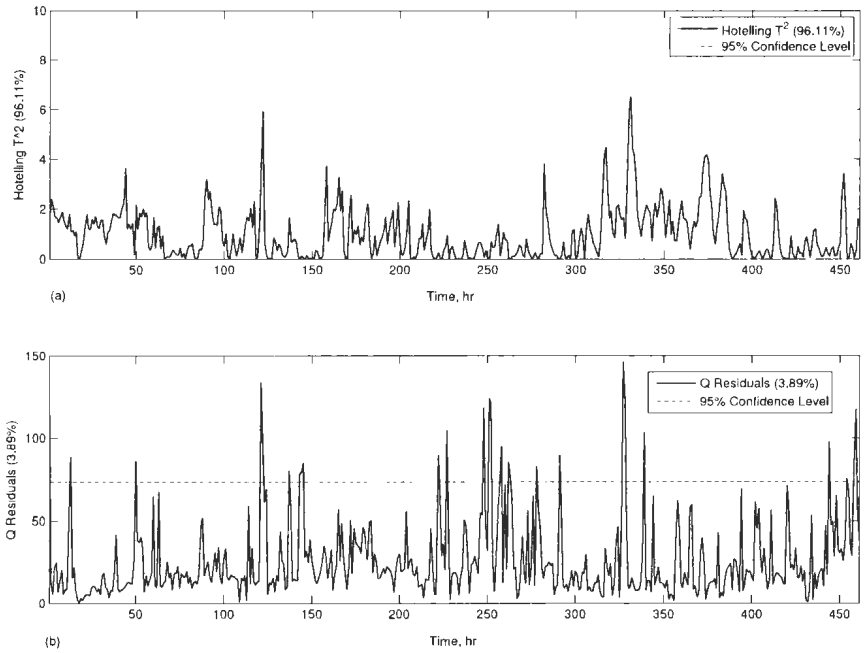

Figure 5.7: (a) Hotelling's T^2 plot and (b) Q-residuals plot of PCA model for LRT thickener

Validation data set 1:

Figures 5.8 (a) and (b) report the Hotelling's T^2 and Q-residuals plots of the first validation data-set, respectively. At t=6.8 hrs, the value of Q residual crossed the confidence limit and remained outside the limit upto t=15.0 hrs when the unit tripped.

Figure 5.9 illustrates the residual contribution of each variable over time using a color plot. It clearly shows that influent temperature has the most residual contribution for fault occurrence at t=6.8 hrs. In order to ascertain the root cause, each variable was further investigated. Figure 5.10 (e) shows at t=6.8 hrs influent temperature started to decrease. A decrease in temperature lowers the settling velocity as it reduces the rate of diffusion of flocculant and rate of collision of particles. Due to low settling velocity, the bed level decreases, which lowers the bed weight as shown

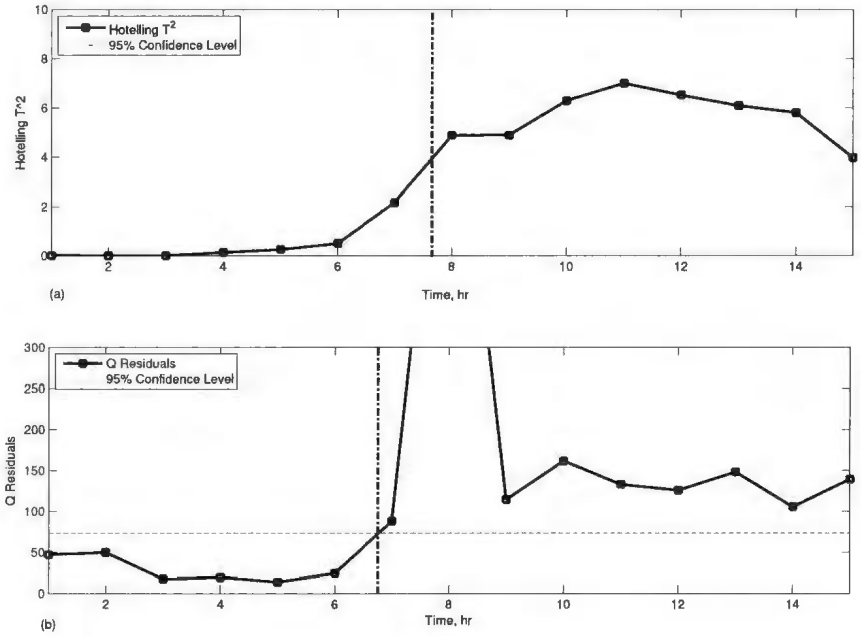

Figure 5.8: (a) Hotelling's T^2 plot and (b) Q-residuals plot of validation data set 1 for LRT thickener

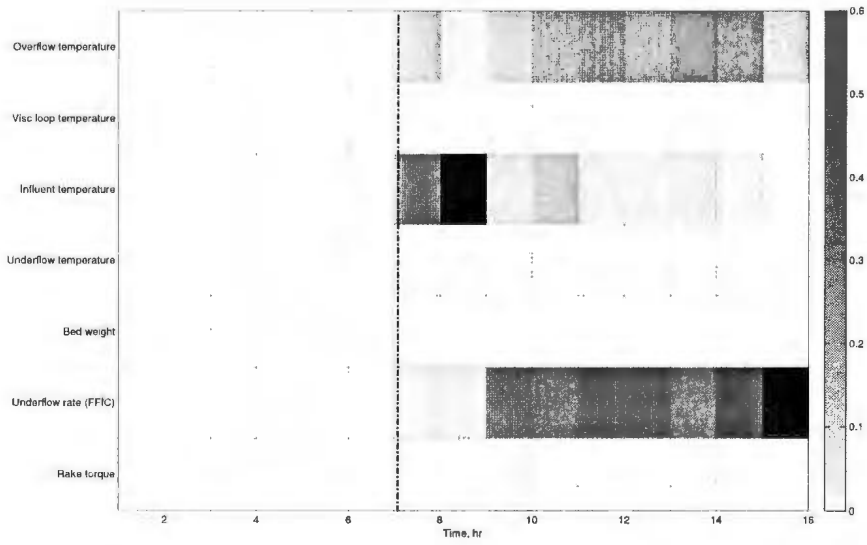

Figure 5.9: Color plot of validation data set 1 for LRT thickener

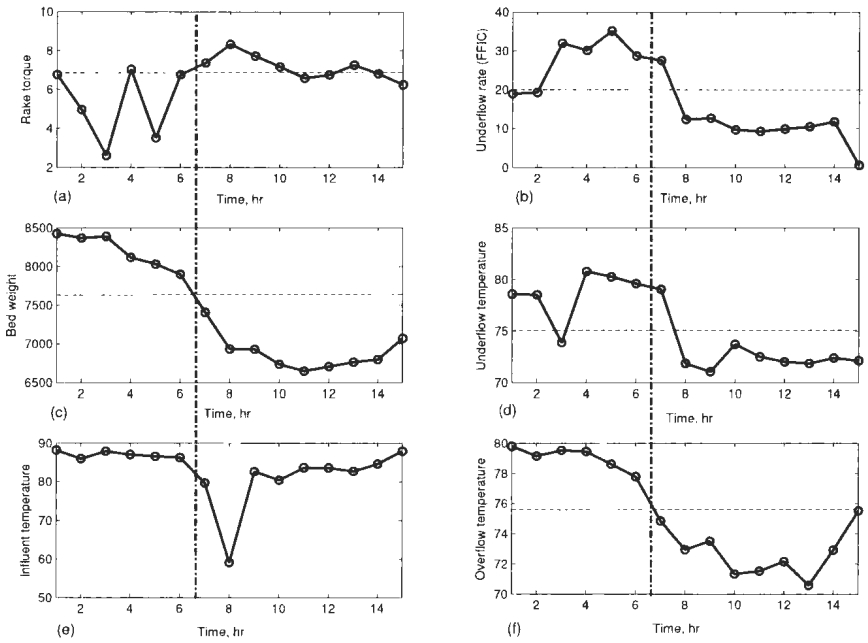

Figure 5.10: Trend plots of different variables of validation data set 1 for LRT thickener

in 5.10 (c). In order to maintain bed weight, the underflow rate was decreased to a minimum level as shown in Figure 5.10 (b). However, it still was not sufficient to recover the system and the underflow was completely shut down.

Validation data set 2:

Figures 5.11 (a) and (b) show Hotelling's T^2 and Q-residuals plot, respectively, for the second validation data-set. Figure 5.11 (b) shows that at t=13.0 hrs, value of Q-residual increased sharply indicating that the process became unstable.

The contribution plot in Figure 5.12 shows that influent temperature has the most contribution to fault occurrence at t=13.0 hrs. Figure 5.13 (e) clearly shows that, upto t=13.0 hours, influent temperature was constant and then it started to decrease sharply. In Figure 5.13 (a), if a threshold is drawn based on the previous peak at t=10.0 hrs, at t=15.0 hrs it will provide an indication to the operator that

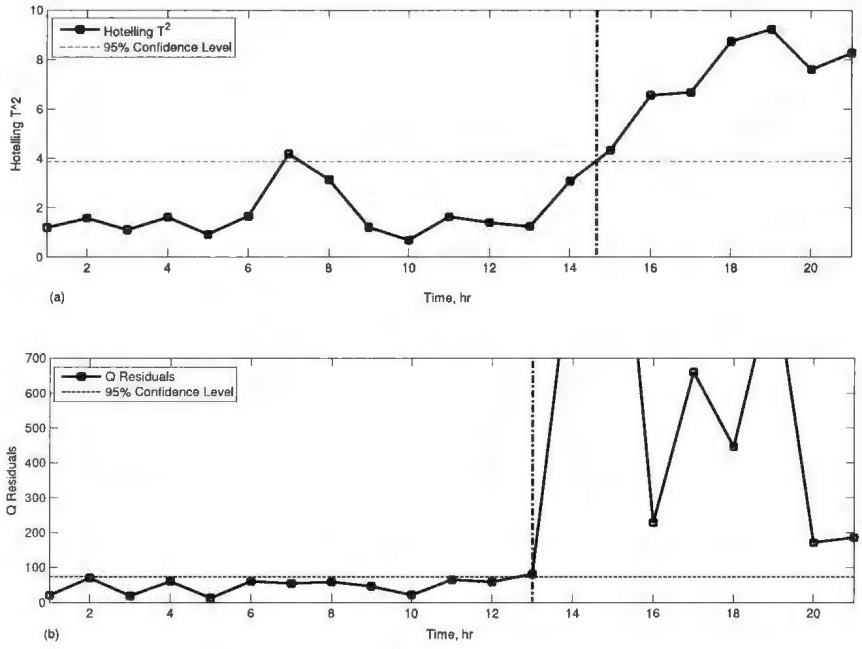

Figure 5.11: (a) Hotelling's T^2 plot and (b) Q-residuals plot of validation data set 2 for LRT thickener

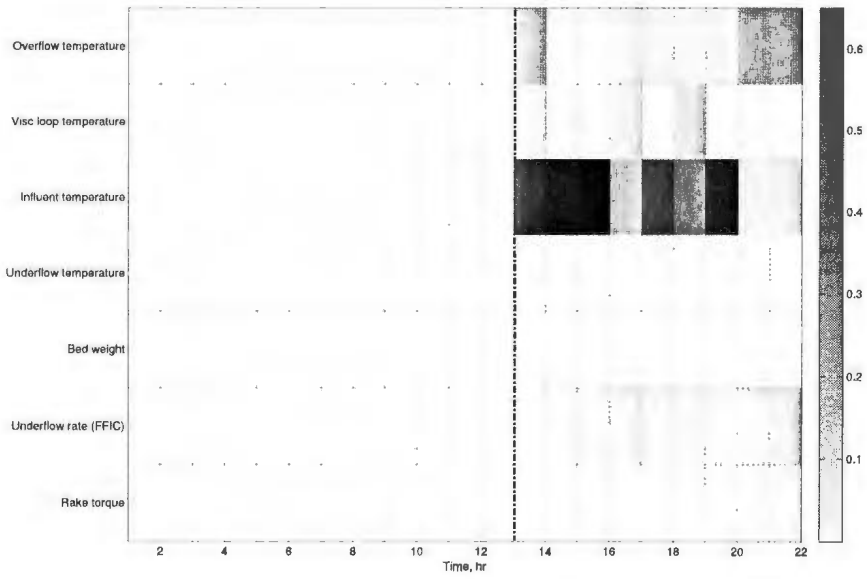

Figure 5.12: Color plot of validation data set 2 for LRT thickener

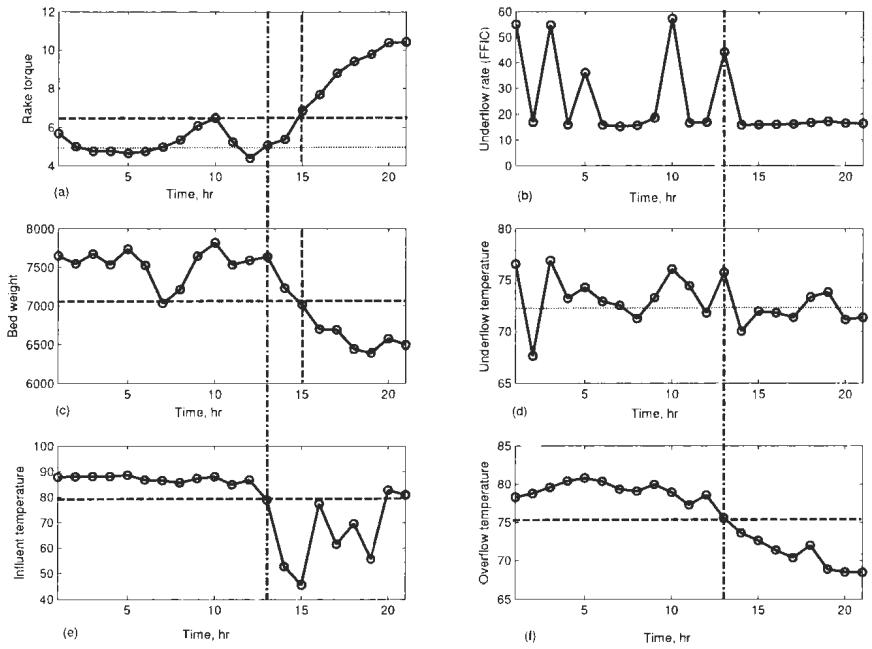

Figure 5.13: Trend plots of different variables of validation data set 2 for LRT thickener

rake torque has reached a critical limit. On the other hand, PCA model detects this fault two hours early at t=13.0 hrs, as well as it precisely indicates that influent temperature is the root cause of the fault. So the operator will have sufficient time to take corrective actions.

Effectiveness of variable selection:

The effectiveness of the proposed variable selection method is demonstrated by building two alternate models. The first one is built using all eleven variables and the second one is built using seven randomly chosen variables where the first four important variables selected by the proposed method are not included. These two models are then applied to validation data set 1 to compare the prediction abilities of the models. Figures 5.14 show (a) Hotelling's T^2 plot and (b) Q residuals plot of val-

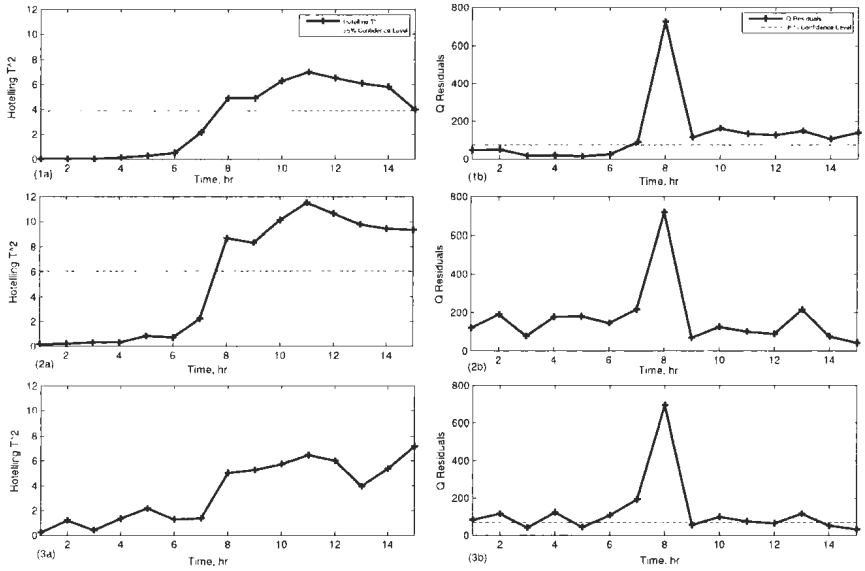

Figure 5.14: Comparison of proposed variable selection method for LRT thickener: [(a) Hotelling's T^2 plot and (b) Q residuals plot], where model is built (1) using seven variables selected from proposed variable selection method (2) using all eleven variables (3) using random seven variables

idation data-set 1 calculated using the three models described above. Comparison of Figure 5.14 (1b) with (2b) and (3b) clearly shows that prediction performance by the model using the proposed variable selection method is consistent, whereas in the other two cases, most of the time, Q-residual is above the confidence limit showing false fault detection.

5.6.2 CCD 1 thickener

5.6.2.1 Data Description

Figure 5.15 shows a schematic diagram of CCD 1 thickener with sensor locations of the preliminary selected fifteen process variables. The description of each variable is summarized in Table 5.6. These fifteen variables were used as input to the variable selection process.

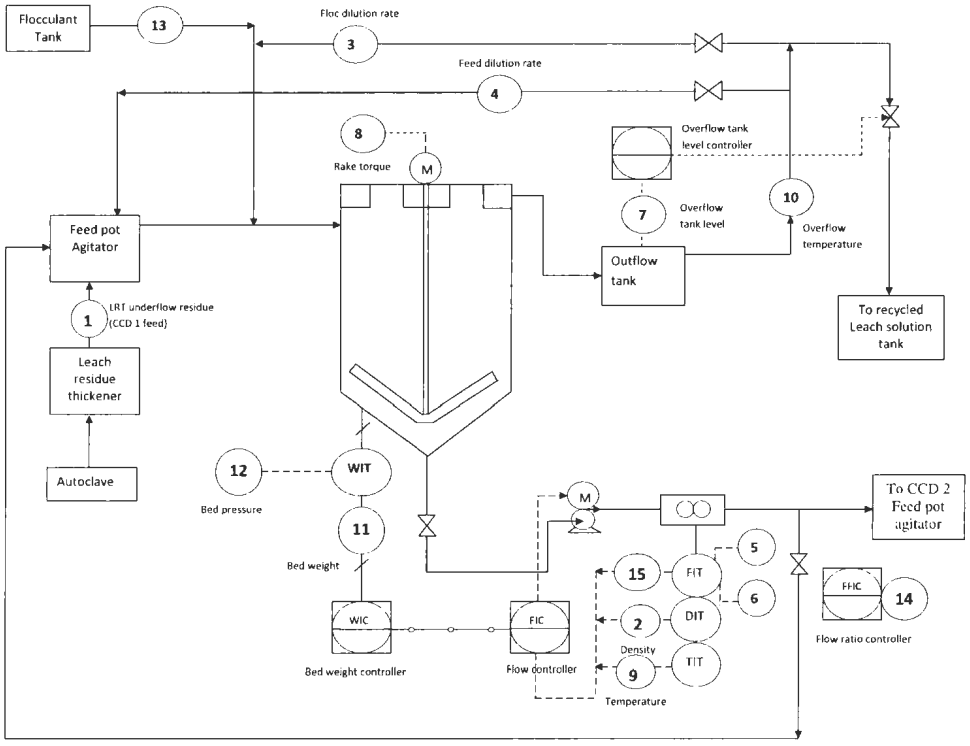

Figure 5.15: Schematic diagram of CCD 1 thickener

Table 5.6: List of preliminarily selected variables for CCD 1 thickener used in proposed variable selection process

No.	Variable name	Description
1	Feed rate	LRT underflow FFIC controller's output (%)
2	Underflow density	Underflow slurry density(gm/cc)
3	Floc dilution rate	Recycled overflow volume to dilute floc (m3/hr)
4	Feed dilution rate	Recycled overflow volume to dilute feed (m3/hr)
5	Underflow slurry flow	Solids mass flow from FIT transmitter (kg/hr)
6	Underflow solids %	Solids % in underflow slurry going to CCD 2 (%)
7	Overflow tank level	Meausred PV in overflow tank level controller (%)
8	Rake torque	Torque measured from power requirement (%)
9	Underflow temperature	Underflow slurry temperature (°C)
10	Overflow temperature	Measured overflow temperature (°C)
11	Bed weight	Measured PV in bed weight controller (kg)
12	Bed pressure	Solids loading of thickener bed(kPa)
13	Floc volume	Floc volume added to feed stream (m3/h)
14	Underflow rate (FFIC)	Measured PV in ratio controller (FFIC)
15	Underflow flow rate (FIC)	Measured PV in flow-controller (FIC) (kg/h)

5.6.2.2 Variable selection

Table 5.7: Selection of levels of data for designing Taguchi orthogonal array using historical data for CCD 1 thickener

No.	Variable	Min	Max	Range	Limit[a]	LL[b]	HL[c]
1	Feed rate	0.103	62.75	62.65	25.06	25.2	37.7
2	Underflow density	1.329	1.89	0.56	0.22	1.6	1.7
3	Floc dilution rate	0.00001	1.001	1.001	0.40	0.4	0.6
4	Feed dilution rate	0.003	5.02	5.02	2.01	2.01	3.02
5	Underflow slurry flow	17.31	587.2	569.9	227.97	245.3	359.3
6	Underflow solids %	13.37	55.44	42.07	16.83	30.2	38.6
7	Overflow tank level	21.31	103.06	81.75	32.70	54.0	70.4
8	Rake torque	0.014	18.61	18.59	7.44	7.5	11.2
9	Underflow temperature	26.74	51.93	25.19	10.07	36.8	41.9
10	Overflow temperature	23.65	70.35	46.70	18.68	42.3	51.7
11	Bed weight	3567.37	6699.0	3131.7	1252.7	4820	5446
12	Bed pressure	34.65	37.92	3.27	1.31	35.9	36.6
13	Floc volume	0.00002	0.014	0.0138	0.0055	0.006	0.008
14	Underflow rate (FFIC)	0.121	47.57	47.45	18.98	19.1	28.6
15	Underflow flow rate (FIC)	36.84	1088.5	1051.6	420.66	457.5	667.8

[a]Limit is 40 percent of range value
[b]Calculated by adding limit value with minimum value; low level data range is from min to LL
[c]Calculated by subtracting limit value from maximum value; high level data range is from HL to max

Using fifteen preliminarily selected variables, first, a PCA model was built to calculate the Hotelling's T^2 values. The ranges of the levels of fifteen variables were also calculated as shown in Table 5.7. Table 5.8 reports the standard orthogonal array for fifteen variables filled in with low and high level values. Table 5.9 reports the corresponding Hotelling's T^2 values as the outcome variable of nine trials conducted for each experiment. S/N raio was calculated for each experiment as given in Table 5.9. The overall S/N ratio and ranks for the variables are given in Table 5.10.

From the rank of the variables, top ten variables which have S/N ratios >0.5 were selected to build the final model.

Table 5.8: Taguchi orthogonal array with low-level and high-level values of fifteen variables to design the experiments for CCD 1 thickener

	Variable number [a]														
Exp.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	**25.2**	**1.6**	**0.4**	**2.01**	**245.3**	**30.2**	**54.0**	**7.5**	**36.8**	**42.3**	**4820**	**35.9**	**0.006**	**19.1**	**457.5**
2	**25.2**	**1.6**	**0.4**	**2.01**	**245.3**	**30.2**	**54.0**	11.2	41.9	51.7	5446	36.6	0.008	28.6	667.8
3	**25.2**	**1.6**	**0.4**	3.02	359.3	38.6	70.4	**7.5**	**36.8**	**42.3**	**4820**	36.6	0.008	28.6	667.8
4	**25.2**	**1.6**	**0.4**	3.02	359.3	38.6	70.4	11.2	41.9	51.7	5446	**35.9**	**0.006**	**19.1**	**457.5**
5	**25.2**	1.7	0.6	**2.01**	**245.3**	38.6	70.4	**7.5**	**36.8**	51.7	5446	**35.9**	**0.006**	28.6	667.8
6	**25.2**	1.7	0.6	**2.01**	**245.3**	38.6	70.4	11.2	41.9	**42.3**	**4820**	36.6	0.008	**19.1**	**457.5**
7	**25.2**	1.7	0.6	**2.01**	**245.3**	38.6	70.4	11.2	41.9	**42.3**	**4820**	36.6	0.008	**19.1**	**457.5**
8	**25.2**	1.7	0.6	3.02	359.3	**30.2**	**54.0**	11.2	41.9	**42.3**	**4820**	**35.9**	**0.006**	28.6	667.8
9	37.7	**1.6**	0.6	**2.01**	359.3	**30.2**	70.4	**7.5**	41.9	**42.3**	5446	**35.9**	0.008	**19.1**	667.8
10	37.7	**1.6**	0.6	**2.01**	359.3	**30.2**	70.4	11.2	**36.8**	51.7	**4820**	36.6	**0.006**	28.6	**457.5**
11	37.7	**1.6**	0.6	3.02	**245.3**	38.6	**54.0**	**7.5**	41.9	**42.3**	5446	36.6	**0.006**	28.6	**457.5**
12	37.7	**1.6**	0.6	3.02	**245.3**	38.6	**54.0**	11.2	**36.8**	51.7	**4820**	**35.9**	0.008	**19.1**	667.8
13	37.7	1.7	**0.4**	**2.01**	359.3	38.6	**54.0**	**7.5**	41.9	51.7	**4820**	**35.9**	0.008	28.6	**457.5**
14	37.7	1.7	**0.4**	**2.01**	359.3	38.6	**54.0**	11.2	**36.8**	**42.3**	5446	36.6	**0.006**	**19.1**	667.8
15	37.7	1.7	**0.4**	3.02	**245.3**	**30.2**	70.4	**7.5**	41.9	51.7	**4820**	36.6	**0.006**	**19.1**	667.8
16	37.7	1.7	**0.4**	3.02	**245.3**	**30.2**	70.4	11.2	**36.8**	**42.3**	5446	**35.9**	0.008	28.6	**457.5**

[a]Bold faced-low level values; normal-high level values

Table 5.9: Calculation of S/N ratio of each experiment in the orthogonal array for CCD 1 thickener

	Y_i= Hotteling T^2 value for each trial														
Exp.	1	2	3	4	5	6	7	8	9	N	Sm_1 [a]	ST_1 [b]	Se_1 [c]	Ve_1 [d]	S/N [e]
1	2.2	2.0	2.1	3.8	4.9	5.2	2.1	4.8	2.3	9	96.2	111.5	15.3	1.9	7.38
2	0.8	0.7	1.6	12.6	0.6	1.0	0.8	0.6	0.8	9	42.0	164.3	122.3	15.3	-7.11
3	0.9	0.8	2.9	2.2	3.3	1.9	3.3	1.4	0.7	9	33.7	42.5	8.8	1.1	5.20
4	5.5	1.9	1.8	0.5	2.5	2.1	0.8	1.2	1.5	9	35.2	52.3	17.1	2.1	2.34
5	5.0	2.6	0.4	1.1	2.9	2.2	1.4	7.0	7.2	9	98.7	149.0	50.3	6.3	2.13
6	2.6	1.2	1.8	4.9	1.8	4.5	9.5	7.2	0.6	9	129.6	202.0	72.4	9.1	1.70
7	2.7	2.6	1.7	1.8	2.4	5.5	0.6	1.5	0.8	9	42.6	59.1	16.6	2.1	3.37
8	1.7	1.4	1.7	1.0	2.4	4.9	0.5	3.1	0.6	9	33.9	49.6	15.7	2.0	2.56
9	0.0	2.2	1.7	2.4	6.5	2.2	1.4	1.2	1.3	9	39.9	65.3	25.4	3.2	1.09
10	1.4	3.4	1.9	1.0	1.9	0.4	0.8	0.6	0.8	9	16.5	23.3	6.8	0.9	3.08
11	5.1	12.6	4.2	2.6	2.3	2.6	2.1	1.2	2.9	9	141.4	234.6	93.3	11.7	0.92
12	1.0	0.4	2.0	1.9	0.5	2.3	1.5	1.2	2.2	9	19.1	23.0	3.9	0.5	6.31
13	1.2	1.9	2.0	2.4	1.9	4.1	6.5	4.5	3.3	9	86.6	109.5	22.9	2.9	5.12
14	3.3	0.8	1.0	5.5	3.0	2.2	0.8	0.6	0.8	9	36.2	58.3	22.1	2.8	1.30
15	2.1	0.8	2.4	1.0	4.3	3.1	2.4	0.6	2.2	9	39.4	50.8	11.4	1.4	4.72
16	2.2	1.0	1.9	0.6	0.6	1.5	0.4	5.9	1.0	9	25.2	48.4	23.2	2.9	-0.67

[a] $Sm_1 = (\sum_1^9 Y_i)/N$

[b] $ST_1 = \sum_1^9 Y_i^2$

[c] $Se_1 = ST_1 - Sm_1$

[d] $Ve_1 - Se_1/(N\text{-}1)$

[e] $S/N = 10 \log \left[\frac{(Sm_1 - Ve_1)}{N \times Ve_1}\right]$

Table 5.10: Calculation of S/N ratio of each variable for CCD 1 thickener

No.	Variable	LLC[a]	HLC [b]	S/N ratio [c]	Rank
1	Feed rate	2.197	2.732	0.54	10
2	Underflow density	2.401	2.528	0.13	15
3	Floc dilution rate	2.284	2.644	0.36	12
4	Feed dilution rate	2.005	3.055	1.05	8
5	Underflow slurry flow	2.082	1.604	0.48	11
6	Underflow solids %	1.578	3.154	1.58	6
7	Overflow tank level	2.354	2.550	0.20	13
8	Rake torque	3.793	1.430	2.36	2
9	Underflow temperature	3.531	1.634	1.90	4
10	Overflow temperature	2.539	2.368	0.17	14
11	Bed weight	4.382	-0.001	4.38	1
12	Bed pressure	3.282	1.647	1.64	5
13	Floc volume	3.054	1.874	1.18	7
14	Underflow rate (FFIC)	3.526	1.402	2.12	3
15	Underflow flow rate (FIC)	2.904	2.024	0.88	9

[a]'Low level contribution' for each variable is calculated as average of S/N ratios of those experiments in the orthogonal array where the variable is contributing as low level

[b]'High level contribution' for each variable is calculated as average of S/N ratios of those experiments in the orthogonal array where the variable is contributing as high level

[c]Absolute difference between HLC and LLC

5.6.2.3 Fault detection model

The selected variables were mean-centered and scaled by the reciprocal of the square root of standard deviation. Figure 5.16 (a) shows the eigenvalue plot for ten principal components. From the cumulative variance captured (%) plot shown in Figure 5.16 (b), two principal components were selected, capturing 97.64 % of total variance. Figures 5.17 (a) and (b) show T^2 and Q-residuals plots of the model, respectively.

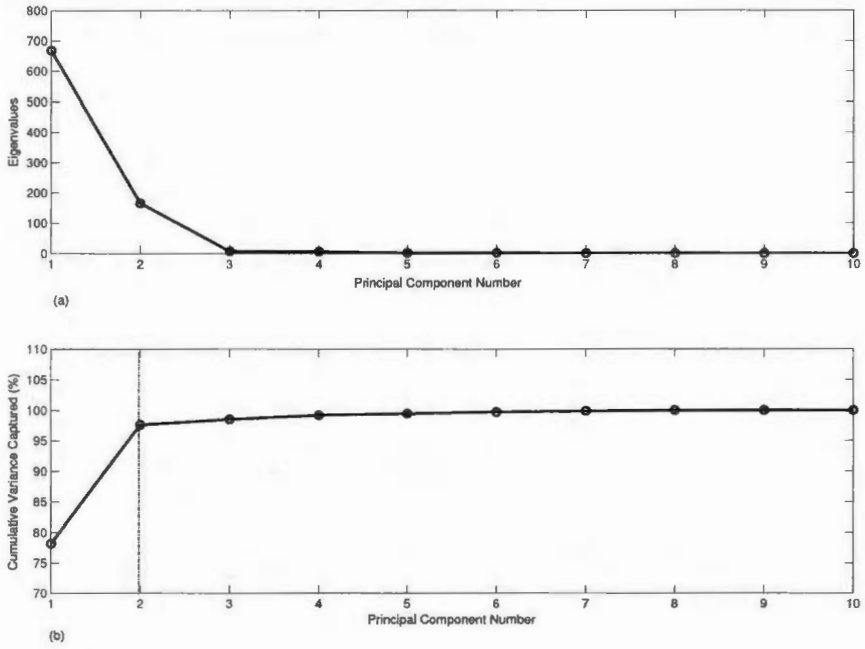

Figure 5.16: (a) Eigen-value plot and (b) cumulative variance captured (%) plot for CCD 1 thickener

5.6.2.4 Validation

In order to validate the fault detection and diagnosis capability of the model, two faulty data-sets were selected where the process was impacted by fault which eventually led to a temporary shutdown of the unit.

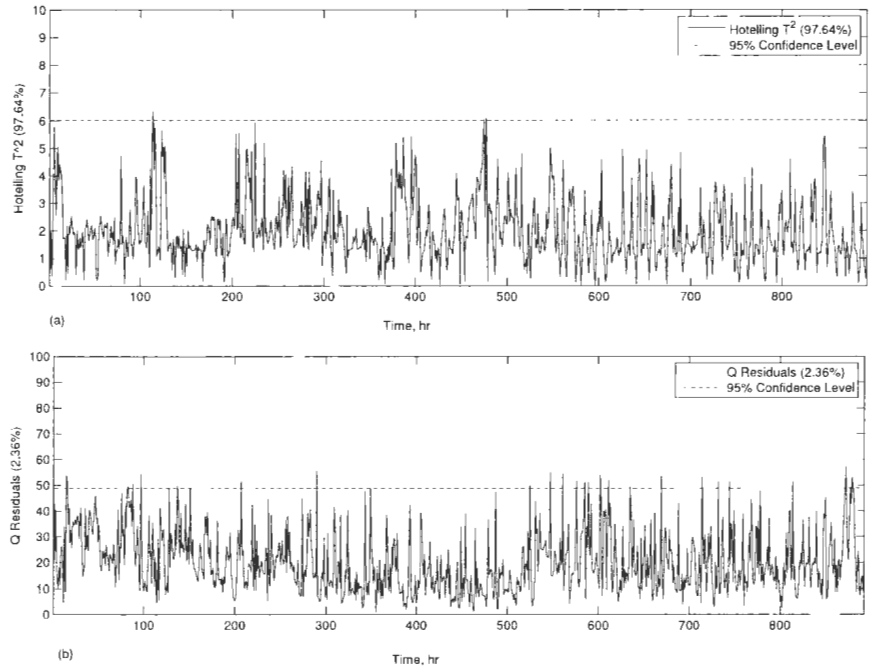

Figure 5.17: (a) Hotelling's T^2 plot and (b) Q statistics plot of PCA model for CCD 1 thickener

Validation data-set 1:

Figures 5.18 (a) and (b) report the T^2 and Q-residuals plots of validation data-set 1, respectively. Figure 5.18 (a) indicates that at t=13.4 hrs value of T^2 crossed the confidence limit and remained outside the limit upto t=20.0 hrs. Figure 5.19 illustrates the residual contribution of each variable over time using a color plot. The plot shows that underflow solids % has the most contribution for fault occurrence at t=13.4 hrs.

In order to ascertain the root cause, each variable was further investigated. Figure 5.20 (e) shows that at t=17.0 hrs, underflow solids % started to decrease. The PCA model detected the fault 3.6 hrs early at t=13.4 hrs. A decrease in underflow solids % is an indication of poor flocculation. The disturbance in the flocculation process causes direct channeling of feed through underflow, which eventually lowers the solids content of the underflow slurry. Figure 5.20 (g) shows that, at t=16.0 hrs,

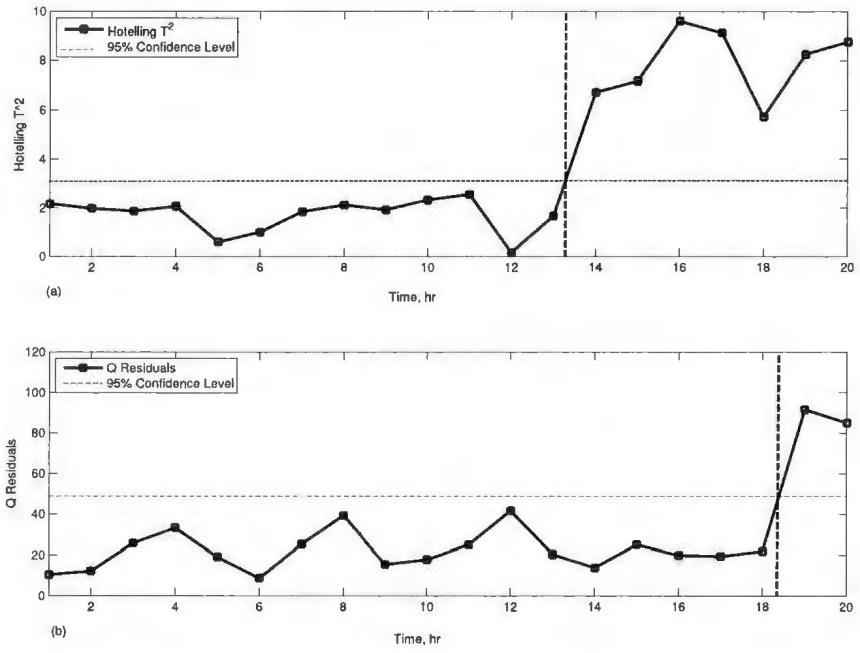

Figure 5.18: (a) Hotelling's T^2 plot and (b) Q residuals plot of validation data set 1 for CCD 1 thickener

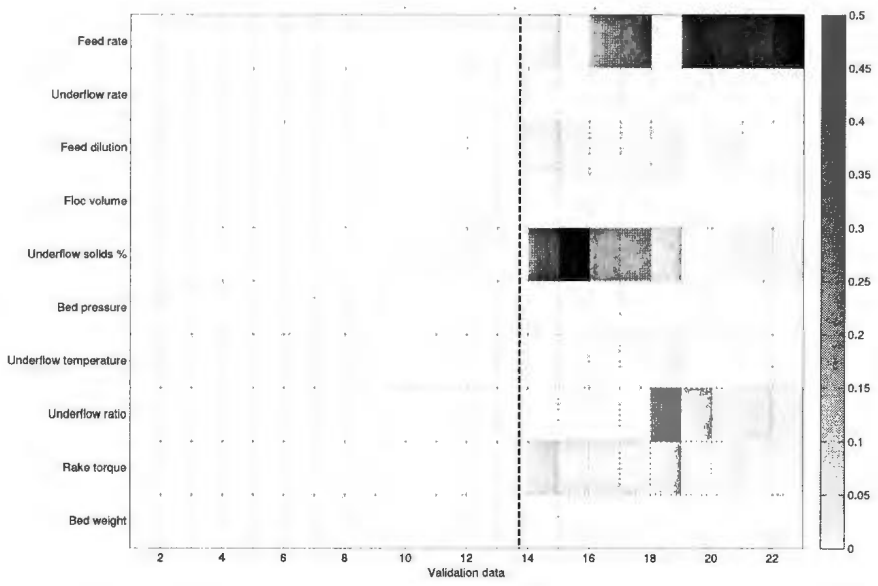

Figure 5.19: Color plot of validation data set 1 for CCD 1 thickener

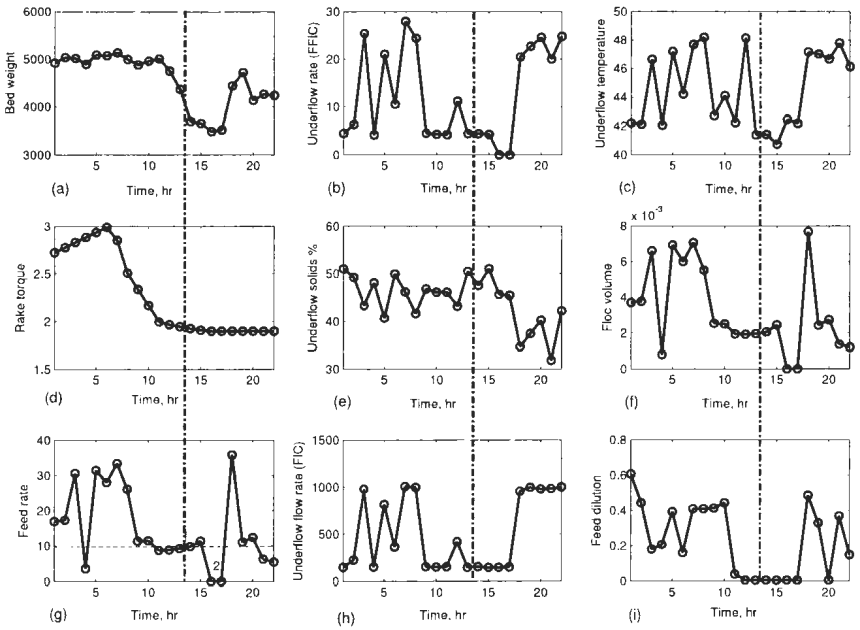

Figure 5.20: Trend plots of different variables of validation data set 1 for CCD 1 thickener

feed was stopped due to disturbance in the flocculation process. Attempts were taken to stabilize the process by increasing the feed rate at t=17.0 hrs; however, it still was not sufficient to recover the system and the unit had to stop at t=22.0 hrs.

Validation data-set 2:

Figures 5.21 (a) and (b) show T^2 and Q-residuals plots, respectively, for validation data-set 2. Figure 5.21 (a) shows that at t=7.4 hrs T^2 value crossed the confidence limit and remained outside the limit upto t=22.0 hrs. Figure 5.22 illustrates the residual contribution of each variable over time using a color plot. It clearly shows that underflow solids % has the most residual contribution for fault occurrence at t=7.4 hrs. In order to ascertain the root cause, each variable was further investigated. Figure 5.23 (c) shows that at t=11.0 hrs underflow solids % started to decrease. Figure 5.23 (a) & (d) show that both bed weight and rake torque started to decrease at t=10.0

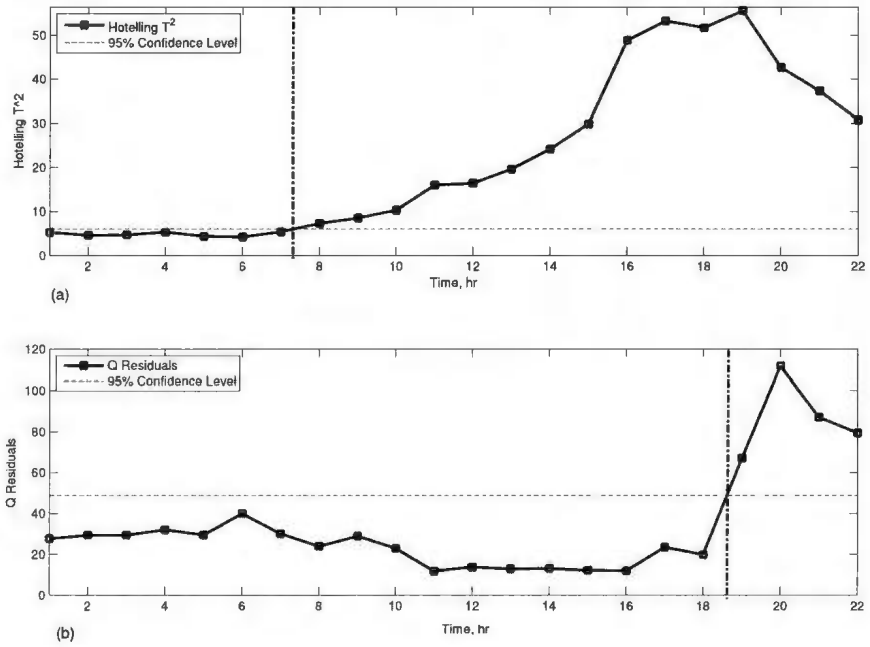

Figure 5.21: (a) Hotelling's T^2 plot and (b) Q residuals plot of validation data-set 2 for CCD 1 thickener

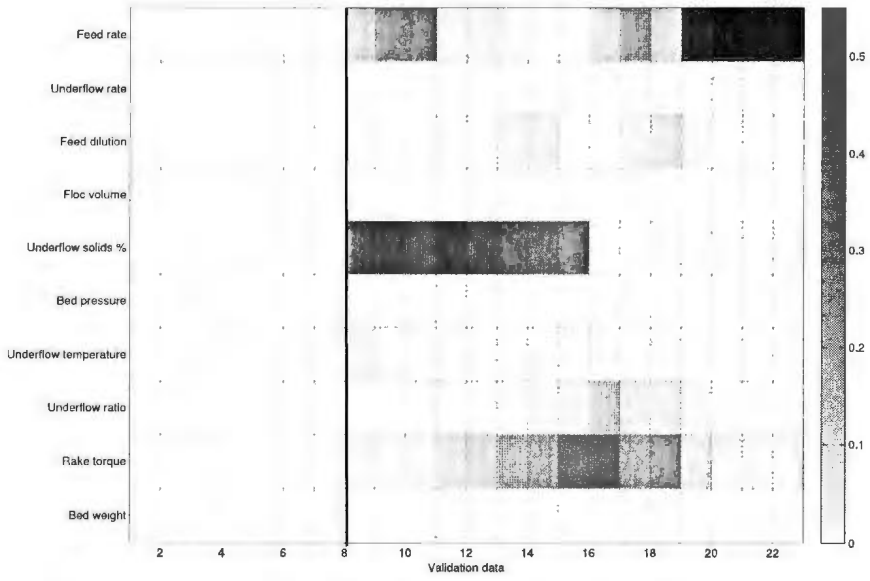

Figure 5.22: Color plot of CCD 1 thickener validation data set 2

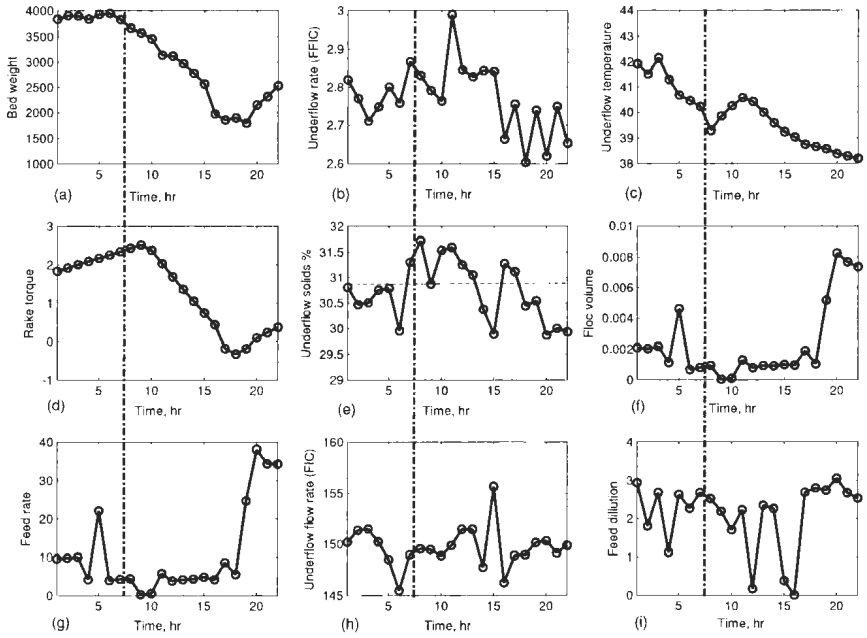

Figure 5.23: Trend plots of different variables of validation data set 2 for CCD 1 thickener

hrs. This indicates that improper flocculation resulted in direct channeling of slurry. The PCA model detects the fault 2.6 hrs early.

5.6.3 CCD 2 thickener

5.6.3.1 Data Description

In a CCD circuit, each thickener has an identical list of variables and control schemes. A similar set of variables selected for CCD 1 PCA model was used to build the PCA model for CCD 2 thickener.

5.6.3.2 Fault detection model

The selected variables were mean-centered and scaled by the reciprocal of square root of standard deviation. Figure 5.24 (a) shows the eigenvalue plot for nine principal

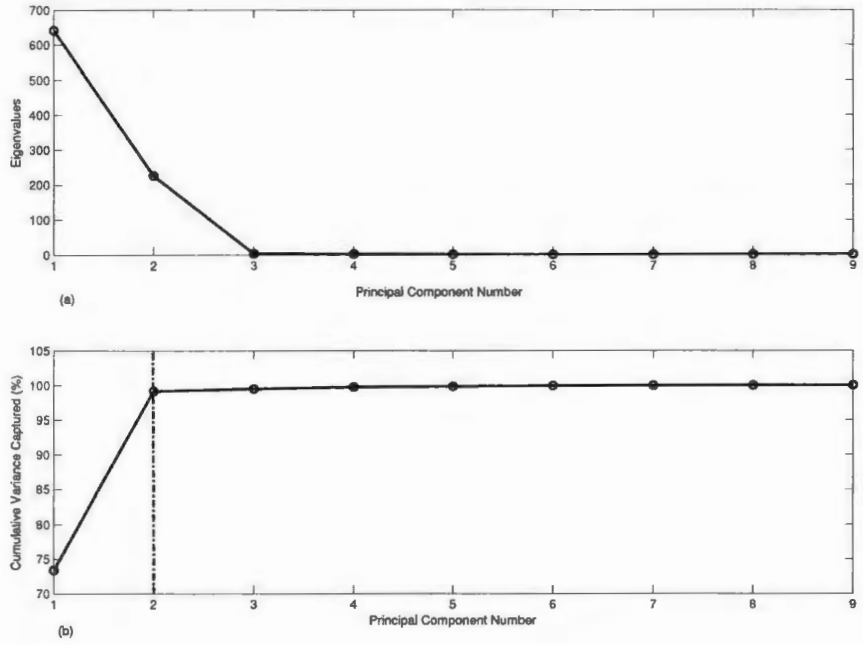

Figure 5.24: (a) Eigen-value plot and (b) cumulative variance captured (%) plot for CCD 2 thickener

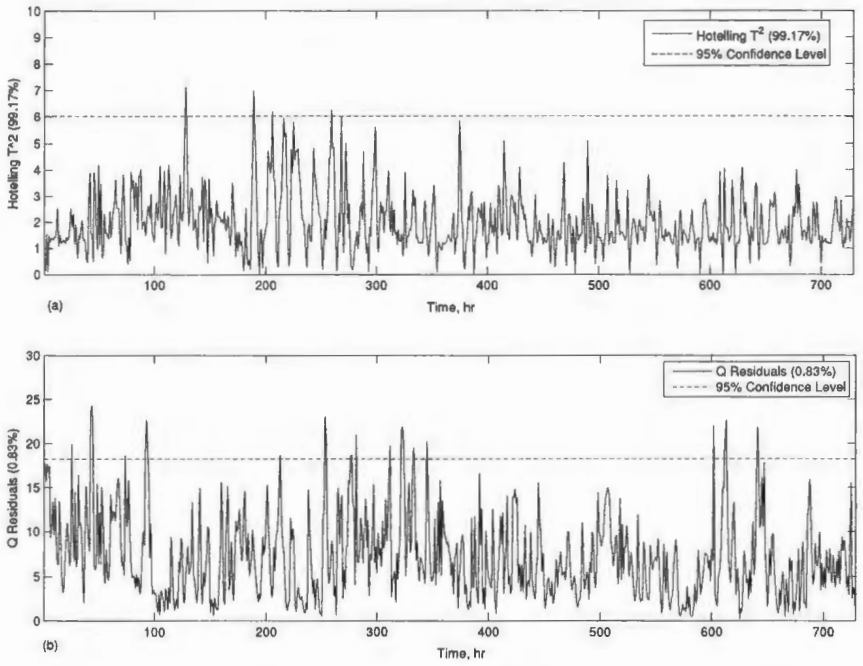

Figure 5.25: (a) Hotelling's T^2 plot and (b) Q residuals plot of PCA model for CCD 2 thickener

components. From the cumulative variance captured (%) plot as shown in Figure 5.24 (b), two principal components are selected capturing 99.17 % of total variance. Figures 5.25 (a) and (b) show Hotelling's T^2 and Q residuals plots, respectively, of the final model.

5.6.3.3 Validation

The model was validated on two known faulty data-sets where the process was impacted by fault which eventually led to a temporary shutdown of the unit.

Validation data-set 1:

Figures 5.26 (a) and (b) report the T^2 and Q residuals plots of first validation data set, respectively. Figure 5.26 (b) indicates that at t=9.0 hrs value of Q residual crossed the confidence limit and remained outside the limit upto t=13.0 hrs. Figure 5.27 illustrates the residual contribution of each variable over time using a color plot. The plot shows that rake torque and feed dilution rate have major residual contributions for fault occurrence at t=9.0 hrs. In order to ascertain the root cause, each variable was further investigated. Figure 5.28 (d) & (a) show that at t=9.0 hrs rake torque started to increase, whereas bed weight was decreasing. Figure 5.28 (c) shows that feed dilution rate became unavailable during fault occurrence, which essentially deteriorated the flocculation process. The feed became relatively viscous due to lack of dilution. The bed material lost its fluidity, which eventually increased rake torque and led to the shutdown.

Validation data-set 2:

Figures 5.29 (a) and (b) show T^2 and Q residuals plots of the second validation data-set, respectively. Figure 5.29 (b) indicates that at t=10.5 hrs the value of Q residuals

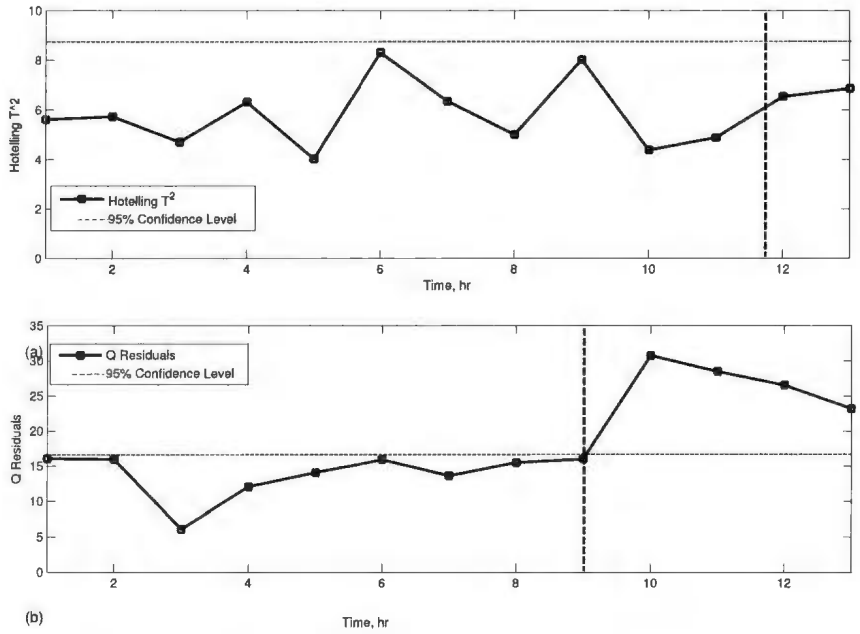

Figure 5.26: (a) Hotelling's T^2 plot and (b) Q residuals plot of validation data-set 1 for CCD 2 thickener

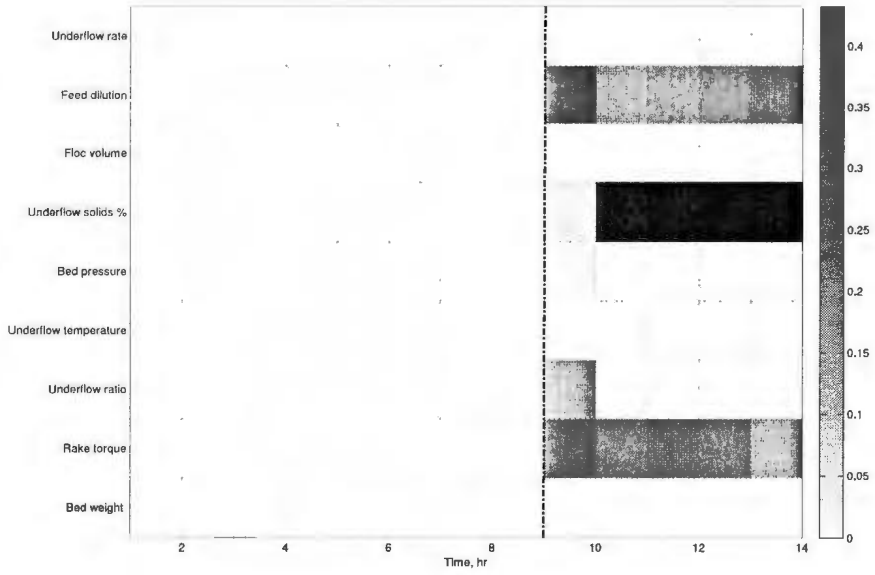

Figure 5.27: Color plot of validation data set 1 for CCD 2 thickener

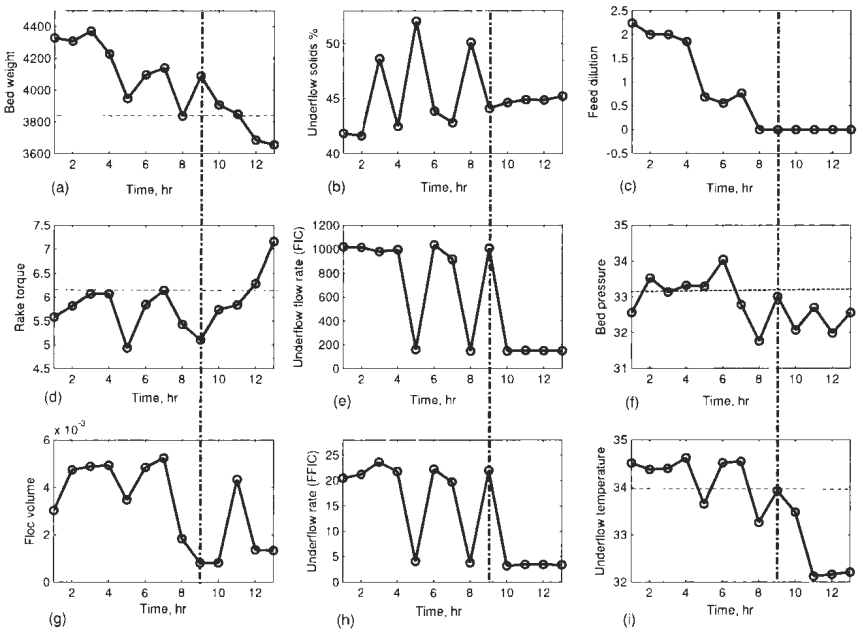

Figure 5.28: Trend plots of different variables of validation data set 1 for CCD 2 thickener

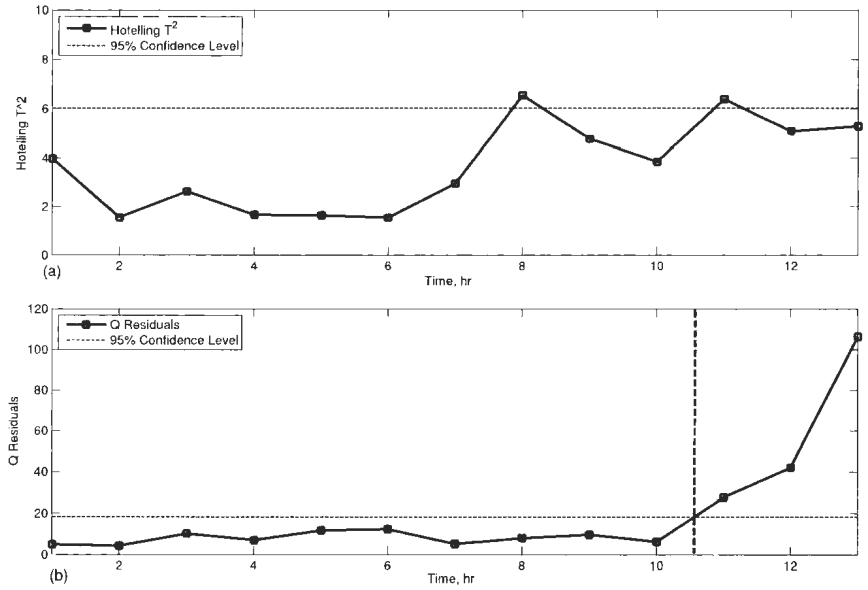

Figure 5.29: (a) Hotelling's T^2 plot and (b) Q residuals plot of validation data set 2 for CCD 2 thickener

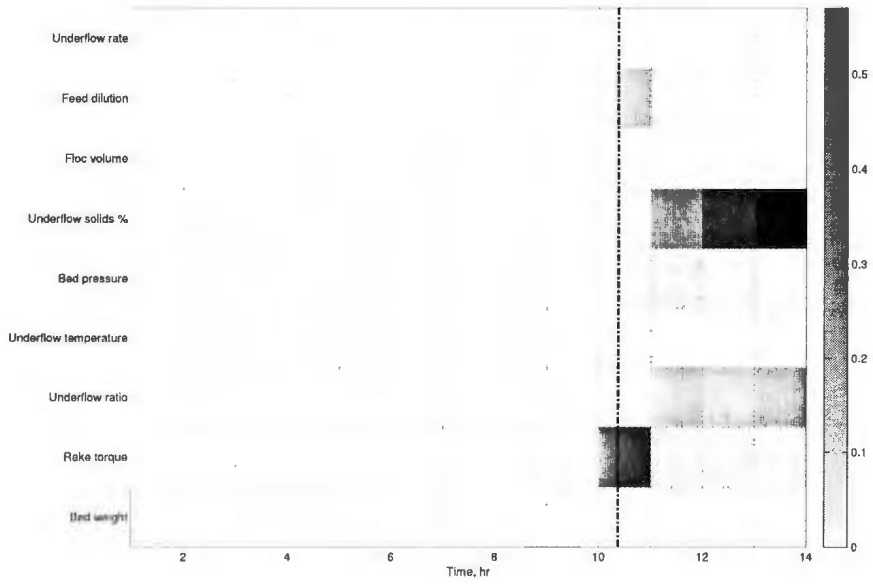

Figure 5.30: Color plot of validation data-set 2 for CCD 2 thickener

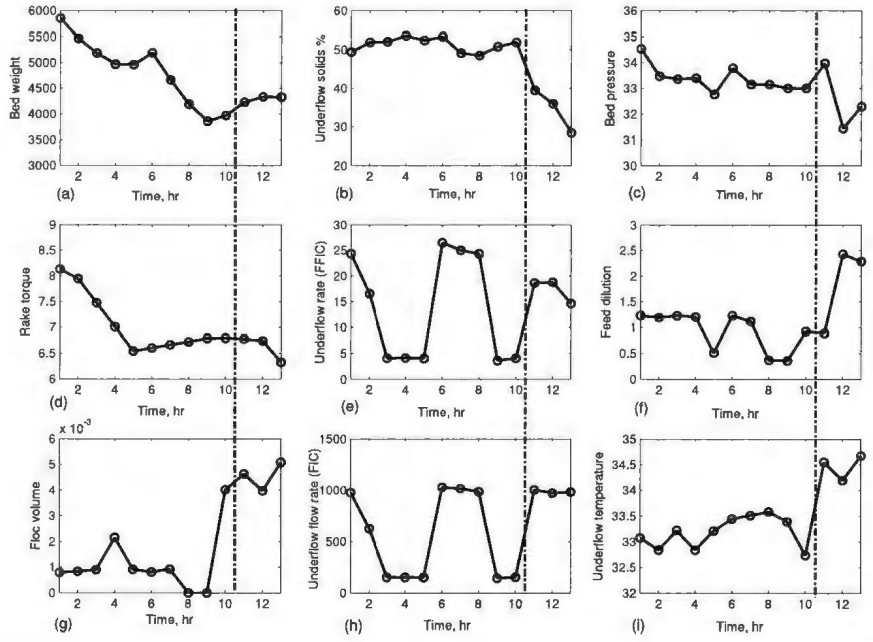

Figure 5.31: Trend plots of different variables of validation data set 2 for CCD 2 thickener

crossed the confidence limit and remained outside the limit up to t=13.0 hrs.

Figure 5.30 illustrates the residual contribution of each variable over time using a color plot. The plot shows that rake torque has the most residual contributions for fault occurrence at t=10.5 hrs. In order to ascertain the root cause, each variable was further investigated. Figure 5.31 (g) shows that floc volume rate started to increase at t=9.0 hrs. As floc addition is proportional to feed rate, this indicates that feed rate was also increased. Figure 5.31 (d) shows that rake torque did not increase with the feed rate. Figure 5.31 (b) also reflects this fact where a decrease in underflow solids % is observed. This indicates that a disturbance was created in the settling process resulting in direct channeling of feed material. Figure 5.31 (f) also indicates that feed dilution rate was increased at t=11.0 hrs to improve flocculation. This action could have been taken an half hour before at t=10.5 hrs when PCA model detected the fault.

5.7 Conclusions

The present chapter discussed a systematic quantitative approach to identify important input variables for a PCA model. The method is based on Taguchi's experimental design and employs historical process data to select important variables. The technique considers the calculated Hotelling's T^2 value as an outcome variable for each designed experiment of the orthogonal array. The proposed Taguchi-based methodology is applied in combination with PCA for building monitoring scheme for LRT thickener, CCD 1 thickener and CCD 2 thickener of a nickel hydromet process. The models are validated using process data with known faults. Contribution plots were used to diagnose the root cause of fault. The major advantages of the proposed method are stated below:

- The proposed Taguchi based methodology offers a quantitative and systematic way of selecting input variables for PCA. It can significantly reduce the models development time for PCA based monitoring tool.

- PCA based monitoring technique can be effectively used for detection and diagnosis of faults in the thickener units of a hydro-metallurgy process. It provides early warnings and is able to diagnose the root cause of a fault effectively.

- The effectiveness of the proposed variable selection method is demonstrated on LRT model. The results show that the variable selection method improves PCA model quality which provides consistency in detecting faults and avoids false fault detection.

Chapter 6

Conclusions

6.1 Contributions

The major contributions of this thesis are listed below:

- Taguchi's experimental design method has been adapted for selecting input variables for process monitoring tools. Detailed methodologies have been developed to select input variables for SVR based inferential predictor and PCA based fault detection and diagnosis method.

- Implementation difficulties in applying Taguchi method to process data were addressed. Taguchi's experimental design array was originally developed for uncorrelated factors. Since process variables are correlated, it becomes difficult to fill the design array using historical process data. In order to overcome this difficulty a classification algorithm was used to classify process variables into different uncorrelated groups. A representative variable was selected from each group. Taguchi's experimental design array was designed considering the representative variables as factors. Since these representative variables are uncorrelated, the design array can be easily filled using historical data.

- A SVR based inferential predictor was developed to predict the 4-CBA concentration of a PTA process. The SVR based predictor successfully modeled process non-linearly and showed better prediction capability compared to a PLS model.

- Input variables for the SVR predictor was selected using the proposed Taguchi's experimental design based variable selection method. Prediction performance of the Taguchi-SVR model was compared with VIP-SVR model, which used VIP method to select input variables. Results show that Taguchi-SVR has less prediction error (RMSE) compared to VIP-SVR.

- A variable selection methodology based on Taguchi's experimental design method is developed for a PCA based monitoring scheme. This systematic, quantitative method can replace the trial and error and significantly minimize the modeling time for PCA. Also the results showed that the PCA model with variables selected using the proposed method has less false alarm compared to the PCA model with larger set of variables.

- PCA based monitoring technique was effectively used to detect and diagnose faults in the thickener units of a hydro-metallurgy process. It was demonstrated using industrial data that the monitoring scheme provided early warnings and capable to diagnose the root cause of a fault effectively.

6.2 Future Recommendations

- The proposed Taguchi based variable selection methodology has been used to select variables for a SVR predictor. The methodology is demonstrated through a case study from a petrochemical process. The main advantage of the method is that it is not dependent on any learning algorithm. Therefore, the proposed

method can also be applied to other multivariate regression methods.

- PCA model with the proposed variable selection methodology is successfully applied for detection and diagnoses of faults in three different thickener units of a hydro-met process. The method can be further validated by applying it to other units of the process, e.g. flotation unit, grinding unit, autoclave, metal extraction unit.

- For filling the Taguchi's experimental design array, data search was carried out manually. This can be sometimes a tedious process. An automated data search method will definitely improve the usability of the method.

Chapter 7

References

Bibliography

[Abrahamsson et al., 2003] Abrahamsson, C., Johansson, J., Sparen, A., and Lindgren, F. (2003). Comparison of different variable selection methods conducted on nir transmission measurements on intact tablets. *Chemometrics and Intelligent Laboratory Systems*, 69:3–12.

[Andersen and Bro, 2010] Andersen, C. M. and Bro, R. (2010). Variable selection in regression-a tutorial. *Journal of Chemometrics*, 24(11-12):728–737.

[Antony and Antony, 2001] Antony, J. and Antony, F. J. (2001). Teaching the taguchi method to industrial engineers. *Work Study*, 50(4):141–149.

[Bakshi, 1998] Bakshi, B. R. (1998). Multiscale PCA with application to multivariate statistical process monitoring. *AIChE Journal*, 44(7):1596–1610.

[Bergh and Acosta, 2009] Bergh, L. and Acosta, S. (2009). Online fault detection on a pilot flotation column using linear PCA models. *Computer Aided Chemical Engineering*, 27:1437–1442.

[Cobb and Clarkson, 1994] Cobb, B. D. and Clarkson, J. M. (1994). A simple procedure for optimising the polymerase chain reaction (PCR) using modified Taguchi methods. *Nucleic Acids Research*, 22(18):3801–3805.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

[DeBoer, 1990] DeBoer, D. E. (1990). *The role of suspension characteristics in continuous gravity thickening.* PhD thesis, Iowa State University.

[Desai et al., 2006] Desai, K., Badhe, Y., Tambe, S. S., and Kulkarni, B. D. (2006). Soft-sensor development for fed-batch bioreacors using support vector regression. *Biochemical Engineering Journal*, 27:225 239.

[Dixon, 1979] Dixon, D. (1979). Theory of gravity thickening. *Progress in Filtration and Separation-1, ed. RJ Wakeman, Elsevier*, pages 113–178.

[Drucker et al., 1997] Drucker, H., Burges, C., Kaufman, L., Smola, A., and Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, 9:155–161.

[Eriksson et al., 2001] Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikstrom, C., and Wold, S. (2001). *Multi–and Megavariate data analysis, Part 2, Advanced application and method extensions.* Umetrics Academy.

[George, 2000] George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95(452):1304–1308.

[Ghani et al., 2004] Ghani, J., Choudhury, I., and Hassan, H. (2004). Application of Taguchi method in the optimization of end milling parameters. *Journal of Materials Processing Technology*, 145(1):84 92.

[Glasrud et al., 1993] Glasrud, G., Navarrete, R., Scriven, L., and Macosko, C. (1993). Settling behaviors of iron oxide suspensions. *AIChE journal*, 39(4):560–568.

[Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

[Hand et al., 2000] Hand, D. J., Fisher, D. H., and Berthold, M. R. (2000). Advances in intelligent data analysis. *Intelligent Data Analysis*, 4(2):93–95.

[Hatonen et al., 1999] Hatonen, J., Hyotyniemi, H., Miettunen, J., and Carlsson, L.-E. (1999). Using image information and partial least squares method to estimate mineral concentrations in mineral flotation. In *Intelligent Processing and Manufacturing of Materials, 1999. IPMM'99. Proceedings of the Second International Conference on*, volume 1, pages 459–464. IEEE.

[Herrera et al., 2006] Herrera, L., Pomares, H., Rojas, I., Verleysen, M., and Guilen, A. (2006). Effective input variable selection for function approximation. *Artificial Neural Networks-ICANN 2006*, pages 41–50.

[Hodouin, 2011] Hodouin, D. (2011). Methods for automatic control, observation, and optimization in mineral processing plants. *Journal of Process Control*, 21:211–225.

[Hodouin et al., 2001] Hodouin, D., Jämsä-Jounela, S.-L., Carvalho, M., and Bergh, L. (2001). State of the art and challenges in mineral processing control. *Control Engineering Practice*, 9(9):995–1005.

[Hongqiu et al., 2010] Hongqiu, Z., Chunhua, Y., and Weihua, G. (2010). Modeling of cobalt removal purification process in zinc hydrometallurgy based on fcm-svm. *Chinese High Technology Letters*, 10:016.

[Hoskuldsson, 2001] Hoskuldsson, A. (2001). Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 55:23–28.

[Imtiaz et al., 2007] Imtiaz, S., Shah, S., Patwardhan, R., Palizban, H., and Ruppenstein, J. (2007). Detection, diagnosis and root cause analysis of sheet-break in a pulp and paper mill with economic impact analysis. *The Canadian Journal of Chemical Engineering*, 85(4):512–525.

[Jackson and Mudholkar, 1979] Jackson, J. E. and Mudholkar, G. S. (1979). Control procedures for residuals associated with principal component analysis. *Technometrics*, 21:341–349.

[Jalali-Heravi et al., 2007] Jalali-Heravi, M., Shahbazikhah, P., Zekavat, B., and Ardejani, M. S. (2007). Principal component analysis-ranking as a variable selection method for the simulation of 13C nuclear magnetic resonance spectra of Xanthones using artificial neural networks. *QSAR & Combinatorial Science*, 26(6):764–772.

[Jemwa and Aldrich, 2006] Jemwa, G. T. and Aldrich, C. (2006). Kernel-based fault diagnosis on mineral processing plants. *Minerals Engineering*, 19:1149–1162.

[Khaw et al., 1995] Khaw, J. F., Lim, B., and Lim, L. E. (1995). Optimal design of neural networks using the Taguchi method. *Neurocomputing*, 7(3):225–245.

[Khoei et al., 2002] Khoei, A., Masters, I., and Gethin, D. (2002). Design optimisation of aluminium recycling processes using Taguchi technique. *Journal of Materials Processing Technology*, 127:96–106.

[Kourti and MacGregor, 1995] Kourti, T. and MacGregor, J. F. (1995). Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, 28:3–21.

[Kresta et al., 1991] Kresta, J. V., MacGregor, J. F., and Marlin, T. E. (1991). Multivariate statistical monitoring of process operating performance. *CJChE*, 69 (1):35–47.

[Kurt, 2006] Kurt, N. (2006). *A study of channelling behaviour in batch sedimentation.* PhD thesis, RMIT University.

[Lahiri and Ghanta, 2008] Lahiri, S. K. and Ghanta, K. C. (2008). The support vector regression with the parameter tuning assisted by a differential evolution technique: study of the critical velocity of a slurry flow in a pipeline. *Chemical Industry & Chemical Engineering Quarterly.* 14(3):191–203.

[Leardi, 2000] Leardi, R. (2000). Application of genetic algorithm PLS for feature selection in spectral data sets. *Journal of Chemometrics*, 14(5-6):643–655.

[Leardi and Lupianez Gonzalez, 1998] Leardi, R. and Lupianez Gonzalez, A. (1998). Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemometrics and Intelligent Laboratory Systems*, 41(2):195–207.

[Leardi et al., 2002] Leardi, R., Seasholtz, M. B., and Pell, R. J. (2002). Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from fourier transform-infrared spectral data. *Analytica Chimica Acta*, 461(2):189 200.

[Lin and Lin, 2002] Lin, J. and Lin, C. (2002). The use of the orthogonal array with grey relational analysis to optimize the electrical discharge machining process with multiple performance characteristics. *International Journal of Machine Tools and Manufacture*, 42(2):237–244.

[Moss, 1978] Moss, N. (1978). Theory of flocculation. *Mineral Eng. Handbook, Flocculation Section, Pile Not FldO1/0.*

[Osborne et al., 1997] Osborne, S. D., Kunnemeyer, R., and Jordan, R. B. (1997). Method of wavelength selection for partial least squares. *Analyst*, 122(12):1531–1537.

[Parvin et al., 2010] Parvin, H., Alizadeh, H., and Minaei-Bidgoli, B. (2010). A modification on k-nearest neighbor classifier. *Global Journal of Computer Science and Technology*, 10(14):37–41.

[Pierna et al., 2009] Pierna, F. A. F., , Abbas, O., Baeten, V., and Dardenne, P. (2009). A backward variable selection method for PLS regression (BVSPLS). *Analytica chimica acta*, 642(1):89-93.

[Qin, 2003] Qin, S. J. (2003). Statistical process monitoring : basics and beyond. *Journal of Chemometrics*, 17 (8-9):480–502.

[Rakotomamonjy, 2003] Rakotomamonjy, A. (2003). Variable selection using SVM based criteria. *The Journal of Machine Learning Research*, 3:1357–1370.

[Rossi et al., 2006] Rossi, F., Lendasse, A., Francois, D., Wertz, V., and Verleysen, M. (2006). Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and intelligent laboratory systems*, 80(2):215–226.

[Rowlands et al., 2000] Rowlands, H., Antony, J., and Knowles, G. (2000). An application of experimental design for process optimisation. *The TQM Magazine*, 12(2):78–84.

[Sanin et al., 2011] Sanin, F. D., Clarkson, W. W., and Vesilind, P. A. (2011). *Sludge engineering-the treatment and disposal of wastewater sludge*. DEStech publications, Inc.

[Sukthomya and Tannock, 2005] Sukthomya, W. and Tannock, J. D. (2005). Taguchi experimental design for manufacturing process optimisation using historical data and a neural network process model. *International Journal of Quality & Reliability Management*, Volume 22 issue 5:pp.485 – 502.

[Vapnik, 1995] Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* Springer Verlag, New York.

[Vapnik, 1998] Vapnik, V. (1998). *Statistical Learning Theory.* John Wiley, New York.

[Vesilind, 1968] Vesilind, P. A. (1968). Design of prototype thickeners from batch settling tests. *Water Sewage Works*, 115(7):302–307.

[Vlachogiannis and Roy, 2005] Vlachogiannis, J. G. and Roy, R. K. (2005). Robust PID controllers by taguchi's method. *The TQM magazine*, 17(5):456–466.

[Weston et al., 2001] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. (2001). Feature selection for SVMs. *Advances in neural information processing systems*, pages 668–674.

[Wise et al., 2007] Wise, B. M., Gallagher, N., Bro, R., Shaver, J., Windig, W., and Koch, R. S. (2007). PLS Toolbox 4.0. *Eigenvector Research Incorporated*, 3905.

[Wise and Gallagher, 1996] Wise, B. M. and Gallagher, N. B. (1996). The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, 6(6):329–348.

[Wu et al., 2001] Wu, M., She, J.-H., and Nakano, M. (2001). An expert control system using neural networks for the electrolytic process in zinc hydrometallurgy. *Engineering Applications of Artificial Intelligence*, 14(5):589–598.

[Wu et al., 2002] Wu, M., She, J.-H., Nakano, M., and Gui, W. (2002). Expert control and fault diagnosis of the leaching process in a zinc hydrometallurgy plant. *Control Engineering Practice*, 10(4):433–442.