The Development of Multivariate Statistical Process Monitoring (MSPM) Tools using Microsoft® Excel

by

Mohd Syaufi Bin Che Elliaziz

Project Dissertation submitted in partial fulfillment of
the requirements for the
Bachelor of Engineering (Hons)
(Chemical Engineering)

JANUARY 2009

Universiti Teknologi PETRONAS
Bandar Seri Iskandar
31750 Tronoh
Perak Darul Ridzuan

CERTIFICATION OF APPROVAL

The Development of Multivariate Statistical Process Monitoring (MSPM) Tools Using Microsoft® Excel

by

Mohd Syaufi Bin Che Elliaziz

A project dissertation submitted to the

Chemical Engineering Programme

Universiti Teknologi PETRONAS

in partial fulfilment of the requirement for the

BACHELOR OF ENGINEERING (Hons)

(CHEMICAL ENGINEERING)

Approved by,

(Ir Dr. Abdul Halim Shah Bin Maulud)

UNIVERSITI TEKNOLOGI PETRONAS

TRONOH, PERAK

January 2009

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.

(MOHD SYAUFI BIN CHE ELLIAZIZ)

ABSTRACT

Multivariate statistical process control methods have been proven in the process industries to be an effective tool for process monitoring, modelling and fault detection.This paper describes the approach used by the writer in the development of a Multivariate Statistical Process Monitoring (MSPM) tools using Microsoft Excel. This developed MSPM tools will act as a process monitoring tools in order to monitor the performance of any equipment or process. In addition, this project will be testing on actual plant data to see the performance of the project. The tool will be developed in Microsoft Excel and Matlab. Microsoft Excel is chosen because of it is easy to use and user-friendly. Furthermore, it has macro function and easier to use when the user wants to develop many tools to the Microsoft Excel. In multivariate statistical process monitoring, a process monitoring model must be developed firstly. The model must be free from any abnormality, fault or outliers. Then the model will be tested on the future data to detect any abnormality in the process by applying the appropriate limits. As a conclusion, the MSPM method can be develop in Microsoft Excel. This tool can help to detect the problem or abnormality of the process and help in diagnoss assignable cause for the process

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

First and foremost, the group would like to thank God for constantly strengthen and enlighten the author every step through the project and make all things possible in the end.

Throughout the progress of Final Year Project, the author would like to acknowledge his family for always being there to provide moral support to him.

Next, the author's utmost gratitude goes to the author's supervisor, Ir. Dr. Abdul Halim Shah Maulud. Without his guidance and patience, the author would not be succeeded to complete the project. To the Final Year Project Coordinator, Mr Tazli Azizan for provide her with all the initial information required to begin the project.

To the entire technician in Chemical Engineering, thank you for assisting the author in completing her project.

To all individuals that has helped the author in any way, but whose name is not mentioned here, the author thank you all.

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

1.1.1 Background

In today's competitive oil and gas industry, the pressure to improve the performance of processing facilities is intense. The advent of modern process measurement, automation, and information systems has resulted in a significant in amount of process data available. Unfortunately, it is often very difficult to monitor such a large amount of data. Multivariate Statistical Process Control (MSPC) methods, and Principal Component Analysis (PCA), have been demonstrated to provide a powerful approach for detection and isolation of abnormal conditions.

Multivariate Statistical Process Control (MSPC) concept and method has become significant in manufacturing and process industrial to control the process. Of these techniques, MSPC methods have been demonstrated to provide a powerful approach for detection and isolation of abnormal conditions. To perform this method, it's required an expensive commercial software or research computing software (e.g Matlab) to process the data. In this project, MSPC will be develop in Microsoft Excel in such the software can be widely used and shared with Microsoft Excel platform.

1.2 Objective and Scope of Study

1.2.1 Objective

The objectives of this study are stated below:-

1. To develop the multivariate statistical process monitoring tool by using Microsoft Excel.
2. To monitor and analysis the performance using the developed Multivariate Statistical Process Control method.
3. To test the developed software using actual plant data

1.2.2 Scope of Study

The project would concentrate on development of monitoring tools based on multivariate statistical method.

1. To study about the fundamental concept of Multivariate Statistical Process Monitoring.
2. To learn more about the software uses in develop the monitoring tools.

CHAPTER 2

LITERATURE REVIEW

2.1 Theory

2.1.1 Statistical Process Control Chart

Statistical process control (SPC) involves using statistical techniques to measure and analyze the variation in processes. Most often used for manufacturing processes and process industries, the intent of SPC is to monitor product quality and maintain processes to fixed targets. Statistical quality control refers to using statistical techniques for measuring and improving the quality of processes. Their objective is to monitoring the performance of a process over time in order to verify that the process is remaining in a "state of statistical control". Such a state of control is said to exist if certain process or product variables remain close to their desired values and only source of variation is "common-cause" variation, that is, variation which affect the all process the time and is essentially unavoidable within the current process. (J.F MacGregor and T. Kourti, 1995).

Shewhart, CUSUM and EWMA charts which SPC chart used to monitor key product variables in order to detect the occurrence of any event having a "special" or "assignable" cause. SPC monitoring methods should be applied on top of the process and its automatic control system in order to detect process behavior that indicates and occurrence of a special event. By diagnosing cause for the event and removing, the process is improved.

Unfortunately, most SPC methods are based on charting only small number of variables, usually the final product quality variables(Y). Many industrial processes involve a set of input variables and quality variables, which are highly correlated. If one of the variable changes, it will affect the other correlated variables. Thus, ignoring the cross-correlation between the variables can lead to misinterpretation of the process behavior. (M.W. Yee and Kamarul A.I.). Therefore, it is very difficult to diagnosis and makes interpretation, as though the variables were independent. Such methods only look at the magnitude of the deviation in each variable independently to each others.

The multivariate method is the only way to treat all the data simultaneously and also extract information on the directionality of the process variations. In addition, when important events occur in progress they are often difficult to detect due to the signal to noise ratio is very low in each variable. But, the multivariate method can extract the information from observations on many variables and can reduce the noise level through averaging.

2.1.2 Multivariate method for monitoring product quality

In most cases, the traditional SPC charts (Shewhart, CUSUM and EWMA) are used to separately monitor key measurement on the final product which define the quality of the product. On this approach, the difficulty is to determine which one of the variables defines the product quality. The product quality only can be defined by correct simultaneous values of all the measured properties, that is, it is a multivariate property.

2.1.3 Multivariate method for process monitoring

The main approach of statistical quality control (SQC) method are only monitor the product quality data (Y) and all of the data on process variables (X) are being ignored.

To perform the SPC, all the data must be look and analyst. The process variables are much more frequently measured than the product quality data. Furthermore, any special event which occur will also have their fingerprints in these process data (J.F MacGregor and T. Kourti, 1995). It will use useful to know if the product is good before using it. Monitoring the process would help early in detection of poor-quality product.

The most practical approaches to multivariate SPC appear to be those based on multivariate statistical projection method such as PCA and PLS. the methods are ideal for handling the large number of highly correlated and noisy process variable measurement that being collected by process computer.

2.2 Principal Component Analysis (PCA)

Although there may be hundreds of plant variables that measured in any given process, there tend to be only a small number of underlying characteristics that actually drive the process. The purpose of PCA is to identify a new set of variables that reflect these characteristics. These new variables, termed scores or latent variables are linear combinations of the original process variables. The expectation is that there will be fewer scores than plant variables and therefore the plant can be monitored with much greater ease by simply analyzing these new variables.(A. AlGhazzawi and B. Lennox, 2007).

Principal component analysis (PCA) was first introduced by Karl Pearson in the early 1900's. The other main advantage of PCA is that once have found these patterns in the data, and compress the data, i.e. by reducing the number of dimensions, without much loss of information.

In mathematical term, PCA decomposes the data matrix **X** of size [m,n]. Consider an m-dimensional data set

$$\mathbf{X} = [\ \mathbf{x}_1, \mathbf{x}_2, \ldots., \mathbf{x}_m\ ] \qquad (1)$$

The principal component decomposition of X can be defined as

$$X = TP^T + E = \sum_{i=1}^{i} t_i p_i^T + E \quad (i < \min(m,n)) \qquad (2)$$

Where n is number of samples,

$$\mathbf{T} = [\ t_1, t_2, \ldots., t_i\ ] \qquad (3)$$

Is the matrix of the principal component scores,

$$\mathbf{P} = [\ p_1, p_2, \ldots., p_i] \qquad (4)$$

Is the matrix of principal component loading and E is the residual matrix in the sense of minimum Euclidean norm and ***i*** is the number of significant component retained.

To monitor the process using a PCA model, a data set of representative normal process operation is used to identify a reference model. When the new data are available, it is projected onto this reference model according to

$$T_{new} = x_{new}\, P + e \qquad (5)$$

Where P is the loading matrix, and two complementary control charts are typically used to assess if the new data are consistent with that from the normal process condition: the Hotteling's T^2 and the Squared Prediction Error (SPE). The hotelling's T2 statistic will detect deviations within then model, where as the SPE statistic will detect deviations from the model. These two statistics will be proceeding for the next semester.

2.2.1 Using Standardized variables

Investigators frequently prefer to standardize the x variables prior to performing the principal component analysis. Standardization is achieved by dividing each variable by its sample standard deviation. This analysis is then equivalent to analyzing the correlation matrix instead of the covariance matrix.

2.3 The Control Chart

The control chart was invented by Walter A. Shewhart in the 1920s. The control chart, (also known as the 'Shewhart chart' or 'process-behaviour chart') is a tool used to determine whether a manufacturing process is in a state of statistical control or not. The figure 1 shows the example of control chart. There are many type of control chart such as X- chart, R-chart and S-chart. But for this, it uses the X-chart to detect the statistical in control or not.

A control chart consists of the following:

- Points representing measurements of a quality characteristic in samples taken from the process at different times [the data]
- A centre line, drawn at the process characteristic mean which is calculated from the data
- Upper and lower control limits that indicate the threshold at which the process output is considered statistically 'unlikely'

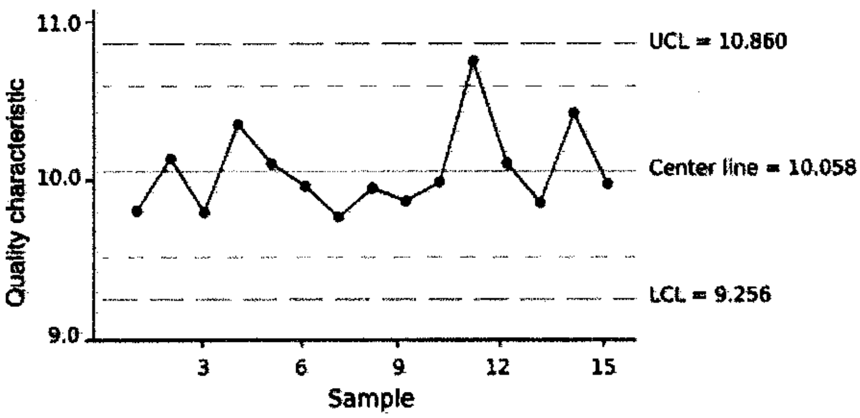

Figure 1: The Shewhart Chart or Process-Behaviour Chart or Control Chart

2.4 Limit

Natural extensions of the Shewhart chart to situations where one observes a vector of ***k*** variables y,, 1 at each time period are the multivariate x^2 and T^2 charts. Given a ***(k x*** 1) vector of measurements y on ***k*** normally distributed variables with an in-control covariance matrix 2 one can test whether the mean μ of these variables is at its desired target **τ** by computing the statistic.

$$x^2 = (y - \tau)^T \sum {}^{-1} (y - \tau) \qquad (6)$$

This statistic will be distributed as a central x2 distribution with ***k*** degrees of freedom if μ = **τ**. A multivariate x^2 control chart can be constructed by plotting x^2 vs. time with an upper control limit (UCL) given by $\chi^2_\alpha(k)$ where (Y is an appropriate level of significance for performing the test (e.g. α = 0.01).

Note that this multivariate test overcomes the difficulty .The x^2 statistic in Eq. (6) represents the directed or weighted distance (Mahalanobis distance) of any point from the target **τ**. All points lying on the ellipse would have the same value of x^2. (The ellipse is the solution to Eq. (8) for $x^2 = x_\alpha(k)$ for two variables). Hence, a x^2 chart would detect as a special event any point lying outside of the ellipse. When the in-control covariance matrix $\sum$ is not known, it must be estimated from a sample of n past multivariate observations as

$$S = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \overline{y})(y - \overline{y})^T \qquad (7)$$

When new multivariate observations (y) are obtained, then Hotelling's T^2 statistics given by

$$T^2 = (y - \tau)^T S^{-1}(y - \tau) \qquad (8)$$

can be plotted against time. An upper control limit (UCL) on this chart is given by:

$$T^2_{UCL} = \frac{(n-1)(n+1)k}{n(n-k)} F_\alpha(q, n-q) \qquad (9)$$

where $F_\alpha(q,n-q)$ is the upper 100α% critical point of the F distribution with k and $n-q$ degrees of freedom (T. Kourti and J F. MacGregor,1994).

2.5 Biplots

The biplot is based on the idea that any data matrix, Y (n x p), can be represented approximately in d dimensions (d is usually 2 or 3) as the product of a two matrices, A (n x d), and B (p x d).

The rows of A represent the observations in a two- (or three-) dimensional space, and the columns of B prime represent the variables in the same space. The prefix "bi" in the name biplot stems from the fact that both the observations and variables are represented in the same plot, rather than to the fact that a two-dimensional representation is usually used.

For the principal component analysis, the axes in the biplot represent the principal components or latent factors and the observed variables are represented as vectors. Below, the figure 2 shows the sample of biplots

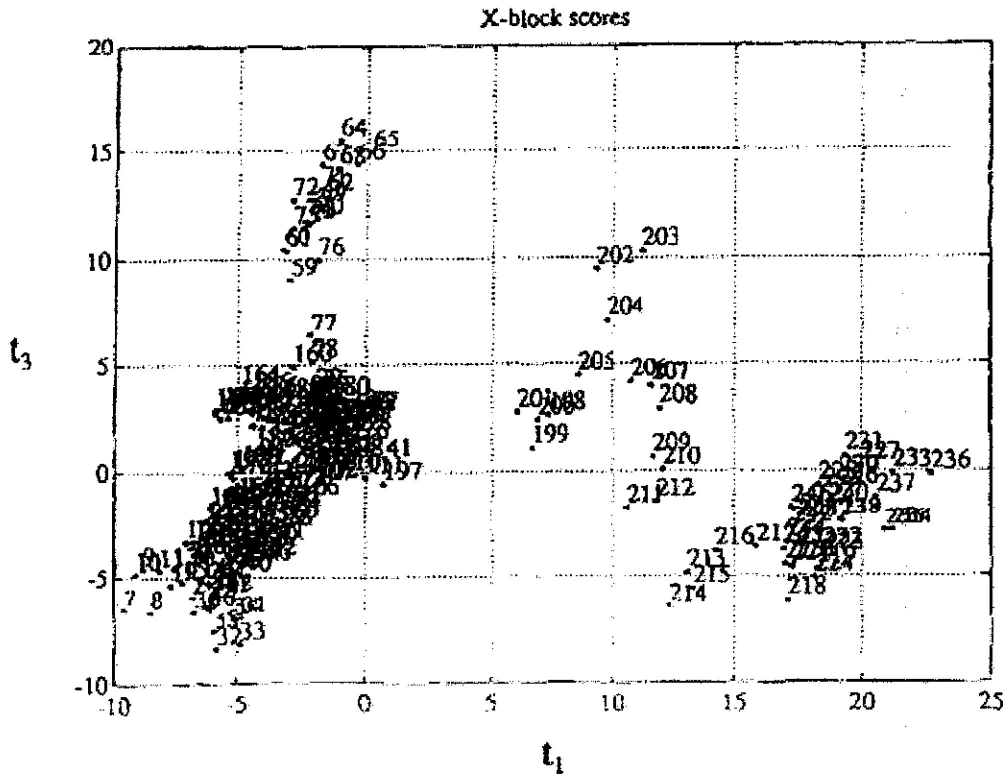

Figure 2 : The Biplots

CHAPTER 3

METHODOLOGY

In this section, it will be explain about the work will be done for this project. For this project, it uses Microsoft Excel 2007 and also Matlab 7.1.

3.1 Steps in Multivariate Statistical Process Control (MSPC)

3.1.1 Data Loading

The data-load process reads a source data file, converts the data to a different format, and inserts the converted data into a database table. The source data can come from one or more of the following sources. During conversion, the source data is often manipulated so that the converted data displays different characteristics.

3.1.2 Data processing/normalization

Normalization is the process of removing statistical error in repeated measured data. This normalization data will be doing before and after the outlier detection. There two goals of the normalization:-

1. Eliminating the redundant data
2. ensuring data dependencies make sense

3.1.3 Outlier detection

Before sending the data to PCA, the outlier must be removing because the PCA is very sensitive to the present of outlier. These outliers are based on the control chart. The data which exceed the limit will be removing.

3.1.4 Principal Component Analysis

Principal Component Analysis (PCA) is to identify patterns in data, and expressing the data in such a way as to highlight their similarities and differences without losing the original information.

3.1.5 Limit Determination

For conventional Shewchart Control Chart, the Upper Control Limit (UCL) and Lower Control Limit (LCL) for mean-centered and variance-scaled variables are +3 and -3 respectively (McNeese and Klein,1991). By using the equations from section 2.4 limit of chart are calculated.

3.2 Concept of idea in Microsoft Excel

Before start doing the coding and interface, it must have concept of idea what will be happening for whole of the program from start until end of the program.

Before doing any calculation or construct a graph, it required a bunch of data at least 2 set of range data. The user will be uploading the data into input interface or windows. The user also must rename their variable to make sure their do not confuse. After the data are uploading into windows, the user will click the button to proceed. There will some instruction on the first interface.

Data from first interface will be paste on second interface. In this interface, the data will be normalizing. Before that, it must find the mean and standard deviations for each range of data which will be use for normalization. The data will be arrange that the data user it will be on left hand site while the normalize data on the right hand side. The mean and standard deviation will be calculated by using this formula:

Mean

= AVERAGE (Number1; number2 :...) (10)

Standard Deviation (SD)

= STDEV (Number1; Number2 ;...) (11)

and, to calculate for normalization data, it will be use this formula

$$nomal = \frac{data - mean}{SD}$$

After data being normalize, the normalize data will be use to construct the control chart on other interface. Before construct, it must calculate the mean and standard deviation for normalize data. By using Upper Limit Control and Lower Limit Control formula, the control chart will be constructing to see the behavior of the data. Make sure all range of data must do the control chart.

If, in the control chart shows there are possible outlier, the outliers must be eliminate first before enter next interface which for Principal Component Analysis. After being remove the outliers, the data have to go back for normalization data because the mean and standard deviation have been change and construct back the control chart.

After there are no outliers, the user cans proceed to next interface for Principal Component Analysis. In this interface, the data will be sent to Matlab for Principal Component Analysis calculation.

After doing the calculation for PCA, the data and figure which generate by the calculation, will be sending back to Microsoft Excel to show the result to the user. Then, for next step, the user will insert the future data. Future data is used to see whether the limits that had been calculated before is fixed or not to it. The future data will be normalizing by using mean and standard deviation from previous normalization.

Then, the normalize future data will be send to Matlab for the calculation and matlab will send back the result to Excel.

CHAPTER 4

RESULT AND DISCUSSION

4.1 Interface

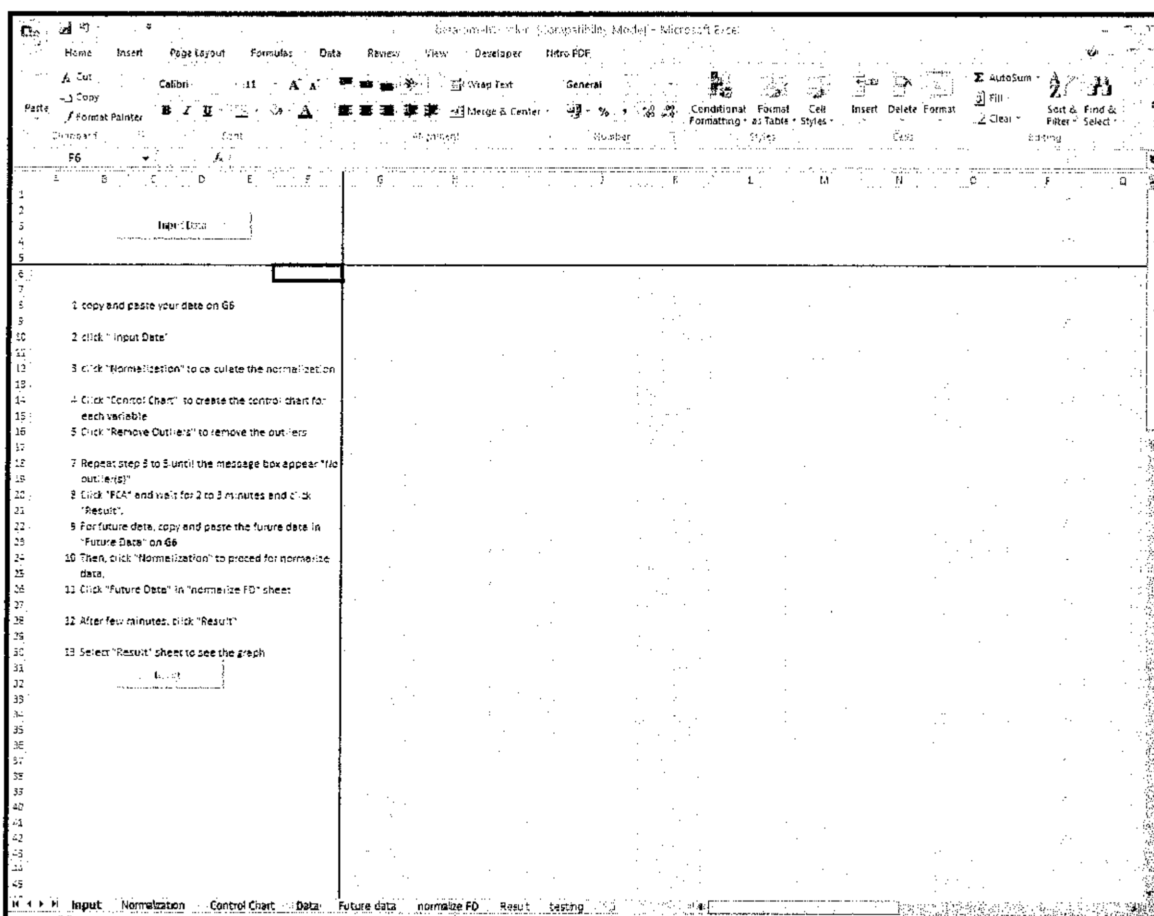

Figure 3: Input interface

Figure 3 above shows the first interface that will be seen by the user. In this interface, the data inside the selected area will be copied and sent to the Normalization sheet. The user must put their data inside the selected area. Then, click the “Input Data” button which the button is at the top left corner. The user must make sure that the ranges of data are equal. At the left hand side, there are procedures on how to use this.

On the bottom, there is "Reset" button to reset or remove all data inside this worksheet.

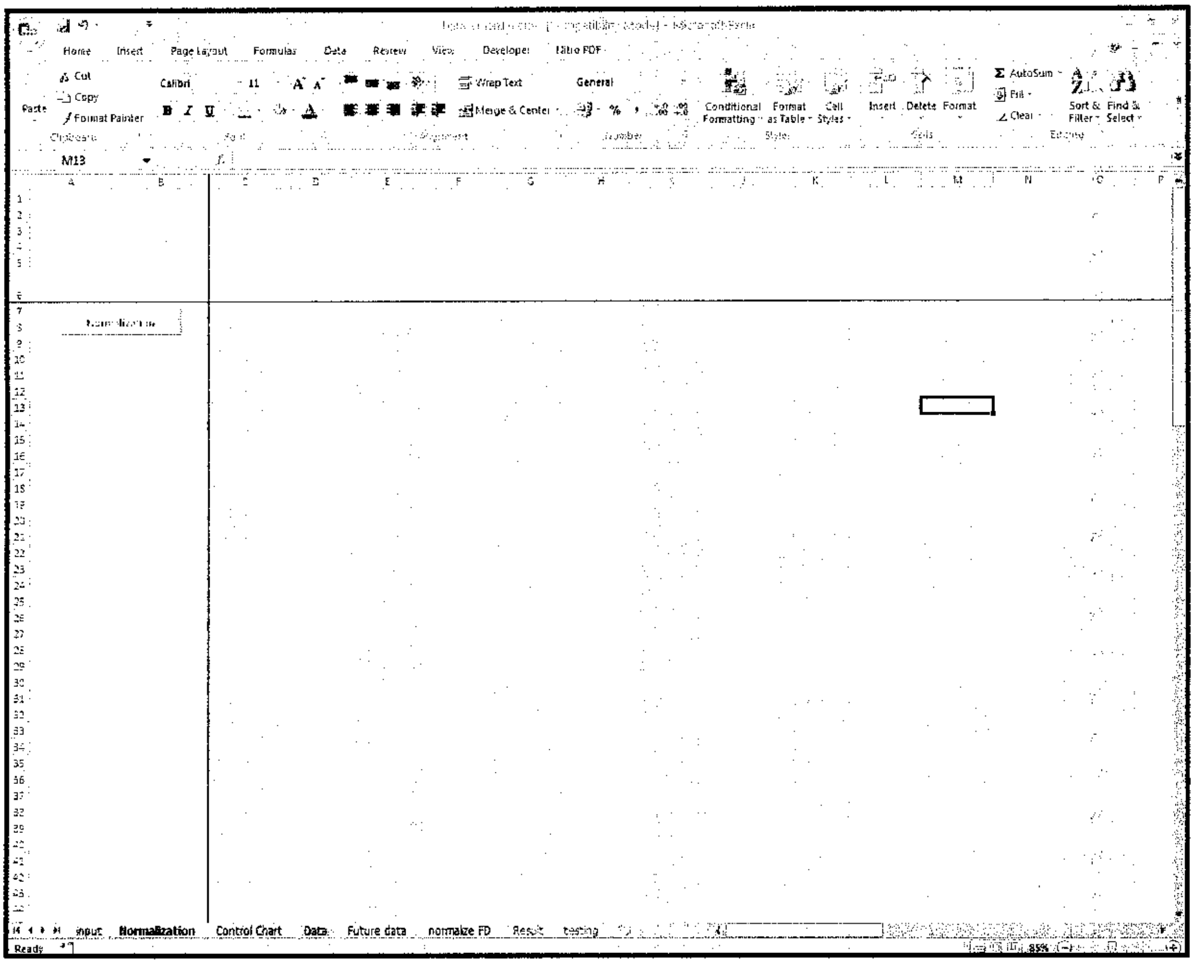

Figure 4: Normalization interface

As illustrated in Figure 4, this is the second interface. The function of this interface is to normalize the data before constructing the control chart. The "Normalization" button is to run the calculation for normalization. First, it will calculate the mean and standard deviation.

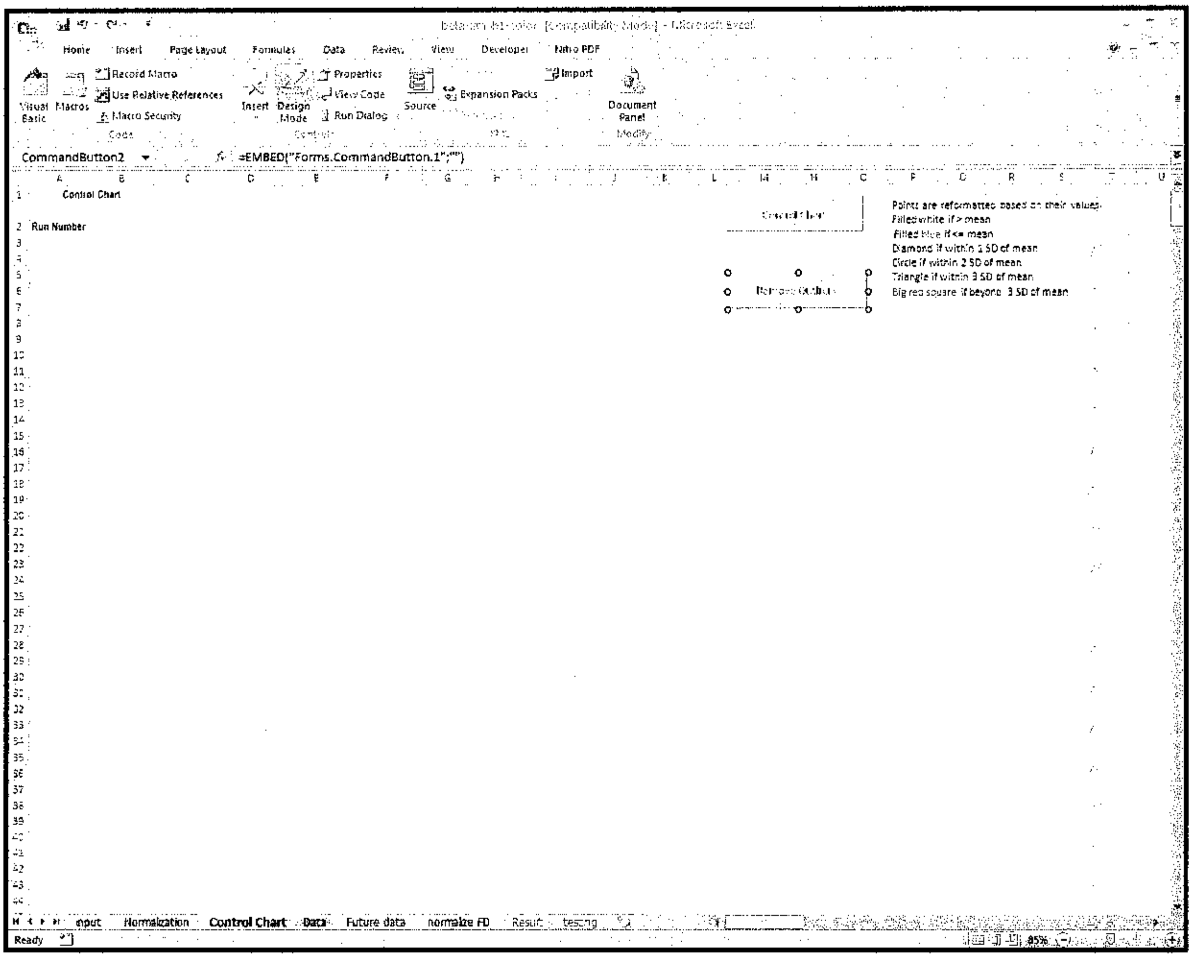

Figure 5: Control Chart Interface

Figure 5 shows the interface for control chart graph. The data will be inserted under the measurement column. The 'Control Chart' button is to create the control chart for each variable. Below the control chart button, there is 'Remove Outlier' button which use to remove outlier from the data. At the top right corner, there are legends for the control chart.

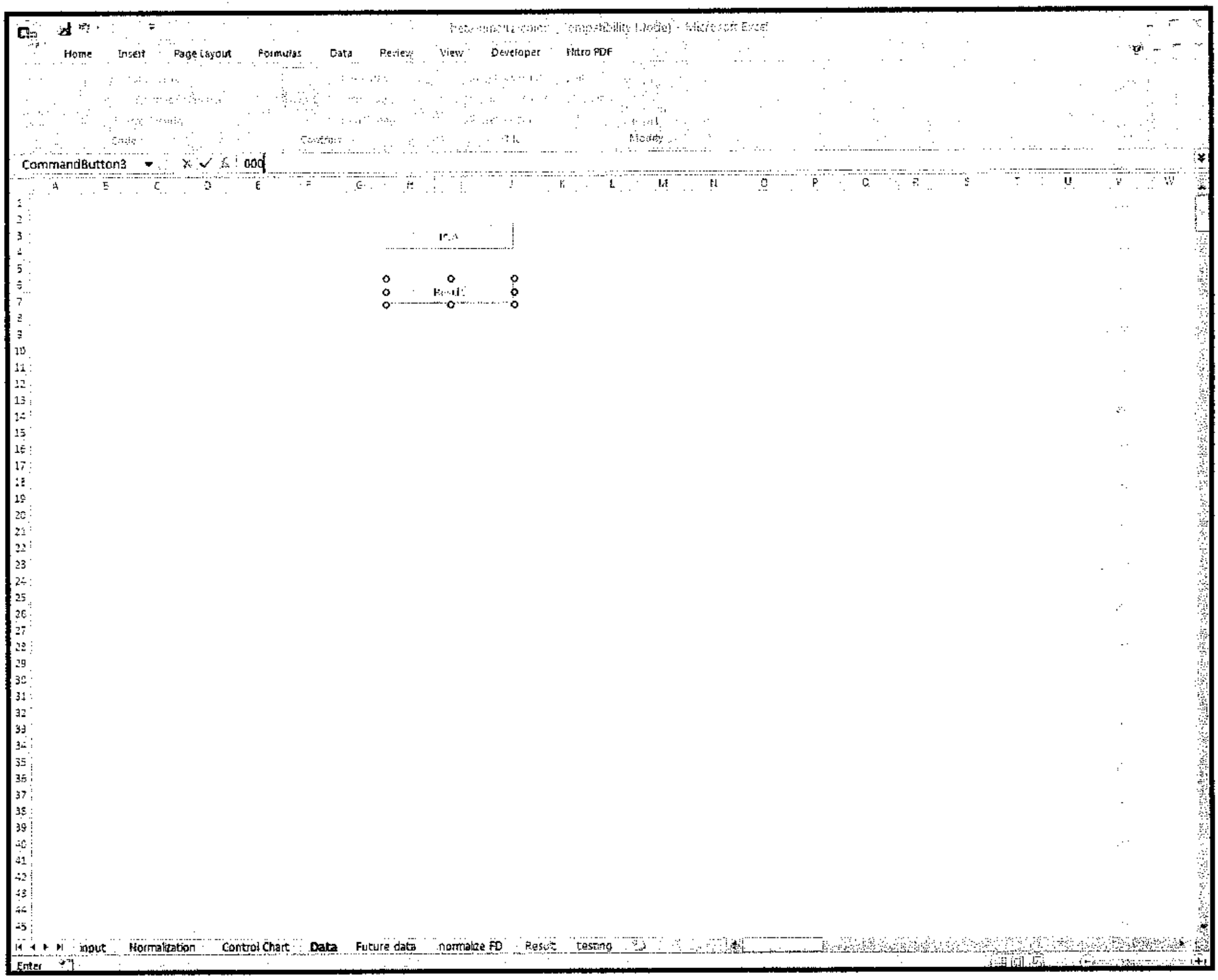

Figure 6: Data Interface

Figure 6 shows the interface for data. On this interface, data which have been removed the outliers will show on this interface. The 'PCA' button on this interface is used to send the data to Matlab for PCA calculation and generate figure. This figure will be sent back to Excel by using 'Result' button. The figure will be shown in result interface.

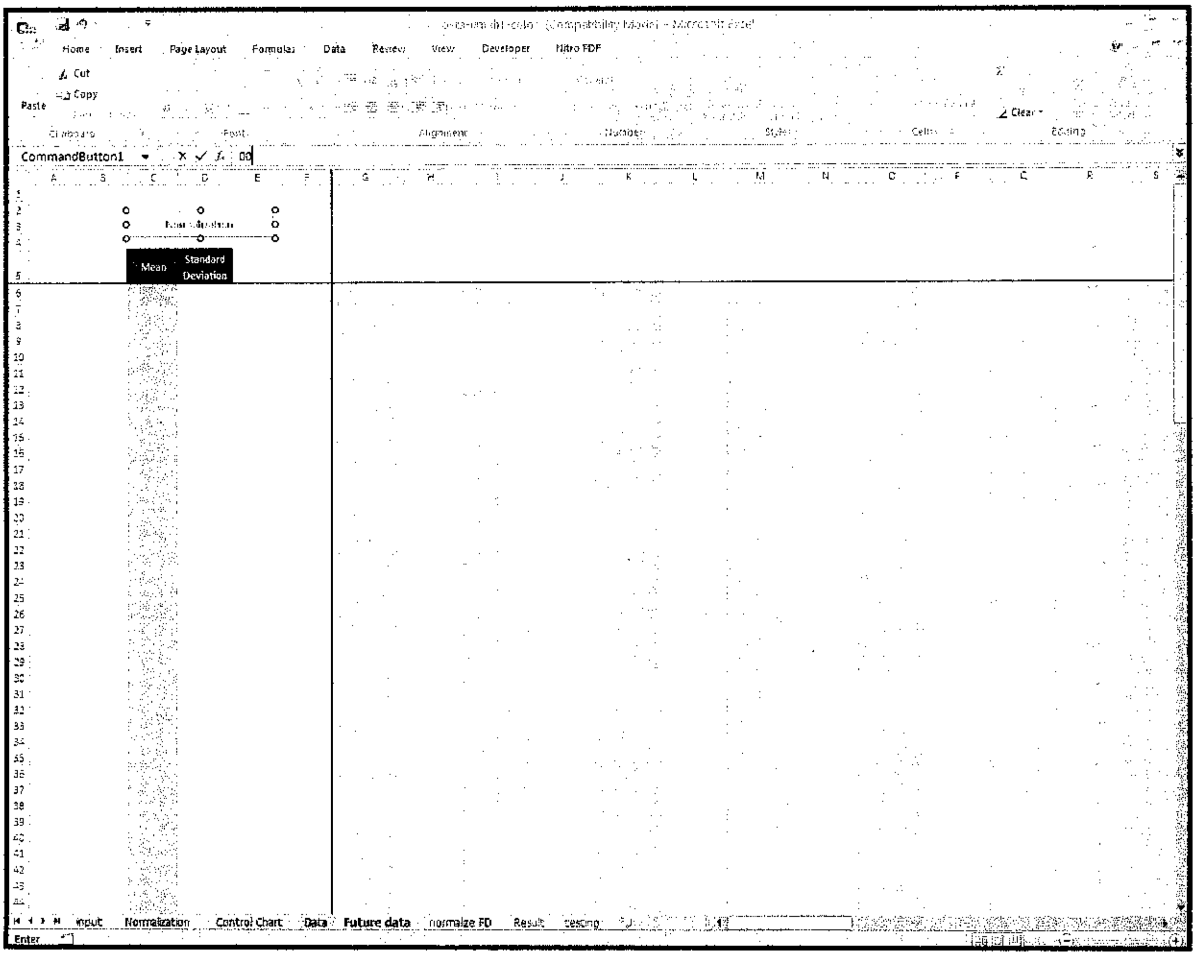

Figure 7: Future Data interface

Figure 7 shows the future data interface. For this interface, the user will place the future data at G7. Future data is used to see whether or not the limits that had been calculated before is fixed to it. On the left hand side, there are mean and standard deviation columns which are taken from previous data normalization. This mean and standard deviation will be used to normalize the future data. On the top left hand side, there is the 'Normalization' button. This button will function to normalize the future data.

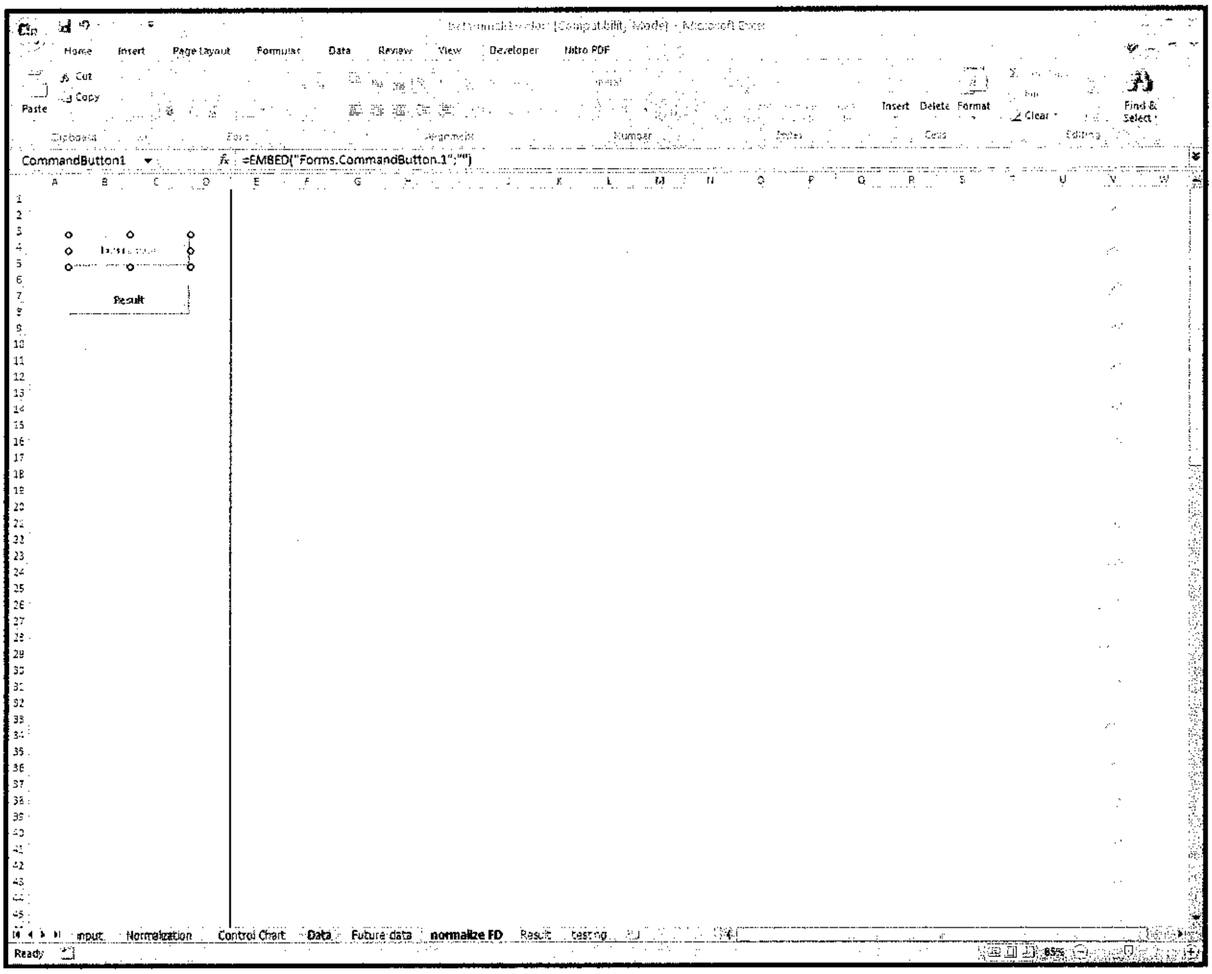

Figure 8: Normalization of Future Data Interface

On this figure8, the normalized data from future data will be displayed on this interface. There are two buttons; 'Future Data' and 'Result'. 'Future Data' button is used to send the normalized future data to Matlab. The 'Result' will be used after the calculation in Matlab. Function for 'Result' is to display the result from the Matlab. The result will be shown in result interface.

4.2 Example

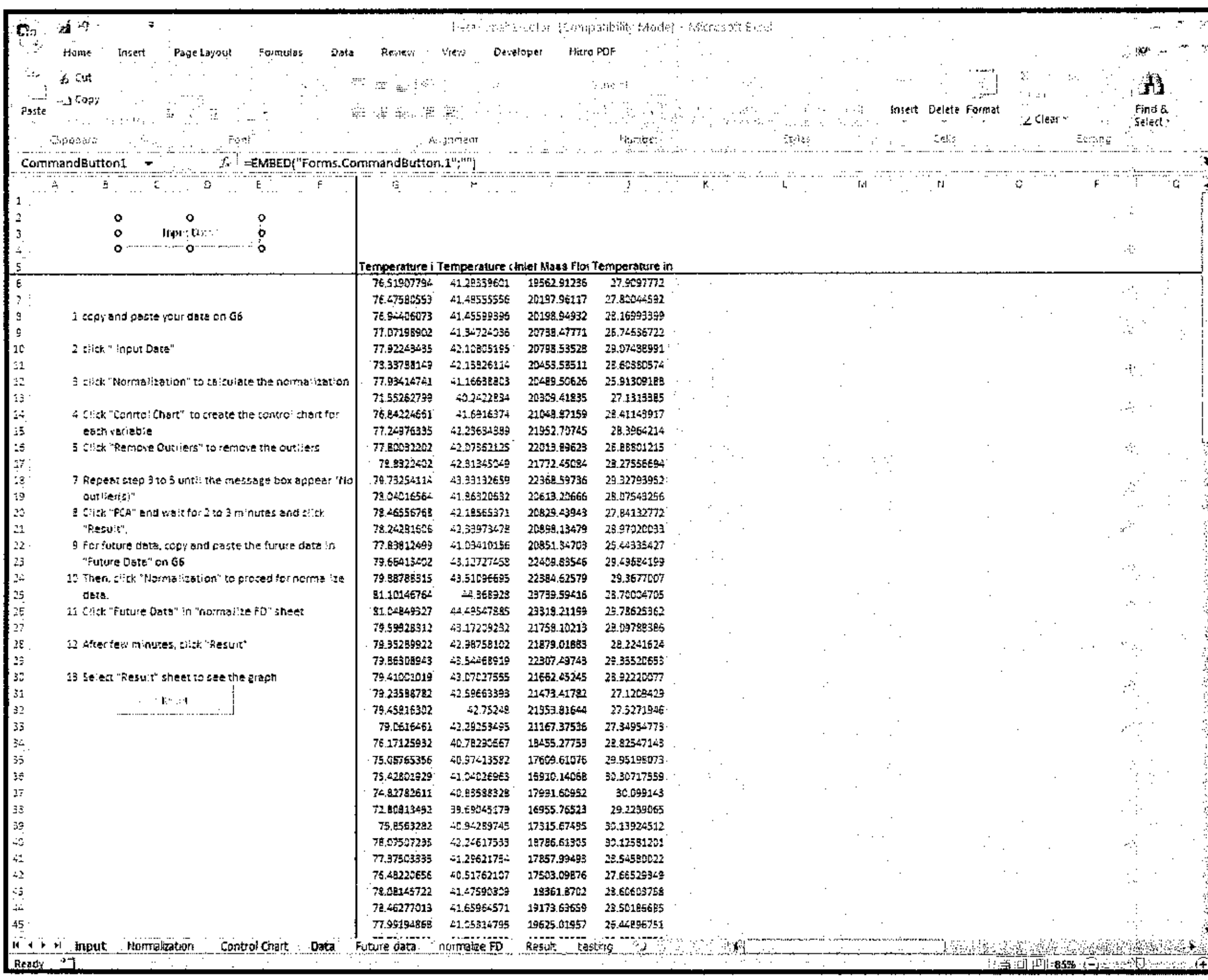

Figure 9: Input interface with sample data

From figure 9, the user has to paste their data into the selected area. The user can also put the variable name on the top of the data. After the date has been pasted, click the button 'input data' to proceed. The data will be selected, and then the user can move to the next sheet.

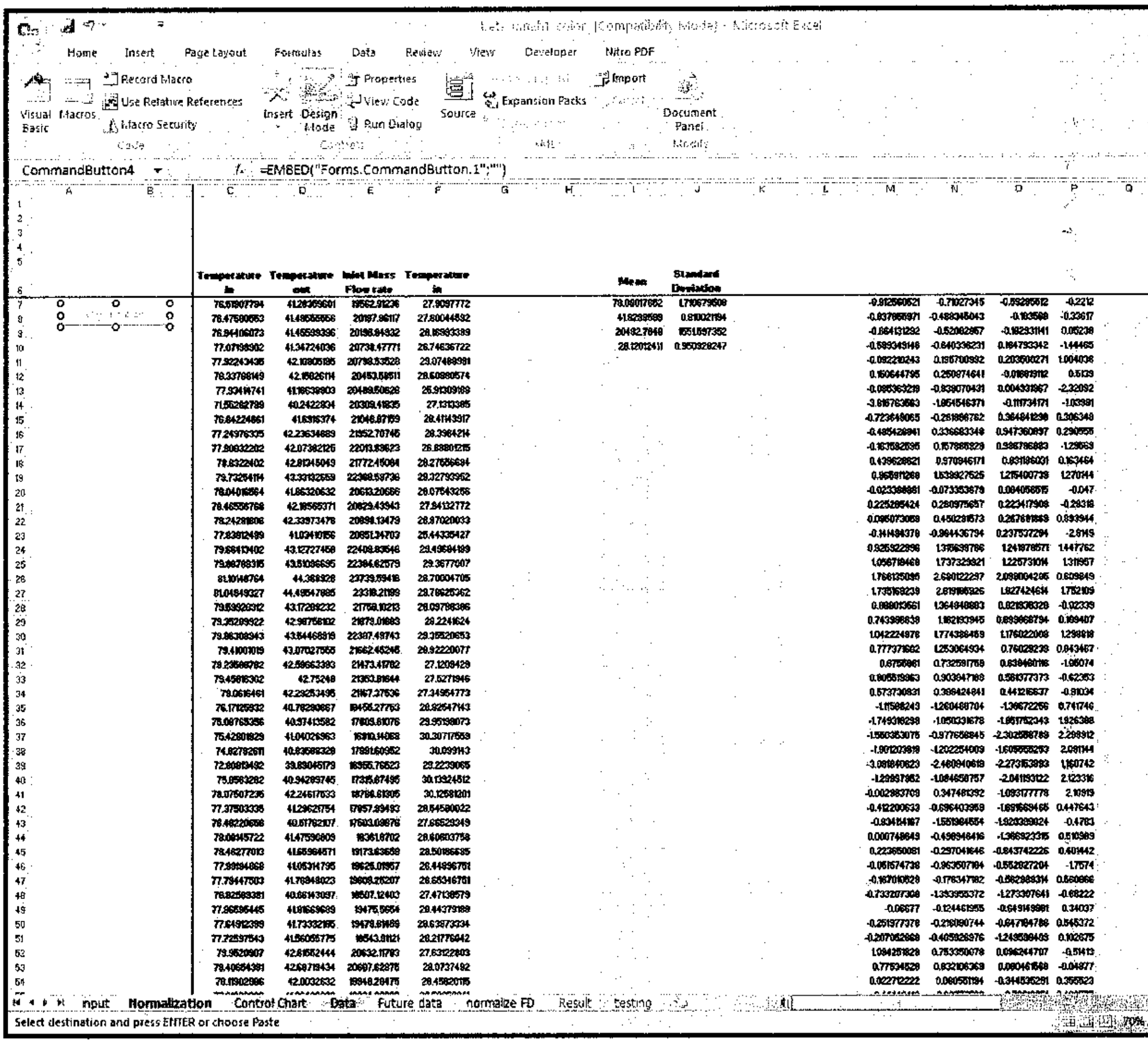

Figure 10: Normalization interface with sample data

In figure 10, the data from the "Input" sheet have been pasted at C7. By clicking the "Normalization" button, Microsoft Excel calculates the mean and standard deviation value from each variable. The normalized data will be under normalization column. The normalized data shows that the mean for each data will be close to zero while for standard deviation it will be close to 1.

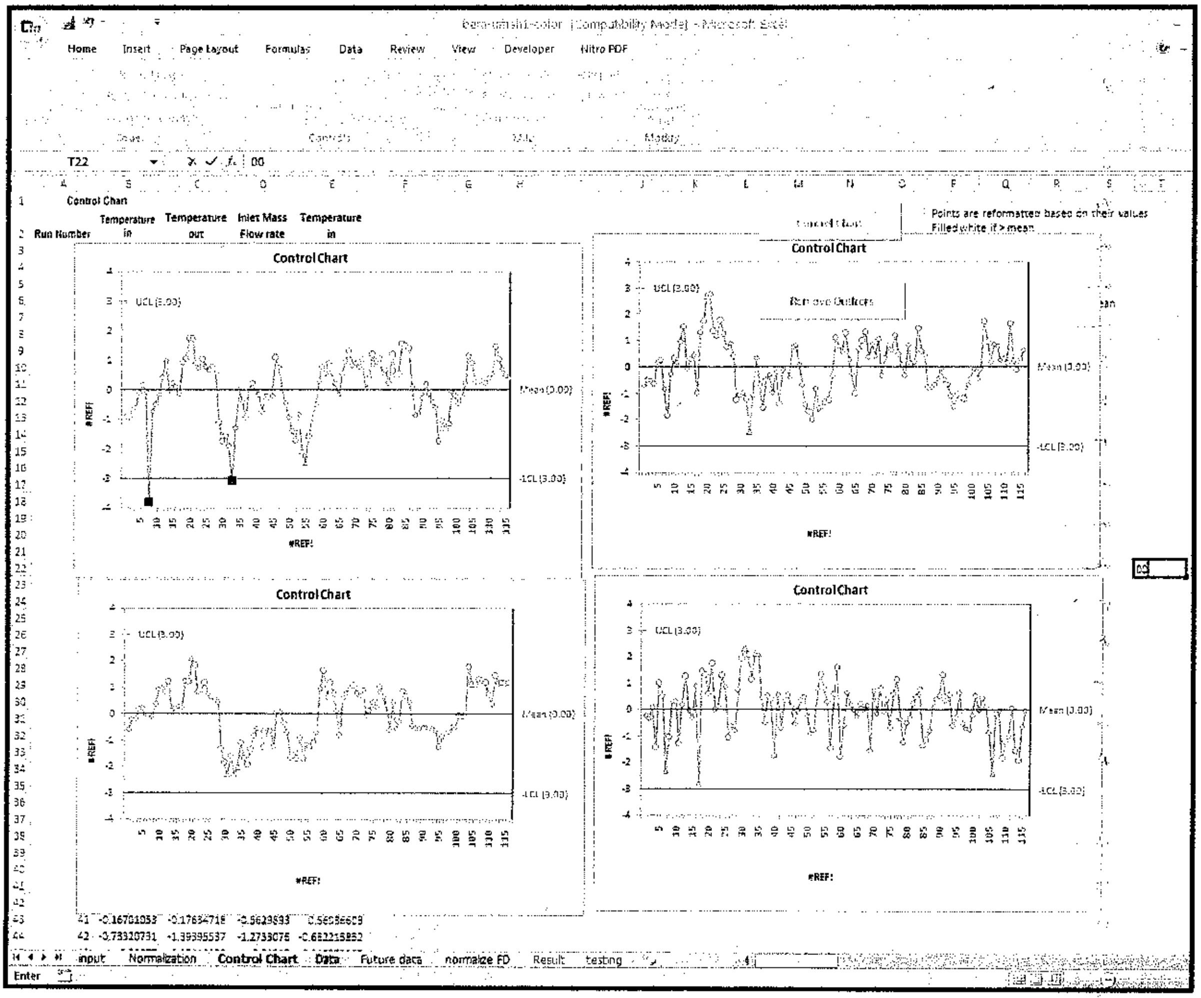

Figure 11: Control Chart interface with sample data

On this figure 9, the interface shows the control chart graph which most all data value inside range which the process remain close to their desired values. For this example, there are 4 variables and 4 control charts for each variable. The small red squares indicate that, there are two potential of the outliers existed in the process. The outliers have to be excluded from the graph. Therefore, the button 'Remove Outliers' will be used to remove the outliers from the data.

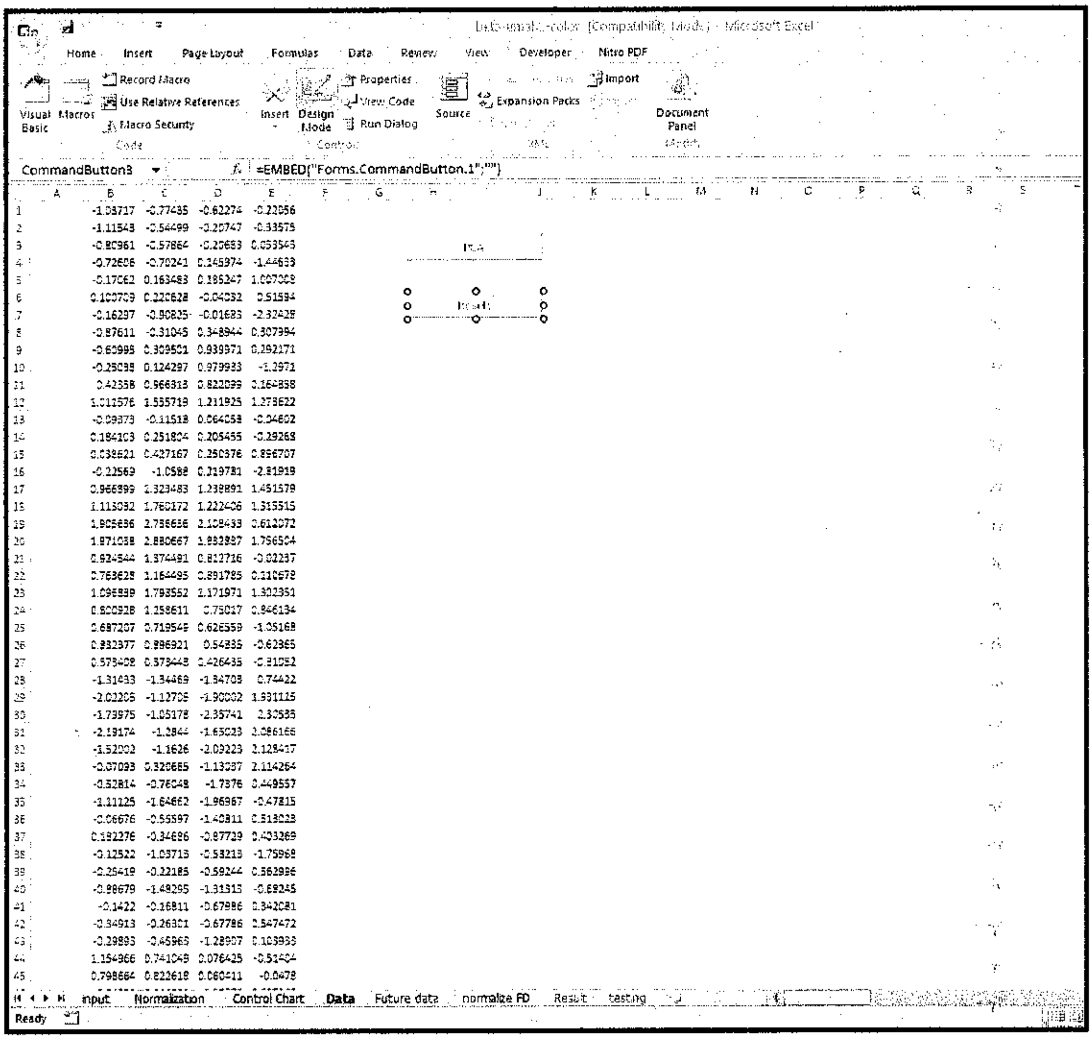

Figure 12: Remove Outlier interface with sample data

On this figure 12, the normalized data has been removed from the outlier. After this, the user must go back to "Normalization" sheet to repeat the step back until no outliers in control chart or appear message box says "No Outlier(s)".

Then, this data will be sent to Matlab for calculation of Principal Component Analysis (PCA) by clicking the "PCA" button. The calculation will take a few minutes. Then, to see the result, the user must click the 'Result' button which will refer to the result sheet.

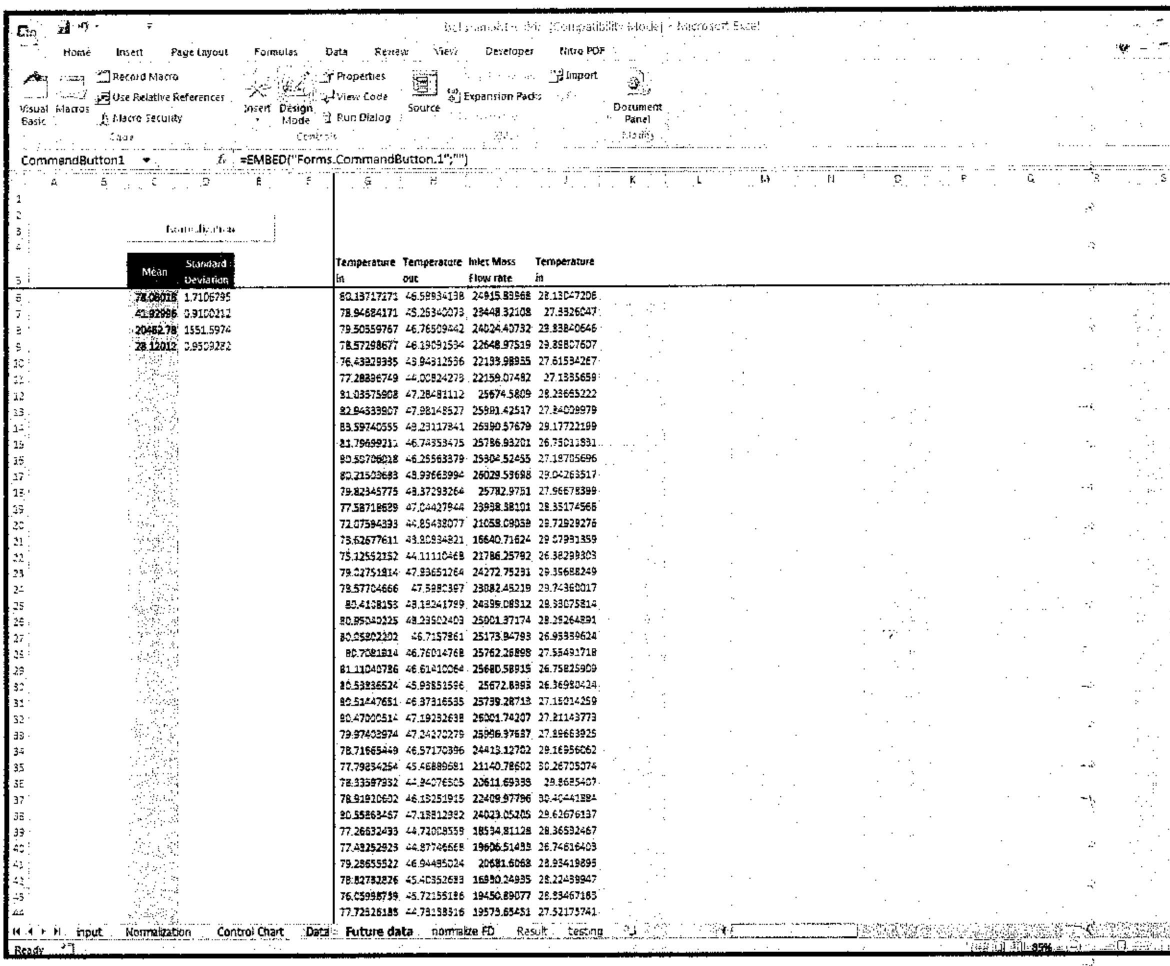

Figure 13: Future Data Interface with sample data

On figure 13, the interface shows the future data in blue box. The sequence for each column must be the same with the sequence in previous data which data modeling. This is to make sure the mean and standard deviation are related to each data. On left side, there are two columns representing the mean and standard deviation. By clicking the 'Normalization' button the 1st column will be normalized by 1st row in mean and standard deviation columns. The normalized data will appear in Normalize FD sheet.

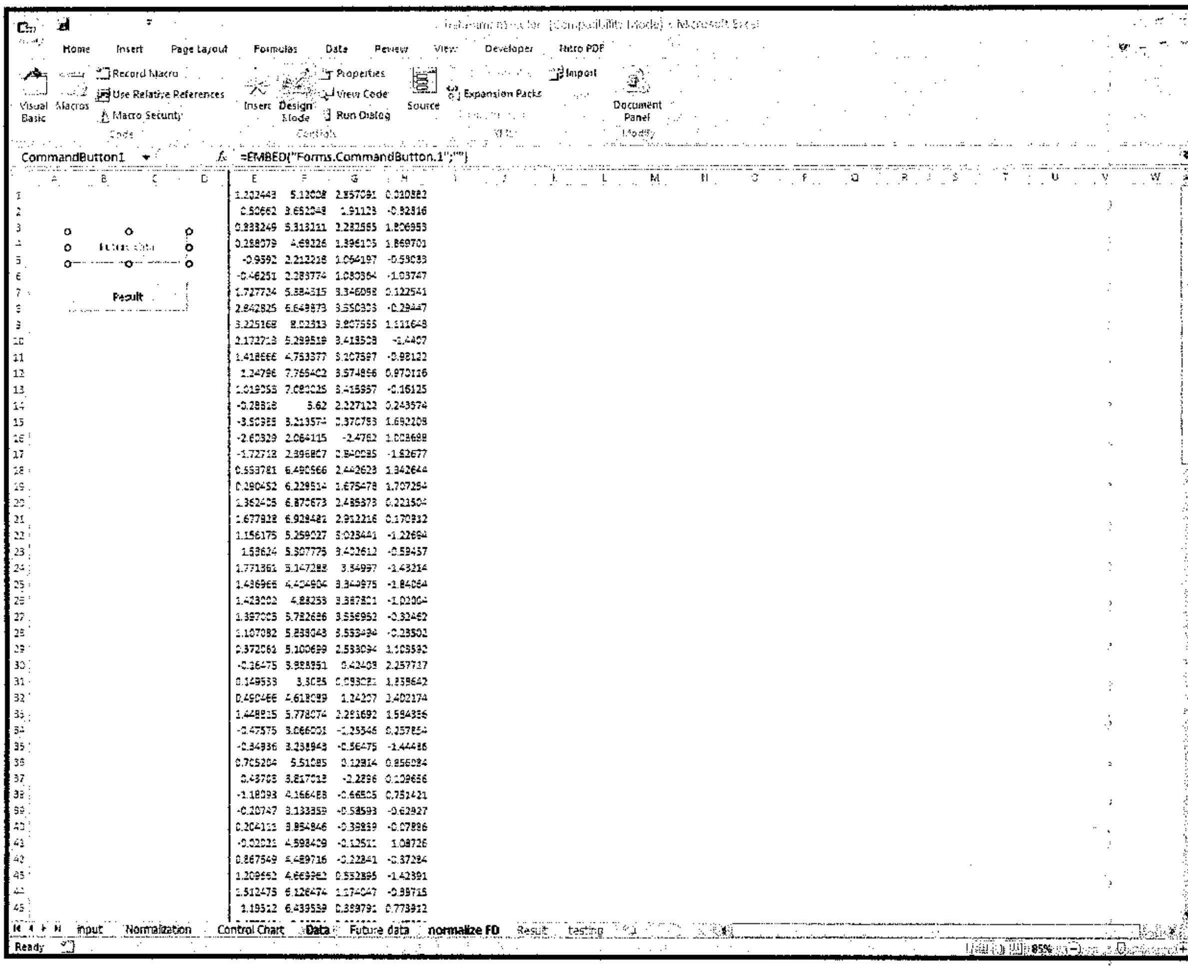

Figure 14: Normalization of Future Data interface with sample data

In figure 14, the future data have been normalized by using mean and standard deviation from previous data. From this, the user can see that the data are mostly ranged between -3 until 3. But, if the values exceed the range, there are possibilities of equipment malfunction. For the next step, the user must click the 'Future data' button. After a few minutes, the calculation is done. Then, the user should click the 'Result' button to show the result from Matlab.

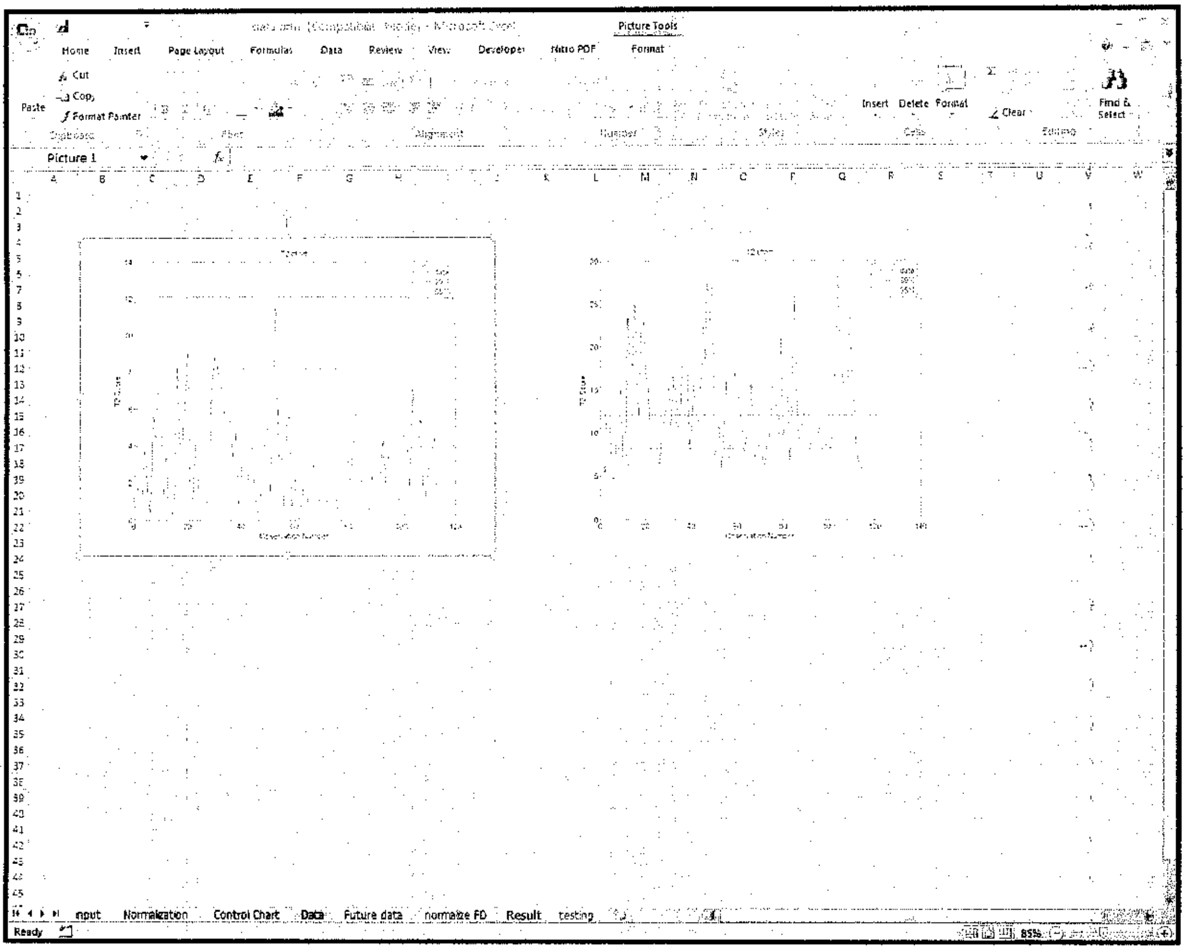

Figure 15: Result Interface with sample data

For figure 15, the interface shows the result which came from the Matlab. On left hand side, the figure shows the T^2 chart for modeling data and right hand side shows the T^2 chart for future data. T^2 chart for future data used the limit identical for the modeling data. From this, the user can see that the future data are beyond the limit. It is understood that something happened in the process which leads the process technologist to find the causes.

CHAPTER 5

CONCLUSION

This project has presented the tool which develop by using Microsoft Excel and matlab. The overall aim of the study was to develop a process monitoring tool using multivariate method that would enable process operators to quickly and easily identify any sources of abnormality in the process. This paper also provided an overview of the concepts behind multivariate statistical process control. The multivariate method can easily detect the abnormality of the process and diagnostics assignable cause. This tools also can share widely with other Microsoft Excel platform.

REFERENCES

1. T. Kourti and J F. MacGregor, *Process analysis, monitoring and diagnosis, using multivariate projection methods*,(1994)
2. A. AlGhazzawi and B. Lennox, *Monitoring a complex refining process using multivariate statistics*, (2006)
3. M. K. Yee and Kamarul A. Ibrahim, *Fault Detection for distillation column using multivariate statistical process control (MSPC)*, (2003).
4. Sung H.Park and G.Geoffrey Vining, *Statistical Process Monitoring and Optimization (*2000*)*
5. Harald Martens and Magni Martens, *Multivariate Analysis of Quality (*2001*)*
6. L.H. Chiang, E.L. Russell and R.D. Braatz, *Fault Detection and Diagnosis in Industrial system, Springer-Verlag London, 2001*
7. D. Aguado and C. Rosen, *Multivariate Statistical Monitoring of continuous wastewater treatment plants*,(2006)
8. L. I. Smith, *A tutorial on Principal Component Analysis*,(2002)
9. M. Kano, S. Hasebe, I. Hashimoto and H. Ohno, *A new multivariate statistical process monitoring method using principal component analysis*, (2001)
10. R. Jacobson, *Step by Step : Microsoft Office Excel 2007 Visual Basic for Application, 2007, Microsoft*
11. J. Walkenbach, *Microsoft Office Excel 2003 Power Programming with VBA,2004, Wiley Publishing*