Department of Chemical Engineering

Monitoring, Diagnostics and Improvement of Process Performance

Muhammad T Rafique

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

July 2009

Declaration

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

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

Signature:

Date:

Dedicated to my Father, my Wife and my Son for their love and affection

ABSTRACT

The data generated in a chemical industry is a reflection of the process. With the modern computer control systems and data logging facilities, there is an increasing ability to collect large amounts of data. As there are many underlying aspects of the process in that data, with its proper utilization, it is possible to obtain useful information for process monitoring and fault diagnosis in addition to many other decision making activities. The purpose of this research is to utilize the data driven multivariate techniques of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) for the estimation of process parameters. This research also includes analysis and comparison of these techniques for fault detection and diagnosis along with introduction, explanation and results from a new methodology developed in this research work namely Hybrid Independent Component Analysis (HICA).

The first part of this research is the utilization of models of PCA and ICA for estimation of process parameters. The individual techniques of PCA and ICA are applied separately to the original data set of a waste water treatment plant (WWTP) and the process parameters for the unknown conditions of the process are calculated. For each of the techniques (PCA and ICA), the validation of the calculated parameters is carried out by construction of Decision Trees on WWTP dataset using inductive data mining and See 5.0. Both individual techniques were able to estimate all parameters successfully. The minor limitation in the validation of all results may be due to the strict application of these techniques to Gaussian and non-Gaussian data sets respectively. Using statistical analysis it was shown that the data set used in this work exhibits Gaussian and non-Gaussian behaviour.

In the second part of this work multivariate techniques of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) have been used for fault detection and diagnosis of a process along with introduction of the new technique, Hybrid Independent Component Analysis (HICA). The techniques are applied to two case studies, the waste water treatment plant (WWTP) and an Air pollution data set. As reported in literature, PCA and ICA proved to be useful tools for process monitoring on both data set, but a comparison of PCA and ICA along with the newly developed technique (HICA) illustrated the superiority of HICA over PCA and ICA. It is evident from the fact that PCA detected 74% and 67% of the faults in the WWTP data and Air pollution data set respectively. ICA successfully detected 61.3% and 62% of the faults from these datasets. Finally HICA showed improved results by the detection of 90% and 81% of the faults in both case studies. This showed that the new developed algorithm is more effective than the individual techniques of PCA and ICA. For fault diagnosis using PCA, ICA and HICA, contribution plots are constructed leading to the identification of responsible variable/s for a particular fault. This part also includes the work done for the estimation of process parameters using HICA technique as was done with PCA and ICA in the first part of the research. As expected HICA technique was more successful in estimation of parameters than PCA and ICA in line with its working for process monitoring.

ACKNOWLEDGEMENT

First of all I would like to thank my God for giving me strengths to do this work and face the ups and downs during this work. Then it is my cherish privilege to acknowledge the moral support of my wife who was always there for me during all this work. I would also like to mention my son here whose smile always motivated to move forward in my life.

I extend my thanks to my Supervisors Professor Moses Tade and Dr. Nicoleta Maynard for their support during my stay in University and giving me opportunity to accomplish this work. Special thanks to Professor Ming Ang and Dr. Gordon Ingram for their feedback about my work.

I would also say special thanks to my colleague Imran who was always there for listening my ideas and to discuss about them not only for technical work but practical life works. Special thanks to Nadeem for good discussions, smiles and ideas.

I would also like to acknowledge support of my friends in the office and university. Foremost I would like to mention Chirayu, Hisham and Faye who were there to help me by any means and for providing good working environment in the office. I also acknowledge the help and company given by Pradeep, Kailash and Johan whenever I needed.

I would also like to acknowledge the help of all faculty members and administrative staff of the Department of Chemical Engineering for their active support.

Finally I am thankful to my Parents, Brothers and Sisters who motivated me to come to Australia and enhance further studies. Finally I place my regards to all those who helped me knowingly or unknowingly.

Table of Contents

Contents **Page #**

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

A	*Mixing Matrix*
C	Mean data
E	*Eigen Vectors*
EV	*Eigen Values*
H	*HICA values*
M	*Mean Centered data*
PC	*Principal Components*
S	*Independent Components*
W	Separating Matrix
X	Observed variables

ABBREVIATIONS

AEM	Abnormal Event Management
ANOVA	Univariate analysis of variance
ARL	Average Run Length
BPNN	Back-propagation neural network
BSS	Blind Source Separation
cMSPC	conventional Multivariate SPC
CSTR	Continuous Stirrer Tank Reactor
HICA	Hybrid Independent Component Analysis
ICA	Independent Component Analysis
GT	Generalized T distribution
IC	Independent Component
LCL	lower control limit
MANOVA	Multivariate Analysis of variance
MIMT	Modified Iterative Measurement Test
MSPC	Multivariate Statistical Process Control
MSPCA	multi-scale PCA
MPCA	Multi-way PCA
ODE	Ordinary Differential Equation
PC	Principal Component
PCA	Principal Component Analysis
PLS	Partial Least Square
UCL	Upper Control Limit
USPC	Univariate Statistical Process Control
WWTP	Waste Water Treatment Plant
pdf	probability density function

PSO	Particle Swarm Optimization
TE	Tennessee Eastman

Chapter # 1

Introduction and Overview

1.1 Background

In recent times industrial environment has become very competitive and for increasing efficiency of the production facilities it is very important to have better control and monitoring of the chemical processes. Waste minimization, plant safety, downtime reduction, consistent product quality, compliance with environmental regulations, better utilization of the plant in line with its designed capabilities (plant optimization) and development of new products are a few of the desirable reasons for better process monitoring. Process monitoring has become very critical as it is not specific to any particular industry rather it is applicable to different industries like waste water treatment plants, mining industry, oil and gas industry, packaging industry etc.

Process monitoring is more related to fault detection and diagnosis. Abnormal Event Management (AEM) which deals with detection of faults at the right time, leading to diagnosis and correction of these abnormal conditions in a process has fault detection and diagnosis as its central part. In petrochemical industries there is a loss of an estimated $20 billion every year so AEM is rated as their main problem to be solved. Considering these challenges on a large scale there is a lot of interest for researchers in both industry and academics to find out ways to overcome these problems. There is an abundance of literature on ways to investigate these issues including individual methods as well as hybrid methods (Venkatsubramanian et al. 2003).

Although there are different techniques which are under consideration for fault detection and diagnosis there is not a single technique which can answer all of the problems faced and come up as a comprehensive fault detection and diagnosis method. Fault detection and diagnosis methods include Data Driven, Analytical and Knowledge-based methods. Following is a table with strengths and weakness of each technique analysed individually (Chiang et al. 2001) , (Choi et al. 2005).

Table 1.1: Strengths and Weakness of Process Monitoring Techniques

Technique	Strengths	Weakness
Data Driven Technique	Derived from process data. Transforms from high dimension to low dimension.	Output mainly depends on quantity and quality of the data.
Analytical Approach	Use mathematical model derived from first principles.	Applicable to small systems. Expensive for large scale systems.
Knowledge-based Approach	Use Qualitative models for process monitoring.	Suitable for systems without detailed mathematical models. Applicable to small systems.

If we compare these techniques and consider the recent advancements in the introduction of computer based control systems which have led to the generation of more and more data set for any process, the data driven techniques will appear as one of the most used technique. Regarding data driven techniques, the behaviour of the data set has also become very important for researchers for fault detection and diagnosis perspective. Based on behaviour of the data, Principal Component Analysis (PCA) and Independent Component Analysis (ICA) have been used to find out faults in the process. In many cases the generated data is highly correlated which represents redundancy in the multivariate data and utilizing PCA we can explore this aspect of data. Along with this there are points where one variable has no correlation with another so in this part of the data set we utilize ICA to find out the underlying information in the data.

1.2 Scope, Goals and Approach

The scope of this work is focussed on Data Driven techniques of PCA and ICA for process monitoring. The overall goal of the work includes application of these techniques to waste water treatment data and air pollution data in order to find out the faults in the system. As both PCA and ICA have got their limitations in process monitoring, effort has been made in developing a new technique based on existing models of PCA and ICA for better fault detection and diagnosis. This new technique has also been validated on waste water treatment plant data set. Work is also done on estimation of process parameters by utilizing already existing techniques of PCA and ICA and the new technique to explore this aspect of these process monitoring algorithms.

1.3 Contributions and Significance

One of the contributions of this thesis is to explore the multivariate process monitoring techniques of PCA and ICA. The techniques of PCA and ICA have been utilized in a new dimension for process parameters estimation based on historical data set. Both techniques are utilized independently and the results are presented separately.

Another contribution of this thesis is to develop a process monitoring technique for better diagnosis of process faults based on historical data set. Both PCA and ICA have been used for monitoring of historical data set but they have proved their partial inability to find out the faults in the system. A new technique is proposed to identify these missed faults in order to enhance further fault detection and diagnosis techniques. Following are the steps for the representation of this contribution.

1. PCA is applied to data set of a Waste Water Treatment Plant (WWTP) to find faults.
2. ICA is applied to find faults using the same data set.
3. A hybrid of both PCA and ICA is developed and applied to the same data set to find out faults and reasons behind these faults.
4. Comparison of all the results obtained is done.

1.4 Thesis Organization

The remainder of the thesis is organized as follows.

Chapter 2 presents a detailed description of process monitoring techniques ranging from Univariate Process monitoring to Multivariate Process monitoring. Further to this detailed description regarding Multivariate techniques, their types and research work done in literature is discussed. Chapter 2 also involves description regarding the data set used for the case studies in this thesis and statistical analysis of this data set.

Chapter 3 presents the work done for estimation of Process Parameters using PCA and the validation of these calculated parameters.

Chapter 4 describes the dimension of ICA technique utilization for estimation of process parameters.

Chapter 5 discusses PCA, ICA and Hybrid of these techniques along with application of these techniques to WWTP and Air pollution data for validation of results and their comparison. It also includes utilization of Hybrid technique for the estimation of process parameters.

Finally in Chapter 6 the conclusion and recommendations are given for future work.

A flow diagram linking different chapters in this thesis is presented on the next page.

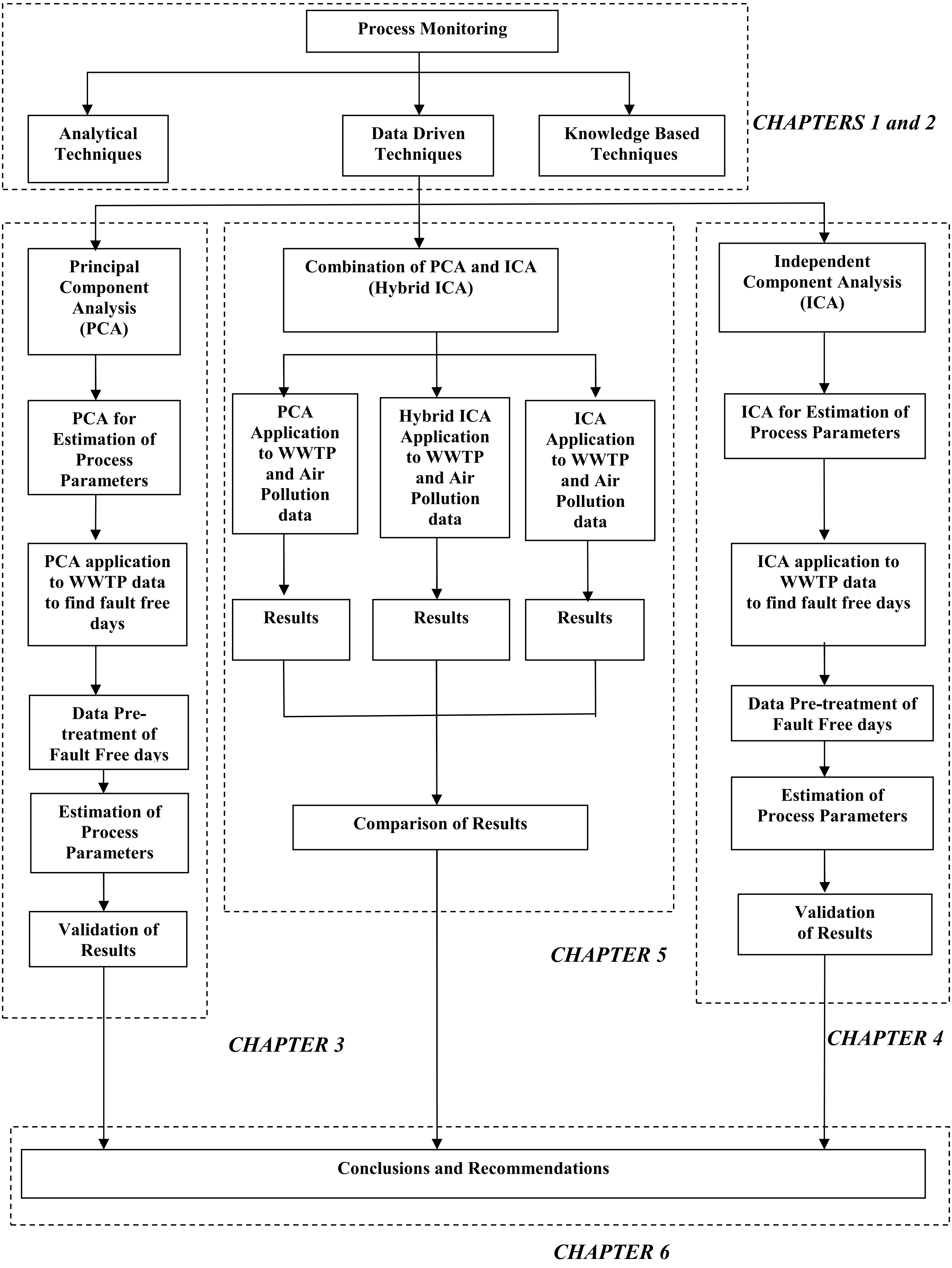

Process Monitoring
Analytical Techniques
Data Driven Techniques
Knowledge Based Techniques
CHAPTERS 1 and 2
Principal Component Analysis (PCA)
PCA for Estimation of Process Parameters
PCA application to WWTP data to find fault free days
Data Pre-treatment of Fault Free days
Estimation of Process Parameters
Validation of Results
CHAPTER 3
Combination of PCA and ICA (Hybrid ICA)
PCA Application to WWTP and Air Pollution data
Hybrid ICA Application to WWTP and Air Pollution data
ICA Application to WWTP and Air Pollution data
Results
Results
Results
Comparison of Results
CHAPTER 5
Independent Component Analysis (ICA)
ICA for Estimation of Process Parameters
ICA application to WWTP data to find fault free days
Data Pre-treatment of Fault Free days
Estimation of Process Parameters
Validation of Results
CHAPTER 4
Conclusions and Recommendations
CHAPTER 6

Figure 1.1: Flow Diagram of Thesis Chapters

Chapter # 2

Introduction of Process Monitoring

The meaning of process monitoring is to represent how the process is behaving under a given set of conditions in order to get desired outcomes. The idea behind process monitoring is to convert the online or offline process details so that we can get information from the process to provide meaningful measures and find the status of the process, leading to rectification of faults in the system. To get the required outputs, limits are applied to the system to keep it in the desired state. Faults diagnosis can also be carried out by developing and comparing measures that accurately represent the different faults in the process.

Process monitoring can be classified by three approaches namely Data Driven, Analytical and Knowledge-based (Chiang et al. 2001).

The Data Driven approach is, as expected, directly derived from the data of the plant. It is applied to both online data and off-line processes. Data driven techniques have become very important as we have large amounts of data being generated by plants due to modern computer aided manufacturing processes. The strength of the data driven technique is to transform high dimension data to low dimension or to give instant visual information about the behaviour of the process. The data driven technique is very effective in process monitoring but has the weakness that it depends on the amount and quality of the available data. Outliers in the data set may lead to wrong results.

In comparison to data-driven techniques, the Analytical Approach uses a mathematical model for the fault diagnosis but has the drawback that it cannot be applied to large scale systems because it requires a lot of models to represent the whole system. Its advantage includes giving better results than data-driven measures but is not normally used as it is quite expensive for large scale operations.

Knowledge-based approaches are suitable for systems for which we do not have detailed mathematical models, as they are mostly based on Causal Analysis, Expert Systems etc. Like the analytical approach, most applications of the knowledge-based approach are applicable to small systems having small number of inputs, outputs etc. (Chiang et al. 2001).

There are some other techniques as well as combinations of these three techniques which include Neural Networks, Expert Systems, Fuzzy Logic, Fuzzy Expert System, Fuzzy Neural

Network, Fuzzy signed Direction Graph and Fuzzy Logic etc. (Chiang et al. 2001), (Gertler 1998).

2.1 Research in Process Monitoring

There has been a lot of research in the field of Process Monitoring. Many scholars are working on different aspects of process monitoring techniques in order to utilize these established techniques more effectively. The comparison of these process monitoring techniques is also one of the areas of research. The fault diagnosis methods can be categorized into three types: Quantitative Model-based Methods, Qualitative Model based Methods and Process History based Methods. Venkatsubramanian et al. (2003) have written a series of papers on process monitoring and diagnosis. In Venkatsubramanian et al. (2003) the process diagnosis using Quantitative models (Called Analytical approach by Chiang et al. 2001) are reviewed. The authors have proposed that the Quantitative model based approaches can be evaluated by comparison of model characteristics. These include: how quickly system detects faults, how it detects the faults with minimum misclassification, its ability to handle noise, adaptability, uncertainties in the system, explanation facility modelling effort and computational requirements etc. Although analytical approaches give good results, there are limitations in the methods that they can only handle linear models. Another problem with model based methods is that if a fault is not specifically modelled then there is not much certainty that the fault can be detected. Therefore, the model based approaches have got their certain limitations to come up as a most comprehensive method for fault detection and diagnosis.

In the second part of the review paper by Venkatasubramanian et al. (2003a) they have discussed the Qualitative Models (Knowledge Based by Chiang et al. 2001). The fundamental understanding of physics and chemistry of the process plays an important role in the development of Qualitative models. The different forms of Qualitative models such as causal models and abstraction hierarchies are discussed by the authors. Although there are certain advantages of Qualitative methods, which are discussed in this paper, generating spurious solutions remains a problem. Different strategies are proposed in order to reduce the generation of these spurious solutions.

In the final review paper Venkatasubramanian et al. (2003b) have discussed fault diagnosis methods based on process historical knowledge. They have also reviewed the methodologies discussed in the earlier papers. By discussion of different data driven techniques and other process monitoring techniques it was revealed that none of the techniques were able to show a comprehensive diagnostic technique to outperform all of the existing techniques. It was also recommended that integration of these different techniques can be a way to develop a hybrid system that could overcome the individual constraints of the techniques. However, there are a lot of challenges in the design and implementation of hybrid systems such as incomplete information about the system, primary role of the operator, validity of assumptions about the process etc.

Venkatasubramanian (2005) have again presented the challenges and overview of these challenges related to fault diagnostic by considering relevance of automated processes hazard analysis to abnormal event management in product lifecycle management. The description about the prognostic and diagnostics systems is given along with desirable features of the diagnostics system. Also classification of diagnostics approaches is given and their comparison is carried out.

2.2 Data Driven Techniques

The data driven Process Monitoring techniques can be broadly classified as Univariate Process Monitoring and Multivariate Process Monitoring.

2.2.1 Univariate Process Monitoring Techniques

In the industry there are different process related parameters/variables which are under considerations. As some of the parameters have more importance than the others due to their impact on the process so mostly we are only concerned with these parameters. For the exploration of the behaviour of these parameters we do univariate analysis of the parameters or data set. Univariate Analysis is the type of parameter analysis where we explore each of the data independently. Here we have a bunch of values or data set and by the central tendency of the data set we can describe the response of the process.

The univariate monitoring technique is normally used on historical data sets to analyse the quality aspects of the products formed. It is important tool in industry because using the visual representation of the data we can find out more information in less time related to the data set. Some of the purposes achieved from univariate data analysis include further information exploration, easy comparison, data summarizing and generating thoughts related to the data set and consequently getting faults rectified. There are different types of univariate data analysis. Some of the important techniques are as follows: Bar Graph, Histogram, Pie Chart, Frequency Polygon and Shewart Chart etc. Their details can be accessed easily in literature (Newmark 1997), (Chiang et al. 2001), (Montgomery 1991), (Barnes 1998), (Mason et al. 1989) etc.

To sum up we can say that univariate control charts are very important way of analysis of both online and offline data. It is still being used abundantly with good results in industries having simple processes. The purpose of its use is not only to find out of control process parameters and bad quality of the product formed but is also used to find out the capabilities of the process. The univariate process control charts are essentially used with one variable into consideration. If there are different variables in the system then in this case we do not consider the effect of one variable on another whereas in practice there are many variables in the process which are related to each other. This is the area where univariate control charts are unable to explain different aspects of the process.

2.2.2 Multivariate Process Monitoring Techniques

Considering only one variable as the focus of our attention, we are able to analyse the data in univariate process monitoring. In most of the statistical process control approaches we are considering only small number of variables which normally include final product quality. Here (in univariate process monitoring) we are examining one aspect of the product quality at a time. In the industry we are having hundreds of different variables so it is not appropriate for most processes to apply univariate process monitoring. Consequently we are compelled to consider all the process parameters for complete process monitoring and to get good information inside the process. In order to analyse such large number of variables we use multivariate process monitoring techniques. It is very helpful to know that in practice only a few events/variables are driving a process at any one time: different combinations of these measurements are simply reflections of the same underlying events (Bersimis et al. 2007), (Bendwell 2002), (Yoon and MacGregor 2001).

With the advent of more and more complex processes in the industry there are more variables involved in the process. Also with the advent of computer aided control systems, the recording of variables after specific period has also become quite easy and usual practice. So it is necessary to analyse the process with so many variables into consideration. As compared to Univariate Process Monitoring a Multivariate Statistical Process Control (MSPC) is a process monitoring technique where we consider different variables and the affect of these different variables on each other and to the whole process (DeVor et al. 1992), (Flury 1997).

MSPC refers to a set of advanced techniques which are used for the monitoring and control of both continuous and batch processes. Through the application of statistical modelling MSPC techniques we can reduce the number of critical parameters from hundreds to two or three composite matrices. These composite matrices are in fact providing a framework for continuous improvements of the process operations as these matrices can be easily monitored in real time in order to benchmark process performance and highlight potential problems (Bersimis et al. 2007), (Aguado et al. 2007), (Kano et al. 2000)

Before we explain different types of multivariate monitoring techniques and their extensions, it would be a good idea to mention the goals of multivariate quality control. Any multivariate control should fulfil the following four conditions (Jackson 2003):

1. The technique should be able to a answer single question: “Is the process in control?”
2. An overall amount of the error should be specified .i.e. how much is the percentage of the error in the system.
3. The procedure should take into account the relationships among the variables.
4. Procedures should be available to answer the question: “If the process is out-of-control, what is the problem?”

The condition 4 is much more difficult to answer than the other three, due to an increased number of variables.

There are different techniques used in multivariate process analysis. These include Factor Analysis, Cluster Analysis, Multidimensional Scaling, T^2 Statistics, PCA and ICA. As the work done in this thesis is mainly on PCA and ICA so initially these two techniques will be explained in detail. After discussion about them there is very brief description of other techniques as well.

Principal Component Analysis

In the present times, operators find themselves compelled to be dealing with many variables as there are hundreds or thousands of variables entering into or out of the system at any single instant (Lattin et al. 2003). Plant economy, environment or complexities of the process are some of the reasons why plant operators have to deal with so many variables. As the data set has got so many dimensions, it is very difficult for the people at the plant to comprehend or even visualize the associated patterns with the data points. Also as there are so many recycling streams present in the system, it makes substantial redundancy among dimensions of the data set thus making the process further complicated. Reasons like these are leading to high levels of correlation and multicollinearity (Lattin et al. 2003).

PCA is a dimensionality reduction technique which makes the process monitoring greatly simple by utilizing the idea of projecting the data into lower dimensional space. Because of lower dimensional representation along with the preservation of the correlation structure between the process variables, it is optimal in terms of capturing the variability in the data (Chiang et al. 2001), (Dunteman 1989), (Brauner and Shacham 2000) .

By applying PCA we are retaining as much of the variations present in the data as possible. (AlGhazzawi and Lennox 2008). Multiple regression and discrimination analysis result in the loss of one or more important dimensions because they use variable selection procedures to reduce the dimensionality so they have this limitation in their application to the data set Draper and Smith (1981). In the case of PCA, the PCA approach uses all the original variables to obtain a smaller set of new variables and these variables are called Principal Components, often written as PCs. The approximation of the original variables can also be carried out using these new variables. The number of principal components required to explain the data depends on the degree of correlation between the data set - the greater the degree of correlation between the original variables, the smaller the number of new variables required. PCs are uncorrelated and are ordered so that the first few retain most of the variation present in the original set (Wang et al. 2004).

PCA is a technique that gives us a new data set with more meaningful information from data as compared to original data. Thus when we do PCA on the data the number of variables obtained would be exactly the same as the number of input variables but the data will be transformed in such a way that if we find the variance of the data then the initial components will account for most of the variability in the data which in most of the cases is more than 90% for the first few principal components but it strictly depends on the nature of the data. As the first few PCs account for most of the variance, we only analyse that part of the data and leave the remaining data.

For the analysis of the process we observe values of all the retained PCs. So those PCs which have values (numeric values) different from the other components are the focus of attention. We first observe the PC1 values and if any of it shows a value significantly different from the other PCs then we analyse it. As the first principal component is a linear combination of the original variables and indicates the greatest variation in the data so we investigate and find the factors behind these variations. Similarly just like PC1, PC2 is also a linear combination of the original variables. i.e. it tells us about the next dominant direction of variation. Again we investigate the variables and find out the variables responsible for this abnormal value (very high or low in comparison to other values of PCs) of PCs and consequently find the fault in the system. So Principal Component Analysis is mainly used to reduce the dimensionality of the data and also to identify the new meaningful underlying variables for each data set.

For the cases when most of the data variations cannot be captured in two or three dimensions, methods have been developed to automate the process monitoring procedures (Macgregor and Kourti 1995). The application of PCA in these methods is motivated by one or more of the following three factors. Firstly, PCA can produce lower dimensional representations of the data which better generalize the data independent of the training set than that using the entire dimensionality of the observation or data set. It therefore improves the proficiency of detecting and diagnosing faults. Secondly, the structure abstracted by PCA can be useful in identifying either the variables responsible for the fault and /or the variables most affected by the fault. Thirdly, PCA can separate the observation space in two subspaces. The first subspace is used for capturing the systematic trends in the process. The second subspace contains essentially the random noise which is normally ignored for the purpose of dimension reduction (Dunia et al. 1996).

Further discussion about algorithm, calculation and research discussion about PCA are as follows.

Algorithm and Calculation of PCA

PCA can be calculated by two methods.

1. Correlation Method
2. Covariance Method

The covariance method is the most used method to find out the PCs and hence this method is described below (Jolliffe 2002), (Smith 2002).

Determination of PCA by Covariance Method

In the calculation of PCA using covariance method the goal is to transform a data set “A” of dimension N to an alternative data set “B” of smaller dimension M so the purpose would be to calculate data set B. Following are the steps for the calculation of PCA.

1. **Getting and Organizing the data set**

Assume that we have got some data set with “N” dimensions (Observations) called “A” and we want to transform that data set in a way to reduce the data so that for each point in data set we end up with only “M” variables. Also M<N (As we are reducing the dimension). Also we assume that the data are arranged as a set of L data vectors so x_1,x_2,x_3,….,x_L with each x

representing a single observation of N variables. Now we have $x_1,x_2,\ldots..,x_L$ as column vectors each of which has N rows.

2. Calculation of Mean

Now the next step is the calculation of mean along with each dimension of data set A.

3. Deviation of Data from mean

Now for the PCA to work properly we have to subtract the mean from each of the data set. So mean subtracted is the average across each dimension. So for x_1, x_2 , x_3,-----,x_L if $\overline{x}$ is the mean then it is subtracted. This produces a data set whose mean is zero. So we get another matrix which we can call “C” which will have same dimensions as for dimensions of “A” but it will have zero mean. So we store mean-subtracted data in the form of matrix which has got same dimensions as matrix “A”. The matrix “C” can be obtained from each column of the data matrix “A” as under.

$$C = A- \overline{x}\, j \qquad (2.1)$$

Where $\overline{x}$ represents mean vector from each column and “j” is is 1xN row vector.

4. Calculation of Covariance Matrix

Now once we have got the mean centred data so the next step is to find out the covariance matrix of this new data set. The general formula for calculation of covariance matrix of two variables “x” and “y” is as under.

$$\mathrm{Cov}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1} \qquad (2.2)$$

5. Eigen vectors and Eigen values calculations of Covariance matrix

After having calculated the covariance matrix we calculate the Eigen values and Eigen vectors. For the calculations of Eigen Vectors and Eigen Values we normally require computer based algorithms and we can do it by using MATLAB. Also matrix “D” of Eigen values will be square matrix. Also the calculated Eigen values and Eigen vectors are ordered and paired so for mth Eigen Values we will have mth Eigen vector.

6. Choosing the Eigen vectors and Eigen values

Now as the purpose of PCA is to reduce the dimension of the data set, so at this stage we apply a methodology to achieve this purpose. So from the column of the Eigen vector matrix “E” and the Eigen value matrix “D” we arrange them in order of decreasing Eigen values. It is also important to maintain correct pairing between the columns in each matrix.

It is also important because Eigen values represent the distribution of the initial data’s energy among each of Eigen vectors. Also as the calculated Eigen vectors form a basis for the data. So if we have “f” as the cumulative energy content for all the Eigen vectors then sum of the energy content in all the data set from 1 to m will be represented as

$$f[m]=\sum_{q=1}^{m} D[P,q] \tag{2.3}$$

For P=q and m= 1, 2, ------, m

So we form a feature vector by taking Eigen vectors that we want to keep from our list of Eigen vectors and thus we form a matrix with these Eigen vectors in the column so

$$\text{Feature Vector}= (\text{Eig}_1\ \text{Eig}_2\ \text{Eig}_3 \text{ ------ } \text{Eig}_n) \tag{2.4}$$

We can also choose vector “f” as a measure for choosing correct number of Eigen vectors. As here we are interested to keep high value of “f” to know as much variance in data as possible so if we want to retain 80% of the variance then we can write

$$F\ [m=L] \geq 80\%$$

7. Determination of Principal components

Now as we choose the Eigen vectors that we want to keep in our data set and have formed a feature vector, we take the transpose of the vector and multiply it with the original data set transposed so

$$\text{Principal Components= Row Feature Vector}^{T} \text{ x Row Data Adjusted}^{T}$$

Where “Row Feature vector” represents matrix with Eigen vectors in the column transposed so the most significant Eigen vectors are now in rows and Row Data adjusted is the mean-adjusted data transposed .i.e. the data items are in each column, with each row holding a separate dimension.

PCA Calculation -An Example

In order to explain the Principal Component Analysis we apply the technique of PCA to a small data set. From this data set we calculate the Principal Components to explain the methodology. We take a data set represented as under in Table 2.1 with four different variables as A, B, C and D with ten observations of each variable.

Table 2.1: Original Values of variables for small data set

Observation Number	A	B	C	D
1	6.2	7.5	6.6	5.8
2	2.8	0.5	0.7	2.6
3	3.3	3.5	2.9	3.6
4	2.7	2.9	2.2	2.7
5	2.5	8.5	3	1.7
6	3.3	2.3	2.7	3.8
7	1.2	3.5	1.6	0.7
8	1.6	7.7	1.1	2.5
9	2.4	5.8	1.6	1.3
10	3.3	2.3	0.9	1.8

The first step is the subtraction of the mean of each column from the corresponding values of data set. So the mean of each column is subtracted and the resulting data in the form of mean adjusted data is written in Table 2.2 as below.

Table 2.2: Mean Centered values of variables for small data set

Observation Number	A	B	C	D
1	3.27	3.05	4.27	3.15
2	-0.13	-3.95	-1.63	-0.05
3	0.37	-0.95	0.57	0.95
4	-0.23	-1.55	-0.13	0.05
5	-0.43	4.05	0.67	-0.95
6	0.37	-2.15	0.37	1.15
7	-1.73	-0.95	-0.73	-1.95
8	-1.33	3.25	-1.23	-0.15
9	-0.53	1.35	-0.73	-1.35
10	0.37	-2.15	-1.43	-0.85

As we get the mean centered data then the next step is the measurement of covariance matrix. The general formula for calculation of covariance between two dimensions 'x' and 'y' is given by eq.2.2 and the values of the covariance are calculated by using the same formula.

After the calculation of Covariance matrix we calculate the Eigen values and Eigen vectors of the data set. The Eigen values and Eigen vectors can be calculated by utilization of any statistical software. In this example we calculate the Eigen values and Eigen vectors by using MATLAB (2009). The calculated Eigen values and the Eigen vectors are given as below.

$$\text{Eigen Values}=\begin{bmatrix} 8.9204 \\ 4.8676 \\ 0.3634 \\ 0.2276 \end{bmatrix}$$

Similarly the Eigen vectors are as below.

$$\text{Eigen Vectors}=\begin{bmatrix} -0.2289 & 0.4955 & -0.0427 & -0.8368 \\ -0.8251 & -0.5316 & 0.1646 & -0.0975 \\ -0.4559 & 0.4265 & -0.6643 & 0.4111 \\ -0.243 & 0.5385 & 0.7278 & 0.3482 \end{bmatrix}$$

After the Eigen values and Eigen vectors calculations if we want to reduce the dimensions of the data we only retain Eigen vectors with highest values. In our case as we are retaining all principal components so we retain all the Eigen vectors as well.

Now in order to get the new data (or Principal Components) we take the Eigen vectors in transposed form and multiply them to the mean adjusted data in transposed form. So the formula for it would be as follows.

$$\text{Principal Components} = \text{Eigen Vectors}^{T} \text{ x Mean Adjusted data}^{T}$$

So by doing this operation we get the following data which is transformation of original data and represented in Table 2.3.

Table 2.3: Principal Components values for small data set

Observation Number	A	B	C	D
1	-5.9768	3.516	-0.1817	-0.1819
2	4.0441	1.3135	0.402	-0.1935
3	0.2085	1.4431	0.1406	0.3481
4	1.3786	0.6816	-0.1225	0.3076
5	-3.3178	-2.5921	-0.4517	-0.0904
6	1.2412	2.1034	0.2216	0.4525
7	1.9864	-1.7135	-1.0168	0.5614
8	-1.78	-2.9921	1.2995	0.2383
9	-0.3318	-2.0186	-0.2528	-0.4582
10	2.5477	0.2588	-0.0382	-0.9837

From this new data set we can get more information regarding any abnormal values or faults etc. For better understanding normal practice is to draw a graph between the calculated values and the number of values (Between number of Principal Components and values of Principal Components) and the abnormal values are identified visually. This is shown in Figure 2.1.

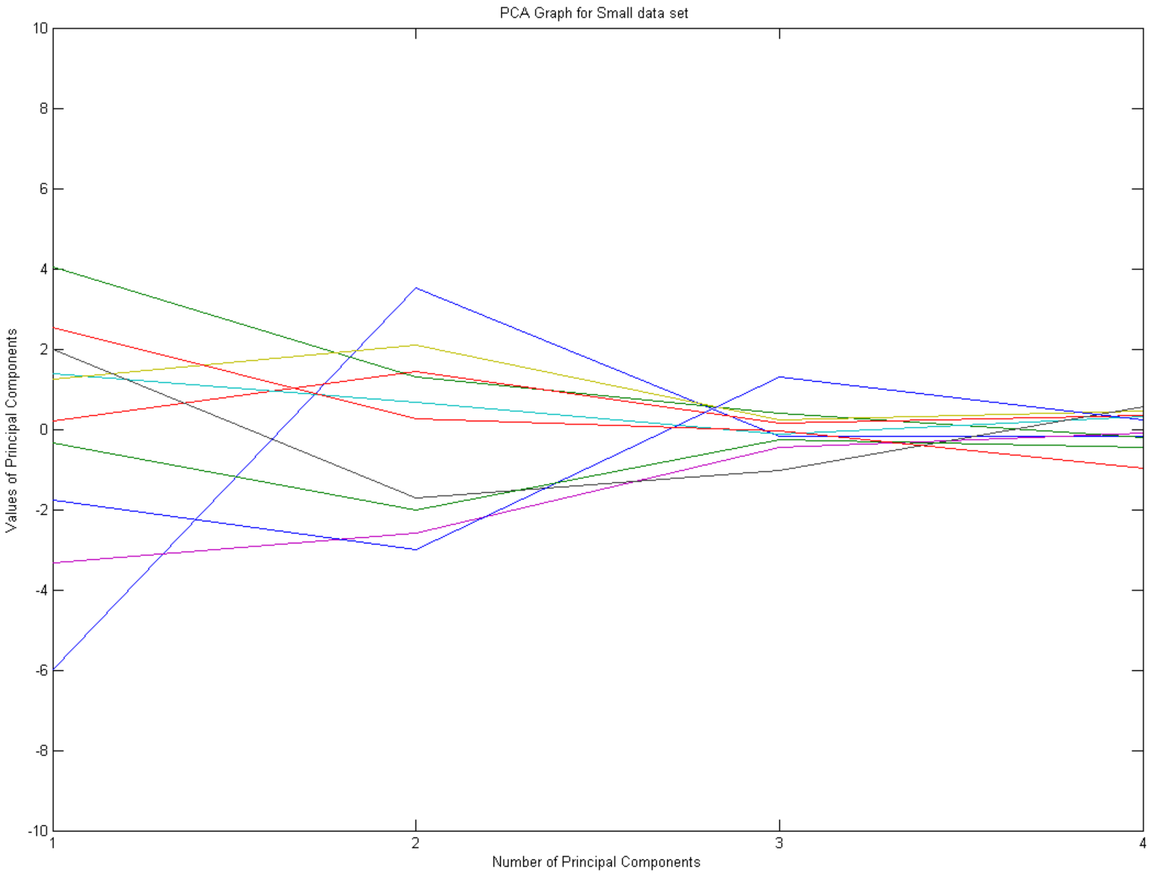


Figure 2.1: Graph between number of Principal Components and values of Principal Components for Small data set

As we know that PCA is visualization technique so in the graph those values of PCs which are quite different (higher or lower) than the others are the focus of attention. As we see in the graph that PC1 has got different values than the other Principal Components (For others graph is converging to show lower values). Here we see that PC1 has got the values of -5.9768 and 4.0441 which are different than rest of PCs. These can be investigated to find out the reasons behind these different values from the rest of the Principal Components. This aspect is more evaluated in our WWTP and Air pollution data sets in section 5.4.

Also it will be clearer if we compare it with our initial data set. The graph of all the four parameters is shown in Figure 2.2.

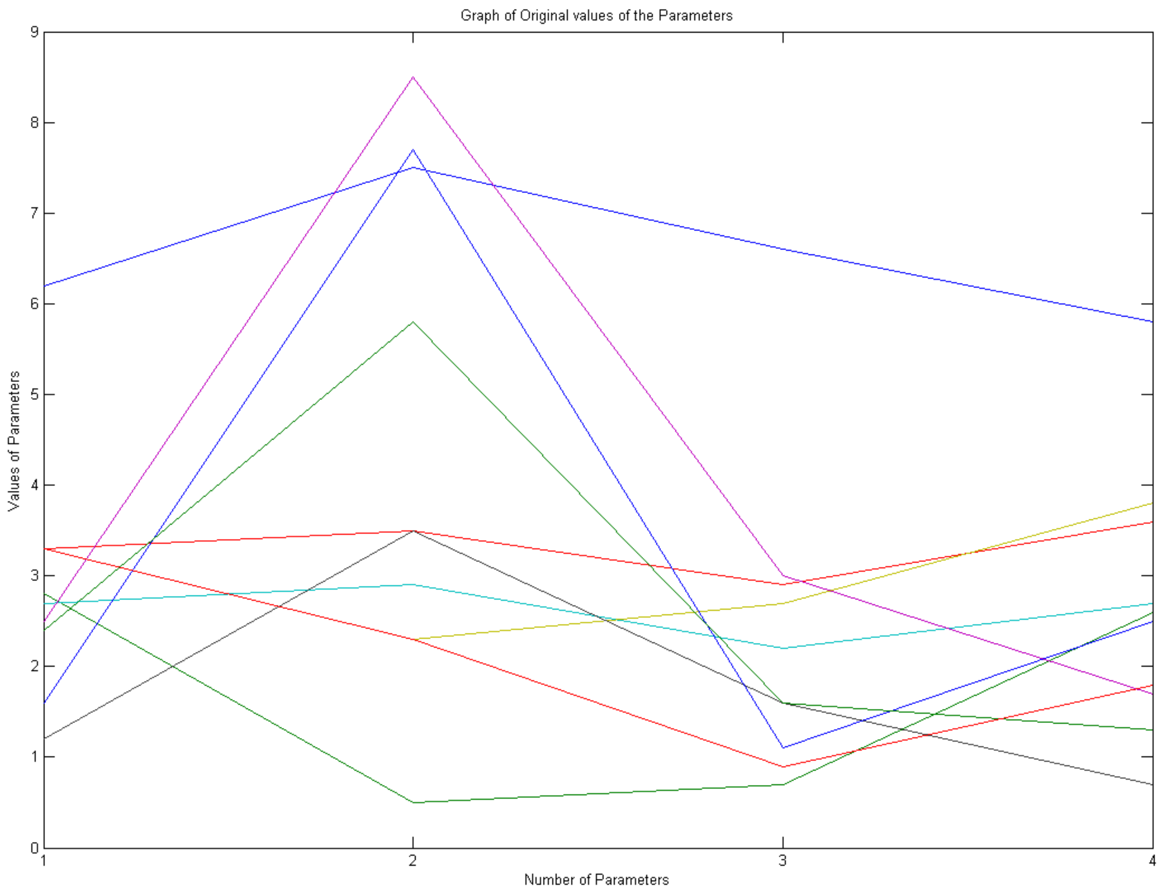


Figure 2.2: Graph between original values of variables and number of variables/parameters

From the graph we can see that there are fluctuations in the original data set but these fluctuations give us no information about the behaviour of data.

So based on all this work we find out the underlying aspects in the data set by visualization of data by projecting it in a new way.

Research and Extensions in PCA

There has been a lot of research regarding PCA as a process monitoring technique along with its applications in other area of interests. Many researchers have shown good results for process monitoring using PCA both for off-line process monitoring and on-line process monitoring. Albazzaz et al. (2005) and Wang et al. (2004) have done comparison of different multivariate techniques to find out faults in a waste water treatment plant data set. They have applied techniques of PCA, T^2, SPE and Conceptual Clustering to determine faults in the data set. By their work they have shown that PCA is more successful in detecting the faults from the historical data set. In this work they have not considered the other techniques like ICA to explore other aspects of the data set. Also their work is only limited to historical data analysis and does not give any information how this work can be used for online process monitoring.

He et al. (2009) used PCA for doing machine condition monitoring, by using time and frequency domains statistical features of the measured signals. Their work is quite diversified but by PCA application they have assumed the data to follow only Gaussian distribution. There is a need to further explore it because in those areas where we have non-Gaussian distribution it cannot give good results. Shinde and Khadse (2009) have used PCA for assessing multivariate process capability based on the empirical probability distribution of PC but again in the areas which do not strictly follow normal distribution we cannot get true process capability so there is need for better methodology which should address this part of the process as well.

Sun et al. (2005) have used PCA for application to detect boiler leakage. They have suggested a fault detection scheme designed for Hotelling's T^2 as well as the squared prediction error. A dynamic PCA model is also developed for boiler leak detection. Their proposed method has shown success to effectively reduce false alarm rate. Narasimhan and Shah (2008) have used PCA for model identification and error covariance matrix estimation from noisy data developing an iterative algorithm for model identification using PCA. They have applied the technique for the case when measurement errors in different variables are unequal and are correlated. Although the proposed technique gives good results, there are three conditions which have to be satisfied. These include that the underlying relationships relating variables have to be linear, the measurement errors from two samples have to be mutually independent and the measured samples should follow a specific equation of the relationship between the variables. As in real practice it is not possible to follow these conditions so these restrictions in the work have to be removed.

Wang et al. (2002) has shown another dimension of PCA application by using PCA to find out actual sensor locations in order to efficiently perform fault detection and diagnosis. This is done by using graph based techniques to optimize sensor location to ensure the identification of faults and consequently fault resolution. The methodology has shown good results for detection of weak process changes and insight to the root cause of the problem malfunction. The methodology is shown by the simulation results of a CSTR process. Although they have shown good results it has limitations that it does not work in areas where we have a multiple-fault situation. Li et al. (2004) have proposed an approach for fault detection and isolation based on abnormal sub-regions using PCA. The effectiveness of the proposed technique is shown by the results on PVC making process. Li et al. (2000) have presented recursive PCA for adaptive process monitoring. As recursive PCA takes into account the continuous changes taking place in the data, hence this approach is more important for online monitoring. They have proposed two recursive PCA algorithms for adaptive process monitoring. The two algorithms, based on rank-one modification and Lanczos tridiagonalization are proposed, and their compatibility is compared after defining an approach to update the correlation matrix recursively. The determination of PCs and the confidence limits for process monitoring is also done recursively. The algorithm is proved by applying it to a rapid thermal annealing process in a semiconductor processing system.

Ning He et al. (2004) have proposed a new method of combined Independent Component Analysis (ICA) and Multi-way PCA (MPCA). In their work they have not assumed that the latent variables are subject to Gaussian distribution. Also their work is based on ICA method that has independent variables as linear combination of MPCA latent variables. The combined ICA and MPCA method is capable of describing non-Gaussian distributed data. The algorithm is evaluated on Penicillin Fermentation benchmark process and is compared with traditional MPCA. Although they have shown good results but their work has the limitation in its application that it cannot be applied to data which does not follow a Gaussian distribution but they have effectively expanded the MPCA monitoring method by their work.

Luo et al. (1999) have proposed a technique for sensor fault detection via multi-scale analysis and dynamic PCA. In their work first wavelet decomposition of a dynamic sensor signal is done and then PCA is applied. Also sequential testing for real-time sensor fault detection is carried out using only the sensor signal itself. They demonstrated that the signals were not able to meet the requirement of PCA but the decomposed sensor signals were able to fulfil the PCA requirements. Also T^2 statistics was able to detect the simulated sensor failures i.e. changes in the sensor mode but were not detected by Q statistics. It was also mentioned that this T^2 statistics was able to do this for details only instead of the original signal. Although their suggested technique is robust for practical environment but it has the limitation that it is very expensive and cannot be applied due to cost or nature of the sensor fault. Lee et al. (2005) have proposed an adaptive Multi-scale Principal Component Analysis (MPCA) for online monitoring. Lee et al. (2005) decomposed the individual variables into wavelet coefficient at each scale. These wavelet coefficients were then used recursively to develop adaptive MPCA to extract correlations at each scale. Process monitoring and diagnosis is carried out by only retaining significant scales and combining them to construct a uni-scale batch data set in the time domain and developing a MPCA model. This can also be used for diagnosis of the fault to find the physical cause as it gives information on the time scale under which a fault affects a process.

Ganesan et al. (2004) have done a comparative literature review on wavelet-based multi-scale statistical process monitoring. Their work included multi-scale methods consisting of different statistical tools such as wavelet decomposition, de-noising, PCA charting and wavelet reconstruction. Their work also includes suggestions regarding research extensions such as average run length (ARL) performance study, Sensitivity Analysis, Monitoring of non-stationary processes and examination of the online performances. Although they have done a good comparison of the techniques, they have not included the recent developed technique of ICA. Wang and Romagnoli (2005) have proposed a robust multi-scale Principal Component Analysis for application to process monitoring. Their presented, robust multi-scale PCA modelling method is based on Generalized T distribution (GT) in score space using adaptive robust estimator. In their work they have shown that compared with conventional multi-scale PCA, the proposed approach incorporates a robustness feature which eliminates the effects of outliers in the training data. In their proposed work, advantages of both multi-scale and robust approaches are exploited so that more accurate models could be obtained for process monitoring purposes. Although the work is good however, this robust method has the drawback of increase computational cost.

Misra et al. (2002) have presented their work on multivariate process monitoring and fault diagnosis by multi-scale PCA (MSPCA). In their proposed method, cross-correlation across the sensors is extracted using a PCA approach, and autocorrelation within a sensor is determined using the wavelet approach. They have collected the contribution from each scale after the decomposition of individual signals using the wavelet approach. Finally PCA is applied to determine the faults in the system. For the validation of the technique data from two different industries is used. The technique has the drawback that it is strictly applicable to on-line process monitoring.

Jeng et al. (2006) have proposed dynamic process monitoring using predictive PCA. Jeng et al. (2006) have divided the data in different groups and one of them is evaluated with PCA in order to find faults. Based on that, other parts of the data are projected by same PCA model. They have built time series models to interpret the operating centers of the projected part and thus operating region is estimated for future monitoring. They have shown that based on this proposed monitoring scheme false alarms will be reduced. The effectiveness of this proposed method is demonstrated by simulation results. This work has the limitation of that the data in different group is assumed to have same behaviour whereas based on the nature of the process changes there can be different behaviours of the different data groups. Also the results are only demonstrated by simulation.

Ge and Song (2007) have proposed process monitoring based on ICA and PCA. They have used the fact that PCA or Partial Least Square (PLS) does not utilize the non-Gaussian information of the process data. So PCA is used to detect faults which follow Gaussian distribution and ICA is used to detect faults which follow non-Gaussian distribution. They have proposed a methodology to detect faults using this information. They have also proposed a mixed similarity factor which has the purpose to detect fault mode. A "main angle" is also proposed to calculate ICA-based similarity factor due to non-orthogonal nature of the extracted independent components. They have evaluated the proposed methodology in Tennessee Eastman (TE) process. Their results show superior power of fault detection and identification, compared with alternative PCA-based methods. These methods have the limitation in capturing the linear nature of the process and work has to be done to extend it to non-linear processes which can be done by the introduction of dynamics of the process under consideration.

Kim et al. (1997) have used nonlinear programming to improve the robustness and performance of Modified Iterative Measurement Test (MIMT) gross error detection algorithm. They have shown that the nonlinear programming technique can be used for both data reconciliation and estimation of gross error in MIMT. Their method has been evaluated on CSTR and has shown improved robustness compared to existing gross error detection algorithms so their enhanced algorithm have shown to be quite promising for data reconciliation and gross error detection of highly nonlinear processes.

Amand et al. (2001) have proposed the combination of data reconciliation and principal component analysis for increase efficiency in fault detection. They have used data reconciliation for the determination of projection matrix for principal component analysis. After that they have applied principal component analysis to raw data for monitoring purpose.

The combined use of these techniques leads to enhanced efficiency in fault detection. Their technique also has the advantage that it relies on lower number of components for process monitoring. It has been evaluated on modelled ammonia synthesis loop to show better process monitoring results.

Independent Component Analysis

In different fields the researchers are working in finding out the methods and techniques for the extraction or separation of useful information from some signals corrupted by noise and interferences. The identification of original signals or factors from a given set of data is the focus of blind source separation (BSS). The term is called 'blind' in this particular definition because both the original factors and process of mixing them are unknown. With the assumption that the available data is given by a linear combination of mutually independent factors, we can apply the ICA to solve this BSS problem (Scholz 2006). Therefore, the Independent Component Analysis (ICA) is a method for finding underlying factors or components from multivariate or multidimensional statistical data. What distinguishes ICA from other methods is that it looks for components that are both statistically independent and non-Gaussian (Hyvärinen et al. 2001), (Helsinki 2006), (Song 2007).

The ICA is very important in chemical plants as there are a large number of measured variables associated with any chemical process. These measured variables are driven by very small number of essential variables. Measurement of these variables is very important as monitoring them can greatly improve the process monitoring and consequently overall process. Independent Component Analysis is new emerging technique which is used to find several independent variables in the form of linear combination of measured variables (Kano et al. 2003). In Independent Component Analysis, the goal is finding the linear representation of non-Gaussian data so that the components are statistically independent or as independent as possible and lead to hidden aspects of the data set as this approach can capture the most essential structure of the data set (Hyvarinen and Oja 2000).

Mostly before the ICA algorithm is applied to the data set, the mixed signals are often pre-processed to remove the correlation between the observed variables which is called whitening. Several methods have been developed to achieve this purpose including PCA. As compared to PCA where the first few PCs explain the most of the data variance and PC1 is more important than PC2 and PC2 is more important than PC3, in ICA, IC1 is not more important than IC2, rather all have their own importance. In case if there is requirement of dimensionality reduction generally the number of PCs explaining the certain percentage of variance can be matched with the ICs and thus the number of ICs to be retained can be decided, which is heuristics and is mostly used.

There are many explanations of ICA given in the literature. Hyvarinen and Oja (2000) have explained it by a Cocktail-party problem which is given below.

For the explanation we assume that we are in a room where two people are speaking simultaneously. We have two microphones which are held at different locations. The microphones record different time signals and they are denoted by $x_1(t)$ and $x_2(t)$, with x_1 and x_2 being the amplitudes and the "t" time index. These recorded signals represent the weighted sum of the speech signals, which can be denoted by $s_1(t)$ and $s_2(t)$. We can represent this by a linear equation as follows.

$$x_1(t)=a_{11}s_1+a_{12}s_2 \tag{2.5}$$

$$x_2(t)=a_{21}s_1+a_{22}s_2 \tag{2.6}$$

Where a_{11}, a_{12}, a_{21} and a_{22} represent parameters and depend on the microphones distance from the speakers. Now the determination of original speech signals $s_1(t)$ and $s_2(t)$ would be very useful by only using the recorded signals $x_1(t)$ and $x_2(t)$. This is called the Cocktail-party problem. We can omit any extra factors (e.g. external noises etc.) from the missing model.

For further explanation we can consider the Figures 2.3 and 2.4.we can assume for the sake of this explanation that original speech signal could look something like those in Figure 2.3 and the mixed signals could look something like Figure 2.4. Now the cocktail-party problem is the estimation of the original signals using only the signals in Figure 2.4.

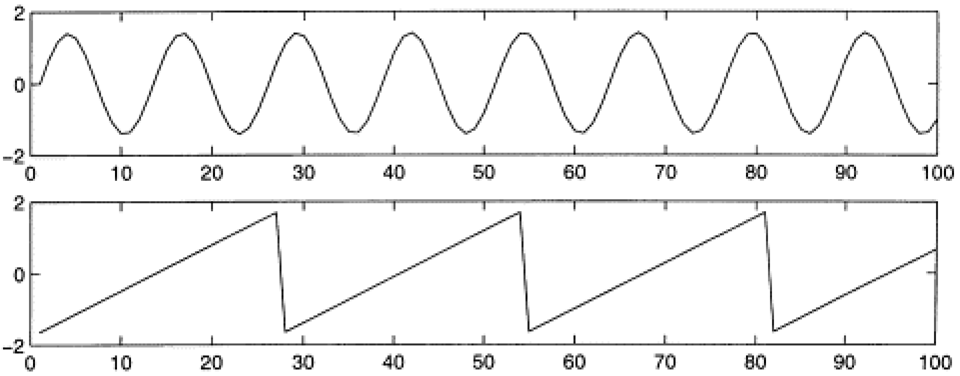


Figure 2.3: Original Speech Signals (Hyvarinen and Oja 2000)

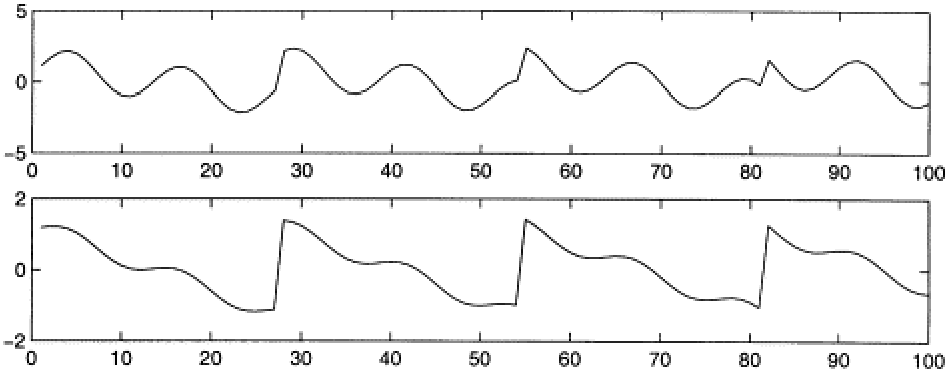


Figure 2.4: Observed mixed speech signals from the source (Hyvarinen and Oja 2000)

The problem is quite difficult as we do not know a_{11}, a_{12}, a_{21} and a_{22}. If we knew them then simply using the linear equation (2.5) or (2.6) by classical methods we could find out the original signals but as we do not have their values, we cannot solve the problem by classical methods.

For the solution of the problem we assume that both $s_1(t)$ and $s_2(t)$ are statistically independent. So by using the statistical properties of $s_i(t)$ we can solve this problem. This assumption does not need to be exactly true in practice so it turns out that this is not an unrealistic assumption as well. Therefore, the developed technique of ICA can be used to estimate the a_{ij}'s based on the information of their dependence which allows us to separate the two original source signals $s_1(t)$ and $s_2(t)$ from their mixtures $x_1(t)$ and x_2 t). Figure 2.5 gives us the representation of the two signals estimated by ICA method described in this example. It may be important to note that the calculated signals are having reversed signs (but it is of no significance as their value are important enough for the estimation of original signal).

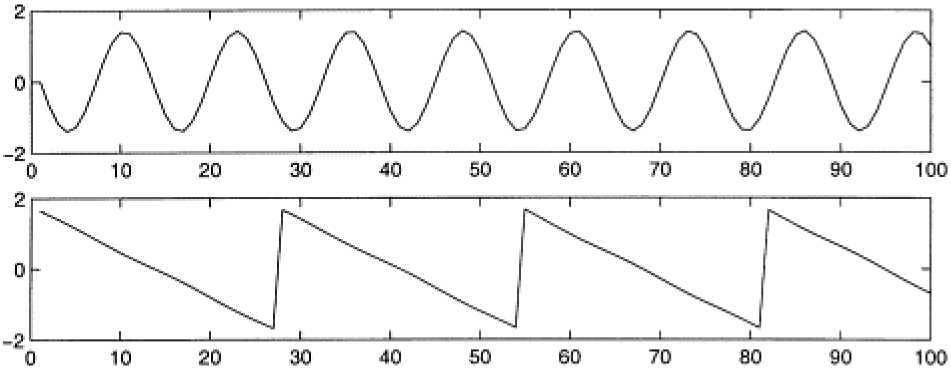


Figure 2.5: The estimated speech signals using Independent Component Analysis

(Hyvarinen and Oja 2000)

Although Independent Component Analysis was developed to solve problems like related to this cocktail-party problem, it has got many other applications, including process fault detection and diagnosis, finding hidden factors in financial data, reducing noise in natural images etc.

Algorithm of ICA

To define the algorithm of ICA if $y_1,y_2,\text{-----},y_m$ denote the random variables with joint probability density function (pdf) of $P(y_1,y_2,y_3,\text{-----},y_m)$ and we also assume that these variables have zero mean, then they will be mutually statistically independent if the following condition is valid.

$$P(y_1,y_2,y_3,\text{--------})= P_1(y_1)P_2(y_2),P_3(y_3)\text{-------}P_m(y_m) \qquad (2.7)$$

Where $P_i(y_i)$ (i=1,2,3,-----,m) denotes the marginal pdf of y_i .i.e. pdf of y_i when it is considered alone. In typical ICA algorithm, we assume that there are 'l' linear mixtures $x_1,x_2,\text{----},x_l$ of m independent (source) components $s_1,s_2,s_3,\text{----},s_m$ so we can write it as

$$X_i=m_{i1}s_1+m_{i2}s_2+\text{---------}+m_{im}s_m \qquad i=1,2,\ l\ (l\geq m) \qquad (2.8)$$

If we have 'n' as the number of sample, then $X=[X(1),X(2),\text{------},X(n)]\ \epsilon\ R^{lxn}$ can be represented as the matrix of observed variables, $S=[S(1),S(2),\text{-------},S(n)]\ R^{mxn}$ is the independent component matrix and A as matrix of elements A=[a_1,a_2,------,a_m] ϵR^{lxm} which is often called mixing matrix, so the above equation can be written as

$$X=AS \tag{2.9}$$

In literature A is also represented by M. So in ICA we estimate the mixing matrix A and/or the independent source vector S from the observed mixed matrix X. So we can also write the above equation as

$$S=WX \tag{2.10}$$

Where W is the separating matrix and X is the observed data matrix and S is the matrix of Independent Components.

ICA Calculations - An Example

In order to explain the calculations of Independent Component Analysis (ICA) we use the same data set as used for explanation of PCA in Table 3.1. For the ICA we use the algorithm developed by (HyvÃ¤rinen and Oja 1997) and using MATLAB we calculate the values of Independent Components of each signal.

As we know that the basic equation of ICA is given as.

$$X=AS$$

So we take the first signal of the data set. With the first signal we obtain following values of the Independent Components represented by “S”.

$$S=\begin{bmatrix} -4.8412 \\ -2.1864 \\ -2.5768 \\ -2.1083 \\ -1.9521 \\ -2.5768 \\ -0.937 \\ -1.2494 \\ -1.874 \\ -2.5768 \end{bmatrix}$$

Also the calculated value of mixing matrix “A” is as below.

A= -1.2807

So using the algorithm we calculate all the Independent Components of the data set and we get new data set shown in Table 2.4.

Table 2.4: Values of calculated Independent Components for small data set

Observation Number	A	B	C	D
1	-4.8412	2.8944	4.0589	4.1615
2	-2.1864	0.193	0.4305	1.8655
3	-2.5768	1.3507	1.7834	2.583
4	-2.1083	1.1192	1.353	1.9372
5	-1.9521	3.2803	1.8449	1.2197
6	-2.5768	0.8876	1.6604	2.7265
7	-0.937	1.3507	0.984	0.5022
8	-1.2494	2.9716	0.6765	1.7937
9	-1.874	2.2383	0.984	0.9327
10	-2.5768	0.8876	0.5535	1.2915

As ICA is also a data visualization technique so we can get more information about the data by construction of graph between the number of Independent Components and values of Independent Components. The constructed graph is shown in Figure 2.6.

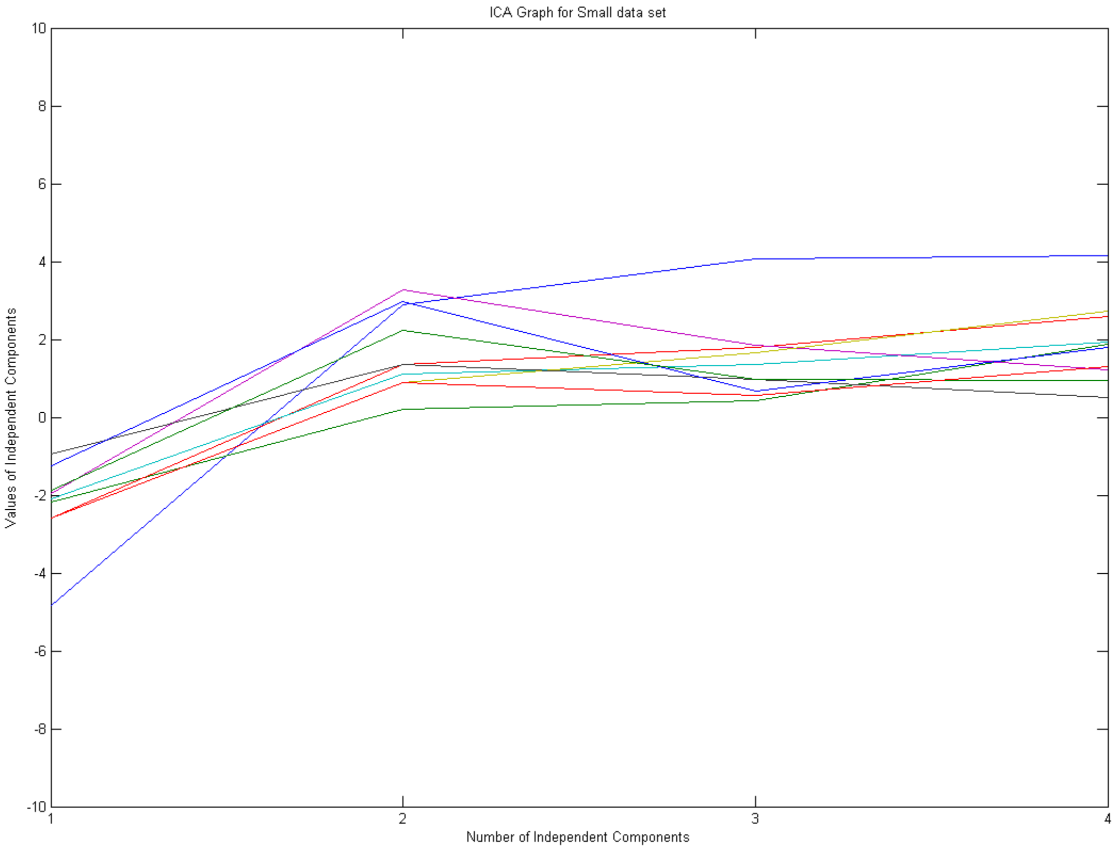


Figure 2.6: Graph between the number of Independent Components and values of Independent Components

Form the graph it is clear that IC1 and IC4 have got quite different values of -4.812 and 4.1615 respectively indicating that they are different than the rest of the data set. Another important point is that just like PC1 (Shown in section 3.3); IC1 also has value different than other components so they both complement each other by indicating problem at same point in the data set thus unveiling hidden information about the data set. The data points like these are investigated to find the factors behind these different values.

Applications of Independent Component Analysis

BSS and ICA have got a lot of attention in many fields such as geophysical data processing, data mining, chemical process data processing, image recognition, biomedical signal analysis and processing etc. (Cichocki and Amari 2002). Other applications include brain imaging and econometrics, where parallel time series is decomposed by ICA to get insight of the structure of the data set (Hyvärinen et al. 2001). In all these cases the main objective is to transform the observations in such a way that outputs correspond to the separate primary source signals form a number of observations of sensor signals from different independent sources (Cichocki and Amari 2002).

Research Work in Independent Component Analysis

There has been a lot of research related to ICA. As it is recently established technique, there is less published work on it as compared to PCA. Significant work include (Hyvarinen and Oja 2000) where they have given a lot of detail about the algorithms and basic uses of ICA. They have also presented FastICA algorithms and have given quite detailed applications of this ICA technique. HyvÃ¤rinen and Oja (1997) have shown the application of Fast Fixed-point algorithm to neural network. By using the algorithm all non-Gaussian Independent Components are found regardless of their probability distributions. The results are obtained by applying simulation to four source signals from four observed matrixes. Their work is dedicated to only this aspect of multivariate data analysis with no address to the data which follow the Gaussian distribution so their work lack an address to this aspect of the data set.

Kano et al. (2003) have applied ICA for monitoring purposes. The simulated results obtained on continuous stirrer tank reactor (CSTR) show its superiority over both Univariate Statistical Process Control (USPC) and conventional Multivariate SPC (cMSPC). Although they have shown good results, they have not made any comparison with other ICA algorithms. In their work they have only utilized Fast-Fixed point algorithm.

Lee et al. (2003) have used ICA for monitoring of a wastewater treatment process. In their work they have used the ICA algorithm with kernel density estimation to get better results than PCA. The work is validated by simulation data. Their work cannot account in the scenario when the data set follows Gaussian distribution.

Albazzaz et al. (2005, Albazzaz and Wang (2006) have applied ICA to reduce dimensions of the data. They have applied upper control limit (UCL) and lower control limit (LCL) for separating abnormal data. Before the calculation of upper limit (UL) and lower limit (LL) they have proposed Box-Cox transformation to transfer the Non-Gaussian co-ordinates to Gaussian distribution. Based on this work abnormal data points are determined. The Box-Cox transformation approach is recommended more than the percentile approach by them but in their work they have not addressed any other method for the transformation of non-Gaussian data points to Gaussian data. Albazzaz and Wang (2004) have also proposed a method for detecting faults using ICA. To overcome the non-Gaussianity of ICA they have made SPC charts with time varying upper and lower control SPC limits. The method has the limitation that it is more specific for batch runs so for continuous running of the plant it does not give good results.

Albazzaz and Wang (2007) have also carried out a study on Dynamic ICA, Static ICA, Dynamic PCA and Static PCA. They have introduced lag shifts to include process dynamics in the ICA model. The validation of the model is carried out on two batch processes. It is shown that dynamic ICA detects faults more clearly and precisely as compared to other techniques of Static ICA, Dynamic PCA and Static PCA.

Factor Analysis

Factor Analysis is a statistical approach that can be used to analyse interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions (factors). Factor Analysis is the way of estimating this interdependence as we determine the part of the data set where some of the variables are overlapping on each other. As we are combining two or more variables using one common factor, so this technique also has the purpose and advantage of dimension reduction of the data set under consideration (Hair et al. 1987).

Although both PCA and Factor Analysis serve the purpose of dimension reduction of the multivariate data, there is a distinction between them. In Factor Analysis we are concerned with identifying underlying sources of variance common to two or more variables (called common factors). An assumption explicit to this common factor model is that the observed variation in each variable is attributable to the underlying common factors and to a specific factor (often interpretable as measurement error). By contrast, there is no underlying measurement model with principal component; each principal component is an exact linear combination (i.e. weighted sum) of the original variables. Therefore, if the error in the factor analysis model are assumed to have the same variance then the Factor Analysis method becomes equivalent to the PCA (Lattin et al. 2003).

Multidimensional Scaling

PCA and Factor Analysis are scaling methods in a way that we use direct observations to create a map of new observations in reduced number of observations. In comparison to these two techniques, we use multidimensional scaling when our only information is an assessment of the relative proximity or similarity between the pairs of objects in the data set. A map of appropriate dimensionality is created by using this information about relative proximity such that the distance in the map correspond to the proximities used to create it (Lattin et al. 2003).

The goal of visualizing data is to find out the distance between the points in the data. It can be done with scatter plot but here we have data in the form of pairs. It is done to find out the data which is different from the rest of the data set and hence can lead to determination of abnormal values.

T^2 Statistics

Hotelling's T^2 is a statistical measure to find out the distance for each observation from the centre of data. It is used to determine abnormal points in the data set.

Wang et al. (2004) did a comparison of different multivariate techniques and T^2 was successful in detecting most of the faults but was not able to detect all the faults. Albazzaz et al. (2005) also showed the same results for their comparative study for PCA, T^2, SPE and Conceptual Clustering. For both cases PCA showed better results among all the techniques.

Multiple Regression

Multiple Regression is useful method of analysis when the research problems involve a single dependent variable presumed to be related to two or more independent variables. The objective of multiple regression analysis is to predict the changes in the dependent variable in response to changes in the independent variables. This objective is mostly achieved by statistical rule or least squares (Hair et al. 1990).

Canonical Correlation

Canonical correlation is used for analysing the relationships between two sets of multiple variables. It is different to PCA because in canonical correlation we find two sets of multiple variables (i.e. one linear combination of variables in one set and second linear combination of the variables in the other set) that exhibit highest possible correlation (i.e. covariance). We can use this technique for dimension reduction of the large data sets (Lattin et al. 2003).

With canonical analysis the objective is to correlate simultaneously several dependent variables and several independent variables. So it can be regarded as a logical extension to multiple regression analysis. As multiple regression involves a single dependent variable, canonical correlation involves multiple dependent variables. The objective is to develop a linear combination of each set of variables both dependent and independent) to maximize the correlation between the two sets (Hair et al. 1990).

Multivariate Analysis of Variance and Covariance

Multivariate Analysis of Variance (MANOVA) is a statistical technique that can be used to simultaneously explore the relationship between several categorical independent variables and two or more dependent variables. As such, it represents an extension of univariate analysis of variance (ANOVA), where we find whether the mean of a variable differs significantly between groups. So in this we determine whether the entire set of means is different from one group to the next (Hair et al. 1990).

Cluster Analysis

Cluster Analysis is a method of creating groups of objects or clusters, in such a way that objects in one cluster are very similar and objects in other cluster are quite distinct (Gan et al. 2007). Cluster Analysis objective is to make groups which are mutually exclusive based on the similarities among the data (Hair et al. 1990). Cluster Analysis is also called Segmentation Analysis, Taxonomy Analysis, Unsupervised Classification (Gan et al. 2007).

Cluster Analysis usually involves three steps. Measurement of form of similarity or association among the entities to determine number of groups in the sample is the first step. The second step involves making groups or clusters of the sample. Determination of the composition of variables by making profile of the variables is the final step. Many times this may be accomplished by applying discriminant analysis to the groups identified by the cluster technique (Hair et al. 1990).

There are different algorithms used for Cluster Analysis. Among them the significant ones involve Hierarchical Clustering, Fuzzy Clustering, Centre-based Clustering, Search-based Clustering, Graph-based clustering, Grid-based Clustering, Density-based Clustering, Model based-clustering, Subspace Clustering and Kmeans clustering (Gan et al. 2007).

In K-means clustering the classification or grouping of attributes is done into K number of groups. The term K is a positive integer. The minimization of sum of squares of distance between data and the corresponding cluster centroid is used to from the groups. Thus the main purpose of the K-mean clustering is to classify the data (Gan et al. 2007).

2.3 Concluding Remarks

PCA and ICA have been used extensively in chemical and process industries in different dimensions (fault detection, signal processing etc.) and have proved to be fruitful tools for process monitoring. The main limitations in these techniques are their strict application on Gaussian and non-Gaussian data independently. We have reviewed work of different researchers in different dimensions of PCA and ICA. Most of them have given good results with the assumptions of Gaussian and non-Gaussian behaviour of the data. So there is need to explore the application of these techniques in other dimensions like parameters estimation etc. along with development of a technique which considers both Gaussian and non-Gaussian aspects of the data set. This research work is carried on the basis of this understanding that if these techniques (PCA and ICA) are merged together; the possibility is that it will produce better results than the individual technique due to the fact that industrial data possess both Gaussian and non-Gaussian characteristics. Along with exploration of PCA and ICA for fault detection there is also the probability of utilization of these models for the estimation of process parameters. If they are utilized in this way then it would be another contribution of these techniques.

2.4 Introduction to the Data Set Used

To represent a case study and explanation of the techniques used in this work two data sets have been used. The details of the data sets are as follows.

2.4.1 Waste Water Treatment Plant data set

The data base consists of 527 data cases representing 527 days of operational data for a WWTP. Each data case is represented by 38 variables (Table 1.1). The data was collected by Poch and made publicly available by Bejar and Corts of the University of Catalonia, Spain. The data can be found at the Machine Learning Databases Repository on the internet (Wastewaterdatabase 2008). The plant is an activated sludge process located in Manresa, a town sited near Barcelona (Catalonia) population of 100,000 inhabitants. The plant treats a daily flow of 35,000 m^3 mainly domestic wastewater although wastewater from industries located inside the town is received by the plant too. It consists of three stages: pre-treatment, primary treatment by clarification and secondary treatment by means of activated sludge. The flow diagram of the plant operation is represented by Figure 2.7.

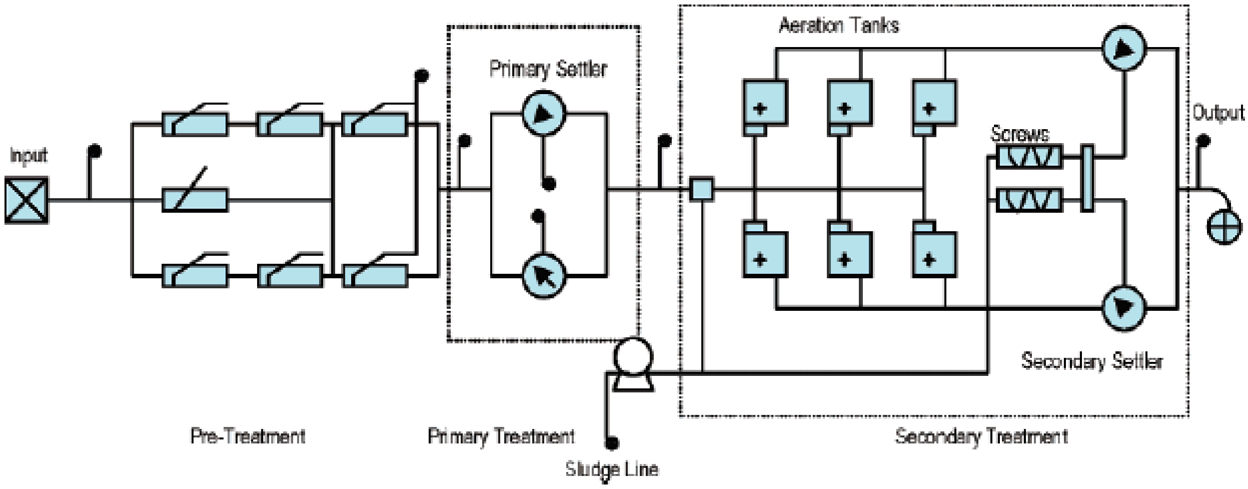


Figure 2.7: Wastewater Treatment Plant's structure Albazzaz et al. (2005)

The database has been used for studies in classification by Sanchez et al. (1997), where two methods, the Kmeans clustering method and Linneo+ metholodgy — a knowledge acquisition tool with an unsupervised learning strategy were investigated. Also Albazzaz et al. (2005) have used the data for their work in the comparative study with multivariate statistical process control. Wang et al. (2004) used the same data for multidimensional visualization of PCs for Process historical data. Albazzaz and Wang (2006) used the same data for historical data analysis based on plots of independent and parallel co-ordinates and statistical control limits. The original data had some missing values which were filled by (Albazzaz et al. 2005) using SPSS software and in this work the same values have been used. Albazzaz et al. (2005) in multidimensional visualisation of process historical data analysis has discussed about these missing values and further information about missing values can be obtained from this work. Also they have identified the faulty days in their work and these faulty days are the basis of our work as well.

Table 2.5: Description of different variables for the Waste Water Plant

No.	Parameter Abbreviation	Parameter Description
1	Q-E	Input flow to plant
2	ZN-E	Input Zinc to plant
3	PH-E	Input pH to plant
4	DBO-E	Input Biological demand of oxygen to plant
5	DQO-E	Input chemical demand of oxygen to plant
6	SS-E	Input suspended solids to plant
7	SSV-E	Input volatile suspended solids to plant
8	SED-E	Input sediments to plant
9	COND-E	Input conductivity to plant
10	PH-P	Input pH to primary settler
11	DBO-P	Input Biological demand of oxygen to primary settler
12	SS-P	Input suspended solids to primary settler
13	SSV-P	Input volatile suspended solids to primary settler
14	SED-P	Input sediments to primary settler
15	COND-P	Input conductivity to primary settler
16	PH-D	Input pH to secondary settler
17	DBO-D	Input Biological demand of oxygen to secondary settler
18	DQO-D	Input chemical demand of oxygen to secondary settler
19	SS-D	Input suspended solids to secondary settler
20	SSV-D	Input volatile suspended solids to secondary settler
21	SED-D	Input sediments to secondary settler
22	COND-D	Input conductivity to secondary settler
23	PH-S	Output pH
24	DBO-S	Output Biological demand of oxygen
25	DQO-S	Output chemical demand of oxygen
26	SS-S	Output suspended solids
27	SSV-S	Output volatile suspended solids
28	SED-S	Output sediments
29	COND-S	Output conductivity
30	RD-DBO-P	Performance input Biological demand of oxygen in primary settler
31	RD-SS-P	Performance input suspended solids to primary settler
32	RD-SED-P	Performance input sediments to primary settler
33	RD-DBO-S	Performance input Biological demand of oxygen to secondary settler
34	RD-DQO-S	Performance input chemical demand of oxygen to secondary settler
35	RD-DBO-G	Global performance input Biological demand of oxygen
36	RD-DQO-G	Global performance input chemical demand of oxygen
37	RD-SS-G	Global performance input suspended solids
38	RD-SED-G	Global performance input sediments

The different statistics relating to the data set are listed in Table 2.6 as below.

Table 2.6: Statistics of different Variables for Waster Water Plant

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

Statistical Analysis of the Data Set

In this work we are considering the multivariate data analysis techniques of PCA and ICA. As these are data driven techniques so proper investigation of the data set will be very helpful in the interpretation of the results. The WWTP data set is investigated to find out whether it follows Gaussian distribution or otherwise.

Histogram and Normal probability plots NIST/SEMATECH (2009) are the two most common types of methods used for the determination of the nature of the data set.

A Histogram is constructed by determining the population of the variables in regular spaced cells and plotting their frequency versus the center of the cell. A probability graph is more sensitive than histogram as it explains more about the data in the graph. For the construction of the graph the variables are sorted into ascending order and the cumulative probability of the variables is calculated. Then graph is plotted between these probability values and the original values of the variables. If the data is Gaussian, it will result in an approximate straight line. A clearly "S" shaped curve on the graph would indicate that the data is not Gaussian although there can be a small departure from the straight line. The presence of small breaks near the middle of the graph also shows abnormalities in the distribution of the data.

For the WWTP data set all the 38 variables have been investigated by construction of probability graphs. Minitab (ver 15) has been used for the construction of the Probability graphs of all the parameters. Form the graphs it was proved that the data set is a mixture of both Gaussian and non-Gaussian variables. Figures 2.8 and 2.9 show the probability graphs of DBO-D and DQO-D. From these graphs it is quite obvious that we get a straight line between the parameter values and the commulative probability which proves that these are normally distributed variables.

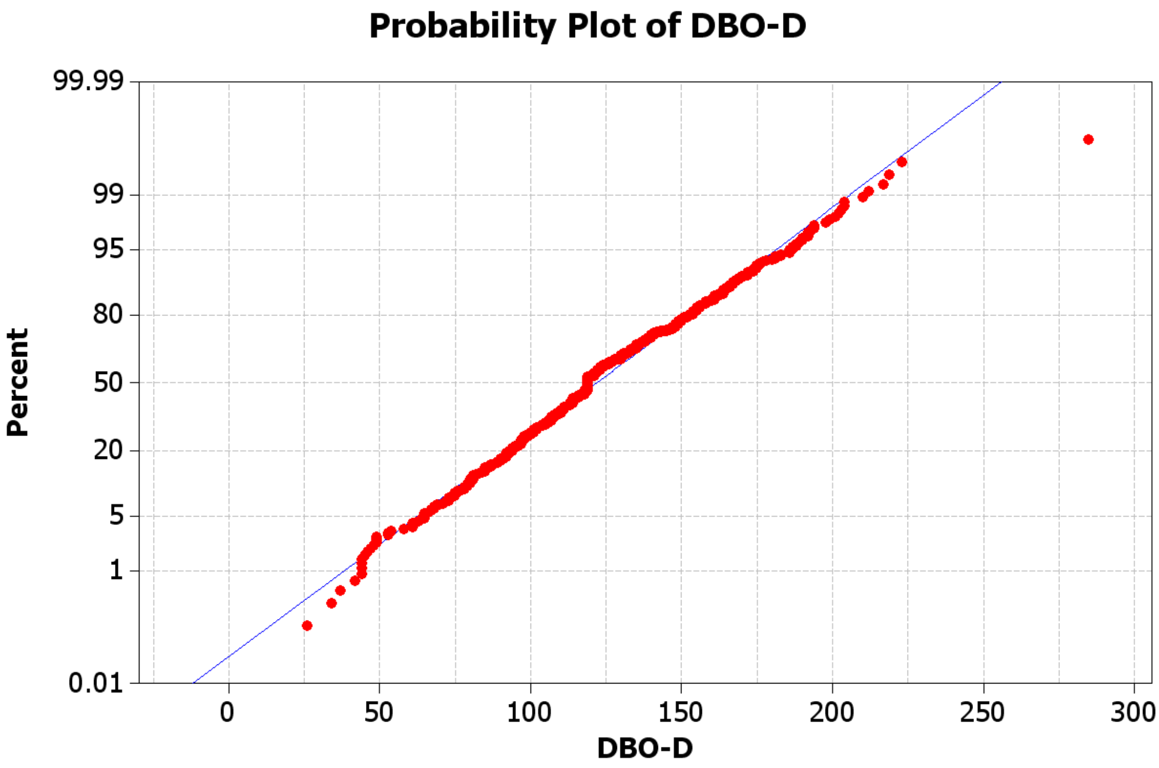


Figure 2.8: Probability plot of DBO-D showing Gaussian distribution

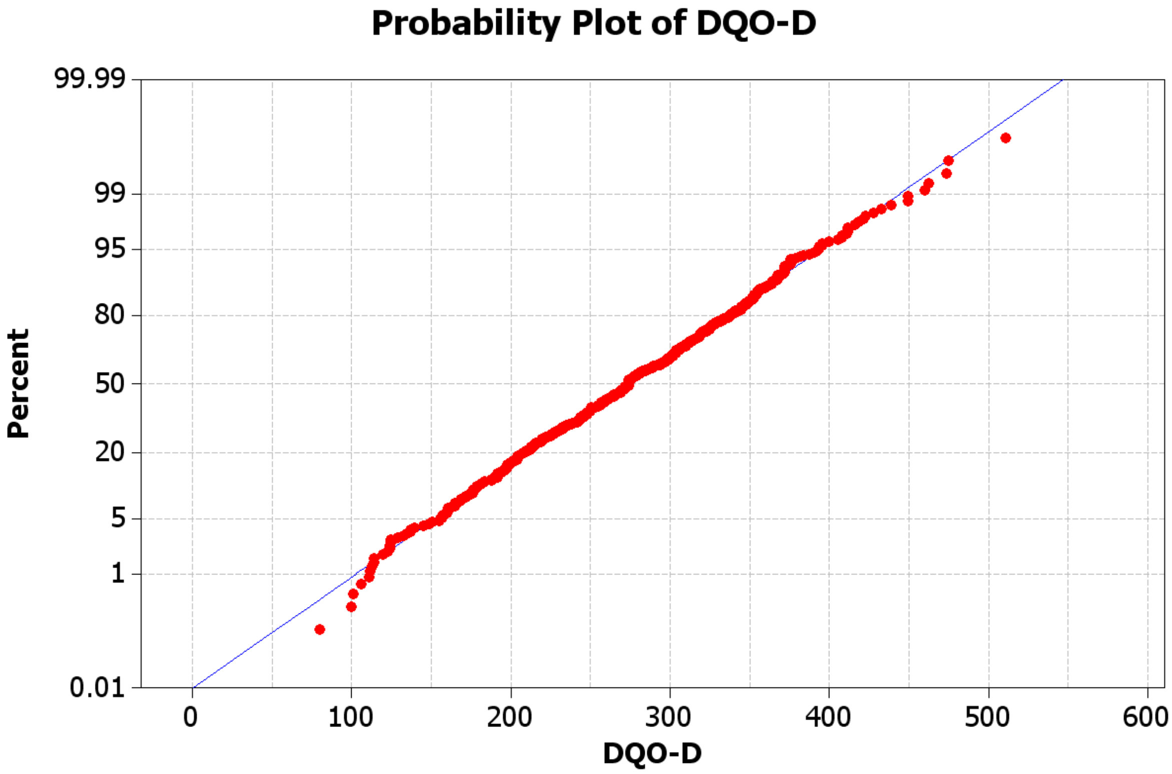


Figure 2.9: Probability plot of DQO-D showing non-Gaussian distribution

Also in Figures 2.10 and 2.11 we have the probability graphs for RD-SED-P and SED-S which do not follow straight line and indicate that the data does not follow Gaussian distribution.

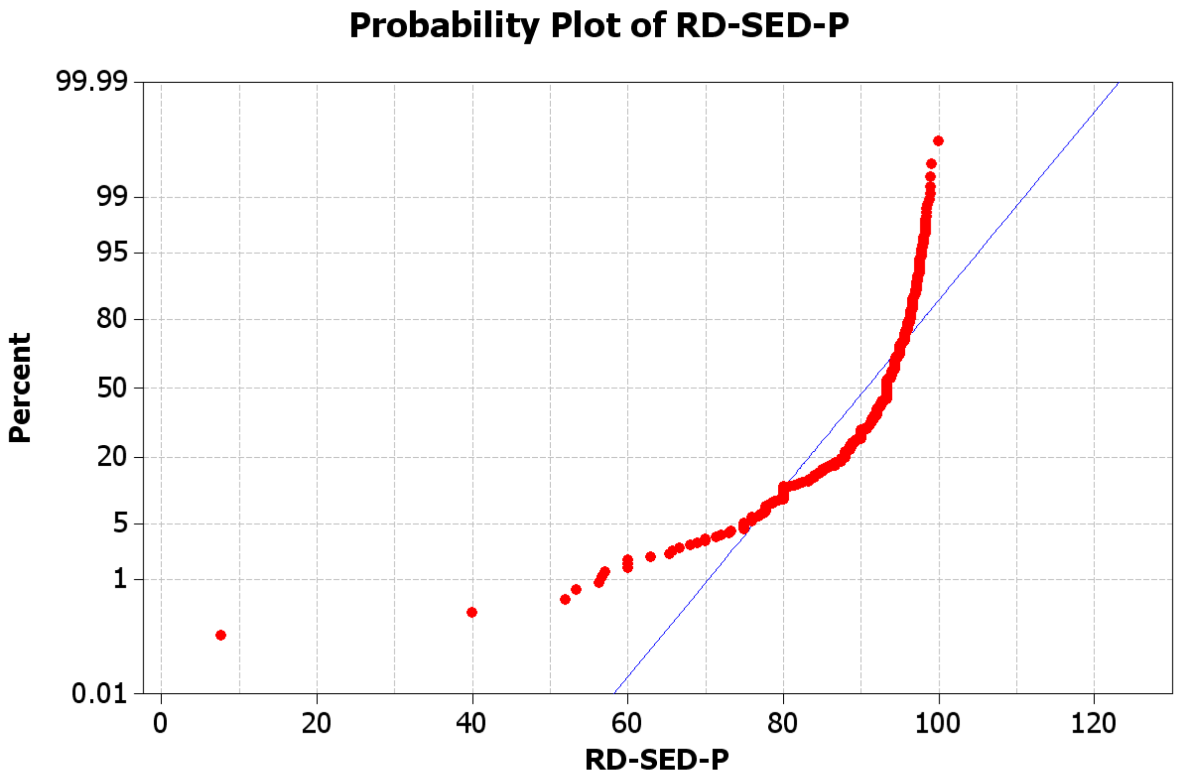


Figure 2.10: Probability plot of RD-SED-P showing non-Gaussian distribution

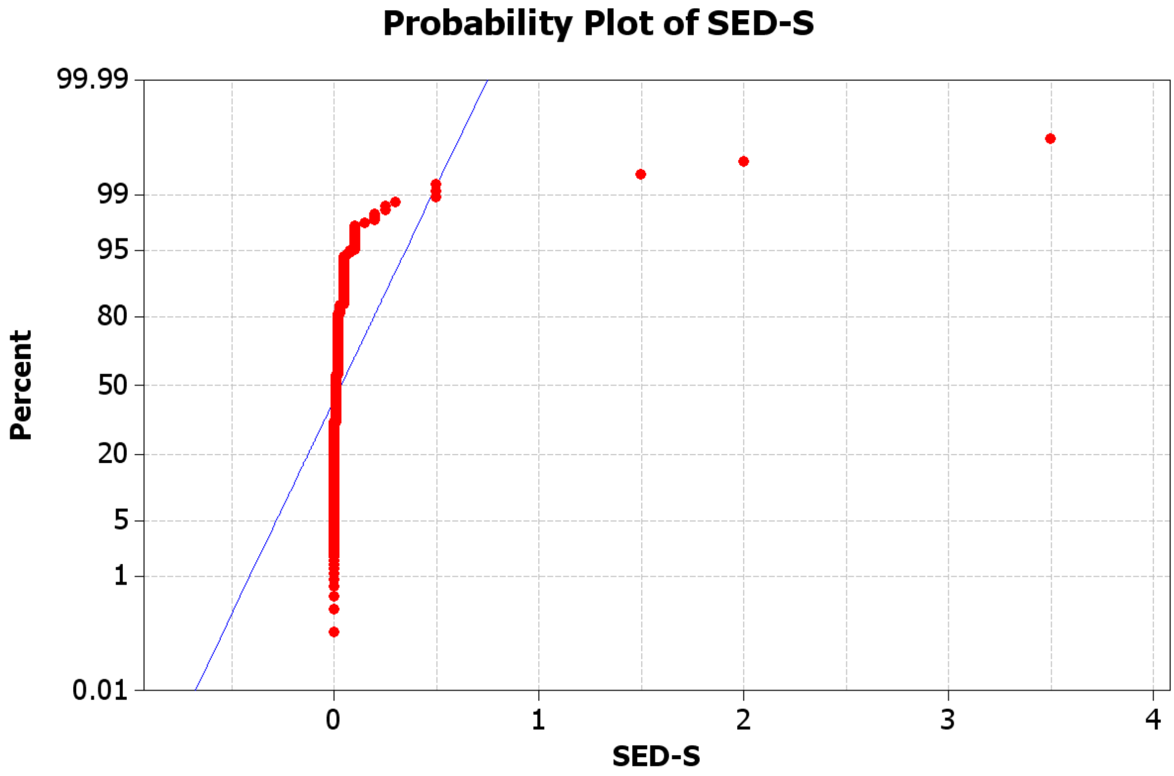


Figure 2.11: Probability plot of SED-S showing non-Gaussian distribution

Probability graphs of all the parameters are included in Appendix "A" and the Gaussian or non-Gaussian behaviour of the data is summarized in Table 2.7.

Table 2.7: Statistical distribution of different variables for the Waste Water Plant

No.of Attribute	Attribute	Distribution
1	Q-E	Gaussian
2	ZN-E	Non- Gaussian
3	PH-E	Non- Gaussian
4	DBO-E	Gaussian
5	DQO-E	Gaussian
6	SS-E	Non-Gaussian
7	SSV-E	Gaussian
8	SED-E	Non-Gaussian
9	COND-E	Gaussian
10	PH-P	Non-Gaussian
11	DBO-P	Gaussian
12	SS-P	Non-Gaussian
13	SSV-P	Gaussian
14	SED-P	Gaussian
15	COND-P	Gaussian
16	PH-D	Non-Gaussian
17	DBO-D	Gaussian
18	DQO-D	Gaussian
19	SS-D	Gaussian
20	SSV-D	Gaussian
21	SED-D	Non-Gaussian
22	COND-D	Gaussian
23	PH-S	Non-Gaussian
24	DBO-S	Non-Gaussian
25	DQO-S	Gaussian
26	SS-S	Non-Gaussian
27	SSV-S	Gaussian
28	SED-S	Non-Gaussian
29	COND-S	Gaussian
30	RD-DBO-P	Gaussian
31	RD-SS-P	Gaussian
32	RD-SED-P	Non-Gaussian
33	RD-DBO-S	Non-Gaussian
34	RD-DQO-S	Non-Gaussian
35	RD-DBO-G	Non-Gaussian
36	RD-DQO-G	Non-Gaussian
37	RD-SS-G	Non-Gaussian
38	RD-SED-G	Non-Gaussian

2.4.2 Air Pollution Data Set

To compare the results obtained on WWTP data, a new data set is utilized as a second case study. This case study is about an air pollution data set with 60 observations and 16 variables. McDonald and Schwing (1973) have used this data set for their work on instabilities of regression estimates relating air pollution to mortality. The data set can be found at (Machine.learning.database 2008). Further description about the variables involved is as follows in Table 2.8.

Table 2.8: Description of different variables of Air pollution data

No. of Observation	Abbreviation	Parameter Description
1	PREC	Average annual precipitation in inches
2	JANT	Average January temperature in degrees F
3	JULT	Average July temperature in degrees F
4	OVR65	% of 1960 SMSA population aged 65 or older
5	POPN	Average household size
6	EDUC	Median school years completed by those over 22
7	HOUS	% of housing units which are sound & with all facilities
8	DENS	Population per sq. mile in urbanized areas, 1960
9	NONW	% non-white population in urbanized areas, 1960
10	WWDRK	% employed in white collar occupations
11	POOR	% of families with income < $3000
12	HC	Relative hydrocarbon pollution potential
13	NOX	Relative nitric oxides pollution potential
14	SO@	Relative sulphur dioxide pollution potential
15	HUMID	Annual average % relative humidity at 1pm
16	MORT	Total age-adjusted mortality rate per 100,000

The different statistics related to the data set are given in Table 2.9.

Table 2.9: Statistics of different Variables for Air Pollution data

No.of Attribute	Attribute	Minimum value	Maximum value	Mean	Standard Deviation
1	PREC	10	60	37.36667	9.984678
2	JANT	12	67	33.98333	10.1689
3	JULT	63	85	74.58333	4.763177
4	OVR65	5.6	11.8	8.798333	1.464552
5	POPN	2.92	3.53	3.263167	0.135252
6	EDUC	9	12.3	10.97333	0.845299
7	HOUS	66.8	90.7	80.91333	5.141373
8	DENS	1441	9699	3876.05	1454.102
9	NONW	0.8	38.5	11.87	8.921148
10	WWDRK	33.8	59.7	46.08167	4.613043
11	POOR	9.4	26.4	14.37333	4.160096
12	HC	1	648	37.85	91.97767
13	NOX	1	319	22.65	46.33329
14	SO@	1	278	53.76667	63.39047
15	HUMID	38	73	57.66667	5.369931
16	MORT	790.73	1113.16	940.3585	62.20669

Statistical Analysis of the Data Set

The statistical analysis of the Air Pollution data set is also done. It is important as it will help to interpret our results. The description about the ways to do the statistical analysis of the data set and their importance has already been discussed. For the analysis of the data set probability graphs have been constructed for all the attributes of the data set. From Figure 2.12 it is clear that the attribute "MORT" follows Gaussian distribution. Similarly from Figure 2.13 it is shown that attribute "HC" follows non-Gaussian distribution. All other attributes have also been evaluated and the probability graphs are added in "APPENDIX B". The statistical distribution of all the attributes is summarized in Table 2.10.

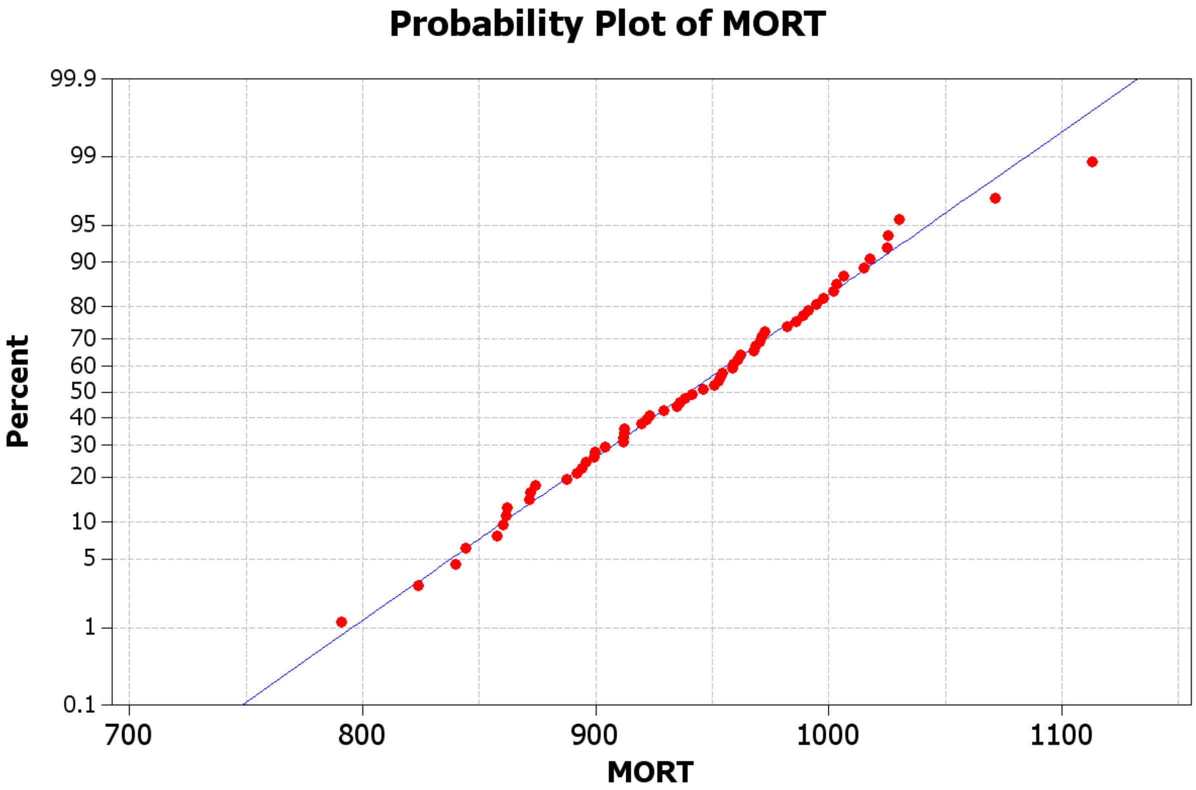


Figure 2.12: Probability Plot of MORT showing Gaussian distribution

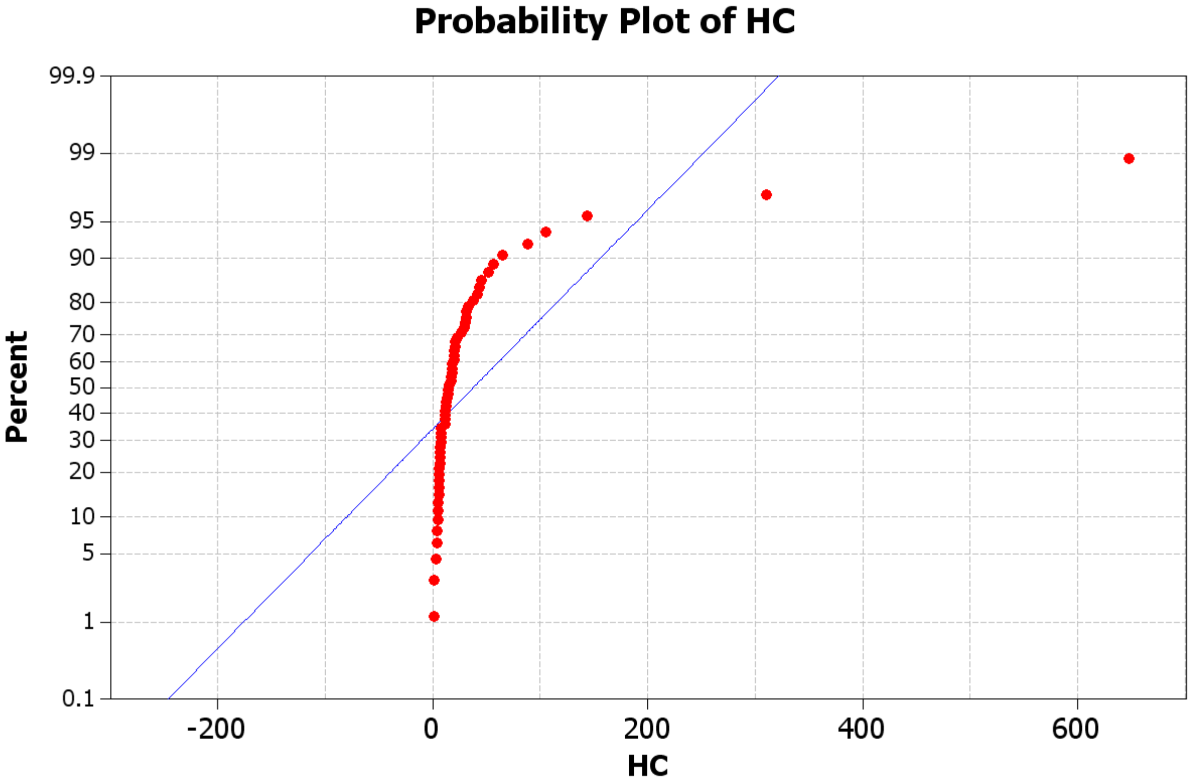


Figure 2.13: Probability Plot of HC showing non-Gaussian distribution

Table 2.10: Statistical distribution of different Variables for Air Pollution data

No. of Attribute	Attribute	Distribution
1	PREC	Non-Gaussian
2	JANT	Non- Gaussian
3	JULT	Non- Gaussian
4	OVR65	Gaussian
5	POPN	Gaussian
6	EDUC	Non-Gaussian
7	HOUS	Gaussian
8	DENS	Gaussian
9	NONW	Gaussian
10	WWDRK	Gaussian
11	POOR	Non-Gaussian
12	HC	Non-Gaussian
13	NOX	Non-Gaussian
14	SO@	Non-Gaussian
15	HUMID	Non-Gaussian
16	MORT	Gaussian

Chapter # 3

Estimation of Process Parameters using Principal Component Analysis

For fault detection in a system PCA has been successfully used in recent times. PCA has proved its superiority for both historical data sets and online systems for multivariate data analysis. Also for the smooth operation of any plant, it is very important for the process parameters to be in range so there are no violating values of the different parameters. If we can use PCA for the determination of these process parameters, it will lead to another area of PCA usage in industry. In this work PCA is used for the estimation of process parameters using historical data set. First PCA is used for the detection of faulty days. After having predicted the abnormal days (days during which plant had problem due to some operating parameters) in the data set, PCA technique is used on the fault free days to utilize it for determination of process parameters by using the Eigen vectors of the fault free data set.

After process parameters are estimated, the validation of the obtained results is carried out. Data mining software is used to construct a Decision Tree. After considering the Decision Tree it is found that it validates most of the results obtained. For the parameter which was not validated by Decision Tree it was observed that is had huge variations in its values which led to inability of Decision Tree to compute its validation. The other reason for non-validation could be that PCA is applicable to Gaussian data set and here we have mixture of both Gaussian and non-Gaussian data set.

3.1 Introduction

With the evolution of more and more computer aided systems, there is an increase in the ability to record operational parameters in any process. As the number of variables increase the recording also becomes very important. The increased number of recycling streams going into the system makes the parameter recording even more important in order to get in depth to the behaviour of the process. Hence the recent focus is more towards MSPC rather than USPC.

As there are many parameters involved in the operation of a plant, so parameter estimation can be one of the areas of work in the field of multivariate data. There are different models which can be used for the estimation of parameters. These include linear models, Gauss-

Newton method for algebraic models and Ordinary Differential Equation (ODE) models with linear dependence on parameters etc. Mathematical models are mostly utilized for the estimation of process parameters. However in many of the mathematical models certain constraints of the model also have to be fulfilled. Mostly these constraints are equality or inequality in the predictive model which also have to be considered for the utilization of the model (Englezos and Kalogerakis 2001).

Issanchou et al. (2003) have developed a model for slow and continuous stirred batch reactions in a liquid-liquid medium. The parameters for the process have been identified by considering it as a non-linear least squares problem. The reaction conditions are such that there is a very slow chemical reaction taking place. The parameters estimated include both physical kinetic parameters and chemical kinetic parameters and are calculated simultaneously. Although there are three different criteria used for precise parameter estimation still there is need for single criteria which is relative to these ones and has to be tested for further accuracy in parameter estimation.

Lohmann et al. (1992) have utilized a method for process parameters estimation based on multiple shooting algorithms and a generalized Gauss-Newton method. It is mainly applicable to non-linear problems. It also helps predict the parameters for different unstable systems and reactions which take place in chemistry and chemical engineering. They have presented their work by its application to kinetic model in coal pyrolysis process.

Schwaab et al. (2008) have used Particle Swarm Optimization (PSO) method to overcome the problem of non-linearity in estimation of process parameters and their statistical analysis. PSO is shown to be precise for minimization and construction of confidence region of parameters estimated. It is applicable to systems with high parameter correlations, system having low sensitivity of objective function to model parameters and systems with discontinuous objective function. Although it gives good results but has the limitation that high computational time is required for it.

PCA is one of the techniques which are used for the monitoring of process parameters, with specific focus to the multivariate nature of the process data. PCA model has been used for process monitoring quite successfully for both historical data analysis and for online monitoring of the system. A lot of work is being done on further development of PCA to get even more accurate results for process monitoring. Although PCA is more related to utilization of the parameters to find out faults in the process but in this work PCA model is utilized to estimate the parameters of the process. Historical data set of WWTP discussed and statistically evaluated in section 2.4 of chapter 2 is used for the achievement of this purpose.

The rest of the chapter describes all the steps of this work. In section 3.2 the methodology utilized for this work is explained. In section 3.3 application of this methodology is carried out which includes identification of faulty days and the data pre-treatment techniques used. In section 3.4 validation of the results obtained is done which also includes introduction to data mining and generation of Decision Tree. Finally the conclusions are added in section 3.4.

3.2 Methodology Used for the Parameters Estimation

In the determination of PCs first the mean centered data is calculated by subtracting the mean of each data set from the original data set. Once we have calculated the mean centered data set then the Eigen vectors of this data set are calculated. Finally for the determination of Principal Components these Eigen vectors and mean centered data sets are multiplied in the transposed form. The mathematical form of the above description can be written as eq. (3.1)

$$\mathrm{PC} = (\mathrm{EV})^{T}(\mathrm{M})^{T} \tag{3.1}$$

This methodology of calculating Principal Components is used in this work. Here “PC” represents the principal components, “EV” is the Eigen vectors of the data set and “M” is the mean centered data. The estimation of parameters can be divided into four sections which are discussed as follows.

3.2.1 Identification of Faulty Days

The first step for the estimation of the process parameters is the identification of faulty days. For this purpose PCs are calculated. Based on the values of PCs the faulty days are determined. It is done because after the determination of faulty days these will be removed from the data set to get fault free days. These fault free days will be utilized to find out the Eigen vectors of the data. Finally these Eigen vectors will be used to estimate process parameters. If we do not remove the faulty days then the Eigen vectors calculated will represent the faulty days as well and lead to inaccurate estimation of process parameters.

3.2.2 Calculation of Eigen Vectors of Fault Free Days

After removing the faulty days we are left with fault free days with all having their values of PCs in certain range. Hence we get a specific range of PCs values representing fault free days. It is done because in the determination of new parameters we will be using fault free PCs values within the range of these calculated PCs.

After the determination of fault free days we again calculate the Eigen vectors and PCs of these fault free days.

3.2.3 Calculation of New Parameters

Once we have got the Eigen vectors of fault free days the next step is the calculation of new parameters. We know the calculation of PCs can be represented by eq. (3.1) as follows.

$$PC = EV^{T} \text{ x } M^{T}$$

As we have the Eigen vectors of unknown parameters we can arrange eq. (3.1) to estimate unknown parameters. We utilize the Eigen vectors and assume values of PCs (based on the work done in section 3.2.2). We will see in the coming section by application of the methodology on a data set that we get PC in the range of '1' and '-1'. We can utilize any value of PCs within this range and hence calculate the unknown parameters. This process can be represented by eq. (3.2)

$$M^{T}=\text{Inv } (EV^{T}) \text{ } (PC) \qquad (3.2)$$

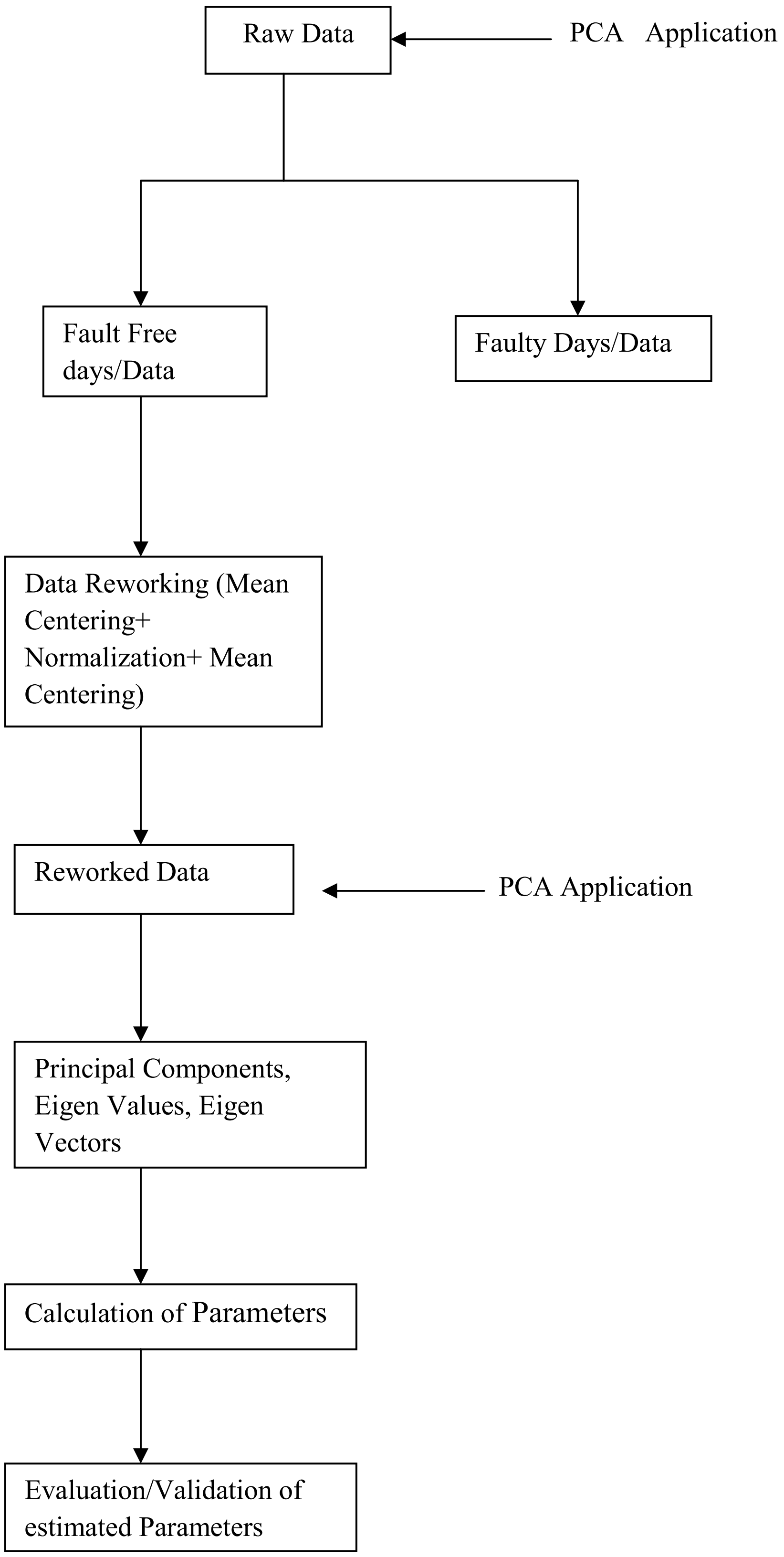
Here "M" is the mean centred data, "EV" represents the Eigen vectors (with "Inv" representing that we are using inverse of Eigen vectors) and "PC" is the assumed Principal Components.

3.2.4 Validation of Calculated Parameters

After the calculation of the new parameters, the final step is the validation/evaluation of these parameters. Evaluation of the parameters is carried out using the data mining software See 5.0.

For the validation of the results we take one of the parameters as the primary parameter (which is the requirement of the software for construction of Decision Tree). Based on this primary parameter the values of the rest of the parameters are determined. Out of all the parameters any parameter can be taken as a primary parameter. A Decision Tree is constructed by the software which gives an idea that with certain value of the primary parameter what should be the values of other parameters. For the validation of results, the values calculated using PCA model and the values determined by the software (in the form of decision Tree) are compared. For the validation of parameters data mining and Decision Tree are explained in section 3.4.1.

The whole methodology can be represented as in Figure 3.1.



Raw Data
PCA Application
Fault Free days/Data
Faulty Days/Data
Data Reworking (Mean Centering+ Normalization+ Mean Centering)
Reworked Data
PCA Application
Principal Components, Eigen Values, Eigen Vectors
Calculation of Parameters
Evaluation/Validation of estimated Parameters

Figure 3.1: Flow Diagram of the Methodology used for Parameter Estimation

3.3 Application of Methodology to Data Set

The methodology is applied to waste water treatment plant (WWTP) data set. The description about the data set and process involved is already given in Chapter 2 of this thesis in section 2.4. The steps and the results for the application of methodology are discussed below.

3.3.1 Identification of Faulty Days

As explained in section 3.3 we know the estimation of process parameters is dependent on Eigen vectors. Here the Eigen vectors are calculated using the historical data set. We have 527 days of historical data. At the start, this data set included both fault free days and faulty days. We are interested to find out fault free days for this work because we want to estimate the parameters for smooth running of the plant. The technique of PCA is utilized to find out faulty days in the data set. After having calculated the faulty days, they are removed from the 527 days data set to get fault free data set. Now these fault free days are utilized to calculate the Principal Components and Eigen vectors. Then using these Eigen vectors the process parameters for smooth running of plant are calculated. The determination of faulty days is very important because if they are not removed then the final results will include parameters representing the out of control plant as well.

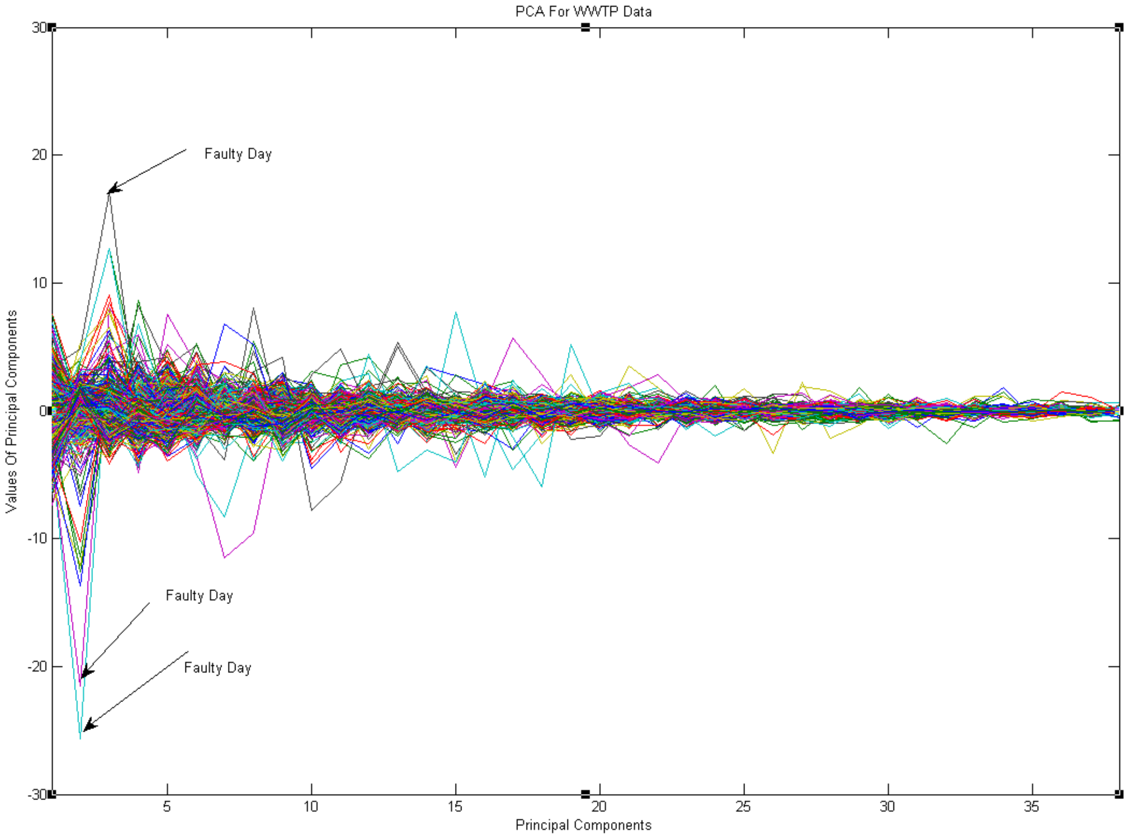


Figure 3.2: PCA Indicating faulty days based on quite different values of PCs

Principal Component Analysis is applied on wastewater data set and as PCA is a data visualization technique it is quite evident from Figure 3.2 that the first few Principal Components (first 15 in this particular case) identify most of the faulty days. As in Figure 3.2 for PCA we have many values of PC1 which are quite higher or lower than the other Principal Components. We can easily investigate these Principal Components and based on the different values of the PCs we can identify the faulty days .e.g. in PC2 we have one of the PC as -12.1108 which is quite different from the other Principal Components so based on this we can easily say that 15/03/1990 is one of the faulty days in the data set of 527 days.

Albazzaz et al. (2005) have also confirmed the presence of these faulty days. In their work they have also utilized T^2 for comparison of results with other multidimensional visualization techniques and T^2 has also confirmed 14 days as abnormal days from these 23 days. In this work these 23 faulty days are removed from the original data set and the fault free data set is retained for the determination of Eigen vectors and Principal Components and consequently process parameters.

3.3.2 Data Pre-treatment and Parameters Estimation

There are different data pre-treatment techniques which can be utilized on the data before the application of PCA on the fault free data set. These include data standardization, mean adjusted data utilization and data normalization etc. The application of each technique to this work has got its own limitations. We cannot use any single data pre-treatment technique mainly because of getting very high range of principal component values. These high values of PCs lead to inaccurate parameter values. e. g. if we use mean adjusted data then with PC value of 10000 we get flow of 28073.34. With this flow we get Input Biological demand of oxygen to plant (DBO-E) as -1858.52 which is highly impractical.

In order to find out the process parameters we have used combination of data pre-treatment techniques. The raw data is first mean centered to get the mean of the data set as zero and then this data set is normalized in order to get values of all the parameters in the range of 0.01 to 0.99. The following formula has been used.

$$\frac{\text{y-LL}}{\text{UL-LL}} = \frac{\text{x-x}_{\text{min}}}{\text{x}_{\text{max}}\text{-x}_{\text{min}}} \tag{3.3}$$

Where,

LL= Lower limit set for the data to be scaled to (in our case, it is 0.01)

UL= Upper limit set for the data to be scaled to (in our case, it is 0.99)

X $_{\text{min}}$ = Minimum value of the attribute

X $_{\text{max}}$ = Maximum value of the attribute

X =Input value of the parameter

Y= Output value of Parameter in the range of 0.01 to 0.99

Also as mean centering is one of the pre-requisite of PCA calculations so the normalized data is again mean centered in order to get the mean of this data set as zero.

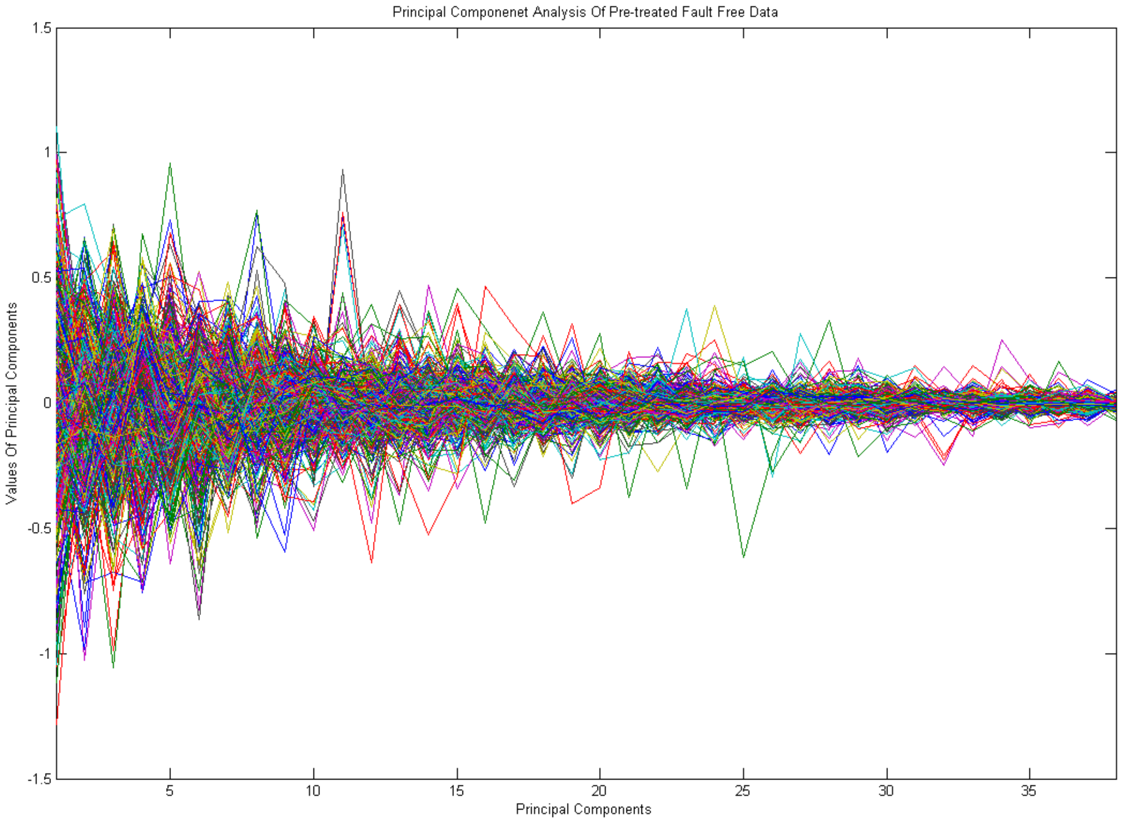


Figure 3.3: PCA Indicating Fault Free and pre-treated data with PCA values in small range

The results obtained by using this pre-treated data set are quite promising, although for the calculations of original parameters values it is required to redo/undo (de-normalization and de-mean centering) all the calculations in order to find out the original values of process parameters.

Principal Component Analysis is carried out on this treated data set and is shown in Figure 3.3. By applying PCA on this we get the values of PCs in the range of “1” to “-1” with clearly most of the PCs lying in the range “0.5” and “-0.5” so we can use these values for the estimation of the unknown parameters.

As discussed before PCA technique is used to remove faulty days. Data pre-treatment is carried out after the removal of faulty days. Calculation of PCs and Eigen vectors is the last step before parameters estimation. Once we have done all these steps then we are in a position to calculate the parameters for smooth running of the plant. The calculation of parameters is done by using eq.3.2.

Here we utilize the calculated values of the Eigen vectors. For the PCs we have got a range of PCs (already described in this section) and utilizing any values of PC we can calculate the parameters. For this case study different values of PCs are taken and all the 38 parameters are calculated. The details of the parameters estimated and validation of results by Decision Tree for WWTP is described in section 3.4.2.

3.4 Validation of Results and Discussion

3.4.1 Introduction to Inductive Data Mining and See 5.0

In this section we introduce the concepts and methodologies for inductive data mining and See 5.0. As the results in this section are validated by this software so the basic concepts of data mining and working of See 5.0 is explained.

a) General Concept of Inductive data mining

Inductive data mining is a technique for generation of Decision Trees and production rules from data cases.

The appeal for construction of Decision Trees for data analysis originates primarily from three inherent properties:

1. Ability to model non linear relationships

2. Ease of interpretability

3. Non-metric nature of the Decision Tree.

Decision Trees have been found to be able to handle large scale problems due to their computational efficiency, to provide interpretable results and in particular, to identify the most representative attributes for a given task.

Decision Trees are very important tools in data mining because they have the capacity to model complex data spaces and unlike traditional methods such as linear discriminate analysis, decision trees are capable of capturing nonlinear relationships within representative data. On comparison with many statistical techniques, they are non parametric and hence make no assumption about the underlying distribution of the data.

Decision Trees are not only capable of modelling nonlinear relationships but they also have a high level of interpretability. A typical Decision Tree consists of a root node linked to two or more child nodes which may or may not link to further child nodes. Every nominal node within the tree represents a point of decision or data splitting based upon the data (DeLisle and Dixon 2004)

Most inductive data mining methods for Decision Tree generation use supervised learning i.e. learning from a set of pre-classified cases. The most well known method probably is See 5.0. It was developed by (Quinlan 1993a, 1986, 1996) that produces decision tree with the following requirements:

Attribute Value Description: The data to be analysed must be a flat file. All information about one object must be expressible in terms of a fixed collection of properties or attributes. Each attribute may have either discrete or numerical value, but the attributes used to describe a case must not vary from one case to another. This restriction rules out domains in which objects have inherently variable structure.

Predefined Classes: The categories to which cases are to be assigned must have been established prior to the construction of Decision Tree.

Discrete Classes: It represents the requirement that the class must be sharply delineated, a case either does or does not belong to particular class and there must be far more cases than classes.

Sufficient data: Inductive generation proceeds by identifying patterns in data as noted above. The amount of data required is affected by factors such as the numbers of properties and classes and the complexity of the classification model, as these increase, more data will be needed to construct a reliable model.

Logical' classification model: The programs construct only classifiers that can be expressed as Decision Trees or sets of production rules. These forms illustrated later, essentially restrict the description of a class to a logical expression whose primitives are statements about the value of particular attributes (Quinlan 1993a).

See 5.0 draws Decision Trees with the help of gain criterion ratio; hence it is inevitable to understand the logic of gain criterion ratio. We take a simple example given in Table 3.1 to illustrate the function of gain criterion ratio.

Table 3.1: An example for explanation of Decision Tree

Outlook	Temp(^{o}F)	Humidity (%)	Windy?	Class
Sunny	75	70	True	Play
Sunny	80	90	True	Don't Play
Sunny	85	85	False	Don't Play
Sunny	72	95	False	Don't Play
Sunny	69	70	False	Play
Overcast	72	90	True	Play
Overcast	83	78	False	Play
Overcast	64	65	True	Play
Overcast	81	75	False	Play
Rain	71	80	True	Don't Play
Rain	65	70	True	Don't Play
Rain	75	80	False	Play
Rain	68	80	False	Play
Rain	70	96	False	Play

Table 3.1 consists of a set of weather data with 14 different cases and two classes that classifies the data. Each case in Table 3.1 is pre-classified in some class. It is used here to establish the cause for the classification of any particular case in some class.

Let,

S = Set of Cases

Then

freq (Ci, S) = number of cases in S belonging to the class Ci.

|S| = total number of cases in S.

Now imagine selecting one case at random from the set S and announcing that this case belongs to C_i. This statement has the probability

$$\text{Probability}=\frac{\text{freq}\left(C_i, S\right)}{|S|} \tag{3.4}$$

And the information conveyed by above message depends upon its probability and is measured in bits as minus the log to base 2 of that probability

$$\text{Information Conveyed} = -\log_2 \left\{ \frac{\text{freq}(C_i, S)}{|S|} \right\} \tag{3.5}$$

As an example, if there are 8 equal probable messages, then,

$$\text{Probability} = 1/8$$

And the information conveyed by each message is

$$\text{Information conveyed} = -\log_2 (1/8) = 3 \text{ bits}$$

To find the expected information from such a message pertaining to class membership, we sum over the classes in proportion to their frequencies in the set S using

$$\text{info } (S) = \sum_{i=1}^{n} (\text{Probability} \times \text{Information Conveyed}) \tag{3.6}$$

$$\text{info } (S) = -\sum_{i=1}^{n} \left[\frac{\text{freq}(C_i, S)}{|S|} \times -\log_2 \left\{ \frac{\text{freq}(C_i, S)}{|S|} \right\} \right] \tag{3.7}$$

This is also known as the entropy of a set.

When applied to the set of training cases, info (T) measures the average amount of information needed to identify the class of a case in T. The expected information required as analogues to Eq. 3.8.

$$\text{info } (T) = -\sum_{i=1}^{n} \left[\frac{\text{freq}(C_i, T)}{|T|} \times -\log_2 \left\{ \frac{\text{freq}(C_i, T)}{|T|} \right\} \right] \tag{3.8}$$

where ‘i’ in Equation 3.9 is the number of classes of the data. In our example, we have only two classes of the weather data .i.e. Play and Don’t play.

If T is portioned according with ‘n’ outcomes of a test. The expected information requirement can be found as weighted sum over the subset as

$$\text{info}_X(T)=\sum_{i=1}^{n}\left\{\left(\frac{\text{freq}(C_i,T)}{|T|}\right)\times\text{info}(T_i)\right\} \tag{3.9}$$

Where

$$\text{info}\ (T_i)=-\sum_{i=1}^{n}\left[\frac{\text{freq}(C_i,T)}{|T|}\times\text{-log}_2\left\{\frac{\text{freq}(C_i,T)}{|T|}\right\}\right] \tag{3.10}$$

and here ‘i’ is the number of classes in the attribute used to measure the expected information for an attribute to use for splitting of the data. If we use the attribute ‘Outlook’ in equation 3.6, then i = 3 as the attribute Outlook has three subclasses i.e. Sunny, Overcast and Rain.

Gain measures the information that is achieved by partitioning T in accordance with the test X. This is the Gain Criteria and it then selects the test to maximize this information gain (known as the mutual information between the test X and the class).

The gain here is

Eq 3.8 – Eq3.9

$$\text{Gain}(X)=\text{info}(T)\text{-info}_X(T) \tag{3.11}$$

The attribute with the higher gain is then used as splitting attribute for portioning the cases into different classes.

We can explain the Gain Criterion ratio by the illustration of the example in Table 3.4

As in the example

No. of Classes	= 2 (i.e. Play & Don’t Play)
Total No. of Cases	$\|T\|$ = 14
No of cases belonging to Class Play	$\|T_P\| = 9$
No of cases belonging to Class Don’t Play	$\|T_D\| = 5$

The average information required to identify the class of a case can be calculated by Eq 3.8,

$$\text{info}(T) = -(9/14) \times \log_2(9/14) - (5/14) \times \log_2(5/14)$$

$$= 0.940 \text{ bits} \qquad \text{(i)}$$

Taking Outlook as a test to divide T into three subsets i.e. Sunny, Overcast and Rain and using Eq 3.9 i.e.

$$\text{info}_X(T) = \sum_{i=1}^{n} \left\{ \left(\frac{\text{freq}(C_i, T)}{|T|} \right) \times \text{info}(T_i) \right\}$$

$$\text{info}_x(T) = \{5/14\ (-3/5 \times \log_2(3/5) - 2/5 \times \log_2(2/5))\}$$

$$+ \{4/14\ (-4/4 \times \log_2(4/4) - 0/4 \times \log_2(0/4))\}$$

$$+ \{5/14\ (-3/5 \times \log_2(3/5) - 2/5 \times \log_2(2/5))\}$$

$$= 0.694 \text{ bits} \qquad \text{(ii)}$$

Subtracting (ii) from (i), we get

(i) - (ii) = 0.246

i.e. Gain = 0.246

Now repeat with the attribute Windy, that will give two subsets i.e. True and False, we will get

$$\text{info}_x(T) = \{6/14\ (-3/6 \times \log_2(3/6) - 3/6 \times \log_2(3/6))\}$$

$$+ \{8/14\ (-6/8 \times \log_2(6/8) - 2/8 \times \log_2(2/8))\} = 0.892 \text{ bits} \qquad \text{(iii)}$$

Again, Subtracting (iii) from (i), we get

(i) - (iii) = 0.048

Gain = 0.048

Hence, the Gain Criterion prefer the test on the attribute Outlook over the test on attribute Windy as the gain from attribute outlook i.e. 0.246> 0.048

Gain Criterion has good results but it has a serious deficiency, it has a strong bias in favour of tests with many outcomes. Considering a hypothetical medical diagnosis task in which one of the attribute contains patient's identification. Since every such identification is intended to be unique, partitioning any set of training cases on the value of this attribute will lead to a large number of subsets, each subset containing one class. Since all of these one-case subclass contain cases of a single class, then

$$\text{info}_X(T) = 0$$

And information gain using this attribute to partition the set of training cases is maximal. From the point of view of prediction, such a division is quite useless.

Now consider the information content of a message that indicate not the class to which the case belongs but the outcome of the test.

By analoging the definition of info(S) i.e. Eq 3.9 we get

$$\text{Split info }(X) = -\sum_{i=1}^{n}\left[\frac{\text{freq}(C_i, T)}{|T|} \times \log_2 \left\{\frac{\text{freq}(C_i, T)}{|T|}\right\}\right] \tag{3.12}$$

This represents the potential information generated by dividing T into n subsets. The information gain i.e. Eq. 3.12 measures the information relevant to classification that arises from same division, then

$$\text{Gain Ratio} = \text{Gain}(X) \Big/ \text{Split Info}(X) \tag{3.13}$$

The gain ratio criterion selects a test to maximize the ratio above, subject to the constraint that the information gain must be large to avoid the bias inherent of gain criterion.

If the split is nontrivial, split info will be small and this ratio will be unstable. To avoid, gain ratio criterion will then select the test to maximize the ration above subject to the condition that the information gain must be large Quinlan (1993a).

This is how gain criterion ratio is used in inductive data mining using See5.0 to construct the decision trees. Interested readers, for detailed information are referred to Quinlan (1993b).

b) An Example

Below is the structure of the Decision Tree drawn from the example of Table 3.1.

Decision tree:

```
Outlook = Overcast: Play (4)
Outlook = Sunny:
:...Humidity <= 75: Play (2)
:...Humidity > 75: Don't Play (3)
Outlook = Rain:
:...Windy = TRUE: Don't Play (2)
…Windy = FALSE: Play (3)
```

The above is the output format from See 5. It can be re-drawn as a more common decision tree structure as Figure 3.4.

This is how a Decision Tree classifies the data into different cases. The See 5.0 first chooses the best splitting node (Outlook in this case) and then classifies the data with respect to the attribute outlook. It classifies that if the attribute Outlook is overcast, then the cases belong to the class Play and if the Outlook is Sunny, we have to look at the values of Humidity. If Humidity is less than or equal to 75, the cases belong to the class Play and if greater than 75, it belongs to the class Don’t Play. Also if the Outlook is Rain, we have take into consideration the attribute windy in order to classify the cases into classes as if it is Windy, then class is Don’t Play else Play.

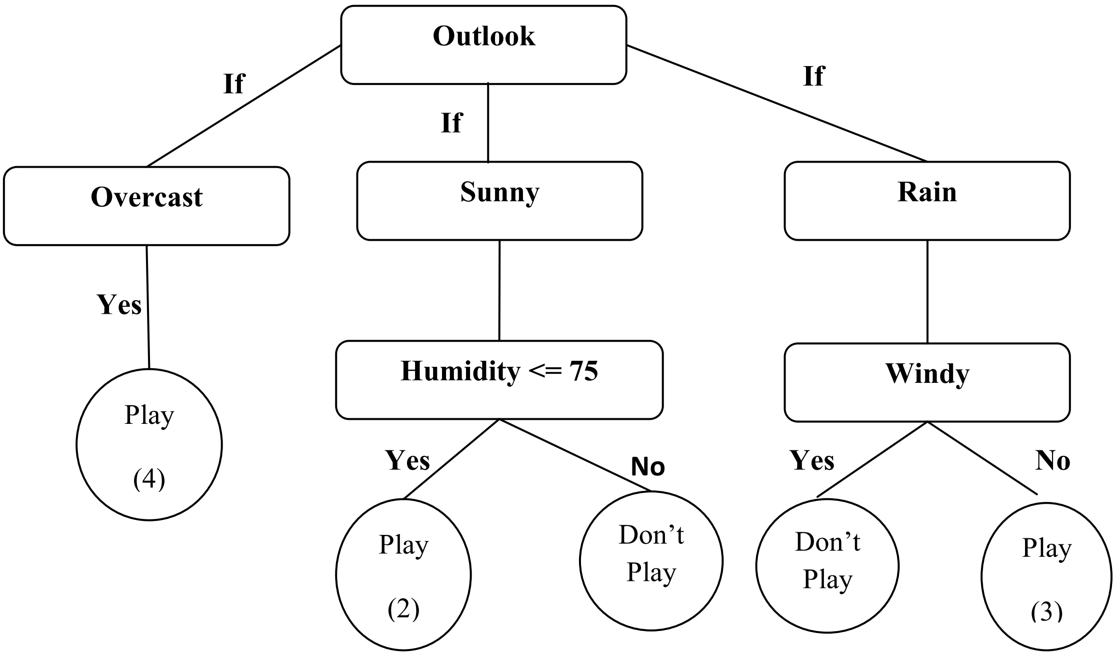


Figure 3.4: The graphical presentation of the Decision Tree generated by See 5. The numeric values in ( ) shows the no of cases classified into each class with this Decision Tree.

C) Applications of Decision Tree

Many applications and usefulness of data mining and Decision Tree are discussed in literature. Murthy (1998) have done a multi-disciplinary survey on automatic construction of Decision Tree from data. They have covered the application of Decision Tree in areas such as statistics, pattern recognition, decision theory, signal processing, machine learning and artificial neural networks. Their survey involves existing work on decision tree construction, identification of important issues involved and the directions of work in this area.

Sun et al. (2007) have introduced a data mining technology in fault diagnosis field. They have proposed a new method based on C4.5 decision tree and PCA. They have used PCA to reduce features after data collection, pre-processing and feature extraction. Then a Decision Tree is generated with diagnosis knowledge. Finally the tree model is used to make diagnosis analysis. The technique is applied for fault diagnosis of rotating machinery. The results are compared with back-propagation neural network (BPNN). The results show that C4.5 and PCA-based diagnosis method has higher accuracy than BPNN.

3.4.2 Validation of Results and Discussion

Using the eq. (3.2) we utilize the calculated Eigen vectors from fault free data set and take PCs as -0.005 to calculate all the 38 parameters. Once all the parameters are calculated the next step is the validation of the parameters. In this case the validation is done by utilization of the data mining software, See 5.0 which has been explained in detail in section 3.5.1. Demo version of See 5.0 is freely available on the internet and in this work validation of the obtained results is carried out using this software. As per the requirement of the data mining software one of the parameters had to be taken as primary parameter and based on this all others are calculated. In this case the calculated flow rate is 38,028.86 (with PC value of -0.005) so it is taken as the primary parameter and for all the other parameters data mining is done for their validation.

After having calculated all the 38 parameters, data mining is done by constructing Decision Tree. As we see from the Decision Tree in Figure 3.6 it cannot give us exact information about all the 38 parameters, but it can give us a range of values for any of the parameters so the Decision Tree has two main parts. The results by Decision Tree sets can be evaluated as follows.

a) Results discussion of First node of Decision Tree

For the first part as we see that for flow rate of 38,028.86 as primary parameter we get seven other parameters. For these parameters, out of seven, six are fully validated. The details of this validation can be summarized by the comparison of the results obtained by the calculations utilizing PCA model and the results given by the Decision Tree. This is discussed below.

Input Sediments to Plant (SED-E)

For flow of 38,028.86 the value of Input sediments to plant (SED-E) predicted by Decision Tree should be greater than 2.7 and our calculated value is 4.66 which is validated.

Input chemical Oxygen demand to Secondary Settler (DQO-D)

The Input chemical demand of oxygen to secondary settler (DQO-D) is supposed to be in the range of 206 and 450. We are having the value of DQO-D as 282.76 which is within the predicted range of the Decision Tree.

Input Biological demand of oxygen to plant (DBO-E)

Input Biological demand of oxygen to plant (DBO-E) calculates value as 189.279 whereas the Decision Tree recommends the value to be more than 182.5 which verifies the calculated value of DBO-E.

Output volatile suspended solids (SSV-S)

The calculated value of output volatile suspended solids (SSV-S) is 79.91 whereas recommended value of SSV-S by the Decision Tree is greater than 74.1.

Input volatile suspended solids to plant (SSV-E)

For input volatile suspended solids to plant (SSV-E) the recommended value by Decision Tree is greater than 57.4 and the value from the data is 61.93 so it is again validated by the software.

Output chemical demand of oxygen (DQO-S)

For output chemical demand of oxygen (DQO-S) we have calculated the value as 87.27 using PCA model whereas the Decision Tree recommends that the value should be less than or equal to 140.

Input suspended solids to secondary settler (SS-D)

There is only one parameter which is almost validated by the software. Input suspended solids to secondary settler (SS-D) whose value should be around 90 whereas we have calculated the value as 94.20. It was investigated and it was found that this parameter had a lot of fluctuations in its original values in the original data set with the values moving from 54 to 230. This is shown in Figure 3.5 as follows.

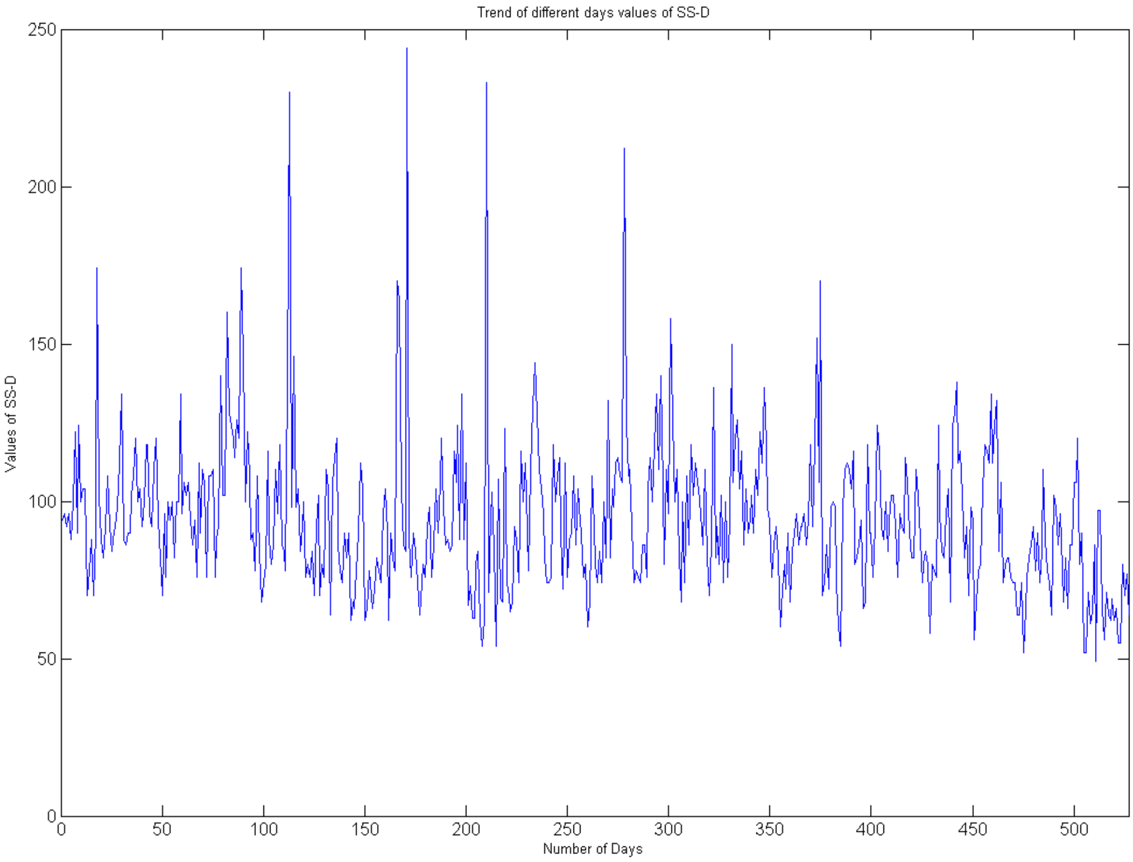
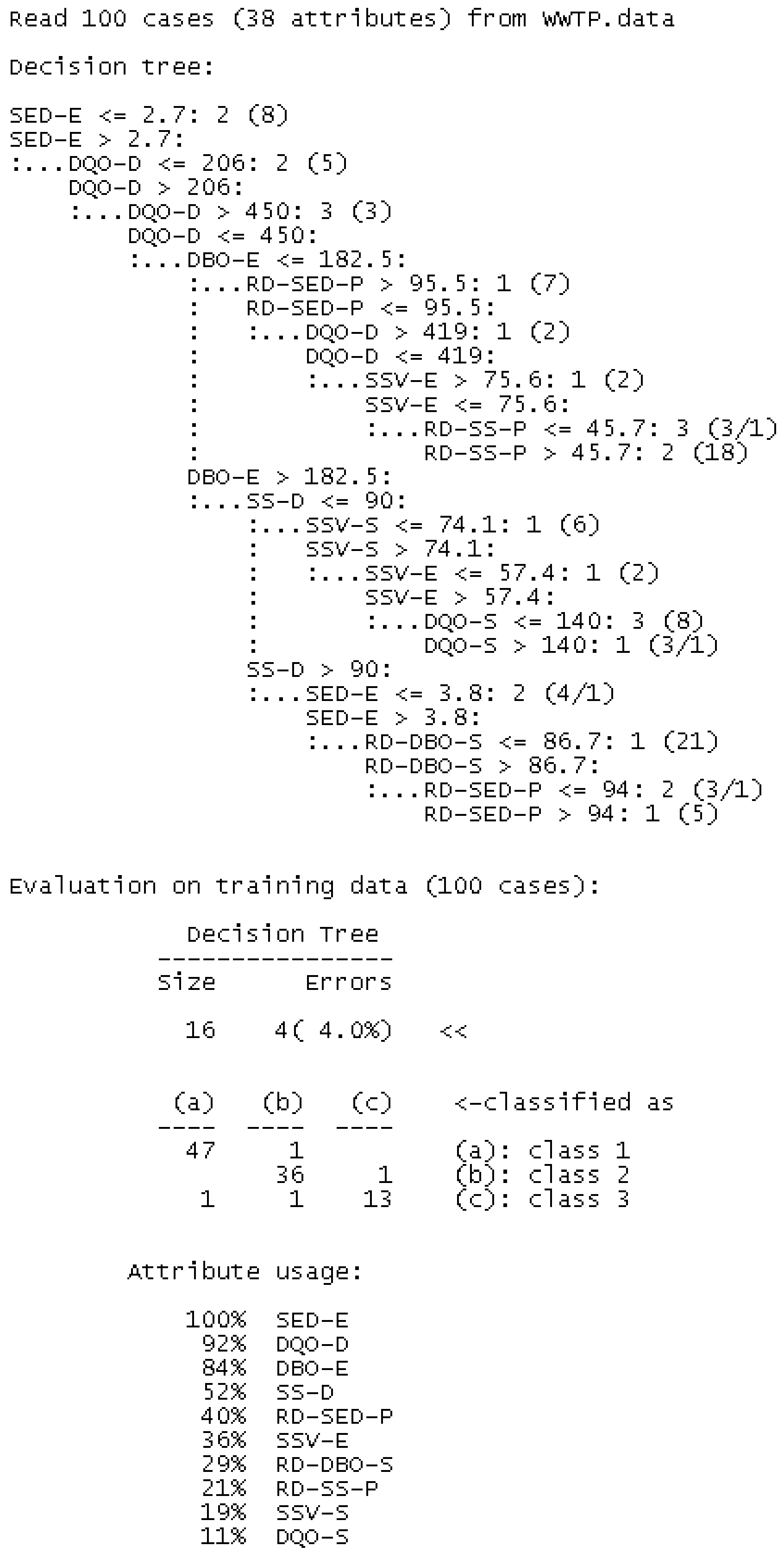


Figure 3.5: Graph showing variation in the different days values for SS-D

A summary of the results obtained by this part of Decision Tree is also given in Table 3.2

Table 3.2: Summary of the Results Obtained by Decision Tree

Numbers	Parameters	Values determined by PCA model	Values by See 5.0	Validated
1	SED-E	4.66	>2.7	Yes
2	DQO-D	282.76	>206≤450	Yes
3	DBO-E	189.279	>182.5	Yes
4	SSV-S	79.91	>74.1	Yes
5	SSV-E	61.93	>57.4	Yes
6	DQO-S	87.27	≤140	Yes
7	SS-D	94.20	≤90	Almost



```
Read 100 cases (38 attributes) from WWTP.data

Decision tree:

SED-E <= 2.7: 2 (8)
SED-E > 2.7:
:...DQO-D <= 206: 2 (5)
    DQO-D > 206:
    :...DQO-D > 450: 3 (3)
        DQO-D <= 450:
        :...DBO-E <= 182.5:
            :...RD-SED-P > 95.5: 1 (7)
            :   RD-SED-P <= 95.5:
            :   :...DQO-D > 419: 1 (2)
            :       DQO-D <= 419:
            :       :...SSV-E > 75.6: 1 (2)
            :           SSV-E <= 75.6:
            :           :...RD-SS-P <= 45.7: 3 (3/1)
            :               RD-SS-P > 45.7: 2 (18)
            DBO-E > 182.5:
            :...SS-D <= 90:
                :...SSV-S <= 74.1: 1 (6)
                :   SSV-S > 74.1:
                :   :...SSV-E <= 57.4: 1 (2)
                :       SSV-E > 57.4:
                :       :...DQO-S <= 140: 3 (8)
                :           DQO-S > 140: 1 (3/1)
                SS-D > 90:
                :...SED-E <= 3.8: 2 (4/1)
                    SED-E > 3.8:
                    :...RD-DBO-S <= 86.7: 1 (21)
                        RD-DBO-S > 86.7:
                        :...RD-SED-P <= 94: 2 (3/1)
                            RD-SED-P > 94: 1 (5)


Evaluation on training data (100 cases):

          Decision Tree
        ----------------
        Size      Errors

          16    4( 4.0%)   <<


         (a)   (b)   (c)    <-classified as
        ----  ----  ----
          47     1          (a): class 1
                36     1    (b): class 2
           1     1    13    (c): class 3


        Attribute usage:

            100%  SED-E
             92%  DQO-D
             84%  DBO-E
             52%  SS-D
             40%  RD-SED-P
             36%  SSV-E
             29%  RD-DBO-S
             21%  RD-SS-P
             19%  SSV-S
             11%  DQO-S
```

Figure 3.6: Decision Tree validating the Results obtained for the Process Parameters Estimation

b) Results discussion of the Second node of Decision Tree

For the second node of the Decision Tree we get data set information which again validates the obtained results. In this case we have six parameters with information from the Decision Tree. For these values, out of six parameters four are fully verified whereas two are almost verified. Details of these results are as follows.

Input sediments to plant (SED-E)

For this branch we have the value of Input sediments to plant (SED-E) to be 4.66 whereas the Decision Tree recommends value to be more than 2.7, so it is validated.

Input chemical demand of oxygen to secondary settler (DQO-D)

The value of Input chemical demand of oxygen to secondary settler (DQO-D) is recommended to be in the range of 206 to 450 and our calculated value is 282.76 which is within the recommended range.

Performance input sediments to primary settler (RD-SED-P)

The value of performance input sediments to primary settler (RD-SED-P) is recommended to be less than 95.5 by the Decision Tree and we obtained the value as 91.05 so it is again validated by the software.

Input volatile suspended solids to plant (SSV-E)

Input volatile suspended solids to plant (SSV-E) is also validated by the Decision Tree as it gives the value to be less than or equal to 75.6 as compared to our calculated value of 61.93.

Input Biological demand of oxygen to plant (DBO-E)

For input biological demand of oxygen to plant (DBO-E) it is almost validated in this case as the calculated value is 189.279 as compared to recommended value of less than 182.5.

Performance input suspended solids to primary settler (RD-SS-P)

For performance input suspended solids to primary settler (RD-SS-P) we calculated the value as 58.27 as compared to 45.7 by the Decision Tree.

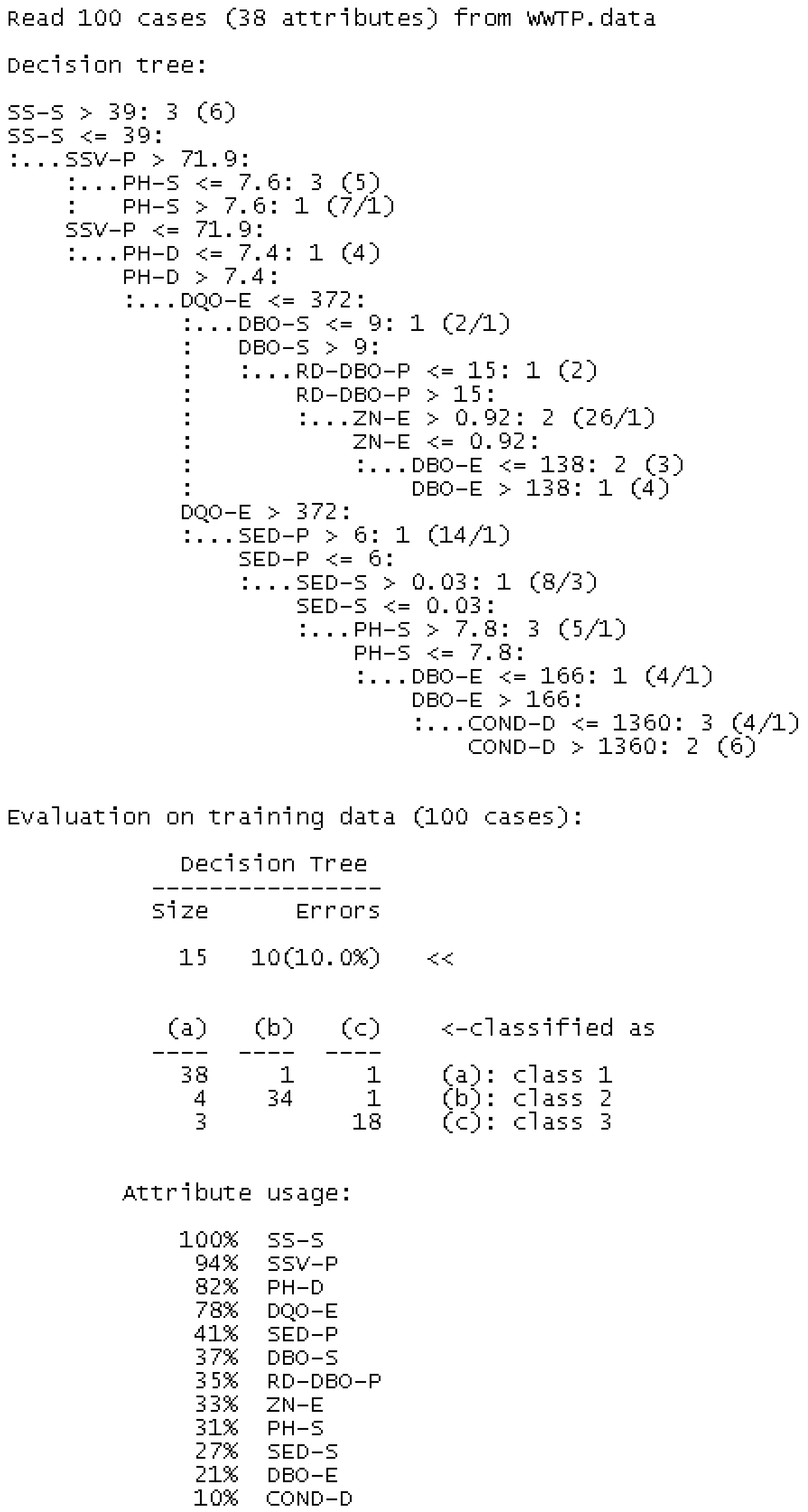
Summary of the results obtained by the Decision Tree is also given in Table 3.3

Table 3.3: Summary of the Results Obtained by Decision Tree

Numbers	Parameters	Values determined by PCA model	Values by See 5.0	Validated
1	SED-E	4.66	>2.7	Yes
2	DQO-D	282.76	>206≤450	Yes
3	RD-SED-P	91.05	≤95.5	Yes
4	SSV-E	61.93	≤75.6	Yes
5	DBO-E	189.279	≤182.5	Almost
6	RD-SS-P	58.27	≤45.7	Almost

c) Results discussion for another Decision Tree

To further verify the estimated parameters by PCA model another Decision Tree is constructed. The data set used for the construction of this tree is taken randomly keeping in view that it is quite different from the initial data used for construction of first Decision Tree in Figure 3.6. It is done for further consolidation of the obtained results. The second Decision Tree is shown in Figure 3.7.



```
Read 100 cases (38 attributes) from WWTP.data

Decision tree:

SS-S > 39: 3 (6)
SS-S <= 39:
:...SSV-P > 71.9:
    :...PH-S <= 7.6: 3 (5)
    :   PH-S > 7.6: 1 (7/1)
    SSV-P <= 71.9:
    :...PH-D <= 7.4: 1 (4)
        PH-D > 7.4:
        :...DQO-E <= 372:
            :...DBO-S <= 9: 1 (2/1)
            :   DBO-S > 9:
            :   :...RD-DBO-P <= 15: 1 (2)
            :       RD-DBO-P > 15:
            :       :...ZN-E > 0.92: 2 (26/1)
            :           ZN-E <= 0.92:
            :           :...DBO-E <= 138: 2 (3)
            :               DBO-E > 138: 1 (4)
            DQO-E > 372:
            :...SED-P > 6: 1 (14/1)
                SED-P <= 6:
                :...SED-S > 0.03: 1 (8/3)
                    SED-S <= 0.03:
                    :...PH-S > 7.8: 3 (5/1)
                        PH-S <= 7.8:
                        :...DBO-E <= 166: 1 (4/1)
                            DBO-E > 166:
                            :...COND-D <= 1360: 3 (4/1)
                                COND-D > 1360: 2 (6)


Evaluation on training data (100 cases):

            Decision Tree
          ----------------
          Size      Errors

            15   10(10.0%)   <<


           (a)   (b)   (c)    <-classified as
          ----  ----  ----
            38     1     1    (a): class 1
             4    34     1    (b): class 2
             3          18    (c): class 3


        Attribute usage:

            100%  SS-S
             94%  SSV-P
             82%  PH-D
             78%  DQO-E
             41%  SED-P
             37%  DBO-S
             35%  RD-DBO-P
             33%  ZN-E
             31%  PH-S
             27%  SED-S
             21%  DBO-E
             10%  COND-D
```

Figure 3.7: Decision Tree validating the Results obtained for the Process Parameters Estimation (With different data set)

As we can see in Figure 3.7, from the Decision Tree that it validated six out of a total nine parameters. This result also validates the working of PCA model for estimation of process parameters. The summary of the results obtained is given in Table 3.4.

Table 3.4: Summary of the Results Obtained by Decision Tree

Numbers	Parameters	Values determined by PCA model	Values by See 5.0	Validated
1	DBO-E	189.28	>166	Yes
2	SSV-P	60	≤71.9	Yes
3	PH-D	7.8	>7.4	Yes
4	DQO-E	401.68	>372	Yes
5	SED-P	4.892	≤6	Yes
6	SED-S	0.0025	≤0.03	Yes
7	PH-S	7.70	>7.8	Almost
8	SS-S	20.64	>39	Almost
9	COND-D	1497.7	≤1360	Almost

3.5 Conclusions

Principal Component Analysis has been used for Process monitoring and is now a well established technique. In this work a new application of Principal Component Analysis is done for the estimation of process parameters. The results have been validated by means of data mining.

Although it gives good results but this method of calculation of the process parameter has its limitations. As the Eigen vectors are calculated using the specific range of parameters so we do have the restriction that the calculations can only be done with in this range .i.e. this method is unable to calculate parameters outside this range. This was expected since we are using the Eigen vectors within a specific range. Although utilization of PCA model has got the limitation that it can only give results within the range of the original data set from where the Eigen vectors and Eigen values are calculated but it may be very useful if we are interested in finding the parameters in this range. Also by this method we get quite good idea about the parameters and for the start up of a plant so it can be a useful tool. Obviously when the plant is in operation there is always fine tuning of the plant required depending on the variations taking place in the plant. Never the less it gives us good information about the plant operating parameters.

Chapter # 4

Estimation of Process Parameters using Independent Component Analysis

Independent Componenet Analysis is a recently developed technique for fault detection and diagnosis. The technique is more focussed for fault detection where data set follows non-Gaussian distribution. Along with ICA, PCA has also shown good results for fault detection in historical data set. In chapter 3 of this thesis, PCA has been used for the estimation of process parameters. In this chapter the technique of ICA is utilized for the estimation of Process parameters for WWTP. ICA is applied to data set to find out faulty days. After the determination of faulty days the fault free data set is utilized to find the ICs values. Once the ICs values range for fault free data is determined alongwith mixing matrix "A" then we utilize it for the estimaton of process parameters, which is explained in section 4.2.

After the estimation of process parameters, the calculated parameters are validated by construction of Decision Tree via inductive data mining. For this particular case two Decision Trees are constructed. In the first Decision Tree, five out of six parameters have been validated. Similarly the results for second Decision Tree are consistent with the first one (five out of six parameters are validated). The reason for non-validation of one of the parameter could be due to the strict application of ICA to non-Gaussian data set and the nature of these parameters .i.e. variations in their values.

4.1 Introduction

Measurement of certain variables in any process is very important. During any process or any experiment, we have a mixture of both fixed and variable prameters. Also in many cases the experimentalists or engineers involved in the process formulate a mathematical model of the data set in order to describe the behaviour of the running process or the expected behaviour shown by the process.

So there is always some model involved to describe the process and the precise explanation of the process depends on the precise selection and utilization of the process model. This implies that with prior knowledge of model, we can estimate the process parameters for a new operational condition with the support of process historical data. In chapter 3 the PCA model has been used to estimate the parameters of a WWTP. PCA demonstrated good results for the estimation of the parameters but still was unable to describe all the parameters accurately. In this chapter we have utilized another multivariate process monitoring technique of Independent Componenet Analysis for the estimation of process parameters. The ICA model is utilized for the estimation and it has shown comparative results.

The rest of the chapter is arranged as follows. In section 4.2 the methodology used for the parameters estimation is discussed. Section 4.3 describes the application of the methodology to the waste water data set. In section 4.4 validation of the results is done along with discussion. Finally conclusions are made in section 4.5.

4.2 Methodology used for the Parameters Estimation

The methodology used for the estimation of the process parameters is based on the ICA model. The ICA model can be described by the equation 2.9 (already discussed in Chapter 2)

We already know that here we have 'X' as the mixed signal or the output signals,'A' is the mixing matrix and 'S' is the matrix of Independent Components. Now utilizing this model of ICA we estimate the process parameters. The following are the steps involved;

1. Identification of faulty days.
2. Calculation of mixing matrix and Independant Componenets.
3. Calculation of new parameters.
4. Validation of calculated parameters.

4.2.1 Identification of Faults in the System

The first step in the utilization of ICA as the model for the parameter estimation is the identificatuion of faults in the system. It is done by the utilization of ICA technique. This step is important because if we do not remove the faults then the calculated mixing matrix will be representing faults. When we calculate the parameters uisng this mixing matrix then it will give us inaccurate results.

4.2.2 Calculation of Mixing Matrix and Independent Components

Once we have got the data set which is free from any faults then we again apply ICA to the data set. This is done to calcualte the mixing matrix of this data set. Alongwith this we also calculate the Independent Componenets. As we have removed the faults from the system so these Independent Components will represent the values where we do not have any faults.

Now in the next step this range of Independnet Components (ICs) values and mixing matrix will be used for the estimation of process parameters.

4.2.3 Calculations of New Parameters

The model equation (e.q 2.9) for ICA is given as

$$X=AS$$

Here ‘X’ is the data set which we get as an outcome, ‘A’ is the mixing matrix and ‘S’ is the matrix of Independent Components. Now if we already have the values of mixing matrix from historical and fault free data set (calculated in section 4.2.2) and the range of ICs then we can utilize these values to find out the process parameters.

4.2.4 Validation of Calculated Parameters

The final step is the validation of process parameters which we have calculated in step 4.2.3. The calculated parameters can be validated by operating the plant on these parameters or by data mining.

In this work data mining is done using software See 5.0 (explained in detail in section 3.5.1). For the data mining one of the parameters is taken as primary parameter and the rest are taken as dependent parameters. Finally the results obtained by the Decision Tree and the results obtained by the ICA model are compared and thus validation or non-validation of the estimated parameters is done.

The whole methodology can be explained by a flow diagram as in Figure 4.1 below.

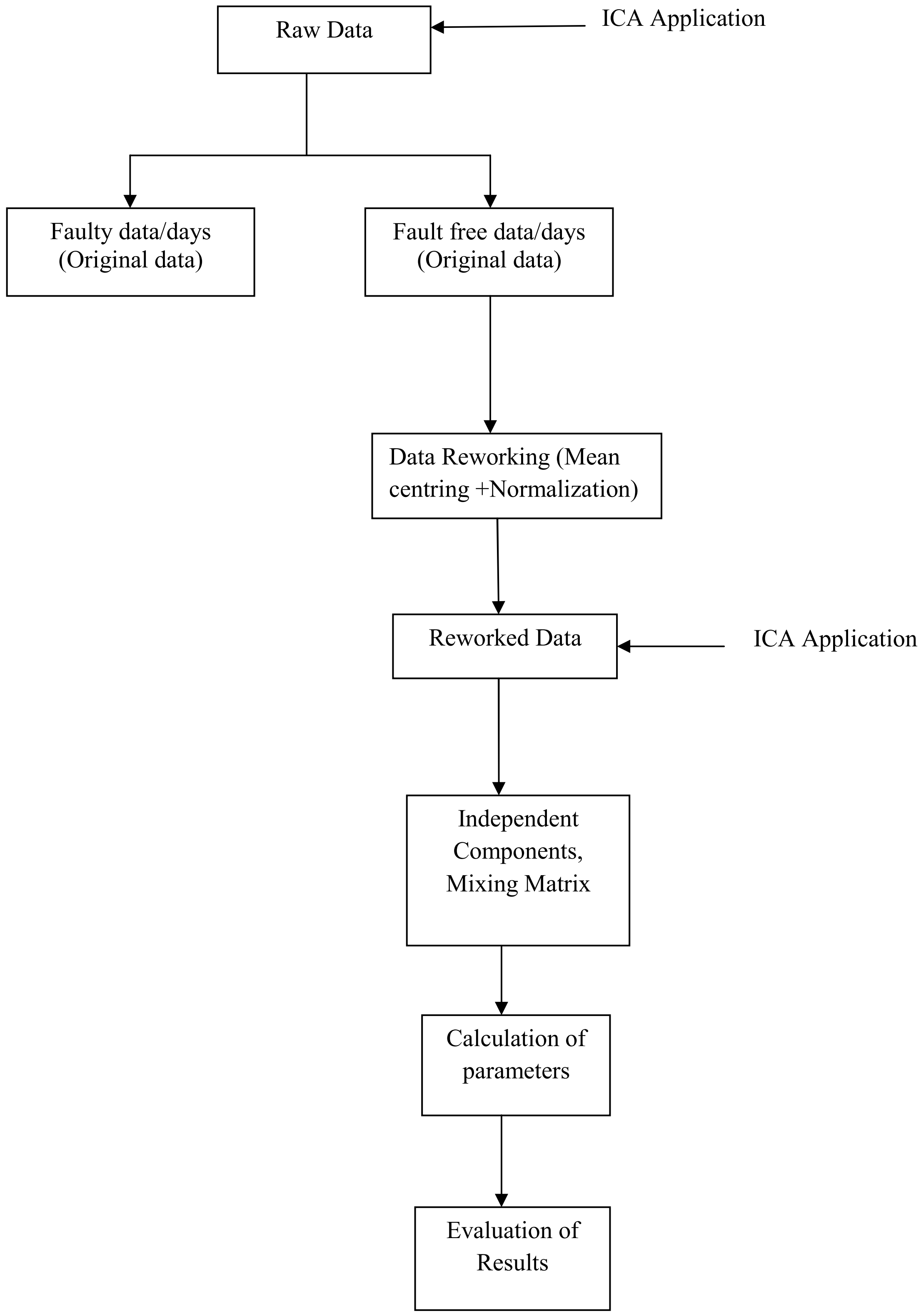

Raw Data
ICA Application
Faulty data/days (Original data)
Fault free data/days (Original data)
Data Reworking (Mean centring +Normalization)
Reworked Data
ICA Application
Independent Components, Mixing Matrix
Calculation of parameters
Evaluation of Results

Figure 4.1: Flow Diagram of the Methodology used for Parameter Estimation

4.3 Application of Methodology to WWTP Data Set

The methodology is applied to waste water treatment plant (WWTP) data set. The description about the data set and process involved is already given in Chapter 2 of this thesis in section 2.4. The steps and the results for the application of methodology are as under.

4.3.1 Identification of Faulty Days

The first step in the estimation of the process parameters is the identification of faulty days. The technique of ICA is applied on the original data set for WWTP. After the application of technique, we identify the faulty days based on the values of ICA. ICA is done on the wastewater treatment plant data using the algorithm developed by Hyvarinen and Oja (2000). The graph is plotted against the number of ICs and the values of ICs and is shown in Figure 4.2.

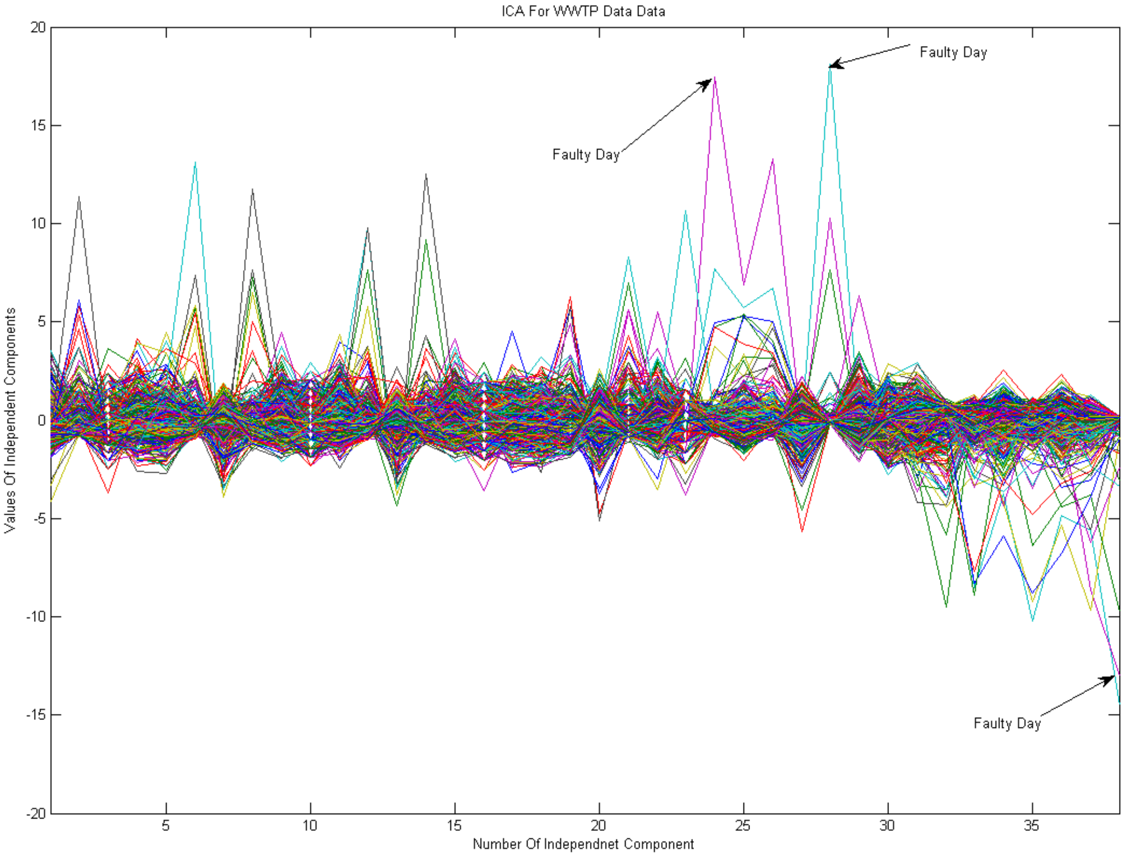


Figure 4.2: ICA showing the faulty days in the data set

As we analyse Figure 4.3, the values of the ICs which are quite different from normal ICs are focus of attention. Based on these values we can find out the faulty day. Also there is a lot of fluctuations shown by the values of ICs but in many of the cases if we identify more than one IC it will lead to the same faulty day (fault) which further verify the presence of faulty day (fault). e.g. if we can analyse for IC 28 on 13/03/1990 then it has a value of 18.1, similarly IC 35 and IC 38 have values of -10.25 and -14.512, respectively so all these values indicate that 13/03/1990 was a faulty day. Similarly for many other days we get more than one abnormal values of Independent Components which prove that the day has some fault.

Based on this work 19 days were identified as faulty. So from the total of 527 days we are left with 508 days as the data set which does not contain any faulty days in the estimation of the process parameters. It was important to remove the faulty days because if we do not remove them then they can lead to inaccurate results for the estimation of process parameters.

4.3.2 Data pre-treatment

The different data pre-treatment techniques have already been discussed in section 3.3.2 of chapter 3. In this work again we have to do some data pre-treatment work. It involves the mean centering of the original data set and then normalization of this mean centered data. Again the restrictions in non-utilization of the original data or just mean centered data are the high values of independent components leading to wrong and impractical results as was the case in parameter estimation using PCA in section 3.3.2 of chapter 3.

In this work we have used combination of data pre-treatment techniques. The raw data is first mean centered to get the mean of the data set as zero and then this data set is normalized in order to get values of all the parameters in the range of 0.01 to 0.99. We have used eq. (3.3) which we used in section 3.3.2 of Chapter 3.

As there is no pre-requisite for ICA to do mean centering again so as compared to the work done for estimation of process parameters using PCA we do not apply mean centering again.

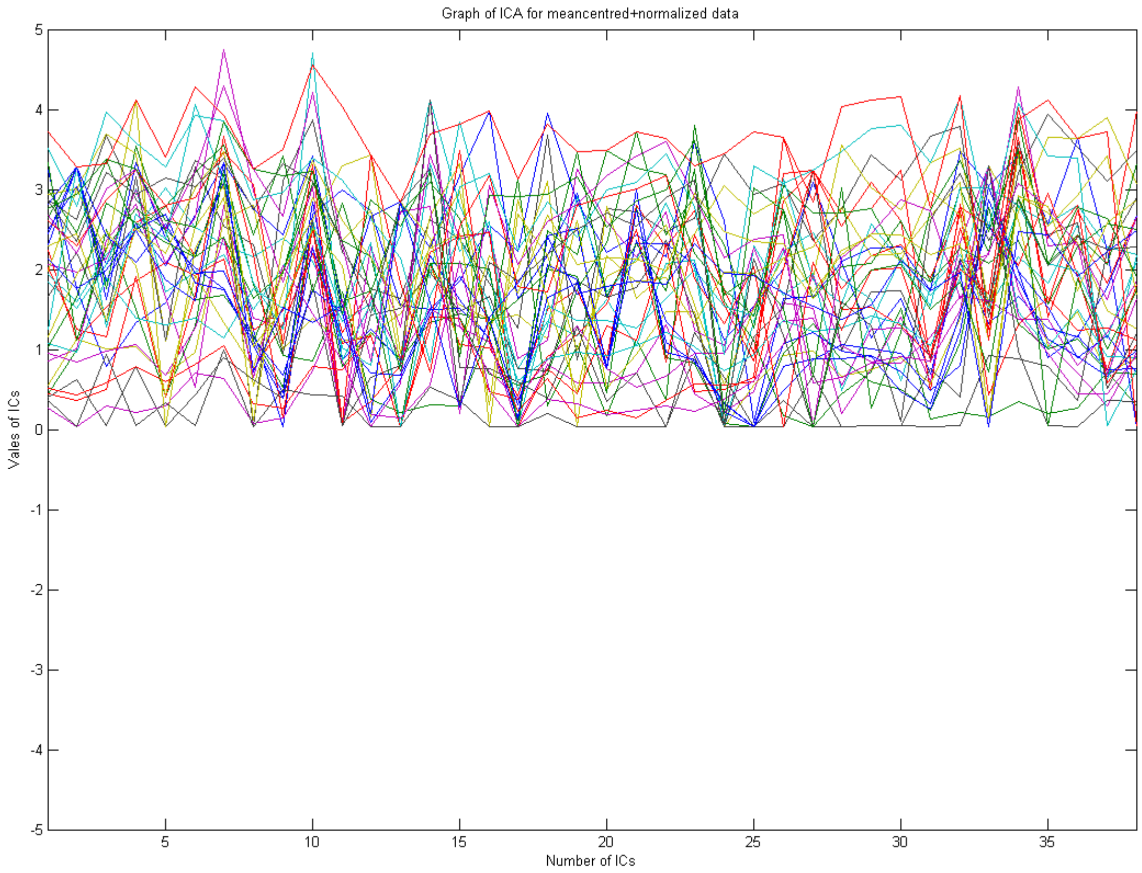


Figure 4.3: ICA on pre-treated data indicating ICs values in small range

The ICA is carried out on the data set and we get values of Independent Components in the range of '4.7' and '0.02' as shown in Figure 4.3. We can use any value of independent components in this range to calculate the parameters.

4.3.3 Parameters Estimation

After calculating the mixing matrix by utilization of fault free historical data and appropriate range of Independent Components we utilize the ICA model for the estimation of process parameters. The model we use for the estimation is represented by eq. (2.9) as below:

$$X=AS$$

Here we have calculated the mixing matrix based on the historical data. The Independent Components values will be used within range of '0.02' and '4.7' as determined in section 4.3.2. The value of Independent Components assumed in this work is '4' which is just to get flow in the range of 47000 to make a case study otherwise we can use any value of ICs to get our desired parameters. Also as in the initial work for the calculation of ICs and mixing matrix in section 4.3.2 the data set was mean centred and normalized prior to application of ICA. Hence our results will also be in the form of mean centred and normalized data. In order to get parameters in their true/original form we do the reverse calculations and thus do de-normalization and de-mean centering to get the true values of the parameters.

4.4 Validation of Results and Discussion

As already discussed we use the basic model of ICA to estimate the process parameters. Now utilizing the ICA model as in eq. (2.9) we calculate the process parameters. The parameters are calculated by utilization of mixing matrix and assuming values of Independent Components in the range of 0.02 to 4.7. In our case we have assumed the Independent Components value to be '4' and based on this we had a flow value of 47687.66. Along with this all the other 37 parameters are also calculated. The results obtained can be validated by data mining or by actual operation of the plant. For the validation of the results in this case we utilize See 5.0 (discussed in section 3.5.1) assuming the flow as the primary parameter since we know that for the data mining using See 5.0 we have to take one parameter as primary parameter.

Based on the historical data and taking flow as the primary parameter we have constructed two Decision Trees which are shown in Figure 4.5 and Figure 4.6. The results for the first Decision Tree in Figure 4.5 are discussed below.

4.4.1 Results Discussion for first Decision Tree

A Decision Tree has been constructed using See 5.0, shown in Figure 4.5. As we see from the Figure 4.5 that we get information about values of six variables. If we do the comparison of the values of these six variables with values calculated by the ICA model then we see that five parameters have been validated from these six variables. A table presenting the calculated parameters values using ICA and data mining is given below and discussed thereafter.

Table 4.1: Summary of the Results Obtained by Decision Tree (Figure 4.5)

Numbers	Parameters	Values determined by ICA model	Values by See 5.0	Validated
1	PH-S	6.4	≤7.9	Yes
2	SSV-E	70.8	>56.7	Yes
3	RD-SS-G	47.58	≤89	Yes
4	SSV-D	39.32	≤66.7	Yes
5	PH-E	8.1	>7.8	Yes
6	SS-E	350.8	≤166	No

a) Output pH (PH-S)

The value of Output pH (PH-S) calculated by the ICA model is 6.4 whereas the recommended value by data mining is less than 7.9 so it is validated.

b) Input volatile suspended solids to plant (SSV-E)

The value of Input volatile suspended solids to plant **(**SSV-E) recommended by the software is more than 56.7. In our calculations by the model we have its value as 70.8 so it is validated.

c) Global performance input suspended solids (RD-SS-G)

The ICA model has recommended Global performance input suspended solids (RD-SS-G) to have value as 47.58 whereas the value obtained by data mining is less than 89. This comparison shows the validation of the results obtained by the ICA model.

d) Input volatile suspended solids to secondary settler (SSV-D)

Input volatile suspended solids to secondary settler (SSV-D) values is obtained to be 39.3 by the ICA model whereas the value recommended by the data mining is less than or equal to 66.7. This proves that the results obtained are correct.

e) Input pH to plant (PH-E)

The input PH to the plant (PH-E) value result given by data mining software is greater than 7.8. The value calculated by ICA model is 8.1 so it is validated.

f) Input suspended solids to plant (SS-E)

The input suspended solids to plant (SS-E) value calculated by the ICA model are 350.8 as compared to recommended value of 166 by the data mining. The reason for its non-validation could be the large range as it has fluctuations from 98 to 2008, as shown in Figure 4.4. The other reason for its non-validation is the application of ICA to non-Gaussian data only.

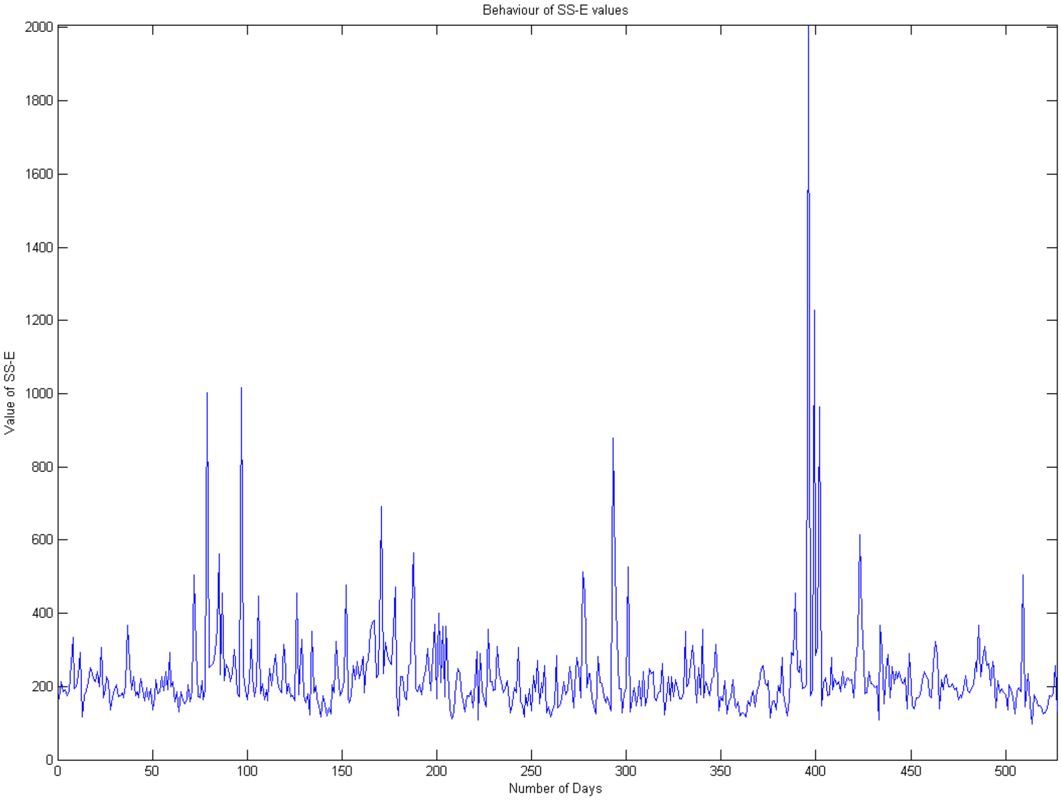
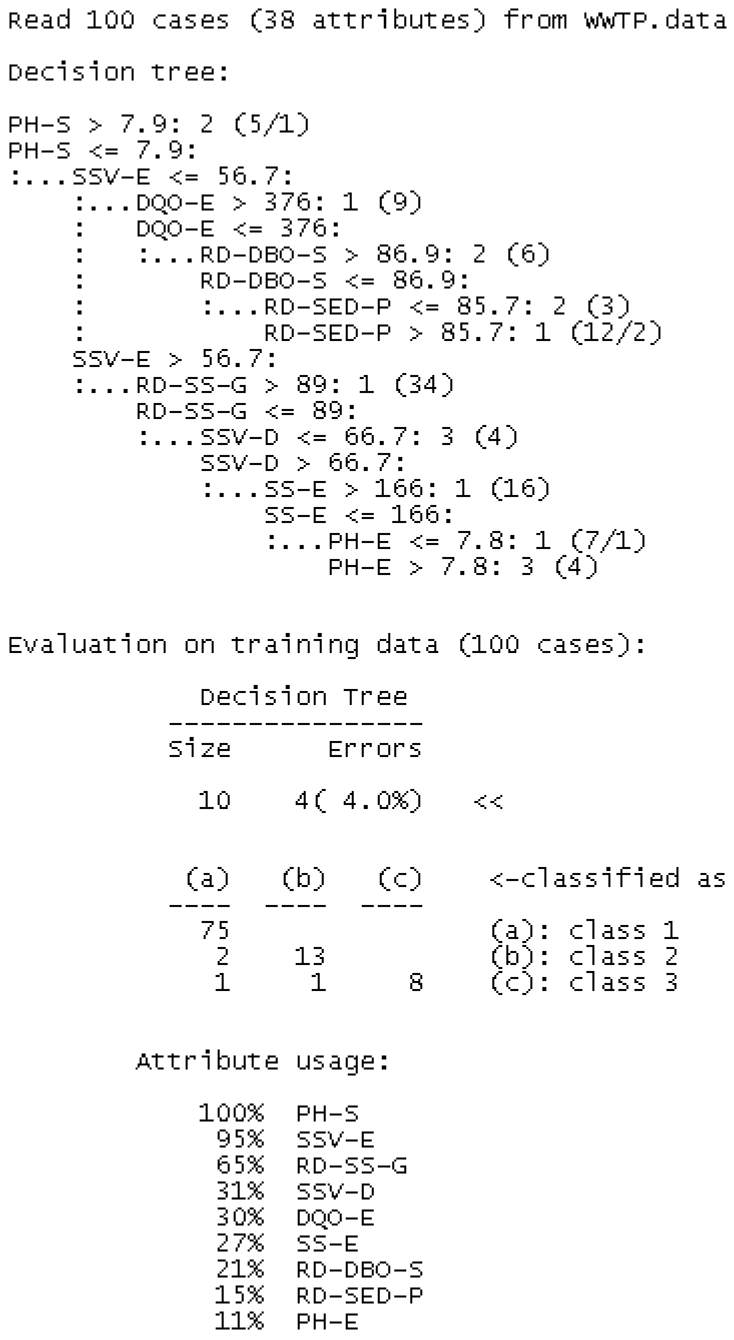


Figure 4.4: Graph showing trend of the different days values for SS-E



```
Read 100 cases (38 attributes) from WWTP.data

Decision tree:

PH-S > 7.9: 2 (5/1)
PH-S <= 7.9:
:...SSV-E <= 56.7:
    :...DQO-E > 376: 1 (9)
    :   DQO-E <= 376:
    :   :...RD-DBO-S > 86.9: 2 (6)
    :       RD-DBO-S <= 86.9:
    :       :...RD-SED-P <= 85.7: 2 (3)
    :           RD-SED-P > 85.7: 1 (12/2)
    SSV-E > 56.7:
    :...RD-SS-G > 89: 1 (34)
        RD-SS-G <= 89:
        :...SSV-D <= 66.7: 3 (4)
            SSV-D > 66.7:
            :...SS-E > 166: 1 (16)
                SS-E <= 166:
                :...PH-E <= 7.8: 1 (7/1)
                    PH-E > 7.8: 3 (4)


Evaluation on training data (100 cases):

	    Decision Tree
	  ----------------
	  Size      Errors

	    10    4( 4.0%)   <<


	   (a)   (b)   (c)    <-classified as
	  ----  ----  ----
	    75                (a): class 1
	     2    13          (b): class 2
	     1     1     8    (c): class 3


	Attribute usage:

	    100%  PH-S
	     95%  SSV-E
	     65%  RD-SS-G
	     31%  SSV-D
	     30%  DQO-E
	     27%  SS-E
	     21%  RD-DBO-S
	     15%  RD-SED-P
	     11%  PH-E
```

Figure 4.5: Decision Tree validating the Results obtained for the Process Parameter Estimation using ICA

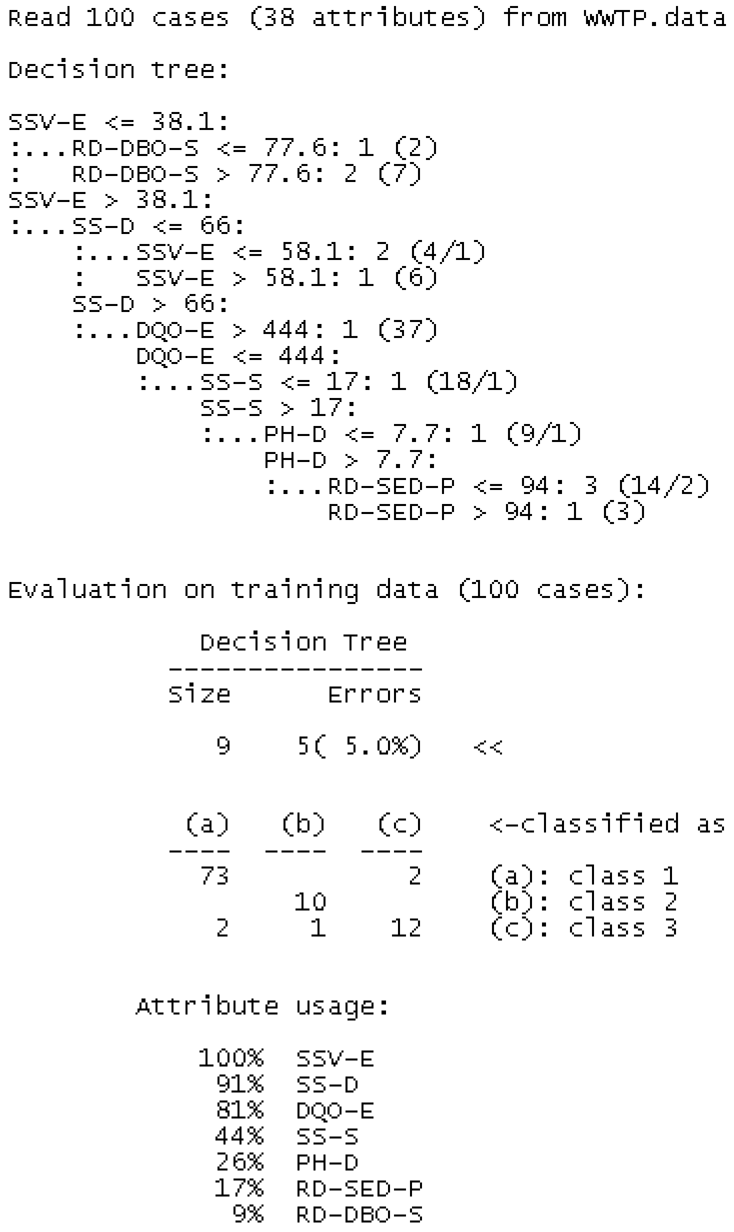
4.4.2 Results Discussion with another Decision Tree

To further verify the results obtained from the work done by ICA model another Decision Tree is constructed. For the construction of Decision Tree, the historical data is used. It is chosen with the view that it is quite different than the data set chosen for the construction of first Decision Tree so that we can have good evaluation of the results obtained by ICA model. The constructed Decision Tree is shown in Figure 4.6.

The Decision Tree indicates information about six parameters. If we do the comparison of the results obtained by the Decision Tree with the results of ICA model then we find that out of six, five parameters are fully validated and one is not validated. Thus it further consolidates the results obtained by the first Decision Tree and ability of the ICA model to estimate the process parameters. One parameter not validated by the Decision Tree is again due to the fact that ICA is applicable to non-Gaussian data set whereas this data set is real plant data representing mixture of both Gaussian and non-Gaussian behaviour. The summary of the results obtained is also given in Table 4.2.

Table 4.2: Summary of the Results Obtained by Decision Tree (Figure 4.6)

Numbers	Parameters	Values determined by ICA model	Values by See 5.0	Validated
1	SSV-E	70.8	>38.1	Yes
2	SS-D	190	>66	Yes
3	SS-S	94.75	>17	Yes
4	PH-D	8.0	>7.7	Yes
5	RD-SED-P	50.2	≤94	Yes
6	DQO-E	738.6	≤444	No



```
Read 100 cases (38 attributes) from WWTP.data

Decision tree:

SSV-E <= 38.1:
:...RD-DBO-S <= 77.6: 1 (2)
:   RD-DBO-S > 77.6: 2 (7)
SSV-E > 38.1:
:...SS-D <= 66:
    :...SSV-E <= 58.1: 2 (4/1)
    :   SSV-E > 58.1: 1 (6)
    SS-D > 66:
    :...DQO-E > 444: 1 (37)
        DQO-E <= 444:
        :...SS-S <= 17: 1 (18/1)
            SS-S > 17:
            :...PH-D <= 7.7: 1 (9/1)
                PH-D > 7.7:
                :...RD-SED-P <= 94: 3 (14/2)
                    RD-SED-P > 94: 1 (3)


Evaluation on training data (100 cases):

            Decision Tree
          ----------------
          Size      Errors

             9    5( 5.0%)   <<


           (a)   (b)   (c)    <-classified as
          ----  ----  ----
            73           2    (a): class 1
                  10          (b): class 2
             2     1    12    (c): class 3


        Attribute usage:

            100%  SSV-E
             91%  SS-D
             81%  DQO-E
             44%  SS-S
             26%  PH-D
             17%  RD-SED-P
              9%  RD-DBO-S
```

Figure 4.6: Decision Tree validating the Results obtained for the Process Parameter Estimation using ICA

4.5 Conclusions

ICA has provided good results for the monitoring of a process where the data strictly follows non-Gaussian distribution. Here we have utilized this technique for the estimation of the process parameters using historical data set. Over all the technique gave accurate results but it has limitations as well, due to its strict application to non-Gaussian data whereas the data we have used is plant operational data which is mixture of both Gaussian and non-Gaussian data. It is the prime reason for not estimating all the parameters correctly. Also for the data set just like we had in the case of PCA, here again we have the restriction that we can only calculate parameters within a certain range due to their dependence on the calculated mixing matrix from historical data. So if we are interested to find out parameters outside this range then this technique is not applicable. So the technique has its importance if we are interested in estimation of parameters within a specific range, for which we can calculate the mixing matrix.

Chapter # 5

Improved Process Monitoring using Hybrid Independent Component Analysis (HICA)

There has been continuous research on advancement of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) for process monitoring in recent years. PCA gives good results where the data follows Gaussian distribution but in most of the industrial data, the operational/historical data usually do not strictly follow Gaussian distribution. For performing analysis of non-Gaussian data, Independent Component Analysis has shown good results but its results are not impressive when the data follows Gaussian distribution. Thus both Principal Component Analysis and Independent Component Analysis have their limitations in comprehensive fault detection.

In this chapter a Hybrid Independent Component Analysis (HICA) technique is proposed and applied which overcomes the independent limitations of both techniques. In order to validate the superiority of HICA over both ICA and PCA we have made comparison of the results by using each of the techniques. Conventional PCA is applied to a waste water treatment plant (WWTP) data and Air pollution data for fault detection; similarly ICA is applied to the same (original) data set to get the faults in these two data sets. Finally the data set is analysed using HICA. The comparison of the faults detection by HICA, ICA and PCA on both data sets are reported and discussed in this chapter. The results obtained by HICA were better than PCA and ICA. It is supported by the fact that for two case studies of waste water treatment plant data and air pollution data, PCA detected 74% and 67% faults, ICA detected 61.3% and 62 % and HICA detected 90% and 80 % respectively. Another aspect of HICA technique is that we use one technique to find out faults rather than using two different techniques to find out all the faults in a historical data set.

5.1 Introduction

In recent times a lot of work has been done in terms of automation and due to increase of computer automation in the chemical plants there are thousands of data points being generated from the process and all of these are being recorded. For univariate data sets most of the faults are identified and rectified at once by the plant operators. Currently as chemical processes become more and more complex with more recycling streams going into the system due to economical and environmental constraints, many more multivariate data sets are generated which need to be analysed in order to identify any faults in the process. It is also important to analyse the process so that safer and efficient process operations can be established. This leads to the need for the analysis of historical data as well. Analysis of the historical data also leads us to operations which are more efficient in terms of reliability; productivity and safety. Data analysis and fault detection and diagnostics are applicable and important for every industry as every industry is generating data sets and their evaluation can always lead to better understanding of the process to rectify the process problems.

PCA gives good results with data following Gaussian distribution and ICA gives good results for data following non-Gaussian distribution. As real life data is mixture of both Gaussian and non-Gaussian data there is need of a technique which can cover these aspects of data set. This work is an attempt to address these issues of the data set. So it is based on the analysis done on PCA algorithm and then merging part of the PCA with ICA algorithm to get better monitoring results. As the Eigen values and Eigen vectors account for the Gaussian behaviour of the data set so this work utilizes these Eigen values and Eigen vectors to account for the Gaussian part of the data and Independent Components to account for the non-Gaussian part of the data. This chapter gives a comparative study between existing techniques of PCA and ICA and their comparison with a new developed technique which is hybrid of these techniques. This is done by application of all of these techniques to a waste water treatment plant data and air pollution data set and comparison of the results. The rest of the chapter is arranged as follows.

The calculation of HICA technique is explained with an example in section 5.2. Theoretical discussion about the behaviour of PCA, ICA and HICA for the data set is given in section 5.3. In section 5.4 all the techniques are applied to two case studies of Waste Water Treatment Plant (WWTP) and Air Pollution data sets to find out faults and a comparison is performed. In section 5.5 utilization of HICA technique for the estimation of process parameters is carried out. Finally in section 5.6 conclusions are given for the results obtained by utilization of these techniques.

5.2 HICA Calculations - An Example

As in this chapter we have made comparison of PCA, ICA and HICA so to explain calculation of HICA we take a small data set. The calculation of PCA and ICA has already been explained in Chapter 2. Here for the calculation explanation of HICA we use the same data set as was used for the explanation of PCA and ICA in section 2.2.2 and is represented in

Table 2.1. As HICA is a combination of PCA and ICA we use both of them partly. The equation for HICA is given as below (further explained in section 5.3).

$$H= (WX) (EV) \tag{5.1}$$

Where "WX" represent Independent Components (Calculated from ICA in section 4.3) and "EV" represents the Eigen vectors (Calculated from PCA in section 2.2). Here we utilize both of them and using eq. 5.1 the resultant values of HICA corresponding to each signal are shown in Table 5.1.

Table 5.1: Calculated values of HICA for small data set

Observation Number	A	B	C	D
1	-4.1416	0.0345	1.0153	6.8864
2	-0.3084	0.0023	1.1969	2.6373
3	-1.9653	0.1567	1.0275	3.6571
4	-1.5284	-0.0194	0.7853	2.8858
5	-3.3971	-1.2676	0.2852	2.4967
6	-1.562	0.4277	1.1374	3.7016
7	-1.4706	-0.4923	-0.0259	1.2317
8	-2.9101	-0.9445	1.3984	1.6584
9	-2.0931	-1.1966	0.4734	2.0792
10	-0.7087	-0.8171	0.8283	2.747

As it is combination of PCA and ICA so just like PCA and ICA the graph between number of HICA and values of HICA can give us valuable information so a graph is plotted which is shown in Figure 5.1.

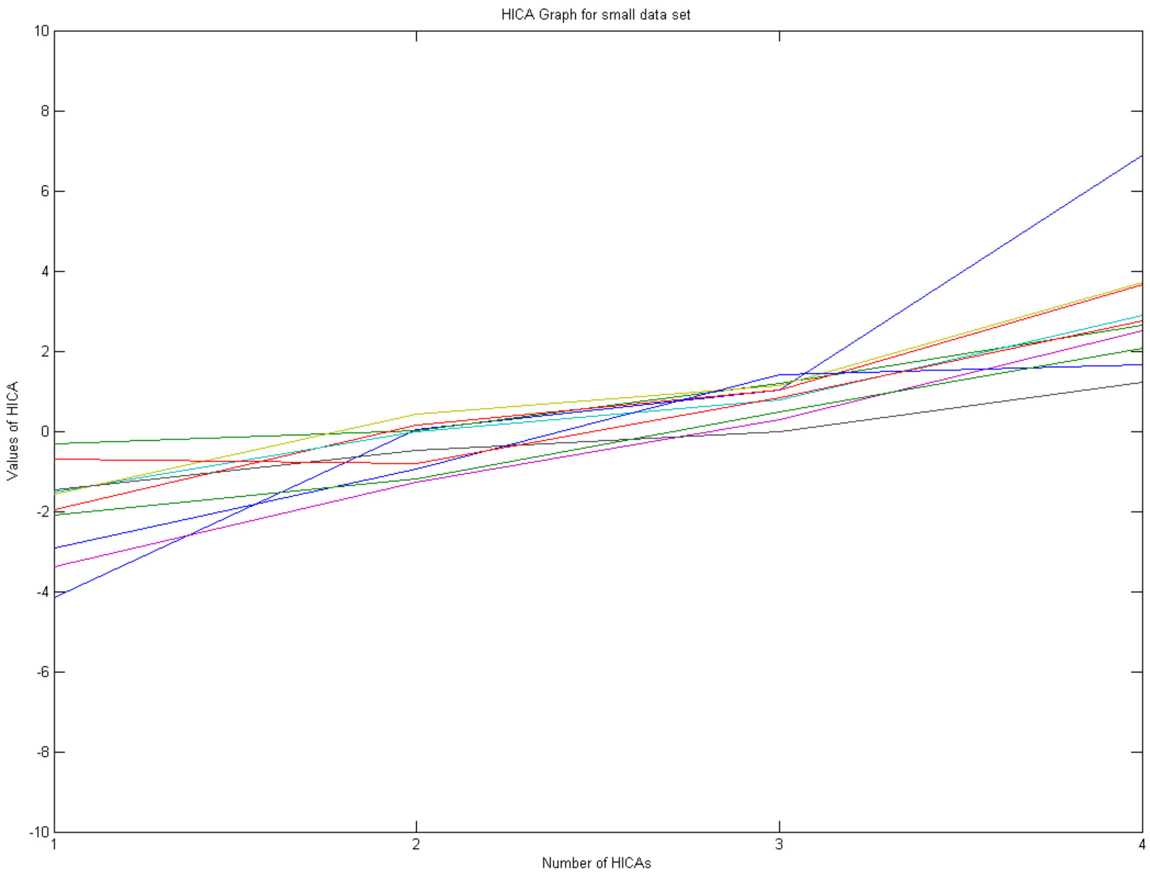


Figure 5.1: Graph between the values of HICAs and Number of HICAs

From the Figure 5.1 we find that HICA 1 and HICA 4 have got values quite different than other HICAs as HICA 1 and HICA 4 have values -4.1416 and 6.8864 respectively. Again in this case just like PC1 and IC1 (As discussed in section 2.2.2) we have HICA1 value different than others indicating problem which shows that it has ability to find hidden factors in the data set. We can evaluate the abnormal values of HICAs and find out the reasons behind these abnormal values i.e. the variables responsible for these abnormal values. This is done in the case of WWTP data and Air pollution data sets.

5.3 Theoretical Description and Comparison of Techniques

In this section we analyse the techniques of PCA, ICA and HICA theoretically. A comparison of the behaviour shown by each technique for different data sets is also done. Before we discuss behaviour of the different monitoring techniques, we can summarize different properties of these techniques in Table 5.2.

As we see from the Table 5.2 that PCA can explain the amount of variance present in the data set so the first few components explain most of the variance of the data set. It also becomes important when we want to reduce the dimension of the data set. As compared to PCA both ICA and HICA do not have ability to explain the variance of the data set so we cannot use these to reduce the dimensionality of the data set. PCA can explain the Gaussian aspect of the data set but ICA does not address the Gaussian part of the data set whereas HICA also has the advantage that it explains the Gaussian part of the data set. For the non-Gaussian part of the data set ICA gives good results but PCA cannot explain this part of the data. HICA does consider the non-Gaussian part of the data set as well. Finally the order of the components is important only in the case of PCA as compared to ICA and HICA. We have retained all the ICs and HICAs because we are making comparison of techniques for results in fault detection. This aspect of techniques along with all others i.e. variance, Gaussian and non-Gaussian distribution are discussed in more details in the coming sections.

Table 5.2: Comparison of Different techniques

Property	**PCA**	**ICA**	**HICA**
Variance of Data	Explained	Not explained	Not explained
Gaussian Data	Explained	Not explained	Explained
Non-Gaussian Data	Not explained	Explained	Explained
Order of Components	Important	Not important	Not important

5.3.1 Variance of Data Set

a) Variance explanation with PCA

As it is already discussed unlike ICs in Principal Components, PC1 is more important than PC2 and PC2 is more important than PC3 and so on. It is because in the determination of PCs, Eigen values are arranged in order of their relative magnitude. This implies that the highest Eigen value will make the first principal component, then second and so on. Also the sum of the total variance in the original variables will be equal to the sum of the Eigen values of the covariance matrix Refaat (2007) so with this discussion we can easily say that with principal components we can explain large variance about the data set so after few Principal Components we have noise in the data set. Once the variance explained by the Principal

Components is obtained then we can easily find out the percentage of total variability explained by each principal component. It can be represented by a Scree plot as follows.

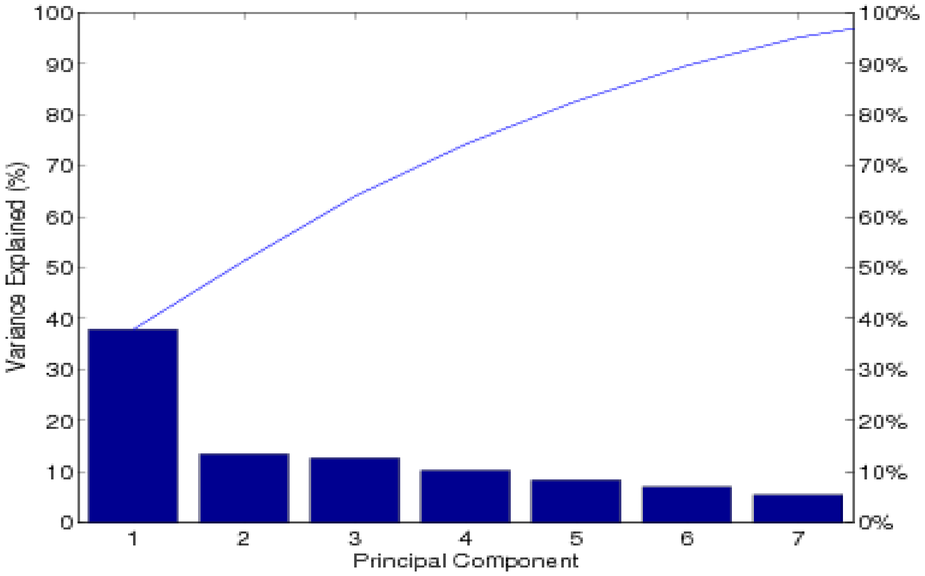


Figure 5.2: Variance of each Principal Component showing that first few have most of the variance (MATLAB 2009)

b) Variance explanation with ICA

In the principal components we can find out the variances so that the first few Principal Components describe about the variances more than the proceeding ones but in the case of ICs we are unable to find out the variances of the Independent Components. It is because in the ICA model we have both 'A' and 'S' as unknown. Now we know that independent components are random variables so to get good results we assume by convention that each independent component has unit variance .i.e. $E\{s_j^2\}=1$ so that mixing matrix 'A' will be worked out in such a way that we take into account this restriction. To further elaborate we can write the ICA model as eq. (5.2)

$$X=AS= (AM^{-1}) (MS) =AS \qquad (5.2)$$

Where 'M' is a diagonal scaling matrix such that $a_{ii}s_i$ results in a column whose elements have variance of one. Also the scaling matrix 'M' could be positive or negative and due to this reason ICA can produce answer with any sign but it is not very significant in most of the cases because it still serves the purpose of finding the latent information about the data Hyvarinen and Oja (2000).

c) Variance explanation with HICA

In ICA as both ‘S’ and ‘A’ are unknown so we cannot find the variance and by convention it is assumed to have unit variance for each of the independent components. In the case of Principal Components we can find out the variance because the Eigen values are arranged in order of their magnitudes. As in HICA we are using the values of Eigen vectors and Independent Components so the variance mainly depends on the behaviour of these Eigen vectors and Independent Components. The Independent Components do not explain any variance in the data set (already discussed in previous section). Similarly the Eigen vectors are not arranged in order of their magnitude so the final values of HICA do not represent any order of preference .i.e. HICA cannot explain the variance in the data set.

Explanation of variance is very important if we are interested in reduction of the data set. If we keep all the data set then it is not very important for the technique to explain the variance. In our case as we are retaining all the data sets after the application of PCA, ICA and HICA in order to make good comparison of these techniques so it is not affecting our results.

5.3.2 Techniques Applications to Different Data

a) Application of PCA to Gaussian data

In the calculation of principal components we find out the mean and variance of the data set. The normal distribution or Gaussian distribution is the one in which the variables tend to cluster around the mean of the data set so it is the basis of finding the Principal Components so we assume that principal components also follow Gaussian distribution. It has also been proved in literature that PCA gives good results with data following Gaussian distribution.

Mathematically if we have ‘X’ as the data matrix than if $x_1,x_2,x_3,.......,x_n$ are the data points then for this data PCA can explain more about data if $x_1,x_2,x_3,.....,x_n$ follow Gaussian distribution. It is also true because if $x_1,x_2,x_3,.......,x_n$ are not Gaussian distributed then diagonalizing a covariance matrix might also not produce good results leading to inadequate explanation of the data set. This assumption also helps in a way that covariance matrix finds out the noise and redundancies which is the prime goal of PCA (Shlens 2005).

b) Application of ICA to Non-Gaussian data

One of the most important restriction in ICA is that the Independent Components must be non-Gaussian and the columns of 'S' are statistically independent. Now for the statistical independence if we have $z_1,z_2,z_{3,}$ ……, z_m as random variables with the joint probability density function (pdf) of $P(z_1,z_2,z_3,------,z_m)$ with the assumption of zero mean then they are mutually statistically independent and we have the condition as described by eq. (5.3)

$$P(z_1,z_2,z_3, \ldots\ldots,z_m)=P_1(z_1).P_2(z_2).P_3(z_3)\ldots\ldots P_m(z_m) \tag{5.3}$$

So P_i (z_i) (where i=1,2,3, ……,m) is the marginal pdf of z_i where it is considered alone. Also the un-corelatedness shows that E $\{z_iz_j\}$ =E$\{z_i\}$. E $\{z_j\}$ for $i \neq j$ and here E {…..} is the mathematical expectation.

So if z_i and z_j are independent, for any functions f_i and f_j they must have

$$E\{f_i(z_i)f_m(z_m)\}=\iint f_i(z_i)f_m(z_m)P(z_i,z_m)dz_idz_m$$

$$=\iint f_i(z_i)f_m(z_m)P_i(z_i)P_m(z_m)dz_idz_m$$

$$= \int f_i(z_i)P_i(z_i)dz_i \int f_m(z_m)P_m(z_m)dz_m$$

$$=E\{f_i(z_i)\}E\{f_m(z_m)\}$$

So if the variables in any system are independent then they are uncorrelated as well (Albazzaz and Wang 2004). Now to further elaborate that the Gaussian variables are not permitted in ICA we assume in the model that the mixing matrix 'A' is orthogonal and the Independent Components are Gaussian. Then the observed variables i.e. x_1 and x_2 are also Gaussian, have unit variance and are uncorrelated.

Their joint density can be written by eq. (5.4)

$$p\left(x_1,x_2\right)=\frac{1}{2\pi}\exp\left(-\frac{x_1^2+x_2^2}{2}\right) \tag{5.4}$$

The joint probability distribution can be illustrated by Fig.5.3.

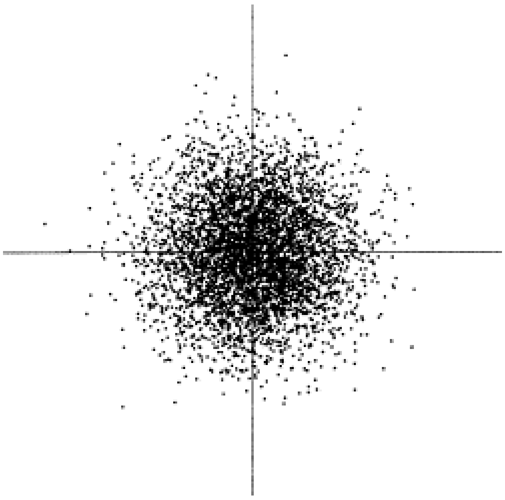


Figure 5.3: Multivariate Distribution of Two Independent Gaussian Variables (Hyvärinen et al. 2001)

From the Figure we can see that as the density is completely symmetric so we do not get any information on the directions of columns of matrix A. It can also be proved that x_1 and x_2 are independent because the distribution of any orthogonal transformation of the Gaussian (x_1,x_2) has got exactly the same distribution as (x_1,x_2). It also confirms that for Gaussian variables we can only estimate the ICA model up to an orthogonal transformation. So the mixing matrix 'A' is not identifiable for Gaussian independent components (Hyvärinen et al. 2001).

c) Application of HICA on data set

PCA gives good results for Gaussian data and ICA gives good results with non-Gaussian data. For ICA if $z_1,z_2,z_3,........,z_m$ are random variables with joint probability density function (pdf) of $P(z_1,z_2,z_3,.......,z_m)$ with zero mean assumption then they are mutually independent with the following condition.

$$P(z_1,z_2,z_3,.........,z_m)=P_1(z_1).P_2(z_2).P_3(z_3)..........P_m(z_m)$$

For PCA if x_1, x_2, x_3,......., x_m are the variables and if they follow Gaussian distribution then we get good results. As practical data is a mixture of both Gaussian and non-Gaussian behaviour, in order to get good results we combine these two by the multiplication of Eigen values (following Gaussian part of the data) and Independent Components (following non-Gaussian part of the data) we get good results using HICA with this approach.

Mathematically the HICA technique has been represented by eq. (5.1). The results are shown in the section 5.4 with two case studies and comparison of the techniques of PCA, ICA and HICA.

5.3.3 Order of Components

a) Order of Principal Components

As discussed in the variance explanation of PCA that first few PCs explain most of the variance of the data set so it also proves that the order of the Principal Components is very important. As the first few PCs can explain most of the problems in the data set so the order of the PCs is very important.

b) Order of Independent Components

As compared to PCA there is restriction in ICA that the order of the independent components cannot be defined in terms of their importance. It is due to the fact that both 'A' and 'S' are unknown. For further illustration we can introduce a permutation matrix P and its inverse in the model which can be represented by eq. (5.5)

$$X=AS= (AP^{-1}) (PS) =AS \tag{5.5}$$

Here the 'PS' have the original independent variables but they are in a different order. Also AP^{-1} is an unknown mixing matrix and is solved by ICA algorithm (Hyvarinen and Oja 2000).

c) Order of HICA

Although in PCA the order is quite important but for HICA we combine the ICs of the data set with the Eigen vectors. So they are not arranged in order of their importance. In other words the reason for HICA of not having importance of their order is that the Eigen vectors are not arranged in any specific order .i.e. in terms of their increasing or decreasing values. It is also important to note that the order of HICAs do not make much difference because if we find the faults in the system keeping all the values of HICAs then multiple values will lead to the same fault. This aspect of HICA is further elaborated in the result section with two case studies. In the first case study results are presented for WWTP and in second case study results are presented for Air pollution data.

5.4 Case Studies, Results and Discussion

Two case studies have been utilized here for the validation and comparison of all the techniques .i.e. PCA, ICA and HICA. The first case study is a waste water treatment plant data set and the second case study is an air pollution data set. Application of the methodologies and the results obtained are discussed as under.

5.4.1 Application on WWTP data

The technique of PCA, ICA and HICA are applied to a waste water treatment plant (WWTP) data set. The details of the data set with statistical evaluation and operational setup of the plant are discussed in section 2.3 of literature review. In the following sections the PCA, ICA and HICA techniques are applied to WWTP data and the results of the work done are given and discussed.

a) Results with Principal Component Analysis

The technique of PCA and identification of faulty days have already been explained in section 3.3.1 of chapter 3 with Figure 3.2. Using PCA, a total of 23 faulty days have been identified which are validated from the literature as well Albazzaz and Wang (2006). Once all the faulty days have been identified then we can further investigate the reason for the abnormal values of Principal Components .e.g. the value of PC2 on 15/03/1990 is investigated and the contribution plots are drawn using a commercial software (SIMCA-P) and based on the plot we can easily find out that RD-SS-G, RD-DQO-G and SED-S are the contributing factors for this faulty day which is shown in Figure 5.4. This can also be verified from the experts working on the chemical plant. For this particular case of 15/03/1990 we can validate it from the literature (Sanchez et al. 1997) that 15/03/1990 was an abnormal day which is clearly out of limit due to the problems with Secondary Treatment Plant.

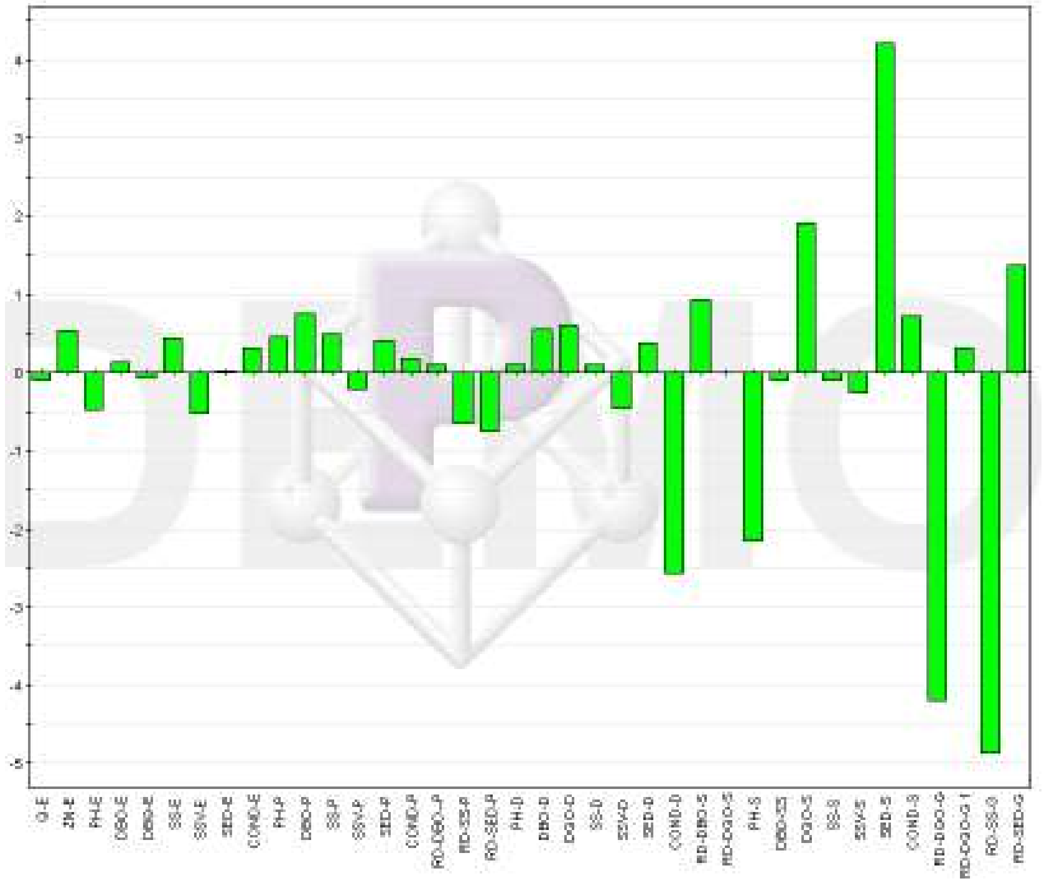


Figure 5.4: The Contribution plot indicating RD-SS-G, RD-DQO-G and SED-S as main factors for faulty days

In this work all the datasets obtained after having done PCA are maintained because of comparison with the results obtained from ICA and HICA as all data points for them are also maintained otherwise the first few PCs (almost 10 give more than 90% variance of the data as shown in Figure 5.5) can explain most of the variance in the data set.

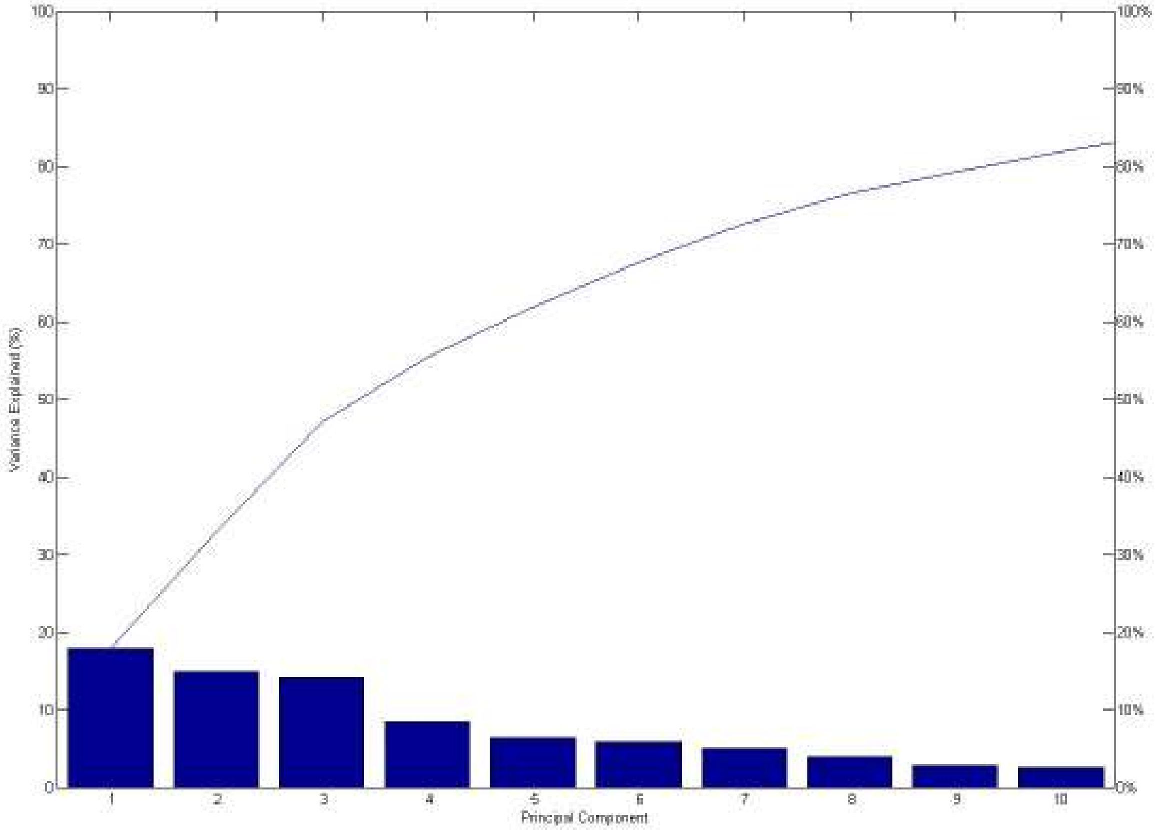


Figure 5.5: First 10 PCs explaining most variance of the WWTP data set

b) Results with Independent Component Analysis

The technique of ICA is applied to the WWTP data set and has been discussed in section 4.4.1 of chapter 4. From chapter 4 we concluded that there were 19 days identified as faulty by ICA. Once it is decided that a day is having fault then contribution plots of the Independent Components on that day can be made. Based on the contribution plots we can find out the responsible parameters for abnormality at that particular day. e.g. Contribution plot for 18/07/91 has been shown in Figure 5.6 which clearly indicates that this day was faulty due to abnormal values of SED-S.

Thus after carrying out the Independent Component Analysis on the data we got 19 days as faulty days. If we do comparison of ICA with PCA then it determined most of the days identified by PCA but missed some days as well. Also it identified some of the days which were missed by PCA.

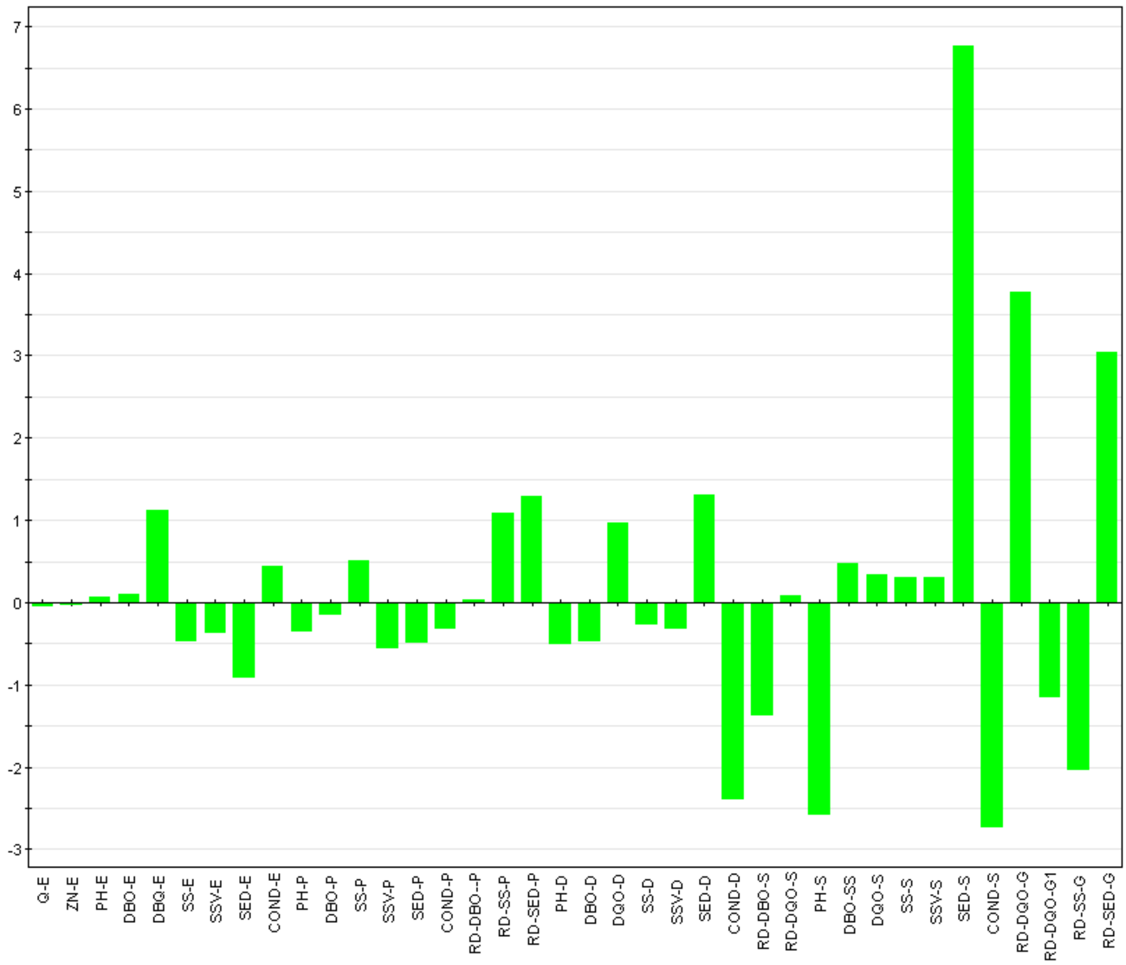


Figure 5.6: The Contribution plot for 18/07/91 indicating SED-S as main factor for faulty day

c) Results with HICA

The technique is applied to the same data set and as it was done just like PCA and ICA, a graph is plotted between the number of HICA and values of HICA. Also in this case all the values are retained. From Figure 5.7 it is quite evident that there are a lot of points which have quite different values. These values are noted and the corresponding day is found to be a faulty day. Again as in case of ICA here different violating values correspond to the same day. e.g. for 17/07/1991 HICA 11 has value of 10.3 and HICA 19 has a value of 10.74, similarly for 13/03/1990 HICA 11 has a value of 11.05, HICA 19 has a value of 8.0 and HICA 27 has a value of 14.09, all these abnormal values correspond to the same faulty day which further strengthen the ability of this technique to determine the faulty day. Also when the faulty days have been identified then analysis of the day can be carried out with the help of the experts working there or using some software and making contribution plots of the factors contributing to the abnormality on that day.

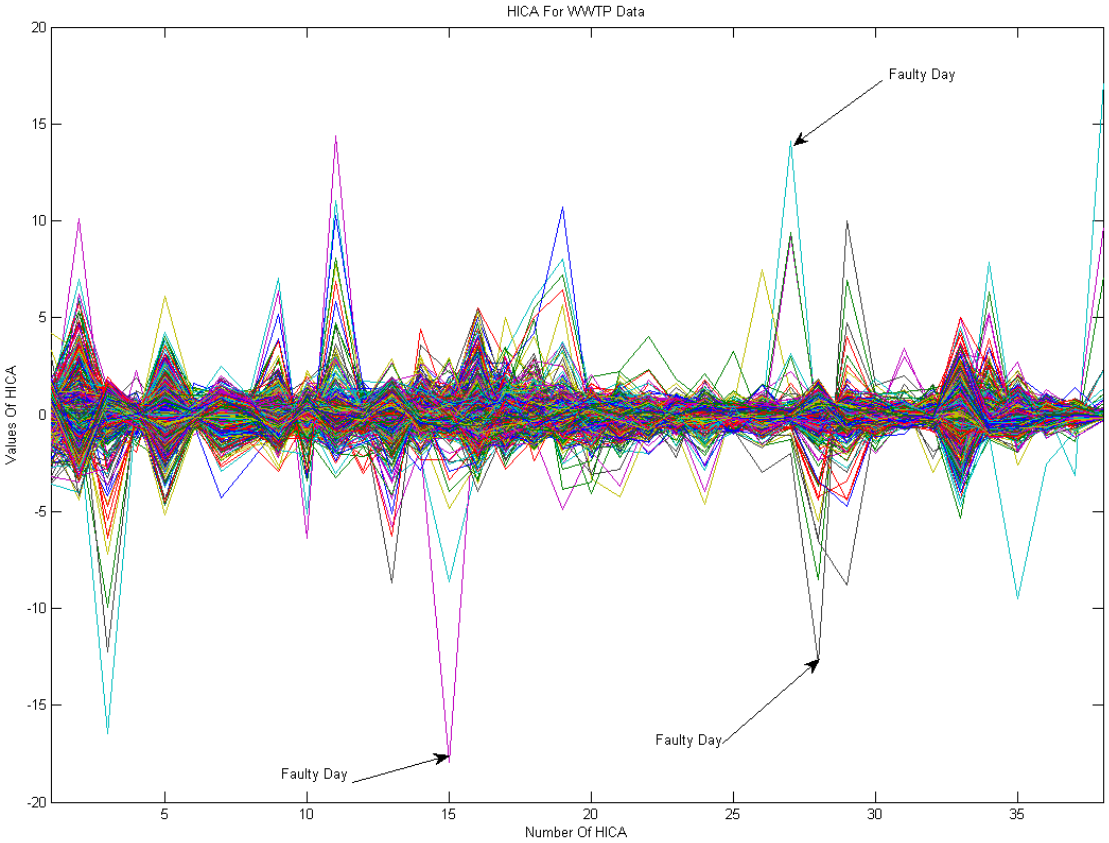


Figure 5.7: HICA showing faulty days in the data set

Here contribution plot is constructed as shown in Figure 5.8. The contribution plot clearly indicates that the abnormality on 18/07/91 is due to violating values of DBO-SS and PH-S.

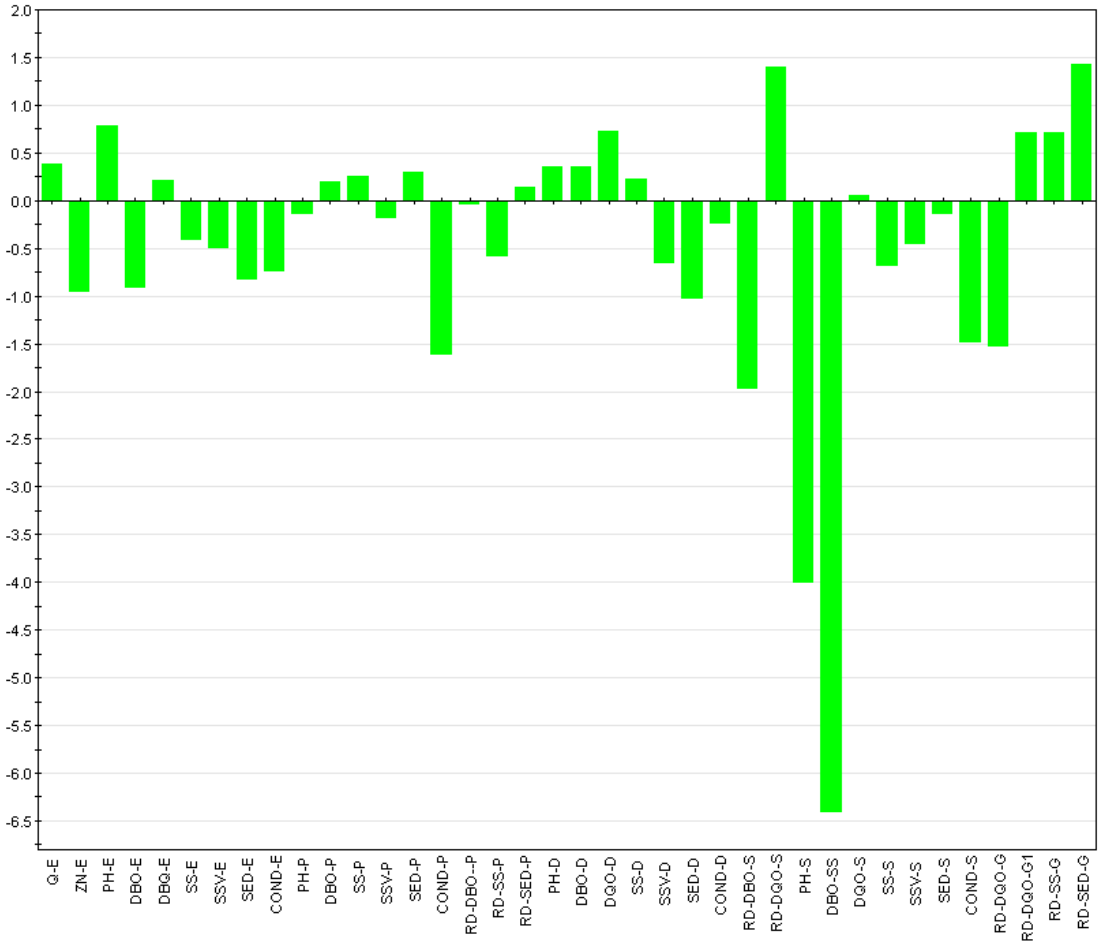


Figure 5.8: The Contribution plot for 8/07/91 indicating DBO-SS and PH-S as main factors for faulty day

d) Comparison of the Results:

The results obtained from all the above techniques are presented in Table 5.3 for comparison and it is quite evident from the table that in the case of PCA it missed some of the faulty days which were detected by ICA and similarly ICA detected faulty days but missed some days which were detected by PCA and when HICA is applied then it was successful in detecting almost all the days which were detected earlier by both PCA and ICA. The days which HICA could not identify were 16/03/1990, 02/12/1990 and 09/07/1991 which were detected by PCA and ICA separately but none of the day was detected by both techniques. Also HICA detected 24/07/1990, 13/09/1990 and 25/01/1991 as faulty days which were missed by both PCA and ICA but if we analyse PCA and ICA then these techniques had them as nearly abnormal days (almost faulty days) whereas HICA has detected them earlier which shows its superiority on both the techniques. Overall PCA detects 23 faulty days, ICA detects 19 faulty days and HICA detects 28 faulty days so if we see the ability of all the techniques then PCA detected almost 74% of the total days, ICA detected 61% of the faulty days and HICA detected 90% of the faulty days. Also for both PCA and ICA together they captured about 84% of the faulty days whereas HICA alone detected 90% of the faulty days which is a good result by HICA.

Table 5.3: Comparison showing faulty days identified by different techniques

Identification of Faulty Days by different Techniques			
Date	**PCA**	**ICA**	**HICA**
13/03/1990	Yes	Yes	Yes
14/03/1990	Yes	Yes	Yes
15/03/1990	Yes	Yes	Yes
16/03/1990	Yes	No	No
27/04/1990	Yes	No	Yes
29/04/1990	Yes	Yes	Yes
5/06/1990	Yes	Yes	Yes
26/06/1990	Yes	Yes	Yes
24/07/1990	No	No	Yes
25/07/1990	Yes	Yes	Yes
12/08/1990	Yes	No	Yes
13/09/1990	No	No	Yes
14/09/1990	Yes	Yes	Yes
3/10/1990	Yes	Yes	Yes
22/10/1990	Yes	No	Yes
16/11/1990	No	No	Yes
2/12/1990	Yes	No	No
16/01/1991	No	No	Yes
25/01/1991	No	No	Yes
29/01/1991	Yes	Yes	Yes
31/01/1991	Yes	Yes	Yes
10/02/1991	No	Yes	Yes
29/04/1991	No	Yes	Yes
24/05/1991	Yes	Yes	Yes
28/05/1991	Yes	Yes	Yes
31/05/1991	Yes	No	Yes
8/07/1991	Yes	Yes	Yes
9/07/1991	No	Yes	No
17/07/1991	Yes	Yes	Yes
18/07/1991	Yes	Yes	Yes
19/07/1991	Yes	No	Yes
Faulty Days Identified	***23***	***19***	***28***
%age	***74***	***61.3***	***90***
Total Faulty days Identified	***31***		

5.4.2 Application on Air Pollution Data Set

The techniques of PCA, ICA and HICA are applied to the Air pollution data set. The objective of the study of this data set is to make another comparison of the techniques of PCA, ICA and HICA. Based on the values of the PCs, ICs and HICAs we establish which observation in this air pollution data set is representing fault. Once we have obtained all of the faulty observations by utilization of each of the techniques, they are then compared with each other. Also in order to further identify the reason behind any of the faulty observation, contribution plots are constructed. The application of PCA, ICA and HICA and their results are discussed as follows.

a) Results with Principal Component Analysis

Principal Component Analysis has been applied to the air pollution data set. The result of the Principal Component Analysis is shown in Figure 5.9.

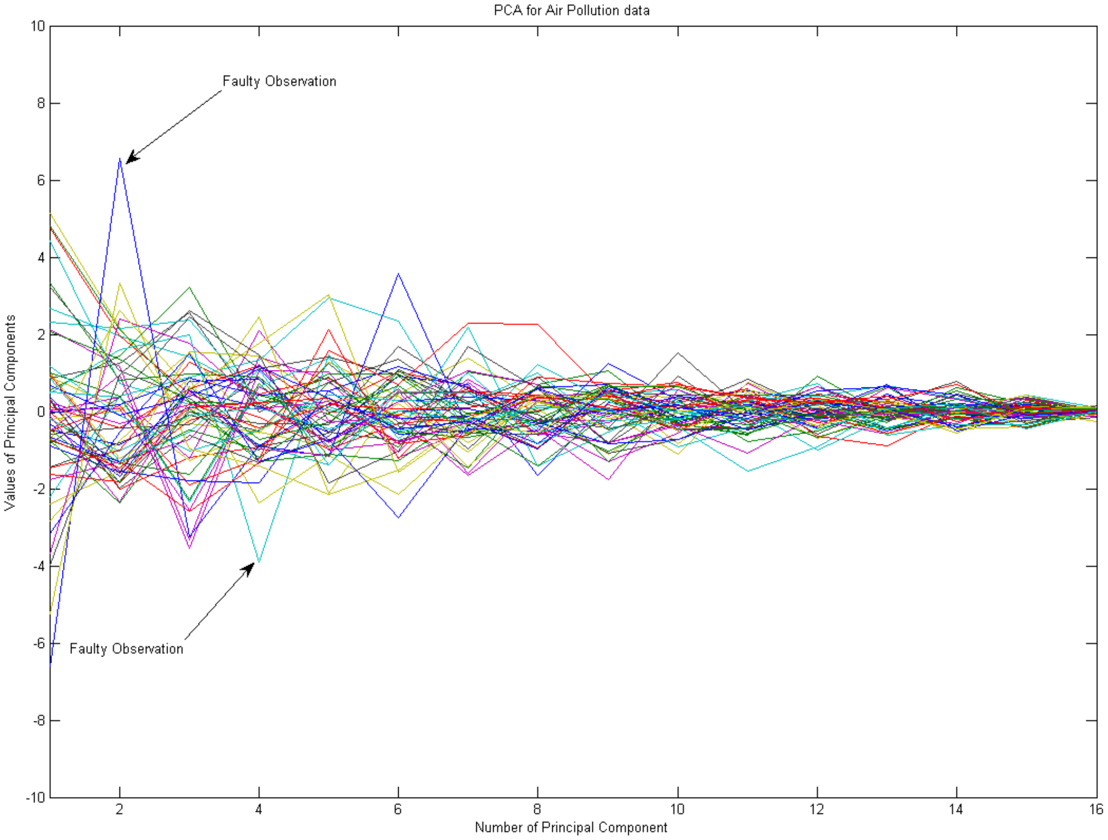


Figure 5.9: PCA in the Pollution data indicating Faulty Observations

From Figure 5.9 it is quite evident that first few PCs describe most of the faulty observations. Just like the WWTP data we identify the faulty observations based on the abnormal values of the calculated PCs. Based on the work done by PCA total 14 observations are identified as abnormal observations. Once abnormal observations are identified then the contribution plots can be constructed for the identification of the factors behind this abnormality. Now in this case contribution plot for observation # 29 is constructed and it was observed that “HC” and “NOX” were the main factors for this abnormal value of PC.

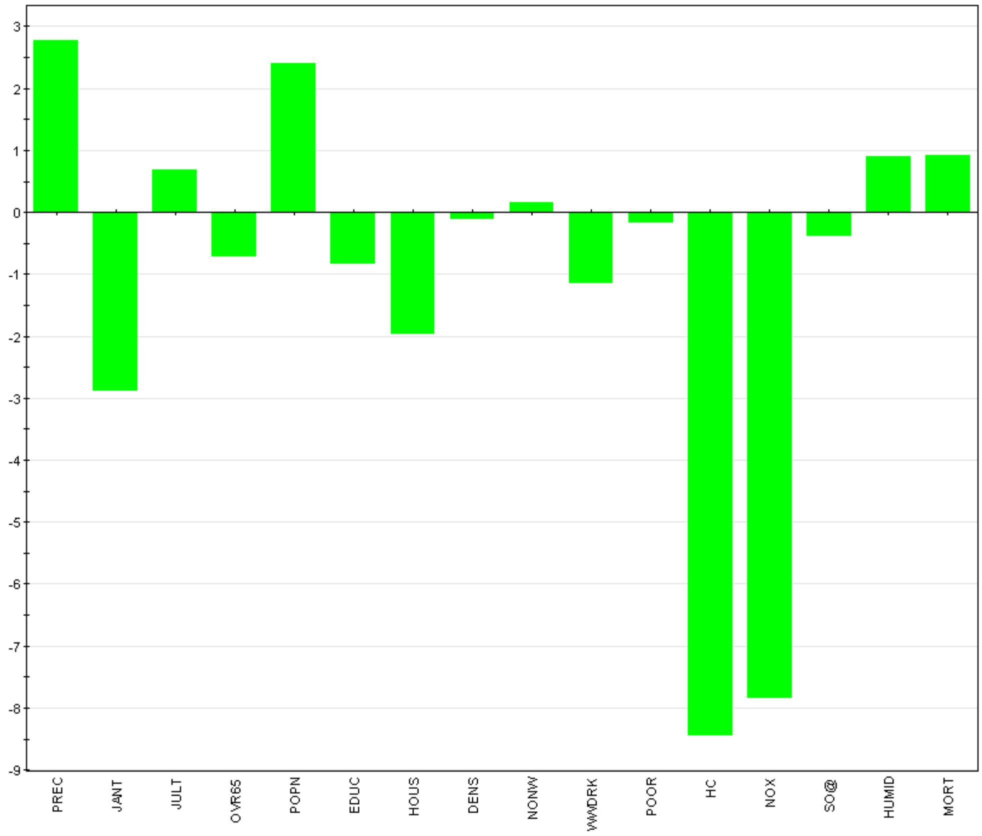


Figure 5.10: The Contribution plot for Observation # 29 indicating HC and NOX as main factors

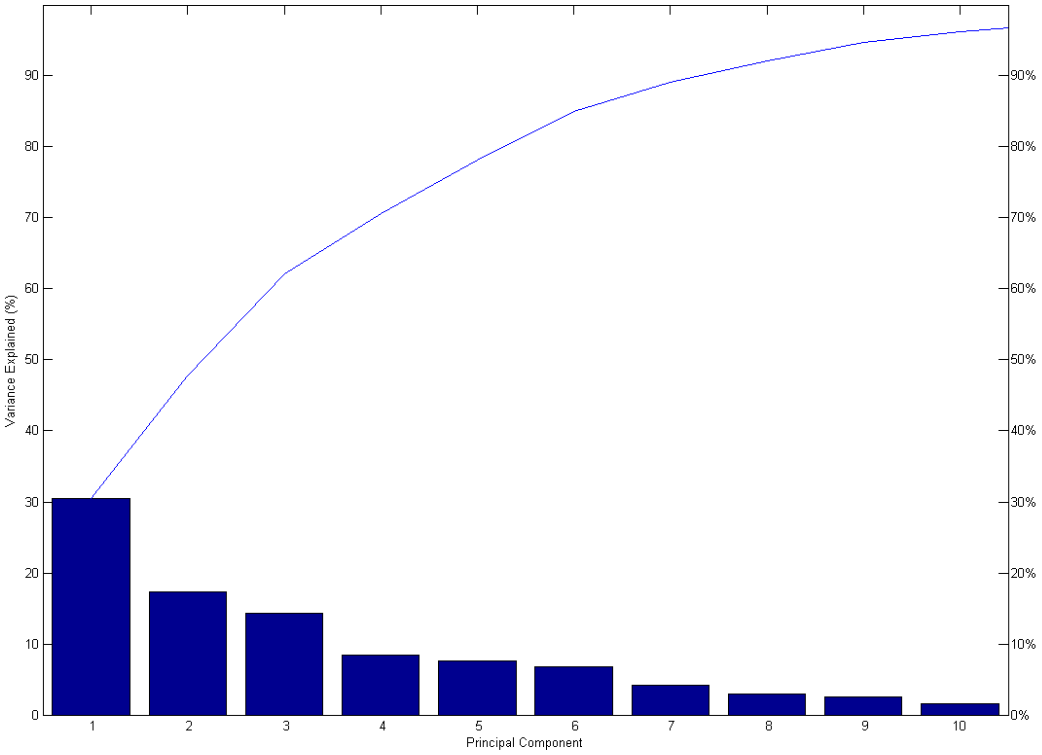


Figure 5.11: First few PCs explaining most variance of the Air Pollution data set

b) Results with Independent Component Analysis

Independent Component Analysis (ICA) is applied on the same data set (Original data set of the Air pollution) and the results obtained are indicated in Figure 5.12.

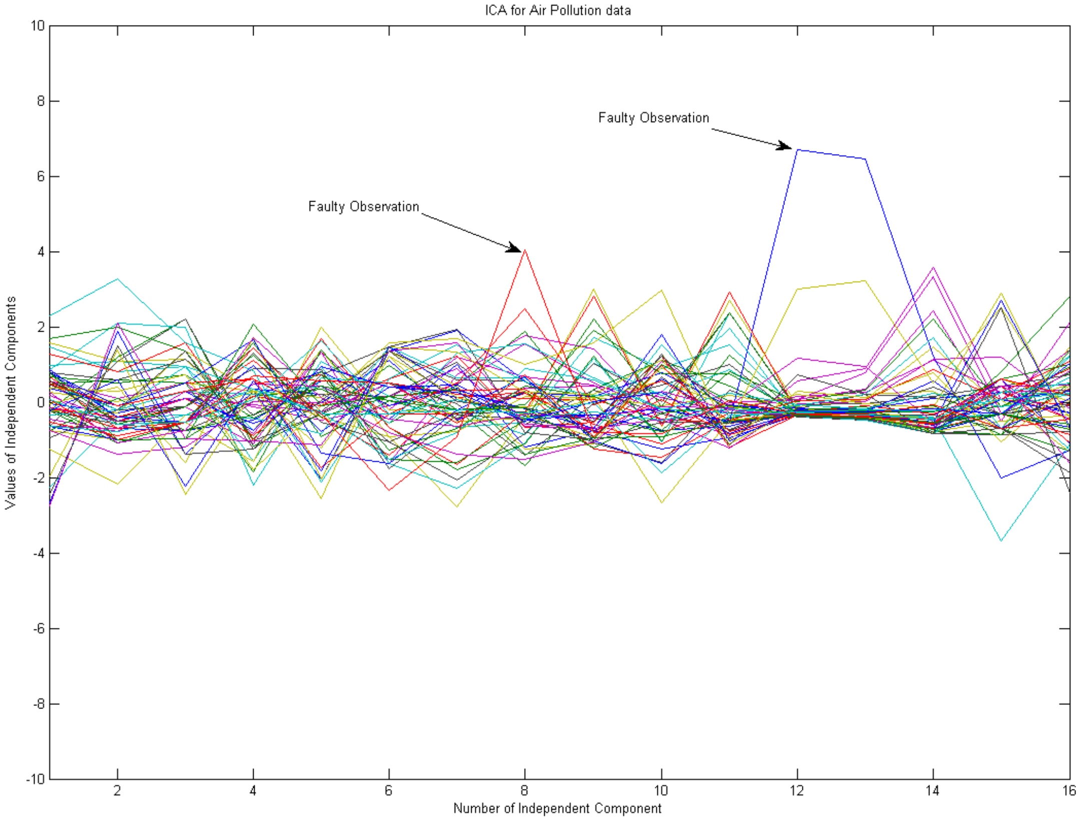


Figure 5.12: ICA in the Pollution data indicating Faulty Observations

It is shown in Figure 5.12 that quite different values of ICs indicate the abnormal observation. Based on the values of ICs, 13 abnormal observations are identified. Again after the determination of abnormal observation we can plot the contribution plots of that IC. The contribution plot for the IC of Observation # 37 have been plotted in Figure 5.13 indicating different factors including MORT, NOX, POOR, HC etc. for this abnormal observation.

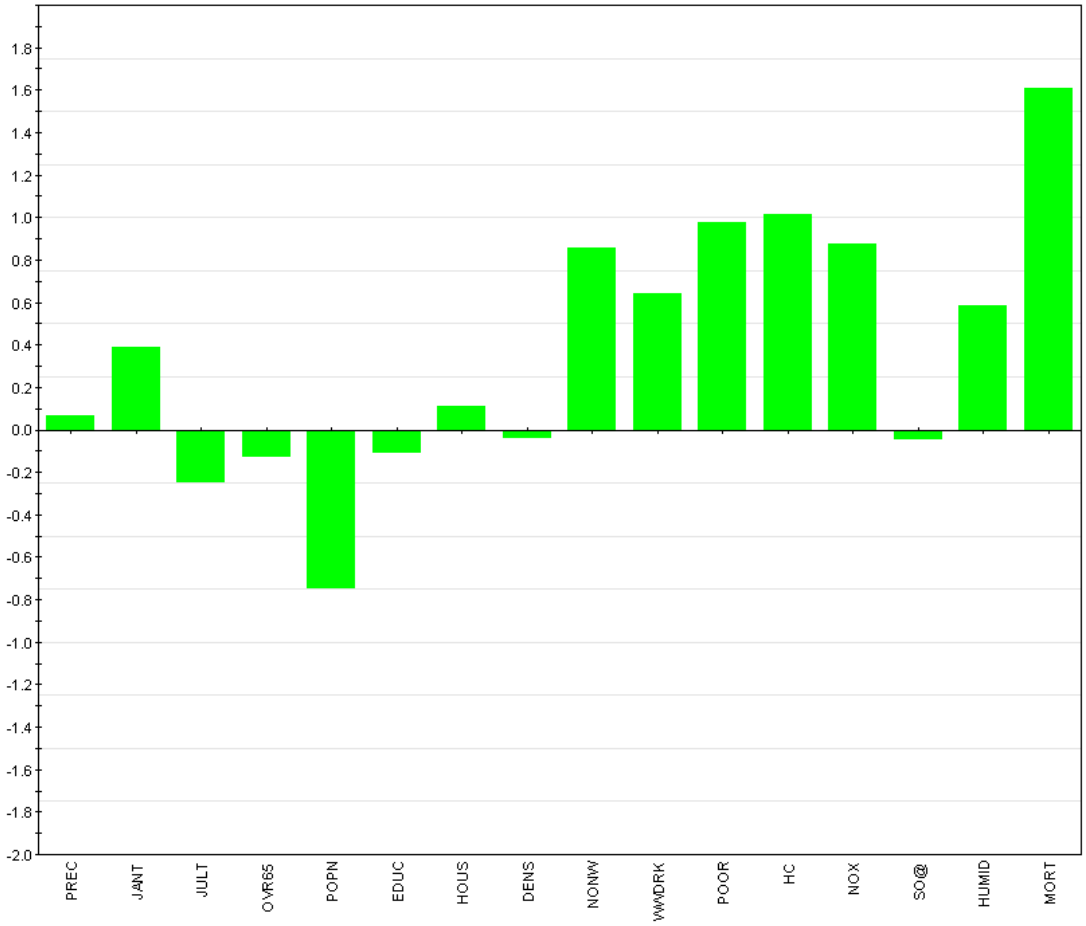


Figure 5.13: Contribution plot for Observation #37 indicating different factors for Fault

c) Results with HICA

HICA technique is also applied to the same data set (Original Air Pollution data set) and the results are indicated in Figure 5.14.

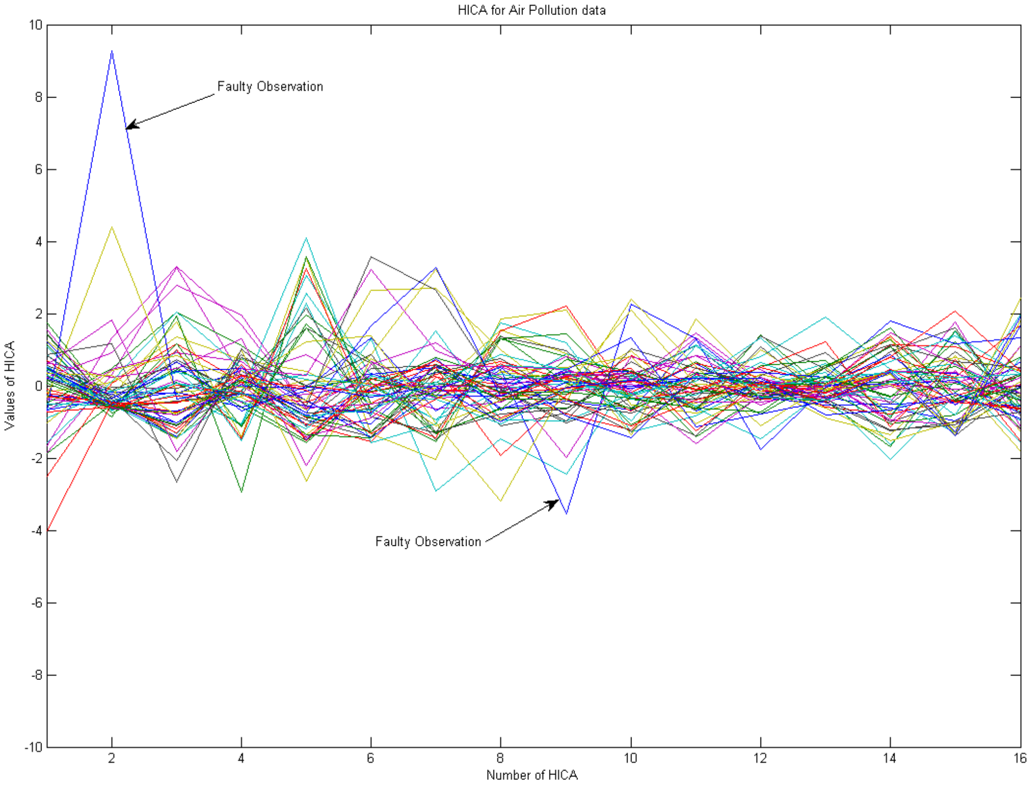


Figure 5.14: HICA in the Pollution data indicating Faulty Observations

As shown in Figure 5.14 we can easily identify the faulty observations based on the values of HICA. Utilizing HICA values total 17 observations are identified as abnormal/faulty. Contribution plot has been constructed for observation # 59 to find out the factors behind its abnormality. The contribution plot is shown in Figure 5.15 for observation #59 indicating “PREC” and “HUMID” as reasons behind its abnormality.

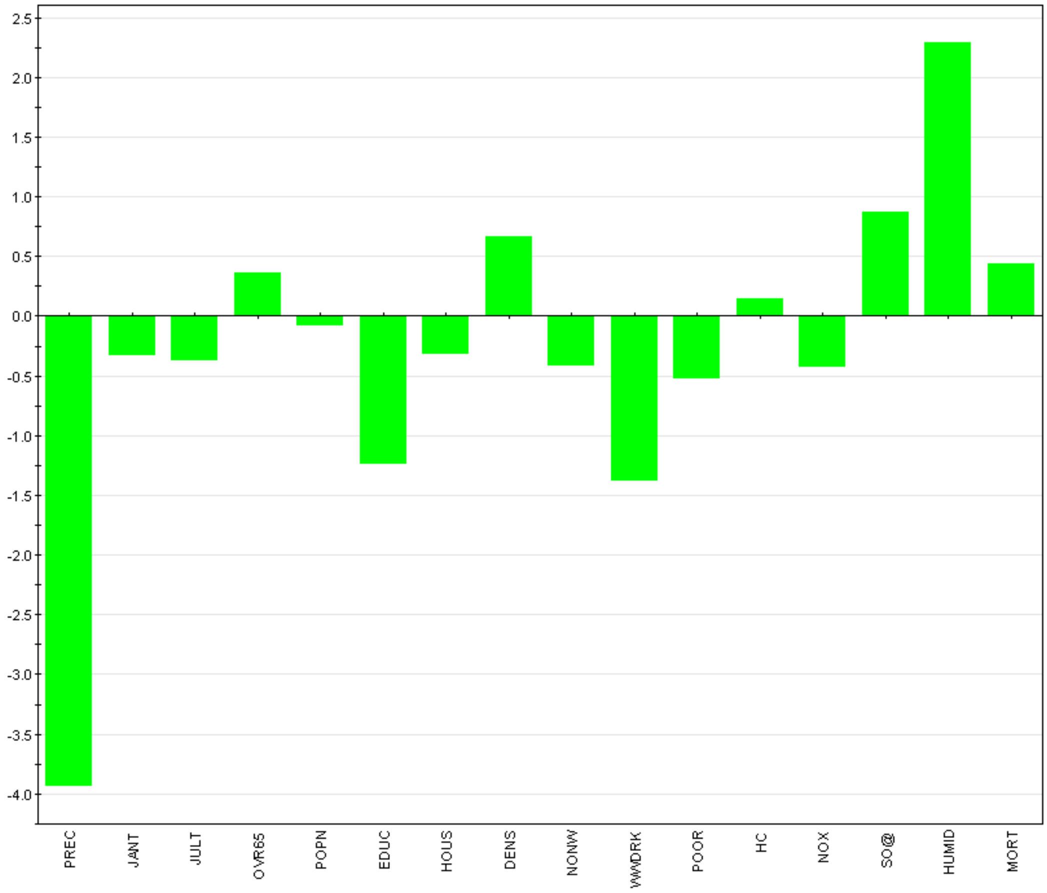


Figure 5.15: Contribution plot for Observation #59 indicating PREC and HUMID as main factors

Table 5.4: Comparison showing faulty Observations identified by different techniques

Identification of Faulty Observations by different Techniques			
Observation Number	**PCA**	**ICA**	**HICA**
5	No	No	Yes
6	Yes	Yes	Yes
11	Yes	No	No
12	No	Yes	Yes
16	Yes	No	No
18	No	Yes	Yes
23	Yes	No	No
25	No	No	Yes
29	Yes	Yes	Yes
31	Yes	Yes	Yes
32	Yes	Yes	Yes
35	Yes	No	No
37	Yes	Yes	Yes
40	No	Yes	Yes
41	No	Yes	Yes
47	Yes	Yes	Yes
48	Yes	Yes	Yes
49	Yes	No	Yes
50	Yes	No	Yes
55	Yes	Yes	Yes
59	No	Yes	Yes
Observations Identified	***14***	***13***	***17***
%age	***67***	***62***	***81***
Total Identified	***21***		

d) Comparison of Results

The techniques of PCA, ICA and HICA are applied to the air pollution data set. All the techniques calculated total of 21 observations as abnormal. PCA was successful in detection of 14 observations out of these 21 observations. Similarly ICA and HICA were successful in detection of 13 and 17 observations as abnormal observations. So these techniques detected 67%, 62% and 81% of the faults respectively. The results obtained from this case study are also quite similar to the case study results of WWTP data with each PCA, ICA and HICA missing some faults individually. Although HICA also missed faults but still it showed better results than PCA and ICA individually this indicates superiority of this technique. Also it verifies the results obtained in the case WWTP data set.

5.5 Estimation of Process Parameters using Hybrid Independent Component Analysis (HICA)

HICA has been shown to be useful technique for fault detection in section 5.4 along with its comparison with already established techniques of PCA and ICA. The utilization of PCA and ICA for the estimation of process parameters has been discussed in chapters 3 and 4. In this section estimation of process parameters using HICA is carried out. Once the parameters are estimated then their validation is carried out by Decision Tree as was done for PCA and ICA in Chapters 3 and 4.

5.5.1 Methodology used for Parameters Estimation

The methodology used for the estimation of process parameters is based on HICA model. It has been described in section 5.2 of this chapter. Following are the steps involved in the estimation and validation of process parameters.

1. Identification of faults using HICA
2. Calculation of HICA values for faults determination
3. Calculation of Eigen values of fault free data
4. Calculations of ICs of data
5. Calculation of mixing matrix of data
6. Parameter estimation
7. Validation of calculated parameter.

The details of these steps are as follows.

a) Identification of faults in the system

The first step is the identification of faults in the system using HICA values. It has already been discussed in chapters 3 and 4 that for parameters estimation it is necessary to remove faults from the data set. It is because the fault free data is further used for parameters estimation and data containing faults can lead to abnormal results.

b) Calculation of HICA values for fault free data

Based on the work done in the previous step HICA values are calculated which represent the fault free data. These values will further be used for the estimation of ICs and consequently the parameters.

c) Calculation of Eigen values of fault free data

The calculation of Eigen values of fault free data is also important as these will be used to calculate ICs along with already determined values of HICAs using HICA model.

d) Calculations of Independent components

As both HICAs and Eigen values are calculated so using HICA model ICs are calculated.

e) Calculation of mixing matrix

The next step is the calculation of the mixing matrix of this fault free data set. The mixing matrix is calculated using eq. 2.9 already discussed in chapter 2.

f) Estimation of process parameters

The final step is the calculation of process parameters using eq.2.9. As the values of ICs and mixing matrix are already calculated in previous steps so using these values the parameters are calculated.

g) Validation of calculated parameters

The final step is the validation of the estimated parameters. In this work the validation is carried out by Decision Tree as was done in Chapters 3 and 4 for estimated parameters using PCA and ICA.

The whole methodology can be explained by a flow diagram by Figure 5.16.

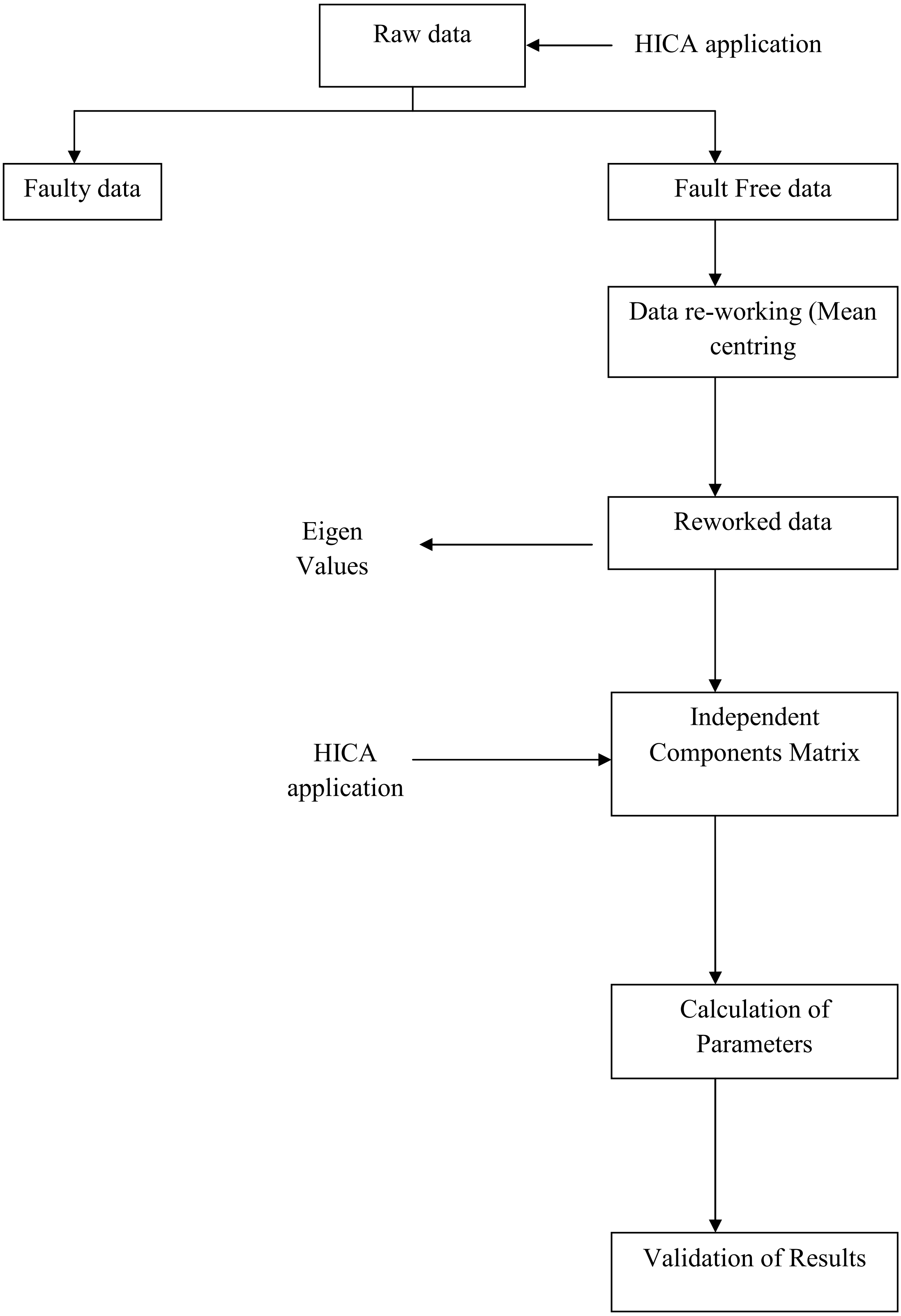


Figure 5.16: Flow Diagram of the Methodology used for Parameter Estimation using HICA

5.5.2 Application of Methodology to WWTP data set

The methodology is applied to WWTP data set as was done for both PCA and ICA for estimation of process parameters. Following are the steps involved in this work.

a) Identification of faulty days

The first step in the estimation of process parameters is the identification of faulty days. The technique is applied to the original data set of WWTP and faults are obtained. This work is already done in section 5.4.1 and the faulty days are removed from the data set.

b) Data pre-treatment

After removing all the faulty days the remaining data is pre-treated. It is done to get the values of HICAs in small range. In this work a combination of data pre-treatment techniques are used which include mean centring and data normalization. This is same as done in section 4.4 of this thesis. After application of these data pre-treatment techniques to the data set HICA is applied again and the graph between the number of HICAs and values of HICAs is plotted which is shown in Figure 5.17.

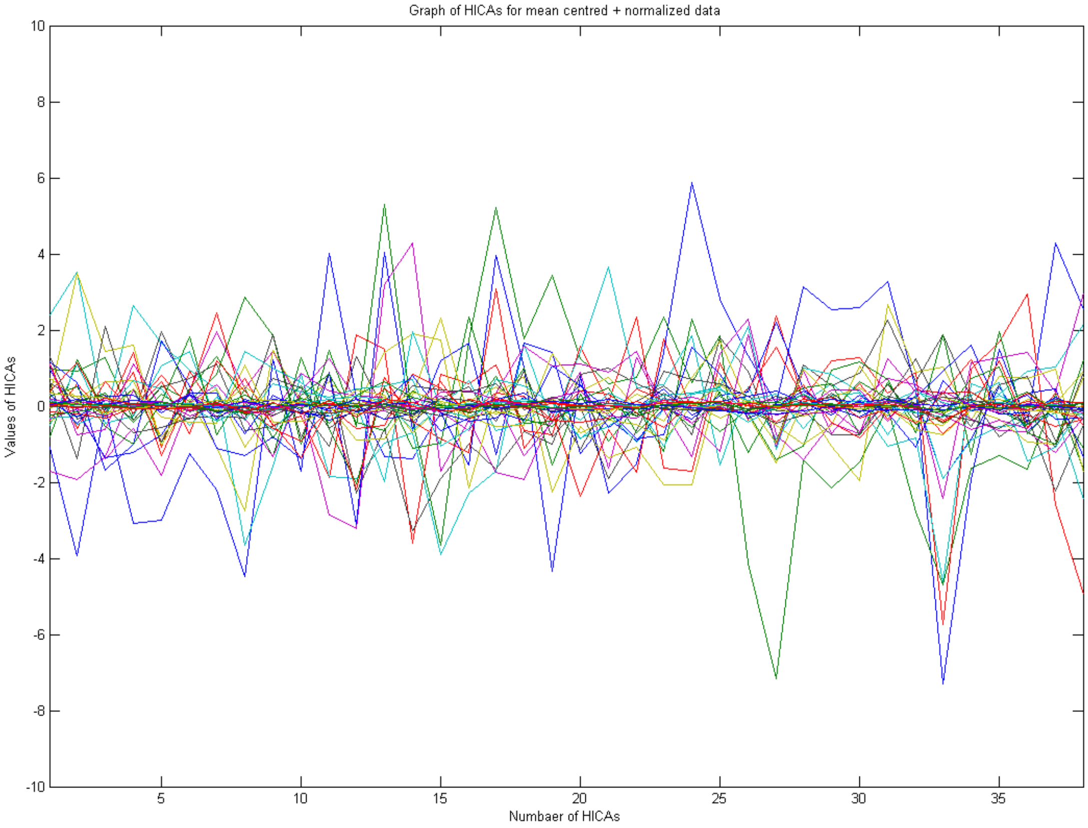


Figure 5.17: Graph between the values of HICAs and number of HICAs

As from the graph it is clear that HICAs have values in the range of “6” and “-6” so any values in this range can be used for the estimation of process parameters.

c) Parameters Estimation

After getting HICAs values in small range the Eigen values of this pre-treated data are calculated. In this work the assumed value of HICA is “5” which is to get input flow in range of 40000 otherwise any value of HICA (within range of “6” and “-6”) can be used. After having got the HICA values and Eigen values the next step is the calculation of ICs using HICA model. Then using the ICA model mixing matrix of this pre-treated data is calculated. The final step is the estimation of process parameters by using the calculated values of ICs (obtained from HICA model) and the mixing matrix in the ICA model. As the data at the start was mean centred and normalized so to get the original values of the parameters the results are de-normalized and de-mean centred.

d) Validation of Results and Discussion

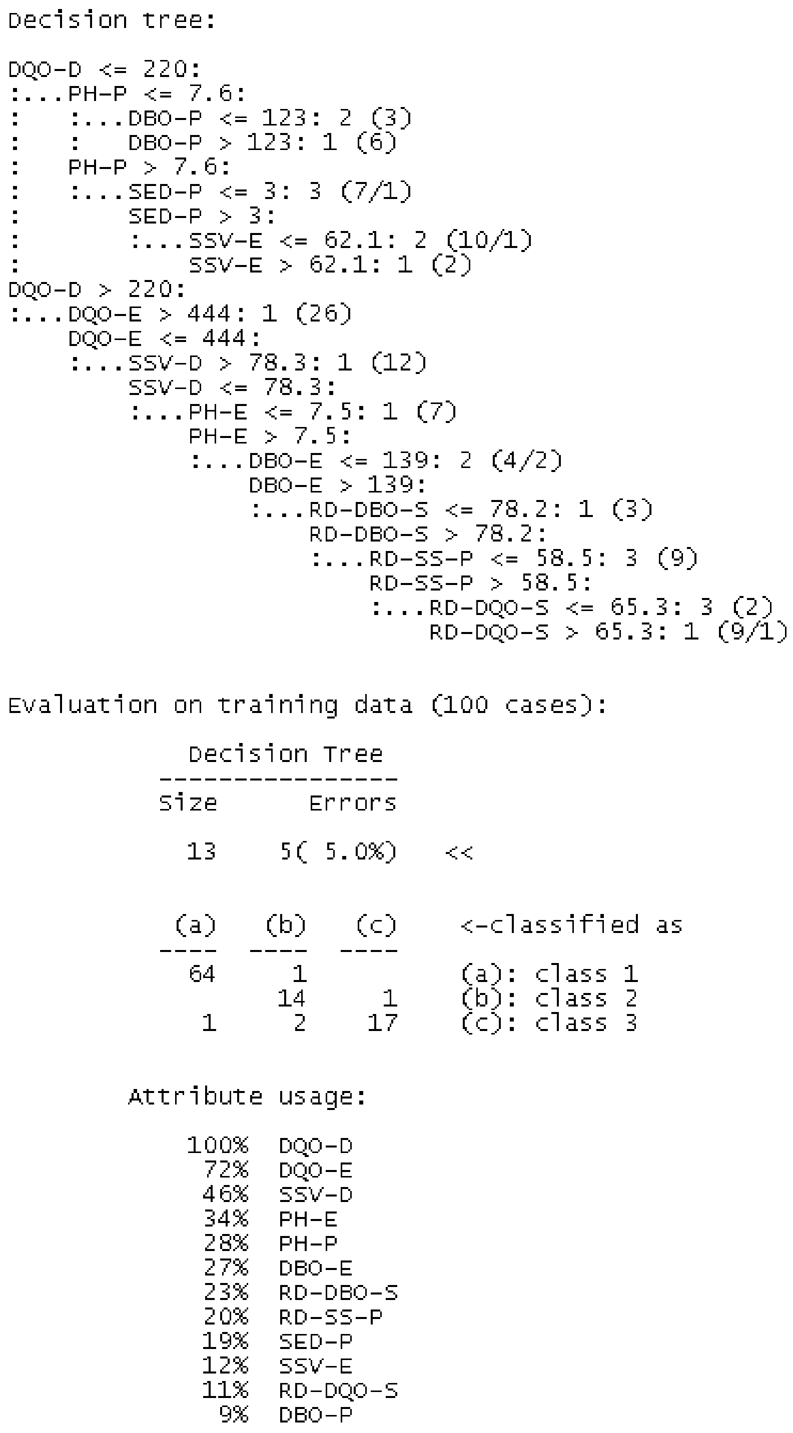
After the calculation of process parameters they are validated by construction of Decision Tree (s was done for PCA and ICA). Two Decision Trees are shown in Figure 5.18 and 5.19. The results for the first Decision Tree in Figure 5.18 are discussed as follows.

Results Discussion for first Decision Tree

The first Decision Tree is shown in Figure 5.18. The result obtained by the HICA model and the Decision Tree are compared. The Decision Tree gives information about seven of the parameters. From the comparison it is found that six out of seven of the parameters were validated by Decision Tree. Table 5.5 is presenting a summary of the calculated parameters values using HICA model and Decision Tree values.

Table 5.5: Summary of the Results obtained by Decision Tree (Fig.5.20)

Numbers	Parameters	Values determined by HICA model	Values by See 5.0	Validated
1	DQO-D	514.03	>220	Yes
2	SSV-D	49.73	≤78.3	Yes
3	PH-E	12.81	>7.5	Yes
4	DBO-E	223.55	>139	Yes
5	RD-DBO-S	27.08	≤78.2	Yes
6	RD-SS-P	130.66	>58.5	Yes
7	DQO-E	867.84	≤444	No

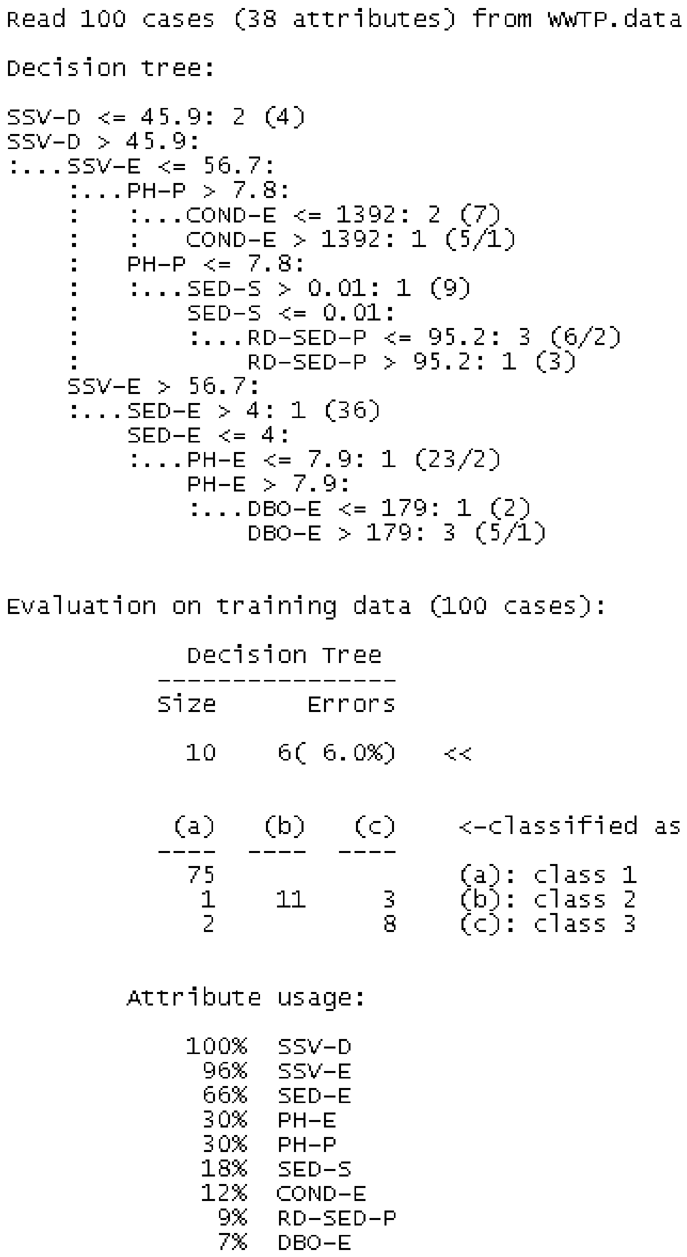


```
Decision tree:

DQO-D <= 220:
:...PH-P <= 7.6:
:   :...DBO-P <= 123: 2 (3)
:   :   DBO-P > 123: 1 (6)
:   PH-P > 7.6:
:   :...SED-P <= 3: 3 (7/1)
:       SED-P > 3:
:       :...SSV-E <= 62.1: 2 (10/1)
:           SSV-E > 62.1: 1 (2)
DQO-D > 220:
:...DQO-E > 444: 1 (26)
    DQO-E <= 444:
    :...SSV-D > 78.3: 1 (12)
        SSV-D <= 78.3:
        :...PH-E <= 7.5: 1 (7)
            PH-E > 7.5:
            :...DBO-E <= 139: 2 (4/2)
                DBO-E > 139:
                :...RD-DBO-S <= 78.2: 1 (3)
                    RD-DBO-S > 78.2:
                    :...RD-SS-P <= 58.5: 3 (9)
                        RD-SS-P > 58.5:
                        :...RD-DQO-S <= 65.3: 3 (2)
                            RD-DQO-S > 65.3: 1 (9/1)


Evaluation on training data (100 cases):

            Decision Tree
          ----------------
          Size      Errors

            13    5( 5.0%)   <<


           (a)   (b)   (c)    <-classified as
          ----  ----  ----
            64     1          (a): class 1
                  14     1    (b): class 2
             1     2    17    (c): class 3


        Attribute usage:

            100%  DQO-D
             72%  DQO-E
             46%  SSV-D
             34%  PH-E
             28%  PH-P
             27%  DBO-E
             23%  RD-DBO-S
             20%  RD-SS-P
             19%  SED-P
             12%  SSV-E
             11%  RD-DQO-S
              9%  DBO-P
```

Figure 5.18: Decision Tree validating the Results obtained for the Process Parameter Estimation using HICA

Results Discussion with another Decision Tree

To further verify the results obtained from the HICA model another Decision Tree is constructed. This Decision Tree also validates most of the results obtained. It has got two nodes and is shown in Figure 5.19.



```
Read 100 cases (38 attributes) from WWTP.data

Decision tree:

SSV-D <= 45.9: 2 (4)
SSV-D > 45.9:
:...SSV-E <= 56.7:
    :...PH-P > 7.8:
    :   :...COND-E <= 1392: 2 (7)
    :   :   COND-E > 1392: 1 (5/1)
    :   PH-P <= 7.8:
    :   :...SED-S > 0.01: 1 (9)
    :       SED-S <= 0.01:
    :       :...RD-SED-P <= 95.2: 3 (6/2)
    :           RD-SED-P > 95.2: 1 (3)
    SSV-E > 56.7:
    :...SED-E > 4: 1 (36)
        SED-E <= 4:
        :...PH-E <= 7.9: 1 (23/2)
            PH-E > 7.9:
            :...DBO-E <= 179: 1 (2)
                DBO-E > 179: 3 (5/1)


Evaluation on training data (100 cases):

            Decision Tree
          ----------------
          Size      Errors

            10    6( 6.0%)   <<


           (a)   (b)   (c)    <-classified as
          ----  ----  ----
            75                (a): class 1
             1    11     3    (b): class 2
             2           8    (c): class 3


        Attribute usage:

            100%  SSV-D
             96%  SSV-E
             66%  SED-E
             30%  PH-E
             30%  PH-P
             18%  SED-S
             12%  COND-E
              9%  RD-SED-P
              7%  DBO-E
```

Figure 5.19: Decision Tree validating the Results obtained for the Process Parameter Estimation using HICA

From the first node we find that it gives information about five of the estimated parameters. From these five parameters four are validated by Decision Tree. It shows the ability of model for correct parameter estimation. Table 5.6 is representing the parameter values for both HICA model and Decision Tree as follows.

Table 5.6: Summary of the Results obtained by Decision Tree (Figure5.19)

Numbers	Parameters	Values determined by HICA model	Values by See 5.0	Validated
1	SSV-D	49.73	>45.9	Yes
2	PH-P	6.46	≤7.8	Yes
3	SED-S	0.0027	≤0.01	Yes
4	RD-SED-P	153.66	>95.2	Yes
5	SSV-E	173.13	≤56.7	No

From the second node of this Decision Tree information about four parameters is obtained. If we compare the values given by the Decision Tree and HICA model then it is found that all the four parameters were correctly estimated by HICA model which indicates its ability to estimate parameters better than PCA and ICA. A summary of the results obtained from both HICA model and Decision Tree is given in Table 5.7.

Table 5.7: Summary of the Results obtained by Decision Tree (Figure 5.19)

Numbers	Parameters	Values determined by HICA model	Values by See 5.0	Validated
1	SSV-D	49.73	>45.9	Yes
2	SSV-E	173.13	>56.7	Yes
3	SED-E	2.37	≤4.0	Yes
4	PH-E	12.8	>7.9	Yes

5.6 Conclusions

A comparison of techniques such as PCA and ICA is made with a proposed technique of HICA in this work. Two case studies have been used for the validation of the results and to show the behaviour of the techniques for different data sets. For both case studies all of the techniques have shown quite good results. For the first case study PCA captured 23 days as abnormal days from 31 days (Total abnormal days identified by all the techniques) which is about 74% of all the abnormal days. Similarly for the second case study PCA identified 14 observations as abnormal observations out of total 21 abnormal observations (Total identified by all the techniques) which is almost 67% of all the abnormal observations. ICA has also shown good results identifying 19 out of 31 abnormal days for the first case study and 13 out of 21 for the second case study which are 61% and 62% respectively. The new proposed technique showed better results than individual techniques of PCA and ICA for both case studies. New technique identified 28 out of 31 faulty days (90%) for the first case study and 17 out of 21 faulty observations (81%) for the second case study. Although for both case studies HICA missed some faults but it was expected as it is hybrid of both PCA and ICA.

Also work has been done for the estimation of process parameters using HICA technique. The technique successfully estimated most of the process parameters which were validated by construction of two Decision Trees. As expected HICA technique was more successful in estimation of process parameter as compared to PCA and ICA by estimating six out of seven parameters correctly for first Decision Tree. For the second Decision Tree it was more successful as it estimated four out of five parameters correctly for first node of the Tree and four out of four for the second node of the Decision Tree.

Chapter # 6

Conclusions and Recommendations

6.1 Conclusions

The multivariate techniques of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) have been explored in this work. The work includes utilization of models of PCA and ICA for the estimation of process parameters. Both PCA and ICA were used independently for the estimation of process parameters by their application to historical data set of waste water treatment plant (WWTP).

The results obtained by the PCA model show its ability for the estimation of process parameters. All the parameters were calculated by PCA model. The calculated parameters were validated by construction of two Decision Trees using data mining software (See 5.0). The basic concept of inductive data mining and technique of Decision Tree has also been explained in this work. From the first Decision Tree six, out of seven parameters were validated. Similarly from the other Decision Tree six out of nine parameters were validated. The reason for not validating all of the parameters may be the strict application of PCA to Gaussian data set. The data set utilized in this study is industrial data and has been statistically evaluated. By the evaluation of the data set it was revealed that the data set is a mixture of both Gaussian and non-Gaussian variables. As we utilized a mixture of Gaussian and non-Gaussian data the PCA model was unable to give accurate results.

Similar to PCA, ICA is also utilized for the estimation of process parameters using the same case study of waste water treatment plant (WWTP). ICA successfully estimated all the parameters and the validation of these are carried out using data mining software (See5.0). Two Decision Trees were constructed for the validation of obtained results. The first Decision Tree validated five out of six parameters. The second Decision Tree also validated five out of six parameters. These results indicate the ability of the model to estimate process parameters. The restriction that ICA could not validate all the estimated parameter is due to the fact that ICA model is strictly valid for non-Gaussian data. The waste water treatment plant data used as a case study in this work is a combination of Gaussian and non-Gaussian data so it could not validate all of the estimated parameters.

Comparison of the techniques of PCA and ICA is also carried out for their ability to identify faults in a system. Along with the comparison of the techniques, a new technique is also proposed which is hybrid of both PCA and ICA called HICA (Hybrid Independent Component Analysis). Two different data sets .i.e. waste water treatment plant data and air pollution data set are used as case studies for comparison of all the results and hence evaluation of these techniques. Overall both PCA and ICA have shown good results but in the comparison of these individual techniques with HICA, the new proposed technique of HICA was much more successful in terms of fault detection. In the two case studies PCA detected 74% and 67% faults, ICA detected 61.3% and 62% and HICA detected 90% and 80% respectively. These results show good comparison of these techniques along with superiority of HICA over PCA and ICA.

This work also includes the evaluation of the faults after they have been detected by PCA, ICA or HICA. By the utilization of any of the techniques, once it is established that there is fault at particular day or observation, the reason behind this fault is also determined. For example with PCA, if there is any particular day or observation classified as a fault, the reason behind this fault is also explored. It is done by the construction of contribution plots of all the variables in the system at that particular day or observation based on the abnormal value of PC. From these contribution plots all the variables contributing to this PC value are determined. From these variables it can be established that one or more of the variables are responsible for the fault. Similar work is carried out with other techniques of ICA and HICA and the variables responsible for the abnormal behaviour of the system are determined. Overall in this work a comprehensive system of fault detection is reported with comparison of different multivariate process monitoring techniques.

The proposed technique of HICA is also used to estimate the process parameters. It was shown that HICA was more successful to estimate the parameters correctly in comparison to PCA or ICA.

6.2 Recommendations for Future work

This study can lead to the following future work.

1. The conversion of the industrial data set (historical or operational) which is usually a mixture of Gaussian and non-Gaussian data set to a Gaussian data set and the application of PCA model for the estimation of the process parameters is required to overcome the limitation of PCA model being used for process parameters estimation. For the conversion of data to Gaussian distribution we can use the technique of Box-Cox transformation as utilized by Albazzaz and Wang (2006) in their work for process monitoring.
2. The application of HICA for online process monitoring is one of the future directions of this work. The following methodology can be used for this work.

   Using HICA faults are determined and contribution plots are constructed. For the online monitoring, the data generated by the system and HICA values are determined. These values are then compared with the benchmark set by the historical data set. The abnormal values of HICA show abnormal event and from these abnormal values of HICA contribution plots for the system are made. Based on the contribution plots rectification action can be carried out.

REFERENCES

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

Albazzaz, H. & Wang, X. Z. (2004) Statistical Process Control Charts for Batch Operations Based on Independent Component Analysis. *Industrial & Engineering Chemistry Research.*

Albazzaz, H. & Wang, X. Z. (2006) Historical data analysis based on plots of independent and parallel coordinates and statistical control limits. *Journal of Process Control,* vol. 16, no. 2, pp. 103-114.

Albazzaz, H. & Wang, X. Z. (2007) Introduction of dynamics to an approach for batch process monitoring using independent component analysis. *Chemical Engineering Communications,* vol. 194, no. 2, pp. 218-233.

Albazzaz, H., Wang, X. Z. & Marhoon, F. (2005) Multidimensional visualisation for process historical data analysis: a comparative study with multivariate statistical process control. *Journal of Process Control,* vol. 15, no. 3, pp. 285-294.

AlGhazzawi, A. & Lennox, B. (2008) Monitoring a complex refining process using multivariate statistics. *Control Engineering Practice,* vol. 16, no. 3, pp. 294-307.

Amand, T., Heyen, G. & Kalitventzeff, B. (2001) Plant monitoring and fault detection: Synergy between data reconciliation and principal component analysis. *Computers & Chemical Engineering,* vol. 25, no. 4-6, pp. 501-507.

Barnes, J. W. (1998) *Statistical Analysis for Engineers and Scientists:A COMPUTER-BASED APPROACH.*

Bendwell, N. (2002) Monitoring of a wastewater-treatment plant with a multivariate model - The benefits of PCA technology are explained. *Pulp & Paper-Canada,* vol. 103, no. 7, pp. 43-46.

Bersimis, S., Psarakis, S. & Panaretos, J. (2007) Multivariate statistical process control charts: An overview. *Quality and Reliability Engineering International,* vol. 23, no. 5, pp. 517-543.

Brauner, N. & Shacham, M. (2000) Considering precision of data in reduction of dimensionality and PCA. *Computers & Chemical Engineering,* vol. 24, no. 12, pp. 2603-2611.

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

Choi, S. W., Lee, C., Lee, J. M., Park, J. H. & Lee, I. B. (2005) Fault detection and identification of nonlinear processes based on kernel PCA. *Chemometrics and Intelligent Laboratory Systems,* vol. 75, no. 1, pp. 55-67.

Cichocki, A. & Amari, S.-i. (2002) *Adaptive blind signal and image processing : learning algorithms and applications,* Chichester ;, J. Wiley.

DeLisle, R. K. & Dixon, S. L. (2004) Induction of Decision Trees via Evolutionary Programming. *Journal of Chemical Information and Computer Sciences,* vol. 44, no. 3, pp. 862-870.

DeVor, R. E., Chang, T.-h. & Sutherland, J. W. (1992) *Statistical quality design and control : contemporary concepts and methods,* New York, Toronto, Macmillan ; Maxwell Macmillan Canada ; Maxwell Macmillan International.

Draper, N. R. & Smith, H. (1981) *Applied regression analysis,* New York, Wiley.

Dunia, R., Qin, S. J., Edgar, T. F. & McAvoy, T. J. (1996) Identification of faulty sensors using principal component analysis. *Aiche Journal,* vol. 42, no. 10, pp. 2797-2812.

Dunteman, G. H. (1989) *Principal components analysis,* Newbury Park, Calif., Sage.

Englezos, P. & Kalogerakis, N. (2001) *Applied parameter estimation for chemical engineers,* New York, Marcel Dekker.

Flury, B. (1997) *A first course in multivariate statistics,* New York, Springer.

Gan, G., Ma, C. & Wu, J. (2007) *Data clustering : theory, algorithms, and applications,* Philadelphia, Pa.,Alexandria, Va., SIAM, American Statistical Association.

Ganesan, R., Das, T. K. & Venkataraman, V. (2004) Wavelet-based multiscale statistical process monitoring: A literature review. *Iie Transactions,* vol. 36**,** no. 9, pp. 787-806.

Ge, Z. Q. & Song, Z. H. (2007) Process monitoring based on independent component analysis-principal component analysis (ICA-PCA) and similarity factors. *Industrial & Engineering Chemistry Research,* vol. 46**,** no. 7, pp. 2054-2063.

Gertler, J. (1998) *Fault detection and diagnosis in engineering systems,* New York, Marcel Dekker.

Hair, J. F., Anderson, R. E. & Tatham, R. L. (1987) *Multivariate data analysis with readings,* New York London, Macmillan ; Collier Macmillan.

Hair, J. F., Anderson, R. E. & Tatham, R. L. (1990) *Multivariate data analysis : with readings,* New York, Macmillan.

He, Q. B., Yan, R. Q., Kong, F. R. & Du, R. X. (2009) Machine condition monitoring using principal component representations. *Mechanical Systems and Signal Processing,* vol. 23**,** no. 2, pp. 446-466.

Helsinki, U. o. (2006) What is Independent Component Analysis? http://www.cs.helsinki.fi/u/ahyvarin/whatisica.shtml

HyvÃ¤rinen, A. & Oja, E. (1997) A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation,* vol. 9, no. 7, pp. 1483-1492.

Hyvarinen, A. & Oja, E. (2000) Independent component analysis: algorithms and applications. *Neural Networks,* vol. 13, no. 4-5, pp. 411-430.

Hyvärinen, A., Oja, E. & Karhunen, J. (2001) *Independent component analysis,* New York ; Chichester, John Wiley & Sons.

Issanchou, S., Cognet, P. & Cabassud, M. (2003) Precise parameter estimation for chemical batch reactions in heterogeneous medium. *Chemical Engineering Science,* vol. 58, no. 9, pp. 1805-1813.

Jackson, J. E. (2003) *A user's guide to principal components,* Hoboken, N.J., Wiley.

Jeng, J. C., Li, C. C. & Huang, H. P. (2006) Dynamic processes monitoring using predictive PCA. *Journal of the Chinese Institute of Engineers,* vol. 29, no. 2, pp. 311-318.

Jolliffe, I. T. (2002) Principal component analysis.

Kano, M., Nagao, K., Hasebe, S., Hashimoto, I., Ohno, H., Strauss, R. & Bakshi, B. (2000) Comparison of statistical process monitoring methods: application to the Eastman challenge problem. *Computers & Chemical Engineering,* vol. 24, no. 2-7, pp. 175-181.

Kano, M., Tanaka, S., Hasebe, S., Hashimoto, I. & Ohno, H. (2003) Monitoring independent components for fault detection. *Aiche Journal,* vol. 49, no. 4, pp. 969-976.

Kim, I.-W., Kang, M. S., Park, S. & Edgar, T. F. (1997) Robust data reconciliation and gross error detection: The modified MIMT using NLP. *Computers & Chemical Engineering,* vol. 21, no. 7, pp. 775-782.

Lattin, J. M., Green, P. E., Carroll, J. D. & Green, P. E. (2003) *Analyzing multivariate data,* Southbank, Vic., Thomson Brooks/Cole.

Lee, D. S., Park, J. M. & Vanrolleghem, P. A. (2005) Adaptive multiscale principal component analysis for on-line monitoring of a sequencing batch reactor. *Journal of Biotechnology,* vol. 116, no. 2, pp. 195-210.

Lee, J. M., Yoo, C. K. & Lee, I. B. (2003) New monitoring technique with an ICA algorithm in the wastewater treatment process. *Water Science and Technology,* vol. 47, no. 12, pp. 49-56.

Li, W. H., Yue, H. H., Valle-Cervantes, S. & Qin, S. J. (2000) Recursive PCA for adaptive process monitoring. *Journal of Process Control,* vol. 10, no. 5, pp. 471-486.

Li, Y., Xie, Z. & Zhou, D. H. (2004) Fault detection and isolation based on abnormal sub-regions using the improved PCA. *Journal of Chemical Engineering of Japan,* vol. 37, no. 4, pp. 514-522.

Lohmann, T., Bock, H. G. & Schloder, J. P. (1992) Numerical-Methods for Parameter-Estimation and Optimal Experiment Design in Chemical-Reaction Systems. *Industrial & Engineering Chemistry Research,* vol. 31, no. 1, pp. 54-57.

Luo, R. F., Misra, M. & Himmelblau, D. M. (1999) Sensor fault detection via multiscale analysis and dynamic PCA. *Industrial & Engineering Chemistry Research,* vol. 38, no. 4, pp. 1489-1495.

Macgregor, J. F. & Kourti, T. (1995) Statistical Process-Control of Multivariate Processes. *Control Engineering Practice,* vol. 3, no. 3, pp. 403-414.

Machine.learning.database (2008) Air Pollution data.

Mason, R. L., Gunst, R. F. & Hess, J. L. (1989) *Statistical design and analysis of experiments : with applications to engineering and science,* New York, Wiley.

MATLAB (2009) Principal Component Analysis.

McDonald, G. C. & Schwing, R. C. (1973) Instabilities of Regression Estimates Relating Air Pollution to Mortality. *Technometrics,* vol. 15, no. 3, pp. 463-481.

Misra, M., Yue, H. H., Qin, S. J. & Ling, C. (2002) Multivariate process monitoring and fault diagnosis by multi-scale PCA. *Computers & Chemical Engineering,* vol. 26, no. 9, pp. 1281-1293.

Montgomery, D. C. (1991) *Introduction to statistical quality control,* New York, Wiley.

Murthy, S. K. (1998) Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey *Data Mining and Knowledge Discovery,* vol., no. Volume 2, Number 4 / December, 1998, pp. 345-389.

Narasimhan, S. & Shah, S. L. (2008) Model identification and error covariance matrix estimation from noisy data using PCA. *Control Engineering Practice,* vol. 16, no. 1, pp. 146-155.

Newmark, J. (1997) *Statistics and probability in modern life,* Fort Worth ; London, Saunders College.

Ning He, Jian-ming Zhang & Wang, S.-q. (2004) Combination of Independent Component Analysis and Multi-way principal component Analysis for batch process monitoring.

NIST/SEMATECH (2009) e-Handbook of Statistical Methods.

Quinlan, J. R. (1986) Induction of decision trees. *Machine Learning,* vol. 1, no. 1, pp. 81-106.

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

Quinlan, J. R. (1993b) *C4.5:Programs for machine learning*, Morgan Kuffmann.

Quinlan, J. R. (1996) Improved use of continuous attributes in C 4. 5. *Journal of Artificial Intelligence Research,* vol. 4, no. 77-90, pp. 325.

Refaat, M. (2007) *Data preparation for data mining using SAS,* San Francisco, Morgan Kaufmann Publishers.

Sanchez, M., Cortes, U., Bejar, J., DeGracia, J., Lafuente, J. & Poch, M. (1997) Concept formation in WWTP by means of classification techniques: A compared study. *Applied Intelligence,* vol. 7, no. 2, pp. 147-165.

Scholz, M. (2006) Approaches to analyse and interpret biological profile data.

Schwaab, M., Biscaia, J. E. C., Monteiro, J. L. & Pinto, J. C. (2008) Nonlinear parameter estimation through particle swarm optimization. *Chemical Engineering Science,* vol. 63, no. 6, pp. 1542-1552.

Shinde, R. L. & Khadse, K. G. (2009) Multivariate Process Capability Using Principal Component Analysis. *Quality and Reliability Engineering International,* vol. 25, no. 1, pp. 69-77.

Shlens, J. (2005) A Tutorial on Principal Component Analysis.

Smith, L. I. (2002) *A Tutorial on Principal Component Analysis*.

Song, Z. G. a. Z. (2007) Process Monitoring Based on Independent Component Analysis-Principle Component Analysis (ICA-PCA) and Similarity Factors.

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

Sun, X., Marquez, H. J., Chen, T. W. & Riaz, M. (2005) An improved PCA method with application to boiler leak detection. *Isa Transactions,* vol. 44, no. 3, pp. 379-397.

Venkatasubramanian, V. (2005) Prognostic and diagnostic monitoring of complex systems for product lifecycle management: Challenges and opportunities. *Computers and Chemical Engineering,* vol. 29, no. 6, pp. 1253-1263.

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

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

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

Wang, D. & Romagnoli, J. A. (2005) Robust multi-scale principal components analysis with applications to process monitoring. *Journal of Process Control,* vol. 15, no. 8, pp. 869-882.

Wang, H. Q., Song, Z. H. & Wang, H. (2002) Statistical process monitoring using improved PCA with optimized sensor locations. *Journal of Process Control,* vol. 12, no. 6, pp. 735-744.

Wang, X. Z., Medasani, S., Marhoon, F. & Albazzaz, H. (2004) Multidimensional visualization of principal component scores for process historical data analysis. *Industrial & Engineering Chemistry Research,* vol. 43, no. 22, pp. 7036-7048.

Wastewaterdatabase (2008) Wastewaterdatabase.

Yoon, S. Y. & MacGregor, J. F. (2001) Fault diagnosis with multivariate statistical models part I: using steady state fault signatures. *Journal of Process Control,* vol. 11, no. 4, pp. 387-400.

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

APPENDIX A

This Appendix represents the probability graphs of the parameters of Waste Water treatment Plant (WWTP) to determine whether the parameter is Gaussian or non-Gaussian.

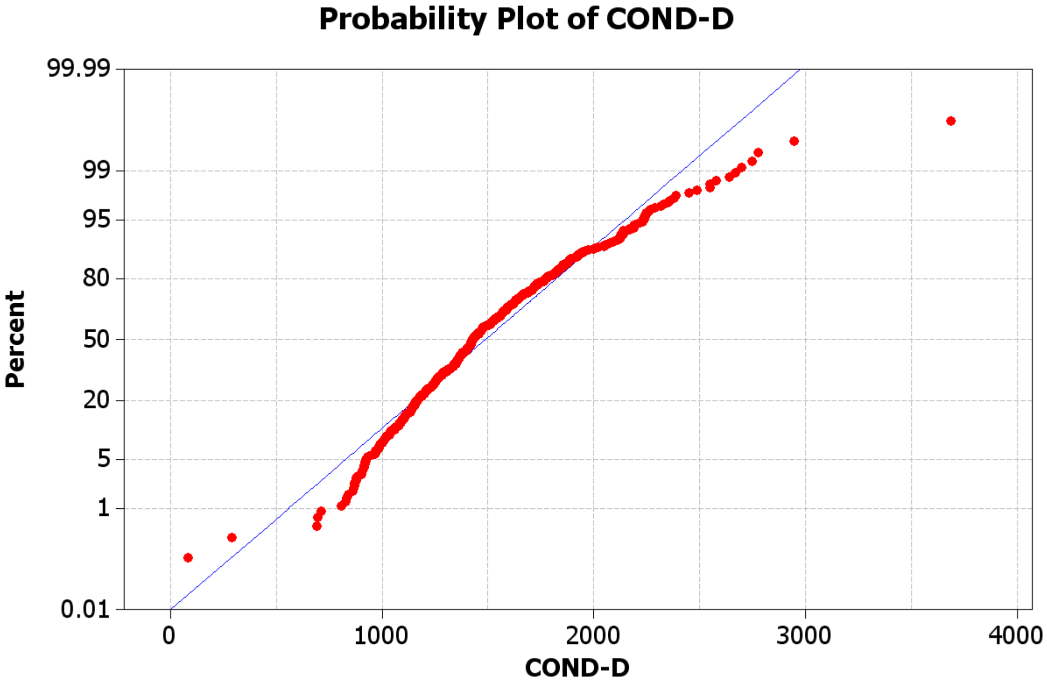


Figure A-1: Probability plot of COND-D showing Gaussian distribution

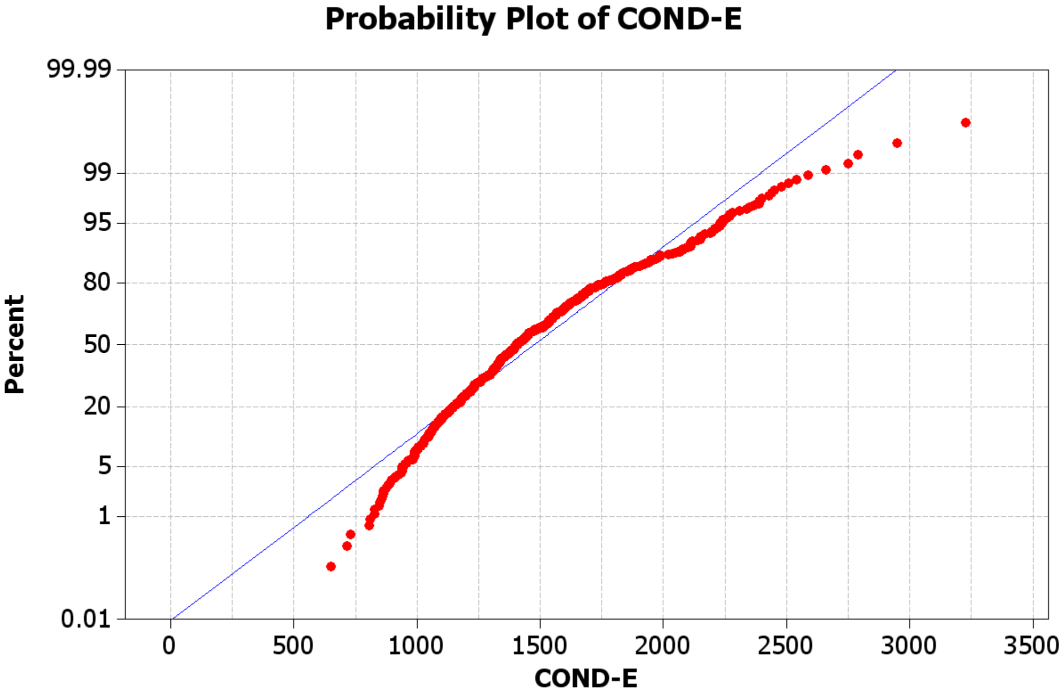


Figure A-2: Probability plot of COND-E showing Gaussian distribution

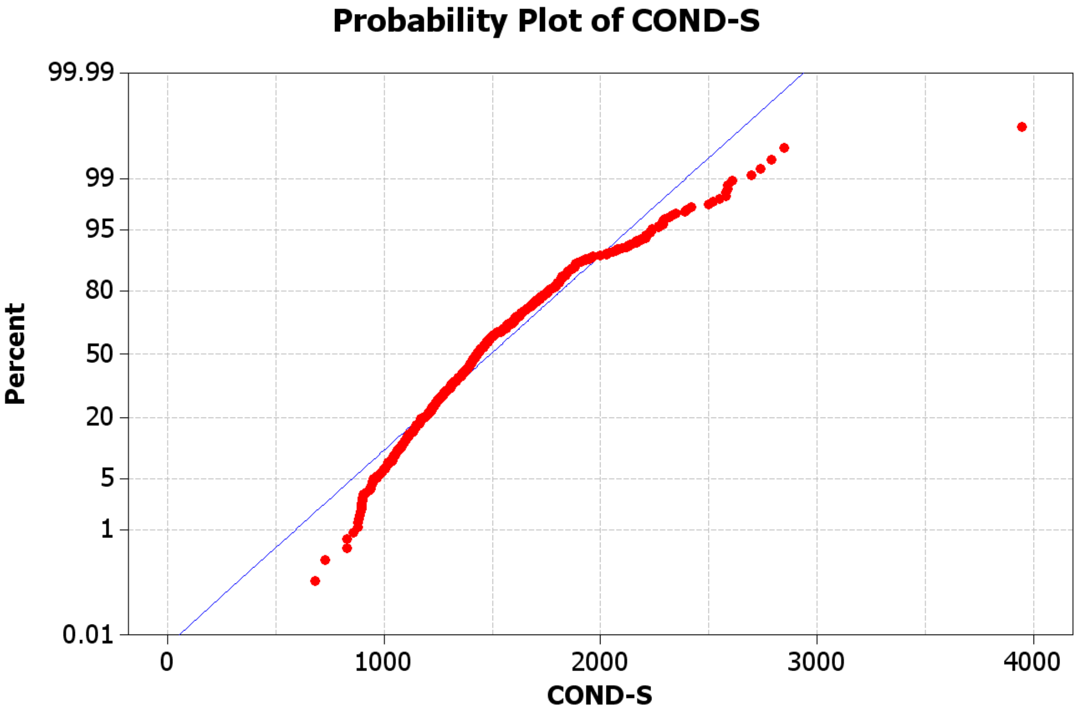


Figure A-3: Probability plot of COND-S showing Gaussian distribution

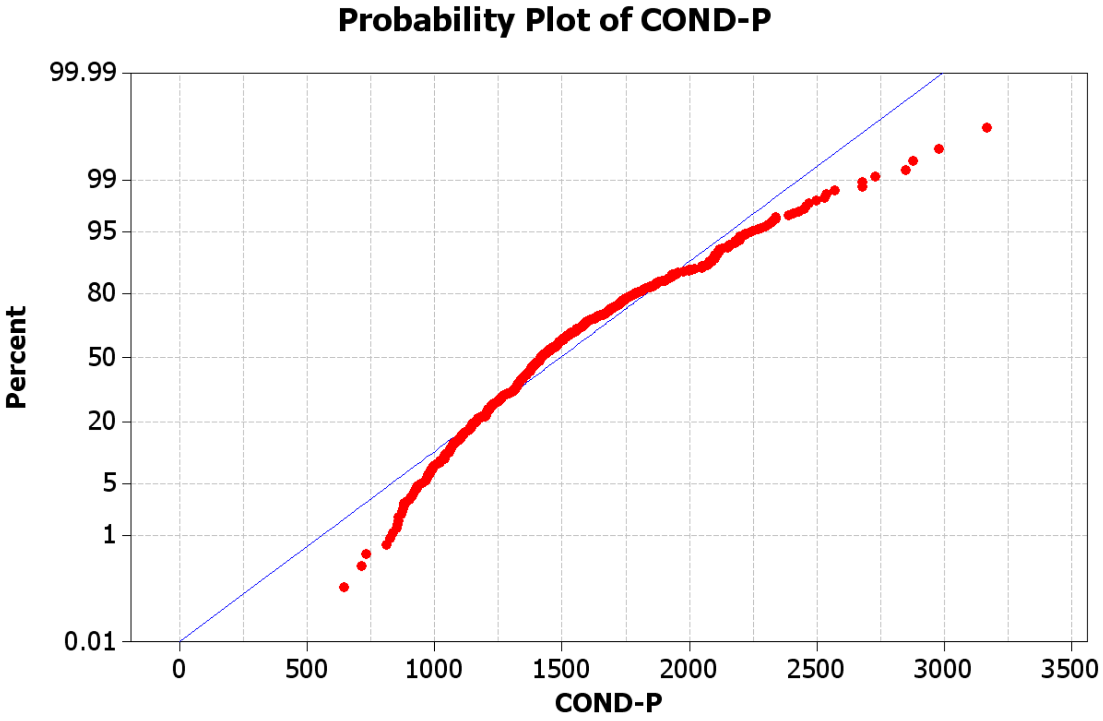


Figure A-4: Probability plot of COND-P showing Gaussian distribution

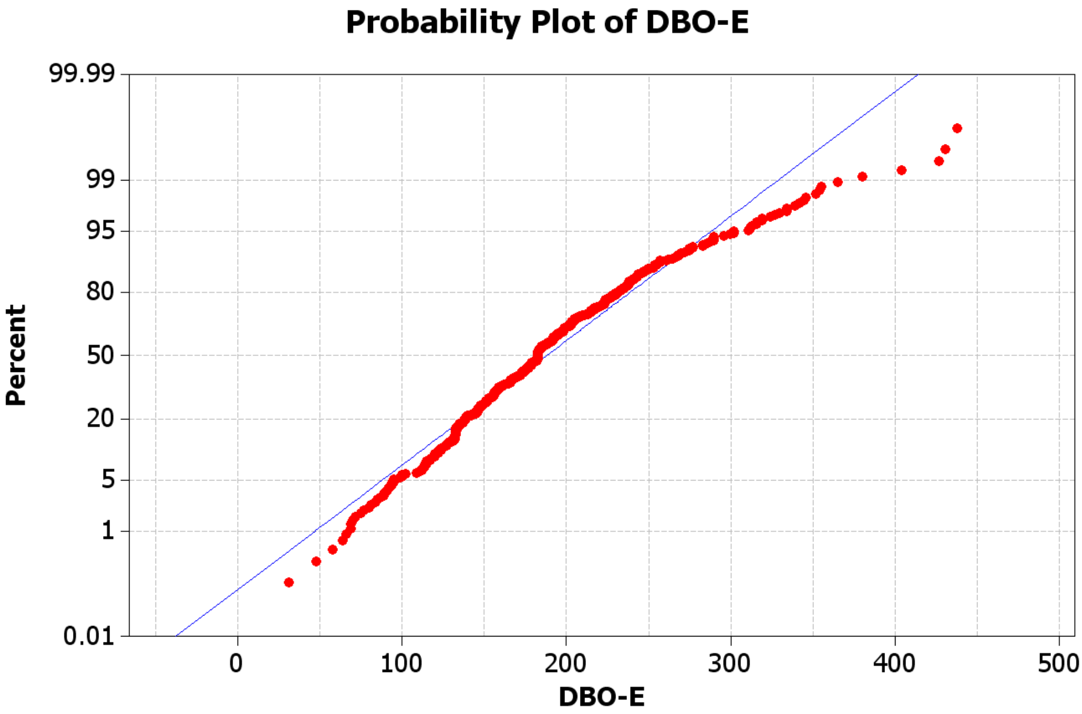


Figure A-5: Probability plot of DBO-E showing Gaussian distribution

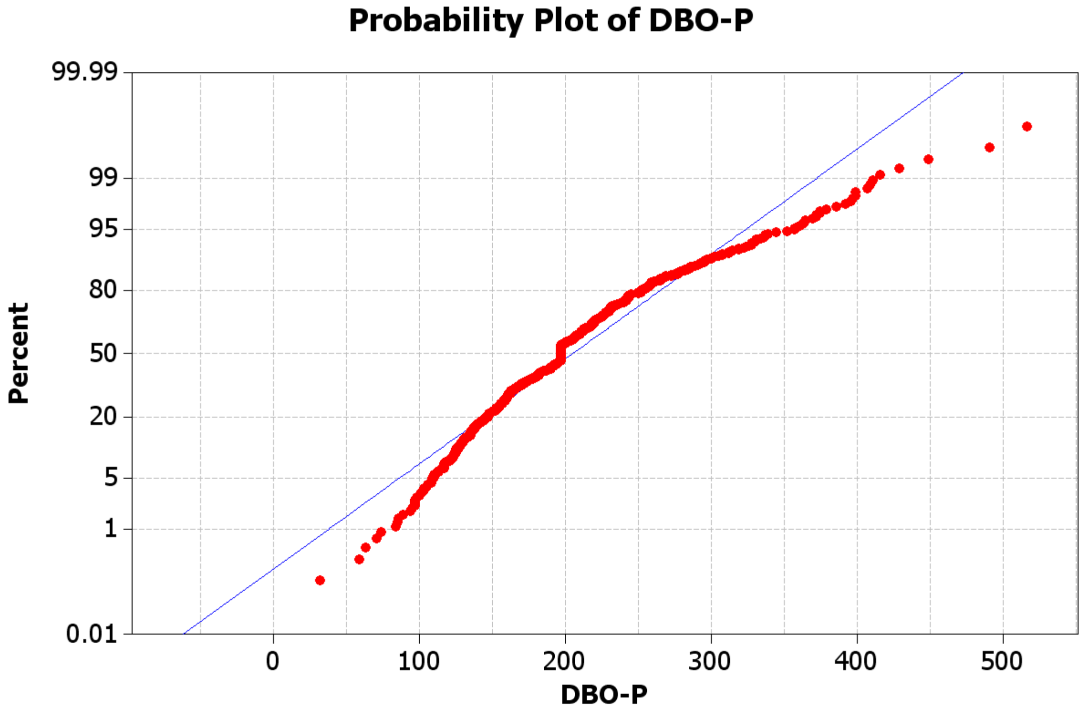


Figure A-6: Probability plot of DBO-P showing Gaussian distribution

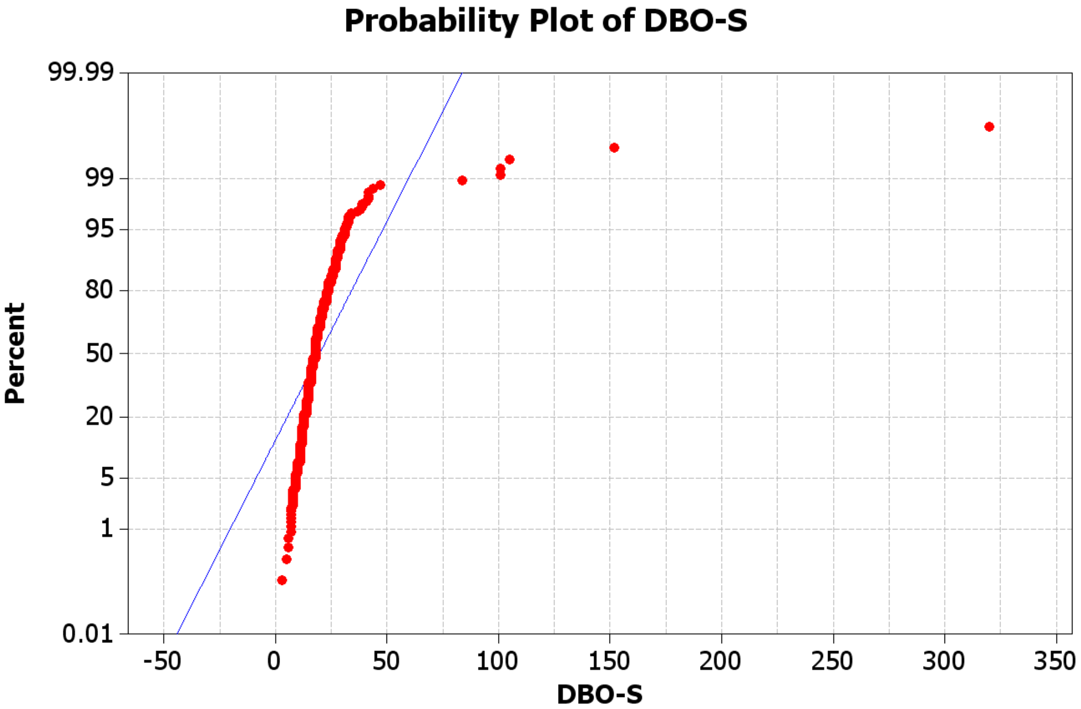


Figure A-7: Probability plot of DBO-S showing non-Gaussian distribution

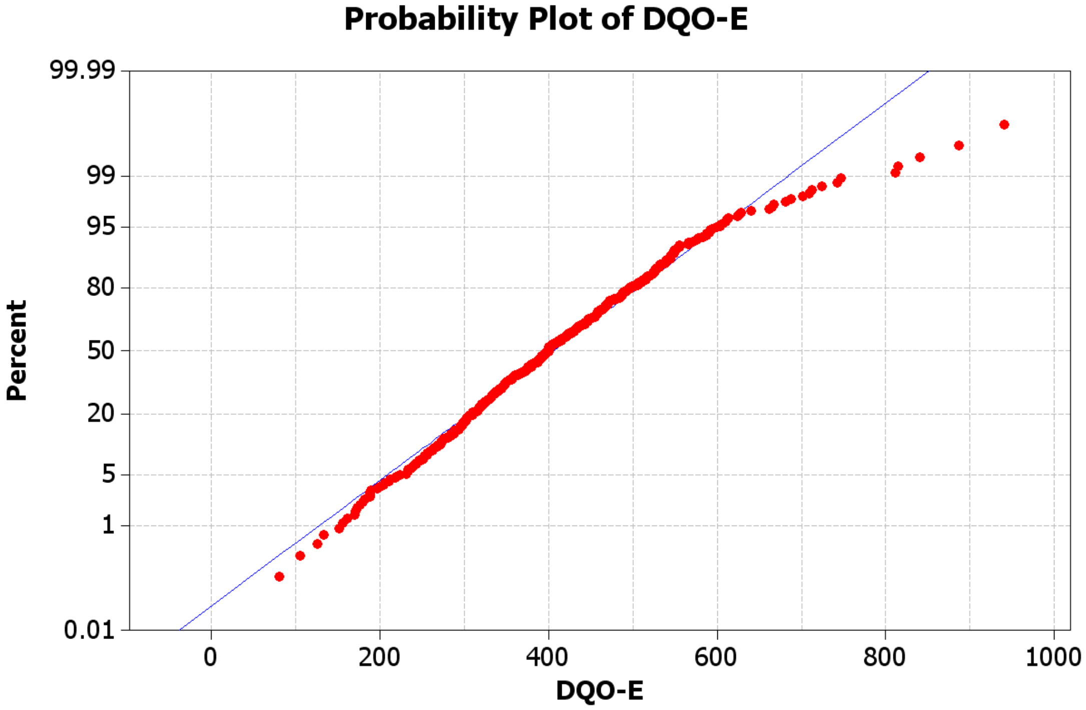


Figure A-8: Probability plot of DQO-E showing Gaussian distribution

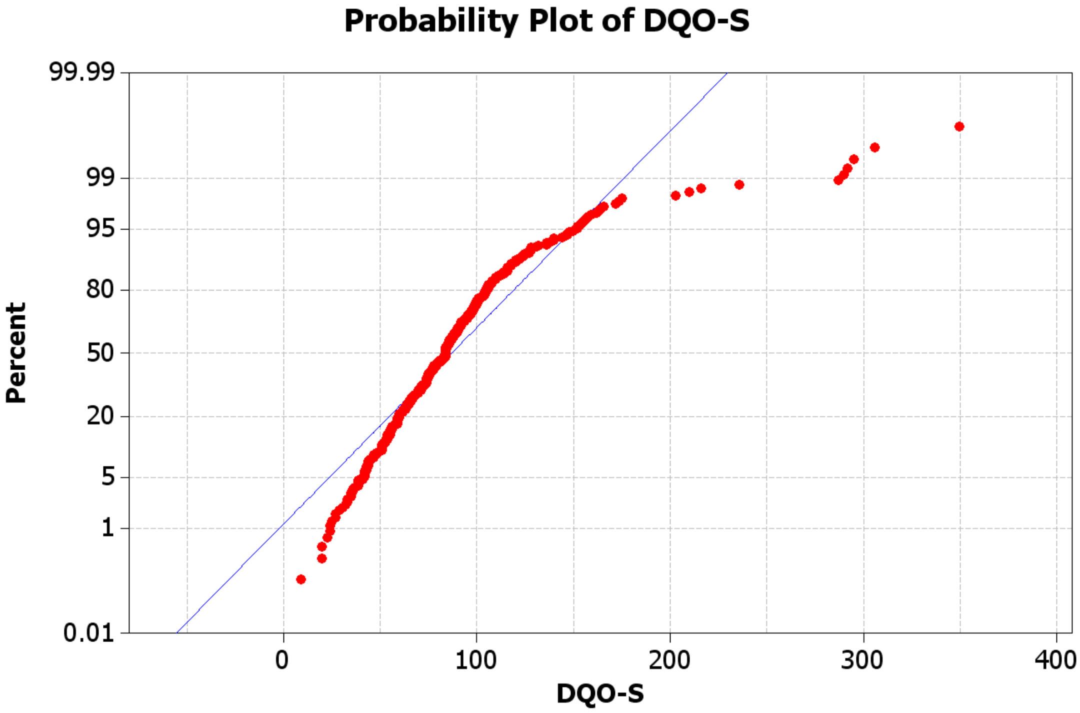


Figure A-9: Probability plot of DQO-S showing Gaussian distribution

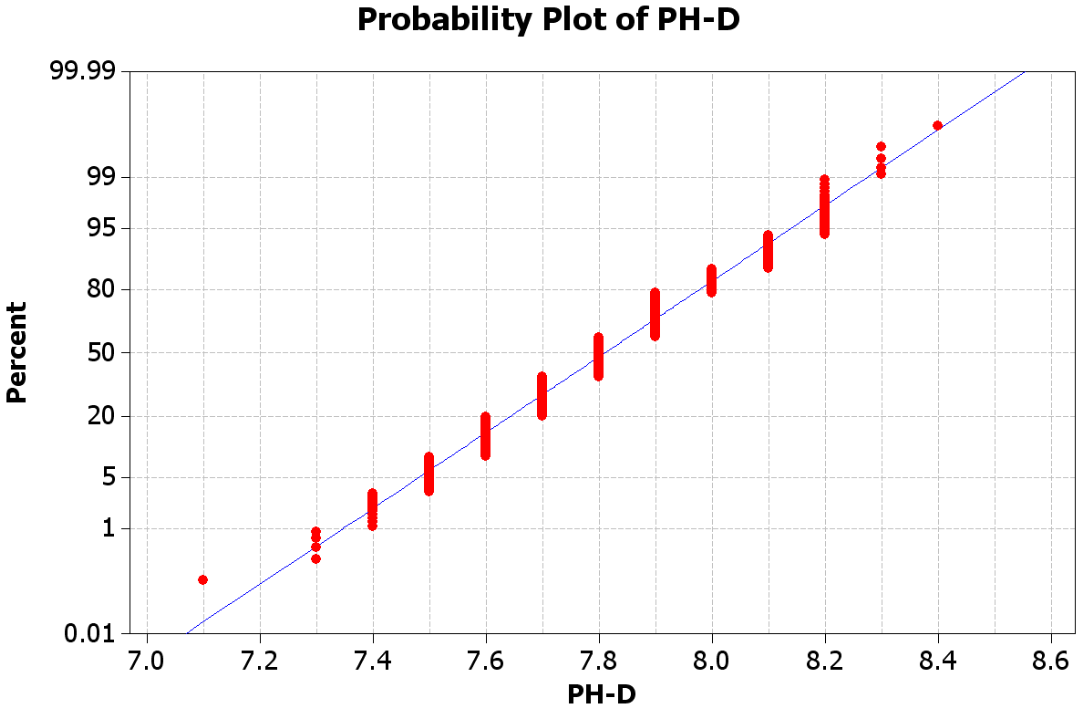


Figure A-10: Probability plot of PH-D showing non-Gaussian distribution

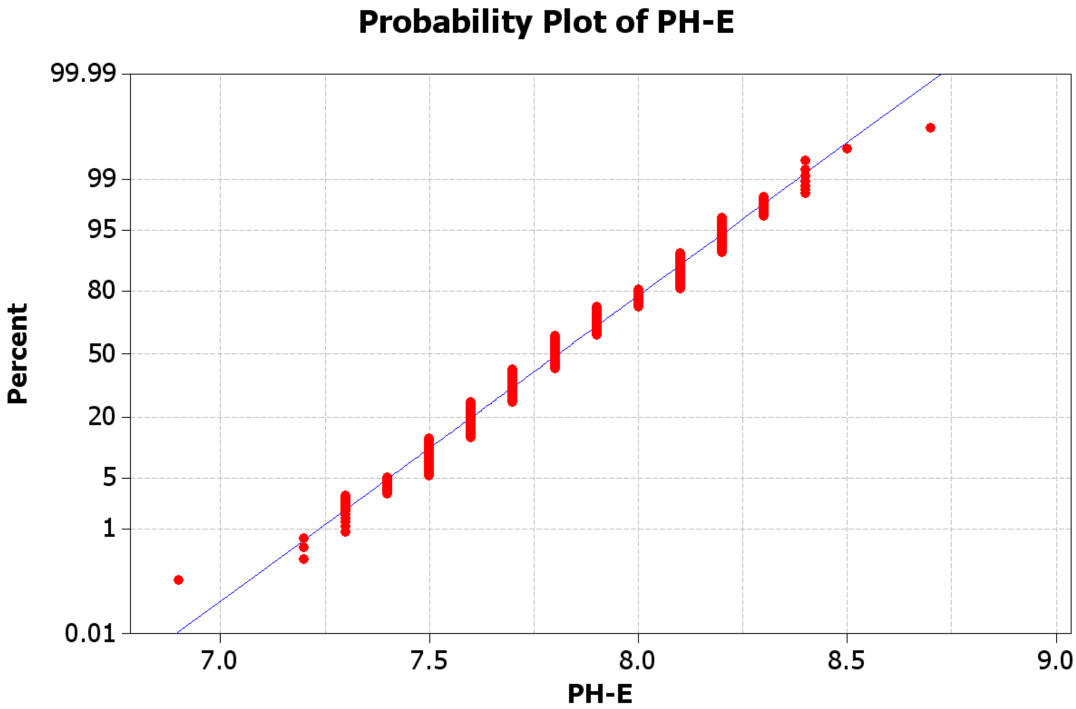


Figure A-11: Probability plot of PH-E showing non-Gaussian distribution

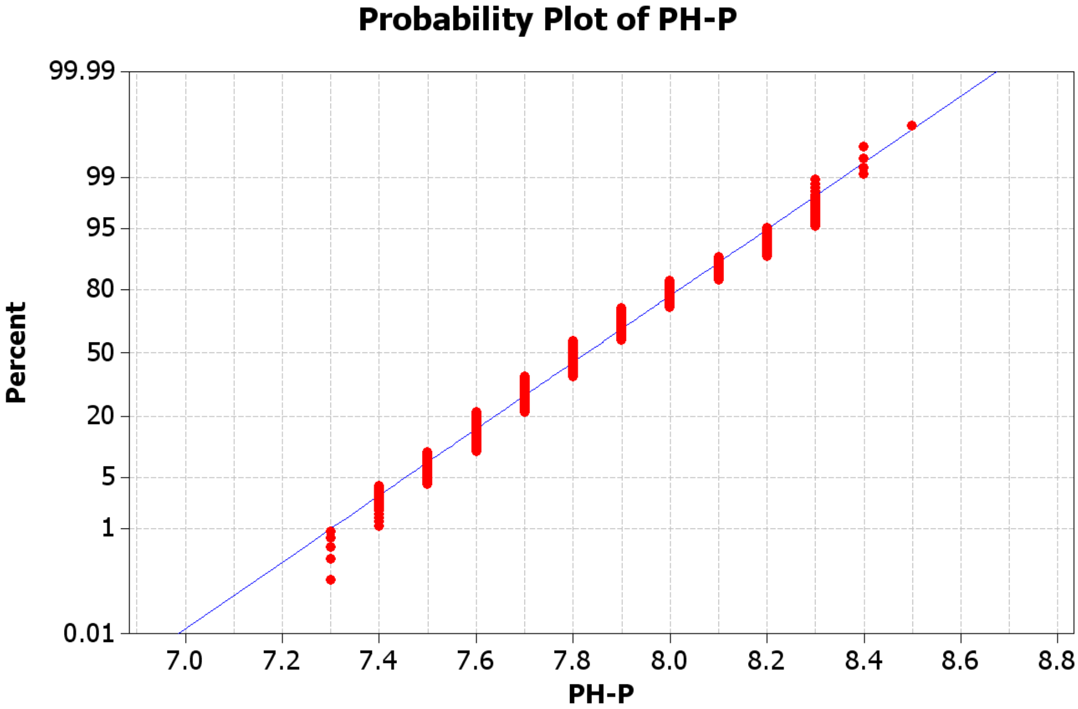


Figure A-12: Probability plot of PH-P showing non-Gaussian distribution

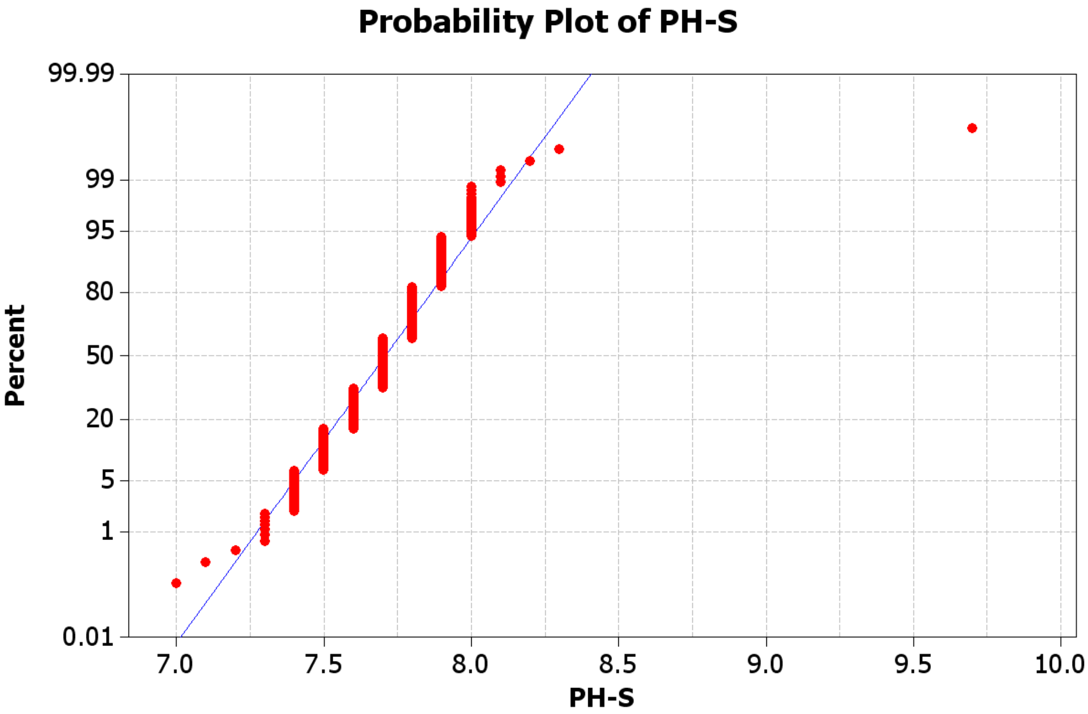


Figure A-13: Probability plot of PH-S showing non-Gaussian distribution

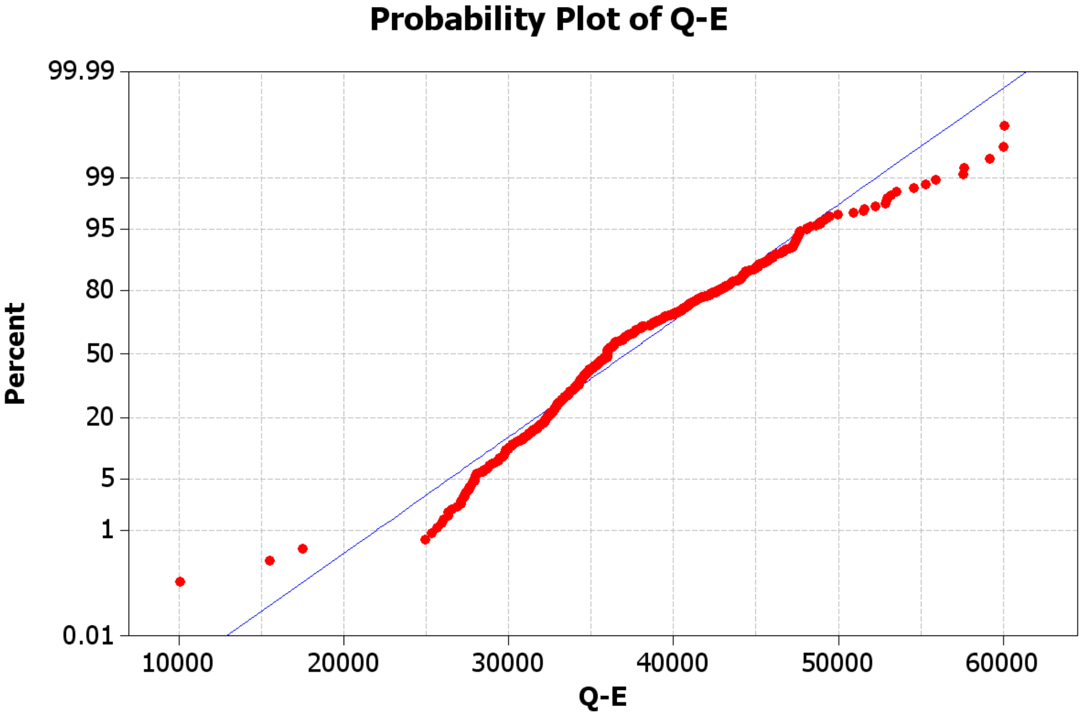


Figure A-14: Probability plot of Q-E showing Gaussian distribution

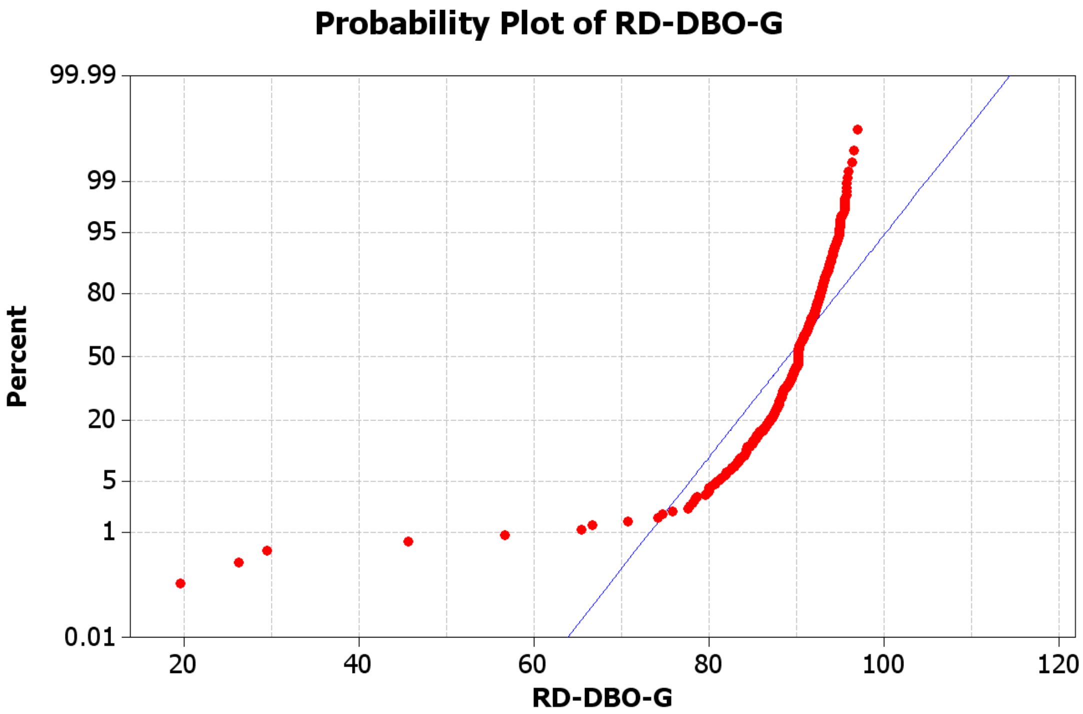


Figure A-15: Probability plot of RD-DBO-G showing non-Gaussian distribution

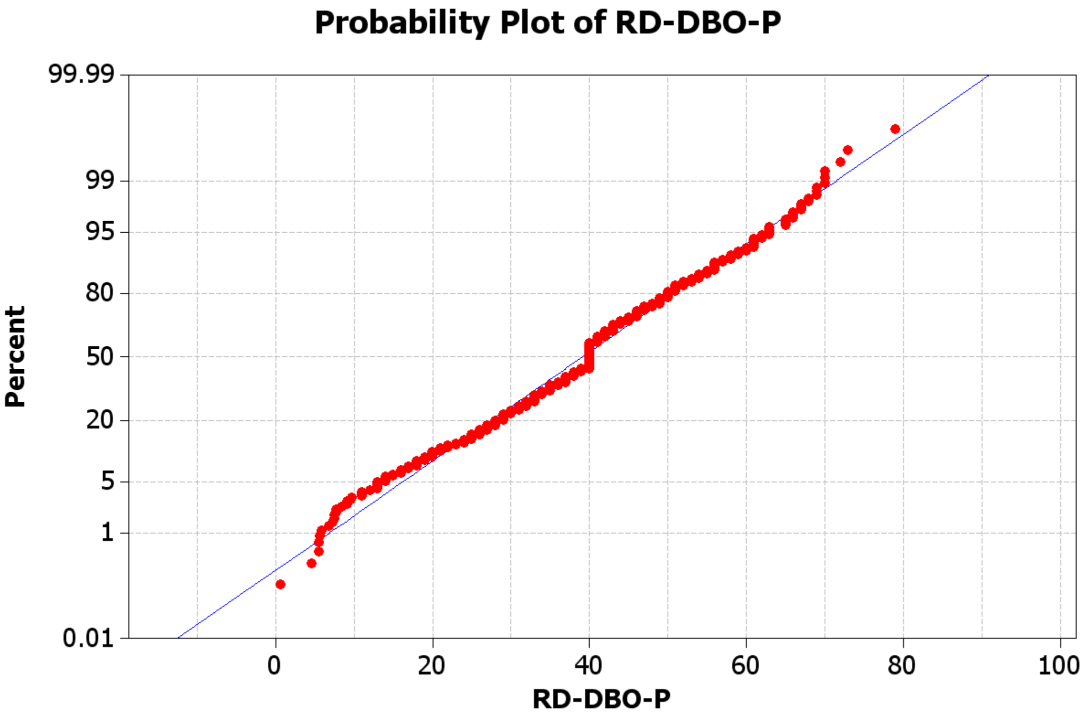


Figure A-16: Probability plot of RD-DBO-P showing non-Gaussian distribution

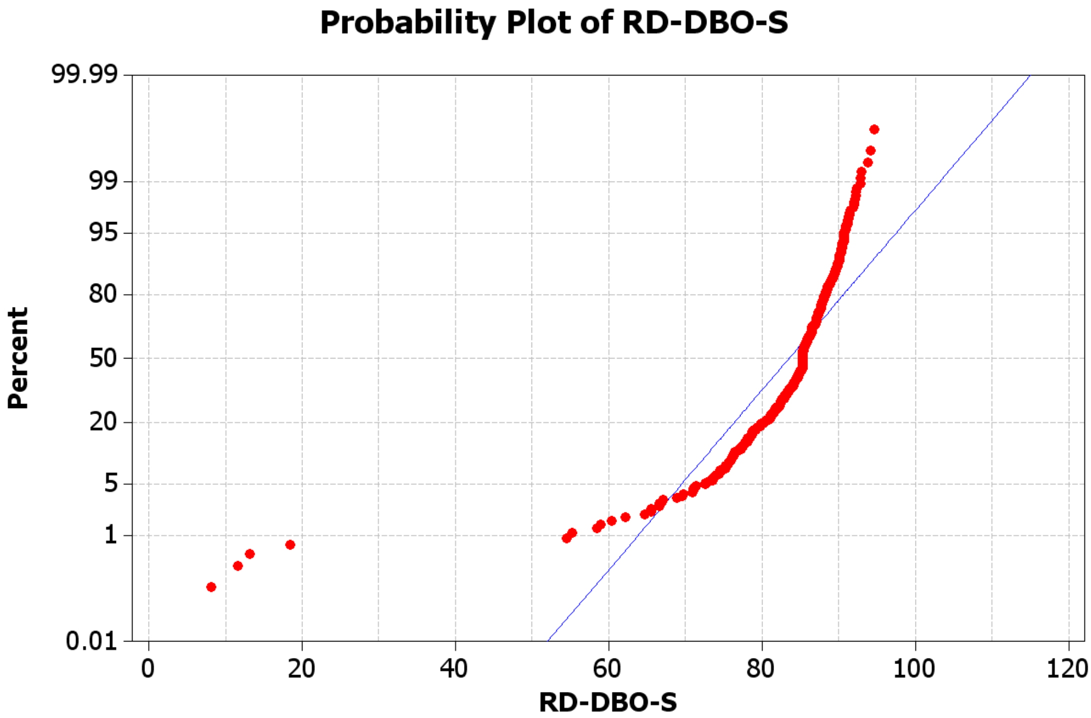


Figure A-17: Probability plot of RD-DBO-S showing non-Gaussian distribution

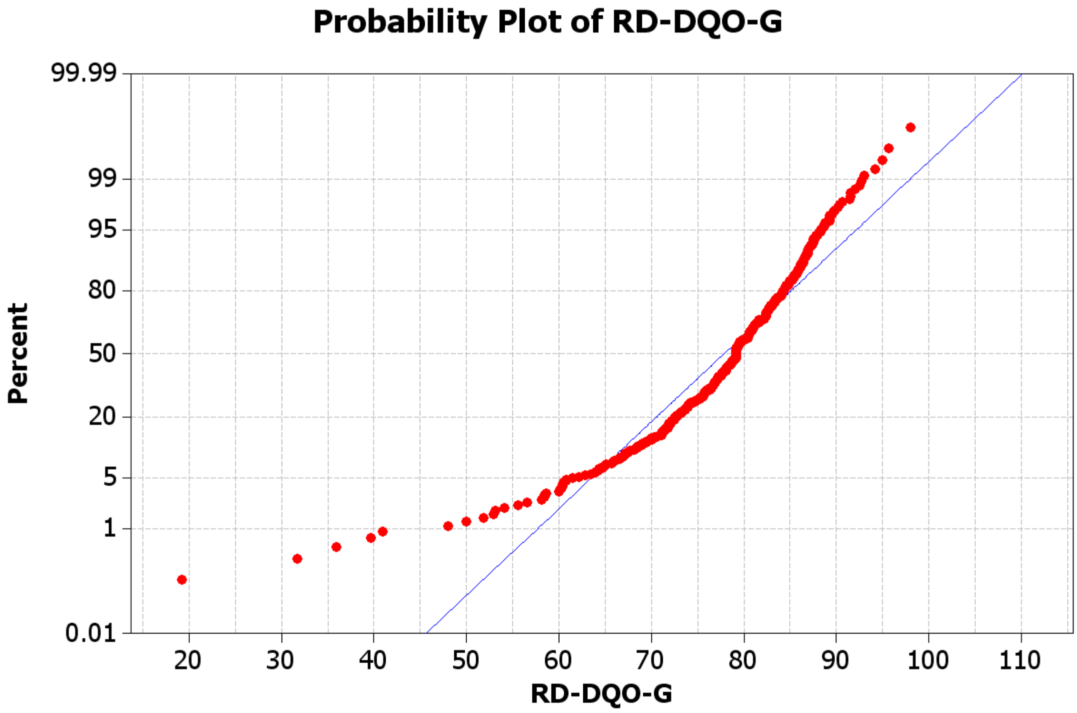


Figure A-18: Probability plot of RD-DQO-G showing non-Gaussian distribution

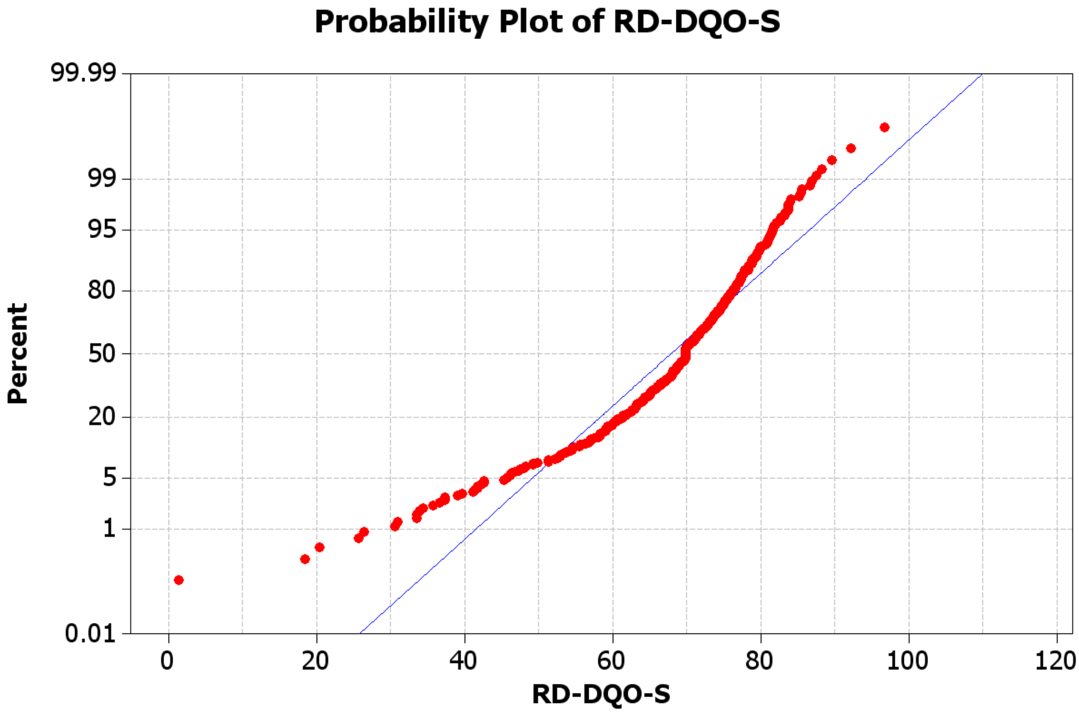


Figure A-19: Probability plot of RD-DQO-S showing non-Gaussian distribution

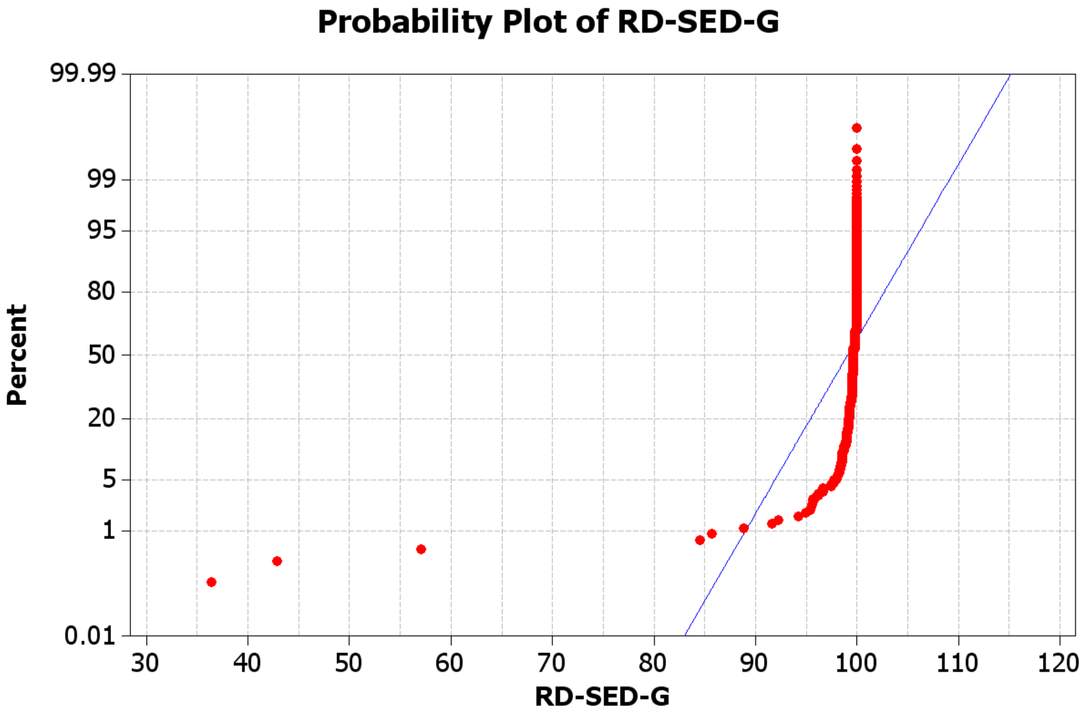


Figure A-20: Probability plot of RD-SED-G showing non-Gaussian distribution

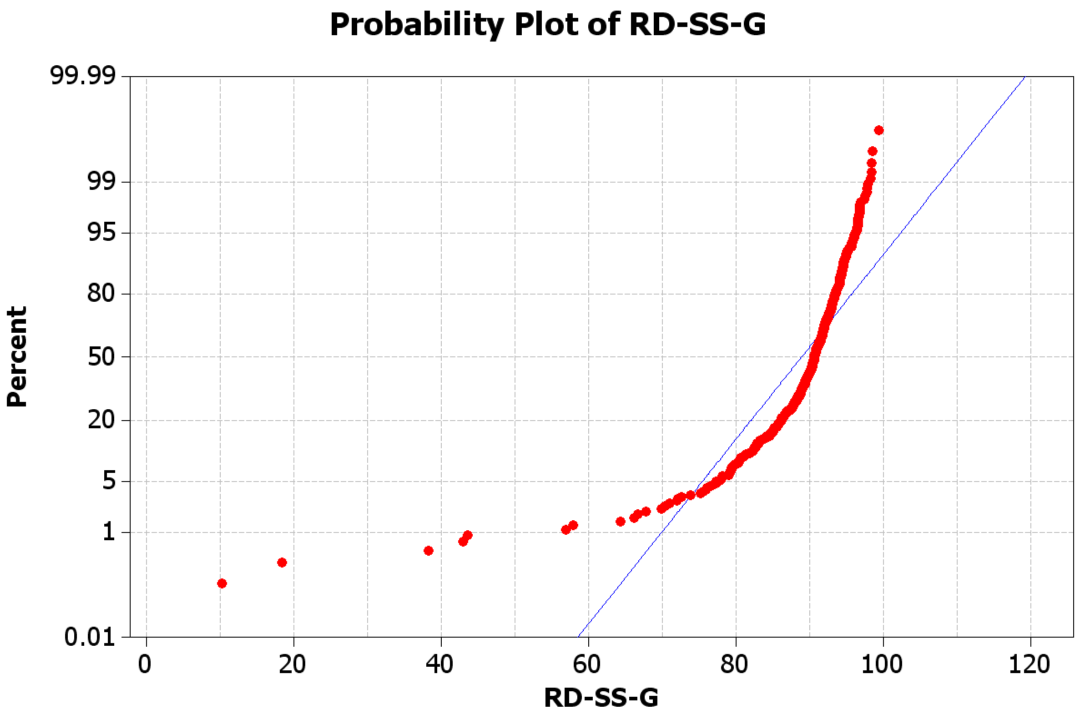


Figure A-21: Probability plot of RD-SS-G showing non-Gaussian distribution

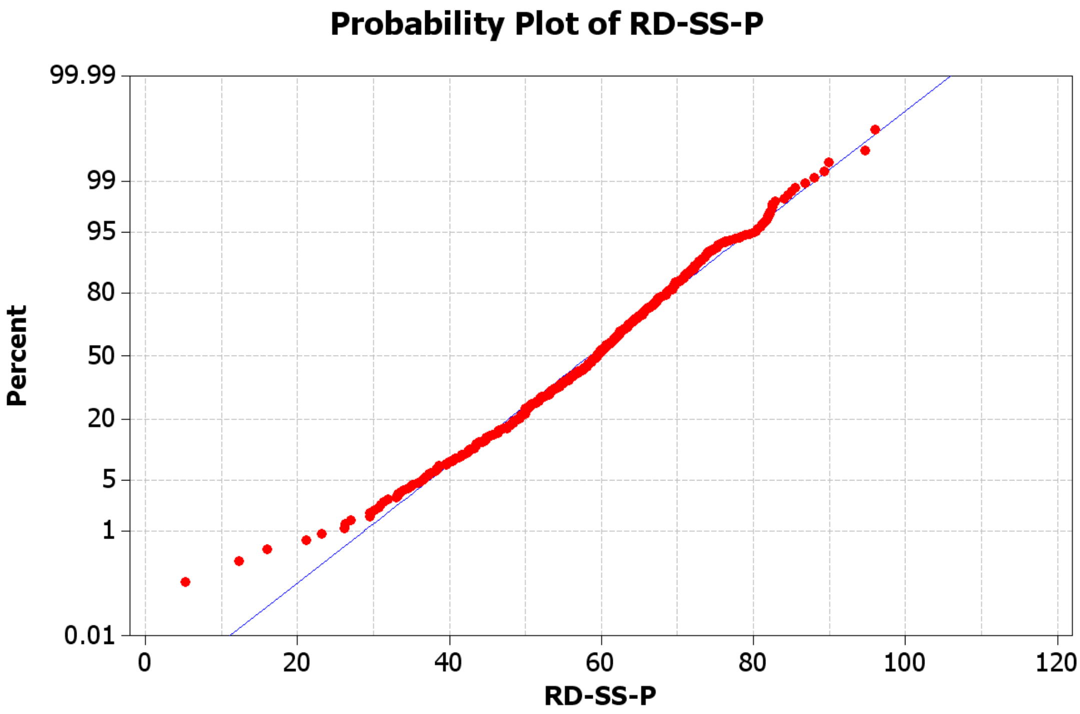


Figure A-22: Probability plot of RD-SS-P showing Gaussian distribution

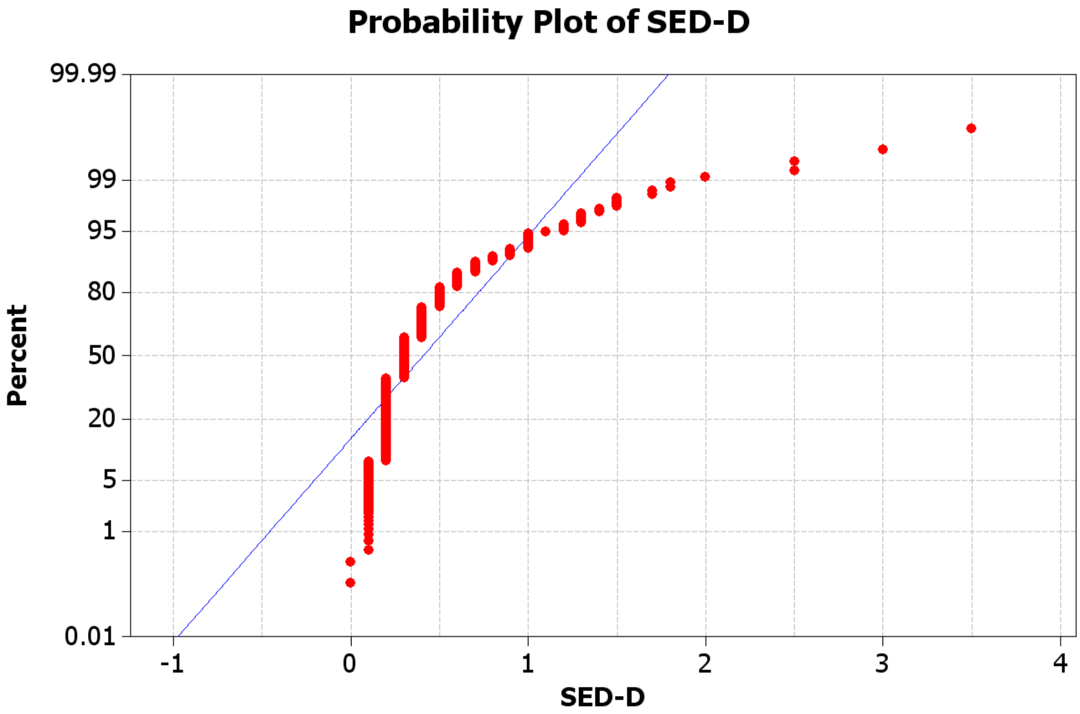


Figure A-23: Probability plot of SED-D showing non-Gaussian distribution

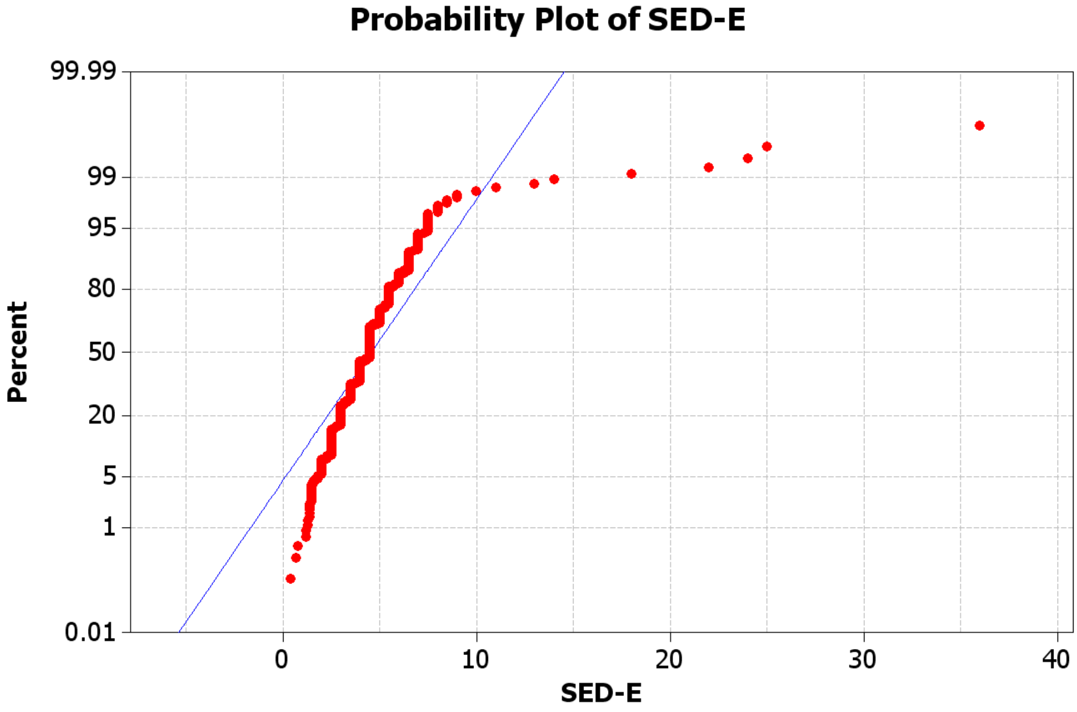


Figure A-24: Probability plot of SED-E showing non-Gaussian distribution

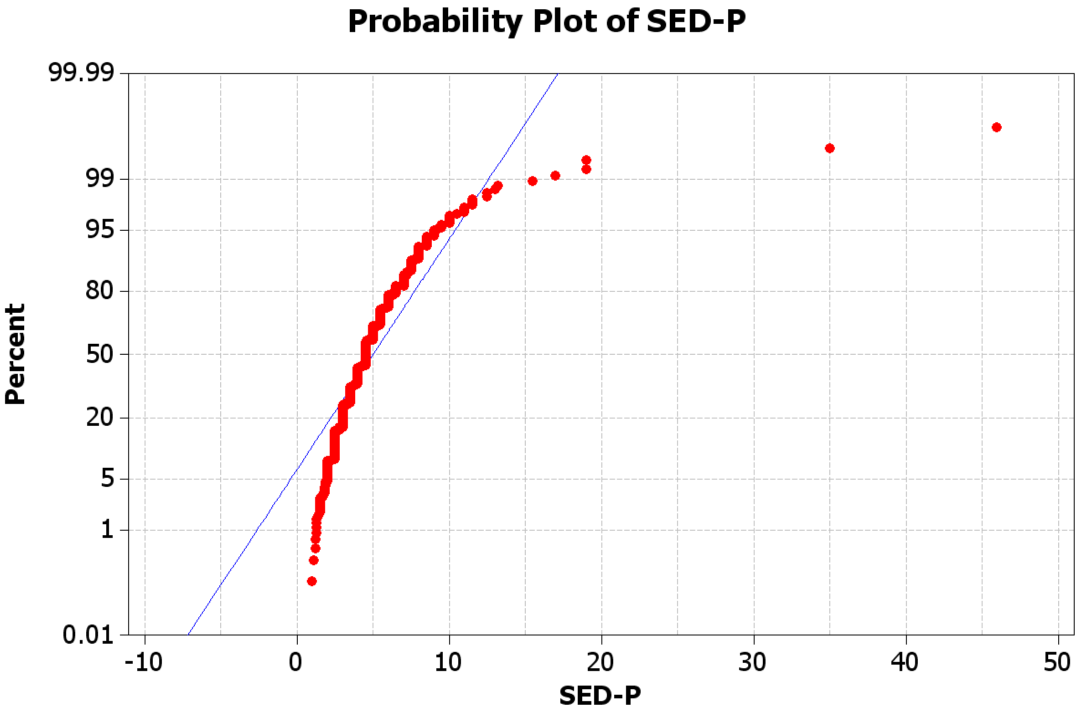


Figure A-25: Probability plot of SED-P showing Gaussian distribution

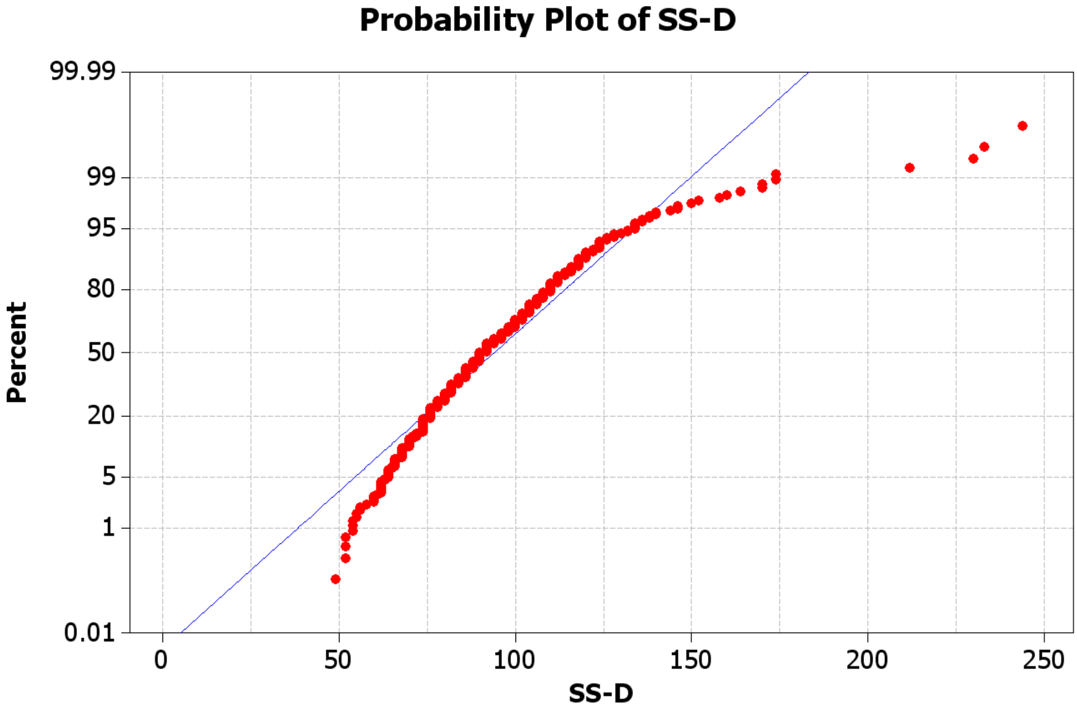


Figure A-26: Probability plot of SS-D showing Gaussian distribution

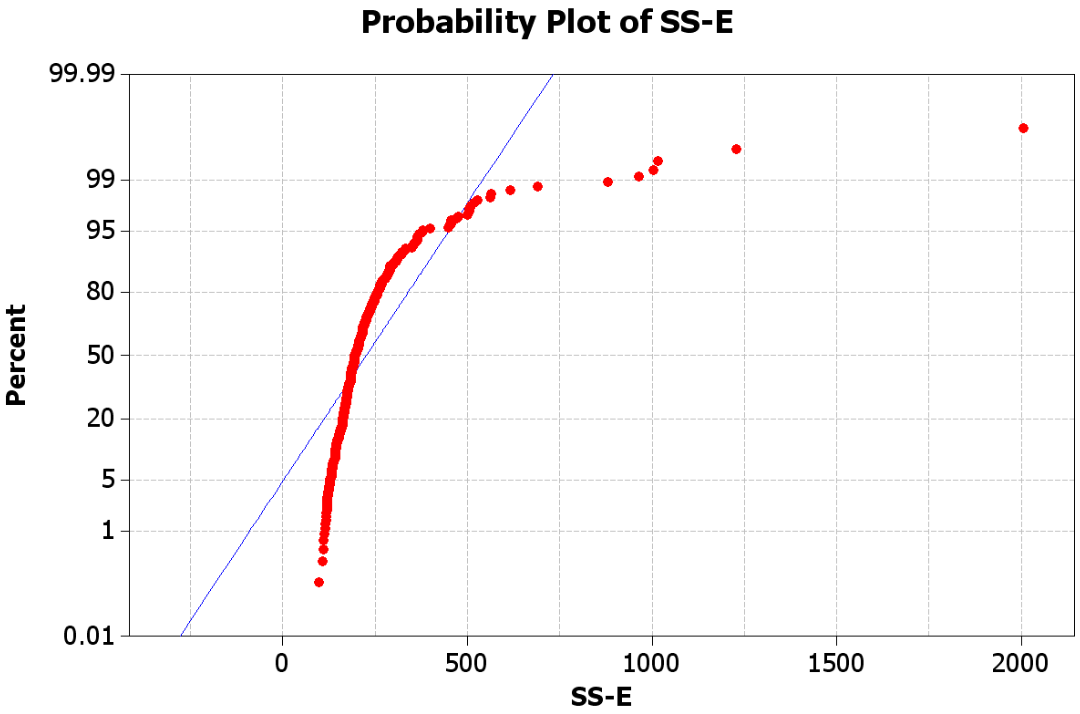


Figure A-27: Probability plot of SS-E showing non-Gaussian distribution

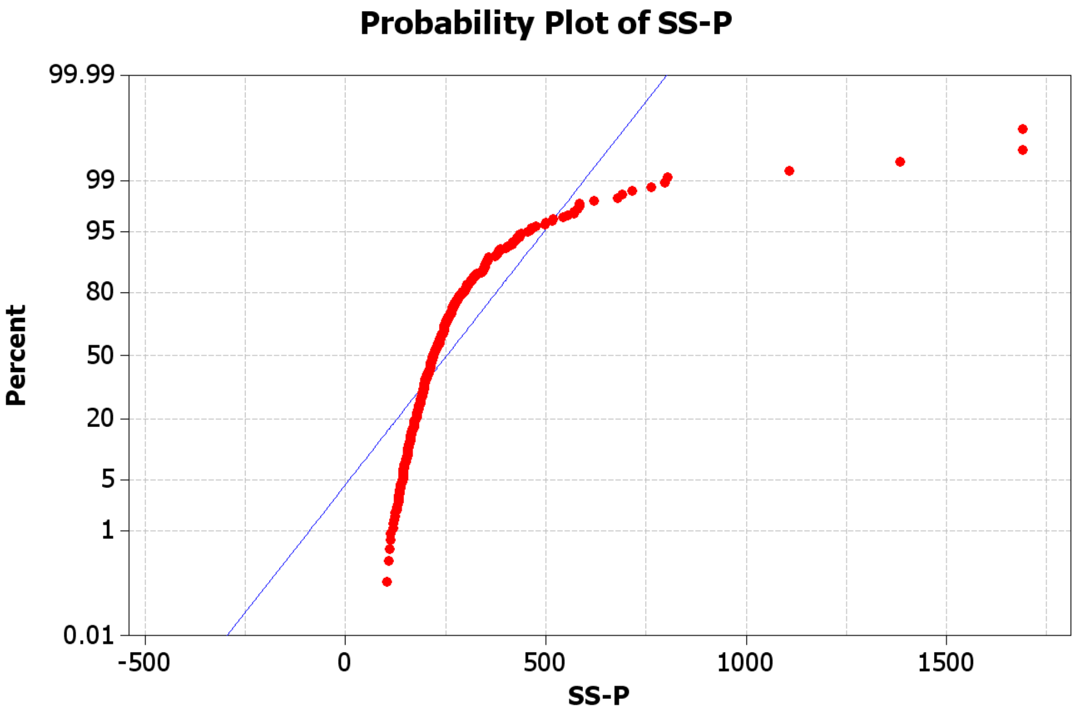


Figure A-28: Probability plot of SS-P showing non-Gaussian distribution

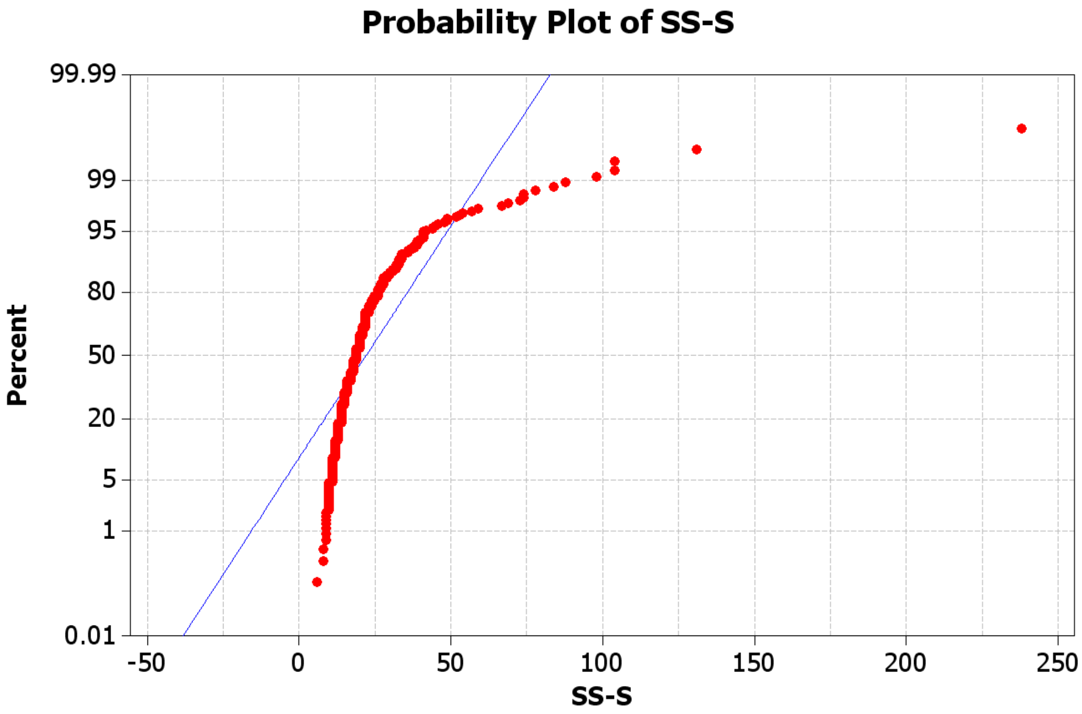


Figure A-29: Probability plot of SS-S showing non-Gaussian distribution

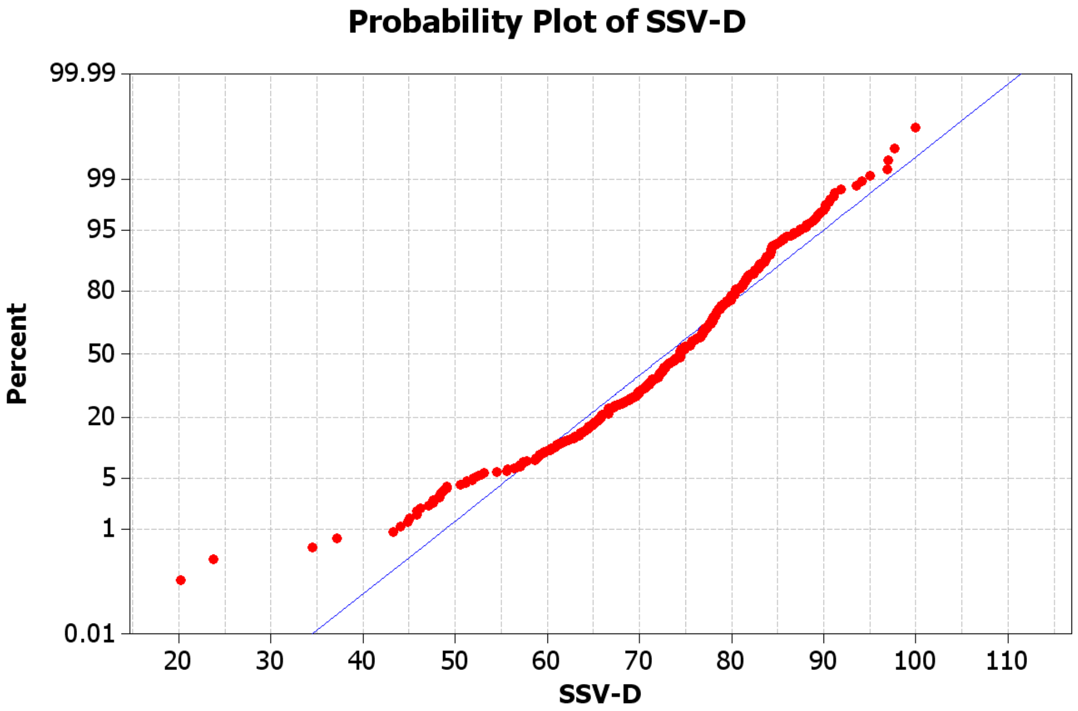


Figure A-30: Probability plot of SSV-D showing Gaussian distribution

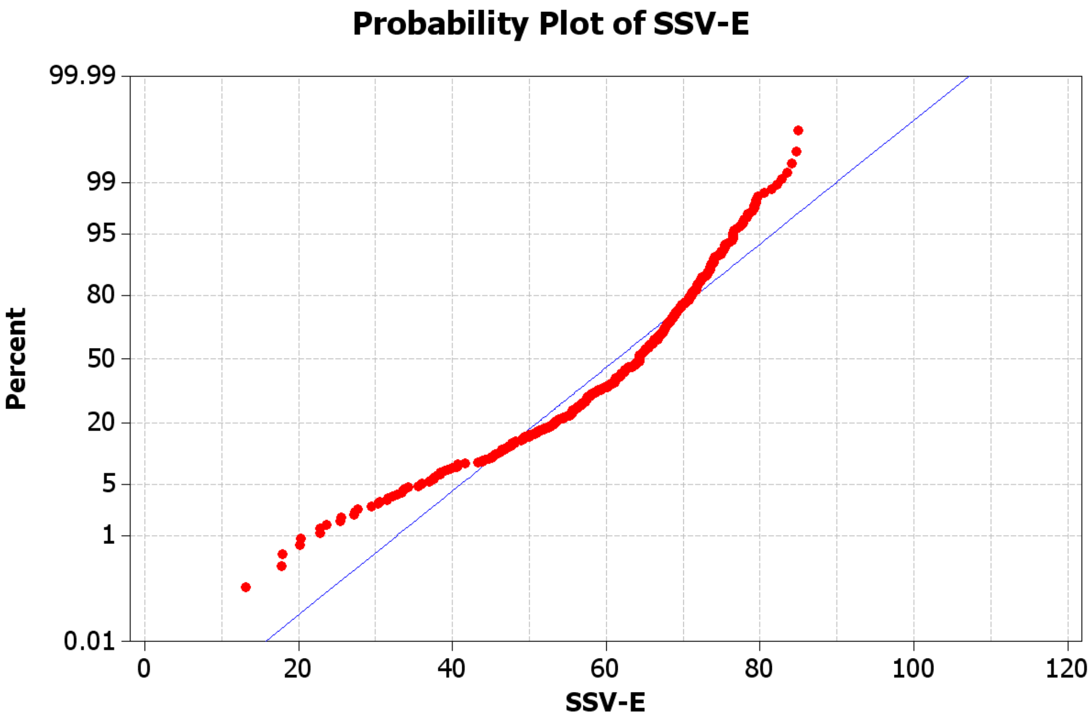


Figure A-31: Probability plot of SSV-E showing Gaussian distribution

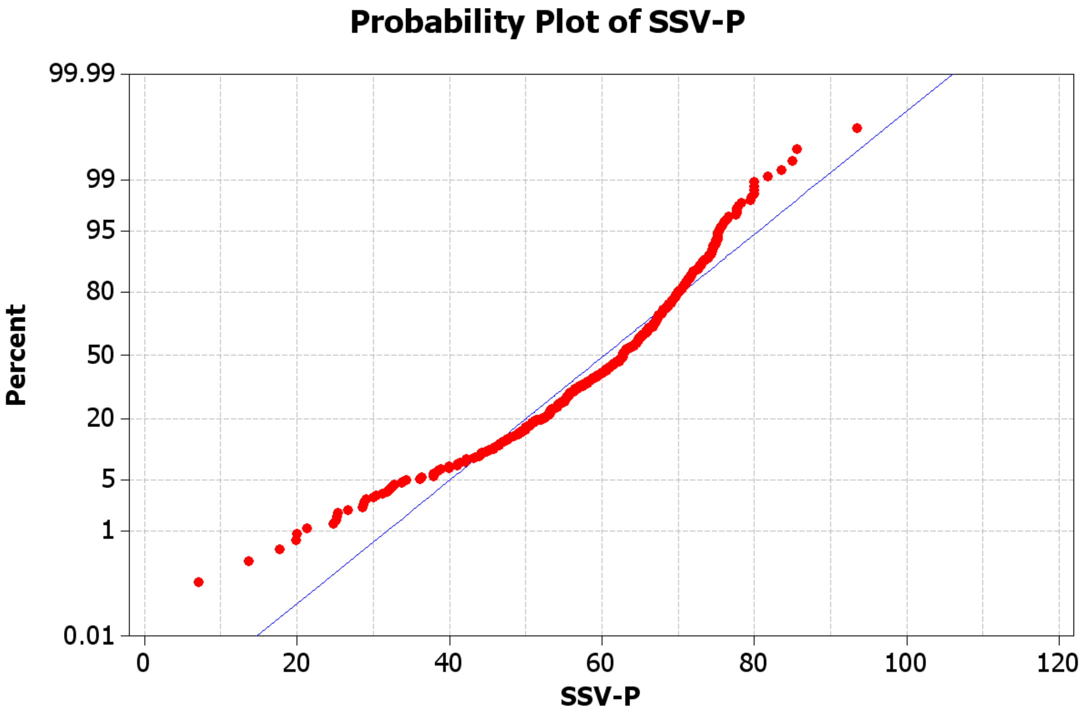


Figure A-32: Probability plot of SSV-P showing Gaussian distribution

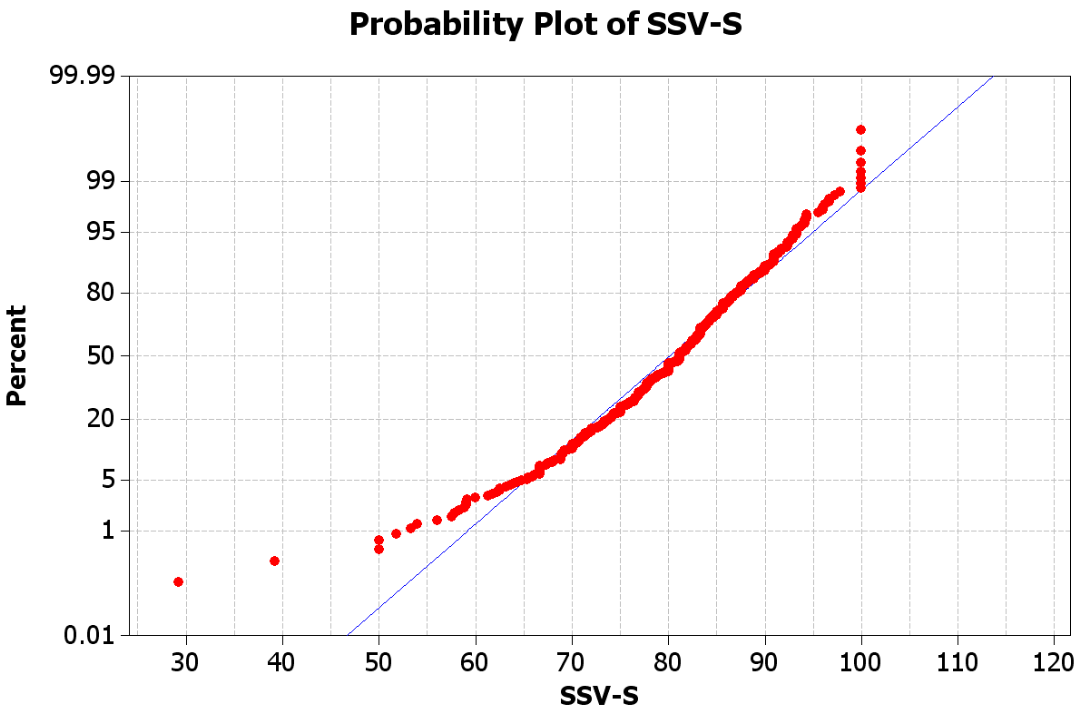


Figure A-33: Probability plot of SSV-S showing Gaussian distribution

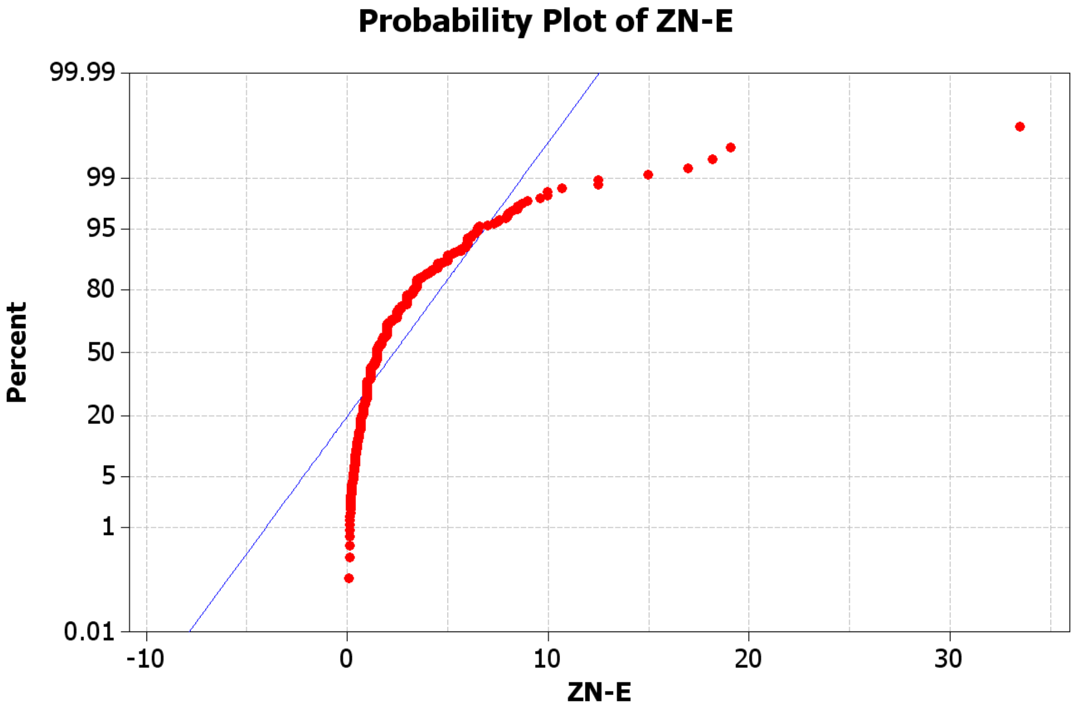


Figure A-34: Probability plot of ZN-E showing non-Gaussian distribution

APPENDIX B

This appendix shows the Statistical analysis of Air Pollution data to find out Gaussian or non-Gaussian nature of the variables.

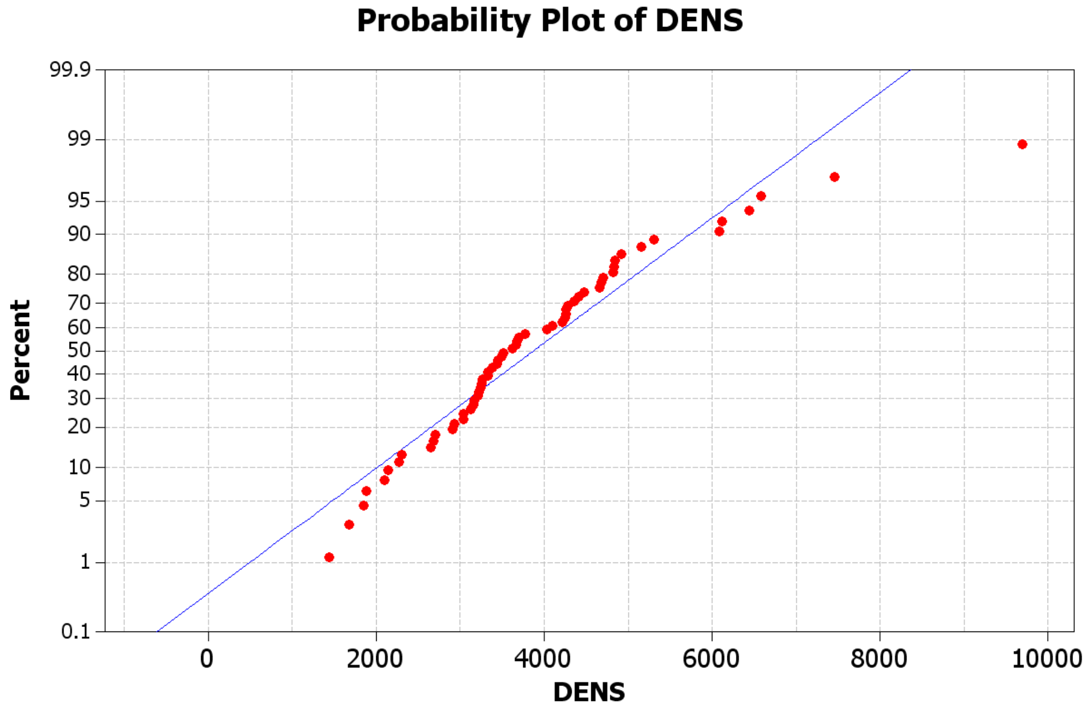


Figure B-1: Probability plot of DENS showing Gaussian distribution

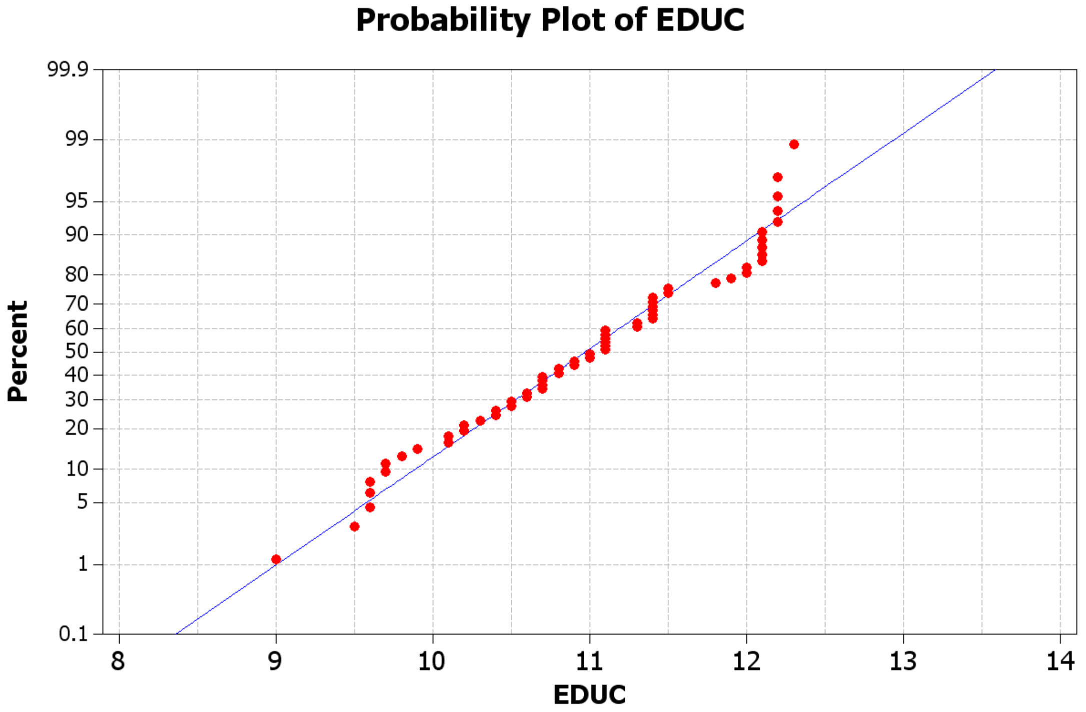


Figure B-2: Probability plot of EDUC showing non-Gaussian distribution

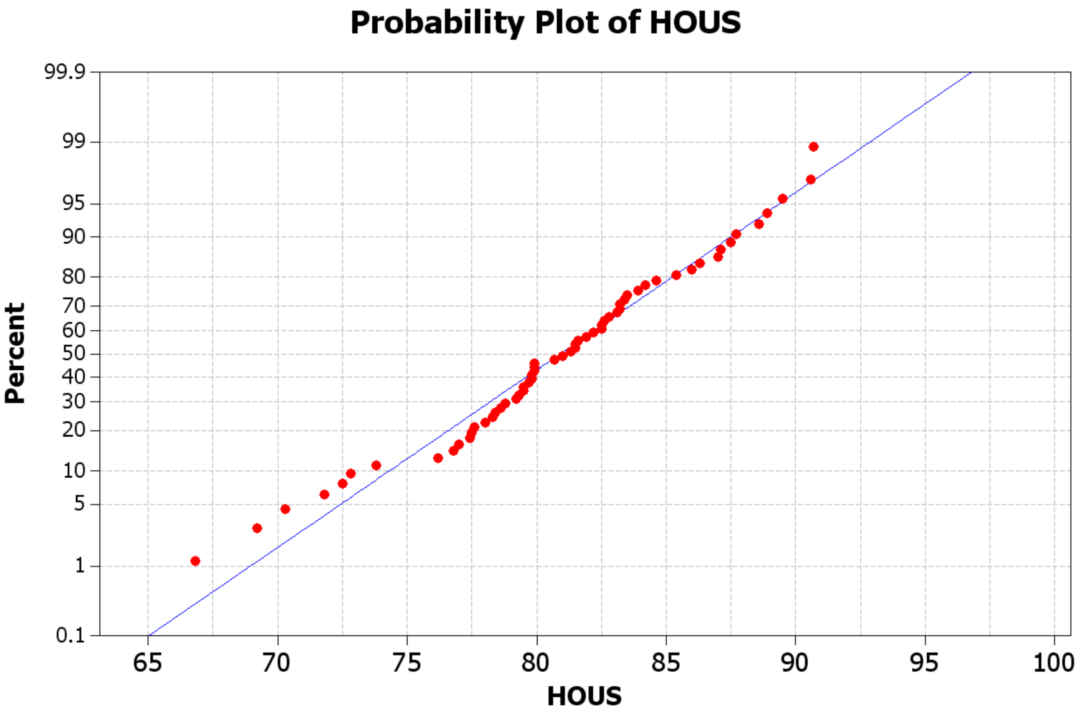


Figure B-3: Probability plot of HOUS showing Gaussian distribution

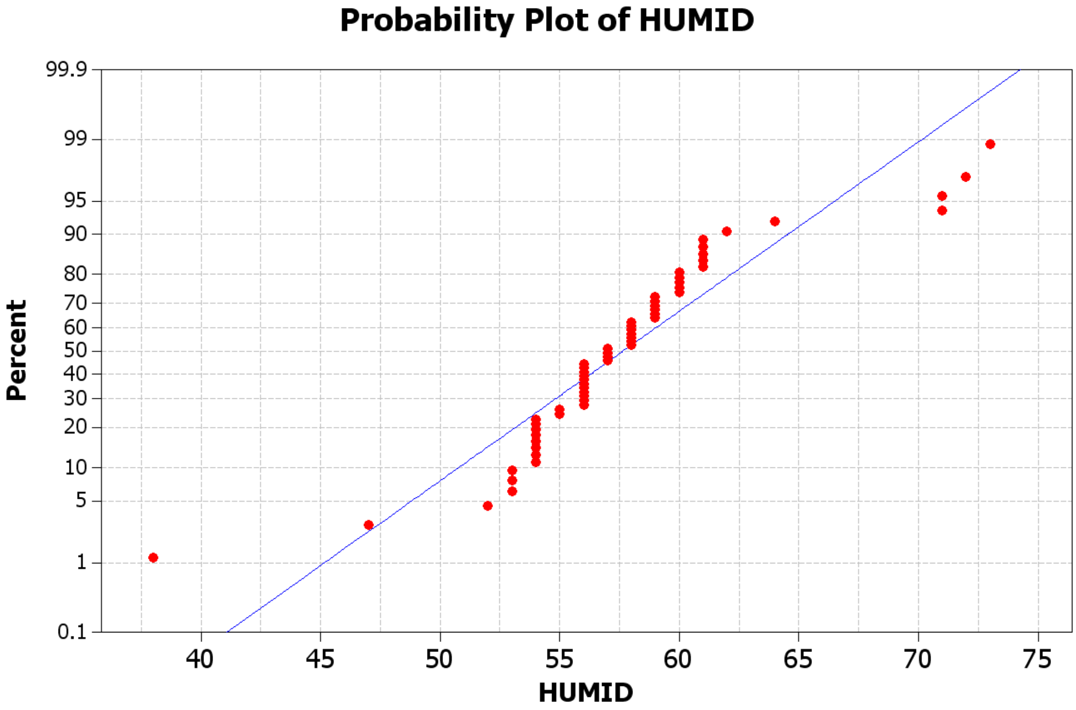


Figure B-4: Probability plot of HUMID showing non-Gaussian distribution

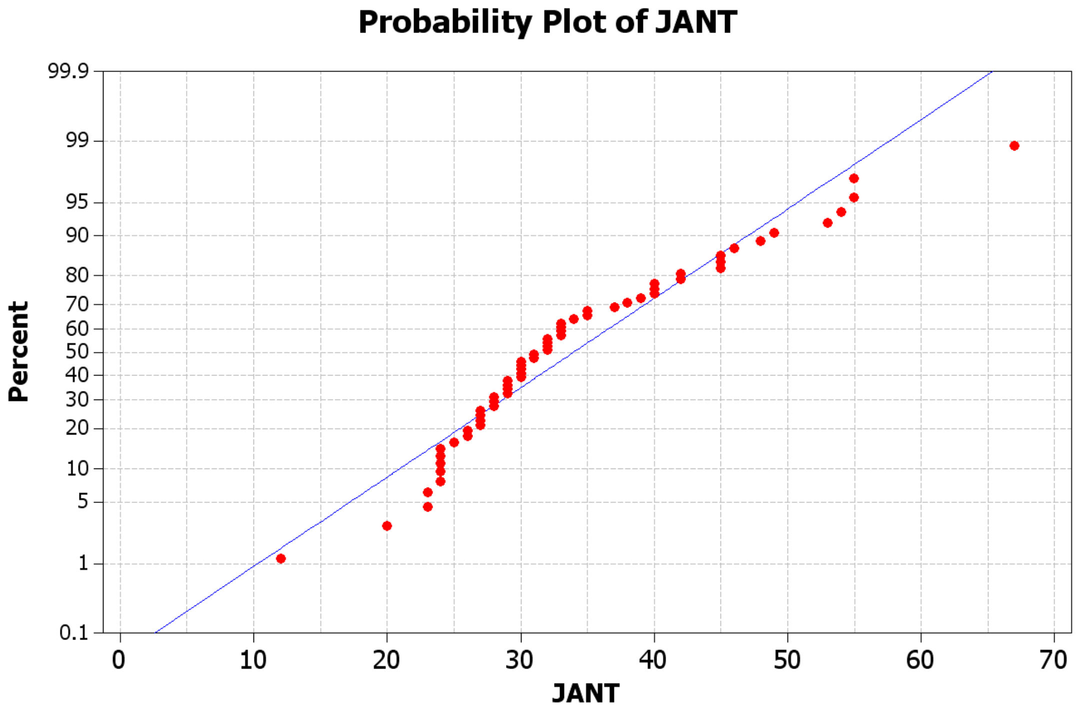


Figure B-5: Probability plot of JANT showing non-Gaussian distribution

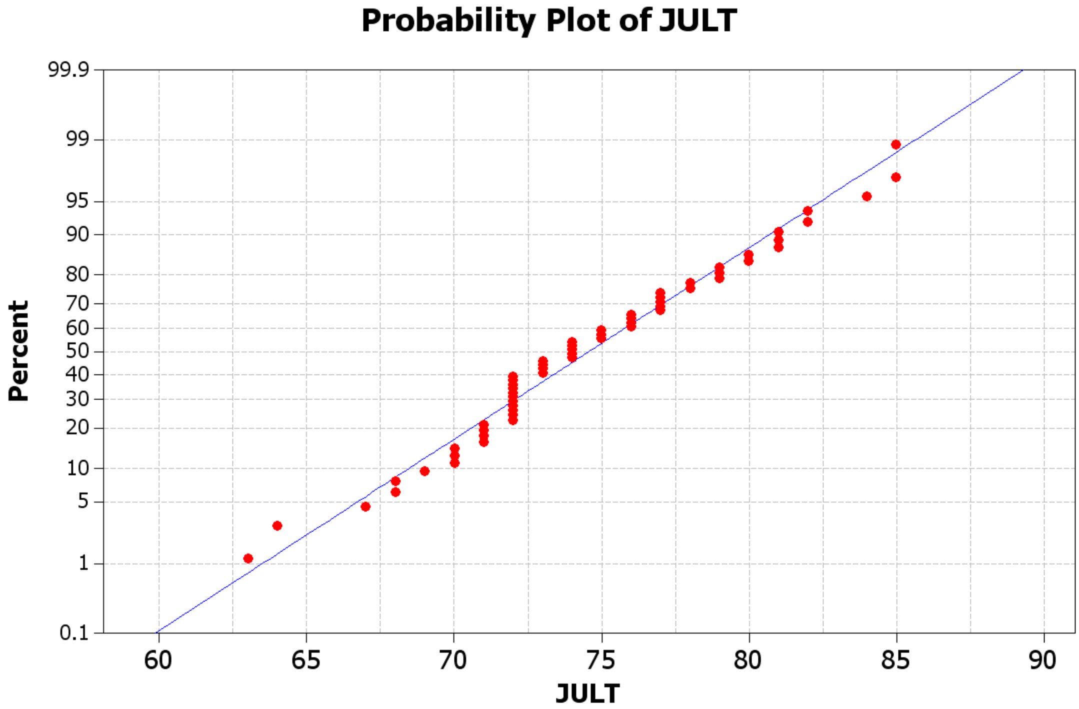


Figure B-6: Probability plot of JULT showing non-Gaussian distribution

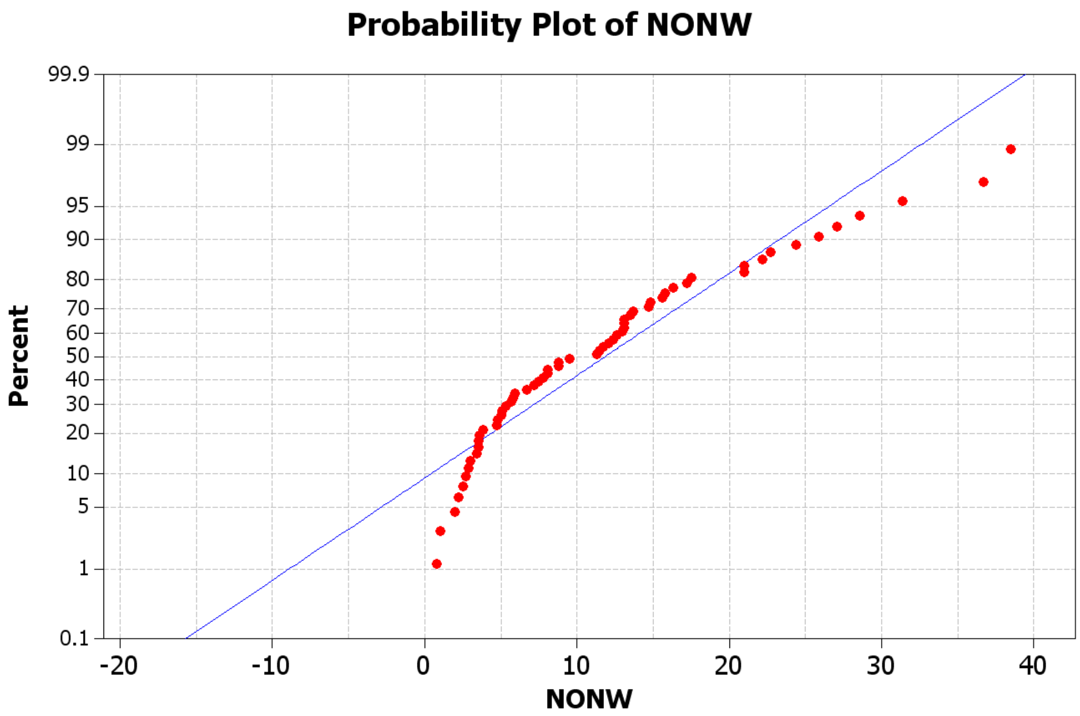


Figure B-7: Probability plot of NONW showing Gaussian distribution

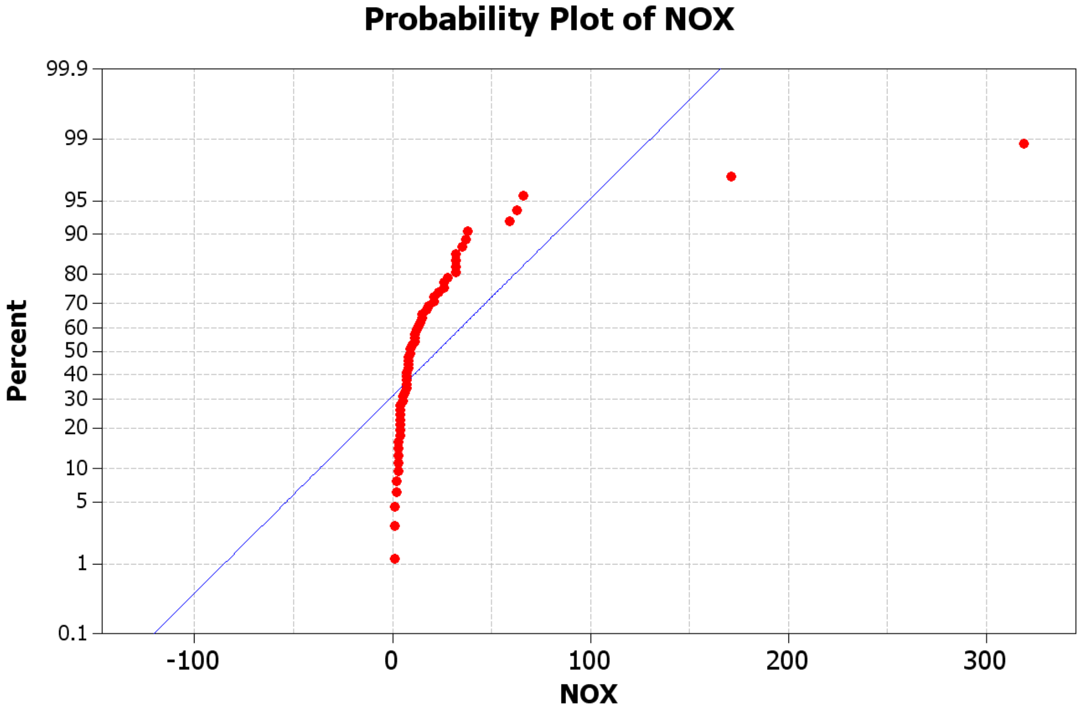


Figure B-8: Probability plot of NOX showing non-Gaussian distribution

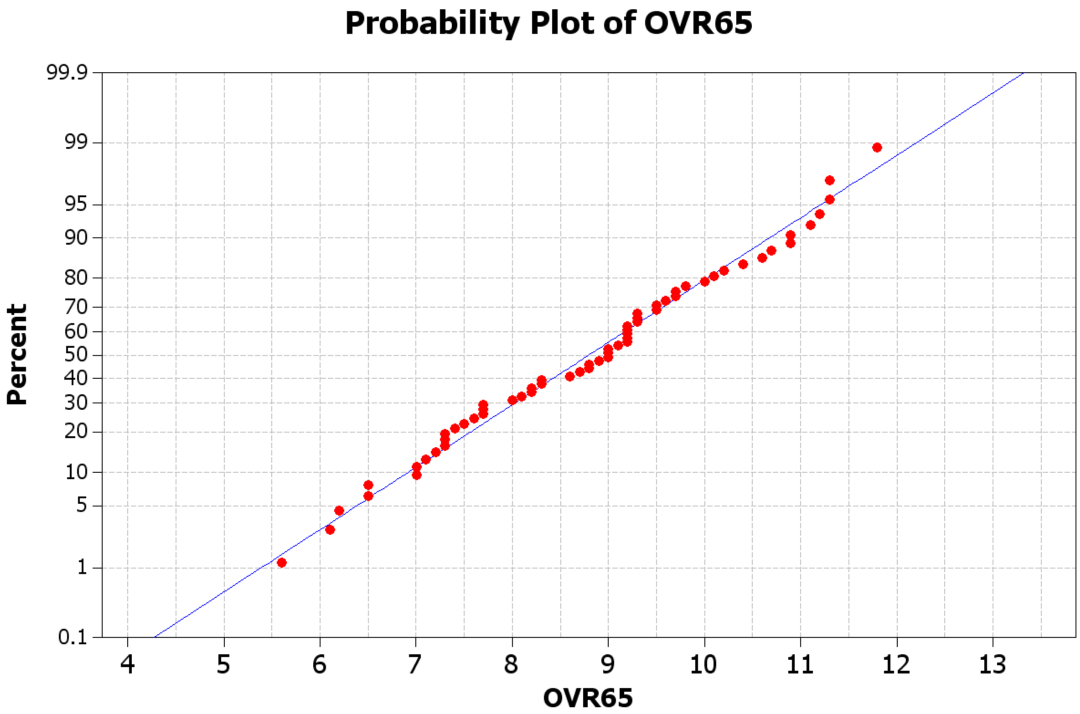


Figure B-9: Probability plot of OVR65 showing Gaussian distribution

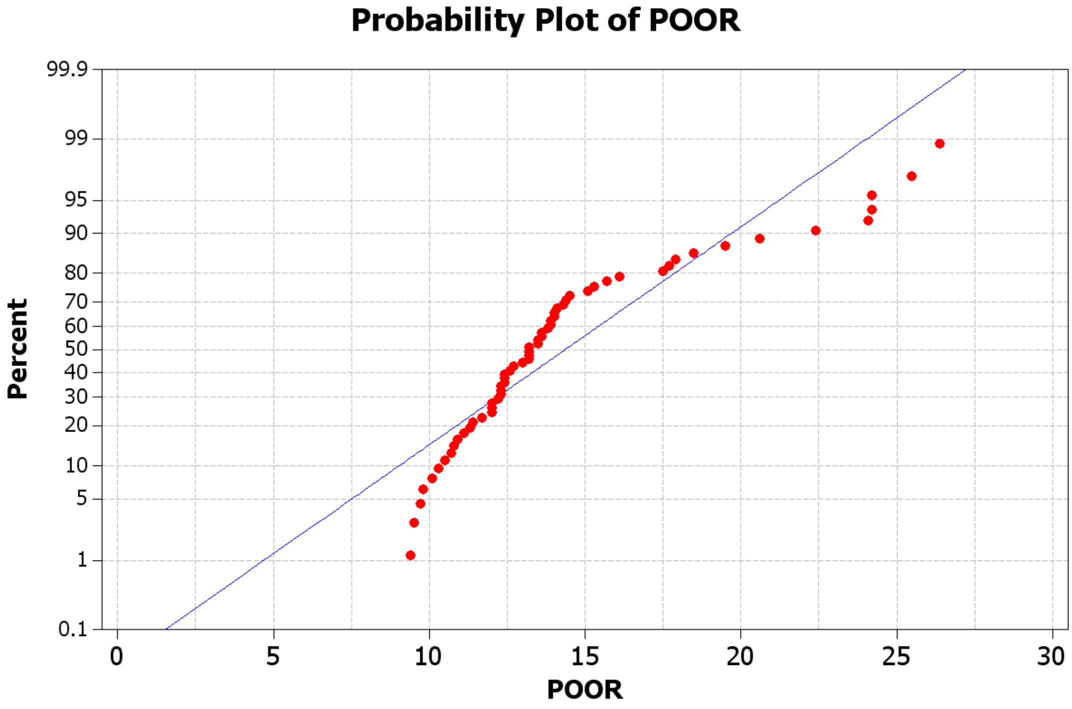


Figure B-10: Probability plot of POOR showing non-Gaussian distribution

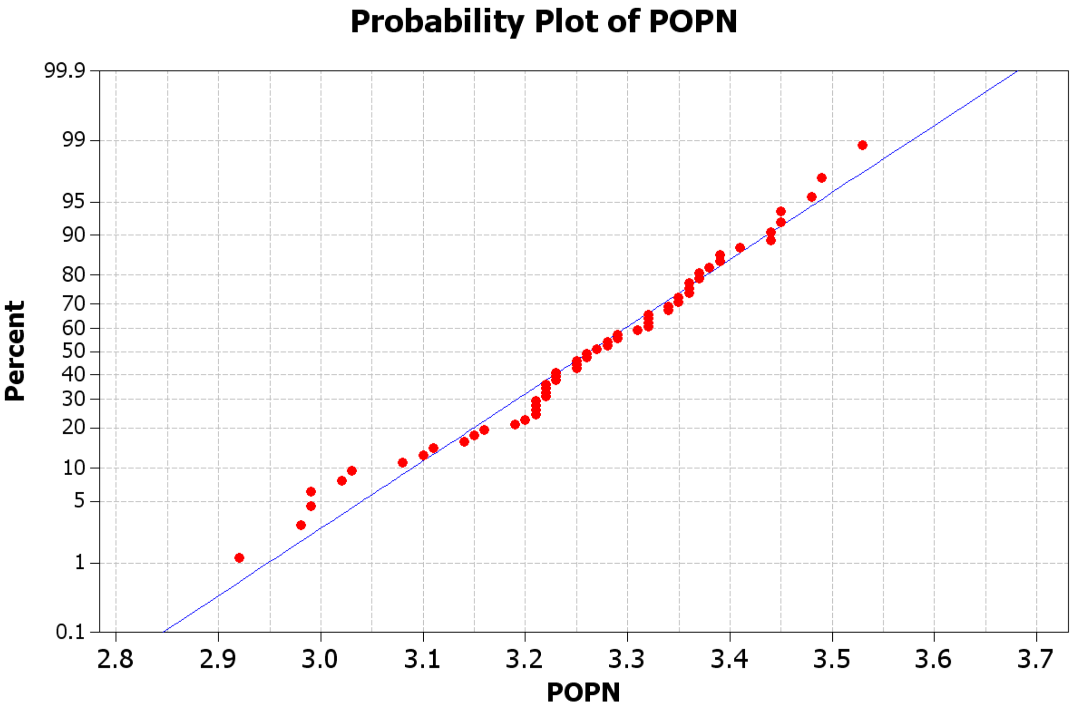


Figure B-11: Probability plot of POPN showing Gaussian distribution

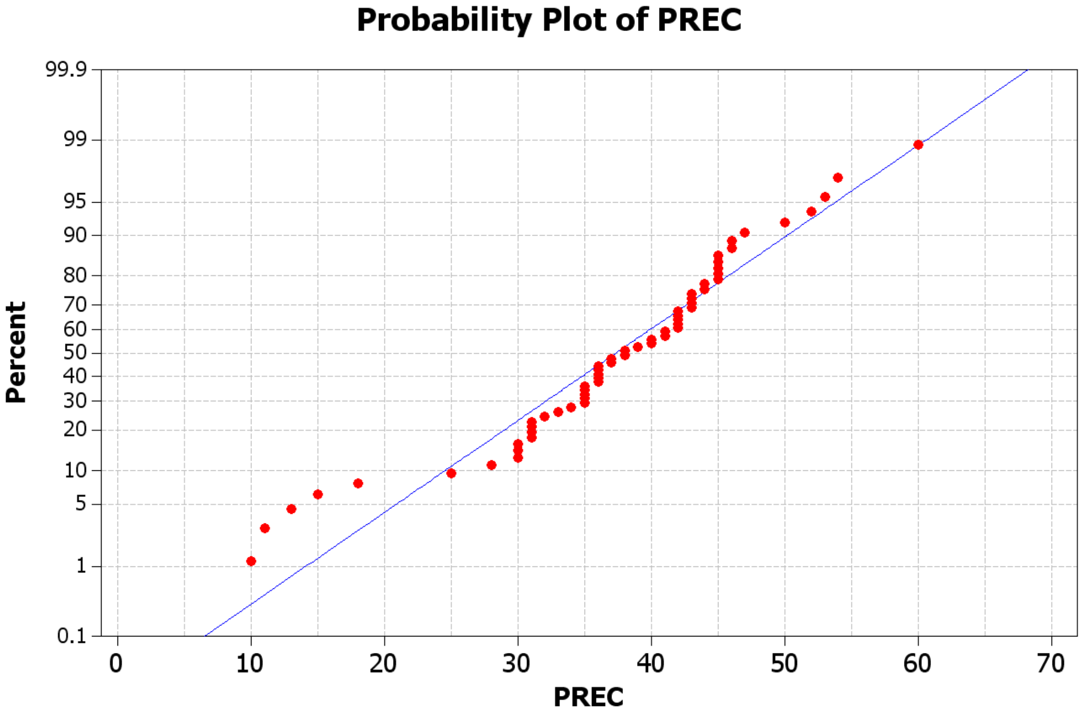


Figure B-12: Probability plot of PREC showing non-Gaussian distribution

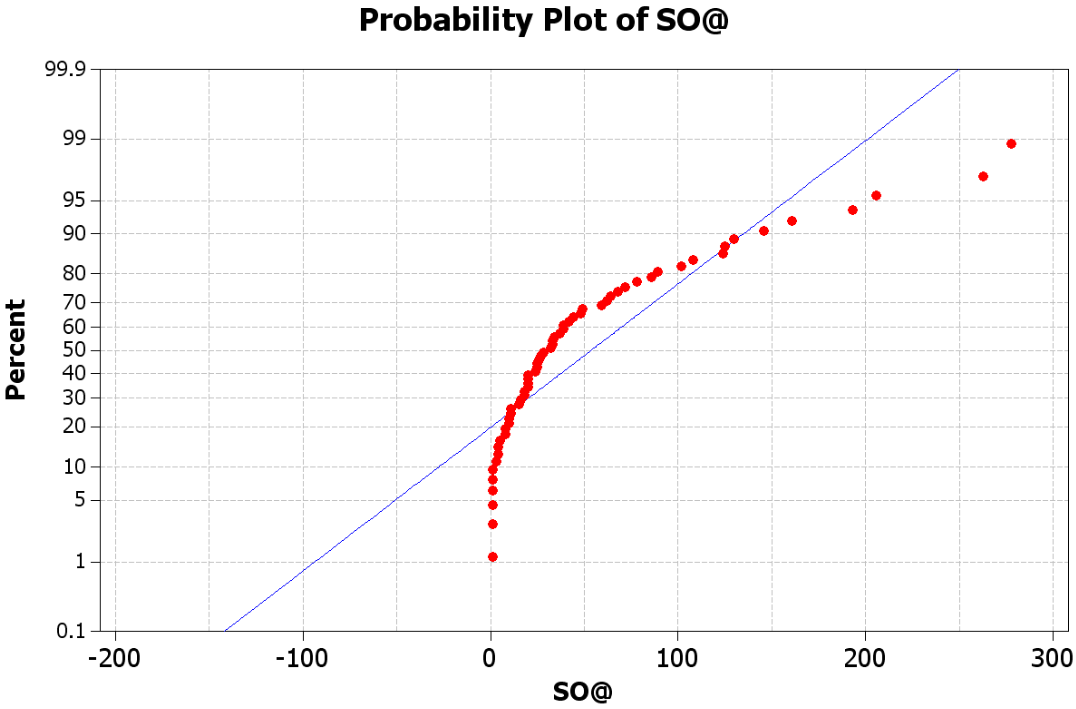


Figure B-13: Probability plot of SO@ showing non-Gaussian distribution

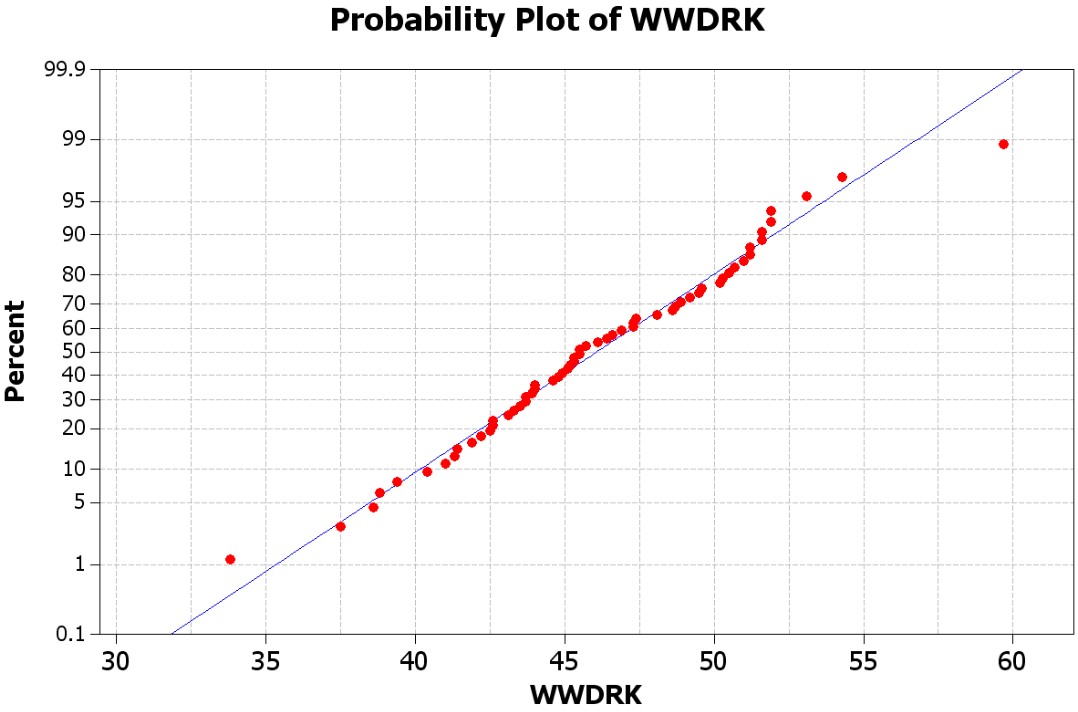


Figure B-14: Probability plot of WWDRK showing Gaussian distribution

APPENDIX C

This appendix gives an idea about the functions used for the determination of HICA. The M-File both for PCA and ICA can be accessed from literature i.e. MATLAB (2009) and HyvÃ¤rinen and Oja (1997).

```
function[H]= HICA(X)

[EV,latent]=princomp(X);

% "X" is the data set.
 % "EV" represents Eigen Vectors of the data set. Also "latent" is the Eigen values of the data set.

 [E, D] = pcamat(B);

% "B" represents the original parameter in the data set."
% "E" represents the Eigenvector of the parameter "B".
% "D" represents the diagonal Eigen value of the parameter "B".

  [nv, wm, dwm] = whitenv(X, E, D);

 % "nv" represents the Independent Components of the parameter.
% "wm" represents the Whitening matrix of the parameter.
% "dwm" represents the De-whitening matrix of the parameter.

  [A, W] = fpica(nv, wm, dwm);

% "W" is the estimated separating matrix.
% "A" is the corresponding mixing matrix.

 H=nv* EV;

% "H" is the matrix value of the HICAs.
```