Application of Multivariate Statistical Process Control to Fuel Cell Manufacturing

Olivia Lannon

MSc. by Research

Institute of Technology, Sligo

Supervisor of Research: Dr. John Donovan

Submitted to the Higher Education and Training Awards Council

August, 2009

Declaration

I declare that I am the sole author of this thesis and that all the work presented in it, unless otherwise referenced, is my own. I also declare that this work has not been submitted, in whole or in part, to any other university or college for any degree or qualification.

I authorise the library of Sligo Institute of Technology to lend this thesis.

Signed: ______________________

Date: ______________________

TABLE OF CONTENTS

ACKNOWLEDGMENTS III

ABSTRACT IV

1. INTRODUCTION 1

1.1 INTRODUCTION 1
1.2 CURRENT STATE OF STATISTICAL PROCESS CONTROL 2
1.3 SCOPE 3
1.4 STRUCTURE 4
1.5 REFERENCES 5

2. STATISTICAL PROCESS CONTROL 6

2.1 INTRODUCTION 6
2.2 DISTANCE STATISTICS 17
2.2.1 Euclidean Distance and Statistical Distance 17
2.3 HOTELLINGS T^2 18
2.3.1 Hotellings T^2 Statistic 18
2.3.2 Hotellings T^2 Distribution 19
2.3.3 T^2 Distributions for Individual observations 20
2.3.4 T^2 Distribution for Subgrouped Data 21
2.4 HOTELLINGS T^2 CONTROL 22
2.4.1 Phase I and Phase II UCL for Individual observations. 22
2.4.2 Phase I and Phase II UCL for Subgrouped Data 30
2.5 AUTOCORRELATION 32
2.6 COLLINEARITY 33
2.7 CONCLUSION 34
2.8 REFERENCES 36

3. PRINCIPAL COMPONENT ANALYSIS 38

3.1 INTRODUCTION 38
3.2 PRINCIPAL COMPONENT ANALYSIS 39
3.3 STANDARDISED PRINCIPAL COMPONENTS 41
3.3.1 Principal Component Scores 41
3.3.2 Standardised PC Scores 45
3.3.3 Retaining Principal Components 47
3.4 PRINCIPAL COMPONENT CONTROL CHARTS 48
3.4.1 PC Scores for a New Observation 51
3.5 T^2 CHARTS AND SQUARED PREDICTION ERROR (SPE) IN PC SPACE 53
3.5.1 T^2 Charts in Principal Component Space 53
3.5.2 Residuals Charts 55
3.5.3 Squared Prediction Error 56
3.5.4 DModX Charts 57
3.6 CONTRIBUTION PLOT 59
3.7 PARTIAL LEAST SQUARES 61
3.8 CONCLUSION 62
3.9 REFERENCES 63

4. MULTIVARIATE SPC FOR BATCH PROCESSES ... 65
4.1 INTRODUCTION ... 65
4.2 BATCH PROCESSING ... 66
4.2.1 T^2 on Batch Observations ... 66
4.2.2 Batch Process Data ... 68
4.2.3 Multiway PCA ... 69
4.2.4 Batch-wise and Variable-wise Unfolding ... 70
4.2.5 Dealing with Incomplete Batches ... 72
4.3 CONCLUSIONS ... 73
4.4 REFERENCES ... 75

5. FUEL CELL TECHNOLOGY AND MANUFACTURING PROCESS ... 77
5.1 INTRODUCTION TO FUEL CELL TECHNOLOGY ... 77
5.2 OVERVIEW OF MANUFACTURING PROCESS ... 79
5.2.1 Stages of Fuel Cell Assembly ... 80
5.3 PERFORMANCE MEASUREMENT TESTS ... 85
5.4 CONCLUSION ... 89
5.5 REFERENCES ... 91

6. MULTIVARIATE SPC TO MONITOR FUEL CELL PERFORMANCE ... 92
6.1 INTRODUCTION ... 92
6.2 MULTIVARIATE ANALYSIS ... 92
6.3 MULTIVARIATE CONTROL CHART ... 93
6.3.1 T^2 Statistic using Individual Observations ... 93
6.3.2 T^2 Statistic on Batch Observations ... 101
6.4 PRINCIPAL COMPONENTS ANALYSIS ... 102
6.4.1 Develop PCA Model ... 102
6.5 COMPARISON OF MULTIVARIATE METHODS ... 113
6.5.1 Phase II using Principal Components Analysis ... 114
6.6 CONCLUSION ... 116
6.7 REFERENCES ... 117

7. DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS ... 118
7.1 INTRODUCTION ... 118
7.2 DISCUSSION ... 119
7.3 CONCLUSIONS ... 120
7.4 RECOMMENDATIONS ... 122
7.4.1 Method ... 122
7.4.2 Parameter Selection ... 122
7.4.3 Characterisation ... 123
7.4.4 Software ... 123
7.4.5 Risk of False Alarms ... 124
7.5 FUTURE RESEARCH ... 124
7.5.1 Partial Least Squares ... 124
7.5.2 Automated Data Collection ... 124
7.5.3 Robust Charting Methods ... 125
7.6 REFERENCES ... 126

ACKNOWLEDGMENTS

I would like to express my gratitude to my research supervisor, Dr. John Donovan. His advice and guidance over the last two years has been exceptional.

This research project was part funded by Enterprise Ireland, as part of an innovation partnership scheme. I am extremely grateful for the opportunity to work on this exciting and innovative project.

ABSTRACT

Application of Multivariate Statistical Process Control to Fuel Cell Manufacturing

Olivia Lannon

Univariate statistical control charts, such as the Shewhart chart, do not satisfy the requirements for process monitoring on a high volume automated fuel cell manufacturing line. This is because of the number of variables that require monitoring. The risk of elevated false alarms, due to the nature of the process being high volume, can present problems if univariate methods are used. Multivariate statistical methods are discussed as an alternative for process monitoring and control.

The research presented is conducted on a manufacturing line which evaluates the performance of a fuel cell. It has three stages of production assembly that contribute to the final end product performance. The product performance is assessed by power and energy measurements, taken at various time points throughout the discharge testing of the fuel cell.

The literature review performed on these multivariate techniques are evaluated using individual and batch observations. Modern techniques using multivariate control charts on Hotellings T^2 are compared to other multivariate methods, such as Principal Components Analysis (PCA).

The latter, PCA, was identified as the most suitable method. Control charts such as, scores, T^2 and DModX charts, are constructed from the PCA model. Diagnostic procedures, using Contribution plots, for out of control points that are detected using these control charts, are also discussed. These plots enable the investigator to perform root cause analysis. Multivariate batch techniques are compared to individual observations typically seen on continuous processes. Recommendations, for the introduction of multivariate techniques that would be appropriate for most high volume processes, are also covered.

CHAPTER ONE

INTRODUCTION

1.1 Introduction

Portable electronic devices have become more advanced and continue to offer greater capabilities and functionality. Device manufacturers, service providers and consumers seek significantly increased and longer lasting power. Since batteries have reached close to their maximum capabilities, a power gap exists between those ever-increasing power demands of electronic applications and the amounts of power in present batteries. (Coll and Quinn, 2006).

Micro Fuel Cells are unique in their composition. They are cheap, lightweight, portable and they provide a power supply that is capable of charging a portable device when there is no wall socket available. For this reason, they are extremely convenient in emergency situations, as they provide an immediate power source.

The overall project involves research on the fuel cell manufacturing process in order to characterise a fuel cell assembly process which is stable and repeatable and can be successfully and safely scaled to meet the volume, yield and throughput targets. Market research has estimated that the potential market for the micro fuel cell is 18 million units per year. Currently no process exists for producing high volume Micro Fuel Cells. This process is new and untested and requires a high degree of technical innovation to deliver an automated process with six sigma quality levels. This is a unique product in that it is totally new; technology at this scale does not exist. This is the first automated line and worldwide volume manufacturing for Micro Fuel Cell line.

The traditional statistical techniques in Six Sigma for monitoring and controlling ongoing production quality have been very effective up to quite recently. These techniques have relied on identifying statistical trends in defective parts that indicate when deterioration in product quality has occurred. In such situations it was relatively easy to observe deterioration, as the ambient number of defective parts was reasonably high.

However, this situation has now changed and these techniques are no longer applicable. For example, in the high volume automated manufacturing sector it is not unusual for manufacturers to have product quality goals in the region of 10 defects in 1 million parts. With such low ambient defect rates, the traditional techniques are no longer effective or workable. Newer techniques have been proposed to detect abnormalities in production processes however very few of these have transferred successfully into industrial practice. These techniques include multivariate data transformation.

The objective of the research is to investigate these multivariate techniques and assess how effectively they perform their intended function bearing in mind the industrial context.

1.2 Current State of Statistical Process Control

In a survey conducted on small and medium enterprises Yusuf and Aspinwall (2000) were extremely surprised to find that that 25% of respondents were not applying Statistical Process Control (SPC) techniques especially when the companies involved tended to be the more advanced in quality practices. This is an unusual situation as the use of SPC is generally considered to represent a gauge of an enterprise's maturity (Montgomery, 2001). Woodall (2000) laments the fact that "some useful advances in control charting methods have not had a sufficient impact in practice". He also identifies a disturbing message that the SPC segment of the Certified Quality Engineer (CQE) exam of American Society for Quality (ASQ) consists almost entirely of material covered in the 1956 Western Electric Handbook.

One reason for this trend away from the application of SPC is that the traditional SPC charts have lost their relevance as identified by Gunter (1998), Woodall (2000) and Stoumbos (2000). These authors note that further research is required to make these techniques more relevant to the needs of modern industry and service environments.

More recently as quality is measured in part per million (ppm), it has become uneconomical to increase sample sizes still further. Because of this, new SPC techniques are being devised to detect changes in the process while retaining economic sample sizes.

1.3 Scope

The objective of the proposed research is to investigate SPC techniques for a high volume automated production process. As this is the first ever high volume automated continuous production project involving manufacture of this particular Fuel Cell technology, there will be constraints from unknown issues that may arise. In order to alleviate them, experimentation will be conducted to identify the process controls required and the optimised process parameters to ensure a stable and repeatable manufacturing process.

In order to establish a reliable and repeatable process which can be scaled successfully to a high volume fuel filling and sealing process for automated assembly process for Micro Fuel Cells the following objectives are required:

1. To investigate the current status of Statistical Process Control (SPC), in particular Multivariate SPC techniques.
2. To propose suitable new/alternative SPC techniques for use with high volume manufacturing.
3. To develop control charts that can also be applied to a higher volume of production.
4. To develop a control chart for a highly automated fuel cell process, which will assess the most important parameter, product performance.
5. To create a single monitoring system that will reduce the amount of work involved in monitoring individual process parameters.
6. Identify suitable diagnostic tools to assist in identifying parameters implicated for root cause of out of control failures.
7. Identify an effective system that can be used to monitor control of future observations, which can be used for increased volumes of production.

8. The system identified must be capable of monitoring both continuous and batch processes.

1.4 Structure

Chapter 2 discusses statistical process control in univariate and multivariate processes. Hotellings T^2 chart is introduced and its application is demonstrated throughout the chapter. Autocorrelation and collinearity considerations are also described.

Chapter 3 examines a multivariate method called Principal Components Analysis (PCA). The standard control charts generated from this method are presented in addition to a T^2 chart on PCA and another chart called DModX. Contribution plots and their usefulness in narrowing down which of the original variables are causing of an out of control signal is introduced. Partial Least Squares (PLS) method is also mentioned.

Chapter 4 discusses batch processing and the differences between it and continuous processing. Various methods for analysing batch process data are reviewed.

Chapter 5 discusses fuel cell technology and the various stages of manufacturing a fuel cell on an automated production line. The testing requirements at each stage of manufacturing and the final performance testing, which will contribute to the overall batch decision, are also described.

Chapter 6 details the research analysis, where the various multivariate methods are applied to real production data. One method is selected as the more suitable, from all the methods presented.

1.5 References

Coll, B., and Quinn, K., (2006), "IP Fuel Cell Proposal_Rev 3.3", Retrieved March 9, 2007 from Celestica Ireland Ltd. database, \\galinf02\data\DEPTS\Engineering\Design Projects\Medis\Enterprise Ireland.

Gunter, B., (1998), "Farewell Fusillade: An Unvarnished Opinion on the State of the Quality Profession". *Quality Progress*, pp. 111–119.

Montgomery, D.C., (2001), *Introduction to Statistical Quality Control*, 4th Edition, Wiley-Interscience, pp. 325-326.

Stoumbos, Z.G., Reynolds, M.R., Ryan, T.P., Woodall, W.H., (2000), "The State of Statistical Process Control as we proceed into the 21st Century". *Journal of the American Statistical Association*, Vol. 95, pp. 992-996.

Western Electric, (1956), *Statistical Quality Control Handbook*. AT& T, Indianapolis, IN.

Woodall, W.H., (2000), "Controversies and Contradictions in Statistical Process Control", *Journal of Quality Technology*, Vol. 32, No. 4, pp. 341-350.

Yusof, S.M., and Aspinwall, E.M., (2000), "Critical Success Factors in Small and Medium Enterprises: Survey Results", *Total Quality Management*, Vol. 11, Nos. 4/5&6, 2000, pp. 448- 462.

CHAPTER TWO

STATISTICAL PROCESS CONTROL

2.1 Introduction

In every industry process data is widely available, Statistical Process Control (SPC) is used to analyse and display data. SPC is widely used to monitor a process over time and improve its performance by reducing variability for the key process parameters. Control charts are the most common form of SPC. These control charts, presented by Walter A. Shewhart (1931), plot the values of the key process variables over time in order to show variation between observations of this variable. Variation is described in two main categories by Woodall (2000), 'Common cause' and 'Special cause' variation. 'Common cause' variation describes the natural variability or "noise" of the process. 'Special cause' variation usually has an assignable cause, that is not part of the natural process variation, and which can be removed. Control charts are used to distinguish between these two types of variation. A control chart is so-called because it has control limits associated with the process and the process is said to be "in control" when the data is between these limits. If something changes in the process then an "out of control" signal will lie outside of these limits, and something must be done to get the process back in control again. The control limits for these charts are referred to as "three sigma" limits and are calculated as ±3 standard errors from the centreline.

This chapter describes both univariate and multivariate data and distributions, what their similarities are and then the differences that make them so distinguishable from one another. Methods for analysing multivariate data are also introduced.

In order to generate a control chart, it is necessary to screen the data of any variation that has assignable cause. Once identified, this data can then be removed before calculating control limits. This is done so that the limits are representative of the process. Screening ensures that only normal process variation is captured. This procedure is referred to as Phase I in process control. It can take a lot of time to characterise and understand a process. In order to achieve a stable historical baseline in which to calculate the control limits from, it is desirable to capture all possible aspects of normal variation that may occur throughout the life cycle of the process.

Once the reference baseline has been characterised and 3 sigma limits are calculated, Phase II of process control then begins. The process is monitored for violations to these pre-defined control limits.

Run rules are applied to control charts to indicate when a process is in an out of control situation.

The Western Electric Handbook (1956) details some of these run rules which are used in order to detect patterns in the data.

Once Phase I upper and lower control limits have been determined then Phase II monitoring of a process can commence. Figure 2.1 shows a typical Shewhart SPC chart for individual measurements. It shows each individual data point plotted on the x axis and the value of the process variable on the y axis. There is an upper control limit (UCL) and a lower control limit (LCL) which act as a signal should there be some change outside of these Phase I "process" limits.

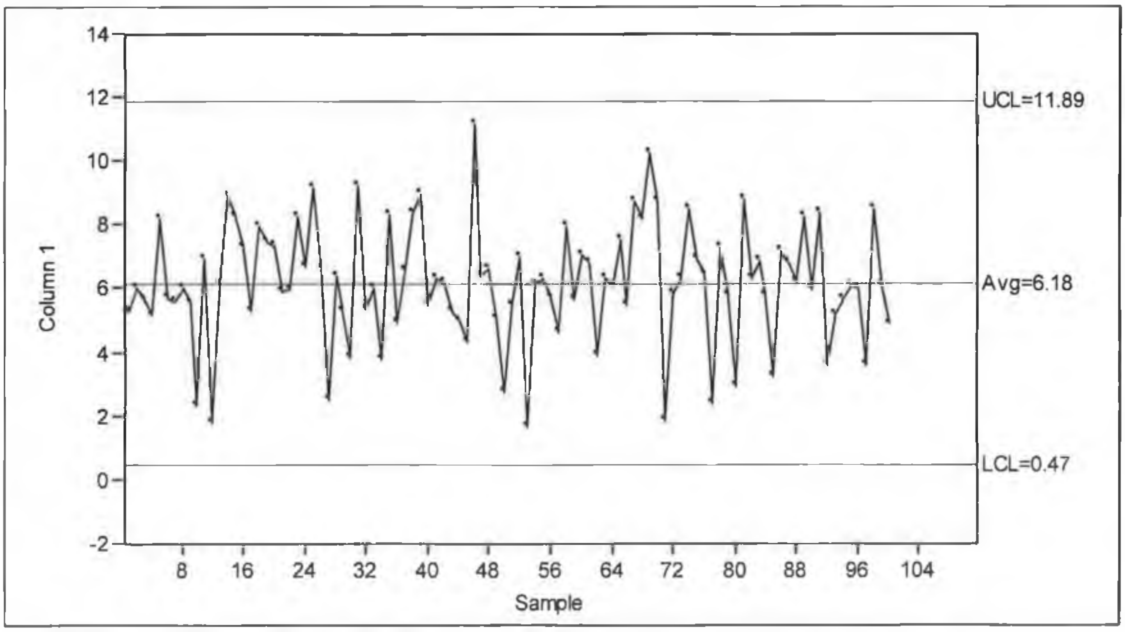


Figure 2.1 Shewhart SPC Chart for individual variables

It is not always possible to monitor every observation generated from a process variable, so it is common practice to take a sample of data points at specified time intervals in order to see if there is any process variation from the last time interval to the current time interval. Taking this into consideration, a variation of the individuals chart is used. A sample of *n* observations is taken at specific time intervals, *d*, for a single variable, the mean of these observations is then plotted on the control chart. This is referred to as an $\bar{x}$ chart. The sample group number is plotted on the x axis and the sample mean of the variable, $\bar{x}$, is plotted on the y axis.

Using the example above, the sample time intervals, *d*, were determined to be every 30 minutes. The $\bar{x}$ control chart, is shown in Figure 2.2. Similarly, the UCL and LCL are used to indicate when a process is out of control.

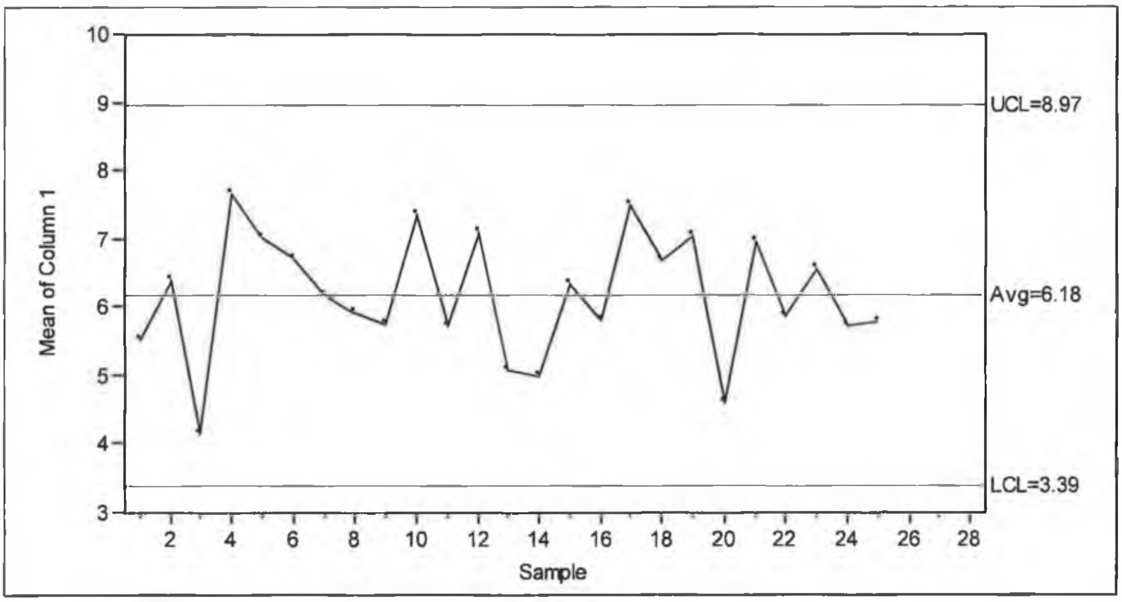


Figure 2.2 Shewhart SPC $\bar{x}$ Chart

In addition to the $\bar{x}$ chart, the dispersion within these sampling groups must also be monitored and this variance is measured and displayed through a standard deviation, S chart or Range, R chart. The dispersion of the data is important to monitor because an $\bar{x}$ chart showing the distance of the mean of a group from its centreline (average) does not describe how the variation of the data is spread across the groups. Figure 2.3 and Figure 2.4 show the example described above plotted on the typical Shewhart control charts used for dispersion.

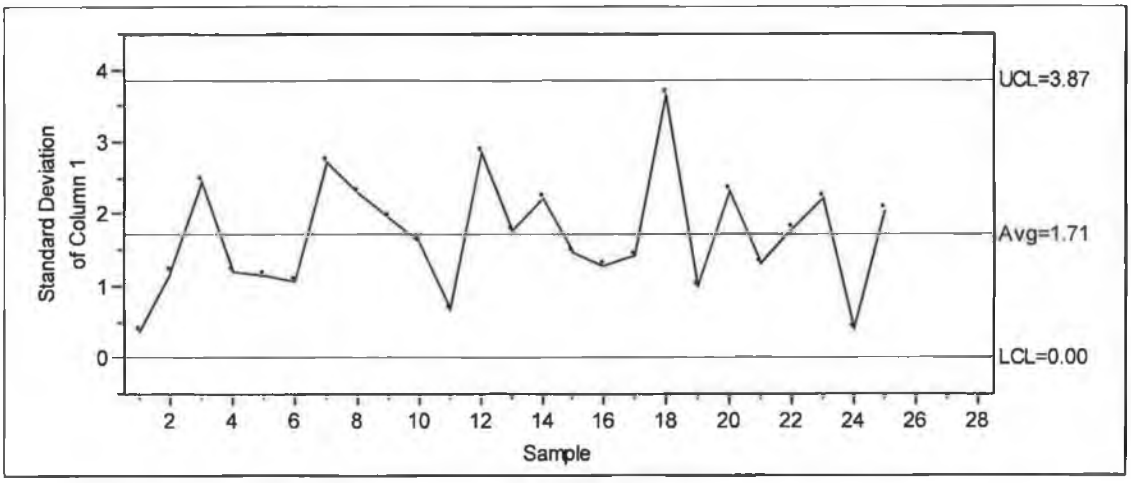


Figure 2.3 S Chart for plotting Standard Deviation within groups

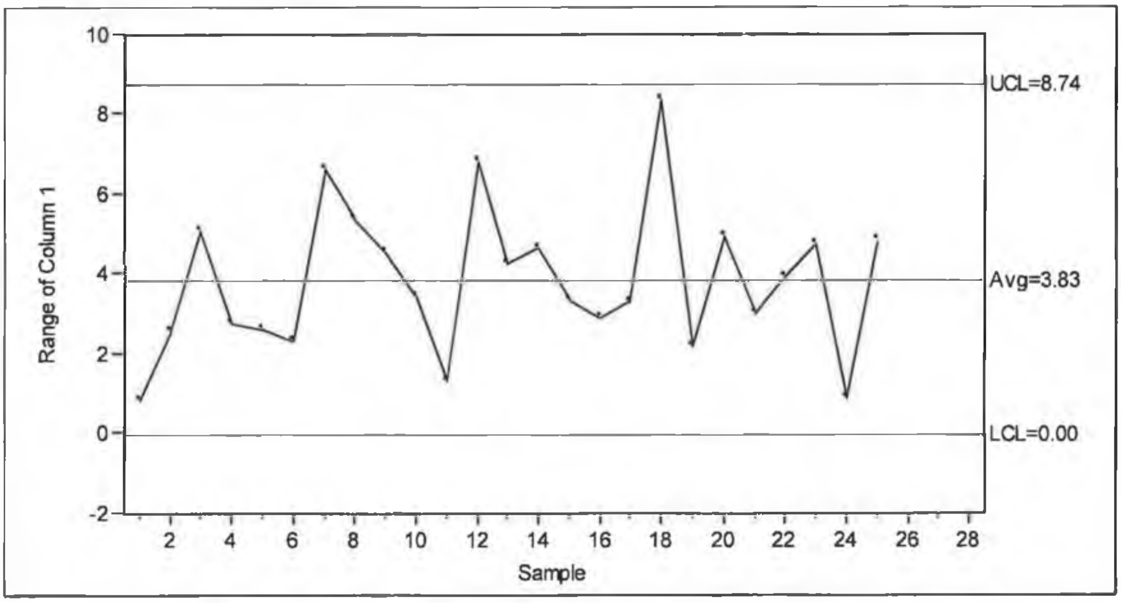


Figure 2.4 R Chart for plotting Range within groups

In industry, pairs of charts, ($\bar{x}$, S) or ($\bar{x}$, R) are monitored simultaneously, for each variable, to alert the user to special causes of variation in the process variable being measured.

These types of charts are referred to as univariate, so-called as the chart contains data where only one process variable is plotted on the chart at a time.

What if two process variables were to be monitored? Typically, univariate charts would be generated to monitor mean and dispersion for each variable, resulting in many control charts being monitored. The number of charts being monitored increases as the number of variables increases. Figure 2.5 and 2.6 shows univariate control charts for two variables, $X1$ and $X2$. Only the means charts are shown but when monitoring the process, the dispersion charts must also be monitored.

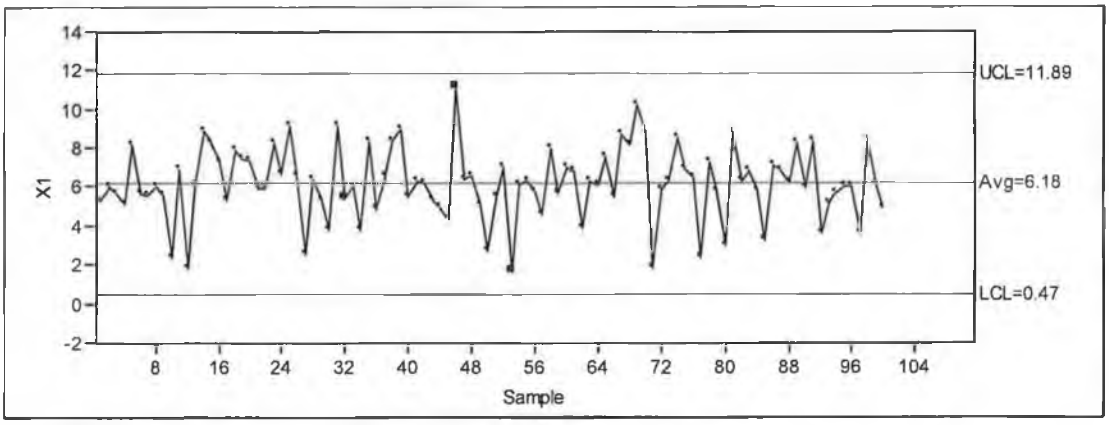


Figure 2.5 Univariate mean control chart for $X1$

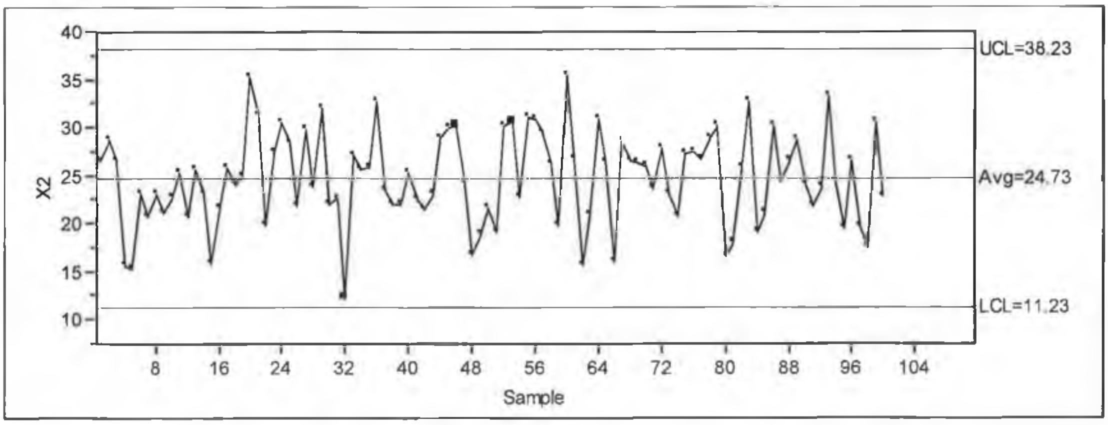


Figure 2.6 Univariate mean control chart for $X2$

When some relationship exists between two variables, this cannot be shown on individual univariate control charts, therefore it is best to use some control region that can assess the relationship between these variables and then determine if they are in-control. Mason and Young (2002) describe that a control ellipse is used to show the true control region for correlated variables. Correlation measures the strength of the relationship between two variables. Correlation will be discussed in more detail later in the chapter, but the equation for constructing an ellipse will be described here.

They described that in order to construct an ellipse for two variables, their correlation measure must be calculated. The variances of each variable, s_1 and $s_{2,}$ is required as well as the covariance of both variables, s_{12}. The covariance measures how the variables vary together. Equation 1.0 is the equation of an ellipse for two variables.

$$\frac{1}{1-r^2}\left[\left(\frac{x_1-\bar{x}_1}{s_1}\right)^2 - 2r\left(\frac{x_1-\bar{x}_1}{s_1}\right)\left(\frac{x_2-\bar{x}_2}{s_2}\right) + \left(\frac{x_2-\bar{x}_2}{s_2}\right)^2\right] \qquad (1.0)$$

where $r = \frac{s_{12}}{s_1 s_2}$ is the sample correlation coefficient.

The more correlated the variables are to one another, the more tilted the elliptical region will be.

Figure 2.7 shows a scatterplot of the two variables, $X1$ and $X2$, and the elliptical region surrounding the data points. The UCL and LCL for $X1$ and $X2$ are also displayed in Figure 2.7 by the box area around the elliptical region. These were the limits determined in the univariate control charts in Figures 2.5 and 2.6 respectively.

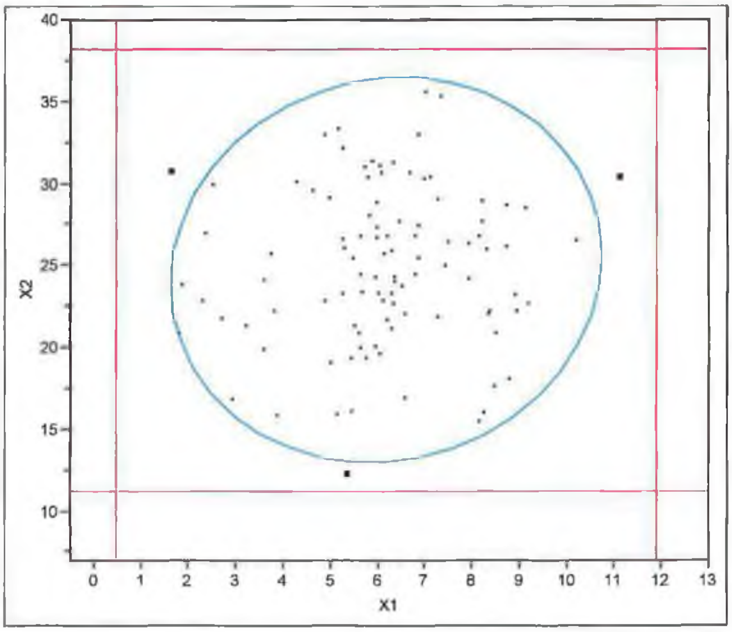


Figure 2.7 Control Ellipse with univariate control limits

All data points fall within the rectangular region, which is determined by their upper and lower control limits. This shows that the process seems in control when treated as univariate data. However there are three points identified that lie outside the elliptical region. This shows a violation of the correlation structure of the $X1$ and $X2$ variables. This relationship is not analysed in the univariate charts. This is seen by the points in bold in Figure 2.5 and 2.6. These points are within the control limits. Therefore, multivariate techniques must be used instead of univariate techniques in order to highlight this type of violation.

In most industries where SPC is utilised, univariate process control methods are more commonly known. This is because it is the process control method that is taught either in university or through training within an industry.

As technology moves further into the future, process data is collected electronically via online and offline computer systems and a lot more frequently than before, even in real time. It would be impossible to look at each individual observation for all process variables. It is also not practical to summarise into groups and look at these control charts as it could still involve hundreds of variables.

To look further into this data requires a little more knowledge and research. The monitoring of two or more process variables is referred to as Multivariate Statistical Process Control (MSPC), and it is possible to monitor several variables simultaneously on one control chart. Most data analysis tools deal primarily with univariate data, however, there are more and more software packages with updates that now include Multivariate data analysis.

A lot of the complicated calculations can be done very quickly through such software packages but you need to have some knowledge about Multivariate data in order to use these.

Multivariate control involves summarising each of the variables in a multivariate space into a univariate statistic and plotting this univariate statistic on a control chart.

CPACT (2008) in Newcastle University, UK, note that some desired characteristics of Multivariate Control are

- Ease of application
- Signal interpretation

- Sensitivity to subtle shifts
- Real time monitoring
- Appropriate software for both detection and interpretation

The application of Multivariate techniques can contribute to an improved understanding of a process, close monitoring of a process performance, early detection of defective product and subsequently cost reduction.

As with most distributions, the primary objective is to describe the mean and variance of a population using certain assumptions about the distribution of the particular process variable being monitored. Montgomery (2005) described that the mean, **μ**, measures the centre of the distribution while the dispersion is measured by variance, $\boldsymbol{\sigma}^2$. In the univariate case, the normal population is described as:

$$X \sim N\ (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$$

Univariate statistics deal with *n* observations of single variables using the normal population distribution.

When using a sample from a population, the more appropriate distribution used to describe the mean and variance of the distribution is known as the Students *t* distribution. A random sample of *n* observations are taken from a population. The sample mean, $\bar{x}$, measures the centre of this distribution while the spread is measured by the sample standard deviation, s. The equation is written as:

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \qquad (1.1)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$.

If Equation 1.1 is squared, then t becomes,

$$t^2 = (\bar{x} - \mu)^2/(s^2/n)$$

$$= n(\bar{x} - \mu)(s^2)^{-1}(\bar{x} - \mu) \qquad (1.2)$$

The univariate t statistic can be extended to the Multivariate case. The Multivariate Normal (MVN) population is described by,

$$X \sim N_p(\mu, \Sigma)$$

where μ is the mean vector and Σ is the covariance matrix.

With Multivariate, the mean is also measured, but it relates to a mean vector, as there are multiple variables, p, and these are given by $x_1, x_2, \dots, x_p$.

$$\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix}$$

The observation vector, X, contains all the information about the variables with mean vector, μ, and the covariance matrix, $\sum$, contains information about the variation of the variables and their relationships.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

which can also be expressed as $X' = (x_1, x_2, \dots, x_p)$.

The dispersion is measured using the variance-covariance matrix, $\sum$, also known as the covariance matrix.

$$\sum = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots\dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots\dots & \sigma_{2p} \\ .. & .. & .. & .. \\ \sigma_{p1} & \sigma_{p2} & \dots\dots & \sigma_{pp} \end{pmatrix}$$

This matrix has diagonals that represent the variance of the i^{th} variable, σ_{ii}. The off-diagonals, σ_{ij}, represent the covariance between the i^{th} and j^{th} variables. This represents how each pair of variables is related.

Suppose $X' = (x_1, x_2, \ldots, x_p)$ is a p-variate normal probability distribution with the vector of the means of X' given by $\mu' = (\mu_1, \mu_2, \ldots, \mu_p)$, and the covariance matrix, Σ.

If there is a sample of n observations, $X_1, X_2, \ldots, X_n$, where $X_i = (x_{i1}, x_{i2} \ \ldots, x_{ip})$, $i = 1, 2, \ldots, n$, with mean vector μ' and a covariance matrix Σ, the Multivariate generalisation of the t^2 statistic in Equation 1.2 becomes

$$T^2 = n(\bar{X} - \mu)' S^{-1} (\bar{X} - \mu) \tag{1.3}$$

where $\bar{X}$ and S are sample estimators of μ and Σ.

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and the sample covariance matrix, $S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$, can also be expressed as,

$$S = \begin{pmatrix} s_{11} & s_{12} & \ldots\ldots & s_{1p} \\ s_{21} & s_{22} & \ldots\ldots & s_{2p} \\ .. & .. & .. & .. \\ s_{p1} & s_{p2} & \ldots\ldots & s_{pp} \end{pmatrix}$$

This matrix has diagonals that represent the sample variance of the i^{th} variable, s_{ii}, and the off-diagonals, s_{ij}, represent the sample covariance between the i^{th} and j^{th} variables. The covariance will only give an indication of their linear relationship, being either positive or negative.

The covariance is often standardised as it is difficult to interpret how strongly the variables are related. The Correlation Coefficient measures the strength of the relationship between the variables. The correlation coefficient measure simplifies the covariance so that the effect of scale is removed. The correlation coefficient is calculated by dividing the standard deviations of the two variables and is denoted by ρ_{ij}.

$$\rho_{ij} = \frac{covariance\ \ (i,j)}{st.dev\,(i)*st.dev\,(j)} = \frac{s^2{}_{ij}}{s_{ii}\ s_{jj}} \qquad (1.4)$$

This can be represented in an *nxp* matrix called the Correlation Matrix,

$$R = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots\cdots & \rho_{1j} & \cdots\cdots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \cdots\cdots & \rho_{2j} & \cdots\cdots & \rho_{2p} \\ .. & .. & .. & .. & .. & .. & .. \\ \rho_{j1} & \rho_{j2} & \cdots\cdots & \cdots\cdots & 1 & \cdots\cdots & \rho_{jp} \\ . & .. & .. & .. & .. & .. & .. \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \cdots\cdots & \rho_{pj} & \cdots\cdots & 1 \end{pmatrix}$$

If two variables are independent then their covariance and hence, their correlation, will be zero. The reverse of this is not necessarily true. Zero correlation does not imply independence, they could possibly be correlated in a non linear way. Transformation of one of the variables could show a linear relationship.

The cross-correlation between two variables measures the linear dependency between them. This can reveal strong relationships between variables.

1. If there are two independent variables that are highly correlated then it can be justified that one of them could be eliminated as they are both providing the same information.

2. A true relationship would be indicated between an independent and a dependent variable.

3. If a relationship is seen between an independent variable and a residual, this would indicate a goodness of fit in a model.

These are just some considerations that should be taken into account when looking at Multivariate data.

2.2 Distance Statistics

This section describes the various distance measurements and how these relate to Multivariate SPC. Statistical distance calculations such as Euclidean (straight line) distance, statistical distance (elliptical distance), Mahalanobis distance are described and one of the most widely used multivariate distributions, Hotellings T^2 distribution which is similar to Mahalanobis distance.

2.2.1 Euclidean Distance and Statistical Distance

In order to understand multivariate control, the concept of statistical distance must be considered. This incorporates the means, variances and covariances of variables and how stable these are as each observation is added over time. For uncorrelated variables, i.e. $\rho = 0$. Consider two observations each with (x, y) coordinates. When plotted, these two data points will have a mean point, and a line through all three data points should show some linear relationship between each of the data points and the mean.

If this mean point truly represents the population mean then the straight line distance between a point (x_1, x_2) and this population point (μ_1, μ_2) is calculated. This is the distance statistic, D and using Pythagoras' Theorem is given by

$$D = \sqrt{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2} \qquad (1.5)$$

This is known as the Euclidean (straight line) distance. Squaring each side, this can be written as

$$D^2 = (x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 \qquad (1.6)$$

If this distance is fixed, then all points that are the same distance from the mean can be represented by a circle around the mean. This centre point is the centroid with radius D.

Mason and Young (2002) highlighted that any point inside this circle has distance to the mean less than D. Any point outside this circle has distance to the mean greater than D. This method ignores the variation between the variables and so it is insufficient in an n-dimensional space.

However, consider the variance of the data, Equation 1.6 can be standardised to be

$$SD^2 = \frac{(x_1 - \mu_1)^2}{\sigma_1{}^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2{}^2} \qquad (1.7)$$

This is known as the Statistical Distance or elliptical distance, as standardised data are enclosed in an ellipse. As with the Euclidean distance, if SD is fixed, then all the points are the same statistical distance from the mean point and so is an ellipse, Figure 2.8.

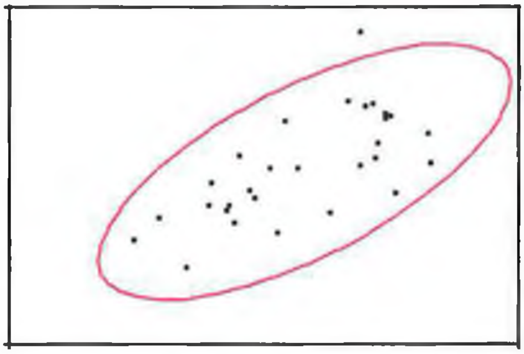


Figure 2.8 - Ellipse

Any point inside the ellipse has statistical distance less than SD and all points outside the ellipse have statistical distance greater than SD. If the variances of x_1 and x_2 were equal then the Euclidean and Statistical Distances would be the same. SD is simply a weighted straight line measure of distance.

An elliptical region of control is easily created for distributions that contain two characteristics, but if more than two characteristics were to be measured it becomes difficult to construct and it will have a multi-dimensional ellipse.

2.3 Hotellings T^2

2.3.1 Hotellings T^2 Statistic

Hotellings T^2 Statistic (1931) is based on a generalisation of the Students t statistic previously seen in Equation 1.3. It is a distance measure that will consider the covariance structure of the MVN. The T^2 distance statistic is given as,

$$T^2 = (X - \bar{X})'S^{-1}(X - \bar{X}) \qquad (1.8)$$

where X is the observation vector,

$\bar{X}$ is the mean vector, and

S is the covariance matrix.

Mason and Young (2002) describe that Equation 1.8 represents the squared distance from each observation to the group mean. Thus, the variability around the group mean can be calculated by summing the T^2 statistic for each observation.

2.3.2 Hotellings T^2 Distribution

Hotellings T^2 distribution was introduced by Harold Hotelling (1947). He was one of the first to design a control chart around the concept of Multivariate SPC. His idea was to use one control chart that would plot one statistic representing information for the dispersion and mean of several quality characteristics. This chart would have an upper control limit, which would indicate when a process goes out of control. However, it wouldn't indicate which characteristic contributed to the out of control condition. This would have to be investigated using additional methods. This section will describe the Hotelling T^2 method for monitoring a Multivariate process and review some methods for detecting the variables that contributed to the out of control situation.

The properties of the T^2 distribution, as detailed by Mason and Young (2002), will depend on a probability function and whether the parameters are known or unknown.

The T^2 is a univariate statistic which describes the corresponding Multivariate distribution, where the p-variate observations must be transformed into a single Hotelling T^2 statistic.

In univariate statistics, the *z* distribution is the population distribution and estimates of population parameters are described in the *t* distribution. The *t* statistic is described by the *t* distribution with (*n*-1) degrees of freedom. The t^2 statistic can be described by an F distribution with 1 and (*n*-1) degrees of freedom.

This is the same for the Multivariate case, if the parameters of the MVN distribution are unknown, a form of the univariate F is used to describe the T^2 statistic.

2.3.3 T^2 Distributions for Individual observations

Mason and Young (2002) have shown that for an individual observation vector X, assuming that $\boldsymbol{\mu}$ **and** $\sum$ **are known** in the MVN distribution, T^2 has similar distribution to the Chi-square distribution, χ^2.

$$T^2 = (X - \mu)'\Sigma^{-1}(X - \mu) \sim \chi^2{}_{(\mathrm{p})}, \qquad (1.9)$$

where $\chi^2{}_{(\mathrm{p})}$ represents a chi-square distribution with p degrees of freedom.

The T^2 distribution depends only on p, the number of variables in the observation vector X. For smaller p, the shape of the distribution is skewed with a long tail to the right. For large p, the distribution looks more symmetrical.

Tracey *et al.* (1992) have shown that if $\boldsymbol{\mu}$ **and** $\sum$ **are unknown**, and the observation vector X is **independent** of $\bar{X}$ and S, i.e. the observation vector X is not included in the calculation of $\bar{X}$ and S, parameter estimates are obtained using a historical baseline consisting of n observations. The T^2 statistic is given by,

$$T^2 = (X - \bar{X})'S^{-1}(X - \bar{X}) \sim \mathrm{F}_{(\mathrm{p,n-p}),} \qquad (1.10)$$

where $\mathrm{F}_{(\mathrm{p,n-p}),}$ is an F distribution with p and (n-p) degrees of freedom.

The T^2 distribution depends on the number of variables, p and the sample size, n. The shape of the distribution is skewed with a long tail to the right.

Sullivan and Woodall (1996) describe that if $\boldsymbol{\mu}$ **and** $\sum$ **are unknown**, and the observation vector X is **not independent** of $\bar{X}$ and S, i.e. the observation vector X is included in the calculation of $\bar{X}$ and S, parameter estimates are obtained using a historical baseline consisting of n observations.

The T^2 statistic is given by,

$$T^2 = (X - \bar{X})' S^{-1} (X - \bar{X}) \sim \left[\frac{(n-1)^2}{n}\right] \beta_{(p/2,(n-p-1)/2)} \qquad (1.11)$$

where $\beta_{(p/2,(n-p-1)/2)}$ is a beta distribution with parameters $p/2$ and $(n - p - 1)/2$.

The T^2 distribution depends on the number of variables, p and the sample size, n. The shape of the beta distribution can look like other familiar distributions such as the normal, chi-square and F distributions.

Mason and Young (2002) pointed out that as the beta distribution can take the form of the F distribution, the beta distribution is generally used to describe the T^2 distribution,

$$\beta_{(p/2,(n-p-1)/2)} = \frac{pF}{(n-p-1)pF}, \qquad (1.12)$$

where

$$F \sim F_{(p,n-p-1)}. \qquad (1.13)$$

If this form of the F distribution is used, the observation vector X is not independent of $\bar{X}$ and S. Either form is acceptable where μ and $\sum$ are unknown.

2.3.4 T^2 Distribution for Subgrouped Data

The above distributions are used for individual observation vectors. Yang and Trewn (2004) consider a distribution for monitoring the mean of a **subgroup** of size m observations taken at k sampling intervals. It is assumed that the observation vector $\bar{X}_i$ is independent of the sample estimates $\bar{X}$ and S, which are obtained using a historical baseline consisting of n observations.

Ryan (1989) defines the Statistical Distance, T^2 statistic, between the sample mean of the i^{th} observation vector, $\bar{X}_i$ and the baseline data mean, $\bar{X}$, as,

$$T^2 = (\bar{X}_i - \bar{X})' S^{-1} (\bar{X}_i - \bar{X}) \sim \frac{(m+n)(n-1)p}{mn\,(n-p)} F_{(p,n-p)} \qquad (1.14)$$

where $\bar{X}_i$ is the sample mean of the i^th^ observation vector, and

$\bar{X}$ and S are the sample estimates.

This T^2 distribution depends on the number of variables, p in $\bar{X}_i$, the sample size of the subgroup, m, and the size of the baseline with n observations.

2.4 Hotellings T^2 Control

For Phase I in process control, a T^2 control procedure where there are unknown parameters will be generated. Data is collected from a historical baseline when the process is in-control. Estimates of these unknown parameters can then be produced from this dataset.

2.4.1 Phase I and Phase II UCL for Individual observations.

Univariate process monitoring uses 3 sigma limits on their control charts for determining upper and lower control limits (UCL & LCL). That is not the case for Multivariate process monitoring. A UCL is determined by a critical value of the probability function used to describe the distribution of the data. A value of α is chosen to minimise the error of a false reject i.e. α error is the probability of stating that a process is out of control when in fact it is not. On univariate control charts, the control limits are set at ±3 sigma. This fixes the false alarm rate, α, at a value of 0.0027. Tracy *et al.* (1992) identify that the LCL can be set to zero in certain situations. This is because, if there is a shift in the mean, the T^2 statistic will increase and so the LCL can be disregarded.

Similar to univariate procedures, there are two phases in Multivariate process monitoring. Woodall (2000) discusses Phase I and Phase II in great detail and what is involved in the transition from one to the other.

Phase I involves the analysis of historical data in order to obtain a baseline and target, or UCL, in which to use in Phase II.

Phase II is where real-time process monitoring is applied, where it is determined if a new observation is part of the baseline dataset characterised in Phase I.

As Phase I involves creating an in-control baseline dataset, which will be used to describe how the process performs in normal operating conditions, the dataset must be constructed by in-control conditions. Therefore, it is necessary to purge the dataset of observations that appear to be inconsistent with the rest of the data, i.e. outliers. Outliers that indicate abnormal conditions are removed, because if these are included, it will inflate the UCL. Outliers will increase the variation of the variables but will have little effect on their correlation. A small number of samples containing outliers will have a large effect on μ and $\sum$.

In a Multivariate system, outliers indicate that an observation is not conforming to the group not just the individual variable being monitored. The removal of outliers must be carefully justified with root cause determined. Not all outliers indicate a nonconforming observation, it may be a genuine observation.

In Phase I application, all observation vectors whose T^2 values are > UCL are removed from the dataset and new estimates of μ and $\sum$ are calculated. This process of purging outliers is repeated until there are no more outliers identified after the recalculation of estimates of μ and $\sum$ and UCL.

Chenouri *et al.* (2009) also explain that observations found to be outside the control limits in Phase I must be investigated. If there is an assignable cause, they are removed, as including them can lead to reduced power for detecting a process change in Phase II as well as inflated control limits. When the unusual observation has been eliminated, the control limits are recalculated.

Mason and Young (1999) describe the importance of creating a representative baseline (model) in Phase I. The better the model fit, the more sensitive the T^2 control method will be to outliers. More emphasis should be placed into constructing a baseline dataset in Phase I. This will ensure that the appropriate process variables are used, ones which characterise the process the best, and in turn a useful model is created from which Phase II can be based upon.

Calculating the UCL for a process also has some additional considerations. Is the process in Phase I or Phase II? Are the population parameters, μ and $\sum$, known, or will sample estimates $\bar{X}$ and S, have to be calculated?

2.4.1.1 *Phase I Control Limits, where μ and $\sum$ are known population parameters*

For a chi-square distribution, where μ and ∑ are known, assuming multivariate normality, from Equation 1.9, the UCL for

$$T^2 = (X - \mu)'\textstyle\sum^{-1}(X - \mu) \text{ is,}$$

$$T^2{}_{UCL} = \chi^2{}_{(\alpha,p)} \qquad (1.15)$$

for chosen α, where $\chi^2{}_{(\alpha,p)}$ is the upper α^{th}quantile of $\chi^2{}_{(p),}$ with p degrees of freedom.

Tracy *et al.* (1992) discussed that for a Phase I "start-up stage", an exact method should be used to construct the control chart. The χ^2 distribution, as in Equation 1.15, and an F distribution are approximated distributions. These approximations and subgroups being small, i.e. n=1 for individual observations, the associated degree of error is unknown.

2.4.1.2 *Phase I Control Limits, where μ and $\sum$ are unknown*

Tracy *et al.* (1992), Lowry and Montgomery (1995) stated that calculating control limits for individual observations of a Phase I process, the exact procedure follows a beta distribution. Assuming multivariate normality, from Equation 1.11, it can be shown that the UCL for

$$T^2 = (X - \bar{X})'S^{-1}(X - \bar{X}) \text{ is,}$$

$$T^2{}_{UCL} = \frac{(n-1)^2}{n}\, \beta_{(\alpha,p/2,(n-p-1)/2)} \qquad (1.16)$$

for chosen α, where $\beta_{(\alpha,p/2,(n-p-1)/2)}$ is the upper α^{th}quantile of $\beta_{(p/2,(n-p-1)/2)}$.

Example 2.1, Phase I

Equation 1.11 is used to calculate the T^2 values and Equation 1.16 to calculate the UCL for the distribution.

This example contains start-up stage data in a Chemical Process generated from Tracy *et al.* (1992). Three variables are used in monitoring the Chemical Process. The three variables are Percentage impurities (X_1), Temperature (X_2) and Concentration (X_3). The initial sample has 14 observations shown in Table 2.1.

Sample No.	%Impurities (X_1)	Temperature (X_2)	Concentration (X_3)
1	14.92	85.77	42.26
2	16.9	83.77	43.44
3	17.38	84.46	42.74
4	16.9	86.27	43.6
5	16.92	85.23	43.18
6	16.71	83.81	43.72
7	17.07	86.08	43.33
8	16.93	85.85	43.41
9	16.71	85.73	43.28
10	16.88	86.27	42.59
11	16.73	83.46	44
12	17.07	85.81	42.78
13	17.6	85.92	43.11
14	16.9	84.23	43.48
$\bar{X}$	**16.83**	**85.19**	**43.21**

Table 2.1 – Chemical Process Data

$$\bar{X} = (16.83, 85.9, 43.21)$$

and the sample covariance matrix is,

$$S = \begin{bmatrix} 0.365 & -0.022 & 0.10 \\ -0.022 & 1.036 & -0.245 \\ 0.10 & -0.245 & 0.224 \end{bmatrix}$$

Using Equation 1.11,

$$T^2 = (X - \bar{X})' S^{-1} (X - \bar{X})$$

$$T^2{}_1 = (14.92 - 16.83, 85.77 - 85.19, 42.26 - 43.21) \begin{bmatrix} 0.365 & -0.022 & 0.10 \\ -0.022 & 1.036 & -0.245 \\ 0.10 & -0.245 & 0.224 \end{bmatrix}^{-1} \begin{bmatrix} 14.92 - 16.83 \\ 85.77 - 85.19 \\ 42.26 - 43.21 \end{bmatrix} = 10.93$$

$T^2{}_2 = 2.01$, $T^2{}_3 = 5.58$, $T^2{}_4 = 3.86$, $T^2{}_5 = 0.04$, $T^2{}_6 = 2.25$, $T^2{}_7 = 1.44$, $T^2{}_8 = 1.21$, $T^2{}_9 = 0.68$, $T^2{}_{10} = 2.17$, $T^2{}_{11} = 4.17$, $T^2{}_{12} = 1.40$, $T^2{}_{13} = 2.33$, $T^2{}_{14} = 0.90$.

Using Equation 1.16 to calculate the UCL, $n = 14$, $p = 3$,

$$T^2{}_{UCL} = \frac{(n-1)^2}{n}\,\beta_{(\alpha,p/2,(n-p-1)/2)}\,,$$

$$T^2{}_{UCL} = \frac{(14-1)^2}{14}\,\beta_{(0.05,3/2,(14-3-1)/2)}$$

$$= 12.07\beta_{(0.05,1.5,5)}$$

$$= 12.07(0.5266) = 6.36$$

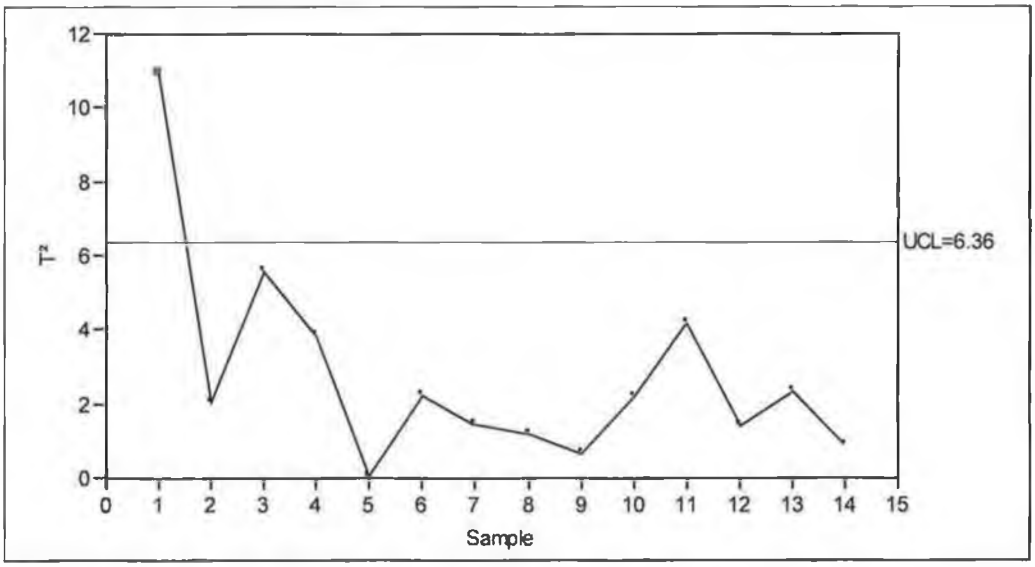


Figure 2.9 – Phase I T^2 Control Chart

From the T^2 calculations above, $T^2{}_1$ shows that Sample 1 is out of control, this is also demonstrated by the graph in Figure 2.9. Sample 1 was determined to be a measurement error and so can be removed from the calculations. There are now 13 samples used to calculate the UCL for Phase I.

$$\bar{\bar{X}} = (16.98, 85.14, 43.28),$$

the sample covariance matrix is,

$$S = \begin{bmatrix} 0.068 & 0.076 & -0.055 \\ 0.076 & 1.092 & -0.216 \\ -0.055 & -0.216 & 0.163 \end{bmatrix}$$

$T^2{}_1 = 1.84$, $T^2{}_2 = 5.33$, $T^2{}_3 = 3.58$, $T^2{}_4 = 0.23$, $T^2{}_5 = 2.17$, $T^2{}_6 = 1.46$, $T^2{}_7 = 1.05$, $T^2{}_8 = 1.91$, $T^2{}_9 = 5.16$, $T^2{}_{10} = 3.84, T^2{}_{11} = 1.65$, $T^2{}_{12} = 7.00$, $T^2{}_{13} = 0.77$.

Note: For calculation purposes, Samples are numbered 1-13. To convert to original Sample numbers add 1 (as Sample 1 has been removed).

$$T^2{}_{UCL} = \frac{(13-1)^2}{13}\,\beta_{(0.05,3/2,(13-3-1)/2)}$$

$$= 11.08\beta_{(0.05,1.5,4.5)}$$

$$= 11.08(0.562) = 6.23\,.$$

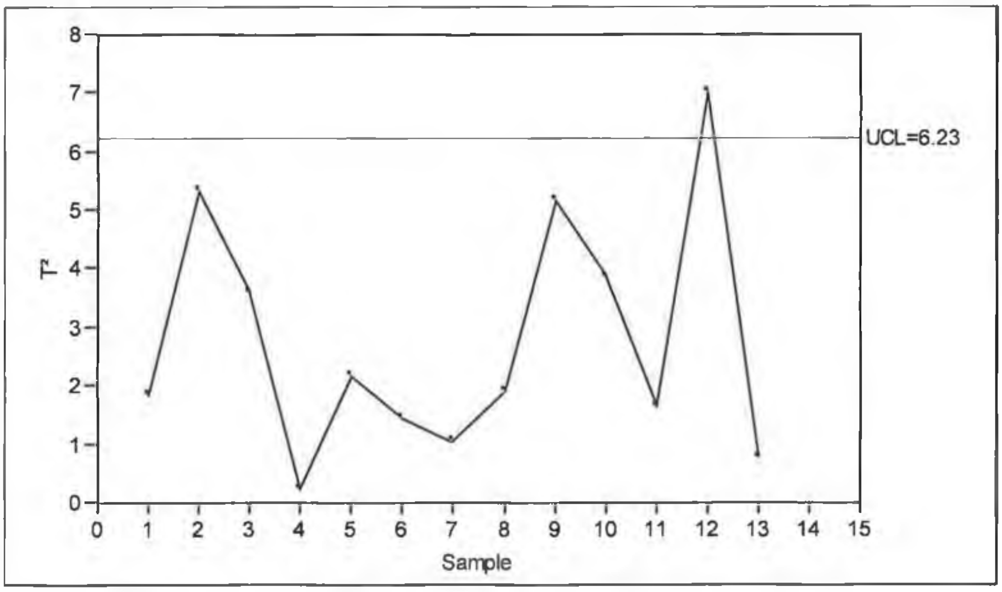


Figure 2.10 – T^2 Control Chart with Sample 1 removed

Although, $T^2{}_{12}$ = 7.00, shows that it is out of control in Figure 2.10, there is no assignable cause for this observation and so it is not removed.

Now that the baseline dataset has been determined, the Phase II UCL is used for monitoring new observations.

2.4.1.3 *Phase II Control Limits, where μ and ∑ are known population parameters*

When μ and ∑ are known for individual observations and there is an existing steady state process, the $T^2{}_{UCL}$ is used as in Phase I (Equation 1.15).

$$T^2 = (X - \mu)'\Sigma^{-1}(X - \mu) \text{ and,}$$

$$T^2{}_{UCL} = \chi^2{}_{(\alpha,p)}$$

To justify use of this equation, Lowry and Montgomery (1995) suggest for Phase I and Phase II control limit calculations, where p is large, n should exceed 500. With a large sample size from Phase I, it is assumed that $\bar{\bar{X}}$ and S are equal to the true population parameters μ and ∑. These parameters are approximations and therefore using this method gives an approximate calculation of the control limit.

The control limit is independent of the size of n used to determine the baseline dataset in Phase I.

This method is not usually recommended, as in industry, it is rarely the case that you will have known population parameters.

2.4.1.4 *Phase II Control Limits, where μ and ∑ are unknown*

The application of the Phase II process monitoring is that each incoming data point is plotted in sequence, and is independent. Equation 1.17 is used, where the incoming observation vector X is not included in the calculation of $\bar{\bar{X}}$ and S.

Where the population parameters, μ and ∑ are unknown, Equation 1.10, is used to calculate the T^2 statistic,

$$T^2 = (X - \bar{\bar{X}})'S^{-1}(X - \bar{\bar{X}})$$

Ryan (1989) defines the exact UCL for this T^2 statistic as,

$$T^2{}_{UCL} = \frac{(n+1)(n-1)p}{n(n-p)} \mathrm{F}_{(\alpha,p,n-p)}, \qquad (1.17)$$

for given α, where n is the size of the baseline dataset from Phase I,

p is the number of variables and

$\mathrm{F}_{(\alpha,p,n-p)}$ is the α^{th}quantile of $\mathrm{F}_{(p,n-p)}$.

This method calculates the exact control limits as the exact distribution of T^2 is obtained from Equation 1.10.

If the new incoming observation is greater than the UCL this will indicate a signal implying that it does not conform to the established baseline dataset. Mason *et al.* (2003) discussed that it is at this point where it is decided, whether to react to each out of control point, or to wait and search for some trend or pattern and declare an out of control when a number of T^2 values are $> UCL$. These are similar to run rules used on Shewhart Control Charts in univariate process monitoring.

Example 2.2

Using the 13 samples from Phase I data from Example 2.1 above, a new observation, sample 14, is collected, $X_{new} = (17.08, 84.08, 43.81)'$.

The T^2 Statistic is calculated using,

$$T^2 = (X_{new} - \bar{X})' S^{-1} (X_{new} - \bar{X})$$

where $\bar{X} = (16.98, 85.14, 43.28)'$, and $S = \begin{bmatrix} 0.068 & 0.076 & -0.055 \\ 0.076 & 1.092 & -0.216 \\ -0.055 & -0.216 & 0.163 \end{bmatrix}$

$$T_{new}^2 = (17.08 - 16.98, 84.08 - 85.14, 43.81 \\ - 43.28) \begin{bmatrix} 0.068 & 0.076 & -0.055 \\ 0.076 & 1.092 & -0.216 \\ -0.055 & -0.216 & 0.163 \end{bmatrix} \begin{bmatrix} 14.92 - 16.83 \\ 85.77 - 85.19 \\ 42.26 - 43.21 \end{bmatrix} = 3.52$$

Using Equation 1.17 to calculate if this new T^2 value is in control,

$$T^2{}_{UCL} = \frac{(n+1)(n-1)p}{n(n-p)} \mathrm{F}_{(\alpha,p,n-p)},$$

$$T^2{}_{UCL} = \frac{(13+1)(13-1)3}{13(13-3)} \mathrm{F}_{(0.05,3,13-3)},$$

$$= 3.88\, \mathrm{F}_{(0.05,3,10)},$$

$$= 3.88(3.708) = 14.38$$

Figure 2.11 shows a graph of the T^2 Statistics from Phase I (samples 1 - 13), the T^2 Statistic for the new observation (sample 14) and the UCL for Phase II.

T^2_{new} is within the control limits, therefore the new observation, X_{new} , is in control.

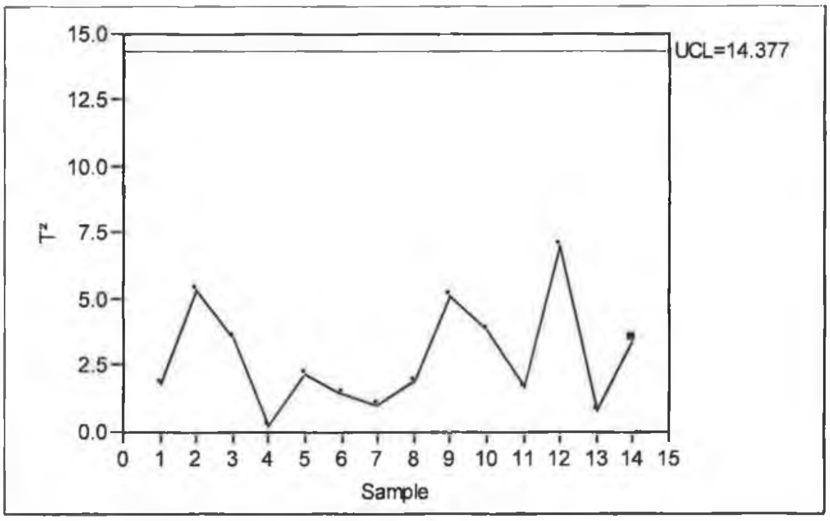


Figure 2.11 – Phase II T^2 Control Chart

2.4.2 Phase I and Phase II UCL for Subgrouped Data

A more common SPC method is to monitor the **subgroup** means.

The mean vector $\bar{\bar{X}}$ of a sample of m observations is distributed as a p-variate normal distribution $N_p(\mu, \Sigma/m)$.

Using subgroup data, the method for calculating the T^2 Statistic and the UCL in Phase I is the same as the method used for individual observations. Mason and Young (2002) noted that the data is in samples of size $\mathrm{m_i}$, $\mathrm{i} = 1,2, \ldots . \mathrm{k}$. The total sample size will be $\mathrm{n} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{m_i}$. The observations can be treated as one group as all the observation vectors come from the same Multivariate Normal distribution.

2.4.2.1 *If μ and ∑ are known,*

For Phase II monitoring of subgroups, where μ and ∑ are known, the T^2 Statistic is calculated by,

$$T^2 = m(\bar{X}_i - \mu)' \, \Sigma^{-1}(\bar{X}_i - \mu) ,$$

where $\bar{X}_i$ is the i^{th} sample mean,

The UCL for a given α is the same as for individual data shown in Equation 1.15,

$$T^2{}_{UCL} = \chi^2{}_{(\alpha,p)}$$

This control limit is independent of the sample size of the subgroups, m and the baseline dataset n.

2.4.2.2 *If μ and ∑ are unknown,*

In Phase II, where μ and ∑ are unknown, the UCL for the T^2 Statistic, given in Equation 1.14,

$$T^2 = (\bar{X}_i - \bar{\bar{X}})' S^{-1}(\bar{X}_i - \bar{\bar{X}}) ,$$

where $\bar{\bar{X}}$ and S are estimates of μ and ∑ from the baseline data, is,

$$T^2{}_{UCL} = \frac{(m+n)(n-1)p}{mn\,(n-p)} F_{(\alpha,p,n-p)} \qquad (1.18)$$

for a given α. n is the size of the baseline dataset.

Example 2.3

Using the new observation data in Example 2.2, $X_{new} = (17.08, 84.08, 43.81)'$, after the baseline dataset has been determined from the 13 samples in Phase I, suppose that the new observation and the 13 samples represent a subgroup mean i.e. each sample is the mean of 5 observations, then $m = 5$. The T^2 value for the new observation would remain the same but the UCL would be calculated as follows,

From Equation 1.18,

$$T^2{}_{UCL} = \frac{(m+n)(n-1)p}{mn(n-p)} F_{(\alpha,p,n-p)}$$

$$= \frac{(5+13)(13-1)3}{5(13)(13-3)} F_{(0.05,3,13-3)}$$

$$= \frac{(18)(12)3}{5(13)(10)} F_{(0.05,3,10)}$$

$$= \frac{648}{650}(3.708) = 3.70$$

T^2_{new} = 3.52, is less than $T^2{}_{UCL}$ = 3.70, and therefore is in control.

2.5 Autocorrelation

If a time dependency exists in the data (such as decay process), it is known as time-correlation or autocorrelation. For data where autocorrelation exists, an adjustment must be made to the T^2 Statistic as it is based on the assumption that independent observations exist in the data i.e. the observation vectors must be uncorrelated over time.

Montgomery and Mastrangelo (1991) discussed that using the T^2 Statistic without appropriately adjusting for a time dependency can result in a false signal.

Mason and Young (2002) described two forms of autocorrelation in a process,

1. Continuous Decay – Correlation exists in the process where the current observation is dependent on an immediate preceding value.

2. Stage Decay – The performance in one stage of the process is dependent on the performance in the previous stage. This can be seen, for example, on equipment wear and tear. There is a step-wise pattern over an extended period of time.

In univariate processes, adjustments are made so that the confounding effect from one variable with correlated variables is removed.

In multivariate processes, an investigation is carried out, to look at how the time variables relate to the other process variables, so they are not confounded. This is performed by adding a variable that is time sequenced, into the dataset, and observing how other process variables relate to it. If they correlate, then it is probable that the process variable will correlate with itself in time.

Bersimis *et al.* (2007) list various papers written on methods used for dealing with autocorrelated multivariate processes.

2.6 Collinearity

Collinearity is where two or more variables are perfectly correlated which results is a singular covariance matrix. This can occur in the data where some variables are computed from some measured variables. The T^2 Statistic is based on the assumption that the covariance matrix is non-singular and can be inverted.

Collinearities in the covariance and the correlation matrix can occur because of

1. Sampling restrictions
2. Theoretical relationships existing in the process
3. Outliers in the data.

By examining the eigenvalues and eigenvectors of the sample covariance matrix, collinearity can be detected. The correlation matrix is usually used for this purpose.

Highly correlated variables are analysed using Principal Component Analysis (PCA). This method is discussed in detail in the next chapter but for the purposes of explaining the detection of collinearity, it can be mentioned that PCA can be used

1. to detect a singular covariance matrix, and
2. to determine variables that are highly correlated

As described by Mason and Young (2002), one of the recommendations for detecting a near singular matrix is given by calculating the condition indices. The formula for calculating these indices is,

$$= \sqrt{\frac{\max \mathit{eigenvalue}}{\mathit{every\ other\ eigenvalue}}}$$

$$= \sqrt{\frac{\lambda_1}{\lambda_i}}$$

where i = 2, ..., n, and n is the total number of eigenvalues.

Eigenvalues are the characteristic roots of a matrix and eigenvectors are the characteristic vectors of the corresponding eigenvalue.

If any of the indices are greater than 30, this indicates that severe collinearity is present in the data. In the case where collinearity exists, the use of the T^2 statistic is not recommended.

There are a number of solutions that can be implemented,

1. One of the variables involved in the collinearity can be removed,
2. The covariance matrix can be reconstructed by excluding the eigenvectors corresponding to the near-zero eigenvalues.

The latter method is used in PCA by reducing the number of Principal Components and will be discussed in the next chapter.

2.7 Conclusion

All techniques discussed in this chapter can be applied for monitoring and statistical process control in continuous processes with multivariate data. Which equation to apply, in order to calculate the T^2 Statistic and UCL, will depend on the type of data (i.e. individuals or subgrouped) and whether the mean and variance are known. This applies for both Phase I and Phase II of the process.

The steps for implementation are summarised as follows,

1. Phase I screens the data for outliers,
    a. The T^2 values for each observation are determined based on their respective distribution.
    b. The UCL is calculated, from which each of the T^2 values are compared against.
    c. Any out of control points with assignable cause are removed.
    d. Step 1 is repeated until there are no outliers with assignable cause.
2. Phase II monitors new observations,
    a. The new T^2 values are calculated using the mean and covariance from the baseline data determined from Phase I.
    b. The UCL is also calculated using the number of observations from Phase I.
    c. The new T^2 values are assessed against the UCL for out of control signals.

2.8 References

Bersimis, S., Psarakis, S., Panaretos, J. (2007), "Multivariate Statistical Process Control Charts: An Overview", *Quality and Reliability Engineering International*, Vol.23, pp. 517-543.

Chenouri, S., Steiner, S.H., Variyath, A.M. (2009), "A Multivariate Robust Control Chart for Individual Observations". *Journal of Quality Technology*, Vol. 41, No. 3, pp. 259-271.

CPACT (Centre for Process Analytics and Control Technologies), School of Chemical Engineering and Advanced Materials, Newcastle University, UK. "Multivariate Data Analysis & Statistical Process Control", 3-day Continuing Professional Development Course, 21-23 October 2008.

Fuchs, C., and Kenett, R. (1998), *Multivariate Quality Control*, Marcel Dekker, Inc.

Hotelling, H. (1931), "The Generalization of Student's Ratio", reprinted in *Multivariate Statistical Methods: Among Group Covariation*, Dowden Hutchinson & Ross Inc.

Hotelling, H. (1947), "Multivariate Quality Control Illustrated by the Air Testing of Sample Bombsights", *Techniques of Statistical Analysis,* edited by Eisenhart, C., Hastay, M.W., and Wallis, W.A., McGraw-Hill, New York, pp. 111-184.

Lowry, C.A., and Montgomery, D.C. (1995), "A Review of Multivariate Control Charts" *IIE Transactions*, Vol. 27, No. 6, pp. 800-810.

Mason, R.L., and Young, J.C. (1999), "Improving the Sensitivity of the T^2 Statistic in Multivariate Process Control". *Journal of Quality Technology*, Vol. 31, No. 2, pp. 155-165.

Mason, R.L., and Young, J.C. (2002), Multivariate Statistical Process Control with Industrial Applications, ASA-SIAM.

Mason, R.L., Chou, Y.M., Sullivan, J.H., Stoumbos, Z.G., Young, J.C. (2003), "Systematic Patterns in T^2 Control Charts", *Journal of Quality Technology*, Vol. 35, No.1. pp. 47-58.

Montgomery, D.C. (2005), *Introduction to Statistical Quality Control*, 5th Edition, John Wiley & Sons, pp.54-62.

Montgomery, D.C., and Mastrangelo, C.M. (1991), "Some Statistical Process Control Methods for Autocorrelated Data (with Discussion)". *Journal of Quality Technology*, Vol. 23, pp. 179-204.

Ryan, T.P. (1989), Statistical Methods for Quality Improvement, John Wiley, New York.

Shewhart, W.A. (1931), *Economic Control of Quality of Manufactured Product*, Van Nostrand, New York, NY.

Sullivan, J.H., and Woodall, W.H. (1996), "A Comparison of Multivariate Control Charts for Individual Observations". *Journal of Quality Technology*, Vol. 28, No.4, pp. 398-408.

Tracy, N.D., Young, J.C., Mason, R.L. (1992), "Multivariate Control Charts for Individual Observations". *Journal of Quality Technology*, Vol. 24, No. 2, pp. 88-95.

Western Electric Co. Inc. (1956), *Statistical Quality Control Handbook*, Delmar Printing Co., Charlotte, NC.

Woodall, W.H. (2000), "Controversies and Contradictions in Statistical Process Control" *Journal of Quality Technology,* Vol. 32, No. 4, pp. 341-350.

Yang, K., and Trewn, J. (2004), *Multivariate Statistical Methods in Quality Management*, McGraw-Hill, New York.

CHAPTER THREE

PRINCIPAL COMPONENT ANALYSIS

3.1 Introduction

Principal Component Analysis (PCA) has been used for many decades for various applications. It is one of the oldest multivariate techniques. It was first described by Pearson (1901) and then Hotelling (1933) described its specific computational technique. As seen with Multivariate analysis, there are many variables to consider and it is not always possible to exclude variables. It is desirable to include as many variables as possible and try not to omit any relevant variables. It is not practical to analyse all variables as the level of correlation between variables is likely to be large.

Jackson (2003) described Principal Component Analysis as a data analysis technique, which describes the multivariate structure of the data. The most common use for Principal Component Analysis is data reduction. This simplifies the data by reducing the dimensionality in the data and so separates the signal from the noise. It does this by extracting a small number of factors that can be used to summarise the data with minimal loss of information about the original variables. This can help during investigation in finding the root cause of a signal. It tries to find a few independent linear combinations of the original variables that will account for most of the variability in the data. Generally two or three Principal Components will account for 80-90% of the data. Once the components are extracted, separate analysis can be performed to help interpret what this means for the process.

Dillon and Goldstein (1984) defined PCA as:

"Principal components analysis transforms the original set of variables into a smaller set of linear combinations that account for most of the variation in the original set. The purpose of PCA is to determine factors (i.e., principal components) in order to explain as much of the total variation in the data as possible."

Fuchs and Kenett (1998) illustrated how Principal Component Analysis plays an important part in Quality Control, it has certain important features.

1. The new variables are uncorrelated,

2. A few of the principal components will determine where most of the variability is captured, so these new variables need only be used for controlling the process.

Principal Component Analysis has many objectives including summarising data, classifying variables, detecting outliers and a warning of faults occurring in the process and traceability of the fault.

If variables are correlated, PCA takes p variables, $x_{1,}x_{2,} \ldots, x_p$, and transforms them into p uncorrelated variables, $t_{1,}t_2, \ldots, t_p$. These new transformed variables are called Principal Components.

3.2 Principal Component Analysis

Principal Component Analysis has three main steps:

1. Calculate the correlation matrix. This is done in order to find groups of variables that are correlated to one another. It can also help to determine which variable to eliminate if one variable is correlated to any of the other variables.

2. Calculate the Principal Components.

3. Calculate the transformed data set to find multivariate relationships in the data.

Principal Components are linear transformations of the original variables and can be used to approximate the original data matrix X.

Let X be an nxp matrix where n corresponds to the number of observations and p corresponds to the number of process variables.

$$X = t_i p_i' + \cdots + t_p p_p' = \sum_{i=1}^{p} t_i p_i{}' \qquad (3.1)$$

where t_i is the ith Principal Component of X also known as the score vector and

p_i is the ith loading vector.

Fuchs and Kenett (1998) pointed out that the loading vector is also known as the eigenvector of covariance matrix, X^TX. The corresponding eigenvalues (λ_i) to the eigenvectors (p_i) are the coefficients of the original variables and show the variance of each principal component (t_i).

The PCA of X is equivalent to the eigenvector analysis of the covariance matrix of X, $X'X$. If the eigenvalues are arranged in order of the largest variation in X to the smallest variation in X, shown as follows,

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p,$$

then their corresponding eigenvectors, $p_1, p_2, \cdots, p_p$ are the loading vectors of X.

The first Principal Component of X is a linear combination of the original variables with the greatest amount of variation, illustrated below.

$$t_1 = p_{11}x_1 + p_{21}x_2 + \ldots + p_{p1}x_p$$

and it has the greatest variance.

The second principal component of X is the linear combination

$$t_2 = p_{12}x_1 + p_{22}x_2 + \ldots + p_{p2}x_p$$

and it has the next largest amount of variation. This Principal Component has a condition whereby it is not correlated (orthogonal) to the first principal component.

Each Principal Component after this is not correlated with any of the other principal components that have been previously defined.

Principal Component Analysis can be performed on the covariance matrix or the correlation matrix. Which to use depends on the nature of the data and what its application will be. As discussed by Yang and Trewn (2004), in industry, when the data is a measurement and it describes a dimensional characteristic i.e. length or distance, it is not recommended to standardise the original data because the original scale is required in order to quantify the actual geometry of an object. When data is normalised, each variable has the same unit and so has the same physical meaning. It would not be appropriate to try and identify components in a "normalised" shape that account for most of the variation in a real shape. For this situation, PCA must be done on the covariance matrix as this deals with the original variables.

MacGregor and Kourti (1995) pointed out that the covariance matrix is not known in practice and so is estimated. Yang and Trewn (2004) also described that if the original variables are in different units or if they have different numerical magnitudes like units of inches and psi (pressure measurement), then the PCA will be more influenced by the larger measurement and so the original variables must be standardised (scaled) and the correlation matrix should be used to perform PCA.

3.3 Standardised Principal Components

If PCA is to be performed on the correlation matrix, as described in the case above, the Principal Components must be standardised.

This first involves standardising the vector $X = (X_1, X_{2,\ldots,} X_p)$, to find the principal component scores.

The next step is to standardise the PC scores, this is done by dividing each PC by their standard deviations.

3.3.1 Principal Component Scores

$X = (X_1, X_{2,\ldots,} X_p)$, is standardised by subtracting the mean from the observation to get the deviations of the variables from their target. This results in standardised X's. In standardising the vector $X = (X_1, X_{2,\ldots,} X_p)$, the resulting matrix Z is written as,

$$Z_i = \frac{X_i - \bar{X}_i}{\sqrt{s_{ii}}}$$

for $i = 1 \ldots p$,

$\bar{X}_i$ is the sample mean for X_i and

s_{ii} is the sample variance for X_i.

Example 3.1

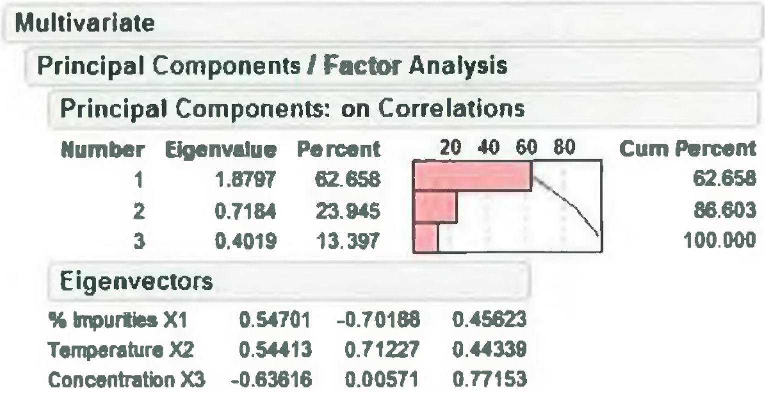
Continuing with the Tracy *et al.* (1992) data used in the previous chapter where the initial dataset had 14 observations. In the Phase I process for T^2 control charting, sample 1 was identified as an out of control point and was excluded.

The next step is to run a principal component analysis on the remaining 13 samples (Table 3.1) using the correlation matrix.

Sample No.	% Impurities X1	Temperature X2	Concentration X3
1	16.9	83.77	43.44
2	17.38	84.46	42.74
3	16.9	86.27	43.6
4	16.92	85.23	43.18
5	16.71	83.81	43.72
6	17.07	86.08	43.33
7	16.93	85.85	43.41
8	16.71	85.73	43.28
9	16.88	86.27	42.59
10	16.73	83.46	44
11	17.07	85.81	42.78
12	17.6	85.92	43.11
13	16.9	84.23	43.48

Table 3.1 – Phase I Historical Data

The Principal Components Analysis platform in JMP, generates the output in Figure 3.1.



Multivariate

Principal Components / Factor Analysis

Principal Components: on Correlations

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	1.8797	62.658		62.658
2	0.7184	23.945		86.603
3	0.4019	13.397		100.000

Eigenvectors

% Impurities X1	0.54701	-0.70188	0.45623
Temperature X2	0.54413	0.71227	0.44339
Concentration X3	-0.63616	0.00571	0.77153

Figure 3.1 – JMP Output for Principal Components Analysis

The cumulative percent of the first two components accounts for 86.6% of the total variation as shown in Figure 3.1, therefore the principal components charts will be generated from two PC's.

The principal component equations are, from Equation 3.1,

$$X = t_i p_i^{'} + \cdots + t_p p_p^{'} = \sum\nolimits_{i=1}^{p} t_i p_i{}'$$

$$t_i = \sum\nolimits_{i=1}^{p} p_i^{'} x_i$$

where t_i is the principal component and p_i are the eigenvectors.

So, the first principal component is,

$$t_1 = p'_1 x_1 + p'_2 x_2 - p'_3 x_3$$

$$t_1 = 0.547x_1 + 0.544x_2 - 0.636x_3$$

which can also be expressed as,

$$t_1 = 0.547\ \%impurities + 0.544\ temperature - 0.636\ concentration$$

The second principal component is,

$$t_2 = -0.702x_1 + 0.712x_2 - 0.006x_3$$

which can also be expressed as,

$$t_2 = -0.702\ \%impurities + \ 0.712\ temperature - 0.006\ concentration$$

After the Principal Components have been identified, the new principal component scores are obtained by substituting the variable values of the original objects into the principal component equations, which are defined by the PC eigenvectors.

$s_{ii} = s_{11}$ = variance of variable x_1. The standard deviations of the variables are, $\sqrt{s_{11}} = 0.2589$, $\sqrt{s_{22}} = 1.0454$ and $\sqrt{s_{33}} = 0.4037$, which are %impurities, temperature and concentration respectively.

The mean vector $\bar{X} = (16.9769, 85.1484, 43.2815)'$.

The principal component scores, T_i, are calculated by,

$$T_i = p_i Z_i = \ p_{1i} Z_1 + p_{2i}\, Z_2 + p_{3i} Z_3$$

for $i \ = \ 1, .. p$

where p_i is the eigenvector and

$Z_i = \frac{x_i - \bar{x}_i}{\sqrt{s_{ii}}}$ is the standardised variable.

The PC1 score for Sample No. 1 is calculated by,

$$T_1 \ = \ 0.547 \frac{x_1 - \bar{x}_1}{\sqrt{s_{11}}} + 0.544\ \frac{x_2 - \bar{x}_2}{\sqrt{s_{22}}} - 0.636 \frac{x_3 - \bar{x}_3}{\sqrt{s_{33}}}$$

$$= \ 0.547 \frac{(16.90 - 16.9769)}{0.2589} + \ 0.544 \frac{(83.77 - 85.1454)}{1.0454} - 0.636 \frac{(43.44 - 43.2815)}{0.4037}$$

$$= 0.547(-0.2970) + 0.544(-1.3156) - 0.636(0.3938)$$

$$= \ -0.1605 - 0.7154 - 0.2504 = -1.13$$

The PC2 score for Sample No. 1 is calculated by,

$$T_2 = -0.7018 \frac{x_1 - \bar{x}_1}{\sqrt{s_{11}}} + \ 0.7122 \frac{x_2 - \bar{x}_2}{\sqrt{s_{22}}} - 0.0057 \frac{x_3 - \bar{x}_3}{\sqrt{s_{33}}}$$

$$= -0.7018\frac{(16.90-16.9769)}{0.2589} + 0.7122\frac{(83.77-85.1454)}{1.0454} - 0.0057\frac{(43.44-43.2815)}{0.4037}$$

$$= -0.7018(-0.2970) + 0.7122(-1.3156) - 0.0057(0.3938)$$

$$= 0.2084 - 0.9369 - 0.0022 = -0.73$$

The rounded results calculated in JMP are shown in Table 3.2.

Sample No.	PC1 Score	PC2 Score
1	-1.13	-0.73
2	1.348	-1.57
3	-0.08	0.979
4	0.084	0.211
5	-1.95	-0.18
6	0.607	0.385
7	0.065	0.609
8	-0.26	1.122
9	1.47	1.019
10	-2.53	-0.47
11	1.333	0.193
12	1.99	-1.16
13	-0.95	-0.41

Table 3.2 – PC Scores for Historical Data

3.3.2 Standardised PC Scores

Recall that PCs are linear combinations of the deviations of the variables from their targets. All PC scores are divided by their standard deviations resulting in standardised PC scores.

Yang and Trewn (2004) describe the standard deviations of the PC scores are the square root of the eigenvalues ($\sqrt{\lambda_i}$), since the eigenvalues are the variance of the PCs.

$$s_i = \sqrt{\lambda_i}$$

$$s_i^2 = \lambda_i = \mathrm{Var}(X_i)$$

The standardised PC scores, also called Latent variables, follow a multivariate normal distribution, so mean = 0 and standard deviation = 1, and so three-sigma limits for UCL and LCL can be applied. UCL = 3, Mean = 0, and LCL = -3.

Each of the Principal Components are divided by their variances, shown as,

$$\frac{T_i}{\sqrt{\lambda_i}} = \frac{p_i z_i}{\sqrt{\lambda_i}} = \frac{p_{1i} z_i + p_{2i} z_i + p_{3i} z_i}{\sqrt{\lambda_i}}$$

for $i = 1, \ldots, p$

Example 3.2

Using the data from Example 3.1, the principal components charts can then be generated from the standardised scores. The eigenvalues for PC1 and PC2 from Table 3.2 in Example 3.1 are, $\lambda_1 = 1.8797$ and $\lambda_2 = 0.7184$.

T_1 = -1.13 and T_2 = -0.73, as calculated in Example 3.1.

The standardised PC1 score for sample 1 is,

$$\frac{T_1}{\sqrt{\lambda_1}} = \frac{-1.13}{\sqrt{1.8797}}$$

$$= \frac{-1.13}{1.37} = -0.82$$

The standardised PC2 score for sample 1 is,

$$\frac{T_2}{\sqrt{\lambda_2}} = \frac{-0.73}{\sqrt{0.7184}}$$

$$= \frac{-0.73}{0.84} = -0.86$$

The standardised PC scores (rounded) for each of the samples, calculated in JMP, are shown in Table 3.3.

As the PC scores are standardised i.e. scaled to unit variance, it is assumed that they follow a standardised normal distribution, N(0,1), so mean = 0 and standard deviation = 1.

Sample No.	Standardised PC1 score	Standardised PC2 score
1	-0.82	-0.86
2	0.983	-1.85
3	-0.06	1.155
4	0.061	0.248
5	-1.42	-0.21
6	0.443	0.454
7	0.048	0.719
8	-0.19	1.324
9	1.072	1.203
10	-1.85	-0.55
11	0.972	0.228
12	1.452	-1.37
13	-0.69	-0.49

Table 3.3 – Standardised PC Scores for Historical Data

There can be as many Principal Components as there are variables. If PCA is performed on the correlation matrix then the number of Principal Components will be equal to the number of variables. How to decide on the number of Principal Components to retain will be discussed in the next section.

3.3.3 Retaining Principal Components

Given a set of n observations on p variables, PCA is used to determine k new variables. A number k, can be chosen which is small relative to p, without the loss of information. These k new variables are called Principal Components (PC's) and will account for most of the variation in the p variables. The set of Principal Components will have the same variation and structure as the original variables.

Jackson (2003) describes various methods for deciding when to optimally determine a number, k.

CPACT (2008) in Newcastle University, UK, discuss a few ways for selecting the number of Principal Components, k, to retain, listed below,

1. Include enough of the components to explain 80-90% of the total variability in the data, as shown in the JMP output in Figure 3.2, by calculating,

$$0.8 < \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} < 0.9$$

2. Exclude the Principal Components whose eigenvalues are less than 1.

3. Cross validate, this means taking approximately three-quarters of the data, calculating the PCs and creating a model. The model is then validated using the remaining one-quarter of the data.

Nomikos and MacGregor (1995) describes that cross validation involves excluding some data from the dataset, the PCA model is then created with the remaining batches. This is done a number of times excluding different sets of data each time. This shows how the PCA models' predictive ability increases by adding more principal components.

3.4 Principal Component Control Charts

Figure 3.2 outlines the method for setting up and using a Principal Component Control Chart, which was described by Yang and Trewn (2003).

The scores from Table 3.3 in Example 3.1 can be plotted on a control chart with UCL= 3 and LCL= -3 as shown in Figure 3.3 and Figure 3.4.

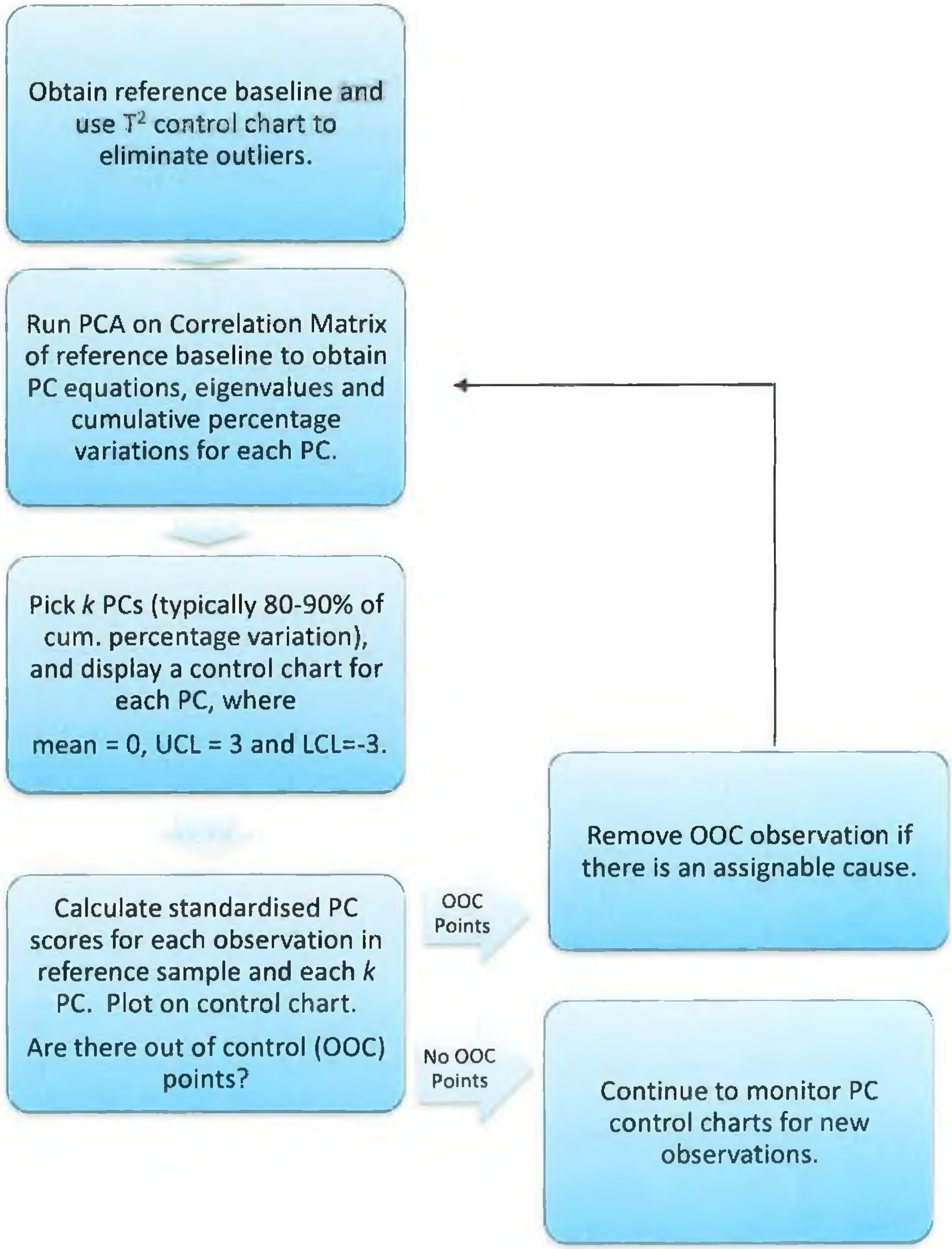

Obtain reference baseline and use T² control chart to eliminate outliers.
Run PCA on Correlation Matrix of reference baseline to obtain PC equations, eigenvalues and cumulative percentage variations for each PC.
Pick *k* PCs (typically 80-90% of cum. percentage variation), and display a control chart for each PC, where
mean = 0, UCL = 3 and LCL=-3.
Calculate standardised PC scores for each observation in reference sample and each *k* PC. Plot on control chart.
Are there out of control (OOC) points?
OOC Points
Remove OOC observation if there is an assignable cause.
No OOC Points
Continue to monitor PC control charts for new observations.

Figure 3.2 Setting up PC Control Charts

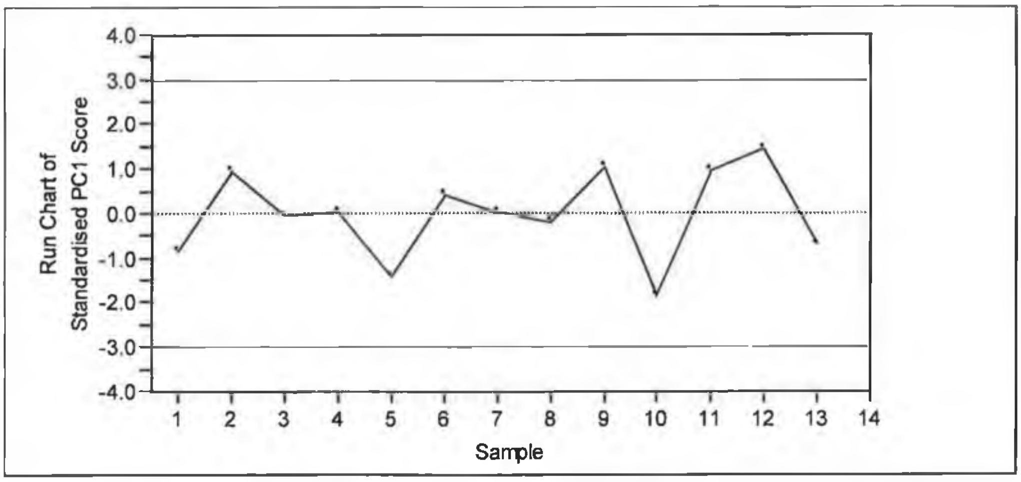


Figure 3.3 – PC Chart for PC1 Score

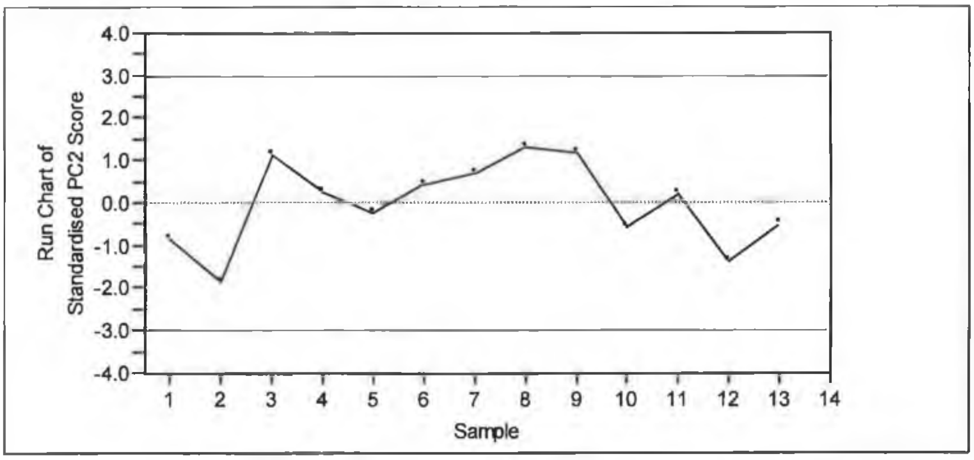


Figure 3.4 – PC Chart for PC2 Score

Run Rules similar to univariate SPC can be used to identify an out of control situation

When plotting PC scores on individual control charts, it is also useful to plot pairs on a scatterplot of two PCs, usually PC1 and PC2 since these are the largest PC's.

Figure 3.5 shows a scatterplot (biplot) of the first two principal components scores.

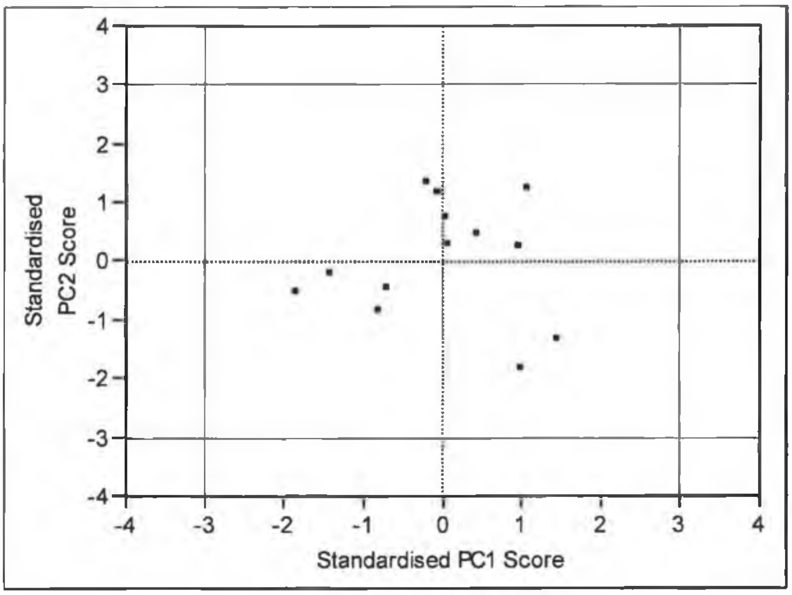


Figure 3.5 – Scatterplot of PC1 and PC2

Transforming the original variables into Principal Components and then plotting the new uncorrelated variables on control charts was proposed by Jackson (1985). In order to do this effectively, the Principal components must have a meaning. It must not be difficult to interpret what this means in terms of the original variables.

As Principal Components represent a special type of correlated variation in the observations, a high PC score indicates an extreme case for that type of variation.

3.4.1 PC Scores for a New Observation

This is also known as Phase II, as new observations are compared against the PCA model developed from Phase I. The standardised scores are obtained for the new observation vector and assessed against the ±3 standard deviations control limits.

Example 3.3

The 13 samples used to develop the model has, $\bar{X}$ and S,

$$\bar{X} = (16.98, 85.14, 43.28)', \quad S = \begin{bmatrix} 0.068 & 0.076 & -0.055 \\ 0.076 & 1.092 & -0.216 \\ -0.055 & -0.216 & 0.163 \end{bmatrix}.$$

The covariance matrix is used in Phase II. This is because $\sum$ is usually unknown, therefore it has to be estimated from the sample. In this case, the sample estimate from Phase I will be used, as it is representative of normal operation.

Using the new observation data in Example 2.2, $X_{\text{new}} = (17.08, 84.08, 43.81)'$,

The standardised values are,

$$Z_1 = \frac{17.08 - 16.98}{\sqrt{0.068}} = 0.38$$

$$Z_2 = \frac{84.08 - 85.14}{\sqrt{1.092}} = -1.01$$

$$Z_3 = \frac{43.81 - 43.28}{\sqrt{0.163}} = 1.31$$

The standardised score for PC1 is,

$$\frac{T_1}{\sqrt{1.8797}} = \frac{-0.547(0.38) - 0.544(-1.01) + 0.636(1.31)}{\sqrt{1.8797}}$$

$$= \frac{-0.20 + 0.54 + 0.83}{1.37} = 0.85$$

The standardised score for PC2 is,

$$\frac{T_2}{\sqrt{0.7184}} = \frac{0.702(0.38) - 0.712(-1.01) - 0.006(1.31)}{\sqrt{0.7184}}$$

$$= \frac{0.26 + 0.71 - 0.007}{0.84} = 1.15$$

Both of these scores are within the control limits, UCL= 3 and LCL= -3, so this observation is in control.

3.5 T^2 Charts and Squared Prediction Error (SPE) in PC Space

3.5.1 T^2 Charts in Principal Component Space

Kourti and MacGregor (1995) described how Hotellings T^2 statistic can also be plotted in the Principal Component space.

Mason and Young (2002) and Fuchs and Kenett (1998) indicate PC Control Charts use the Principal Component on the correlation and so better detection for out of control points is obtained if they are used alongside the T^2 chart. In addition, PC control charts complement T^2 charts since PCA does not measure the deviation of the multivariate data from the norms.

These Hotellings T^2 charts are generated based on the first *k* PCs of the correlation matrix,

$$T_i^{\,2} = \sum_{i=1}^{k} \frac{t_i^{\,2}}{s_{t\;i}^2}$$

where $s_{t\;i}^2 = \lambda_i$, is the estimated variance of t_i, the principal component score.

These are sometimes referred to as the Principal Factors.

$$T^2 = \frac{t_1^{\,2}}{\lambda_1} + \frac{t_2^{\,2}}{\lambda_2}$$

Example 3.4

Using the results from Example 3.3, i.e. the standardised scores, $t_1 = 0.85$ and $t_2 = 1.15$ and the eigenvalues of the correlation matrix, $\lambda_1 = 1.8797$ and $\lambda_2 = 0.718$, the T^2 statistic for the new observation is,

$$T^2 = \frac{0.85^2}{1.8797} + \frac{1.15^2}{0.7184} = 0.384 + 1.840 = 2.22.$$

Ferrer (2007) also showed this derivation of the T^2 statistic as calculated in Example 3.4. He also details that the Phase I and Phase II control limits for the T^2 statistic are similar to the control limits used for individual observations, where μ and ∑

are unknown. This is based on the assumption that the scores follow an MVN distribution.

Phase I is based on the Beta distribution, shown in Equation 1.16 and Phase II UCL is based on an F distribution, shown in Equation 1.17. When the T^2 statistic is being plotted in the score space, the original number of variables, p, changes to the number of retained components in the score space, K, and N is the number of observations. Uppercase is used to denote calculations in the score space. The Phase I UCL, similar to Equation 1.16, is calculated as,

$$UCL(T_K^2) = \frac{(N-1)^2}{N}\beta_{(\alpha,K/2,(N-K-1)/2)}$$

and the Phase 2 UCL is calculated as,

$$UCL(T_K^2) = \frac{K(N^2-1)}{N(N-K)}F_{(\alpha,K,N-K)}$$

which is the same as Equation 1.17, $T^2{}_{UCL} = \frac{(n+1)(n-1)p}{n(n-p)}\mathrm{F}_{(\alpha,p,n-p),}$ as the observations are not independent of the PCA parameters.

Phase II control limits where the population parameters and μ and $\sum$ are unknown, are calculated using Equation 1.17, to determine if this new T^2 value is in control.

$N = n = 13$ and $K = p = 2$ as there are two principal components retained in the model.

$$T^2{}_{UCL} = \frac{(N+1)(N-1)K}{N(N-K)}\mathrm{F}_{(\alpha,K,N-K),}$$

$$T^2{}_{UCL} = \frac{(13+1)(13-1)2}{13(13-2)}\mathrm{F}_{(0.05,2,13-2),}$$

$$= 2.35\ \mathrm{F}_{(0.05,2,11),}$$

$$= 2.35(3.98) = 9.35$$

The $T^2 = 2.22 < T^2{}_{UCL} = 9.35$, therefore the new observation is in control.

Kourti and MacGregor (1995) outline that a high T^2 value suggests that the observation has an abnormal extreme value in one or more of the original variables. This is caused by the observation being a large distance from the origin in the score space, given that all PC scores are in control. This indicates that the cause of the variation cannot be explained by a known Principal Component and it needs to be investigated further.

Similar to T^2 charts, PCA does not identify the variables that are responsible for an out of control situation. They are not original variables and so operators can find them difficult to interpret.

3.5.2 Residuals Charts

Montgomery (2005) describes that for autocorrelated data in a univariate time series model, a useful control chart is one based on the residuals. The model, which describes the correlation structure of the data is used to remove autocorrelation from the data. A control chart based on the residuals can be plotted. Suppose that $\hat{x}_t$ is the fitted value of x_t, the residuals are calculated by,

$$e_t = x_t - \hat{x}_t$$

A residual is effectively the difference between an observed value and a predicted value created by a model.

Traditional control charts can be applied to the residuals. Any abnormal patterns or out of control point would suggest that the parameters used to create the model have changed. This would indicate that the original variable, x_t is out of control.

In Multivariate data, when a model is developed using a set of k Principal Components, based on historical data, the matrix X is defined by,

$$X = TP'$$

For each new observation, the fitted values $\hat{x}_{new}$, can be calculated.

These values are then used to evaluate the SPE (Q Statistic), i.e. the squared distance between the observed and the predicted values from the nominal.

3.5.3 Squared Prediction Error

Jackson (2003) describes SPE (Q Statistic) as a measure of how close the observation is to the k-dimensional space defined by the model. If the new observation fits the model, then the SPE will be small as X and $\hat{X}$ will be similar. If they show differences, a high SPE will emerge, which indicates that the predicted and observed values are not similar, and that something has changed in the process.

$$SPE_{new} = \sum_{i=1}^{k} (x_{new} - \hat{x}_{new})^2$$

The SPE scores are also plotted on a control chart. Jackson and Mudholkar (1979) recommended approximate control limits for a given level of significance, α, for the quadratic residuals as,

$$Q_\alpha = \theta_1 \left[\frac{c_\alpha \sqrt{2\theta_2 h_0^{\,2}}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^{\,2}} \right]^{\frac{1}{h_0}}$$

where c_α is the normal variate with the same sign as h_0.

The remaining quantities are as follows,

$$\theta_1 = \sum_{i=k+1}^{p} \lambda_i, \quad \theta_2 = \sum_{i=k+1}^{p} \lambda_i^{\,2}, \quad \theta_3 = \sum_{i=k+1}^{p} \lambda_i^{\,3}$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$$

where k = retained principal components and p = total number of variables.

Kourti and MacGregor (1995) explain that when a process is in control, the SPE will be small. When the PCA model has been generated from an in control process, the SPE accounts for variation that is not accounted for by the model, as it is based on the residuals. So, when a new type of special event happens and this event was not accounted for when developing the in-control PCA model, new PCs will appear indicating that the structure of the correlation has changed. These new events will be detected by the SPE values. They will be large, which indicates that the model does not fit for that observation.

3.5.4 DModX Charts

Eriksson *et al.* (2001) proposed an alternative method based on the equality of variances from a normal distribution. Their method is also based on the SPE. SIMCA P+, developed by umetrics, calculates the DModX.

Eriksson *et al.* (2001) define two methods for determining DModX,

1. Absolute DModX

2. Normalised DModX, denoted as $\mathrm{DModX_{norm}}$.

The absolute DModX calculates the absolute distance of an observation to the model. It uses a correction factor, c, a function of the number of variables, P, and the number of retained components,K, for use in Phase I.

$$\mathrm{DModX} = \mathrm{s_i} = \mathrm{c}\sqrt{\frac{\sum \mathrm{e_{ip}^2}}{\mathrm{P-K}}}$$

where $\sum \mathrm{e_{ip}^2}$ is the SPE.

The correction factor, c, accounts for the degrees of freedom based on the fact, that the distance to the model (DModX) is likely to be somewhat smaller for an observation in the reference dataset, as it has influenced the model. In Phase II, $c = 1$.

Eriksson *et al.* (2001) show that the normalised DModX, ($\mathrm{DModX_{norm}}$) calculates the normalised distance of an observation to the model. This uses the Absolute DModX and divides it by the pooled residual standard deviation, S_o. This is shown as,

$$DModX_{norm} = \frac{DModX}{s_o} = \frac{s_i}{s_o}$$

where

$$s_o = \frac{\sum_{i=1}^{N}\sum_{p=1}^{P} e_{ip}^2}{(N-K-K_0)(P-K)}$$

and $K_0 = 1$ if the model is centred, so the divisor will be $(N-K-1)(P-K)$ otherwise $K_0 = 0$, where the divisor will be $(N-K)(P-K)$.

The $\left(\frac{s_i}{s_o}\right)^2$ statistic has an approximate F distribution with

$P-K$ and $(N-K-1)(P-K)$ degrees of freedom.

The UCL is calculated as,

$$UCL(SPE) = \frac{P-K}{c^2} s_0^2 F_{\alpha,P-K,(N-K-1)(P-K)}$$

for specified α, where $F_{P-K,(N-K-1)(P-K)}$ has an F distribution with

$P-K$ and $(N-K-1)(P-K)$ degrees of freedom.

Figure 3.6 shows the normalised DModX chart for the 13 samples from Example 3.1.

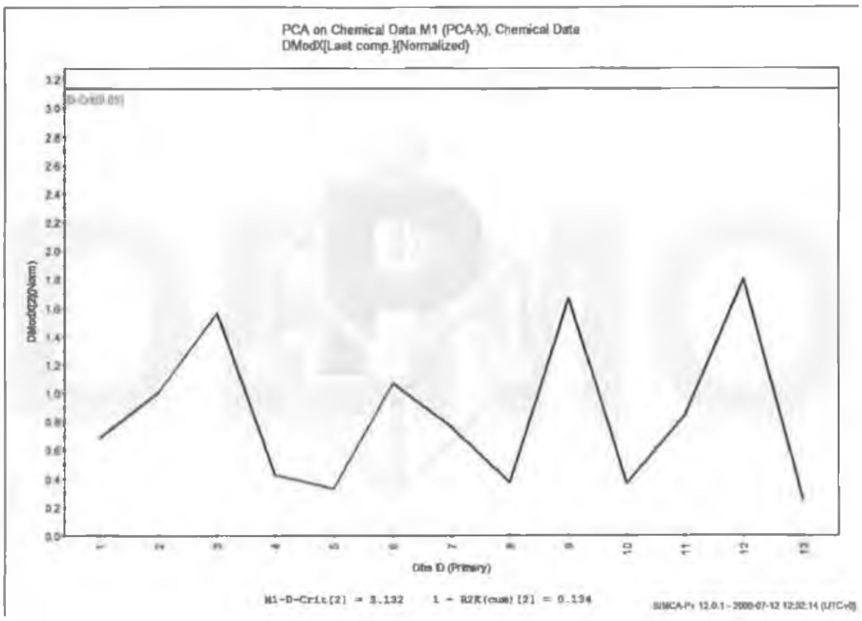


Figure 3.6 - DModX plot in SIMCA P+

The correlation has a large effect on the distribution of a Multivariate statistic. The SPE monitors the correlation structure and the T^2 monitors the magnitude (direction) given that the correlation is ok. Changes in these statistics are indicated on their respective control charts. It is because of this that both charts must be monitored in order to investigate the cause of an out of control signal. If either of these statistics moves outside its control limit then something has changed.

The T^2 control chart is a general multivariate control chart and can be used with or without using Principal Components. The SPE (Q-Statistic) and DModX, however, specifically deal with the residuals from PCA.

Chiang and Colegrove (2007) explain how a PCA model can be used to develop T^2 and Q (SPE) charts in order to detect changes for all variables at the same time.

Wikstrom *et al* (1998) proposed a scores monitoring and residual tracking, (SMART) chart which measures the process over time which consists of a Shewhart type Hotelling T^2 or scores control chart and a DModX chart displayed horizontally. These SMART charts can simultaneously show the systematic variation in the data from the scores and T^2 charts and the unsystematic variation from the DModX chart.

To clearly see which variables contributed to an out of control signal, one must revert back to the original variables. Firstly, their contribution to the calculated scores and the SPE must be assessed. Remember that if PCA is performed on the correlation matrix, the variables are standardised and so interpretations on the variables can be difficult to see immediately.

This contribution can be displayed using a contribution plot. The major advantage to using this type of plot is that it allows the interpretation of information in terms of the original process variables.

3.6 Contribution Plot

A Contribution Plot is very effective in identifying the set of original variables whose contribution has changed from those predicted from the PCA model. MacGregor and Kourti (1995) described it as one of the most common approaches for identifying the original variables responsible for an out-of-control signal.

It will show how each variable contributes to the calculation of the PC score. It shows the change in the new observations relative to their average values calculated from the PCA model.

The PC scores are written as a weighted sum of the data. The loadings are the weights.

As the PC chart, T^2 and SPE charts cannot be interpreted in terms of the original variables, the contribution plot can display the contributions that each of the original variables has on the calculation for the particular statistic. The scores chart and T^2 chart show the variation that is explained by the model. The contribution plot of DModX shows unexplained variation.

Figure 3.7 shows the contributions that sample 1, in Example 3.1, had on the T^2 statistic from each of the variables, %Impurities, Temperature and Concentration.

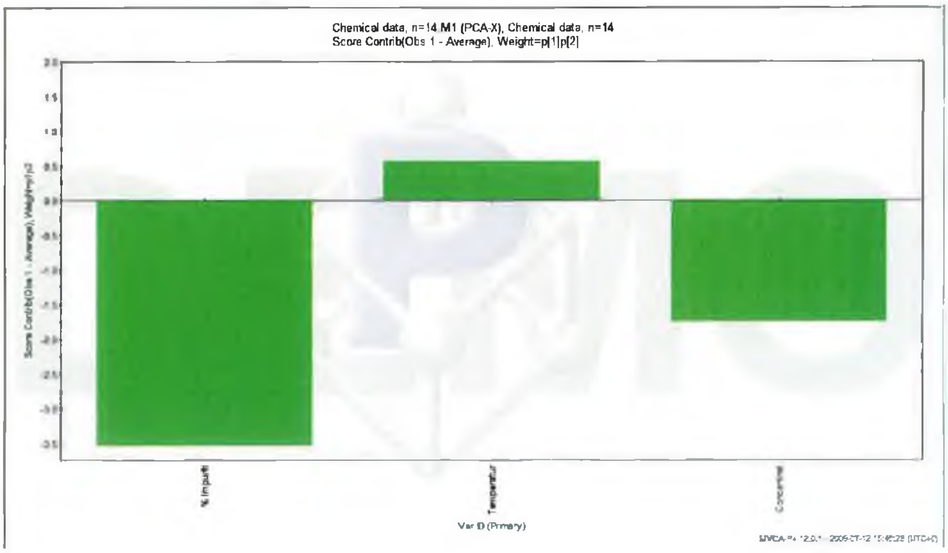


Figure 3.7 - Contribution Plot for sample 1 OOC point.

In fault detection, the contribution chart shows the contribution that each variable made to the abnormal condition. This information can help in finding the particular variables that are the root cause of the problem.

3.7 Partial Least Squares

Suppose that a process measures a large number of input process variables, X, that are highly correlated and the process also measures the output quality variables, Y. It is possible to predict the Y quality variables using the X process variables measurements by developing a predictive model. As with all predictive models, it must be calculated from a stable reference dataset of optimal conditions. This type of model can be generated through Partial Least Squares (PLS). PLS can also be known as Projection to Latent Structures.

AlGhazzawi and Lennox (2009) explain that PLS is similar to Principal Component Regression as it explains the variation in the process data, which can be known as cause data, by reducing it to a set of factors that will maximally explain the variation of the data in X. In addition to this, PLS also explains the variation in this input (cause) data that is most predictive of the quality (effect).

This method has the ability to predict if the process is potentially going to have a negative effect on the output quality. It does so by identifying an unusual event at a certain stage in the process. The earlier this is detected, the better for the process.

In manufacturing, the benefits of having an accurate predictive model of the process have high cost savings and also reduce the potential for distressed inventory of finished product.

The score vectors that are derived from the PLS analysis are different to those that are obtained from PCA.

AlGhazzawi and Lennox (2009) discussed the application of PCA and PLS in a multivariate process control system. They described the breakdown of the PLS algorithm into cause and effect matrices.

The purpose of modelling is to maximally explain the variation of the data, in order to explain the output Y using the input X. The purpose of PLS is to find the relationship between X and Y while making the error matrix in the quality data, as small as possible.

Cross validation helps to select the number of latent variables. A small number of latent variables should explain the greatest variation in the input and output variables.

In PLS, the monitoring statistics that are used are T^2, SPE_x and SPE_y.

For Multivariate SPC, the difference between PCA and PLS is that PCA monitors a process through a single block of information whereas PLS considers the relationship between the process and quality variables. The process is monitored through a model of the quality variables that was developed from the process information. The monitoring statistics of PLS method (T^2, SPE_x and SPE_y) can also help in identifying the root cause of a signal by relating it to the X or Y variable.

Berismis *et al.* (2007) conclude that PCA and PLS techniques are mostly used in the area of chemometrics but that they can be used for any type of multivariate process.

3.8 Conclusion

In addition to the multivariate techniques discussed in Chapter 2, Principal Component Analysis and modelling can be used for monitoring and statistical process control in continuous processes.

The steps for implementation are summarised in Figure 3.8.

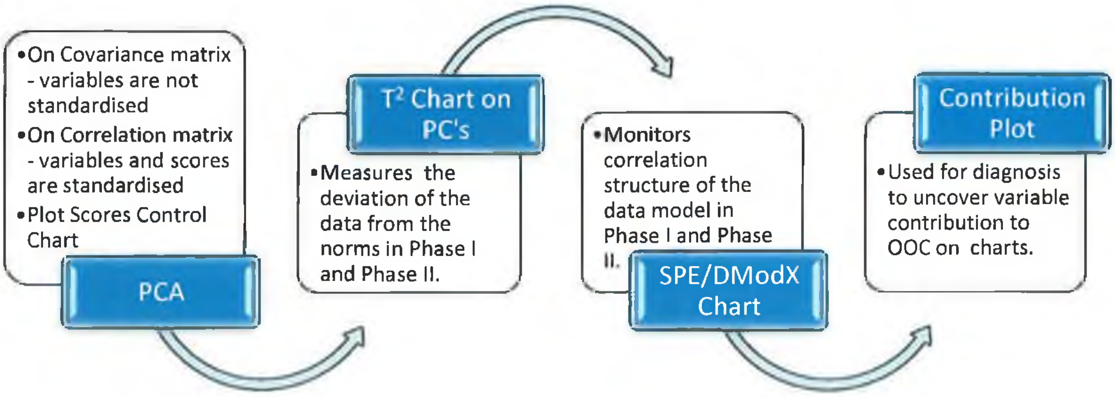

•On Covariance matrix - variables are not standardised
•On Correlation matrix - variables and scores are standardised
•Plot Scores Control Chart
PCA
T² Chart on PC's
•Measures the deviation of the data from the norms in Phase I and Phase II.
•Monitors correlation structure of the data model in Phase I and Phase II.
SPE/DModX Chart
Contribution Plot
•Used for diagnosis to uncover variable contribution to OOC on charts.

Figure 3.8 – Multivariate Principal Component Analysis implementation steps

Not all processes are continuous, some are batch processes. They will be discussed in the next chapter.

3.9 References

AlGhazzawi, A., and Lennox, B. (2009), "Model predictive control monitoring using multivariate statistics", *Journal of Process Control*, 19, pp. 314-327.

Bersimis, S., Psarakis, S., Panaretos, J. (2007), "Multivariate Statistical Process Control Charts: An Overview", *Quality and Reliability Engineering International*, Vol.23, pp. 517-543.

Chiang, L.H., and Colegrove, L.F. (2007), "Industrial Implementation of on-line multivariate quality control". *Chemometrics and Intelligent Laboratory Systems*, Vol. 88, pp. 143-153.

CPACT (Centre for Process Analytics and Control Technologies), School of Chemical Engineering and Advanced Materials, Newcastle University, UK. "Multivariate Data Analysis & Statistical Process Control" 3-day Continuing Professional Development Course, 21-23 October 2008.

Dillon, W.R., and Goldstein, M. (1984), *Multivariate Analysis*, 2nd edition, Wiley, New York.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wold, S. (2001), *Multi- and Megavariate Data Analysis: Principals and Application,* Umetrics AB.

Ferrer, A. (2007), "Multivariate Statistical Process Control Based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process", *Quality Engineering*, Vol.19, No.4, pp. 311-325.

Fuchs, C., and Kenett, R. (1998), *Multivariate Quality Control*, Marcel Dekker, Inc.

Hotelling, H. (1933), "Analysis of a Complex of Statistical Variables into Principal Components", *Journal of Educational Psychology*, Vol. 24, pp. 417-441, pp. 498-520.

Jackson, J.E. (2003), *A User's Guide to Principal Components*, Published by John Wiley & Sons, Inc. Hoboken, NJ.

Jackson, J.E., and Mudholkar, G.S. (1979), “Control Procedures for Residuals Associated with Principal Components Analysis”, *Technometrics*, Vol. 21, No.3, pp. 341-349.

SAS Institute Inc. JMP version 7.0.2.

Kourti, T., and MacGregor, J.F. (1995), “Process analysis, monitoring and diagnosis, using multivariate projection methods”, *Chemometrics and Intelligent Laboratory Systems*, Vol. 28, pp. 3-21.

Kourti, T., and MacGregor, J.F. (1996), “Multivariate SPC Methods for Process and Product Monitoring”, *Journal of Quality Technology*, Vol. 28. No.4, pp. 409-428.

MacGregor, J.F., and Kourti, T. (1995), “Statistical Process Control of Multivariate Processes”, *Control Eng. Practice*, Vol. 3, pp. 403-414.

Montgomery, D.C. (2005), *Introduction to Statistical Quality Control*, 5th Edition, Wiley & Sons.

Nomikos, P., and MacGregor, J.F. (1995), “Multivariate SPC Charts for Monitoring Batch Processes”. *Technometrics*, Vol.37, No.1, pp. 41-59.

Pearson, K. (1901), “On Lines and Planes of Closest Fit to Systems of Points in Space”, *Phil. Mag. Ser. B*, Vol.2, pp. 559-572.

Umetrics Inc, SIMCA P+, version 12.0.1.0.

Wikstrom, C., Albano, C., Eriksson, L., Friden, H., Johansson, E., Nordahl, A., Ranner, S., Sandberg, M., Kettaneh-Wold, N., Wold, S. (1998), “Multivariate process and quality monitoring applied to an electrolysis process Part I. Process Supervision with multivariate control charts”, *Chemometrics and Intelligent Laboratory Systems*, Vol. 42, pp. 221-231.

Yang, K., and Trewn, J. (2004), *Multivariate Statistical Methods in Quality Management*, McGraw-Hill, New York.

CHAPTER FOUR

MULTIVARIATE SPC FOR BATCH PROCESSES

4.1 Introduction

A Multivariate control system needs to be adaptable to all processes in industry, otherwise it has a limited capability. Continuous processes are most common in automated and semi-automated systems where multiple observations are generated and collected every second. This is usually conducted over a certain time period, i.e. day, shift. The individual process control charts are monitored regularly by the operators. For online systems this can be problematic due to the number of false alarms that can be encountered.

Another common process in industry is batch processes. Data is collected after a particular process and a decision can be made to determine what the next steps are.

Slack, Chambers and Johnston (2007) identify the simple differences between a batch and a continuous process. They highlight that continuous process are managed over long periods of time at high volumes. It is an endless run of production where each unit of product is undivided. The effortless flow from one stage of the process to another is the principal characteristic of a continuous process.

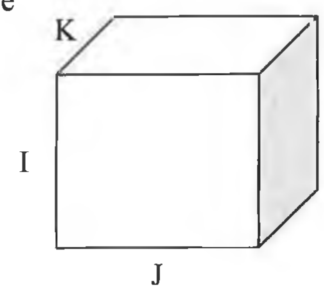
They describe that a batch process has a process route that it follows, where groups of product are treated together. A batch process is repetitive if a batch is of a large volume. The size of a batch can vary from very small to very high volumes. A batch has a beginning and an end.

CPACT (2008) describe the main characteristics of a batch process as,

- They are finite in duration,
- Can be of low volume, so small production runs,
- Usually consist of high value products,
- They may have complex mechanisms.

A batch process can be 3-dimensional; where the

- Batches represent I dimension,
- Variables represent J, and
- Time is represented by K.



The T^2 Statistic has the ability to deal with both continuous and batch processes.

4.2 Batch Processing

Batch process SPC differs from continuous processes as it is a finite process in duration. Batch processing typically uses the data collected from passing batches to create a model from which future batch progress are monitored against.

4.2.1 T^2 on Batch Observations

Mason *et al.* (2001), discuss two categories of batch processes, namely Category 1 and Category 2. Calculations of the T^2 statistic and UCL for the T^2 chart will differ, depending on the batch category. Phase I for removing outliers and finding an appropriate baseline in which to use for Phase II is described for each category.

Mason *et al.* (2001) classifies a process as Category 1 if it is assumed that the observations come from the same p-dimensional normal distribution. The mean vector, μ, and covariance matrix, $\sum$, will be common. This type of batch process will manufacture product similar to a continuous process. This is the type of batch process that will be discussed in this research.

Category 2 batch processes assume that the observations come from different multivariate normal distributions, $N_p(\mu_i, \sum)$. μ_i, $i = 1, 2, \ldots, k$, is the population mean vector of the i^{th} batch. The batch mean vectors are separated but are contained within a defined acceptable region.

In their paper, Mason *et al.* (2001) discuss how batch sizes can affect the overall calculation of the batch means. Suppose data is collected for a number of batches, r, and the batch size is equal to one, so that on each batch, there is only one observation for p variables.

$$n_i = 1, \text{ where } i = 1, \ldots, r.$$

The total sample size is $N = \sum_{i=1}^{r} n_i = r$.

Hence, the method in Phase I, for screening for outliers for individual observations on a continuous process can be applied where each batch in a batch process has one observation.

The equation for calculating the Phase I T^2 Statistic on a batch process, where the batch size = 1 is,

$$T^2 = (X - \bar{X})' S^{-1} (X - \bar{X}) \sim \left[\frac{(\mathrm{N}-1)^2}{\mathrm{N}}\right] \beta_{(\mathrm{p}/2,(\mathrm{N}-\mathrm{p}-1)/2)} \qquad (4.1)$$

where $\beta_{(\mathrm{p}/2,(\mathrm{N}-\mathrm{p}-1)/2)}$ is a beta distribution with parameters $p/2$ and $(N - p - 1)/2$.

The UCL is calculated using,

$$\mathrm{T}^2{}_{\mathrm{UCL}} = \frac{(\mathrm{N}-1)^2}{\mathrm{N}} \beta_{(\alpha,\mathrm{p}/2,(\mathrm{N}-\mathrm{p}-1)/2)} \qquad (4.2)$$

for chosen α, where $\beta_{(\alpha,\mathrm{p}/2,(\mathrm{N}-\mathrm{p}-1)/2)}$ is the upper α^{th} quantile of $\beta_{(\mathrm{p}/2,(\mathrm{N}-\mathrm{p}-1)/2)}$.

Equations 4.1 and 4.2 are similar to Equations 1.11 and 1.16, for calculating the T^2 Statistic and UCL for individual observations.

Phase II is calculated in a similar manner to the individual observation situation. Consider an observation vector, X, with an unknown mean vector, $\bar{X}$. The covariance matrix, S, is taken from Phase I with a baseline size $N = nr$.

The T^2 Statistic, for a future batch containing m observations, where $m = 1$ is given by,

$$T^2 = (X - \bar{X})' S^{-1} (X - \bar{X}) \qquad (4.3)$$

The corresponding UCL is given by,

$$T^2{}_{UCL} = \frac{(N+1)(N-1)p}{N(N-p)} \mathrm{F}_{(\alpha,\mathrm{p},\mathrm{N}-\mathrm{p})}, \qquad (4.4)$$

for given α, where $N = nr$, is the size of the baseline dataset from Phase I,

p is the number of variables, r is the number of batches and $\mathrm{F}_{(\alpha,\mathrm{p},\mathrm{N}-\mathrm{p})}$ is the α^{th}quantile of $\mathrm{F}_{(\mathrm{p},\mathrm{N}-\mathrm{p})}$.

Equation 4.3 is similar to Equation 1.10, for calculating a Phase II T^2 Statistic and Equation 4.4 is similar to Equation 1.17, for calculating the UCL, for individual observations.

These equations are also similar to the ones proposed by Ferrer (2007) in Chapter 3, section 3.5 where he described plotting the T^2 statistic in the score space.

4.2.2 Batch Process Data

Nomikos and MacGregor (1995) discuss an alternative method for analysing and monitoring a batch process. They explain how a batch has a recipe of materials, such as chemicals, that are processed under controlled conditions, according to some specified time trajectories, in which the process variables are varied. Once a batch is completed, the product is then tested to see if it is of good quality. This is done by measuring certain quality variables from a sample of the batch, usually in a laboratory. There will be some batch-to-batch variation which can happen for any number of reasons. This variation will ultimately lead to undesired conditions where the current batch and subsequent batches, will be of poor quality.

On-line monitoring of critical process variables at specific times in the process may alleviate the problem of having poor quality batches detected after a batch has been completed. This will enable earlier detection and correction of issues arising earlier in the process.

The process variables are monitored at certain times in the process, using historical data. It doesn't matter whether they are successful or unsuccessful batches, as they will contribute valuable information in which to build a model. This model can then be used to characterise what conditions a batch requires in order to be successful.

There are many process variables that can be measured, which vary over time and can be highly correlated. As batches are finite and non-linear in nature, building a model can prove to be difficult.

CPACT (2008) describe various methods for multivariate process control. Their research is primarily based on chemical data.

PCA and PLS are discussed in Chapter 3. These are bi-linear techniques. Batch data must be transformed some way in order to convert from a 3-dimensional into a 2-dimensional space. Continuous data is bi-linear and is represented by a 2-dimensional matrix.

The data can be analysed a number of ways,

- Unfolding the data into a 2-dimensional array, which is similar to continuous process, and then applying PCA or PLS.
- Keeping the tri-linear form and applying multi-linear techniques.

4.2.3 Multiway PCA

MacGregor and Nomikos (1992) presented a paper on monitoring batch processes and in another paper they, (Nomikos and MacGregor, 1994), extended multivariate SPC methods used in continuous processes to Multiway Principal Component Analysis (MPCA).

In their 1995 paper, Nomikos and MacGregor discuss the measurement of many variables over the finite duration of a batch process, using MPCA. Control Limits are calculated from distributional information obtained from historical data.

Multi-way PCA decomposes the three-way data array, $\underline{X}$ or X, into a series of principal components that consists of score vectors (t_r) and loading matrices (p_r), or unfolded vectors ($\mathrm{p_r}$), plus a residual E, which is as small as possible, in a least squares sense.

$$\underline{\mathrm{X}} = \sum_{r=1}^{R} t_r \otimes P_r + \underline{\mathrm{E}} \qquad \text{or} \qquad X = \sum_{r=1}^{R} t_r\, p_r{}' + E$$

where $r = 1, \ldots R$, are the retained principal components.

Decomposition of the data block is in two parts. The residual, E, describes the noise in the data. The other part, $\sum_{r=1}^{R} t_r\, p_r{}'$, expresses the dataset as two fractions. One fraction (t_r) related to the batches and the other (p_r) related to the variables and their time variation.

The loading matrix, $P_r(J\ x\ K)$, contains most of the structural information about how the variable measurements deviate from their mean trajectories under normal operating conditions.

To test a new batch for unusual occurrences, the loading matrix can be used by obtaining the predicted t scores and residuals for a new batch, $X_{new}\ (J\ x\ K)$.

If the t-scores of the new batch are within the normal operation region and its residuals are small, then no unusual behaviour is detected and the batch is performing the same as that of the reference dataset.

As with PCA, the t-scores can be plotted, as can Hotellings T^2 and the DModX statistic.

4.2.4 Batch-wise and Variable-wise Unfolding

In their 1995 paper, Nomikos and MacGregor suggested unfolding the batch data (Batch-wise unfolding). By unfolding the three-way array, $\underline{X}\ (I\ x\ J\ x\ K)$, into slices using Multiway PCA. The slices can be rearranged into a two-dimensional matrix where PCA can then be performed. The PC score vectors contain information on batch-to-batch variation and the loading matrices show the variables behaviour over time.

CPACT (2008) simplified the diagram given by Nomikos and MacGregor (1995), Figures 4.1 and 4.2, which shows how the data array, $\underline{X}$, is unfolded slice by slice into a 2-dimensional matrix, X, where PCA can then be applied.

Wold *et al.* (1998) proposed unfolding the variables of the data (variable-wise unfolding). This approach will only capture the covariance structure of the variables and it doesn't account for the behaviour of the batch process. From Figure 4.1, the first few K rows are measurements for the first few batches. If variables are only measured for a portion of a batch, this will lead to missing data. The behaviour of a batch is monitored using the scores. The loading vectors contain information on the variables. No assumptions are needed for future measurements.

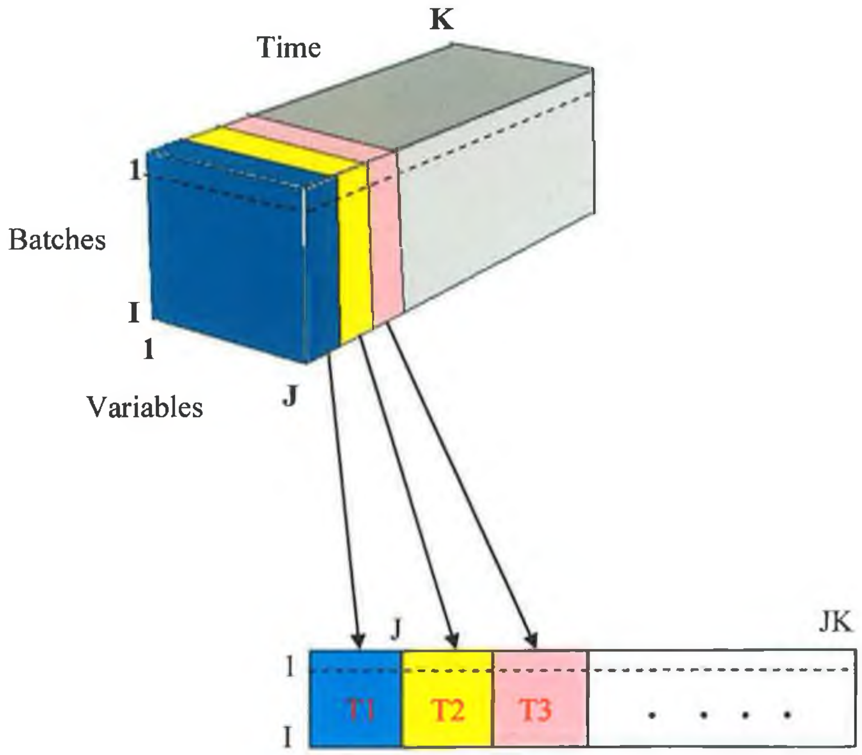


Figure 4.1 – Unfolding a three-way matrix into a two-dimensional matrix

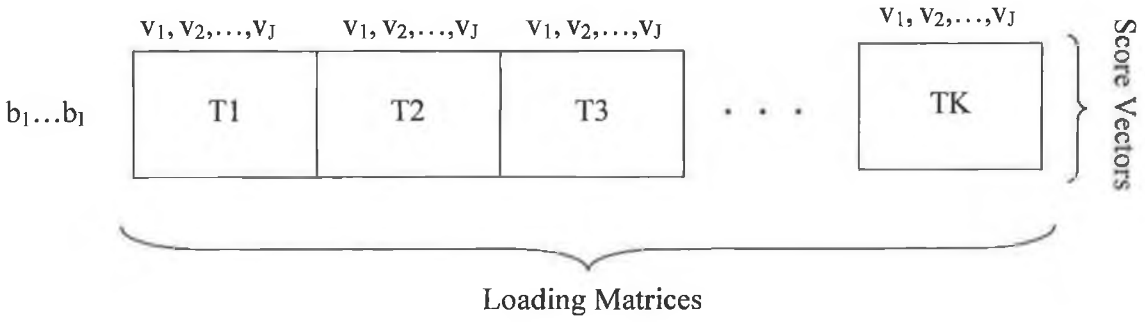


Figure 4.2 – Two-dimensional matrix

The horizontal slices $(J\ x\ K)$ are the loading matrices, which contains information on the batch-to-batch variation i.e. each slice describes the trajectories of all the variables from a single batch (i).

The vertical slices $(I \, x \, J)$ are the score vectors, which reflect the behaviour of the variables over time i.e. each slice describes the values of the variables for all batches at the same time interval (k).

Nomikos and MacGregor (1995) suggested that using the vertical slices to analyse and monitor a batch process is the most meaningful way of unfolding $\underline{X}$.

- The vertical slices must be arranged into a two-dimensional matrix, $X \, (I \, x \, JK)$. These represent each time point, T1 representing the first time point, as shown in Figure 4.2.
- Mean scaling and centring the data must then be completed before carrying out PCA. Unfolding this way, by subtracting the mean trajectory from each process variable, removes the non-linearity associated with batch processes.
- PCA is then performed on this adjusted data, i.e. PCA is applied to the distances instead of the measured trajectory.

4.2.5 Dealing with Incomplete Batches

Nelson *et al.* (2006) discuss a problem with the method presented above. As a batch is finite in duration, the batch must be complete in order to calculate if it performed successfully. As the batch progresses, it would not be appropriate to use the above method to test the data. The matrix X_{new} will not be completed until the batch has finished. Each time interval (k) only has the measurements up until that particular interval in time. The rest is undefined.

Nomikos and MacGregor (1995) suggested three methods in order to predict future unknown observations for an incomplete batch, X_{new}. These methods for in-filling are,

1. Using zero deviations. This assumes that future observations will operate along the mean trajectory as calculated from the reference dataset. However, this method has a sensitivity issue, as it is unable to detect a fault at the start of a batch.
2. Using the current deviations. This assumes that future deviations will continue to operate at the same level as the current values for the remainder of the batch.

3. Using the missing data method. There is no in-filling applied. The future observations are considered as missing values. The principal components of the reference dataset can be used to predict the missing values by using the observed values up to the time interval and using the correlation structure of the measurement variables in the reference dataset defined by the model.

Cho and Kim (2003) proposed another method for predicting future observations of a batch. They propose using past batch trajectories, by choosing a past batch trajectory that is comparable to the current batch, from a library of batch trajectories.

Kourti (2005) strongly recommends against in-filling or using missing data methods. Kourti states that: "The reason that we should not fill with missing data is simple: *these variables are missing at the same time interval for all the batches in the data base; the behaviour of a variable that is never measured at a certain time interval, is not observable at that time interval.*"

Nelson *et al.* (2006) discuss the issues with the use of measurement sets that are incomplete. They analyse the impact that using the missing measurements presents uncertainties in the predictions, scores, Hotellings T^2 and SPE statistics as well as the contributions.

Wurl *et al.* (2001) also discuss batch processing but they focus on a PLS model. They suggest further dividing the process into two stages, namely startup stage and production stage.

4.3 Conclusions

Batch data has its own requirements as it is a finite process with complex systems. It can require a different approach than continuous processes, as batch processes are 3-dimensional and continuous processes are 2-dimensional.

There are various methods identified for analysing batch data. Mason *et al.* (2001) identify a simple technique for modelling batch data which is similar to the continuous method. They propose calculating a T^2 Statistic and UCL based on a particular category of batch data.

Unfolding a 3-dimensional matrix into a 2-dimensional matrix in order to build a model using PCA or PLS was proposed by Nomikos and MacGregor (1995), and Wold *et al.* (1998). These methods are mainly applied to the chemical industries where a process is a batch process from start to finish. There are drawbacks to this approach, which involves dealing with incomplete batch data. Some suggest methods for infilling.

The approach to batch data proposed by Mason *et al.* (2001) will be considered for assessing a batch process as part of this research. The research data is suitable to a Category 1 batch process.

4.4 References

Cho, H.W., and Kim, K.J. (2003), "A Method for Predicting Future Observations in the Monitoring of a Batch Process". *Journal of Quality Technology*, Vol. 35, No. 1, pp. 59-69.

CPACT (Centre for Process Analytics and Control Technologies) (2008), School of Chemical Engineering and Advanced Materials, Newcastle University, UK. "Multivariate Data Analysis & Statistical Process Control", 3-day Continuing Professional Development Course, 21-23 October 2008.

Ferrer, A. (2007), "Multivariate Statistical Process Control Based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process", *Quality Engineering*, Vol.19, No.4, pp. 311-325.

Kourti, T. (2003), "Abnormal situation detection, three-way data and projection methods; robust data archiving and modelling for industrial applications", *Annual Reviews in Control*, Vol. 27, pp. 131-139.

MacGregor, J.F., and Nomikos, P. (1992), Monitoring Batch Processes, *Batch Processing Systems Engineering: Current Status and Future Directions* (NATO ASI Series F), eds. Reklaitis, Rippin, Hortacso, and Sonol, Heidelberg: Springer-Verlag.

Mason, R.L., Chou, Y-M, Young, J.C. (2001), "Applying Hotellings T^2 Statistic to Batch Processes". *Journal of Quality Technology*, Vol. 33, pp. 466-479.

Mason, R.L., and Young, J.C. (2002), Multivariate Statistical Process Control with Industrial Applications, ASA-SIAM.

Nelson, P.R.C., MacGregor, J.F., Taylor, P.A. (2006), "The impact of missing measurements on PCA and PLS prediction and monitoring application", *Chemometrics and Intelligent Laboratory Systems,* Vol. 80. pp. 1-12.

Nomikos, P., and MacGregor, J.F. (1994), "Monitoring of Batch Processes Using Multiway Principal Components Analysis", *AIChE Journal*, Vol. 40, pp. 1361 – 1375.

Nomikos, P., and MacGregor, J.F. (1995), "Multivariate SPC Charts for Monitoring Batch Processes", *Technometrics,* Vol. 37, No.1, pp. 41-59.

Slack, N., Chambers, S., Johnston, R., (2007), *Operations Management*, fifth edition, Pearson Education.

Wold, S., Kettaneh, N., Friden, H., Holmberg, A. (1998), “Modelling and diagnostics of batch processes and analogous kinetic experiments”. *Chemometrics and Intelligent Laboratory Systems*, Vol. 44, pp. 331-340.

Wurl, R.C., Albin, S.L., Shiffer, I.J. (2001), “Multivariate Monitoring of Batch Process Startup”. *Quality and Reliability Engineering International*, Vol. 17, pp. 269-278.

CHAPTER FIVE

FUEL CELL TECHNOLOGY AND MANUFACTURING PROCESS

5.1 Introduction to Fuel Cell Technology

A fuel cell is an electrochemical device that converts the chemical energy of a fuel into electrical energy such as electricity. The fuel cell works like a battery in principle, it generates power and as long as there is fuel, there will be power. Linden and Reddy (2001) explain the essential difference between a battery and a fuel cell is the way the source of energy is supplied. When power is needed in a fuel cell, the fuel and oxidant are provided by an external source. In a battery, the fuel and oxidant are elements of the device. When the reactant is depleted, electrical energy will no longer be generated. Oxygen, or air, is the most common oxidant used in fuel cells. Details on fuel cells can be found on Fuel Cells 2000 website <http://www.fuelcells.org>.

Linden and Reddy (2001) give an overview about how the fuel cell operates. A fuel cell consists of two electrodes, an anode and a cathode. They are catalysts which enable the reaction between the fuel and the oxidant. They are saturated by an electrolyte and divided by a gas barrier. The fuel travels along one electrode while the oxidant travels along the other electrode. This is illustrated by Linden and Reddy (2001), through Figure 5.1.

The ions and electrons generated from the fuel, along with the oxygen, come together on the surface of the electrode which the oxidant travels along. Water is produced as a by-product and, of course, electrical energy is produced as the output.

Fuel Cells come in many different forms, some which include, proton exchange membrane fuel cell, PEMFC, direct methanol fuel cell, DMFC and direct liquid fuel cell, DLFC, along with many more. They are differentiated based on the type of electrolyte used.

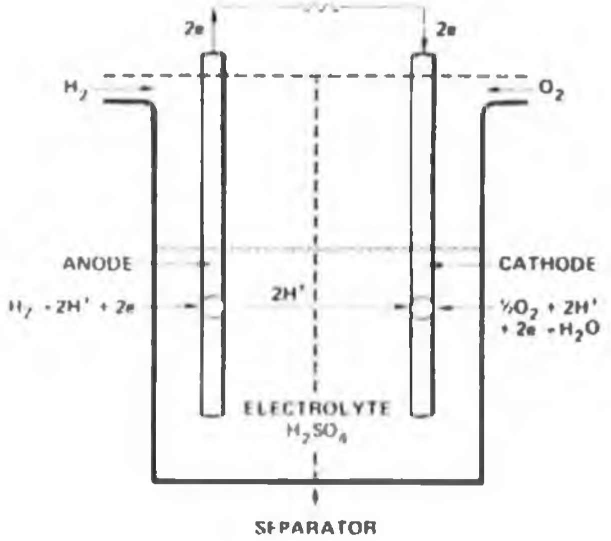

2e
2e
H_2
O_2
ANODE
CATHODE
$H_2 \rightarrow 2H^+ + 2e$
$2H^+$
$\frac{1}{2}O_2 + 2H^+ + 2e \rightarrow H_2O$
ELECTROLYTE
H_2SO_4
SEPARATOR

Figure 5.1 – Typical Fuel Cell (Linden and Reddy, 2001)

Xue *et al.* (2006) discuss the complexity of a polymer electrolyte membrane (PEM) fuel cell. These systems have limitations with respect to the components used in the fuel cell. The assembly of the critical components, the membrane and the electrodes, can create fault conditions. Failure analysis has revealed three forms for failure of a fuel cell. They discuss the effects of,

1. Drying out of the membrane

2. Fuel starvation for electrochemical reaction

3. Leaking of the membrane

The DLFC uses a borohydride technology. This type of fuel cell is the subject of this research. It is a micro-fuel cell based secondary power source. It supports the batteries that are currently used on a device and it is designed to power and charge most common handheld electronic devices such as mobile phones and MP3 players <http://www.medistechnologies.com>. It has an integrated fuel cartridge that generates power immediately after activation. This is achieved by filling the fuel cell with fuel from the cartridge.

The benefits of a Fuel Cell are discussed on Fuel Cells 2000 website <http://www.fuelcells.org>, these include,

- Low to Zero emissions – Fuel cells are very quiet, as there are no fan systems, so there is no noise pollution.

- Highly efficient – Fuel cells do not burn fuel, they obtain their energy electrochemically.

In addition to these benefits, a portable fuel cell is engineered by Medis Technologies. Details on their portable fuel cells can be found on their website <http://www.medistechnologies.com>. They describe them as,

- Easily transportable – They are lightweight and can even be carried and used on an airplane.

- Recyclable – They are recyclable and are also RoHS compliant (Restriction of Hazardous Substances). This ensures that levels of certain chemical elements meet EU standards, which makes it a green environmentally friendly product.

- Safe – They are UL (Underwriters Laboratories) and CE listed.

5.2 Overview of Manufacturing Process

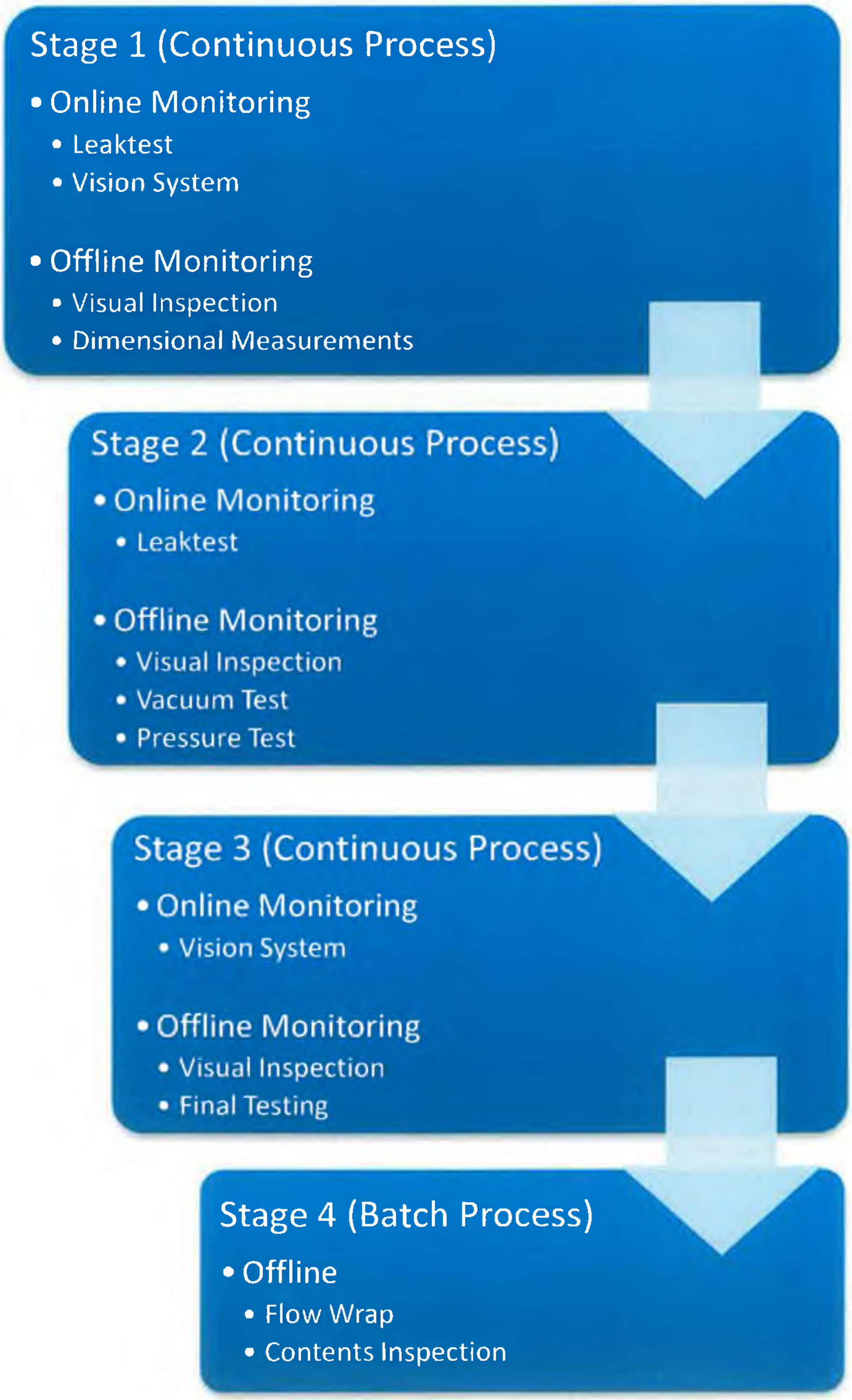
This research was conducted at a contract manufacturing company based in Galway, Ireland. The company provided an upgrade to an existing Fuel Cell manufacturing process based in Israel. The facility in Israel is an R&D based facility, which manufactures approximately one thousand fuel cells monthly. The manufacturing process in Ireland provides a high volume automated facility, which has the ability to produce over one million fuel cells monthly. This line is the world's first high volume, fully automated assembly line, to mass manufacture a fuel cell for portable devices. The equipment required to assemble this fuel cell was custom made in order to achieve this. Due to confidential information, the fuel cell manufacturing process and laboratory testing cannot be discussed in detail.

5.2.1 Stages of Fuel Cell Assembly

The Fuel Cell manufacturing line is divided into four stages, each of which is treated as an independent process, as shown in Figure 5.2.

- Stage 1 manufactures the Cell Core of the fuel cells. This is where the electrical components, anode and cathode are assembled.
- Stage 2 is the Fuel Module manufacture. A container is assembled and then filled with a chemical fuel. The filled container is called the fuel cartridge.
- Stage 3 assembles the Cell Core to the fuel cartridge and places them into an outer casing.
- Stage 4 is a Pack line where the fuel cell and operating instructions are contained in the outer packaging.

The fuel used to power the fuel cell is stored in the cartridge. The cell core and fuel are not mixed until a customer has activated it for the first time after purchase. Once activated, the fuel cell is an immediate power source, so there is no need to warm up before it starts generating power. There are approximately 20 watt hours of power available until the fuel cell is depleted.



Stage 1 (Continuous Process)

- Online Monitoring
  - Leaktest
  - Vision System
- Offline Monitoring
  - Visual Inspection
  - Dimensional Measurements

Stage 2 (Continuous Process)

- Online Monitoring
  - Leaktest
- Offline Monitoring
  - Visual Inspection
  - Vacuum Test
  - Pressure Test

Stage 3 (Continuous Process)

- Online Monitoring
  - Vision System
- Offline Monitoring
  - Visual Inspection
  - Final Testing

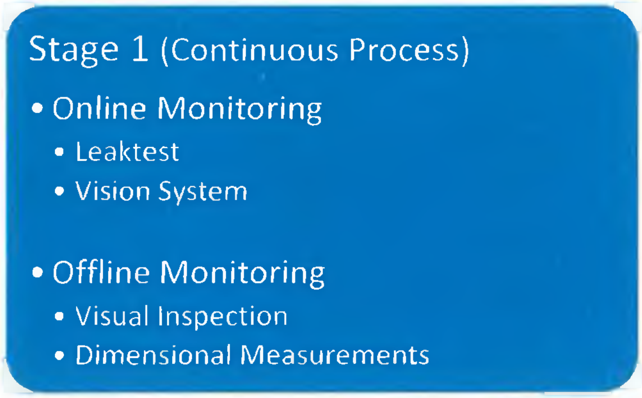
Stage 4 (Batch Process)

- Offline
  - Flow Wrap
  - Contents Inspection

Figure 5.2 – Manufacturing Process Flow

Online and offline monitoring systems are used to check the quality of the product as it moves through the production line. The online monitoring systems check 100% of the product for various attributes. Offline monitoring is facilitated by audit stations, which are located in each of the four stages of the manufacturing line. The offline monitoring is done in a laboratory. Samples are automatically removed from the production line at a predetermined rate using an automated system via the audit station. Various tests are performed on these samples to monitor quality attributes. These tests are detailed in the following sections.

5.2.1.1 *Stage 1*



Stage 1 (Continuous Process)

- Online Monitoring
  - Leaktest
  - Vision System
- Offline Monitoring
  - Visual Inspection
  - Dimensional Measurements

Figure 5.3 – Stage 1 Testing

Stage 1 on the manufacturing line is a continuous process. This stage continually assembles the components using the raw materials. Figure 5.3 details the tests performed at Stage 1 of the process.

Online monitoring in Stage 1 includes,

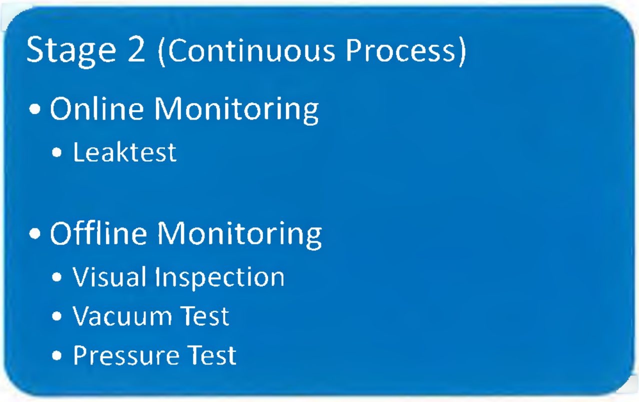
- Leak testing: leak testers are used to test 100% of product during the assembly of the various components. The leak test checks the integrity of the assembled parts.

- Vision systems are used to detect defective product or incorrectly assembled parts, in order to meet requirements.

Stage 1 has four audit stations. As the samples are removed from the line, via the audit stations, visual inspections are performed on each assembly. The visual inspections look for defects associated with that particular stage of the process. Results are logged onto a monitoring system as a good part or bad part. This calculates the number of good and bad parts produced per hour.

These samples are then taken to the laboratory for dimensional checks. This is done via a contact measurement system, which has all the specifications for each assembly pre-programmed into it. This system will display Statistical Process Control charts for each characteristic of each assembly. Quality Inspectors respond to out of specification measurements as detailed in Sampling and Response Plans.

5.2.1.2 *Stage 2*



Stage 2 (Continuous Process)

- Online Monitoring
  - Leaktest
- Offline Monitoring
  - Visual Inspection
  - Vacuum Test
  - Pressure Test

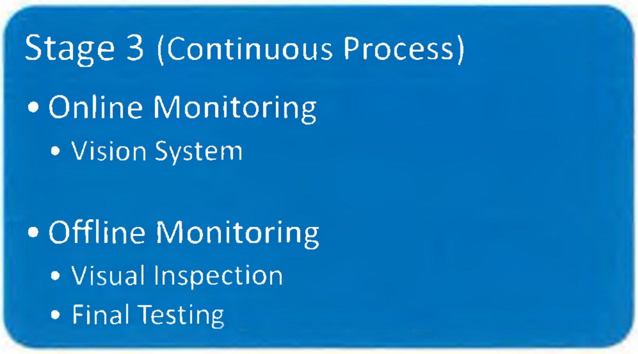
Figure 5.4 – Stage 2 Testing

Stage 2 is also a continuous process. It has two audit stations. Figure 5.4 details the tests performed at Stage 2 of the process.

Leak testing is performed as part of online monitoring in Stage 2. Leak testers are used to test 100% of product during the assembly of the fuel container. The leak test checks the integrity of the assembled parts before and after the container is filled with fuel.

Samples are removed via the audit stations. They undergo a vacuum test and pressure test in addition to a visual inspection. The vacuum and pressure tests are performed on several areas of the fuel container. This is to ensure there will be no chemical leakage resulting from the assembly of the fuel container.

5.2.1.3 *Stage 3*



Stage 3 (Continuous Process)

- Online Monitoring
  - Vision System
- Offline Monitoring
  - Visual Inspection
  - Final Testing

Figure 5.5 – Stage 3 Testing

Stage 3 is a continuous process, and it has two audit stations. Figure 5.5 details the tests performed at Stage 3 of the process.

Online monitoring in Stage 3 involves a vision system, which is used to detect defective product or incorrectly assembled parts.

Once the visual inspections are complete, the samples are taken to the laboratory for Final Testing. The final tests that are performed include,

- Discharge Testing – this is a destructive test that measures the performance of the fuel cell. It does this by measuring the electrical output of a fuel cell, i.e. Power and Energy.

- Safety Testing – this is to ensure there is no leakage of corrosive chemicals.

- Activation Testing – this measures the force required to break a membrane that will activate the fuel cell.

- Shelf Life Testing – this measures the time period from which the fuel cell can be left unactivated and still perform satisfactorily.

Units not meeting the required criteria at Stage 3 Final Testing, are sent for Failure Analysis. Laboratory technicians will then try to determine the cause of failure.

Meanwhile, at the end of Stage 3, the manufacturing line artificially creates batches from a continuous process. Every 2000 units are considered to be a batch. This was determined to be an adequate number in order to contain an issue in the event that one should arise. Each batch has a unique batch number.

5.2.1.4 *Stage 4*

Stage 4 (Batch Process)

- Offline Monitoring
  - Flow Wrap
  - Contents Inspection

Figure 5.6 – Stage 4 Testing

Stage 4 is the pack line. This stage also has two audit stations. Figure 5.6 details the tests performed at Stage 4 of the process.

This packs units for each batch into a flow wrap package. Each unit is then placed into a separate box along with the product information. Samples are inspected to ensure that all required components are contained in the box. Units are then packed into outer boxes and placed on a pallet in the warehouse. They will be shipped directly to the customer, from the warehouse. Their shipping status is dependent on the results of the final testing performed in Stage 3.

Once the Stage 3 final tests are complete, a decision can be made in relation to the quality of each batch, and acceptable batches can be released for shipment.

5.3 Performance Measurement Tests

There are many measurements and tests carried out at various points in the process. This research will investigate the performance measurements and assess the results using Multivariate statistical methods. The discharge test performed at Stage 3, measures the electrical output in terms of Power and Energy. Power and Energy will ultimately be what the customer will get from using the product.

As this is the first high volume automated fuel cell line, purposely built for this technology, qualification production runs were carried out to ensure capability of the process and the product functionality. Once all "Risk production" runs were completed and results achieved, the "Production" phase commenced.

All final testing is destructive, which has a huge impact on costs. Therefore, it is in the best interest of the company to have a minimum amount of sampling in the laboratory.

A batch consisting of 2000 units, from which 80 units are used in a discharge test. The discharge test, for performance, measures two parameters,

- Power (W),
- Energy (Wh)

40 units are used to assess each of these parameters.

The results of testing will determine the status of each batch, i.e. pass, fail or hold.

An automated monitoring system ensures that voltage measurements are performed every minute and converted to power and energy outputs, W and Wh, respectively. These results are then automatically exported to a database for analysis.

Electrical power is calculated from the voltage measurements using Ohms Law (Holzner, 2005),

$$V = IR$$

where V is the voltage measured in volts, I is the current measured in amperes and R is the resistance measured in ohms. This formula can be rearranged to show that,

$$I = \frac{V}{R}$$

Gibilisco (2005) demonstrates that power can be calculated using,

$$P = IV$$

where P is power in watts.

Substituting *I* from Ohms Law,

$$P = \frac{V}{R}V = \frac{V^2}{R}$$

Electrical Energy is expressed as a function of power and time (Gibilsco, 2005),

$$E = Pt$$

where *E* is Energy measured in Watt Hours and *t* is time.

Figure 5.7 is a graph of the product label claim given for a particular fuel cell product <http://www.medistechnologies.com>. It shows the relationship between the voltage and the energy generated by the fuel cell over time.

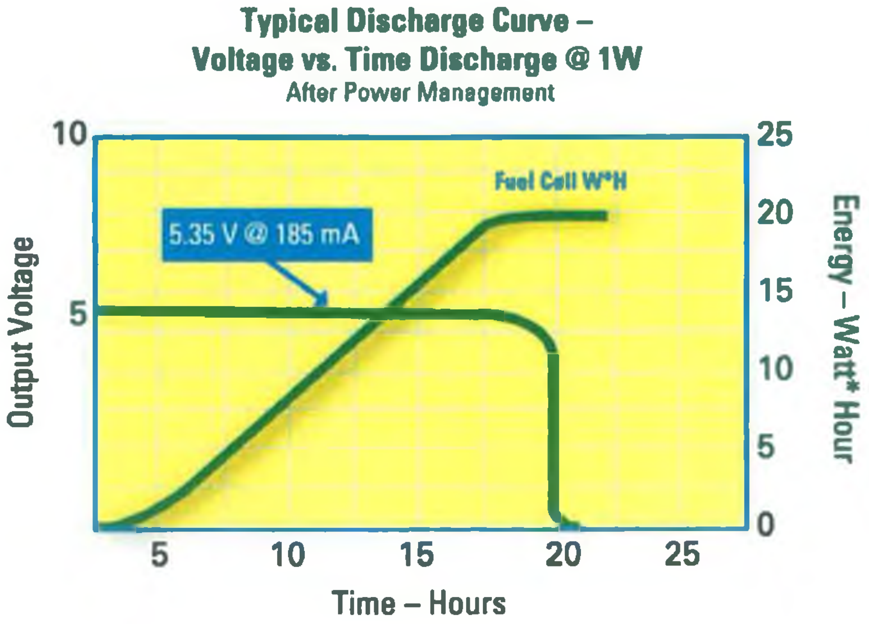


Figure 5.7 – Typical Discharge Graph of a Fuel Cell

Power

Power is measured at the *start up* phase. This stage is defined as the 5 minute period after activation. This is to ensure that the fuel cell generates power immediately after activation, as stated by the product claim.

The *working* power is also measured. This is defined at numerous time points throughout the life of the unit after the start up phase, namely 2h, 6h, and 12h.

Power is assessed at each of these stages, to ensure the reliability of the fuel cell is acceptable throughout the product life.

Figure 5.8 shows a power graph of twenty fuel cells. Each line on the graph shows the trend of the power output over time for a single fuel cell. Time is shown on the *x*-axis and the watts are shown on the *y*-axis. It is easier to see the trend of twenty cells instead of the forty on the one graph. Power and time values have been removed for confidentiality reasons.

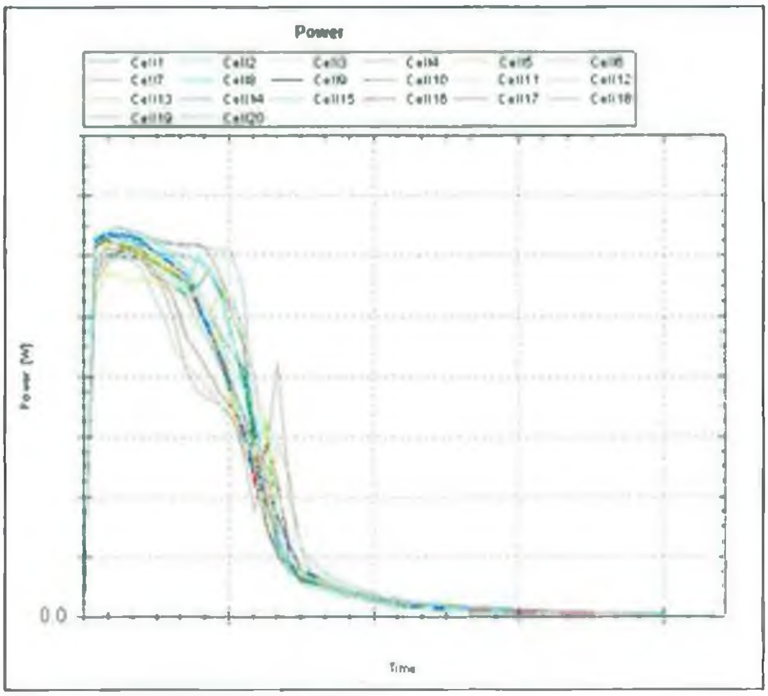


Figure 5.8 –Power Output from twenty Fuel Cells

Energy

Energy is also evaluated from the voltage measurements. Energy testing on a fuel cell is terminated once the energy from the fuel cell reaches a predefined cut-off value. This cut-off value defines when the energy has deteriorated significantly. The total number of Watt Hours (Wh) is then calculated.

Figure 5.9 shows the energy output from twenty fuel cells. Time is shown on the *x*-axis and the watt hours are shown on the *y*-axis for a single fuel cell. Each line on the graph shows the trend of the energy output over time. Energy and time values have been removed for confidentiality reasons.

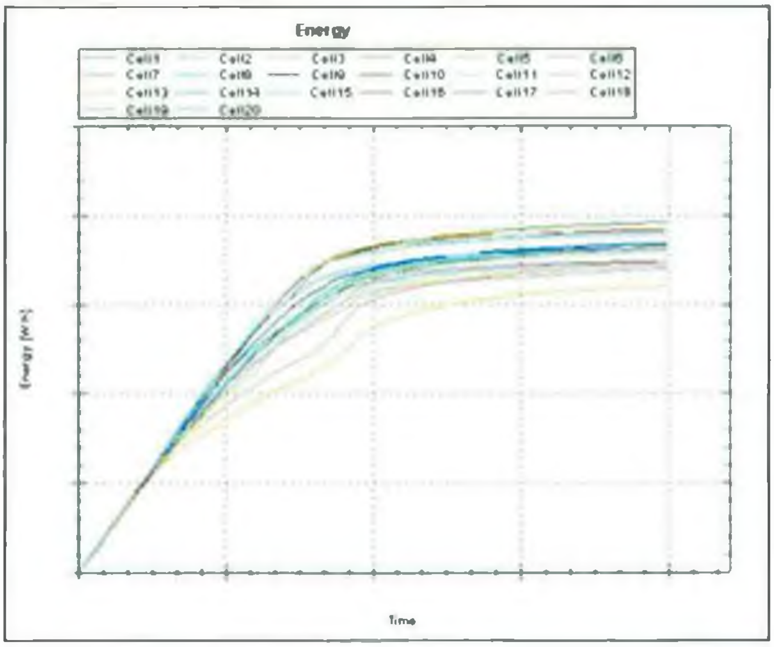


Figure 5.9 – Energy Output from twenty Fuel Cells

A batch average for power and energy is reported based on forty units. A decision is made for batch release based on these results. These are the parameters chosen for this research. The performance results contribute to the decision as to whether a batch receives an acceptable or unacceptable batch status.

It is critical that the customer gets an immediate source of power and also that the fuel cell generates enough power and energy over a period of time that is acceptable for the customer's requirements, based on the label claim, as shown in Figure 5.7. If these requirements are not met then the reputation of the fuel cells reliability is in jeopardy and repeat business would be threatened.

5.4 Conclusion

A description on fuel cell technology and how it works is presented. Some reasons for failures on a fuel cell are also highlighted. The manufacturing process for a new innovative micro fuel cell is described. The first three stages of the process are continuous and this feeds into the final stage where the batches are determined. It is an automated high volume production line that is capable of manufacturing over one million fuel cells a month. This equates to an average of 16 batches per day, operating on 24/7 (24 hours a day, 7 days a week) system.

Various product testing is performed at each stage of the process, which is facilitated by automatic sampling stations distributed throughout the line. The final product performance testing data, which evaluates power & energy, will be used for this research.

5.5 References

Fuel Cells 2000 (July 2009), Fuel Cells 2000, 1100 H Street NW, Suite 800, Washington, DC 20005 USA, viewed 18 July 2009, <http://www.fuelcells.org>.

Medis Technologies (2008), Medis Technologies Ltd., 805 Third Avenue, 15th Floor New York, NY 10022, viewed 18 July 2009, <http://www.medistechnologies.com>.

Gibilisco, S. (2006), *Teach yourself electricity and electronics*, McGraw-Hill, New York.

Holzner, S. (2005). *Physics for Dummies*, Wiley, New York.

Linden, D., and Reddy, T. (2001), *Handbook of Batteries*, McGraw-Hill, New York.

Xue, X., Tang, J., Sammes, N., Ding, Y. (2006), "Model-based condition monitoring of PEM fuel cell using Hotelling T^2 control limit". *Journal of Power Sources*, Vol. 162, pp. 388-399.

CHAPTER SIX

MULTIVARIATE SPC TO MONITOR FUEL CELL PERFORMANCE

6.1 Introduction

The data presented in this research is based on the performance results generated from the power and energy output measurements from each batch. These batch results are calculated from testing a sample of forty units for power and energy, respectively.

Four parameters are used to assess power and one parameter is used to assess energy. Therefore, five parameters are used to evaluate the performance of a batch.

A Multivariate process control method will be presented for Phase I and Phase II of the process. This method will be used as an alternative to the traditional univariate method of assessing five individual control charts for the means.

The Phase I data presented was generated from the "Production" phase of the process. After it was determined that this phase had been characterised, Phase II data collection began. This involved ramping up production, by increasing the number of batches manufactured in a 24 hour period.

The data used for Phase II monitoring contains seventeen new batch observations. They were sent for final testing once all samples were gathered from the Stage 3 audit station.

All batch samples that are for discharge testing must be started at the same time. As the batch results from each test become available, they are exported to a database for analysis.

6.2 Multivariate Analysis

This chapter will investigate various multivariate SPC methods through charting and model building. It will identify which is a more suitable method for fuel cell manufacturing. It will also determine which method identifies an out of control situation the fastest.

The objectives of this chapter are to,

- Develop a multivariate control chart using both parameters (power and energy) as inputs. Charts that will be covered include,
  - The T^2 Statistic using individual observations,
  - The T^2 Statistic on batch observations
- Develop a model using power and energy parameters by the use of,
  - Principal Components Analysis
  - T^2 and DModX control charts on PCA
- Identify the criteria to determine the most appropriate multivariate control chart/method for this process.

6.3 Multivariate Control Chart

6.3.1 T^2 Statistic using Individual Observations

Each batch result is treated as an individual observation. In Phase I, the T^2 Statistic and UCL for individual observations, are calculated using Equations 1.11 and 1.16 respectively.

In Phase II, the UCL for new individual observations is calculated using Equation 1.17.

The targets for the *start up* power, *working* power and energy for each batch are given in Table 6.1.

P5min (W)	P2h (W)	P6h (W)	P12h (W)	Energy (Wh)
0.85	1.00	1.00	1.00	17.00

Table 6.1 - Batch Targets

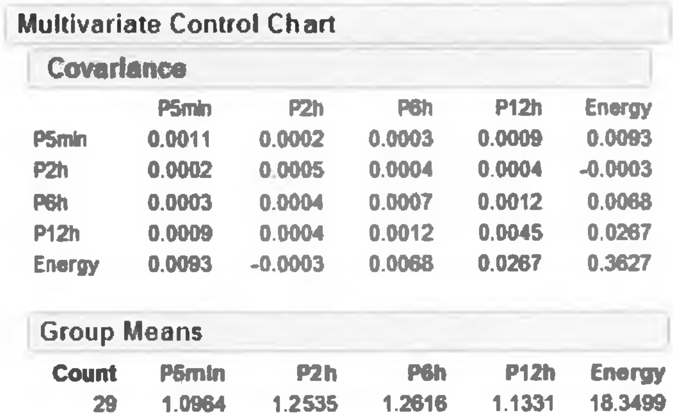
6.3.1.1 *Multivariate Phase I – Screening Data for Outliers*

Data generated from the power and energy performance testing, as identified in Chapter 5, will be assessed. Each power timepoint, (5 minutes, 2 hours, 6 hours, 12 hours) is measured in Watts (W) and the energy is measured in Watt hours (Wh). Twenty nine batch results for Phase I, are given in Table 6.2.

Batch	P5min	P2h	P6h	P12h	Energy
1	1.09	1.23	1.27	1.19	18.71
2	1.11	1.22	1.25	1.21	18.69
3	1.10	1.26	1.26	1.13	18.08
4	1.14	1.26	1.25	1.06	18.12
5	1.11	1.25	1.24	1.06	18.48
6	1.11	1.25	1.23	1.08	17.73
7	1.11	1.23	1.27	1.08	18.69
8	1.09	1.27	1.28	1.12	18.08
9	1.13	1.25	1.26	1.11	18.71
10	1.11	1.24	1.24	1.07	18.49
11	1.03	1.22	1.23	1.11	18.83
12	1.08	1.22	1.24	1.07	17.79
13	1.07	1.25	1.25	1.10	18.14
14	1.11	1.23	1.26	1.24	18.85
15	1.11	1.28	1.29	1.17	18.57
16	1.10	1.24	1.26	1.18	18.72
17	1.11	1.27	1.29	1.20	18.34
18	1.12	1.26	1.29	1.20	18.45
19	1.12	1.28	1.29	1.19	18.85
20	1.14	1.29	1.29	1.18	18.43
21	1.07	1.30	1.33	1.18	18.65
22	1.09	1.25	1.23	1.00	17.61
23	1.07	1.25	1.21	1.04	18.00
24	1.03	1.26	1.25	1.09	17.58
25	1.00	1.24	1.24	1.02	16.49
26	1.09	1.22	1.24	1.16	18.56
27	1.12	1.26	1.29	1.20	19.29
28	1.09	1.26	1.28	1.23	19.58
29	1.13	1.29	1.27	1.19	17.63

Table 6.2 – Batch Data for Phase I

JMP 7.0 software, developed by SAS Institute, is used to plot the multivariate control charts. Applying the multivariate control chart platform in JMP gives the following results shown in Figure 6.1.



Multivariate Control Chart

Covariance

	P5min	P2h	P6h	P12h	Energy
P5min	0.0011	0.0002	0.0003	0.0009	0.0093
P2h	0.0002	0.0005	0.0004	0.0004	-0.0003
P6h	0.0003	0.0004	0.0007	0.0012	0.0068
P12h	0.0009	0.0004	0.0012	0.0045	0.0267
Energy	0.0093	-0.0003	0.0068	0.0267	0.3627

Group Means

Count	P5min	P2h	P6h	P12h	Energy
29	1.0964	1.2535	1.2616	1.1331	18.3499

Figure 6.1 – Phase I Hotellings T^2 Output

The parameter means and variance from Figure 6.1 are consistent with manual calculations given by,

$$\bar{X} = (1.096, 1.254, 1.262, 1.133, 18.350)$$

and the sample covariance matrix, S,

$$\begin{array}{ccccc}
0.0011 & 0.0002 & 0.0003 & 0.0009 & 0.0093 \\
0.0002 & 0.0005 & 0.0004 & 0.0004 & -0.0003 \\
0.0003 & 0.0004 & 0.0007 & 0.0012 & 0.0068 \\
0.0009 & 0.0004 & 0.0012 & 0.0045 & 0.0267 \\
0.0093 & -0.0003 & 0.0068 & 0.0267 & 0.3627
\end{array}$$

Using Equation 1.11, the T^2 Statistics are calculated,

$$T^2 = (X - \bar{X})' S^{-1} (X - \bar{X})$$

The T^2 values for each batch are contained in Table 6.3.

Batch	T^2
1	4.667
2	5.301
3	0.431
4	4.507
5	3.412
6	2.960
7	8.358
8	2.521
9	2.898
10	2.446
11	9.344
12	4.157
13	0.737
14	4.749
15	2.275
16	0.939
17	2.194
18	2.355
19	2.288
20	4.510
21	12.033
22	4.612
23	8.307
24	5.982
25	13.773
26	2.813
27	2.681
28	7.989
29	10.760

Table 6.3 – T^2 Statistics for each Batch

The Phase I UCL is calculated from Equation 1.16, where $n = 29,\ p = 5$

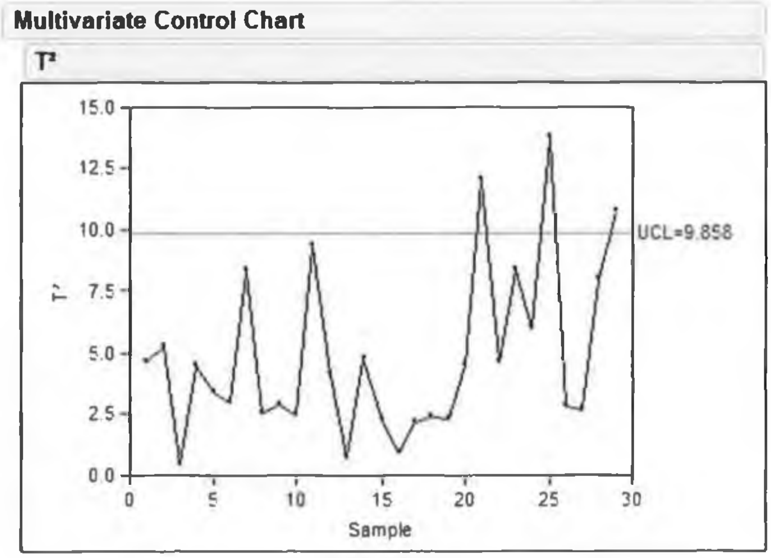
$$T^2{}_{UCL} = \frac{(n-1)^2}{n}\ \beta_{(\alpha,p/2,(n-p-1)/2)}\ ,$$

$$T^2{}_{UCL} = \frac{(29-1)^2}{29}\ \beta_{(0.05,5/2,(29-5-1)/2)}$$

$$= 27.034\beta_{(0.05,2.5,11.5)}$$

$$= 27.034(0.3646) = 9.86$$

The T^2 values from Table 6.3 and the UCL are plotted on a multivariate chart illustrated in Figure 6.2.



Note: UCL is calculated based on Alpha=0.05

Figure 6.2 – Hotellings T^2 Chart

The out of control points identified in Figure 6.2 are measured against the individual observations in Table 6.2. Batch 21, 25 and 29 are assessed against their group means in Figure 6.1. It can be seen that Batch 21 is performing well above the batch average for working power output (2h, 6h and 12h). Batch 25 has a below average energy output. These output values are considered abnormal events. Batch 29, however, does not seem to have any extreme case for either power or energy.

Failure analysis on Batch 25 concluded that defective materials were present on some of the units, resulting in the overall low performance of the batch. However, not all of the units from this batch were defective.

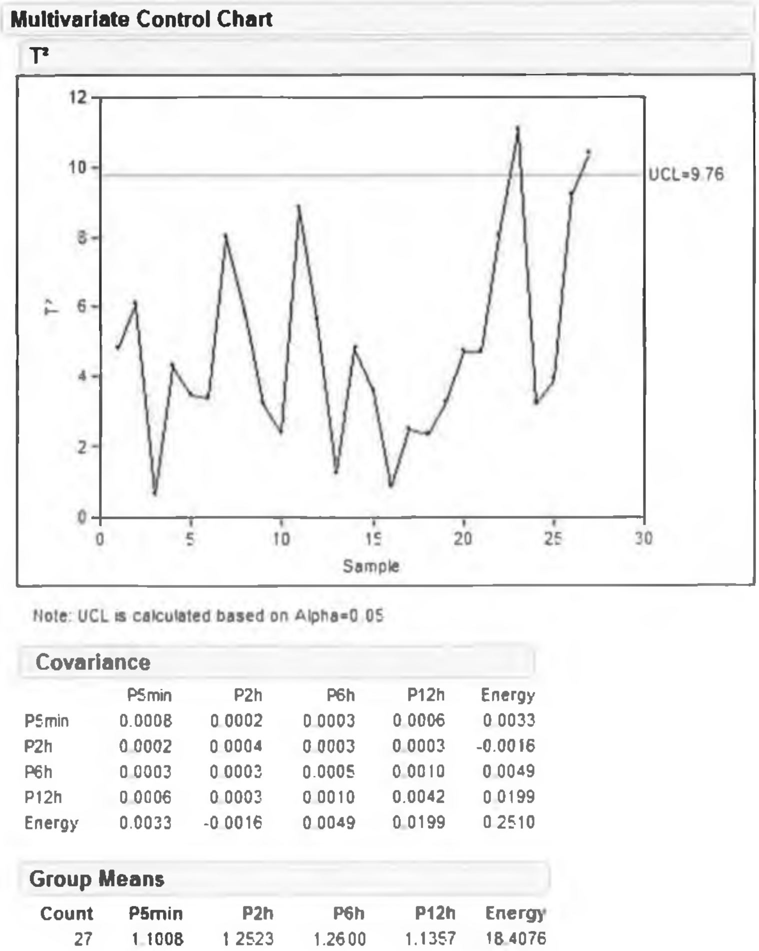
Batch 21 is an atypical batch as its results are high. Failure analysis concluded that some units from this batch were assembled with high performing components. These high performing components are not representative of a normal fuel cell composition. To include this batch in Phase I, would inflate the mean and sample covariance matrix, so this batch was removed for the purposes of characterising a typical baseline for Phase I.

Batches 21 and 25 will be removed from the baseline dataset for Phase II monitoring as it was determined that these batches are not representative of normal batch operation.

Batch 29 will not be removed, as there was no obvious root cause for the high T^2 Statistic.

After removing Batch 21 and 25, a second screening of the data ($n = 27$) is then performed. This reveals an additional batch above the new UCL. This is shown in Figure 6.3. The new UCL was recalculated using $n = 27$.

Batch 24 (sample number 23 in Figure 6.3) has a T^2 Statistic of 11.03, which is above the UCL of 9.76. Upon examination of the raw data, one notices that this batch has the lowest energy output of 17.58Wh. This is below the batch average of 18.41Wh, however it is above the target energy, as listed in Table 6.1. The *start up* power of 1.03W is also below the average, but is well above the 0.85W requirements for start up. It was decided to retain Batch 24 as part of the baseline dataset as there were no abnormal results observed.



Multivariate Control Chart

Note: UCL is calculated based on Alpha=0.05

Covariance

	P5min	P2h	P6h	P12h	Energy
P5min	0.0008	0.0002	0.0003	0.0006	0.0033
P2h	0.0002	0.0004	0.0003	0.0003	-0.0016
P6h	0.0003	0.0003	0.0005	0.0010	0.0049
P12h	0.0006	0.0003	0.0010	0.0042	0.0199
Energy	0.0033	-0.0016	0.0049	0.0199	0.2510

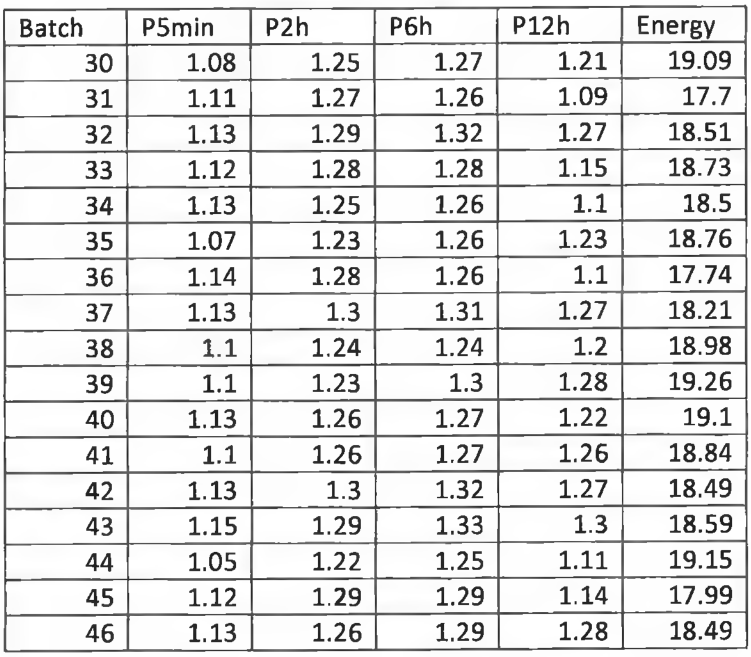
Group Means

Count	P5min	P2h	P6h	P12h	Energy
27	1.1008	1.2523	1.2600	1.1357	18.4076

Figure 6.3 – Second Screening of Phase I Data

6.3.1.2 *Multivariate Phase II – Monitoring New Observations*

Twenty seven batches were used to create the historical baseline for use in Phase II monitoring of the process. Seventeen new batches were manufactured and their performance results are given in Table 6.4.



Batch	P5min	P2h	P6h	P12h	Energy
30	1.08	1.25	1.27	1.21	19.09
31	1.11	1.27	1.26	1.09	17.7
32	1.13	1.29	1.32	1.27	18.51
33	1.12	1.28	1.28	1.15	18.73
34	1.13	1.25	1.26	1.1	18.5
35	1.07	1.23	1.26	1.23	18.76
36	1.14	1.28	1.26	1.1	17.74
37	1.13	1.3	1.31	1.27	18.21
38	1.1	1.24	1.24	1.2	18.98
39	1.1	1.23	1.3	1.28	19.26
40	1.13	1.26	1.27	1.22	19.1
41	1.1	1.26	1.27	1.26	18.84
42	1.13	1.3	1.32	1.27	18.49
43	1.15	1.29	1.33	1.3	18.59
44	1.05	1.22	1.25	1.11	19.15
45	1.12	1.29	1.29	1.14	17.99
46	1.13	1.26	1.29	1.28	18.49

Table 6.4 – New Batch Data

These seventeen additional batches can be plotted on their own T^2 chart to determine if they are within the Phase II UCL. This UCL will be calculated using the statistics from Phase I as described previously in Chapter 2.

The group mean and covariance are determined from Phase I, as shown in Figure 6.3.

Equation 1.11 is used to calculate the T^2 Statistics for the new batches,

$$T^2 = (X_{new} - \bar{X})' S^{-1} (X_{new} - \bar{X})$$

The T^2 values for these seventeen new batches are presented in Table 6.5.

Sample	Batch	T^2
1	30	3.884
2	31	2.675
3	32	8.972
4	33	3.241
5	34	2.442
6	35	5.841
7	36	5.413
8	37	10.260
9	38	8.367
10	39	10.933
11	40	4.830
12	41	6.551
13	42	9.438
14	43	12.410
15	44	9.095
16	45	4.681
17	46	8.115

Table 6.5 – T^2 Statistics for new batches

The Phase II UCL is calculated from Equation 1.17, where $n = 27,\ p = 5$.

$$T^2{}_{UCL} = \frac{(n+1)(n-1)p}{n(n-p)} \mathrm{F}_{(\alpha,\mathrm{p},\mathrm{n}-\mathrm{p}),}$$

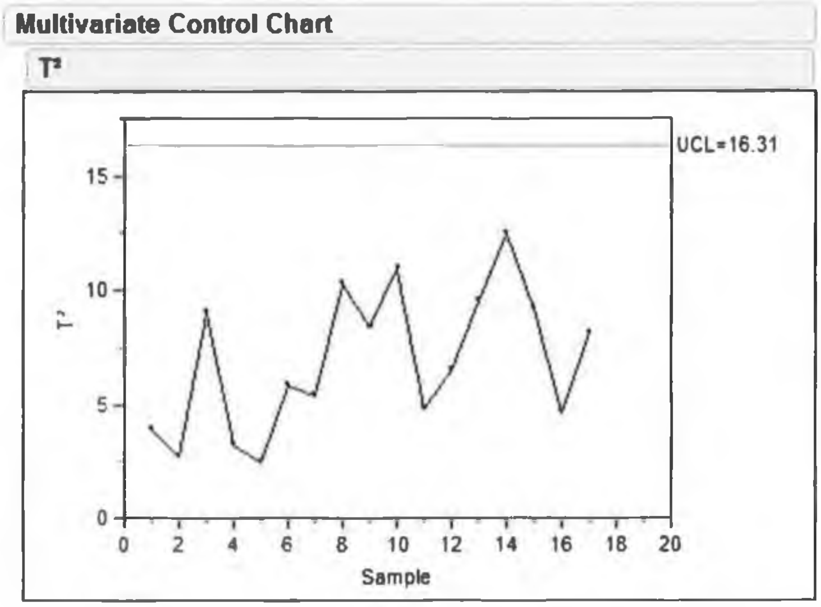
$$= \frac{(27+1)(27-1)5}{27(27-5)} \mathrm{F}_{(0.05,5,27-5),}$$

$$= \frac{(28)(26)5}{27(22)} \mathrm{F}_{(0.05,5,22),}$$

$$= \frac{3640}{594} \mathrm{F}_{(0.05,5,22),}$$

$$= 6.13(2.661) = 16.31$$

The Multivariate control chart platform in JMP produces the results shown in Figure 6.4.



Note: UCL is calculated based on Alpha=0.05

Figure 6.4 – Phase II Hotellings T^2 Output

The multivariate control chart in Figure 6.4, shows that the process is now performing satisfactorily. The T^2 control chart will be monitored for all new observations.

Another application of the multivariate control chart is using the T^2 Statistic on batch observations. This is discussed in the next section.

6.3.2 T^2 Statistic on Batch Observations

The Phase I process for plotting the T^2 Control Chart using the batch data in Table 6.2, is assessed exactly as discussed in Section 6.3.1.

Treating a batch using the equations for a Category 1 with Batch Size = 1, as described by Mason *et al.* (2001), will give the same result as treating data as an individual observation.

Recall from Section 4.2, $n_i = 1$, where $i = 1, \dots, r$ and the the total sample size is $N = \sum_{i=1}^{r} n_i = r$.

The Phase I T^2 Statistic and UCL are calculated from Equations 4.1 and 4.2 respectively, using N = 27. The results for T^2 values and UCL are identical to those presented in Table 6.3 and Figures 6.1 and 6.2 as the equations are equivalent. Batch 21 and 25 are removed using the same reasons for exclusion when treated as individual observations.

The remaining twenty seven batches are used to calculate the mean and covariance matrix applied in Phase II.

The Phase II process uses a baseline size $= nr$, where r is the number of batches as described in section 4.2. The Phase II T^2 Statistic, for new batches containing $m = 1$ observations, and UCL are calculated from Equations 4.3 and 4.4 respectively, where $N = 27$. The results are identical to those shown in Table 6.5 and Figure 6.4.

The T^2 control charts will not give any additional information for an out of control situation. Other methods can be utilised to convert the out of control point back in terms of the original observations.

The next section uses an alternative method, using Principal Components Analysis. This should help to determine, which out of all these methods, is the more appropriate monitoring method for interpreting an out of control point.

6.4 Principal Components Analysis

Principal Component Analysis will be generated using the same raw data as described in Section 6.1. The Phase I process of using a T^2 control chart to screen the data for outliers, is used prior to principal components analysis. Section 6.3.1.1 has already provided the results for this screening.

6.4.1 Develop PCA Model

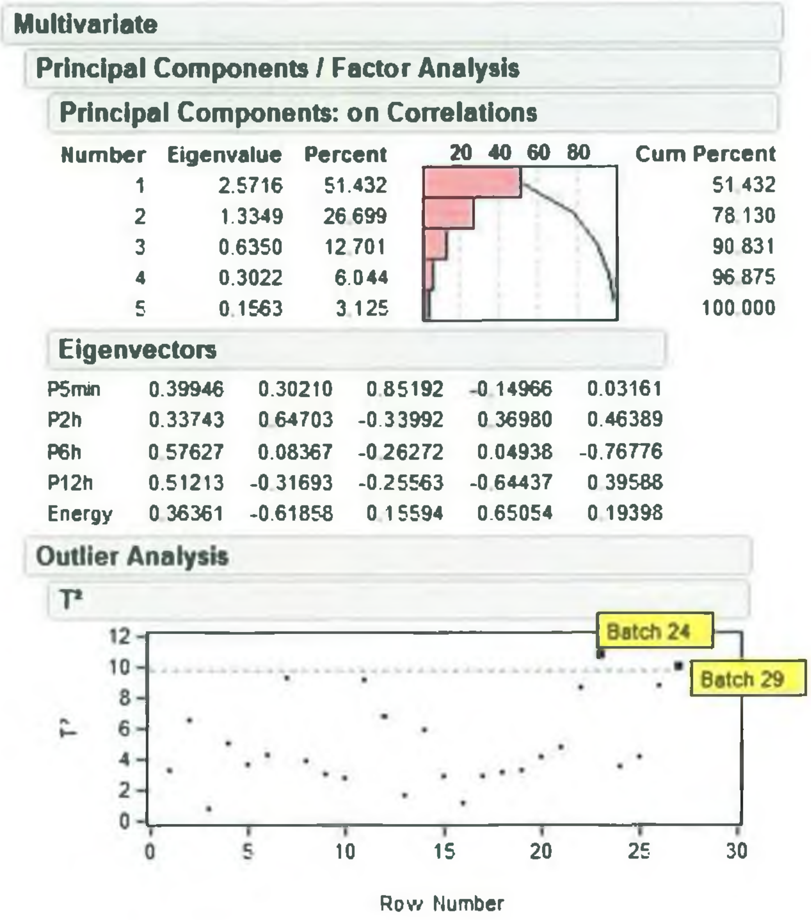
The initial dataset had twenty nine batches. Batches 21 and 25 were removed as part of Phase I, shown by the red arrows in Table 6.6. There are twenty seven batches remaining, for creating the Principal Components Model, shown in Table 6.6.

Batch	P5min	P2h	P6h	P12h	Energy
1	1.09	1.23	1.27	1.19	18.71
2	1.11	1.22	1.25	1.21	18.69
3	1.10	1.26	1.26	1.13	18.08
4	1.14	1.26	1.25	1.06	18.12
5	1.11	1.25	1.24	1.06	18.48
6	1.11	1.25	1.23	1.08	17.73
7	1.11	1.23	1.27	1.08	18.69
8	1.09	1.27	1.28	1.12	18.08
9	1.13	1.25	1.26	1.11	18.71
10	1.11	1.24	1.24	1.07	18.49
11	1.03	1.22	1.23	1.11	18.83
12	1.08	1.22	1.24	1.07	17.79
13	1.07	1.25	1.25	1.10	18.14
14	1.11	1.23	1.26	1.24	18.85
15	1.11	1.28	1.29	1.17	18.57
16	1.10	1.24	1.26	1.18	18.72
17	1.11	1.27	1.29	1.20	18.34
18	1.12	1.26	1.29	1.20	18.45
19	1.12	1.28	1.29	1.19	18.85
20	1.14	1.29	1.29	1.18	18.43
22	1.09	1.25	1.23	1.00	17.61
23	1.07	1.25	1.21	1.04	18.00
24	1.03	1.26	1.25	1.09	17.58
26	1.09	1.22	1.24	1.16	18.56
27	1.12	1.26	1.29	1.2	19.29
28	1.09	1.26	1.28	1.23	19.58
29	1.13	1.29	1.27	1.19	17.63

Table 6.6 – Batch Data for Creating PCA Model, $n = 27$

The correlation matrix is used to calculate the principal components. JMP generates the output, given in Figure 6.5.

One notices in Figure 6.5, that the outlier analysis is calculated using a T^2 control chart. Two out of the twenty seven batches are identified as outliers. These are Batch 24 and Batch 29. Recall that these batches were also identified in Phase I screening of outliers, in Section 6.3.1.1, but remained as part of the baseline dataset as no actual root cause was determined for the out of control situations.



Multivariate

Principal Components / Factor Analysis

Principal Components: on Correlations

Number	Eigenvalue	Percent	Cum Percent
1	2.5716	51.432	51.432
2	1.3349	26.699	78.130
3	0.6350	12.701	90.831
4	0.3022	6.044	96.875
5	0.1563	3.125	100.000

Eigenvectors

P5min	0.39946	0.30210	0.85192	-0.14966	0.03161
P2h	0.33743	0.64703	-0.33992	0.36980	0.46389
P6h	0.57627	0.08367	-0.26272	0.04938	-0.76776
P12h	0.51213	-0.31693	-0.25563	-0.64437	0.39588
Energy	0.36361	-0.61858	0.15594	0.65054	0.19398

Outlier Analysis

T²

Figure 6.5 – PCA Output for $n = 27$ Batches

The number of principal components to retain is based on the rationale that the selected number of PC's should include enough of the components to explain 80-90% of the total variability in the data. The cumulative percent of the first two components accounts for 78% of the total variation. The first three components accounts for 90% of the total variation, therefore three principal components will be retained.

The linear equation for the first principal component is,

$$0.399 * P5\min + 0.337 * P2h + 0.576 * P6h + 0.512 * P12h + 0.3636 * Energy$$

The linear equation for the second principal components is,

$$0.302 * P5min + 0.647 * P2h + 0.083 * P6h - 0.317 * P12h - 0.619 * Energy$$

The linear equation for third principal component is,

$$0.852 * P5min - 0.340 * P2h - 0.263 * P6h - 0.256 * P12h + 0.156 * Energy$$

The PC scores are calculated using these linear equations and the results are shown in Table 6.7.

Batch	PC1 Score	PC2 Score	PC3 Score
1	0.41	-1.38	-0.21
2	0.18	-1.61	0.72
3	-0.14	0.69	-0.23
4	-0.33	1.39	1.41
5	-0.92	0.27	0.87
6	-1.55	1.06	0.67
7	-0.18	-0.60	0.84
8	0.30	1.01	-0.90
9	0.44	0.03	1.14
10	-0.99	-0.11	0.99
11	-2.17	-2.26	-1.11
12	-2.26	-0.19	0.17
13	-1.18	0.09	-0.75
14	0.94	-1.61	0.37
15	1.75	0.73	-0.60
16	0.39	-0.96	0.09
17	1.66	0.56	-0.62
18	1.72	0.23	-0.11
19	2.26	0.40	-0.28
20	2.33	1.50	0.09
22	-2.56	1.37	0.32
23	-2.76	0.40	-0.11
24	-2.09	0.69	-2.30
26	-0.85	-1.47	0.37
27	2.33	-0.81	0.15
28	2.09	-1.68	-0.70
29	1.18	2.25	-0.28

Table 6.7 – Principal Component Scores

The PC scores are standardised by dividing each PC score by the square root of its eigenvalue, resulting in the standardised scores shown in Table 6.8.

These standardised PC scores should follow a standard normal distribution $N(0,1)$. They can be plotted on a control chart with mean = 0 and control limits = +/-3, as shown in Figures 6.6, 6.7 and 6.8.

Batch	Standardised PC1	Standardised PC2	Standardised PC3
1	0.25	-1.19	-0.26
2	0.11	-1.40	0.90
3	-0.09	0.60	-0.29
4	-0.20	1.20	1.77
5	-0.57	0.23	1.09
6	-0.97	0.92	0.84
7	-0.11	-0.52	1.05
8	0.19	0.88	-1.13
9	0.27	0.03	1.43
10	-0.62	-0.09	1.25
11	-1.35	-1.95	-1.39
12	-1.41	-0.17	0.21
13	-0.74	0.07	-0.95
14	0.59	-1.39	0.47
15	1.09	0.63	-0.75
16	0.25	-0.83	0.12
17	1.03	0.49	-0.78
18	1.07	0.20	-0.14
19	1.41	0.35	-0.35
20	1.45	1.30	0.11
22	-1.60	1.19	0.41
23	-1.72	0.35	-0.13
24	-1.30	0.60	-2.88
26	-0.53	-1.27	0.46
27	1.45	-0.70	0.18
28	1.30	-1.46	-0.88
29	0.73	1.95	-0.35

Table 6.8 – Standardised PC Scores

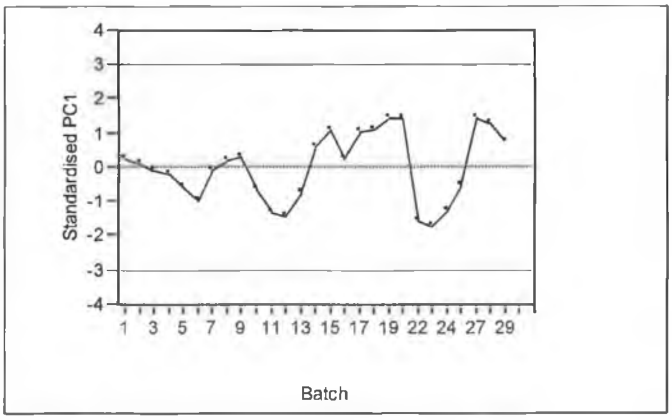


Figure 6.6 – Standardised PC1 Score Control Chart

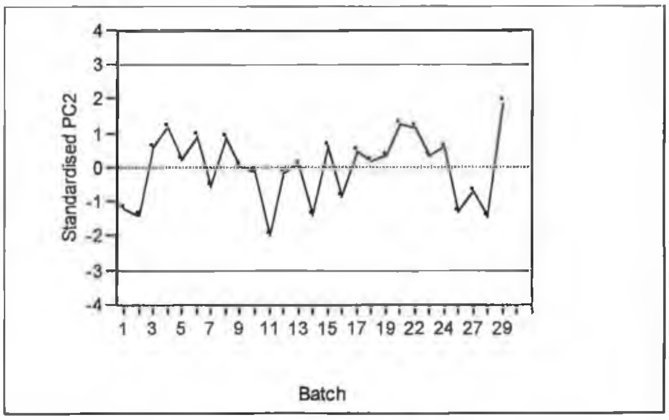


Figure 6.7 – Standardised PC2 Score Control Chart

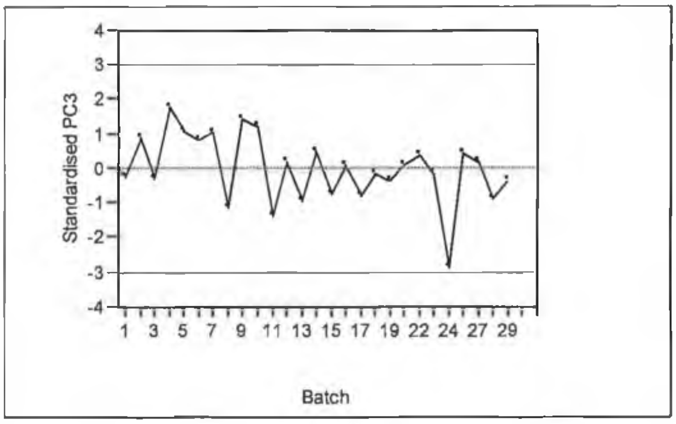


Figure 6.8 – Standardised PC3 Score Control Chart

Scatterplots for the standardised PC scores can also be very informative. These are shown in Figures 6.9, 6.10 and 6.11.

The PC3 score for Batch 24 is very close to the lower control limit, as seen in the score control chart in Figure 6.8, and the scatterplots in Figures 6.10 and 6.11. This is one of the batches that was identified as a possible outlier in Phase I screening, but it was not removed from the baseline dataset as no root cause was determined to warrant removal.

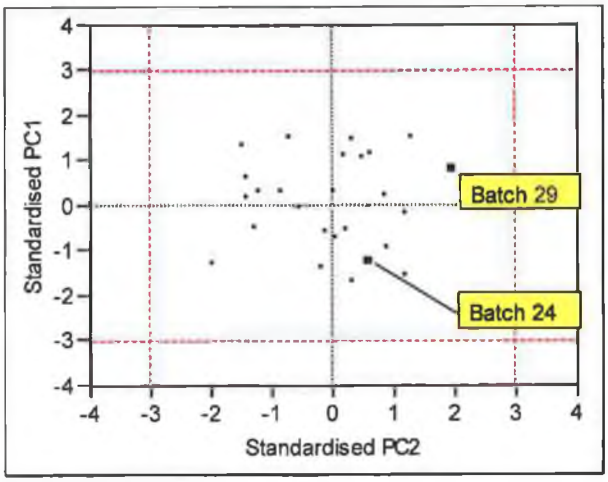


Figure 6.9 – Scatterplot of Standardised PC1 and PC2

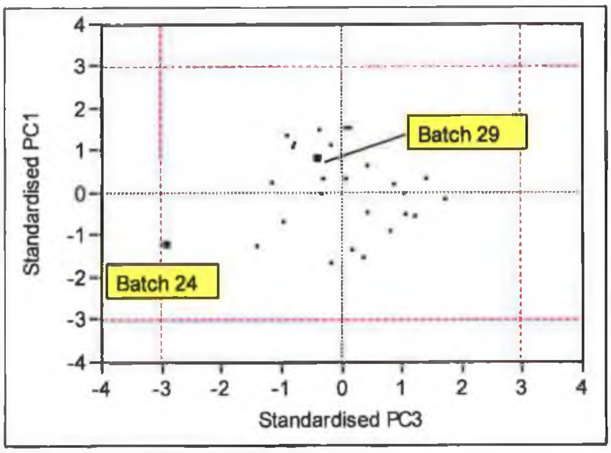


Figure 6.10 – Scatterplot of Standardised PC1 and PC3

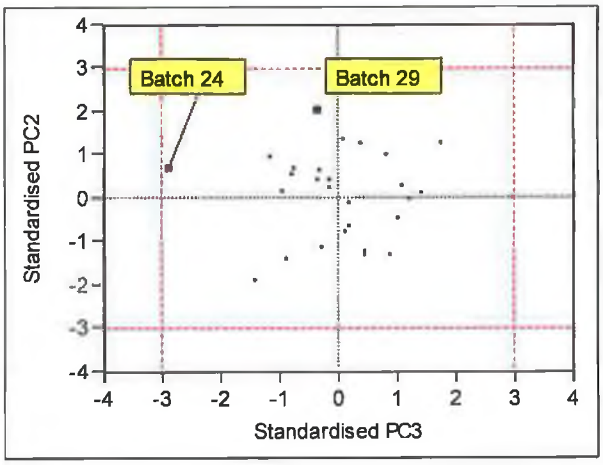


Figure 6.11 – Scatterplot of Standardised PC2 and PC3

The other batch that was identified as a possible outlier was Batch 29. From the score control chart in Figure 6.7 and the scatterplots in Figures 6.9 and 6.11, it seems its PC2 has a heavier weighting towards the UCL, but this is not significant.

6.4.1.1 *T^2 Chart on PCA*

SIMCA P+ plots a T^2 control chart on the principal component scores in addition to scores and loadings plots. The T^2 chart can help to determine if a good model has been developed by identifying if there are any outliers in the model. The T^2 values on Principal Components are calculated using the formula described in Section 3.5,

$$T_i^2 = \sum_{i=1}^{k} \frac{{t_i}^2}{s_{t\ i}^2}$$

where $s_{t\ i}^2 = \lambda_i$, is the estimated variance of t_i, the principal component score.

The T^2 values for the PCA model are shown in Figure 6.12. Batch 24 is showing to be outside the 95% confidence region of the model.

The green line in Figure 6.12 shows the control limit for $\alpha = 0.05$, i.e. 95% confidence limit. The red line indicates the control limit for $\alpha = 0.01$, i.e. 99% confidence limit.

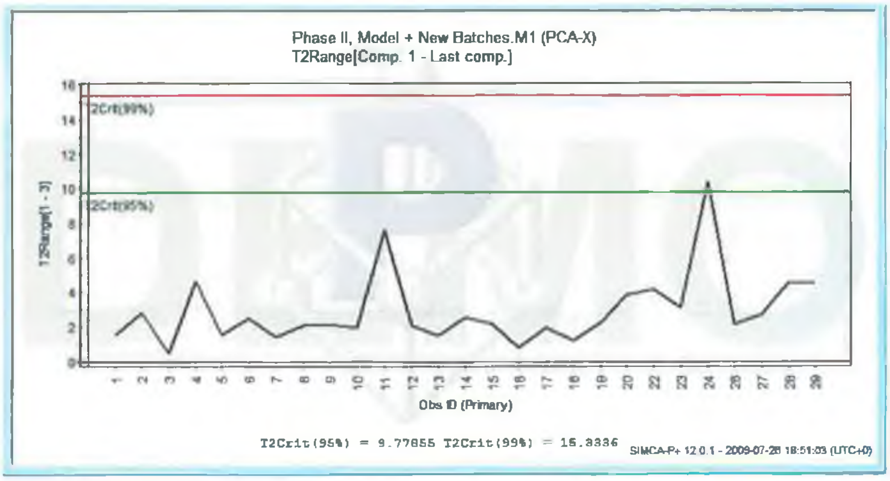


Figure 6.12 – T^2 on PCA Model

Perhaps a contribution plot can provide additional information to the cause of the elevated T^2 value.

A contribution plot can explain why a point on the T^2 chart has moved from the average. Figure 6.13 shows the weighted difference between the data in Batch 24 and the model average. The weights are derived from the loadings. In this case, the loadings are pl to p3. It identifies that the contributions to the average start up power (P5min) has moved from the model average.

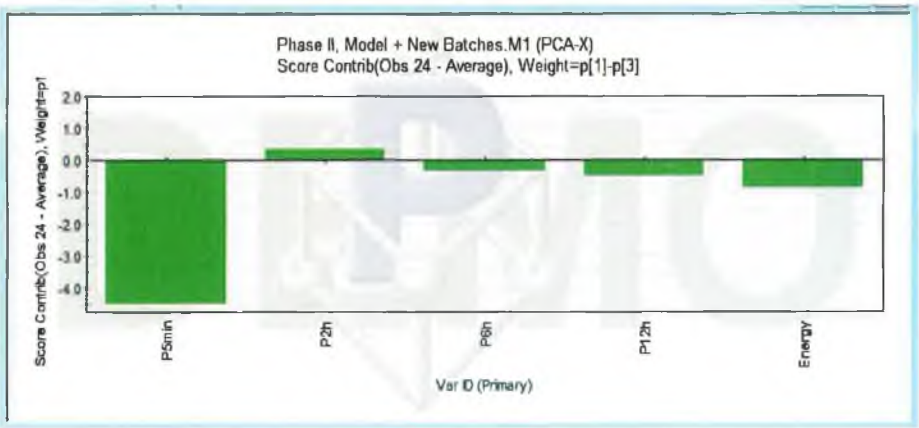


Figure 6.13 – Weighted Contribution plot for T^2, Batch 24

6.4.1.2 *DModX Chart on PCA*

The normalised DModX chart can be used, as described in Section 3.5.1. It can also give an indication if the model has any outliers, as it can identify a change in the correlation structure of the data. It is monitored in addition to the T^2 chart as recommended by Wikstrom *et al.* (1998). The combined charts are called SMART charts. The charts are generated for each principal component to enable the user to determine which principal component has contributed to the out of control situation.

Figures 6.14, 6.15 and 6.16 show the normalised DModX charts for each of the three retained components of the model using SIMCA P+. The contribution plot for a selected batch will display the scaled residuals of all the variables. These scaled residuals are multiplied by the absolute value of the weight parameter. The weights are the principal components.

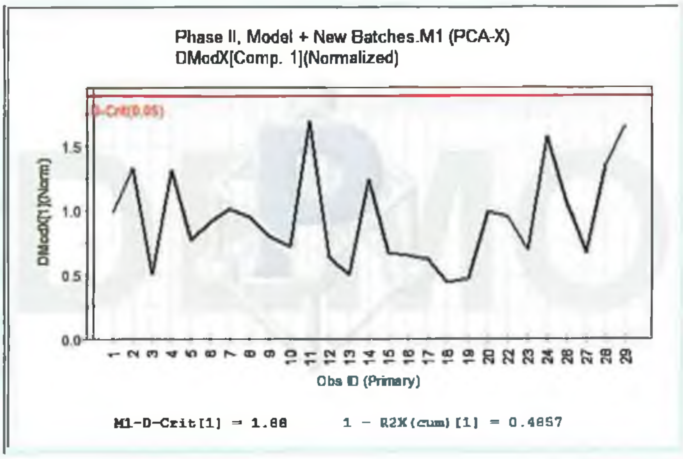


Figure 6.14 – Normalised DModX of PC1

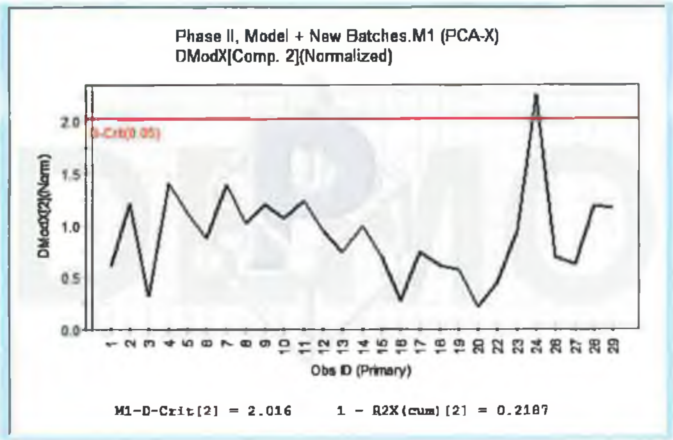


Figure 6.15– Normalised DModX of PC2

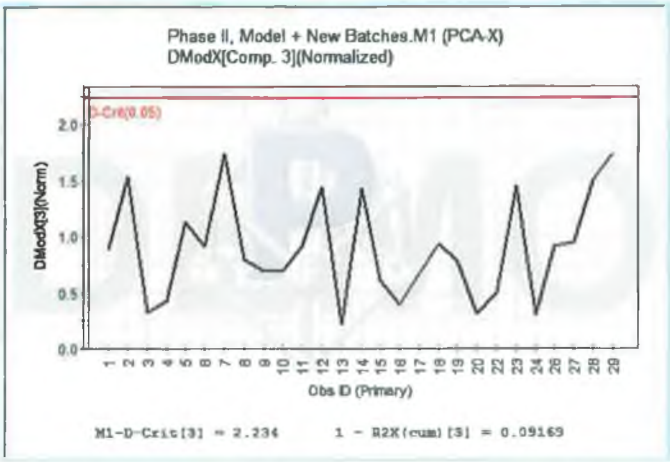


Figure 6.16 – Normalised DModX of PC3

Figure 6.18 identifies that Batch 24 has contributed to an out of control situation for PC2. A contribution plot should help determine which variable(s) are causing the out of control. From Figure 6.15, it seems that the start up power (P5min) is again flagging a signal in the DModX chart.

Batch 24 has the lowest start up power (1.03W) out of all the batches in the model. The model average is 1.10W with a standard deviation of 0.03W. This is not significant for the product performance, as the target for start up power is 0.85W.

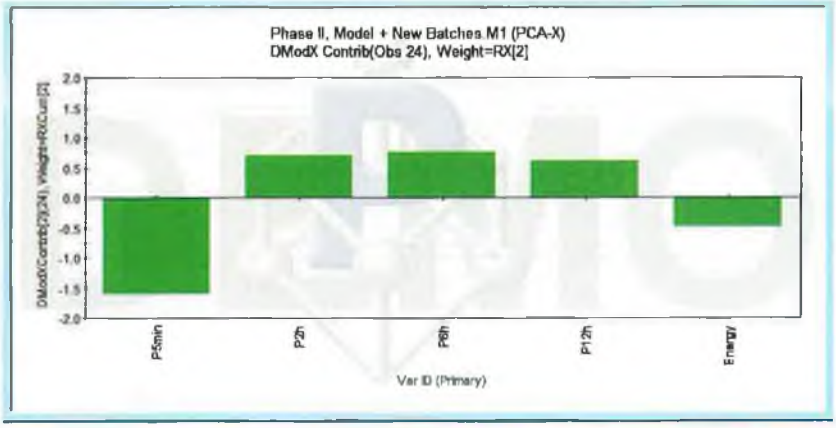


Figure 6.17 – Contribution Plot for DModX, Batch 24, PC2

Figure 6.18 shows a scatterplot of DModX values for PC2 along with its control limit on the *y*-axis, the *x*-axis contains the observation data for start up power (P5min) for all batches is shown. This plot illustrates where Batch 24 fits in the model, in relation to all the other batches in the model.

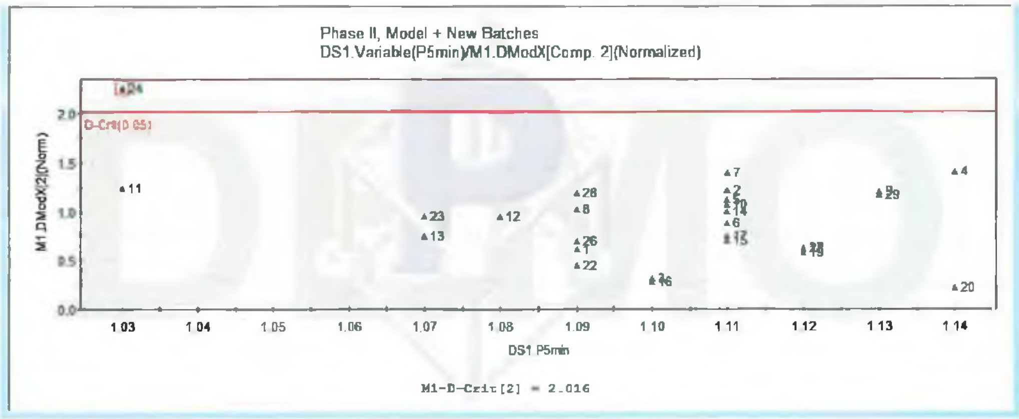


Figure 6.18 – Scatterplot of DModX and Start up Power

There is nothing wrong with the process even though Batch 24 is indicating an out of control. This could be due to the selection of an α level of 0.05, which will naturally indicate an out of control for 1 in 20 samples.

One should look carefully at the chosen level of alpha (α). Figures 6.3, 6.4 and 6.5 show the T^2 charts for $\alpha = 0.05$. This alpha level will result in 1 out of 20 points giving an out of control signal by falling outside the control limits even when the process is in control. Batch 24 indicated an out of control but upon investigation, no root cause was determined.

Figures 6.6, 6.7 and 6.8 show the PC score charts. These charts are standardised and so follow a normal distribution with centre 0 and ±3 standard deviations. The α level for these charts are set at 0.0027, as three standard deviations of a normal distribution falls within 99.73% of the distribution. Batch 24 did not signal an out of control on these charts.

In the T^2 chart on PCA in Figure 6.12, Batch 24 was just outside the green 95% control limits ($\alpha = 0.05$). However, if one was to monitor the process using the red 99% control limits ($\alpha = 0.01$), Batch 24 would not have signalled an out of control. Wikstrom *et al.* (1998) pointed that 95% limits can be used as warning limits and 99% limits as the action limits.

6.5 Comparison of Multivariate Methods

A multivariate method is determined to be the most suitable for this process if the following criteria are satisfied.

- Create a representative baseline dataset,
- Generate relevant control limits for future use,
- The ability to detect an abnormal situation in future observations,
- Easily generate alternative charts for investigative analysis, if required

- The ability to assist in determining what contributed to an abnormal situation,
- Continue to monitor new observations with minor effort.

All the methods described in this chapter have their uses. The T^2 control chart can generate a control limit for characterising a process in Phase I. This information is then used to generate control limits for Phase II of a process. It can also deal with both individual and batch observations. It has the ability to create the different control charts depending on the process. The T^2 control charts are quick to determine an out of control signal, however, it is not easy to interpret how the out of control signal relates to the original variables. It does not have the ability identify a variable or group of variables that could have contributed to the signal.

A baseline model is developed using Principal Components Analysis. This model is used to generate control charts that detect if a future observation does not conform to the model. T^2 and DModX control charts are generated. Should an out of control signal occur, the PC linear equations, loadings and scores, are a good way to investigate how much influence each variable has on the out of control signal. Contribution plots are also a helpful tool to assist one in determining the questionable variables associated with an out of control signal.

Control charts based on PCA is the most appropriate method for detecting, monitoring and interpreting an out of control signal. These control charts have an advantage over the regular T^2 control charts. PCA control charts have the ability to identify the contributions and give more information for the out of control signal. They satisfy all the criteria for the most suitable method as outlined above.

6.5.1 Phase II using Principal Components Analysis

Phase II for Principal Components Analysis was ran on the seventeen new batch observations, as the T^2 and DModX charts on PCA were determined to be the best method for monitoring and interpreting the process data. T^2 and DModX results are shown in Figures 6.19 – 6.22.

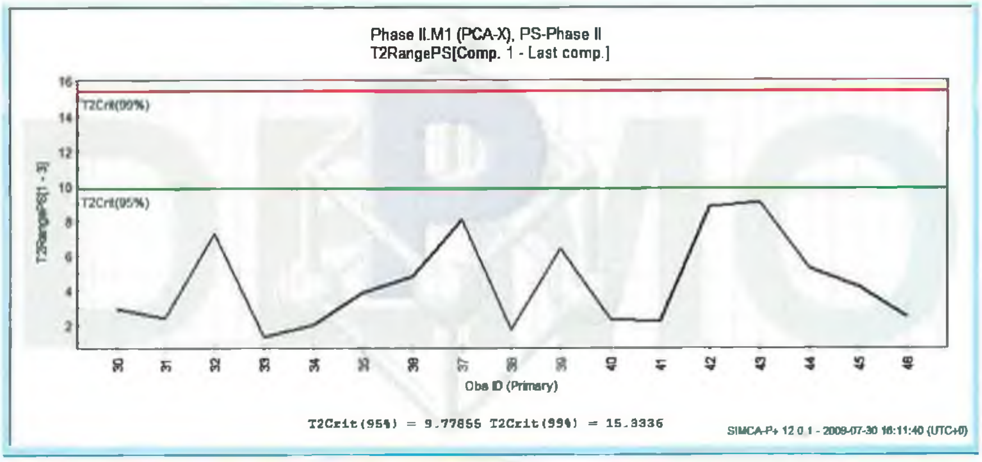


Figure 6.19 – T^2 Control Chart for New Batches

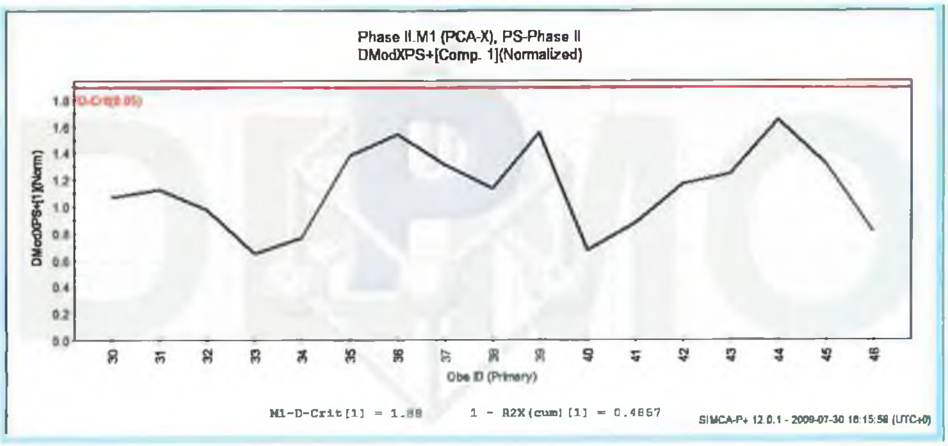


Figure 6.20 – DModX Control Chart on PC1

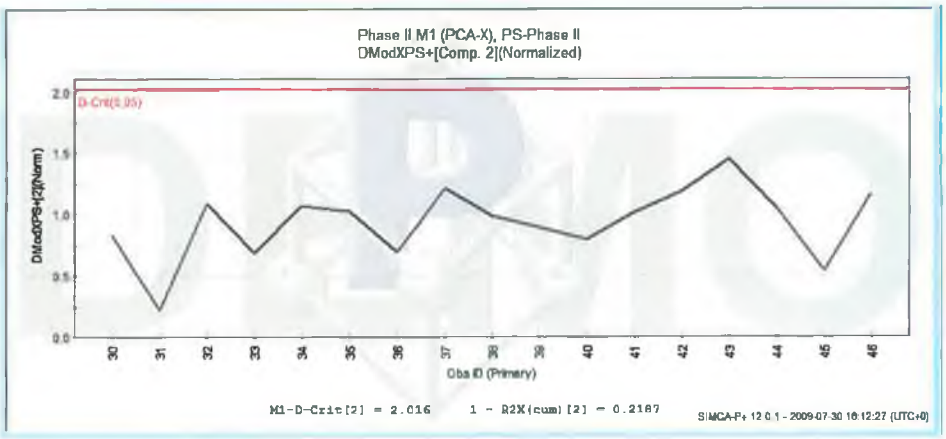


Figure 6.21 – DModX Control Chart on PC2

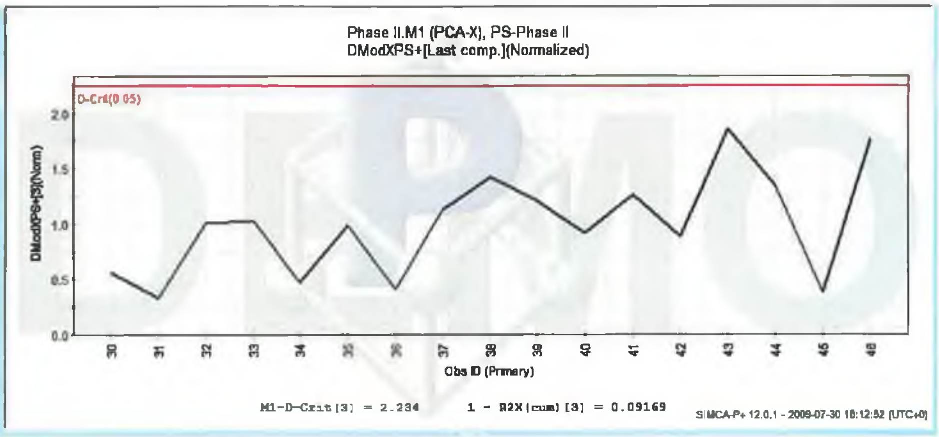


Figure 6.22 – DModX Control Chart on PC3

All new batches are in control therefore there is nothing abnormal in the process. These charts will be used to monitor future observations.

6.6 Conclusion

An objective of this research is to identify suitable methods to be used for monitoring and controlling an automated high volume process. All multivariate methods discussed using T^2 control charts and control charts on PCA satisfy this purpose. The latter, has more advanced features that provide valuable information on the current process. It also identifies how future observations compare to the in-control model and it can assist in interpreting how these future observations do not fit the model.

6.7 References

Mason, R.L., Chou, Y-M, Young, J.C. (2001), "Applying Hotellings T^2 Statistic to Batch Processes". *Journal of Quality Technology*, Vol. 33, pp. 466-479.

Wikstrom, C., Albano, C., Eriksson, L., Friden, H., Johansson, E., Nordahl, A., Ranner, S., Sandberg, M., Kettaneh-Wold, N., Wold, S. (1998), "Multivariate process and quality monitoring applied to an electrolysis process Part I. Process Supervision with multivariate control charts", *Chemometrics and Intelligent Laboratory Systems*, Vol. 42, pp. 221-231.

CHAPTER SEVEN

DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

7.1 Introduction

Modern manufacturing processes contain both online and offline process systems, which are capable of collecting information on hundreds of process variables. These variables are analysed to give reliable and useful information on the process. They must be selected wisely, so that they will provide relevant data to enable the user to monitor and control conditions contributing to the parameter results. Monitoring and controlling this data is performed through Statistical Process Control (SPC). SPC has been used in industry for many decades, through various control charting techniques. Univariate SPC techniques are more commonly used in industry, through the application of traditional Shewhart control charts. These control charts display how each point compares to its average, for a single process parameter. They provide upper and lower control limits, which signal an out of control observation should a point fall outside these limits. However, Multivariate SPC is becoming more popular due to the powerful capability of analysing more than one variable on a single control chart.

This research, through many literature reviews, will investigate the current status of SPC, in particular Multivariate SPC techniques. This should put forward consideration for suitable new or alternative SPC techniques. The most suitable techniques will be considered and applied to performance data generated from a high volume automated manufacturing line of a Fuel Cell. Product performance will be considered for the application of these methods, as this is one of the most important quality characteristics for the end user.

These multivariate techniques will be investigated to create a single monitoring system that will effectively reduce the amount of work involved in monitoring individual process parameters.

Investigative tools to assist in identifying parameters that could be a factor in the root cause for out of control failures, will also be assessed. Consequently, this will identify the most effective system that can be used to monitor and control future observations. This system should have the ability to be applied on a manufacturing line with increased volumes of production. It must also be capable of monitoring both continuous and batch processes.

7.2 Discussion

Xue *et al.* (2006) point out that some of the main challenges in fuel cell industry is the durability and reliability of the fuel cell. Under normal conditions fuel cells can even fail. This can be due to the membrane drying out, no electrochemical reaction because of fuel starvation and leaking of the membrane. A fuel cell system has varying conditions. They state that it would be very expensive and impractical to create a non-model baseline, as a huge amount of data would need to be collected from all these varying conditions.

Statistical Process Control monitoring is divided into two stages. Phase I involves screening the data to create a baseline from which the process parameters are estimated. Phase II monitoring and controlling future data observations. Both of these were considered as part of this research, in order to create an effective monitoring system.

The literature reviews conducted on these multivariate techniques were evaluated. These techniques included Hotellings T^2, Principal Components Analysis (PCA), Partial Least Squares (PLS), Squared Prediction Error (SPE) and Distance to the Model (DModX).

Hotellings T^2 and PCA techniques were applied to the fuel cell performance data. The Hotellings T^2 chart can quickly determine an out of control observation. It does not, however, provide information as to which of the variables created the out of control situation. The T^2 value on the control chart is not presented in the original variable units.

Multivariate PCA differs from the standard control chart procedure. It describes the multivariate structure of the data by determining relationships between the variables in a dataset. PCA takes a small number of factors and creates a baseline model in which future observations are assessed against. The control charts will determine whether or not these observations conform to the model.

This was found to be the most effective method. Not only does it have the capability to create control charts to monitor the process for detecting an out of control situation, it also has the ability to diagnose which of the original variables played a part in producing the signal. Similarly, as with the T^2 control chart, the control chart values are not in terms of the original units. However, PCA can translate the out of control signal back into the original variable contributions.

A single monitoring system, using multivariate PCA, is used to assess product performance of the Fuel Cell. Once the PCA model has been developed, T^2 and DModX control charts are generated. The T^2 control chart is different to the standard T^2 control chart. The values that are plotted on the control chart are not the original variable observations. They are calculated from the scores that were generated by the PCA model. The T^2 and DModX charts are monitored simultaneously. Each chart provides a different perspective on how the data fits the model. The T^2 control chart monitors observations that deviate from the mean, but still remain on the plane. Observations that violate the correlation structure of the model, by moving off the plane, are monitored by the DModX chart. In addition to these, scores and loadings charts can be considered as part of the monitoring system. If a signal occurs on any of the control charts generated from PCA, a contribution plot has the ability to isolate a particular variable or group of variables that were responsible for the out of control situation.

This monitoring system, on PCA control charts, could also be applied to production processes with even higher volumes. The number of charts that are monitored will remain the same. Even increasing the number of variables to be monitored does not affect the number of charts that will be monitored. It is only the number of data points that increases.

7.3 Conclusions

All of the objectives for this project were achieved through the application of multivariate SPC techniques on the fuel cell performance data collected from an automated high volume manufacturing process. Multivariate SPC, an alternative technique to the standard Univariate SPC, was proposed. Multivariate SPC control charts can be created easily enough with the right knowledge.

This research investigated the present status of both univariate and multivariate SPC techniques. The literature reviews have shown that there are many Multivariate SPC methods available, but they are not always used in industry due to the lack of education and resources for the application of these techniques.

The most appropriate techniques were selected and these methods were applied to the performance data from a Fuel Cell manufacturing line.

These multivariate SPC techniques generated control charts that can be applied to both continuous and batch processes. This reduces the labour involved in monitoring individual process parameters, as is done in Univariate SPC.

A single monitoring system was identified as the most effective system. It can monitor and control future observations, and also provides suitable tools to assist in identifying which parameters to investigate for root cause of out of control failures. This system can also be used in processes with increased volumes of production.

This project had a lot of challenging obstacles from a quality perspective. The sampling and testing requirements for a high volume manufacturing line are a lot different to a normal manufacturing environment. This project was developed from an R&D environment. Moving from this to a semi-automated line and then developing into a high volume automated line brought unexpected problems. These problems could not have been anticipated at the start of the project, as it was unknown how mass manufacturing would affect the product performance.

These unexpected problems meant that it took a little longer to get to Phase I, where a model of the process could be characterised. All of these issues were worked through with an acceptable final result, i.e. shippable product.

Phase I is a very important part of the process characterisation. An appropriate and representative model developed from Phase I is vital for monitoring future product. The better the fit of the model, the easier it is to detect abnormal situations.

The consequences surrounding inaccurate screening in Phase I would result in abnormal batch performances that do not cause a signal. The danger is that these abnormal conditions remain in the baseline dataset, as they are not thoroughly investigated for root cause, in order to justify their removal.

Similarly, if the baseline model does not contain enough data to support all conditions that a normal process could encounter, a signal may occur as an out of control. In reality, there is nothing wrong with the process. Should this particular condition have been included as part of the model, no signal for an abnormal condition would have been generated.

Phase II monitoring needs to be an efficient system for high volume manufacturing, in order for SPC to be effective. The monitoring system identified through multivariate control charts on PCA satisfies the requirements for the manufacturing line presented in this research. It also satisfies the requirements should production ramp up resulting in an increase in the number of batches manufactured daily.

The various testing implemented both online and offline in the laboratory was initially unexpected, so it involved expanding the laboratory to facilitate the requirements for all final testing of the product.

7.4 Recommendations

7.4.1 Method

It has been shown that Principal Components Analysis (PCA) for multivariate SPC applications on high volume processes is the most effective method for an automated high volume manufacturing line.

7.4.2 Parameter Selection

There can be hundreds of parameters measured as part of a manufacturing process. The parameters that are selected for analysis should be representative of what one is attempting to demonstrate for end product functionality. The critical quality attributes should be considered as well as the corresponding critical process parameters. Should the incorrect parameters be selected then any analysis and decisions made would be done in vain, as they would not be applicable to the final quality decision. The accurate selection of parameters is critical as to whether they are appropriate for what one is attempting to measure. It is important to thoroughly assess all relevant parameters and determine if they contribute valuable information to the principal quality characteristic of the product.

7.4.3 Characterisation

The characterisation of the baseline dataset is a very important step in process monitoring and control. Phase I for process monitoring involves selecting the observations and conditions that are considered to be representative of normal process operation. These observations should be selected from in-control situations. Therefore, this stage should be approached with caution. Should the baseline not be suitably characterised, the Phase II stage of the process, where future observations are monitored for conformance to the baseline, would be compromised. Abnormal future observations will not be accurately represented as an out of control state. Therefore it is essential that the baseline be captured appropriately.

7.4.4 Software

Multivariate data analysis is an advanced technique. The use of the most appropriate software is essential as hand calculations can be very complex, time-consuming and tedious. Standard software packages, such as Minitab and JMP, only cover certain aspects of multivariate data analysis. For instance, the regular T^2 charts and Principal Components Analysis can be generated on separated platforms, but T^2 charts on PCA and the SPE or DModX charts cannot be created effortlessly.

More advanced software is necessary to plot the T^2 statistic and DModX charts on Principal Components Analysis. SIMCA, developed by umetrics, is one such software package. It is largely used for model building and predictive analysis. It has the ability to generate all the charts necessary in order to build an effective monitoring system. It also has the capability to generate contribution plots. These are an important tool for establishing root cause. SIMCA can be expensive as licences for this type of software are usually costly.

For this reason, it is highly likely that many industries would incorporate the T^2 method for monitoring a process. The importance of pros should be weighed up against the cons in order to determine a company's priorities in relation to quality. No compromise on quality should have to be made.

7.4.5 Risk of False Alarms

In terms of the data analysis, the chosen α levels are crucial in a high volume process. One should be aware for the implications associated with selecting the α level. An α level of 0.05 results in 1 in 20 batches will naturally falling outside the limits. An α level of 0.01 results in 1 in 100 and $\alpha = 0.0027$ results in 1 in 370 out of control signals when there is no abnormal situation occurring. If the α level is not carefully considered, there would be a lot of resources wasted in trying to find a root cause for a problem that doesn't exist.

An intermediate choice of $\alpha = 0.01$ should be selected for a high volume production line. This will result in 1 in 100 batches naturally falling outside the control limit.

7.5 Future Research

7.5.1 Partial Least Squares

Partial Least Squares (PLS) could be considered as an alternative to Principal Components Analysis (PCA). PLS is also capable of providing a monitoring system for multivariate processes. It is based on regression techniques where a predictive model has been characterised from assessing the impact that the input variables have on the output variables.

7.5.2 Automated Data Collection

Systems could be implemented online to contain software that would take parameter measurements to monitor control using PCA or PLS methods. This would enable real-time monitoring, so that a pending batch status could be made before a batch reaches end of line.

Fuel data could be collected and PLS could be used to build a model for the chemical composition, and how it will affect the overall performance parameters. This data could be assessed before the fuel is transferred into fuel cell.

7.5.3 Robust Charting Methods

More robust control charts could be introduced in Phase I, when screening for outliers. The traditional multivariate T^2 control chart can quickly detect extreme outliers. However, moderate outliers will be unsuccessfully detected. Robust control charts are more efficient for detecting outliers during Phase I.

These control charts are determined using robust estimators of the mean vector and covariance matrix. These would replace the mean vector and covariance matrix, estimated using conventional methods.

There are various methods for evaluating these robust estimates. The minimum covariance determinant (MCD) estimators and the minimum volume ellipsoid (MVE) estimators were proposed by Vargas (2003) and Jensen *et al.* (2007). More recently, Chenouri *et al.* (2009) proposed using reweighted minimum covariance determinant (RMCD) estimators, which can be used in Phase II.

7.6 References

Chenouri, S., Steiner, S.H., Variyath, A.M. (2009), "A Multivariate Robust Control Chart for Individual Observations". *Journal of Quality Technology*, Vol. 41, No. 3, pp. 259-271.

Jensen, W.A., Birch, J.B., Woodall, W.H. (2007), "High Breakdown Estimation Methods for Phase I Multivariate Control Charts". *Quality and Reliability Engineering International*, Vol.23, No. 5, pp. 615-629.

Vargas, J.A. (2003), "Robust Estimation in Multivariate Control Charts for Individual Observations". *Journal of Quality Technology*, Vol. 35, pp. 367-376.

Xue, X., Tang, J., Sammes, N., Ding, Y. (2006), "Model-based condition monitoring of PEM fuel cell using Hotelling T^2 control limit". *Journal of Power Sources*, Vol. 162, pp. 388-399.