ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

AN OVERVIEW OF ENGINEERING PROCESS CONTROL IN THE UNIVARIATE AND MULTIVARIATE CASE

By

Spyridon K. Mexas

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
2003

ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

AN OVERVIEW OF ENGINEERING PROCESS CONTROL IN THE UNIVARIATE AND MULTIVARIATE CASE

By

Spyridon K. Mexas

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
July 2003

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΕΠΙΣΚΟΠΗΣΗ ΣΤΟ ΜΗΧΑΝΟΛΟΓΙΚΟ ΈΛΕΓΧΟ ΔΙΕΡΓΑΣΙΩΝ ΣΤΗ ΜΟΝΟΜΕΤΑΒΛΗΤΗ ΚΑΙ ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΠΕΡΙΠΤΩΣΗ

Σπυρίδων Κ. Μέξας

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Ιούλιος 2003

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

A Thesis submitted in partial fulfillment of

the requirements for the degree of

Master of Science

AN OVERVIEW OF ENGINEERING PROCESS CONTROL IN THE UNIVARIATE AND MULTIVARIATE CASE

Spyridon K. Mexas

Approved by the Graduate Committee

S. Psarakis	A. Kostaki	D. Karlis
Assistant Professor	Assistant Professor	Lecturer
Thesis Supervisor	Members of the Commitee	

Athens, October 2003

Michael Zazanis, Associate Professor
Director of the Graduate Program

DEDICATION

To my newborn son and my lovely wife

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my supervisor Dr. Stelios Psarakis for his valuable help and advice during the preparation of this thesis. I wish also to thank my family for their support and encouragement as well as everyone that provided me with anykind of support.

VITA

I was born in Athens in February 1965. In 1982 I entered the Military Academy of Evelpidon in Athens and I graduated in June 1987. In October 2001 I followed the postgraduate program in the Department of Statistics, Athens University of Economics and Business. My research interests are in Statistical and Engineering Process Control.

ABSTRACT

Spyridon Mexas

AN OVERVIEW OF ENGINEERING PROCESS CONTROL IN THE UNIVARIATE AND MULTIVARIATE CASE

July 2003

In a modern manufacturing environment traditional control charts present limited usefulness due to the data which are often autocorrelated. Process adjustments techniques based on Engineering Process Control have become particularly popular the last years among the quality control engineers due to the recent interest on combination of Statistical Process Control techniques with the Engineering Process Control techniques.

The aim of this dissertation is to present process adjustment techniques in order to improve a production process and the combination of Statistical Process Control with the Engineering Process Control in the univariate and multivariate case respectively.

ΠΕΡΙΛΗΨΗ

Σπυρίδων Μέξας

ΕΠΙΣΚΟΠΗΣΗ ΣΤΟ ΜΗΧΑΝΟΛΟΓΙΚΟ ΈΛΕΓΧΟ ΔΙΕΡΓΑΣΙΩΝ ΣΤΗ ΜΟΝΟΜΕΤΑΒΛΗΤΗ ΚΑΙ ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΠΕΡΙΠΤΩΣΗ

Ιούλιος 2003

Σε ένα σύγχρονο βιομηχανικό περιβάλλον τα παραδοσιακά διαγράμματα ελέγχου διεργασιών παρουσιάζουν περιορισμένη χρησιμότητα εξαιτίας των δεδομένων τα οποία συνήθως είναι αυτοσυσχετισμένα. Οι τεχνικές προσαρμογής διεργασιών οι οποίες βασίζονται στον Μηχανολογικό Έλεγχο Διεργασιών έχουν γίνει ιδιαίτερα δημοφιλείς τα τελευταία χρόνια μεταξύ των μηχανικών ελέγχου ποιότητας οφειλόμενες στο πρόσφατο ενδιαφέρον συνδυασμού των τεχνικών του Στατιστικού Ελέγχου Διεργασιών με τις τεχνικές του Μηχανολογικού Έλεγχο Διεργασιών.

Ο σκοπός της συγκεκριμένης διατριβής είναι να παρουσιάσει τις τεχνικές προσαρμογής διεργασιών προκειμένου να βελτιωθεί μια παραγωγική διαδικασία, και στον συνδυασμό του Στατιστικού Ελέγχου Διεργασιών με τον Μηχανολογικό Έλεγχο Διεργασιών στη μονομεταβλητή και στην πολυμεταβλητή περίπτωση αντίστοιχα.

TABLE OF CONTENTS

Chapter 1 Introduction 1

1.1 Statistical Process Control 1

1.1.1 Acceptance Sampling 1

1.1.2 Statistical Process Control 1

1.2 Engineering Process Control 2

1.3 Combining SPC with EPC 2

Chapter 2 Literature Overview 5

2.1 Introduction 5

2.2 Process Monitoring 5

2.3 SPC Charts for IID Data 6

2.3.1 Rational Subgroups 6

2.3.2. Shewart Control Charts 7

2.3.2.1 The $\bar{x}$ and R Control Chart 8

2.3.2.2 The $\bar{x}$ and S Control Chart 9

2.3.2.3 Shewart Control Chart for Individuals 11

2.3.3 Control Charts for Attributes 12

2.3.3.1 Control Chart for Fraction nonconforming (p - chart) 12

2.3.3.2 Control Chart for Nonconformities (c - chart) 14

2.3.3.3 Control Chart for Nonconformities per unit (u - chart) 14

2.3.4 Alternative Control Charts 15

2.3.4.1 The Cumulative Sum (CUSUM) Control Chart 16

2.3.4.2 The Exponentially Weighted Moving Average (EWMA) Control Chart 19

2.4 Control Charts for Autocorrelated Data 20

2.4.1 Standard Control Charts Based on Residuals 21

2.4.1.1 The Moving Centerline EWMA Control Chart 21

2.4.1.2 CUSUM Control Chart for Monitoring Autocorrelated Processes 22

2.4.1.3 Cuscore Charts for Process Monitoring 23

2.4.2 Modified Control Charts 24

2.4.2.1 A Model Free Approach (The Batch means Control Chart) 25

Chapter 3 ARIMA and Transfer Function Models 27
3.1 Introduction 27
3.2 The Necessity of ARIMA Models 27
3.3 Stationarity and Non – Stationary Time Series Models 28
3.4 Basic Properties of Time Series 29
3.5 Stationary ARMA(p,q) models 30
3.5.1 Autoregressive Models 30
3.5.2 Moving Average Models 31
3.5.3 Mixed Autoregressive – Moving Average Models 31
3.5.3.1 Order Selection 32
3.5.3.2 The FPE and AIC and AICC criteria 32
3.5.3.3 Estimating the Coefficients of an ARMA(p,q) model 33
3.5.3.4 Forecasting ARMA processes 33
3.6 Modelling Non – Stationary Models 34
3.6.1 Special cases of the ARIMA model 35
3.6.2 Variance Stationarity and Transformations 35
3.7 Transfer Function Models 38
3.7.1 The Importance of Transfer Function Models in Industrial Processes 38
3.7.2 Basic terminology in transfer functions 38
3.7.3 Continuous and Discrete Dynamic Models 40
3.7.3.1 Continuous Dynamic Models 41
3.7.3.2 Discrete Dynamic Models 41
3.7.4 Transfer Function Models with Added Noise 42
3.7.5 Determination of the Transfer Function Models 42
3.7.6 Forecasting Using Leading Indicators 45
3.7.7 A Detailed example in Transfer Function Modelling 46
3.8 Fundamentals of State Space Modelling 50
3.8.1 Modelling a continuous process 51
3.8.2 Modelling a discrete process 52
3.8.3 Obtaining the transfer function from state space form 52

Chapter 4 Engineering Process Control 53
4.1 Introduction 53

4.2 Process Adjustment 53
4.2.1 Deming's Funnel Experiment 54
4.2.2 When SPC and EPC should be Applied in a Process 57
4.3 Controllers 58
4.3.1 Basic terminology of controllers 59
4.3.2 Controllers Definition 59
4.3.3 A Simple Feedback Adjustment Scheme 60
4.3.4 Discrete PID Controllers 62
4.3.4.1 Parameterization of PID Controllers 62
4.3.4.2 The Proportional (P) Controller 62
4.3.4.3 The Proportional Integral (PI) Controller 63
4.3.4.4 The Proportional Integral Derivative (PID) Controller 65
4.3.5 EWMA Controllers 65
4.3.5.1 Relationship of Integral Control and EWMA Statistic 66
4.3.5.2 The Single EWMA Controller 67
4.3.5.3 The Double EWMA (DEWMA) Controller 68
4.3.5.4 Age – Based Double EWMA Controller 69
4.3.5.5 EWMA Controller with Step Change Disturbance 69
4.3.6 Feedback Adjustment Charts 70
4.3.6.1 The Bounded adjustment chart 70
4.3.7 Other Types of Controllers 72
4.3.7.1 Feedforward Controllers 72
4.3.7.2 Feedforward as a Supplement to Feedback Controllers 73
4.4 Grubb's Adjustment Rules 73
4.4.1 Grubbs extended rule 75
4.5 Self Tuning (Adaptive) Controllers 76
4.5.1 Direct and Indirect Controllers 76
4.5.2 An Example of Indirect ST Controller 76

Chapter 5 Optimal Feedback Controllers 79
5.1 Introduction 79
5.2 Optimal FeedbackControllers 79
5.2.1 Minimum Variance (MMSE) Controllers 80

5.2.2 The Generalized Minimum Variance (GMV) Controller 82
5.2.3 The Constrained MMSE (CMMSE) Controller 84
5.2.4 Optimal PI Controllers 85
5.3 Tuning for Optimization the EWMA Controllers 87
5.3.1 Tuning a Single EWMA Controller 88
5.3.2 Tuning a Double EWMA (DEWMA) Controller 90
5.4 The Variance Constrained PI Self Tuning Controller 92
5.4.1 Derivation of the variance constrained ST PI controller 93

Chapter 6 Integration of SPC – EPC 95
6.1 Introduction 95
6.2 Comparison of SPC and EPC Control 95
6.3 The Algorithmic Statistical Process Control 97
6.3.1 The Concept of ASPC 97
6.3.2 The Batch Polymerization Example 97
6.4 SPC Monitoring of MMSE and PI Controllers 99
6.4.1 SPC Monitoring of MMSE Controlled Processes 100
6.4.2 SPC Monitoring of PI Controlled Processes 101
6.5 Combining SPC and EPC Control when a Transient Disturbance Exists 102
6.5.1 Effective Integrated Process Control in a Transient Period Using Simulation and an ARMA (p,q) Disturbance Model 102
6.5.2 IPC Design for Startup Operations wityh ARIMA(0,1,1) Disturbance Model 104
6.5.3 Transient Disturbances with Cuscore Control Charts 105
6.5.4 An EWMA Bounded Adjustment Scheme with Adaptive Noise Variation 107
6.6 An Integrated SPC/EPC scheme with Feedforward Control 109
6.6.1 The shift scenario 110
6.6.2 The trend scenario 110
6.6.3 Integrating SPC and feedforward control 110
6.7 Integration of SPC/EPC using Adaptive Controllers 111
6.7.1 Guidelines of detection and identification 112
6.7.2 Conclusions 113

Chapter 7 Multivariate SPC – EPC 115

7.1 Introduction 115

7.2 The Multivariate SPC 115

7.2.1 The Hotelling T^2 control chart 116

7.2.2 The Multivariate CUSUM and EWMA Control Charts 117

7.2.2.1 The Multiple Univariate Cusum chart 118

7.2.2.2 The Grosier's MCUSUM charts 119

7.2.2.3 The MC1 and MC2 charts 120

7.2.2.4 An MCUSUM which is based on sequential probability ratio (SPRT) 121

7.2.2.5 Multivariate EWMA Charts 122

7.2.3 Monitoring Multivariate Processes using Regression Adjustment 123

7.2.4 Interpretation of Out – of – Control Signals 124

7.3 The Multivariate SPC 124

7.3.1 A Multivariate EWMA Controller 125

7.3.2 A Multivariate Double EWMA Controller 126

7.4 Integrating the EPC Processes with Joint Monitoring Schemes 128

7.4.1 Joint Monitoring of a PID Controlled Process 128

7.4.2 The dynamic T^2 chart for monitoring feedback controlled processes 130

Chapter 8 Conclusions and Further Research 133

8.1 Introduction 133

8.2 Conclusions and Further Research 133

References 137

LIST OF TABLES

Table		Page
5.1	EWMA Designs	90
5.2	DEWMA Designs	92
6.1	SPC and EPC comparison (Messina (1992))	96

LIST OF FIGURES

Figure	Page
2.1 $\bar{x}$ and R control chart	10
2.2 $\bar{x}$ and S control chart	11
2.3 p-chart	13
2.4 p-chart with revised control limits	13
2.5 c-chart	15
2.6 u-chart	15
2.7 V – mask control scheme	17
2.8 $\bar{x}$ and R control chart vs CUSUM control chart	18
2.9 EWMA control chart with n=5, $\lambda = 0.25$, L=3	20
2.10 Plot of batch means using b = 3	26
2.11 Xbar chart of residuals	26
3.1 Plot of a non stationary time series data	29
3.2 Plot of a stationary time series data	29
3.3 Plot of the original series of airlines passengers with the ACF and PACF plots	36
3.4 Plot of transformed series of airlines passengers with the ACF and PACF plots	36
3.5 Plot of the transformed series of airlines passengers with the ACF and PACF plots exhibits trend and seasonality	37
3.6 Plot of the transformed series of airlines passengers with the ACF and PACF plots seems to be stationary	37
3.7 Input and Output from a dynamic system	39
3.8 Cross correlation plot before prewhiten the input	45
3.9 Cross correlation plot after prewhiten the input	45
3.10 The leading indicator and sales data time serie graphs	46
3.11 The SACF plot of leading indicator and sales data time series	47
3.12 Sample ACF of the X_t time series	48
3.13 Sample cross correlations graph from X_t and Y_t time series	48
3.14 Plot of the cross correlations residuals	50

3.15 Plot of the residuals correlations 50

4.1 Simulation of the Funnel experiment of four adjustment rules 56

4.2 Time series simulation of the four adjustment rules 56

4.3 When SPC and EPC should be applied to a process 58

4.4 Block diagram of a transfer function 60

4.5 Block diagram of a sum of variables 60

4.6 A feedback control 61

4.7 Effect of increasing K_P in a proportional controller 63

4.8 Effect of increasing K_I in an integral controller 64

4.9 Bounded adjustment chart showing the original and adjusted output the EWMA and the real adjustments. The circled EWMA indicate where adjustments are made 71

4.10 A feedforward scheme 72

4.11 Output values of the quality characteristic using indirect ST controller 77

4.12 Values of the manipulated variable suggested by the indirect ST controller 77

4.13 Parameter estimates used by the indirect ST controller 78

5.1 Effects on input and output a) White noise input b) Output of white noise input c) MMSE control input d) output when X_t is an MMSE control 83

7.1 A T^2 control chart with Phase I and II control limits 117

7.2 An MCUSUM based on SPRT 122

7.3 An MEWMA control 123

Chapter 1

Introduction

1.1 Statistical Process Control

Statistical Quality Control has its roots in the decade of 1920 (Bell laboratories) where two major areas were developed:

1.1.1 Acceptance sampling

Acceptance sampling methods are used in industry to make decisions regarding the disposition of "lots" of manufactured items including the acceptance or rejection of individual lots. Harold Dodge and Harry Romig (1959) developed control sampling plans using a random sample of units from a lot. The decision of accepting or rejecting the lot depends on the number of units that don't meet on or more of specific quality characteristics.

1.1.2 Statistical Process Control

Statistical Process Control (SPC) also known as statistical process monitoring consists of techniques to monitor a production process over time in order to detect changes in process performance. The main purpose of SPC is to look for assignable causes (variability) in the process data. If assignable causes are present the production engineer stops the production, eliminates the assignable cause and restarts the production cycle. Walter Shewart was the first who introduced control charts to monitor a production process for assignable causes. Since that time many other techniques were developed in the frame of control charts such as CUSUM and EWMA charts. In SPC, the production process is thought to lie in either of two states: an in-control state and an out-of-control state. Control charts are used to distinguish between these two states. As long as, the chart does not signal the existence of an out-of-control state, the process is thought to be in statistical control. SPC control has been proven to be effective, in the form of control chart use, for the parts industries by over 60 years of successful applications. However, the SPC techniques in the process industries have had limited successes due to the nature of the process data which are often autocorrelated.

1.2 Engineering Process Control

One fundamental assumption behind the use of traditional control charts is that successive values of the quality characteristic through time are not correlated with each other. However, in a modern manufacturing environment the quality data are serially correlated (autocorrelated). The performance of SPC schemes is badly affected in the presence of autocorrelation and an alternative method is necessary in order to compensate for process dynamics. Therefore, engineering process control (EPC) methods are used to monitor the process output, compare it with target value, and make continuous compensatory adjustments to the process input to keep the output on target. The compensation is applied in the form of feedback, feedforward, or a combination of both. However, it is well-known that when a process is in statistical control frequent adjustments will inflate the process variance and thus increase the process output off target.

1.3. Combining SPC with EPC

Statistical process control (SPC) and engineering process control are two different but complementary approaches with common goal the reduction of variation of the process. SPC attempts to remove disturbances using process monitoring, while EPC attempts to compensate them using process adjustment. Therefore, a substantial number of researchers proposed various methods that combine SPC and EPC in an integrated form. The idea of the SPC and EPC combination is simple. EPC schemes should be able to produce uncorrelated data as the output deviation from the target. However, EPC schemes do not seek to isolate and eliminate the assignable causes. These causes can be detected by applying an appropriate SPC scheme along with the EPC adjustment rules.

The aim of this dissertation is to present an overview of the most recent approaches concerning the modern manufacturing environment including SPC, EPC schemes, as well as a combination of them in the univariate and the multivariate case with emphasis to engineering process control.

In the next chapter, a literature review of process monitoring using the most known SPC charts is given. Chapter 3 is referred to ARIMA and transfer functions models. Furthermore, the state space modeling is briefly discussed. Next, in chapter 4, the most known feedback and feedforward controllers as well as a combination of them are presented. In chapter 5, the MMSE, PID and EWMA controllers in the sense of optimization are discussed. In chapter 6, we present the integration of feedback and feedforward control schemes using statistical process control charts. Chapter 7, is concerned with SPC, EPC and SPC/EPC schemes of an industrial process where multiple inputs and multiple outputs (MIMO) define the quality of a product. Finally, in chapter 8, we present some conclusions and further research topics for the SPC and EPC approaches.

Chapter 2

Literature Overview

2.1. Introduction

Monitoring and improving quality is a part from an industrial process. Much engineering effort was put into action to reach a situation where the quality characteristics of a product conform to the predetermined demands. Statistical Process Control (SPC) is a tool to achieve this objective through the reduction of variability. A brief description of process monitoring is given in section 2.2. In section 2.3 the most common charts for independent data are presented while section 2.4 is concerned with control charts for autocorrelated data.

2.2 Process Monitoring

Statistical Process Control (SPC) is a set of techniques such as Shewart control charts, exponentially moving average (EWMA) and cumulative sum (CUSUM) control charts which are widely used in industries to improve the quality giving us the feasibility to monitor the output of a production process. In industries a usual phenomenon is that, independently how well a process designed or maintained, it has an amount of unavoidable variability. Two sources of variability are considered in SPC; common cause that is, inherent variability which is caused by chance where we cannot eliminate it without a deep modification in the process and assignable or special cause due to particular problems variability which disturbs the normal functioning of the process. The latter usually arises from improperly machines, human errors and raw material. If a process runs in the first case we say that it is in statistical control or else is out of control. So, with the term "process monitoring" we try to detect the occurrence of an assignable cause in order to correct it.

2.3 SPC charts for IID data.

A key assumption underlying the use of the traditional control charts in the univariate case is that the observations are independent through time. In the next paragraphs the most usual charts are presented when quality characteristics are measured in numerical scale and when is not convenient to represent this characteristic numerically so to classify it as conforming and nonconforming, named as attribute. These charts are classified into two general types: i) variable control charts when the quality characteristic can be measured in numerical scale where both the mean value and variability are accounted; and ii) attributes control charts which are not measured numerically and we classify each item as conforming and non - conforming. However, in modern manufacturing processes data are serially correlated in time (autocorrelated) violating this assumption. Autocorrelation results in a number of problems affecting the performance of control charts as increasing the false alarm rate Alwan (1992). The same problem extends in the multivariate case Lowry and Montgomery (1995). In the next paragraphs we briefly present the basic SPC charts for IID and autocorrelated data.

2.3.1 Rational Subgroups

The way the observations are sampled and grouped may have a large effect on the behaviour of a control chart. Shewart (1931) introduced the concept of "rational subgroup". Under this principle rational subgroup is a sample where only common causes are responsible of the observed variation. In other words, the within sample variability is due to common cause variability while the between samples variability is due to special causes. In general, two approaches are considered. In the first approach the sample units are produced as close together in time as possible so the chance to observe a special cause is the minimum. This approach provides us with a better estimation of standard estimation when we use variables control charts. In the second approach, each sample consists of units which are representative of all units that have been produced since the last sample was taken. This approach is preferable when we want to decide to accept or not all the units of the product that have been produced since the last sample.

2.3.2 Shewart Control Charts

Shewart charts first introduced by W.A Shewart (1924) plot either the individual process measure or the average value of a small sample (usually 5) depending the sampling frequency along with the target level and control limits. The assumed in control model, known as Shewart's model, is

$$Y_t = \mu + \varepsilon_t \qquad \text{for} \qquad t = 1,2, \ldots$$

where Y_t is the value of the quality characteristic from t-th sample, μ is the mean value of this characteristic and ε_t is the random error where $\varepsilon_t \sim N(0,\sigma^2)$, $Cov(\varepsilon_t \varepsilon_{t+i})=0$ for $i \neq 0$.

Assuming that the quality characteristic is normally distributed with mean μ and standard deviation σ the probability that any sample mean $\bar{x}$ will fall between

$$\mu + z_{1-a/2}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \mu - z_{1-a/2}\frac{\sigma}{\sqrt{n}}$$

is $1-\alpha$. Thus, if μ and σ are known they could be applied to determine the lower and upper control limits for the sample means. Furthermore, we usually replace $z_{1-a/2}$ with 3, so that "three sigma" limits are applied. The three sigma limits correspond to $\alpha=0.0027$, that is, only 27 of 10,000 observations may fall out of the control limits generating an incorrect out of control signal (false alarm rate).

In designing a control chart we must specify both the sample size and the frequency of sampling. One way to weigh up the choice of the sample size and sampling frequency is through the average run length (ARL) of the control chart. The ARL is the average number of points that must be plotted before a point indicates an out-of-control condition. If the observations are uncorrelated then the ARL is,

$$ARL = \frac{1}{p}$$

where p is the probability that any point exceeds the control limits. Therefore, a good control chart should own the property that when the process is in control the ARL is large and when the process is out of control the ARL is small.

2.3.2.1 The $\bar{x}$ and R Control chart

Assuming that previous model is correct and $x_1, x_2, ..., x_n$ is sample with size n the average of the sample is

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} \text{ where } \bar{x} \sim N\left(\mu, \sigma^2/n\right)$$

Usually μ and σ are unknown and we estimate them from preliminary samples when the process is in control. Suppose that m samples are available with n observations each of them and $\bar{x}_1, \bar{x}_2, ..., \bar{x}_m$ the average of each sample the best estimator for μ is

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + ... + \bar{x}_m}{m}$$

and R_i , $i = 1, 2, ..., m$ is the range of each sample given by the formulae

$$R = x_{\max} - x_{\min}$$

and if $R_1, R_2, ..., R_m$ are the ranges of m samples the average range is

$$\bar{R} = \frac{\bar{R}_1 + \bar{R}_2 + ... + \bar{R}_m}{m}$$

An unbiased estimator for σ is $\hat{\sigma} = \frac{\bar{R}}{d_2}$ where d_2 is the mean of the fraction $\frac{R}{\sigma}$.

The center lines and control limits for $\bar{x}\, and\, R$ charts are:

Control limits for $\bar{x}$ Chart

$$UCL = \bar{\bar{x}} + A_2 \bar{R}$$
$$Center\, line = \bar{\bar{x}}$$
$$LCL = \bar{\bar{x}} - A_2 \bar{R}$$

Control limits for R Chart (2-1)

$$UCL = D_4 \bar{R}$$
$$Center\, line = \bar{R}$$
$$LCL = D_3 \bar{R}$$

where $A_2 = d_2\sqrt{n}$, $D_3 = 3d_3 / d_2$, *and* $D_4 = 1 + 3d_3 / d_2$.

The values of d_2, d_3, A_2, D_1, D_4 are easily can be found in most books of SPC (e.g Montgomery 2001) tabulated in tables according to the sample size n. Furthermore, Montgomery (2001) suggests treating the initial control limits as "trial" control limits, subject to subsequent revision. When the R chart is out of control, we often eliminate the out–of–control points and a "revised" value of $\bar{R}$ is recomputed. This value is used to determine the new limits and centerline of the R chart and new limits on the $\bar{x}$ chart.

2.3.2.2 The $\bar{x}$ and S Control chart

Sometimes is preferable to estimate the process standard deviation directly. This leads to $\bar{x}$ and S control chart which we choose when the following practical rules are applied:

a. The sample size n is moderately large (e.g $n > 10$) and,

b. The sample size is variable.

Despite the fact that we know that an unbiased estimator of σ^2 is the sample variance where $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}$, S is not an unbiased estimator of σ. Because the underlying distribution is normal, S estimates the quantity $c_4\sigma$, where c_4 constant depends on sample size n. Furthermore, the standard deviation of S is $\sigma\sqrt{1-c_4^{\,2}}$ .Assuming that m preliminary samples are available, each of size n, and S_i is the standard deviation of the ith sample. The average of m standard deviations is $\bar{S} = \frac{1}{m}\sum_{i=1}^{m} S_i$. The centerlines and control limits for $\bar{x}$ and S charts are:

Control limits for $\bar{x}$ Chart

$$UCL = \bar{\bar{x}} + A_3 S$$
$$Center\,line = \bar{\bar{x}} \tag{2-2}$$
$$LCL = \bar{\bar{x}} - A_3 S$$

Control limits for S Chart

$$UCL=\overline{S}+3\frac{\overline{S}}{c_4}\sqrt{1-c_4^2}$$
$$Center\,line=\overline{S} \qquad (2\text{-}3)$$
$$LCL=\overline{S}-3\frac{\overline{S}}{c_4}\sqrt{1-c_4^2}$$

As an example of the previous charts we have 15 samples with sample size n = 10 and the quality characteristic is the fill volume of beverage bottles. We set $\overline{x}$ and R as well as $\overline{x}$ and S charts on this process where are presented in Figure 2.1and 2.2 respectively.

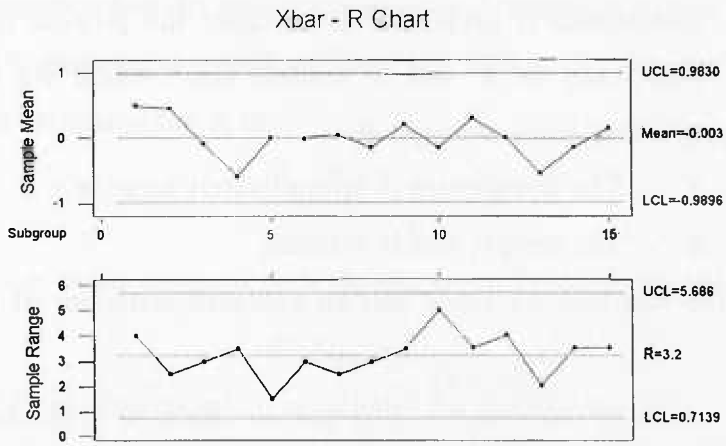


Figure 2.1 $\overline{\mathrm{x}}$ and R control chart

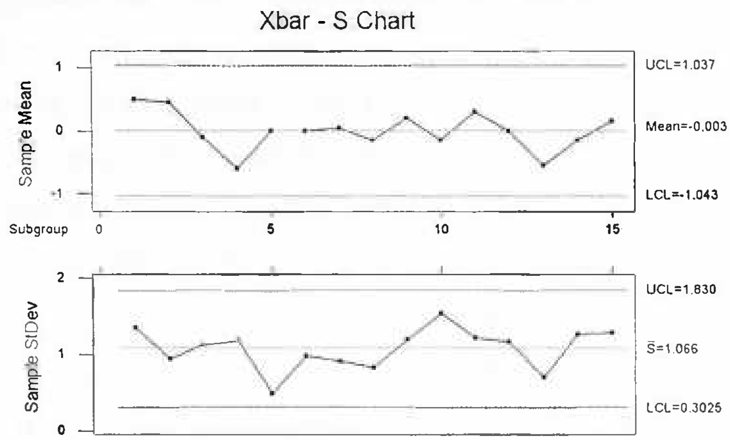


Figure 2.2 $\overline{x}$ and S **control chart**

As we see from the graphs, the process is in statistical control and the limits in S control chart are tighter than limits in the R chart indicating us that S chart in this case, is preferable than R chart provided that sample size n is moderately large. Another one point is that control limits of $\overline{x}$ chart based on $\overline{S}$ are slightly different to the $\overline{x}$ chart based on $\overline{R}$. This is in general a presumable result although there are cases where these limits are almost the same.

2.3.2.3 Shewart Control Chart for individuals

There are cases in industrial processing, such as differ repeated measurements on the process, were the sample size which we use for monitoring is $n=1$. In these cases we use a control chart for individuals based in the moving range of two successive observations that is, $MR=|x_i-x_{i-1}|$. The control limits are given by the following formula:

Control limits for MR Chart

$$UCL=\overline{x}+3\frac{\overline{MR}}{d_2}$$
$$Center\,line=\overline{\overline{x}}$$
$$LCL=\overline{x}-3\frac{\overline{MR}}{d_2} \tag{2-4}$$

2.3.3 Control Charts for Attributes

If we cannot represent a quality characteristic in numerical scale we classify each item as "conforming" or "nonconforming". In this case quality characteristics are called attributes such as the production of malfunctioned parts of an industrial product. The most widely known and used control charts which are presented further are:

i. Control chart for fraction nonconforming or p – chart

ii. Control chart for nonconformities or c – chart

iii. Control chart for nonconformities per unit or u - chart

2.3.3.1 Control chart for fraction / number nonconforming (p – chart)

Fraction nonconforming (p) is the ratio of the number of nonconforming items in a population to the total number of items in this population. Furthermore sample fraction nonconforming ($\hat{p}$) is the ratio of the number of nonconforming units D in the sample with size n where $\hat{p}=D/n$. Because of the distribution which generated the data (nonconforming) assumed to be binomial, the estimated fraction nonconforming from observed data given that we have selected m preliminary samples (m usually is 20 or 25) for the i-th sample is: $\hat{p}_i=D_i/n$ and the average of nonconforming is $\overline{p}=\sum_{i=1}^{m}\hat{p}_i/m$ where $\overline{p}$ is the estimator of the unknown fraction nonconforming p. The center line and control limits for p *chart* are:

Control limits for $p-chart$

$$UCL=\overline{p}+3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

$$Center\,line=\overline{p} \qquad (2\text{-}5)$$

$$LCL=\overline{p}-3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

In the figure 2.3 is presented a usual p – chart where we have nonconforming parts from an industrial process in samples of size 100. As we observe only the

12th sample exceeds the control limits (UCL). Assuming that assignable causes can be found we determine the revised control limits in figure 2.4 where there are not any points out of control.

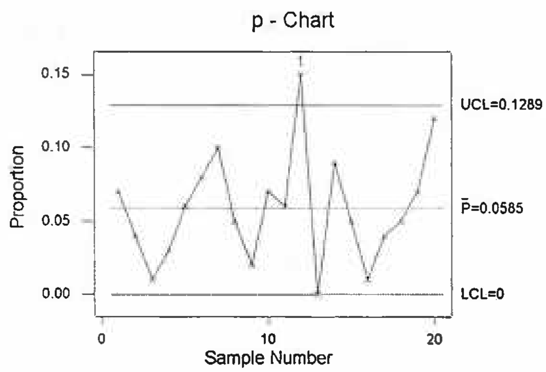


Figure 2.3 p-chart

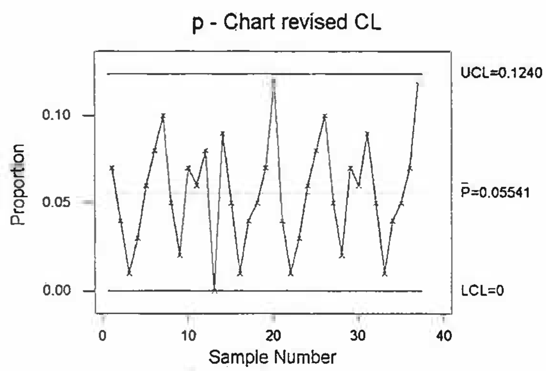


Figure 2.4 p-chart **with revised control limits**

Sometimes is more desirable to construct a control chart on the number of nonconforming. This is called *np chart*. The interpretation of this chart is similar as that of *p chart*. The center line and control limits of this chart are:

Control limits for *np chart*

$$UCL = np + 3\sqrt{np(1-p)}$$
$$Center\,line = np \qquad (2\text{-}6)$$
$$LCL = np - 3\sqrt{np(1-p)}$$

2.3.3.2 Control chart for nonconformities(c – chart)

In the sense, that a nonconforming item of a product does not satisfy one or more of predetermined specifications of this product results in a defect or nonconformity. Suitable control charts to detect nonconformities is the *c chart* where the number of defects (c) is the parameter of the Poisson distribution (Montgomery, 2001). Therefore the control limits and center line for *c chart* are:

Control limits for *c chart*

$$UCL = c + 3\sqrt{c}$$
$$Center\,line = c \qquad (2\text{-}7)$$
$$LCL = c - 3\sqrt{c}$$

When no standard is given we substitute in the previous formulae c by $\bar{c}$ that is the observed average number of nonconformities in a preliminary sample.

2.3.3.3 Control chart for nonconformities per unit (u – chart)

When the sample size is exactly equal to one inspection unit (u) where $\bar{u}$ is the average number of nonconformities per inspection units the parameters of control chart are as follows.

Control limits for *u chart*

$$UCL = \bar{u} + 3\sqrt{\bar{u}/n}$$
$$Center\,line = \bar{u} \qquad (2\text{-}8)$$
$$LCL = \bar{u} - 3\sqrt{\bar{u}/n}$$

In the next figures (2.5 and 2.6) we have 22 samples with sample size 25 where the number of nonconformities varying between 1 and 20. We observe that clearly the points do not fall within the bounds of the 3s control limits, because the 10th, 11th and 22nd sample have unusual number of nonconformities and we conclude that the process is not in statistical control

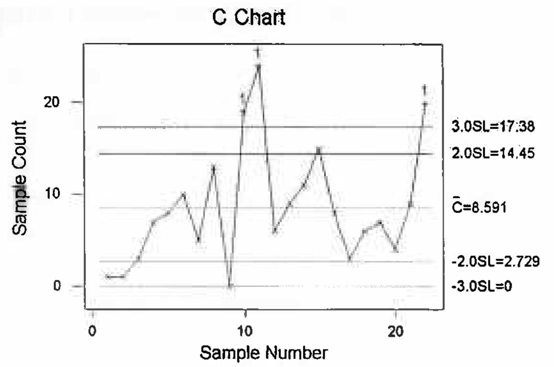


Figure 2.5 c chart

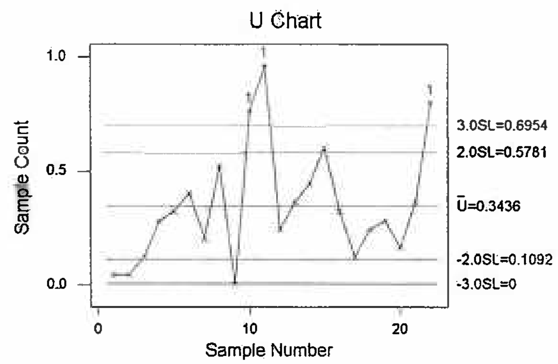


Figure 2.6 u chart

2.3.4 Alternative Control Charts

The above control charts were developed by W.A. Shewart so they named Shewart control charts. These control charts only use the current observation or sample to monitor the process that means they do not have memory. This is a serious drawback, because it ignores any information given by the entire sequence of points. Other criteria like warning limits and tests of runs can be used to include information (some memory) from previous observation but decrease the simplicity of Shewart control chart. Here, we present two very effective control charts that take account for previous observations and essentially used when small shifts from the mean is of interest: The cumulative sum (CUSUM) control chart and the exponentially weighted moving average (EWMA) control chart.

2.3.4.1 The cumulative **sum (CUSUM) control chart**

The CUSUM control chart was introduced by Page (1954) and uses an unweighted sum of all previous observations. This chart has a long memory. The author proposed that if we have collected samples we can plot the quantity $C_i = \sum_{j=1}^{i} (\bar{x}_j - \mu_0)$ where $\bar{x}_j$ is the average of jth sample and μ_0 is the target mean of the process. If the mean of the process is $\mu_t > \mu_0$ then a positive drift will build up in cumulative sum C_i. On the contrary if $\mu_t < \mu_0$ then a negative drift will build up in C_i. Hence, if we observe trend in the plotting points this indicate us that the process mean has shifted and we search for an assignable cause. There are two approaches to represent CUSUMs, the tabular (or algorithmic) CUSUM and the V – mask CUSUM. Some reasons not to use the latter approach are described in Montgomery (2001). These reasons will be discussed later.

(i) The Tabular (CUSUM)

If x_i is an observation of a sample where the process is in control then $x_i \sim N(\mu_0, \sigma^2)$. When there is a shift in the mean,"target" value, we observe that the CUSUM will signal and an adjustment must be made. If we consider C_i^+, C_i^- the statistics which compute the accumulating derivations above and below the target value respectively, we have:

Tabular Cusum

$$C_i^+ = \max\left[0, x_i - (\mu_0 + K) + C_{i-1}^+\right]$$
$$C_i^- = \max\left[0, (\mu_0 - K) - x_i + C_{i-1}^-\right] \quad (2\text{-}9)$$

where K is called reference value and is computed as $K = \frac{\delta}{2}\sigma = \frac{|\mu_1 - \mu_0|}{2}$ and $\delta = \frac{|\mu_1 - \mu_0|}{\sigma}$ represents the shift in standard deviation units and starting values $C_0^+ = C_0^- = 0$. If C_i^+ or C_i^- exceeds the decision interval H the process is considered out of control. Some general recommendations, based in studies about CUSUM ARL performance, for selecting H and K are as follows. Consider $H=h\sigma$ and $K=k\sigma$ where σ is the standard deviation of the sample variable which we use to form the CUSUM. If we use $h=4\ or\ 5$ and $k=1/2$, we will have a CUSUM with good ARL properties against a shift 1σ in the mean of the process (Montgomery (2001)).

Sometimes is better to standardize the variable x_i before any calculation because if we use many CUSUM charts we have the advantage of the same value for h and k that means they are independent of the scale.

(ii) The V-Mask CUSUM)

An alternative approach using CUSUM is the V-mask procedure. The V – mask procedure is applied to successive values of the CUSUM statistic $C_i=\sum_{j=1}^{i} y_i$ where y_i is the standardized observation $y_i=(x_i-\mu_0)/\sigma$. A V-mask is shown in figure 2.7. If any of the cumulative sums lie within the two "arms" the process is in control otherwise is out of control.

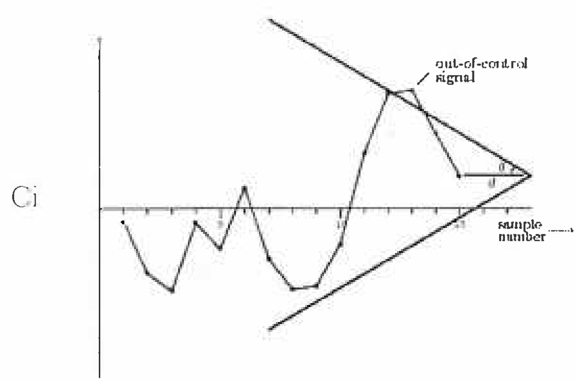


Figure 2.7 V – mask control scheme

As we mentioned before, Montgomery (2001) proposed to not use the V – mask procedure for the following reasons:

i. As the V – mask is a two sided scheme is not useful for one – sided process monitoring.

ii. It is difficult to determine how far backwards would extend the arms of V – mask in so doing interpretation difficult for the practitioner.

(iii) *The Fast Initial Response or Headstart Feature*

This procedure was devised by Lucas and Crosier (1982) to improve the sensitivity of a CUSUM at process start up. The Fast Initial Response (FIR) just sets the starting values C_0^+ and C_0^- equally to a nonzero value, usually $H/2$. If the process is in control the values of C_0^+ and C_0^- are soon unaffected by the headstart because the consecutive observations near the target value set the cusums rapidly to zero. However, if the process is out of control, the headstart will allow the CUSUM to detect the out of control points more quickly, resulting in shorter out of control values.

(iv) *Detecting small shifts – An example*

As we claimed above, in section 2.3.4, CUSUM control charts are essentially used when small shifts of the mean are of interest. In the following example we have 25 samples with sample size n = 5. In figure 2.8 we present the $\overline{x}$ and R and the CUSUM control chart. It is easy to see that CUSUM for samples from 4 through 10 lie outside the upper sigma limit indicating that process is out of control. On the other hand $\overline{x}$ chart failed to detect assignable cause for other samples except 5th sample

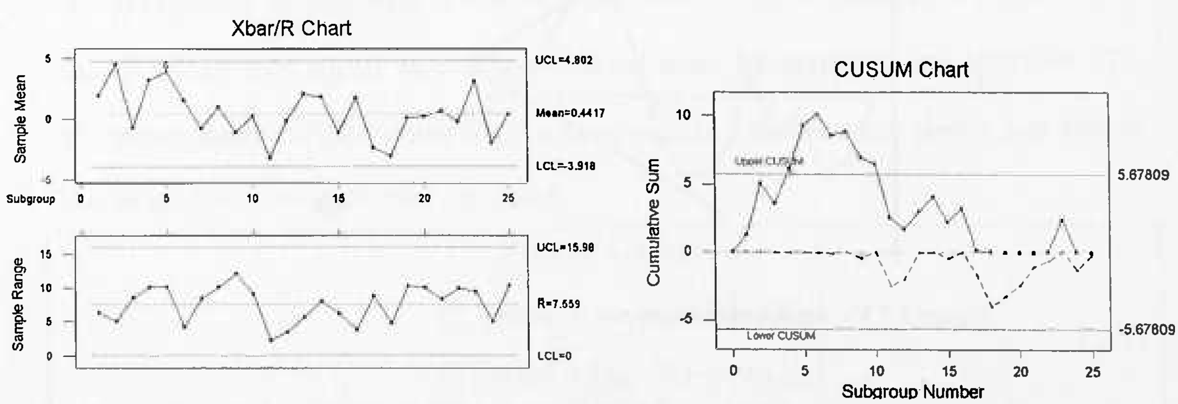


Figure 2.8 $\overline{x}$ *and* R **control chart vs CUSUM control chart**

2.3.4.2 The exponentially weighted moving average (EWMA) control chart

Another control chart is the exponentially weighted moving average (EWMA) control chart which is also effective when our aim is to detect small shifts of the target value. This chart first introduced by Roberts (1959) and is defined as follows:

EWMA statistic

$$z_i = \lambda x_i + (1-\lambda) z_{i-1} \text{ where } 0 < \lambda \leq 1 \text{ or}$$

$$z_i = \lambda \sum_{j=0}^{i-1} (1-\lambda)^j x_{i-j} + (1-\lambda)^i z_0 \qquad (2\text{-}10)$$

When λ increases, that means, approaches to unity more weight is given to recent data whereas when λ decreases more weight is given to older data. EWMA is very insensitive to the normality assumption so we can use it as control chart for individuals in a variety of applications. The control limits and center line are calculated as:

EWMA control chart

$$UCL = \mu_0 + L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}\left[1-(1-\lambda)2^i\right]}$$

$$Centerline = \mu_0 \qquad (2\text{-}11)$$

$$LCL = \mu_0 - L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}\left[1-(1-\lambda)2^i\right]}$$

As a rule of thumb smaller values of λ can detect smaller shifts and larger values of λ can detect larger shifts. Values of λ in the interval $[.05,.25]$ have been found that work practically well, (see, e.g., Montgomery (2001)). EWMA is better than CUSUM for large shifts, in particular when $\lambda \geq 0.1$. Alternative control procedures such as the combined Shewart – EWMA are effective both in large and small shifts. In the next figure an EWMA chart is represented for the data which we used from the previous example. The EWMA chart signals at the 5th sample.

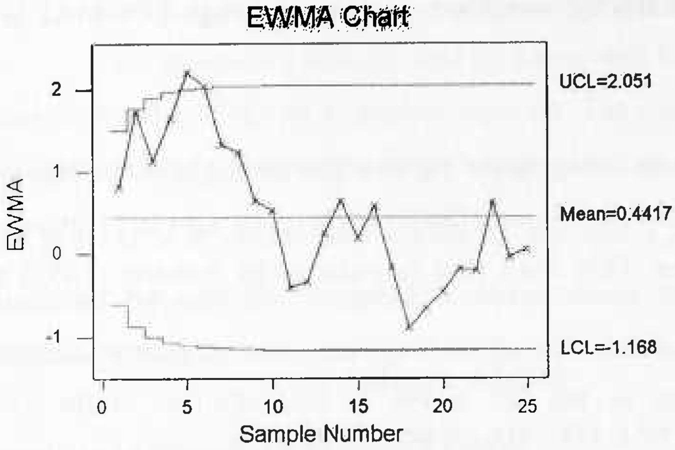


Figure 2.9 EWMA control chart with n=5, λ = 0.25 , L=3

2.4 Control Charts for Autocorrelated Data

A key assumption underlying the use of the traditional control charts in the univariate case is that the observations are independent through time. The assumption of uncorrelated observations is no longer valid in many manufacturing processes in the modern industries. Autocorrelation maybe the result of dynamics that is intrinsic of an industrial process. Usually, autocorrelation is observed in processes when observations are closely spaced in time. A rule of thumb is to sample from the process data stream less frequently considering the drawbacks. However, we make inefficient use of the available data and it takes longer to detect a process shift when it really exists (Montgomery, 2001). Control charts that have been designed under the assumption of independence of observations will be deeply affected with the existence of autocorrelation; e.g positive autocorrelation can produce negative bias in estimators of the process standard deviation with consequence much tighter control limits than desired. Tighter control limits can produce more often false alarms that is, non existent assignable causes with economic consequences in the industrial process.

Two general approaches have developed and studied in recent years. The first forecasts each observation from previous observations and calculates the forecast error (residual), after each observation is obtained. Next, these residuals are plotted with standard control charts. The second approach uses standard control charts that are based in original observations, but adjusts (widen) the control limits trying to reduce the false alarm rate. Furthermore, model free approaches developed and are working reasonably well in some cases.

2.4.1 Standard control charts based on residuals

As we stated above, the first approach for dealing with autocorrelated data applies an appropriate time series model assuming that this model will remove the autocorrelation from data and then apply traditional control charts to the residuals. Some techniques have been proposed in the literature and will be discussed in the next paragraphs.

2.4.1.1 The Moving center line EWMA control chart

An approximate procedure based on EWMA was proposed by Montgomery and Mastrangelo (1991). Assuming that we can model a process using an integrated moving average model IMA (1,1) which has the form $x_t - x_{t-1} = \varepsilon_t - \theta\varepsilon_{t-1}$ and a nonstationary behavior, that means the variable x_t drifts because there is not a fixed value of the process mean. Box et al. (1994) proved that the EWMA statistic with $\lambda = 1-\theta$ is the optimal one – step ahead prediction error for this process. Thus if $\hat{x}_{t+1}$ is the forecast for the observation in period $t+1$ then $\hat{x}_{t+1} = z_t$, where, z_t is the EWMA statistic. So, the sequence of $e_t = x_t - \hat{x}_{t-1}$ is independently and identically distributed (iid) with mean zero. Therefore, usual control charts could be applied to these one step ahead prediction errors. As a result we can approximate an appropriate time series model with EWMA.

Subsequently the usual three – sigma control limits using these errors satisfy the statement $P[-3\sigma \le e_t \le 3\sigma] = 0.9973$. Therefore, the control limits with centerline z_t for period $t+1$ are:

Moving center line EWMA control chart $UCL = z_t + 3\sigma$ $center\ line = z_t$ (2-12) $UCL = z_t - 3\sigma$

where σ^2 is estimated as dividing the sum of squared prediction errors for the optimal λ by n (number of observations).

Furthermore, Montgomery and Mastrangelo (1991) pointed out that it is possible to combine information about the state of statistical control and process dynamics on a single control chart.

2.4.1.2 CUSUM control chart for monitoring autocorrelated processes

Lu and Reynolds (2001) introduced CUSUM control charts based on the residuals or on observations in the case of processes which can be modeled as an AR(1) process plus an additional random error. The AR(1) process with an additional error is equivalent to an ARMA(1,1) process, (Box et al., 1994) that can be written as $(1-\phi)X_k = (1-\phi)\xi + (1-\theta B)\gamma_k$ where γ_k's are independent random variables with mean 0 and variance σ_γ^2. Here we present the residuals case. Assuming that we have a simple model $X_k = \mu_k + \varepsilon_k$ where μ_k is the random process mean at sampling time k. It is assumed that μ_k can be modeled as an AR(1) process, where $\mu_k = (1-\phi)\xi + \phi\mu_{k-1} + \alpha_k$ and ξ is the process mean, α_t is a random shift and $|\phi| < 1$. The authors, also assumed that $\alpha_k \sim N(0, \sigma_a^2)$ and are independent from ε_t .When the process follows the previous equations with $\xi = \xi_0$ where ξ_0 is the target value the process is in – control. When an assignable cause occur, this causes a shift ξ away from ξ_0. For this model the residual at observation k from the minimum square error forecast at observation $k-1$ is $e_t = X_t - \xi_0 - \phi(X_{t-1} - \xi_0) + \theta e_{t-1}$ where ϕ and θ are parameters of ARMA (1,1) model, (Box et al., 1994). Suppose that there is a step change in the process mean from ξ_0 to ξ_1 between the time $k = \tau - 1$ and τ, it can be proved that the asymptotic mean of residuals is $\frac{1-\phi}{1-\theta}(\xi_1 - \xi_0)$.These residuals are independent and normally distributed with variance σ_γ^2. A two sided CUSUM chart based on the residuals has the following statistics:

CUSUM for autocorrelated data

$$CR_k^+ = \max\left\{0, CR_k^+ + \left(e_k - r\sigma_\gamma\right)\right\}$$
$$CR_k^- = \min\left\{0, CR_k^- + \left(e_k - r\sigma_\gamma\right)\right\} \quad (2\text{-}13)$$

It is necessary to choose the value of r and therefore to determine the sensitivity of the chart in various shifts. When observations are independent a good choice of r is approximately the half of the standardized distance between the target and the shift when that is detected. When there is autocorrelation in the process the value of r will not be necessarily the same as with this when observations are independent. In the CUSUM of residuals, the residuals are independent but their mean is not constant after the shift in the process mean. Lu and Reynolds also proposed that the optimal value of r will depend primarily from the asymptotic mean of residuals (given in the previous paragraph) and the performance of the control chart for a range of shifts. Furthermore, they suggested that for relatively low levels of autocorrelation $0.4 < \phi < 0.8$ a value of r around 0.5 would give a reasonable good performance for a large range of shifts. In the case when autocorrelation gets higher, the behavior of the process gets closer to a random walk. With a random walk there is no tendency to return the process to the target, so the usual meaning of an in-control process is not valid. In this case, an adjustment mechanism could provide valuable results in reducing overall process variation. Process control methods based on adjustment mechanisms are labeled as Engineering Process Control (EPC) methods and will be presented in the next chapters.

2.4.1.3 Cuscore charts for process monitoring

The Cuscore statistic Q is a cumulative sum of Fisher's score statistic and is a function of the product of the residuals that is, how the residuals change with the parameter of the model. Considering that we have a time series model $y_t=\theta x_t+\alpha_t$ where the residuals α_t can be written as $\alpha_t=f(y_t,x_t,\theta)$ $t=1,2,...,n$ where y_t are the observed variable of a response, x_t is an independent variable and θ is some unknown parameter. If θ_o is the true value of the parameter it can be shown that $Q\approx(\hat{\theta}-\theta_o)\times\sum dt_o^2$ where $d_{to}=-\partial\alpha_t/\partial\theta\ \alpha_t$ and $\hat{\theta}$ is the maximum likelihood estimator of θ. The approximation is exact when the model is linear.

Ramirez (1998) introduced the Cuscore statistic in process monitoring. In this context the Shewart's model can be written as $y_t=\theta+\alpha_t$ where θ is the mean and α_t's is some random noise. This model transforms the data into noise according to $\alpha_t=y_t-\theta$ which has derivative with respect to θ constant and equal to 1. We can observe, that Cuscore for monitoring changes in the mean, is exactly the Cusum statistic that is, the sum of deviations from the mean value θ:

$$Q\approx\sum(y_t-\theta)$$

Ramirez (1998) proposed that Cuscore statistic can be used to check the assumption that the time series model which we applied to our data will also fit future observations reasonably well and its parameters will remain fairly constant. So, in the previous "Shewart" model after a suitable time series model has been fitted to our data the residuals would conform to the "Shewart" model with $\theta=0$. If this is not happen the time series model is no longer adequate for our data or an assignable cause is affecting the process.

2.4.2 Modified control charts

Due to the existence of autocorrelation in many industrial processes, control charts where designed under the assumption of independence of the successive observations suffer because of the high frequency of false alarms. In order to overcome these problems control charts based on the residuals have been proposed but they face the fact that behave worse when there are processes where the residuals are not independent and if the parameters, of the applied time series model, are unknown and must be estimated. In the next paragraphs other approaches will be presented that are based in the adjustment of control limits trying to decrease the false alarm rate.

2.4.2.1 A model free approach (The Batch means control chart)

Runger and Willemain (1995) proposed a control chart which is based on unweighted batch means (UBM) for monitoring an autocorrelated process. This approach derived from computer simulation models where high correlated data frequently occurs. This chart, gives equal weights to every point in the batch (every batch is composed of sequential observations). Using the subscript j symbolize the j – batch, the mean of this batch is, $\overline{x}_j = \frac{1}{b}\sum_{i=1}^{b} x_{(j-1)b+i}$ $j=1,2,\ldots$.

Although the UBM approach is a model free approach, there is not the necessity of an ARMA model, one has to determine an appropriate batch with size b which is harder than if he had selected an appropriate time series model.

Runger and Willemain suggested that the unweighted batch means could be plotted using a usual individuals control chart. Another one difference with residuals chart is that, it is as simple as a usual control chart. The same authors advocated that the batch size must be selected so as to reduce the lag autocorrelation of the batch means to approximately 0.1. They proposed to start with b=1 and to double b until the lag 1 autocorrelation is small enough accordingly the value 0.1.

In the figures 2.10, 2.11 we present an individual chart plot for 75 observations and a $\overline{x}$ plot of the batch means where b=3 respectively. As we see an assignable cause in the batch means plot was detected more quickly than in individuals plot.

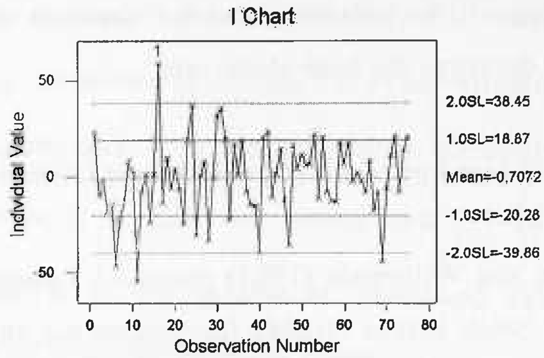


Figure 2.10 Plot of batch means using b = 3

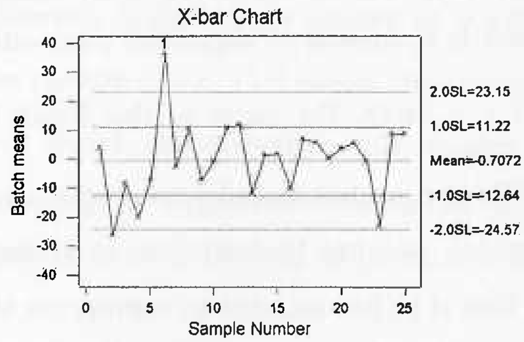


Figure 2.11 Xbar chart of residuals

Chapter 3

ARIMA and Transfer Function Models

3.1 Introduction

In this chapter ARIMA and transfer function models are discussed. ARIMA models are extensively used in industrial processes when autocorrelation exists. ARIMA models are appropriate for autocorrelated processes whose input stream is closely related. On the other hand, there are quality applications (e.g. chemical process industries, semiconductor manufacturing) which we refer to as "dynamic input processes" where the dynamic nature of an input creates an additional source of variability. For these applications it is imperative to model the dynamic relationship between process input and output. These models are called Transfer function models. A brief description why we need ARIMA models is given in 3.2 section. In 3.3 section a comparison of stationarity versus non stationarity is discussed while in 3.4 section basic properties of time series are presented. In 3.5 section we discuss the most known stationary ARMA models while 3.6 section is concerned with the modeling of these models. Furthermore, in 3.7 section transfer function models are stated under the view of industrial the processes. Finally in section 3.8 a basic presentation of state space models is discussed.

3.2 The necessity of ARIMA models

In the previous chapter (section 2.4) we analyzed the situations where the data are not independent but are presenting an autocorrelation structure. In this case Shewart control charts are not appropriate and an effort is made to model this dependency trying to find a time series model that fits the data. Therefore we use that model to remove autocorrelation and apply control charts to residuals. Some of these models are AR(p), MA(q), ARMA(p,q) , IMA(p,q), etc. All of these are belonging to a wider class of models, stationary and non – stationary, which are called ARIMA models.

3.3 Stationary and Non – Stationary Time Series Models.

An important class of stochastic models which we use for describing time series, consists of the stationary models. Stationarity of a process means that the process remains in equilibrium about a constant mean level (Box et al., 1994). More formally a time series is said to be strictly stationary if the joint distribution of $X_1, X_2, \ldots, X_n$ is the same with the joint distribution of $X_{1+h}, X_{2+h}, \ldots, X_{n+h}$ for all integers h and $n > 0$. A less restrictive requirement called weakly stationarity of order f if the moments up to order f depend only on time differences, that is, if the following conditions are valid:

Weakly stationarity	
$E(X_t) = \mu_X(t)$ is independent of t $Cov(X_{t+h}, X_t) = \gamma_X(h)$ is independent of t for each t	(3-1)

where $E(X_t)$ and $Cov(X_{t+h}, X_t)$ are the mean and autocovariance function (ACVF) of (X_t) respectively and $\gamma_X(h)$ is the value of the autocovariance function at lag h.

In the previous paragraph we defined the stationarity of a time series model. However, is a usual phenomenon in industry that many processes are better represented as non - stationary that is, as no having a constant mean level over time. In the figure 3.1 and 3.2 are given examples of a non stationary and a stationary process respectively:

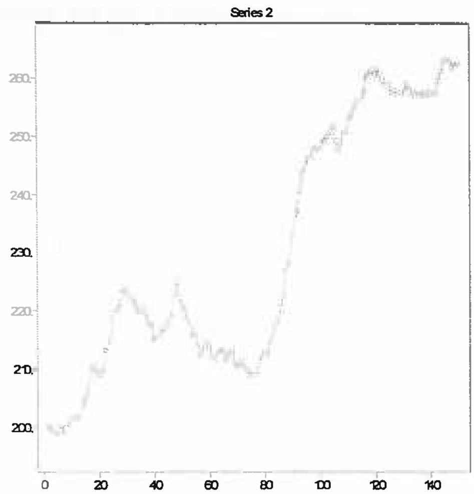


Figure 3.1 **Plot of a non stationary time series** data

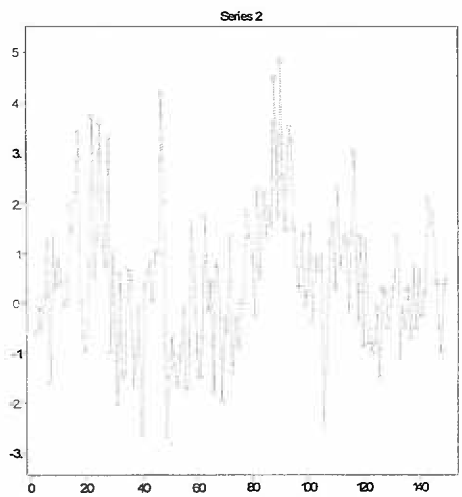


Figure **3.2 Plot of a stationary** time series data

3.4 Basic Properties of **Time series**

Let (X_t) be a stationary time series. The autocorrelation function (ACF) of (X_t) is:

$$\rho_X(h)=\frac{\gamma_X(h)}{\gamma_X(0)} \tag{3-2}$$

where $\gamma_X(0)=Var(X_t)$.

The partial autocorrelation function (PACF) of this time series (X_t) is the function $\alpha(.)$ which is defined by the equations:

$$\alpha(0)=1$$
$$\alpha(h)=\phi_{hh}, \quad h \geq 1 \tag{3-3}$$

where ϕ_{hh} is the component of $\phi_h = \Gamma_h^{-1}\gamma_h\Gamma_h$ and Γ_h is the covariance matrix.

3.5 Stationary ARMA (p,q) models

In a stationary process the ARMA(p,q) models are extensively used where we try to eliminate autocorrelation and to apply control charts to residuals. In the next paragraphs, briefly, ARMA(p,q) will be discussed, as well as special cases of them as AR(p) and MA(q) models.

3.5.1 Autoregressive Models

A very useful model which we usually use trying to remove autocorrelation from a process is the autoregressive model. In this model the current value of the process is expressed as a finite linear aggregate of previous values and a random "shock" ε_t. These random shocks usually assumed to be Normal with mean zero and variance σ_ε^2. A sequence of ε_t is called white noise. The stationary time series (X_t) is called autoregressive process of order p, AR(p), if it satisfies the equation:

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t$$

where $Z_t \sim N(0,\sigma^2)$ and Z_t are uncorrelated with X_s $s<t$ and $\phi_1,\ldots,\phi_p$ are constants (3-4)

Autoregressive models can be stationary or non stationary. Stationarity exists when all the roots of the autoregressive operator, $\Phi(B)=1-\phi_1 B - \ldots - \phi_p B^p$, is considered as polynomial in B of degree p (B is the backward shift operator), lie outside the unit circle. Some basic properties of the AR(p) models are:

- The ACVF of an AR(p) process $(\gamma_X(h))$ can take any value until p. When $h > p$, it declines geometrically.

- The ACF $(\rho_X(h))$ will die down, sometimes with an oscillatory manner. When it is close to a stationarity boundary, the rate at which the ACF decays, may be very slow (e.g., an AR(1) model with $\varphi = .999$). This may be an indication that the process generating the data is nonstationary.

- The PACF value of an AR(p) process is zero after p.

3.5.2 Moving average models

In this model the current value of the process is expressed as a finite linear aggregate of past innovations Z_t. Thus the stationary time series (X_t) is called moving average process of order q, MA(q), if satisfies the equations:

$$X_t = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}$$

$$\text{where } Z_t \sim WN(0,\sigma^2) \text{ and } \theta_1,\ldots,\theta_q \text{ are constants} \qquad (3\text{-}5)$$

Some basic properties of the MA(q) models are:

- The ACVF of a MA(q) process $(\gamma_X(h))$ is zero for $h > q$.

- The ACF $(\rho_X(h))$ can take any value up to q. When $h > q$, then $\rho_X(h) = 0$.

- The PACF value of a MA(q) process declines geometrically up to q, sometimes with an oscillatory manner.

3.5.3 Mixed Autoregressive - Moving average models

When we want to achieve a greater flexibility and accuracy of a time series model, sometimes, is better to include both autoregressive and moving average terms in the model. This leads to the mixed autoregressive – moving average ARMA (p,q) model where,

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}$$

where $Z_t \sim WN(0,\sigma^2)$ and the polynomials

$$1-\phi_1 z - \ldots - \phi_p z^p \text{ and } 1+\theta_1 z + \ldots + \theta_q z^q \text{ have no common factors} \quad (3\text{-}6)$$

The ACVF , ACF and PACF of ARMA(p,q) processes is a mixture of all aforementioned for AR(p) and MA(q) processes.

3.5.3.1 Order selection

It is not beneficial from a forecasting point of view to choose the parameters of the model (p and q) subjectively large. If we fit a very high order model this lead us to a small white noise variance but if forecasting is our major concern the mean squared error of the forecast will depend from the white noise variance and from the errors arising from estimation of the parameters of the model.

The most common criteria which are used in order to achieve a reasonable choice of p and q are the final prediction error (FPE) criterion and the Akaike (AIC) criterion for AR processes and the AICC criterion for ARMA processes which are described as follows:

3.5.3.2 The FPE and AIC and AICC criteria

The FPE criterion was developed by Akaike(1969) in order to select the appropriate order of an AR process to fit to a time series $\{X_1,\ldots,X_n\}$. The idea is to choose the model (X_t) in such a way that minimizes the one step mean square error when the model fitted to X_t is used to predict an independent realization (Y_t) of the same process that generated X_t. So a good choice of p is this, that minimizes the quantity $FPE_p = \hat{\sigma}^2 \frac{n+p}{n-p}$ where $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2. Correspondingly, the AIC criterion consists of minimizing the quantity $\ln(\sigma_p^2)+\frac{2p}{T}$ which is asymptotically equivalent with the FPE criterion (Brockwell and Davis, 1996). Similarly with the previous criteria the bias - corrected Akaike criterion (AICC) for fixed p,q and ϕ_p,θ_q minimizes the quantity:

$$-2\ln L_X\left(\phi_p,\theta_q,S\left(\phi_p,\theta_q\right)/n+2(p+q+1)n/(n-p-q-2)\right)$$

3.5.3.3 Estimating the coefficients of an ARMA(p,q) model

The determination of an appropriate ARMA(p,q) model involves a number of interrelated problems. One of them is the estimation of the coefficients $(\phi_i\, i=1,\ldots,p),(\theta_i\, i=1,\ldots,q)$ and the white noise variance σ^2. Assuming that the data are "mean - corrected" that is, the sample mean have been subtracted a zero mean ARMA model is,

$$\phi(B)X_t=\theta(B)Z_t\ \ Z_t\sim WN\left(0,\sigma^2\right).$$

When p and q are known, good estimators of ϕ and θ can be found maximizing the likelihood with respect to $p+q+1$ parameters $(\phi_1,\ldots,\phi_p),(\theta_1,\ldots,\theta_q)$ and σ^2. The maximum likelihood estimation is not linear; therefore, good initial values are required so as, the algorithm to converge as fast as possible.

For pure autoregressive models Yule – Walker and Burg's algorithm are provided while for models with $q>0$ innovations algorithm and Hannan – Rissannen algorithm are more preferable (Brockwell and Davis, 1996).

3.5.3.4 Forecasting ARMA processes

Forecasting at time t, using the available observations from a time series, for a future time $t+1$ can increase considerably the control and optimization of an industrial process. A forecast is characterized by its origin and lead time. The origin is the time from which the forecast is made (usually the last observation in a realization) and the lead time is the number of steps ahead that the series is forecast. The most common criterion for computing forecasts is the mean square error (MSE) and our aim is to minimize it.

Assuming that (X_t) is a stationary time series model with known mean and autocovariance function in terms of $X_1,...,X_T$ observed values. Our aim is to find the best linear predictor P_n which forecasts X_{T+h} with minimum mean squared error (MMSE). This predictor is,

$$P_n X_{T+h} = a_0 + a_1 X_T + ... + a_n X_{T+1-n}$$

The best $a_0, a_1, ..., a_n$ are depending from the $\gamma(0), \gamma(1), ..., \gamma(n)$ assumed known.

The MMSE is given by the following equation:

$$E\left(X_{T+h} - P_n X_{T+h}\right)^2 = \gamma(0) - \gamma(h)'\Gamma_n \gamma(h)$$

Since $\gamma(h) \to 0$ for $h \to \infty$ then $MMSE \approx \gamma(0)$ that means when "past" is so far that correlations are negligible the prediction is as bad as if "past" was unknown.

3.6. Modelling non stationary models (ARIMA) models

The ARIMA models are appropriate for autocorrelated processes whose input streams are closely related, in other words the autocovariance function is decreased very slowly (West et al., 2002). The definition of the ARIMA model is:

X_t is an ARIMA(p,d,q) model if $Y_t = (1-B)^d X_t$ where d is a non negative integer and Y_t is a causal ARMA(p,q) process (3-7)

The previous definition equivalently means that X_t satisfies the next difference equation:

$$\phi(B)(1-B)^d X_t = \theta(B) Z_t$$

where $\phi(B)$ is a stationary autoregressive operator and $Z_t \sim WN\left(0, \sigma^2\right)$.

The last definition implies that the process can be obtained by summing ("integrating") the stationary process d times (Box et al., 1994).

3.6.1 Special cases of the ARIMA model

The most frequently ARIMA models are the following:

(i) The $(0,1,1)$ process: $(1-B)X_t=\theta(B)Z_t$ where $\theta(B)=(1-\theta_1 B)$

(ii) The $(0,2,2)$ process: $(1-B)^2 X_t=\theta(B)Z_t$ where $\theta(B)=(1-\theta_1 B-\theta_1 B^2)$

(iii) The $(1,1,1)$ process: $\phi(B)(1-B)X_t=\theta(B)Z_t$ where $\phi(B)=(1-\phi_1 B)$ and $\theta(B)=(1-\theta_1 B-\theta_1 B^2)$.

3.6.2 Variance stationarity and transformations

When the time series data exhibit a degree of variability which changes especially, when this variability is increased, it is an indication that these series must be transformed to achieve stabilization of the variance before model these data. A usual transformation is taking the natural logarithms of the series. This transformation is appropriate in case where the variance of the series is proportional to the mean. Another one transformation widely used is the family of Box - Cox transformation proposed by Box and Cox (1964) which is:

$$Y_t=\begin{cases}\dfrac{X_t^{\lambda}-1}{\lambda}, & \lambda\neq 0\\ \ln\lambda, & \lambda=0\end{cases} \tag{3-8}$$

where Y_t is the transformed time series and λ is the transformation parameter.

As an example we have 144 observations of international airline passengers from 1949 to 1960 (monthly). The dataset adapted from Box et al., (1994). In figure 3.3 we can see the time series plot of the original time series data as well as the plots of ACF and PACF. In figure 3.4 the plots show how the increased variability is reduced taking natural logarithms of the original series.

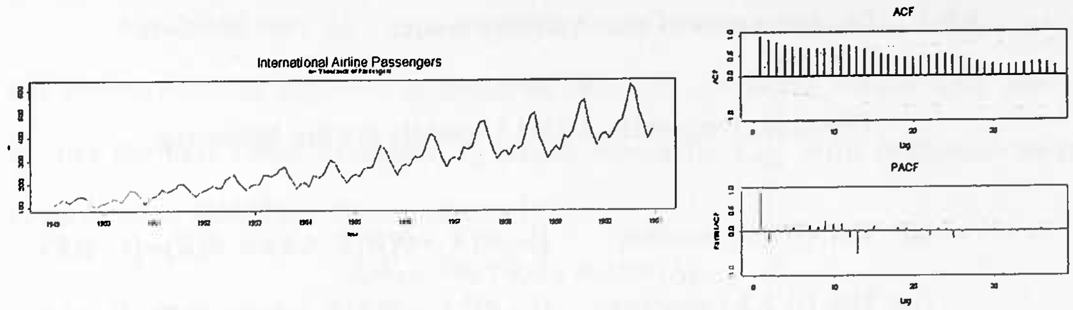


Figure 3.3 Plot of the original series of airlines passengers with the ACF and PACF plots

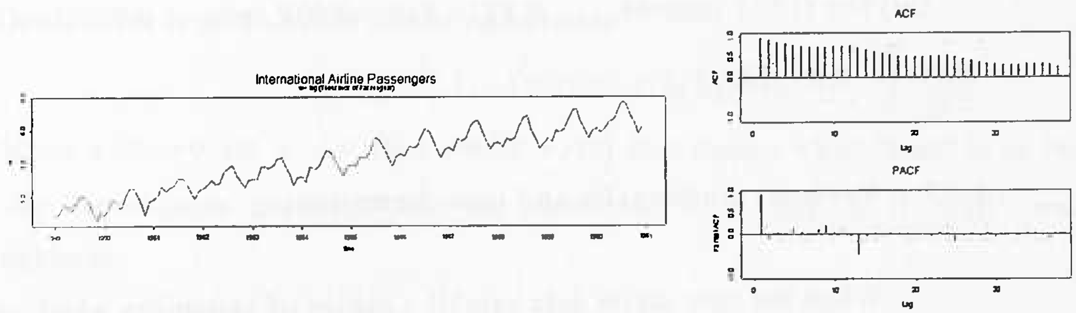


Figure 3.4 Plot of transformed series of airlines passengers with the ACF and PACF plots

If we inspect a plot of the time series and we detect trend and seasonality then there are two methods to confront with this problem:

(i) Classical Decomposition

A common method is to decompose the series into three components: a trend, a seasonal component and a random noise. This method extensively discussed by Brockwell and Davis (1996) and criticised by Box et al. (1994) because there are cases where it produces misleading results. This method proposes to estimate and extract the deterministic components (trend and seasonality) hoping that the noise component will turn out to be a stationary time series. After that, we have to find a satisfactory model in order to analyze its properties and use it in conjunction with the trend and seasonal component for prediction and simulation purposes.

(ii) Differencing

This approach, developed by Box and Jenkins (1976), based on applying repeatedly appropriate differences to the original series until the differenced observations resemble with a stationary time series. Then we can model, analyze and predict this stationary series and hence the original process.

Continuing the previous example we can observe from plot (figure 3.4) that transformed data reveals trend and stationarity. Figure 3.5 gives the plot of these transformed data and figure 3.6 shows the plot when we applied first differencing to eliminate trend and first differencing with period 12 to eliminate seasonality. As we can see from the plot the process appears to be stationary.

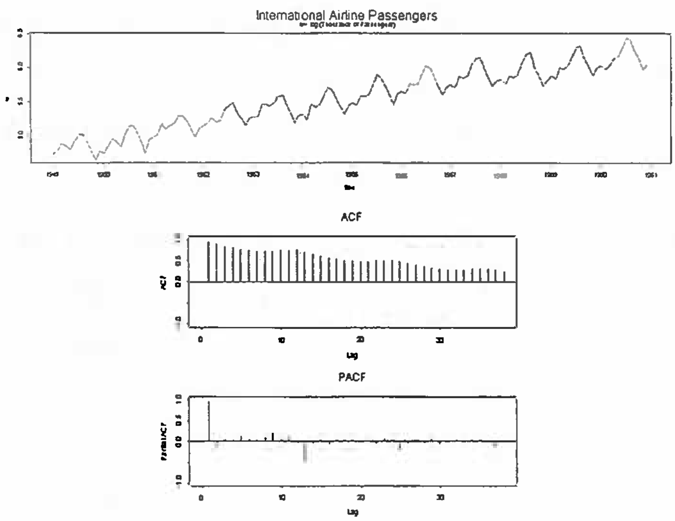


Figure 3.5 Plot of the transformed series of airlines passengers with the ACF and PACF plots exhibits trend and seasonality

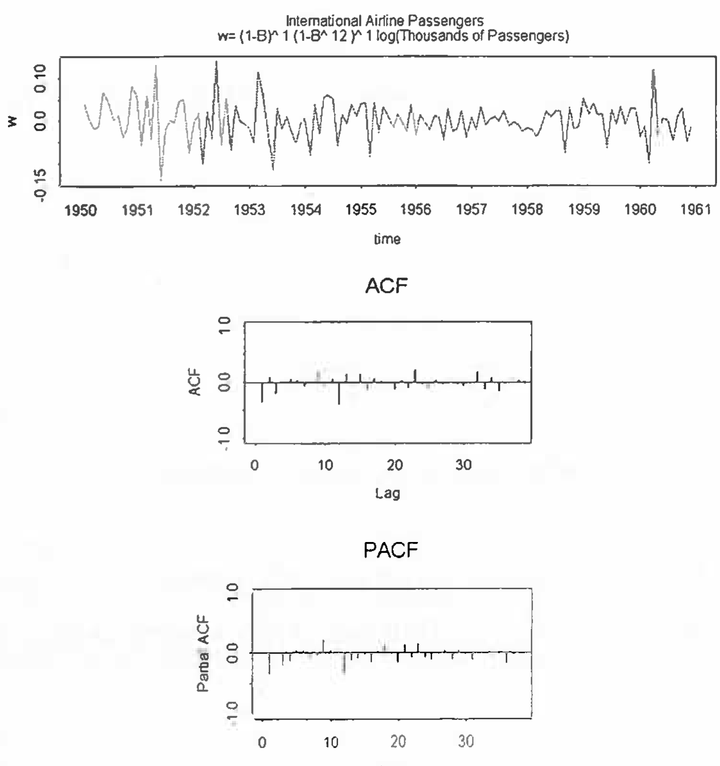


Figure 3.6 Plot of the transformed series of airlines passengers with the ACF and PACF plots seems to be stationary

3.7 Transfer function models

There are industries including chemical processes, semiconductor manufacturing, where ARIMA models are not appropriate to monitor the autocorrelation of observations. The input of these processes namely "dynamic input processes" (West et al., 2002) are highly affected by an additional source of variability. For these processes, other models are suitable to model the relationship between input and output. We assume that X measures the level of an input to a system (e.g the concentration of some ingredient in the feed of a chemical process). Furthermore, we also assume that the level of X, influences the level of a system output Y. Because of the inertia of the system, a change in the level of X will not immediately affect the output, but will produce a *delay* response to Y until coming in balance with the new level. We call this change *dynamic response* and the model that describes it *transfer function* model. However, in all the conditions of an industrial process, independently how well these conditions are controlled, there are influences other than X that will affect Y. All of these uncontrollable effects we call it *noise* or *disturbance*.

3.7.1 The importance of transfer function models in industrial processes

Statistical methods are used widely for estimating transfer function models that taking account of the noise in the system as we described it before. In the next paragraph some basic meanings of the terminology related to transfer function models will be given.

3.7.2 Basic terminology in transfer functions

In the previous paragraph we presented a dynamic system of an industrial process. Now we illustrate it in a scheme as this in figure 3.7.

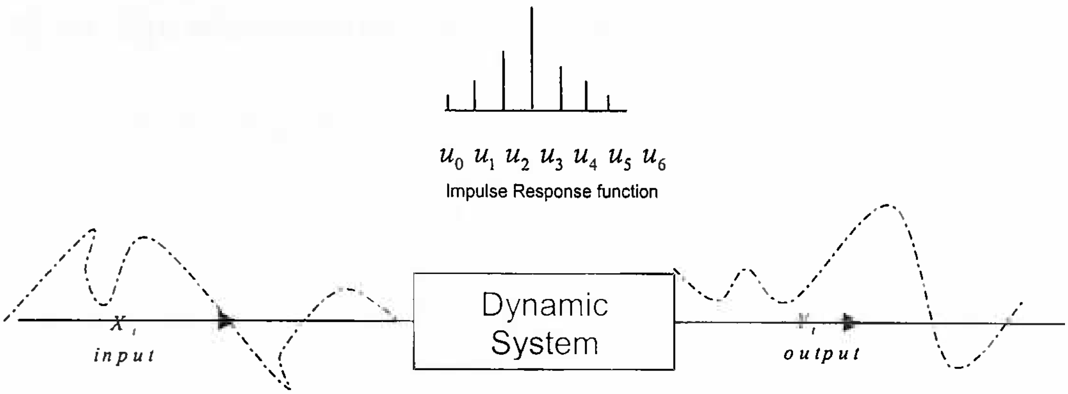


Figure 3.7 Input and Output from a dynamic system

Steady state level: We refer to a steady state level of the output when from time to time the input is held at some fixed value. That is, we mean that the value $Y_\infty(X)$, discrete output, from a stable system comes in balance when the input is held at the fixed level X.

Steady state gain: Many times the relationship between $Y_\infty(X)$ and X is approximately linear. Hence, we can write the steady – state relationship as $Y_\infty = gX$ where g is called steady – state gain.

Transfer function of the filter: Suppose that the input is being varied and X_t and Y_t represent deviations at time t the inertia of the system can be represented by a linear filter as $Y_t = u(B)X_t$. The operator $u(B)$ is called transfer function of the filter.

Impulse response function: The weights $u_0, u_1, \ldots$ are called the impulse response function of the system. That is, u_j can be regarded as the response (output) at times $j \geq 0$ to a unit *pulse* input at time 0 such that $X_t = 1$ if $t = 0$ and $X_t = 0$ otherwise. Suppose that the deviations are in balance initially and at times $t = 1, t = 2, \ldots$produce impulse response patterns of the deviations in the output where add together to produce the overall output response.

Stability: If the infinite series $u_0 + u_1(B) + u_2(B)^2 + \ldots$ converges for $|B| \le 1$ or in other words u_j are absolutely summable, so that $\sum_{j=0}^{\infty} |u_j| < \infty$ then the system is said to be *stable*. It can be proved that $\sum_{j=0}^{\infty} |u_j| = g$ (Box et al., 1994).

Parsimony: Unrestrained use of the parameters u_j, at the estimation stage, could lead to inaccurate and unstable estimation of the transfer function.

Forcing function: In general, the function that is responsible for driving the dynamic system is called forcing function.

Step (unit) function: Suppose that there is a forcing function which was at a steady state level of zero and changed instantaneously to a steady level of unity; this function is called step function.

Step response: The response of a system after applied a step function is called step response.

Pure unit delay or *responsive model*: In any transfer function model the dynamic behaviour of Y_t is due only to disturbance dynamics (noise) and not to process dynamics (there is not dynamical relationship between X and Y) then, this model is called pure unit delay or responsive model.

3.7.3 Continuous and discrete Dynamic models

Supposing that we have a pair of observations (X_t, Y_t) at equispaced intervals of time, of an input X and an output Y from a dynamic system as this presented in figure 3.7. There are cases where both X and Y are continuous but are observed only at discrete times. We consider that a discrete model can be fitted in these data. Where we have continuous and discrete systems it is used the basic sampling interval as the unit of time.

3.7.3.1 Continuous dynamic models

Suppose that we have a continuous dynamic system with a steady – state relationship that is, $Y_\infty = g_1 X$. To relate output and input we use the following differential equation

$$\frac{dY}{dt} = \frac{1}{T_1}\left[g_1 X(t) - Y(t)\right] \; or \; \left(1 + T_1 D\right) Y_1(t) = g_1 X(t) \qquad (3\text{-}9)$$

where, T_1 is constant (called time constant) and $D = d/dt$. The solution of the differential equation (3-9) is, $Y_1(t) = \int_0^\infty u(v) X(t-v) du$ and $u(v) = g_1 T_1^{-1} e^{-v/T_1}$.

Suppose that $X(t)=0$ and abruptly raised to a level $X(t)=1$. The step response of the system solving the differential equation with a unit step input is $Y_1(t) = g_1\left(1 - e^{-t/T_1}\right)$. Sometimes there is an initial period of pure delay or *dead time* before the response of a system after a given input change begins to take effect. In these cases the differential equation (3-9) is modified as

$$\left(1 + T_1 D\right) Y_1(t) = g_1 X(t - \tau) \qquad (3\text{-}10)$$

3.7.3.2 Discrete **dynamic models**

In the case of a discrete dynamic system we usually represent it as

$$\left(1 - \delta_1 B - \delta_2 B^2 - \ldots - \delta_r B^r\right) = \left(\omega_0 - \omega_1 B - \omega_2 B^2 - \ldots - \omega_r B^r\right) X_{t-b}$$

or as

$$\delta(B) Y_t = \omega(B)^b X_t \qquad (3\text{-}11)$$

where the transfer function is $u(B) = \delta^{-1}(B)\Omega(B)$ where $\Omega(B) = \omega(B)^b$ and b represents the delay in periods. Thus the transfer function is the ratio of two polynomials in B.

Recalling from paragraph 3.6, ARIMA models where $\phi(B)X_t=\theta(B)Z_t$, can be written as $X_t=\phi^{-1}(B)\theta(B)Z_t$. Therefore a time series model can be represented as an output from a dynamic system to which the input is white noise and the transfer function is the ratio of two polynomials in B.

Stability for a discrete transfer function model is the same as stationarity for ARMA time series models. Thus, for stability we require that the roots of $\delta(B)=0$ lie outside the unit circle, that is,

for first order model we require that $-1<\delta_1<1$ and

for second order model we require that $\begin{cases}\delta_1+\delta_2<1\\ \delta_2-\delta_1<1\\ -1<\delta_2<1\end{cases}$

3.7.4 Transfer function models with added noise

In reality transfer function models, as we stated in 3.7.1, will not follow the pattern as aforementioned. A disturbance might be originate at any point in the system but usually we consider it on the output Y. Assuming that the noise N_t is independent the level of X and is additive with respect to the influence of X then, $Y_t=\delta^{-1}(B)\omega(B)X_{t-b}+N_t$ and if we represent it as ARIMA(p,d,q) model then, $N_t=\phi^{-1}(B)\theta(B)Z_t$ where $Z_t\sim WN\left(0,\sigma^2\right)$ and finally as:

$$\boxed{Y_t=\delta^{-1}(B)\omega(B)X_{t-b}+\phi^{-1}(B)\theta(B)Z_t \qquad (3\text{-}12)}$$

In the next paragraphs will be discussed methods for identifying fitting and checking transfer function models of the form (3-12).

3.7.5 Determination of the Transfer Function Models

In order one to determine a transfer function model has to follow the following steps:

a) Estimation of the cross correlation function $r_{xy}(k)$ between the input and the output of a discrete dynamic system,

b) Identification of a transfer function model,

c) Fitting and checking the model.

(a) Cross correlation function

Assuming that a bivariate time series $\mathbf{X}_t$ is a series of two dimensional vectors $(X_{t1}, X_{t2})'$ observed at times t where t=1,2,3,... with mean vector

$$\boldsymbol{\mu}_t = E\mathbf{X}_t = \begin{bmatrix} EX_{t1} \\ EX_{t2} \end{bmatrix} \text{ and}$$

covariance matrices

$$\Gamma(t+h,t) = Cov(\mathbf{X}_{t+h}, \mathbf{X}_t) = \begin{bmatrix} \operatorname{cov}(X_{t+h.1}, t_1) & \operatorname{cov}(X_{t+h.1}, t_2) \\ \operatorname{cov}(X_{t+h.2}, t_1) & \operatorname{cov}(X_{t+h.2}, t_2) \end{bmatrix}$$

this series is said to be stationary if the moments $\boldsymbol{\mu}_t$ and $\Gamma(t+h,t)$ are both independent of t.

Suppose that we want to describe an input time series X_t and the corresponding output time series Y_t from a dynamic system. We assume that data are a pair of discrete time series generated by discrete bivariate process and at times $t+h, t+2h, ..., t+Nh$ are represented by $(X_1,Y_1),(X_2,Y_2),...,(X_N,Y_N)$. In general this bivariate process is not stationary. Therefore, appropriate differences (x_t, y_t) could be applied, where $x_t = (1-B)^d X_t$ and $y_t = (1-B)^d Y_t$ are transformed to be stationary. The cross covariance coefficients between x and y at lag $k>0$ is:

$$\gamma_{xy}(k) = E\left[(x_t - \mu_x)(y_{t+k} - \mu_y)\right] \quad k=0,1,2...$$

The cross covariance function describes the relationship between future values of y and current values of x and the cross correlation coefficient at lag k is,

Cross correlation coefficient	
$$\rho_{xy}(k) = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y}$$	(3-13)

The cross correlation function (CCF) gives the correlation between x_t and y_{t+k}

(b) Identification of a transfer function model

Suppose that a transfer model can be written as $Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + N_t$ where $\delta(B) = 1 - \delta_1(B) - \ldots - \delta_r(B)^r$ and $\omega(B) = \omega_0 - \omega_1(B) - \ldots - \omega_s(B)^s$. The identification includes the following

1) Estimate $\hat{u}_j$ (impulse response weights) using cross correlation function.

2) Use these estimates to guess the orders of r and s and the delay parameter b.

3) Substitute the estimates $\hat{u}_j$ in the equation $1 - \delta_1(B) - \ldots - \delta_r(B)^r (u_0 + u_1(B) + \ldots) = (\omega_0 + \omega_1(B) + \ldots + \omega_s(B)^s) B^b$ to obtain initial estimates of δ and ω.

The identification process would be simplified if the input to the system is white noise. When the original input is not white noise, simplification is possible by "prewhitening". This technique is described extensively in Box et al. (1994).

(c) Fitting and checking the model

Generally the same comments should be made as with ARIMA(p,d,q) models. An additional point is that a transfer function models includes, at least one, input variable. This variable could be correlated with the noise component N_t. Therefore, it is essential to prewhiten the input to avoid the correlation between the input and the noise component.

As an example (Box et al., 1994) we have 296 successive pairs of continuous observations, measured at 9 second intervals, from a gas furnace where air and methane combined to form a mixture of gases containing CO_2. The air feed was held constant but the methane feed rate could be varied. The resulting CO_2 concentration was measured to provide information about the dynamics of the system. In the figures 3.8 and 3.9 we see the plots of cross correlation before and after prewhitening (an AR(5) model applied to the input) respectively.

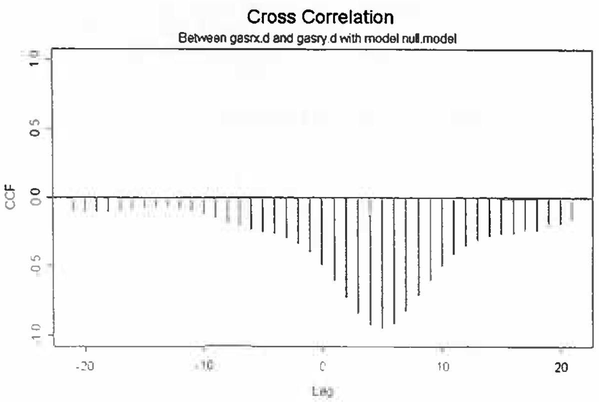


Figure 3.8 Cross correlation plot before prewhiten the input

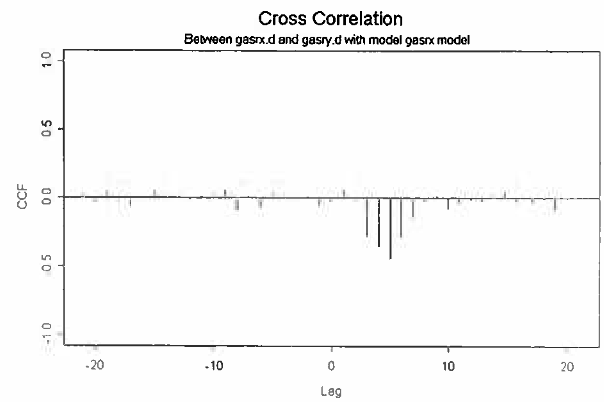


Figure 3.9 Cross correlation plot after prewhiten the input

Furthermore, serious model inadequacy can usually be detected by examining:

1) The autocorrelation function of the residuals from the fitted model. If the autocorrelation function shows correlation pattern this indicates model inadequacy.

2) Certain cross correlation functions involving input and residuals. In particular, cross correlation function between prewhitened input and the residuals.

3.7.6 Forecasting using leading indicators

Prediction of a time series, say Y_t, can be substantially improved by using information from an associated series X_t. This is valid if changes in Y tend to be expected by changes in X. In this case X is called "leading indicator". We build a transfer function noise model as stated in (3.7.3) by $Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + \phi^{-1}(B)\theta(B)Z_t$. Also, we assume that an adequate model for the leading series X_t is $X_t = \phi^{-1}(B)(1-B)^d \theta(B)W_t$ were $W_t \sim WN(0,\sigma_\eta^2)$. Our aim is to find the linear combination of X_t and Y_t that predicts Y_{t+h} with minimum squared error.

We rewrite $Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + \phi^{-1}(B)\theta(B)Z_t$ as $Y_t = \nu(B)W_t + \psi(B)Z_t$ where $\nu(B) = \delta^{-1}(B)\omega(B)B^b\phi^{-1}(B)(1-B)^d\theta(B)$ and $\psi(B) = \phi^{-1}(B)\theta(B)$. The prediction $\hat{Y}_t(h)$ of Y_{t+h} can be written as $\hat{Y}_t(h)=\sum_{j=0}^{\infty} v^0_{h+j} W_{t-j}+\sum_{j=0}^{\infty} \psi^0_{h+j} Z_{t-j}$. It can be proved that the minimum squared error is given by the following equation:

Minimum Squared Error (MMSE) prediction

$$E\left[Y_{t+h}-\hat{Y}_t(h)\right]^2 = \sigma_Z^2 \sum_{j=0}^{h-1} v_j^2 + \sigma_W^2 \sum_{j=0}^{h-1} \psi_j^2 \tag{3-14}$$

3.7.7 A Detailed example in Transfer Function Modelling

As an example (adapted from Brockwell and Davis, 1996) we consider the leading indicator (U_t, $t=1,2,\ldots 149$) and sales (V_t, $t=1,2,\ldots 149$) time series data, given by Box and Jenkins (1994) where the first one is the input and the second the output series respectively. The graphs of the two series and their sample ACF suggest that both series are non stationary (figure 3.10 and 3.11).

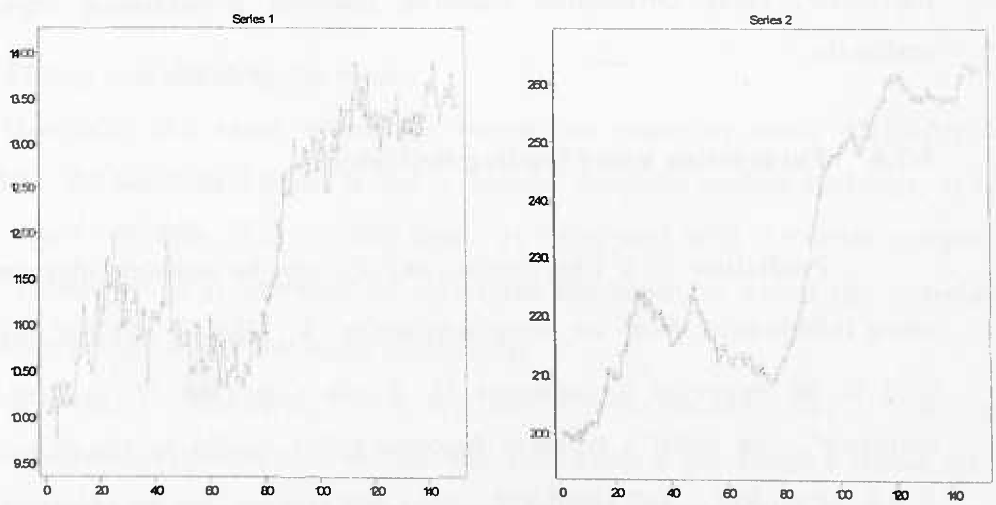


Figure 3.10 The leading indicator and sales data time serie graphs

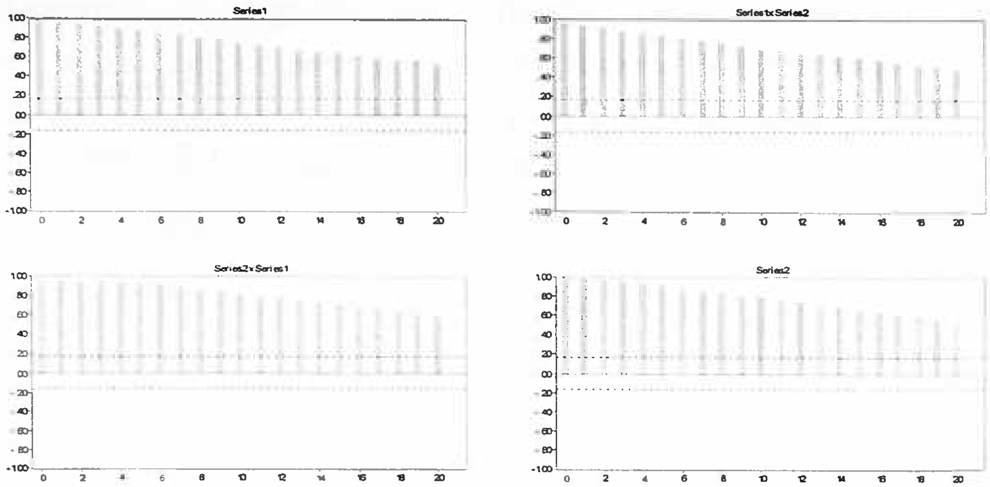


Figure 3.11 The SACF plot of leading indicator and sales data time series

(i) Outline

Before we set up a transfer function model relating to the previous series we start with differencing and mean correction to generate transformed X_t and Y_t, input and output series respectively, which can be modeled as stationary time series. Our goal is to fit an appropriate transfer function model of the form:

$$Y_t = \sum_{j=0}^{\infty} \tau_j X(\tau - j) + N(t)$$

where $N(t)$ is an ARMA process uncorrelated with $X(t)$,

$$\phi_N(B)N(t) = \theta_N(B)W(t), \; W(t) \sim WN(0, \sigma_W^2)$$

and the transfer function $T(B)$ has the form,

$$T(B) = \sum_{j=0}^{\infty} \tau(j)B^j = \frac{B^d\left(w_0 + w_1 B + \ldots + w_q B^q\right)}{1 - v_1 B - \ldots - v_p B^p}$$

and the input process X_t is described as an ARMA process

$$\phi_X(B)X(t) = \theta_X(B)Z(t), \; Z(t) \sim WN(0, \sigma_Z^2)$$

(ii) Modelling X_t

The parameters in the last three equations will be estimated from the given observations from X_t and Y_t. Furthermore, cross correlations can be computed for model checking. Moreover the AIC value is computed for model comparisons and forecasts are computed from the fitted model. Examining the plot of the sample ACF and PACF (figure 3.12) suggests that an appropriate time series model for the X_t is a MA(1) model,

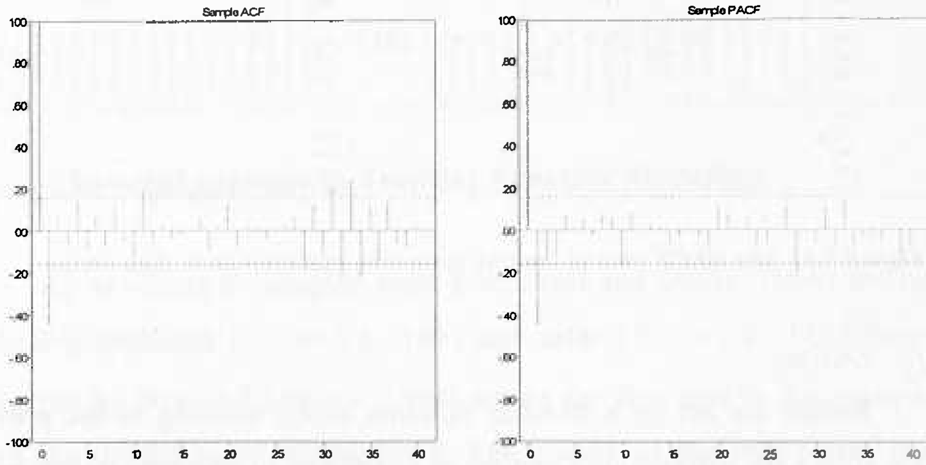


Figure 3.12 Sample ACF of the X_t time series

and using maximum likelihood estimation is formulated as follows:

$$X(t)=Z(t)-0.4744Z(t-1),\ \ Z(t)\sim WN(0,0.7794)$$

(iii) Preliminary coefficient estimates

The sample cross correlations graph is the following:

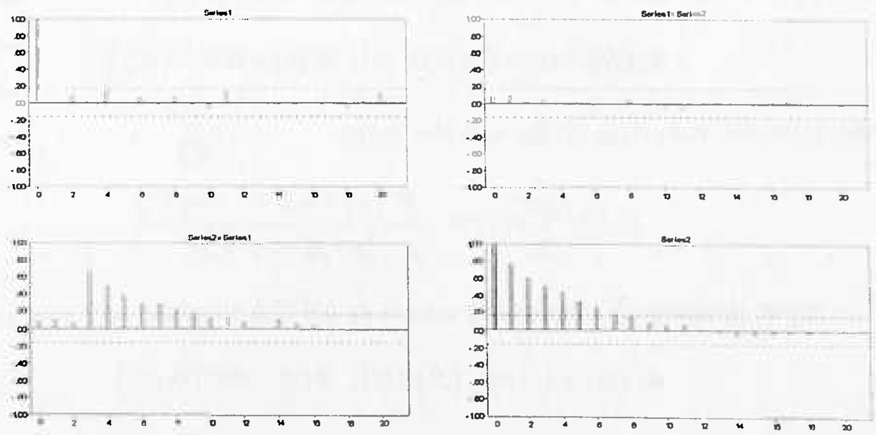


Figure 3.13 Sample cross correlations graph from X_t and Y_t time series

The upper right and lower left graphs suggest that the Y residuals at time t+h are correlated with X residuals at time t for h=3,...,9 but not otherwise. Thus an appropriate transfer function model is this with τ_j non zero for j=3,...,9 and zero otherwise. The estimated values of $\tau_j=\rho(j)\sigma_Y/\sigma_X$ (Brockwell and Davis, 1996), where $\rho(j)$ is the correlation between Y residuals at time t+h and X residuals at time t. Since $\tau_j=\rho(j)\sigma_2/\sigma_1$ is decreasing approximately geometrically for $j\geq3$ we take $T(B)$ which has the form,

$$T(B)=w_0(1-v_1B)^{-1}B^3$$

Thus the preliminary coefficient estimates are given from the functions,

$$w_0=\tau_3 \text{ and } v_1=\tau_4/\tau_3$$

and in this example $w_0=4.86$ and $v_1=0.698$.

(iv) Final estimates

Using the preliminary coefficients as starting values we can use least squares estimation to estimate more efficiently the parameters of our model. Thus,

$$T(B)=\frac{B^3*4.717}{1-0.7248B}$$
$$X(t)=Z(t)-0.474Z(t-1),\ Z(t)\sim WN(0,779)$$
$$N(t)=W(t)-0.5825W(t-1),\ W(t)\sim WN(0,04864)$$
$$AICC=27.664$$

(v) Residuals checking

In the next figure (3.14) the plot of the cross correlations residuals is presented indicating that there are no significant correlations between the input and noise residuals. Furthermore, to test the goodness of fit, the plot in the figure 3.15 shows that the assumption of uncorrelated residuals of the fitted transfer model is also valid.

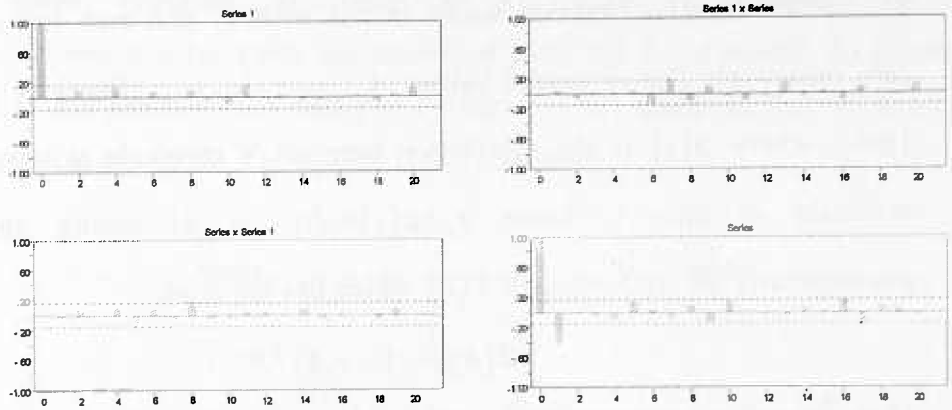


Figure 3.14 Plot of the cross correlations residuals

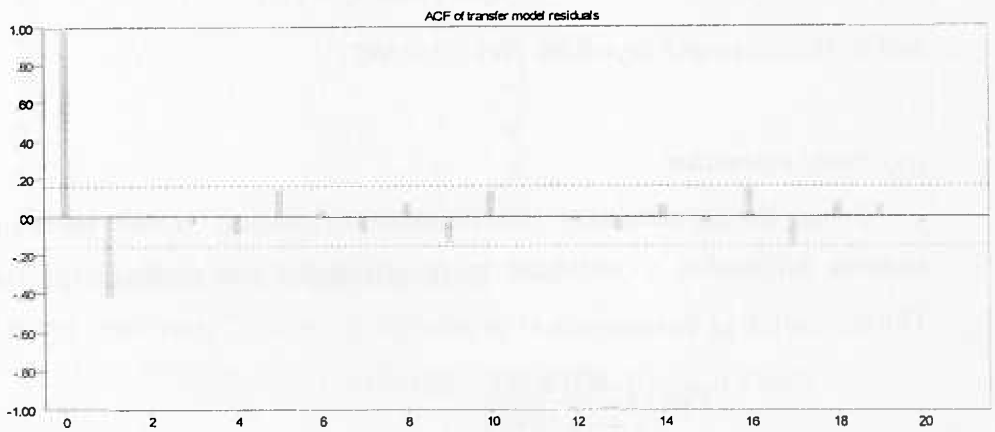


Figure 3.15 Plot of the residuals correlations

3.8. Fundamentals of State space modelling.

The transfer function model, was presented in section 3.7, characterizes the relationship between input and output in a dynamic process. A different description of this process is based on internal models which describe the internal couplings between input and output using a number of variables not directly observable, called state variables and the corresponding model is called state – space model. State variables are internal variables that determine the future behaviour of the process given the input and the current state of the process. We note that when we model an industrial process, using state space form, it is not necessary the state to have a physical meaning. Assuming that we have a dynamic process with input X and output Y.

The state space form consists of the following two equations:

a) The state or system or transition equation which describes the dynamics of the state variables using differential or difference equations and,

b) The observation equation which expresses Y_t as a linear function of a state variable Z_t plus noise at time t. A general form of a state space model is:

$$\begin{aligned} Z_t &= \Phi_t Z_{t-1} + a_t \\ Y_t &= H_t Z_t + N_t \end{aligned} \qquad (3\text{-}15)$$

where Z_t is the unobservable vector $(r \times 1)$ of state variables, Y_t is the number of observations, $\Phi_t (r \times r)$ is a transition matrix and $H_t (1 \times r)$ is a vector that varies with time t, a_t is the white noise, called *process noise* and N_t is called *measurement noise*.

3.8.1 Modelling a continuous process

A state space formulation of a deterministic continuous process can be modelled using differential equations in state and observation equation as:

$$\frac{dZ}{dt} = \Phi Z_t + \Gamma X_t$$
$$Y_t = C Z_t$$

where $Z(r \times 1)$ vector of r unobservable state variables, $X(r \times 1)$ vector of inputs and $Y(p \times 1)$ vector of outputs. If the process sampled periodically at times $t_k = k\Delta t$ it can be proved (DelCastillo, 2002) that the state of the process at discrete points t_k is given using Laplace transforms as,

$$Z_{t+1} = A Z_t + B X_t$$
$$Y_t = C Z_t$$

assuming, that $\Delta_t = t_{k+1} - t_k$ is constant and $A = e^{\Phi \Delta t}$, $B = \int_0^{\Delta t} e^{\Phi s} ds \Gamma$.

3.8.2 Modelling a discrete process

In the case where we have a discrete process, this can be generally modelled solving a first order of linear equations. Assuming that a state space form of a discrete process is described as:

$$Z_{t+1} = AZ_t + BX_t + a_t$$
$$Y_t = CZ_t + N_t$$

where, a_t and N_t are model disturbances. We can solve the state equation by repeated substitution. If time $t=0$ is the initial time after k repeated steps we have:

$$Z_k = A^k Z_0 + \sum_{j=0}^{k-1} A^{k-j-1} B(X_j + a_j)$$

3.8.3 Obtaining the transfer function from state space form

At any time we need to obtain a transfer function from a state space formulation we need to eliminate the state variables. Thus, considering the state equation of a deterministic model as

$$Z_{t+1} = FZ_t = AZ_t + BX_t$$

where F is the forward shift operator, a state space form is given by the following equations:

$$(FI - A)Z_t = BX_t$$
$$Y_t = C(FI - A)^{-1} BX_t$$

where, the transfer function in terms of forward shift operator (F) is

$$H^*(F) = C(FI - A)^{-1} B$$

and the transfer function in terms of backward shift operator B is,

$$H(\mathrm{B}) = C(I - \mathrm{B}A)^{-1} \mathrm{B}B$$

Chapter 4

Engineering Process Control

4.1 Introduction

As we stated in Chapter 2 the process control and variability reduction are of major concern in the modern industrial policy. There are two basic statistical approaches to deal with this problem. The first one is the statistical process monitoring which uses control charts or the widely known statistical process control (SPC). The main concern of this approach is to quickly detect assignable causes so that an action will be taken to correct them. The second approach is based on adjusting the process, using information from the previous observations about the deviation of the target level. This approach is called Engineering Process Control (EPC) or feedback adjustment (Montgomery, 2001). When this control is implemented by measuring devices, sensors and actuators is called Automatic Process Control (APC). In section 4.2 the process adjustment which is applicable through Deming's funnel experiment is presented. In section 4.3, the most widely known discrete feedback controllers including EWMA controllers in Run-to-Run (R2R) control are discussed and is also a brief description of feedforward controllers is given. In section 4.4 the Grubb's adjustment rule which is applied when the process is initially off the target is presented. Finally in section 4.5 adaptive controllers are discussed.

4.2 Process Adjustment

Statistical process control is developed mainly for parts manufacturing, while engineering process control is applied for continuous processes in chemical and semiconductor manufacturing. The latter approach is based on the assumption of compensation and regulation of a manipulated variable that is adjusted in order to keep the process output on target. There are situations where, despite our best efforts, the process has a tendency to "drift" or "wander" away from target. This may due to raw materials, temperature effects or any others unknown causes which impact the process. Thus, a process regulation assumes that there is another variable that can be adjusted in order to compensate for this drift in the process output and after a series of regulation the process output will be close to the target. But when someone will use engineering process control instead of statistical process control? Also could we use both of them? In the next paragraphs examples are given where SPC, or EPC, or both of them should be used.

In section 3.7 we presented the transfer function models. In a transfer function process there is a manipulated (controllable) variable X_t which we consider that has an effect on the output Y_t. In this dissertation we focus on single input single output (SISO) systems. Therefore, the output is determined by a manipulated variable and our major concern is to find the relationship between the input and the output with the presence of inevitable noise (disturbance). Furthermore, in section 3.7.3 was presented an additive model with noise which has a simpler structure of this form

$$Y_t = S_t + N_t$$

where S_t is a signal, that is a transfer function of X_t and the disturbance term N_t is a possibly correlated noise process that is determined by a white noise sequence ε_t. We can think the relationship between N_t and ε_t as a time series where ε_t is the input and the correlated noise N_t is the output.

4.2.1 Deming's Funnel experiment

There are cases where an adjustment could never be applied. Here, we illustrate an extreme case through the Deming's funnel experiment (Del Castillo, 2002). The original version is described by Deming (1986) and it is a classic example of quality improvement. The experiment was carried out by putting a funnel over a target bull's eye, on a flat surface. Marbles were dropped into the funnel and their position with respect to the target can be adjusted from drop to drop. The aim of this experiment is to minimize the deviations of the marbles positions to the target. In this experiment it is assumed that:

i) The process is in statistical control and

ii) The process is initially on target.

One hundred (100) marbles are dropped into the funnel and four adjustments are proposed:

1. Leave the funnel fixed; no adjustment.

2. At drop k (k = 1,2,...) the marble will come at point Y_k from the target. Move the funnel at $-Y_k$ from its last position; one adjustment.

3. Move the funnel $-Y_k$ from the target; one adjustment.

4. Set the funnel for the next drop (k +1) right over where the marble came to rest at the preceding drop; some adjustments.

We can think that the position of the funnel is the manipulated variable in a feedback mechanism and the distance from the target is the deviation of a quality characteristic from the target value. Rules 2 to 4 contain a feedback action given that we have information from previous measurements to set the current position of the funnel. Rule 1, is the same with SPC and finally, rule 4 imitates the operator's actions when try to make each part of the manufacturing process the same with the previous one to achieve consistency. Figures 4.1 and 4.2 illustrate the simulation of the four adjustment rules and a time series simulation of these rules when the process is in statistical control respectively.

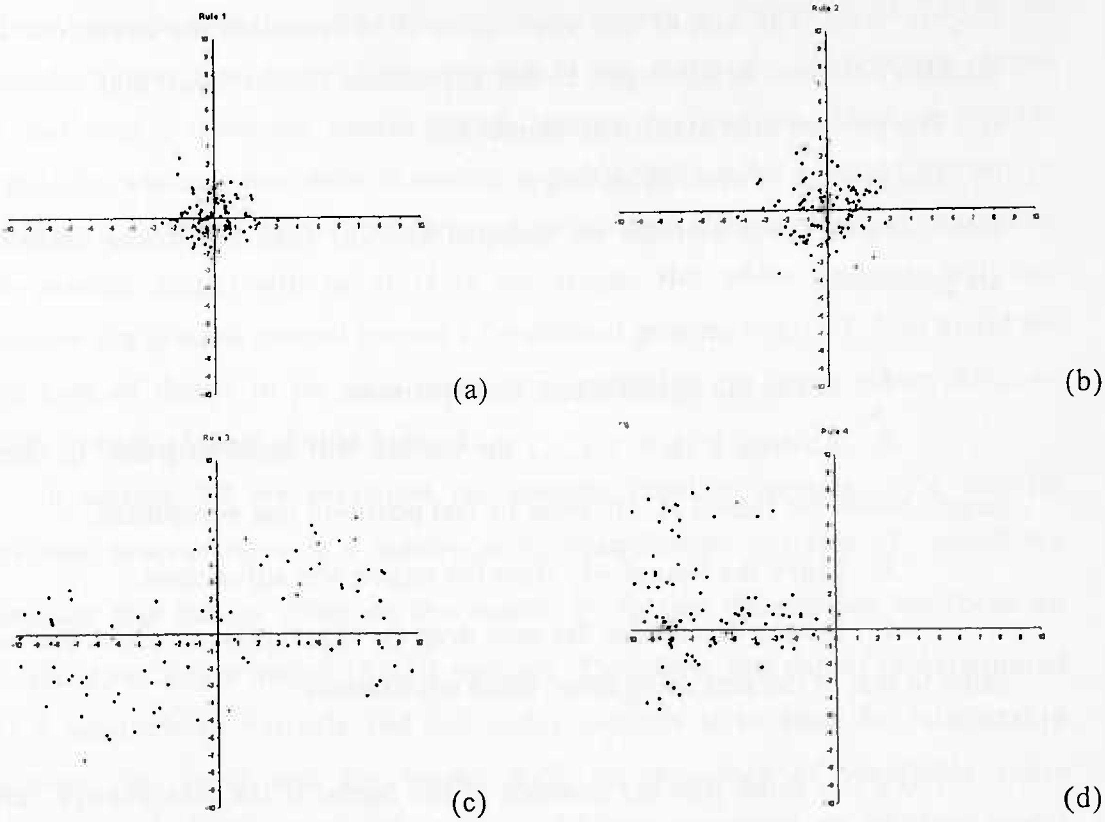


Figure 4.1 Simulation of the Funnel experiment of four adjustment rules

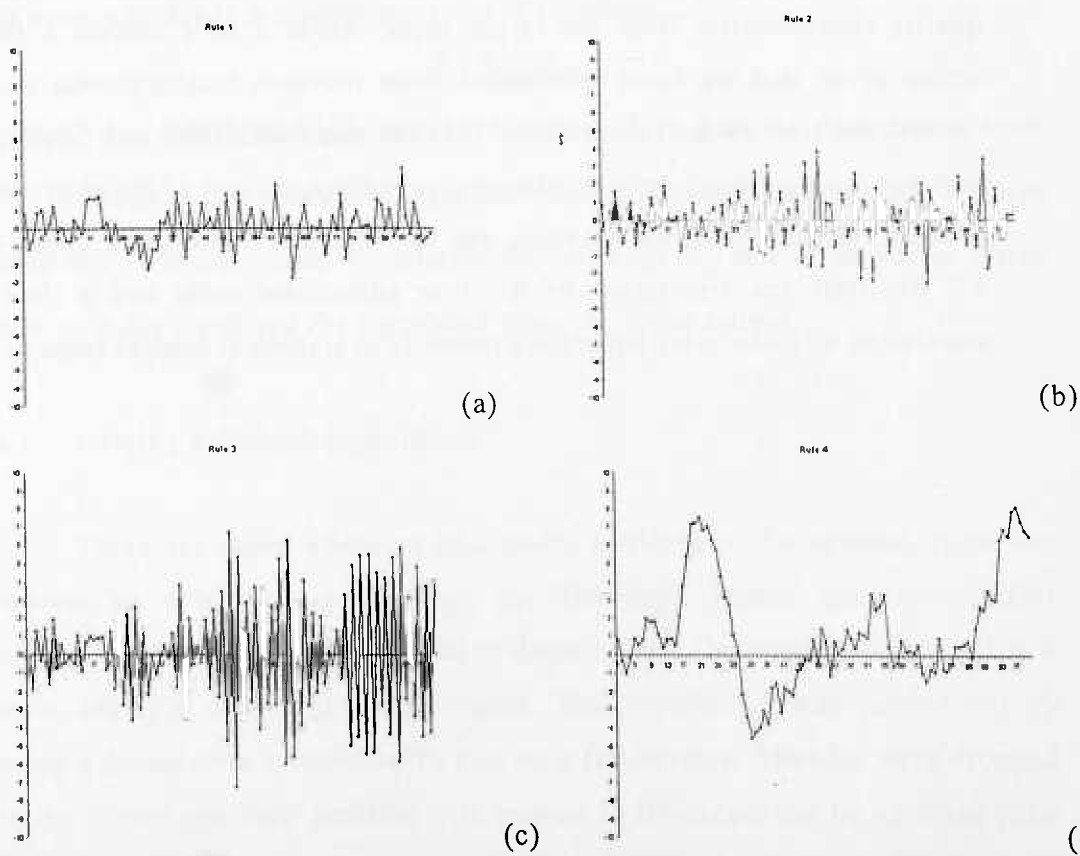


Figure 4.2 Time series simulation of the four adjustment rules

In figure 4.1a, where there is not adjustment, the dropped marbles will be in a small region around the initially position (process is on target; minimum variance). In figure 4.1b, is the result after one adjustment according to the first observation (process is on target; twice the variance of rule 1). In figure 4.1c, the process is non - stable and explodes an oscillatory pattern which is being wider over time. In figure 4.1d, as the time passes, process explodes and the marbles are moving away from the target which is analogous to random walk in time series modeling.

It is obvious that when a process is on target and in statistical control state, is not adjusted. Grubbs (1954) proposed an adjusting mechanism in the case where the process is initially off target. This mechanism will be discussed in section 4.4 in details.

4.2.2 When SPC and EPC should be applied in a process

Hunter (1994) gave an example where he tried to define when someone can use SPC or EPC techniques. We assume that there is the following stochastic model (here a state space model):

$$\begin{aligned} Y_t &= Z_t + a_t \\ Z_t &= \phi Z_{t-1} + u_t \end{aligned} \tag{4-1}$$

where u_t and a_t are white noise sequences uncorrelated each other and Y_t is the deviation of the target of a quality characteristic. The variable Z_t models the dynamic behavior of the deviation of the target that is, how the process mean changes over time, and is directly unobservable. The dynamic behavior of this process is observable under the presence of the measurement error a_t. The parameter ϕ determines how fast the process is relative to the time. Thus, this stochastic model is a state space model as we have seen in previous chapter (3.8). If we solve the state space equation with respect to ϕ, $\phi = \exp\left(\frac{-\Delta t}{T_c}\right)$, where Δ_t is the time between samples assumed constant. The variable T_c denotes the *time constant* of the dynamical system. We conclude that the shorter T_c the faster the process dynamics are. In particular:

a) If $\Delta_t >> T_c$ then $\phi \to 0$ and $Y_t = u_t + a_t \equiv \varepsilon_t$ (Shewart's model). That means if the sampling interval is slow relative to process dynamics the observed process will be uncorrelated and SPC chart will work satisfactorily.

b) If $\Delta_t << T_c$ then $\phi \to 1$ and $Z_t = Z_{t-1} + u_t$. In this case the observation equation $Y_t = Z_t + a_t$ is a random walk plus noise. That means that a rapid sampling, results in a nonstationary process where SPC charts are not appropriate. Therefore an EPC method based on predictions Y_{t+h} is more appropriate.

Summing up in a few words, if in the aforementioned state space model the measurement error a_t is much larger than Z_t then, an SPC chart works much better than an EPC strategy.

In figure 4.3(adapted from Hunter, 1994) we present when SPC and EPC apply better.

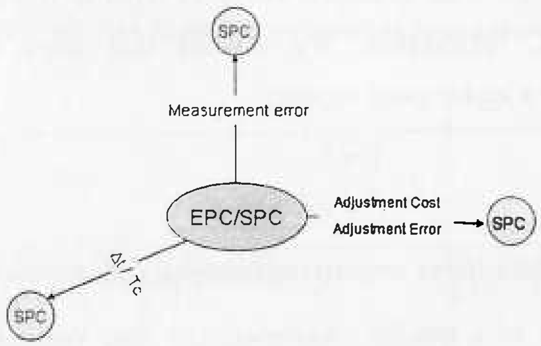


Figure 4.3 When SPC and EPC should be applied to a process

On the other hand there are cases where is not clear when to use SPC or EPC techniques. What this figure implies, is that EPC may be needed or not but SPC is always necessary.

4.3 Controllers

EPC is a set of feedback or feedforward adjustments, or even, a combination of them. Our aim is to find a control rule known also as controller in order to adjust the input of a process in order to achieve a desired performance of the output as close as possible to a target value. In this section the most common feedback controllers including EWMA controllers in R2R situation and a brief description of feedforward controllers are discussed.

4.3.1 Basic terminology of controllers

The most usual terms concerning controllers are illustrated as follows:

Output: Quality characteristic to be controlled.
Input: Compensating variable that can change the level of the output.
Adjustment: Change in input level required to compensate for output deviation.
Disturbance: The time period followed by output if no compensatory adjustments are made.
Feedback control: Using the past output deviations from target to determine a process adjustment.
Feedforward control: Present and past values of some other predictive input variable to determine process adjustment.
Open Loop: An open-loop system is a system with no feedback. In an open-loop system, there is no 'control loop' connecting the output of the system to the input to the system.
Closed loop: A closed-loop system includes feedback. The output from the system is fed back through a controller into the input to the system.
Offset: Output deviation from target if no process adjustment is applied.

4.3.2 Controllers definition

Controller is a function or rule or algorithm that depicts how a manipulated (controllable) variable X_t can be adjusted from observation to observation. In cases where the data are available, controllers are implemented automatically (sensors, electronic controllers) or manually. The magnitude of the adjustments depends on the nature of process. There are processes where adjustments are made even though the output is close the target and others where adjustments are expensive and this practice is not preferable. In the situation where frequent adjustments are expensive or not feasible, a common strategy is to wait until there is a significant deviation from the target in order to adjust them. These strategies, known as deadband policies, are proved to be optimal (Box and Kramer, 1992) wherever there is a considerable cost or adjustment error.

Block diagrams

Block diagrams are graphical representations of information flows and transformations that affect these flows of the process. Block diagrams are very common in engineering but are uncommon in statistics. In a discrete process the information flows (variables) are represented by arrows and transformation that affects these variables by boxes or blocks. In figure 4.4 is illustrated a simple transfer function model as $Y_t = \frac{\omega(B)}{\delta(B)} X_t$ where $\frac{\omega(B)}{\delta(B)}$ is the transfer function as a ratio of two polynomials.

In figure 4.5 a block diagram is presented. This is a case of a transfer function model where two variables are needed to be added ($S_t + N_t = Y_t$). The signs near the end of the incoming arrows indicate if a variable must be added or subtracted.

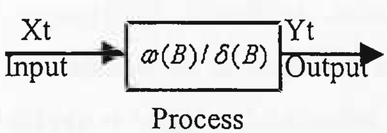


Figure 4.4 Block diagram of a transfer function

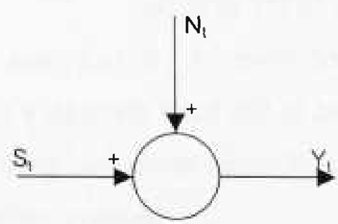


Figure 4.5 Block diagram of a sum of variables

4.3.3 A simple feedback adjustment scheme

Consider a process where a quality characteristic is out of control. Suppose that there is a disturbance N_t from its target value T when no adjustment is made. Suppose also that there is a manipulated variable X which can be used to adjust the process and that a unit change in X will produce g units of change in the output variable Y in one time interval. At time t, X is equal to X_t and at time $t+1$ the deviation from the target is $\varepsilon_{t+1}=Y_{t+1}-T$ and after adjustment,

$$\varepsilon_{t+1}=gX_t+N_{t+1} \qquad (4\text{-}2)$$

Suppose that we can compute an estimate $\hat{N}_t(1)$ at time t of N_{t+1} with error $e_t(1)$ so,

$$N_{t+1}=\hat{N}_t+e_t(1) \qquad (4\text{-}3)$$

then using (4-2) and (4-3)

$$\varepsilon_{t+1}=gX_t+\hat{N}_t(1)+e_t(1) \qquad (4\text{-}4)$$

and if an adjustment of X at time t is $X_t=-\frac{1}{g}\hat{N}_t(1)$ then

$$\varepsilon_{t+1}=e_t(1) \qquad (4\text{-}5)$$

So, the deviation from target ε_{t+1} at time $t+1$ for the adjusted process is equal with the error in forecasting N_{t+1} and the actual adjustment to the manipulated variable X at time t is

$$\boxed{X_t-X_{t-1}=-\frac{1}{g}\left(\hat{N}_{t+1}-\hat{N}_t\right) \qquad (4\text{-}6)}$$

A simple feedback controller is illustrated with a block diagram in the following figure:

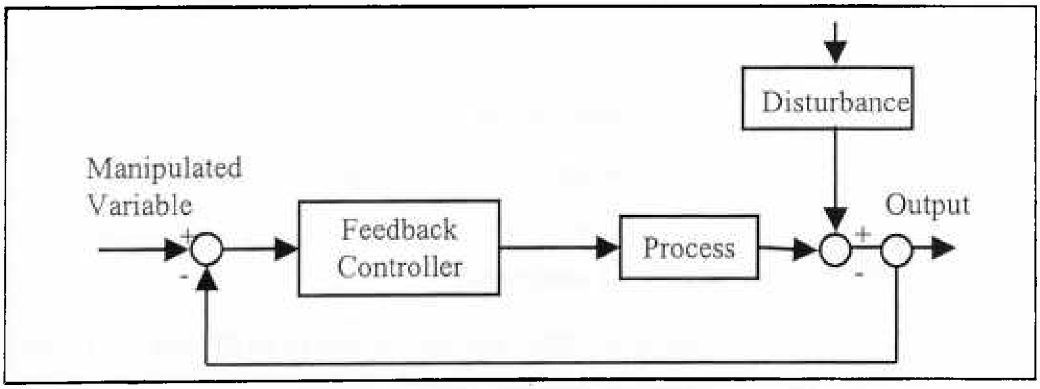


Figure 4.6 A feedback control

The controller takes the deviations from the target as input and adjusts the manipulated variable the next period of time. The value of output variable is fed back to the controller from period to period.

4.3.4 Discrete PID Controllers

The most common type of controller in industrial processes is the proportional integral derivative controller (PID). This is a feedback controller which implies that a feedback adjustment occurs when compensatory changes are made in an adjustment variable X in reaction to output deviations from the target. In this dissertation the controllers that will be discussed are the feedback controllers, although will be also mentioned other types of controllers in the end of the chapter.

4.3.4.1 Parameterization of PID controllers

There are several ways in which PID controllers parameterized. Here two ways are presented. The first is called incremental form and the second parallel form:

Incremental form of PID controller

$$\nabla X_t = K_P \nabla e_t + K_I e_t + K_D \nabla^2 e_t \tag{4-7}$$

Parallel form of PID controller

$$X_t = K_P e_t + K_I \sum_{j=1}^{t} e_j + K_D \nabla e_t \tag{4-8}$$

where K_P, K_I, K_D are the tuning parameters, correspond to proportional (P), integral (I) and derivative (D) term respectively, manipulated to minimize the process variation, e_t is the output deviation from the target and X_t is the manipulated variable. In the case where one or more of these parameters is omitted results in several special cases such as PI and I controllers.

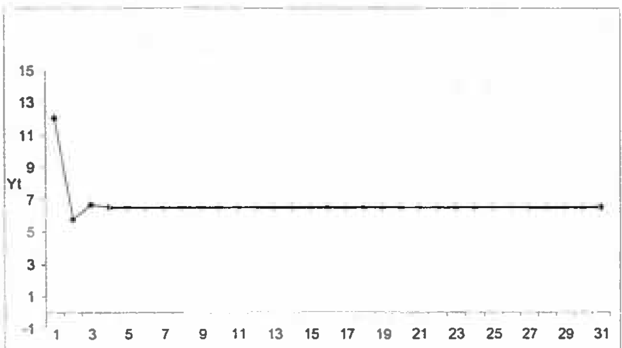
4.3.4.2 The Proportional (P) controller

This type of controller has the generic form,

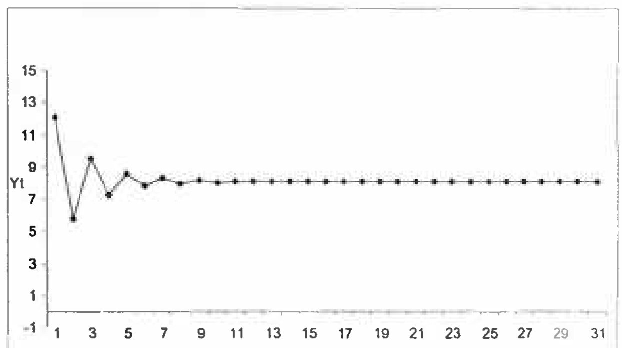
$$X_t = K_P e_t \tag{4-9}$$

Furthermore, the process output deviation e_t from target T can be described by $e_t = T - Y_t$ where Y_t is the output quality characteristic. It can be proved (Castillo

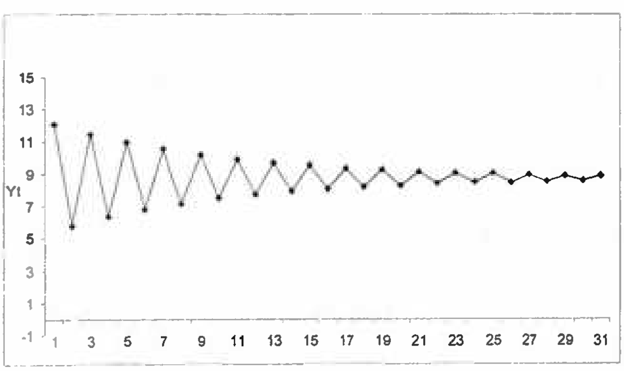
2002) that $Y_t = \alpha + \dfrac{g}{1-gB} X_{t-1}$ where under the actions of the controller $X_t = K_P e_t = K_P (T - Y_t)$. In the figure 4.7, as an example, a proportional controller applied where target value is T=12, offset say, $a = 3$ and different increasing values of K_P are tried in order to study its behavior. We can see that as K_P increases, the offset decreases; but before the process achieves the target value, $T = 12$, oscillations are making the system unstable. Thus, proportional controller reduces offset without never to eliminate it.



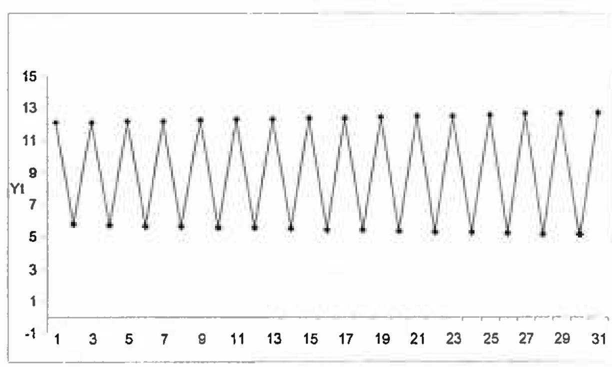
a) K_P**=1.5**



b) K_P**=3**



c) K_P**=4**



d) K_P**=4.35**

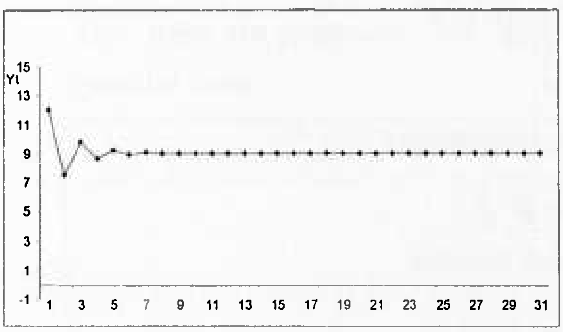
Figure 4.7 Effect of increasing K_P in a proportional controller

4.3.4.3 The Proportional Integral (PI) controller

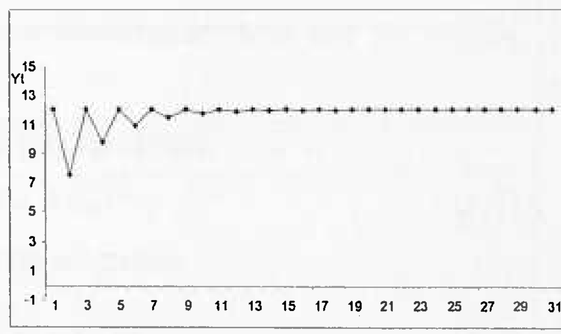
As mentioned before, the proportional controller has not the ability to eliminate the offset. Thus, the integral term is added so that the PI controller form is

$$X_t = K_P e_t + K_I \sum_{j=1}^{t} e_j \qquad (4\text{-}10)$$

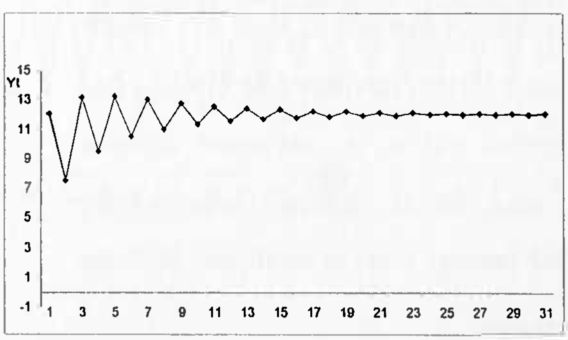
where the pure integral controller is $X_t = K_I \sum_{j=1}^{t} e_j \Rightarrow \nabla X_t = K_I e_t$. An interesting property of the integral term is that after a long time $e_t \to 0$. It can be proved (Astrom and Haglund, 1995)) that always the offset will be eliminated. This is an essential property in SPC if we think that sudden shifts are the most common types of assignable causes in industrial processes. Continuing the previous example of section 4.3.3.2 with the same values of target value, offset and $K_P = 3$ we increased gradually the value of K_I from 0 to 2. As a result, the offset reduced rapidly and eliminated although when large values of K_I are applied the process turns into unstable due to extreme oscillation.



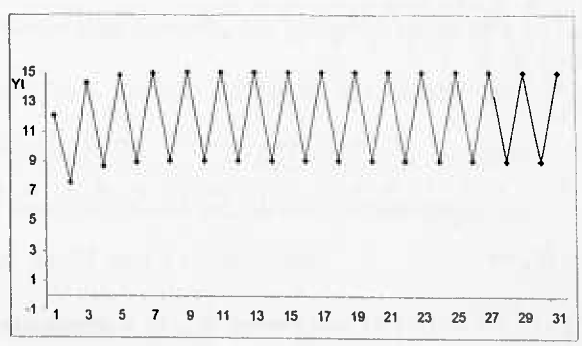
a) K_I=0



b) K_I=1



c) K_I=1.5



d) K_I=2

Figure 4.8 Effect of increasing K_I in an integral controller

4.3.4.4 The Proportional Integral Derivative (PID) controller

In the previous section we saw that a PI controller acts better than the proportional controller. Nevertheless, there are processes where a tighter control is required. In these cases the derivative term is added in a PI controller. The idea behind the derivative term is to predict where the output will be and anticipate it, given that the PI controller will be late to correct the process. Therefore, the derivative action is like a control action proportional to a predicted deviation from the target. A general form of a PID controller is the same with this which was given in section 4.3.3.1. It is worthwhile to observe that the adjustment in a PID controller is a function on the last three observed errors (deviations from the target) while in PI, PD controllers and P, I controllers is a function of the last two and one errors respectively.

As an empirical rule how to design a PID controller to obtain a desired response one could follow the steps shown below:

1. Set your design criteria.
2. Obtain an open loop response and determine what needs to be improved.
3. Add a proportional control to improve the rise time (the time it takes for the output to reach 90% of its final value).
4. Add a derivative control to improve the overshoot (Overcompensating response).
5. Add an integral control to eliminate the steady state (offset) error.
6. Adjust each of K_p, K_i, and K_d until you obtain a desired overall response. There is no need to implement all three controllers if is not necessary.

4.3.5 EWMA Controllers

Early applications of the exponentially weighted moving average (EWMA) appear in quality and process control (Box and Kramer, 1992). Recently, the EWMA controller has received a special attention in semiconductor industries where EWMA feedback controllers are used for compensating against disturbances that affect the run–to–run (R2R) variability in the quality characteristics of silicon wafers at a certain manufacturing step

Castillo and Hurwitz (1997), Moyne et al., (2000). Because semiconductor industry usually involves more than 200 process steps, variation in a certain step will impact subsequent processes and finally the quality of integrated chips. Moreover EWMA controllers have a wide application not only in the semiconductor industry but in any batch process where adjustments are necessary with every batch and there is a substantial drift in the quality characteristic.

4.3.5.1 Relationship of Integral control and EWMA statistic

Consider the simple feedback adjustment scheme which we discussed in section 4.3.3. It is reasonable to use for estimating the $\hat{N}_t(1)$ (4-4) the EWMA statistic in order to give a weight in past values of $\hat{N}_t, \hat{N}_{t-1}, \ldots$. Thus $\hat{N}_t(1) = \lambda\left(N_t + \theta N_{t-1} + \theta^2 N_{t-2}\right)$ where $0 \le \theta \le 1$ the smoothing constant and $\lambda = 1 - \theta$.

Using equation (4-6) the adjustment at time t is given by

$$X_t - X_{t-1} = -\frac{1}{g}\left(\hat{N}_t(1) - \hat{N}_{t-1}(1)\right)$$

Thus the formula for a EWMA forecast can be written

$$\hat{N}_t(1) - \hat{N}_{t-1}(1) = \lambda e_{t-1}(1) \text{ where } e_{t-1}(1) = N_t - \hat{N}_{t-1}$$

Therefore in any feedback scheme the required adjustment to cancel out a EWMA of the noise N_t when a compensatory variable was set is,

$$\boxed{X_t - X_{t-1} = -\frac{\lambda}{g} e_{t-1}(1) = -\frac{\lambda}{g}\varepsilon_t \qquad (4\text{-}11)}$$

Thus, the pure integral controller is equivalent to set the level of the manipulated variable and to cancel the one step ahead forecast error which made with the EWMA statistic.

4.3.5.2 The single EWMA controller

In the area of semiconductor manufacturing an EWMA controller proposed by Sachs et al (1995), which recommends process adjustment at each run of silicon wafers. The authors assumed that the value of the quality characteristic y_n for run number n is described by the relation:

$$Y_n = a + bX_{n-1} + \varepsilon_n \tag{4-12}$$

$\varepsilon_n \sim WN\left(0,\sigma^2\right)$, X_{n-1} is the input of the process that denotes the control action of the manipulated variable at the end of run $n-1$, Y_n is the output variable that is, the deviation from target and a,b are parameters estimated from the data. Assuming that X_n is a scalar, although can be extended to the multiple input case, the relationship between the controllable variable and the target value T is,

$$X_n = \frac{T-a}{b} \tag{4-13}$$

where the estimated $\hat{b}$ of the process gain is available prior to the beginning of the control action and an estimate $\hat{a}$ of a can be provided recursively using the EWMA equation. At run n the estimate of a is denoted by $\hat{a}_n$ is described by the equation

$$\hat{a}_n = \lambda\left(Y_n - \hat{b}X_{n-1}\right) + (1-\lambda)\hat{a}_{n-1} \tag{4-14}$$

where λ gives more weight to the most recently observations of the quality characteristic. The equation (4-13) is the single EWMA controller. It can be shown that the single EWMA controller is a pure I controller, using (4-13) in conjunction with (4-12), where the estimate $\hat{a}_n$ and the controllable variable are given by the equations:

$$\boxed{\begin{aligned} &\hat{a}_n = \lambda\left(Y_n - T\right) + \hat{a}_{n-1} \\ &X_n = -\frac{\lambda}{\hat{b}}\sum_{n=1}^{t-1}\left(Y_n - T\right) + X_0 \qquad (4\text{-}15) \end{aligned}}$$

where X_0 is the initial control factor and the integration constant is $K_P = -\lambda/b$.

When the process does not obey in equation (4-12) but instead obeys the equation

$$Y_n = a + bX_{n-1} + D_n + \varepsilon_n \tag{4-16}$$

where D_n term is a deterministic drift disturbance common in many manufacturing processes, the EWMA controller is not optimal that is, will not achieve the minimum possible variance of σ^2.

4.3.5.3 The double EWMA (DEWMA) controller

Given that a single EWMA controller is not appropriate in cases where there is a severe drift of the quality characteristic, in other words in cases where can be exhibit considerable offset, Butler and Stefani (1994) proposed to extend the single EWMA controller. More specifically, they proposed to add a second EWMA equation to single EWMA controller that would compensate for the offset. Butler and Stefani assumed a model with linear deterministic drift (4-16) and also assumed that a prior estimate $\hat{b}$ of b exists, so, they proposed to use a double EWMA controller (also called predictor – corrector controller) where the controllable variable is given by

$$X_n = \frac{T - \hat{a}_n - D_n}{\hat{b}} \tag{4-17}$$

and

$$\boxed{\begin{aligned} \hat{a}_n &= \lambda_1\left(Y_n - \hat{b}X_{n-1}\right) + \left(1-\lambda_1\right)\hat{a}_{n-1} \\ D_n &= \lambda_2\left(Y_n - \hat{b}X_{n-1} - \hat{a}_{n-1}\right) + \left(1-\lambda_2\right)D_{n-1} \qquad (4\text{-}18) \end{aligned}}$$

where $0 \le \lambda_1, \lambda_2 \le 1$.

In order to use the double EWMA controller, process engineers must select the weights λ_1, λ_2. The selection of controller's weights does not addressed by Butler and Stefani where arbitrarily values were used. The weight selection problem is studied by DelCastillo (1999) and it will be discussed later.

4.3.5.4 Age – Based Double EWMA controller

In the case of a tool wearing in semiconductor processes the EWMA statistic is not adequate; therefore a DEWMA controller was proposed as we discussed earlier by Butler and Stefani (1994). However, this controller does not take into consideration the process "age". An adjusted DEWMA formula was proposed by Chen and Guo (2001) in cases where the sampling time is not equally spaced. Thus the DEWMA equations are modified as

$$\hat{a}_n = \lambda\left(Y_n - \hat{b}X_{n-1}\right) + (1-\lambda_1)\left[\hat{a}_{n-1} + (t_n - t_{n-1})D_{n-1}\right] \tag{4-19}$$

$$D_n = \lambda_2\left(\frac{Y_n - \hat{b}X_{n-1} - a_{n-1}}{t_{n-1} - t_n}\right) + (1-\lambda_2)D_{n-1} \tag{4-20}$$

where t_n denotes the process age at n run. The authors applied this controller in a chemical mechanical polishing (CMP) process where the results showed that the proposed controller improves the control efficiency significantly.

4.3.5.5 EWMA controller with step change disturbance

So far, EWMA controllers are discussed taking into consideration the assumption that EWMA estimator is optimal when the process mean follows an IMA(0,1,1) model (Box et al., 1994). This estimator performs well for various processes and especially in semiconductor fabrication processes. Chen and Elsayed (2002) are proposed an EWMA estimator whose mean follows another kind of disturbance process which is called as step – change model. This model was derived from realistic situations were the process mean is subject to occasionally step changes caused by variations in the physical conditions.

Chen and Elsayed assumed that:

a) The disturbance D_t is normally distributed with mean μ_t and variance σ^2 at time t that is,

$$D_t = \mu_t + \varepsilon_t$$

where $\mu_0 \sim f\left(\xi, \tau^2\right)$ and $\mu_t = \begin{cases} \mu_{t-1} \ \textit{with probability} \quad 1-p \\ \sim f(\xi,\tau^2) \ \textit{with probability } p \end{cases}$

The probability p represents the frequency of changes to the process assuming that the step change occurrence is independent of the prior history of the process.

b) The EWMA estimator here is used to estimate μ_t is,

$$\hat{\mu}_t = \lambda D_{t-1} + (1-\lambda)\hat{\mu}_{t-1}$$

where $\hat{\mu}_0 = D_0$ is an estimate of the overall mean ξ from the historic data. This estimator is beneficial because of simple PID controller implementation. In order to choose appropriate values for λ the authors are proved that λ is given by:

$$\lambda = \frac{-p(1+r^2) + r\sqrt{p^2 r^2 - p^2 + 2p}}{(1-p)}$$

where $r = \tau/\sigma$ that is, the standard deviation ratio. The authors also presented some contours plots for several values of λ. These plots showed that as λ increases the step change increases too or occurs more frequently. Furthermore, the effect of the step change occurrence frequency on λ, decreases, when the step change size gets smaller.

4.3.6 Feedback adjustment charts

The feedback adjustments schemes which described in previous sections can be implemented with the combination of sensors and other devices automatically. However there are processes where feedback adjustments are made manually. In these cases the process operators are using manual adjustments charts.

4.3.6.1 The Bounded adjustment chart

There are cases where the cost of making an adjustment is a concern. In these cases some modifications are made to the feedback adjustment procedure so that less frequent adjustments are necessary. A very simple way to do this is using the bounded adjustment chart. This chart is based on the idea that an adjustment will be made not after every observation but when the absolute value of a forecasted deviation from target is larger than a threshold value (Box and Kramer 1992, Box and Luceno 1997); that is, some pre - specified bounds given by $\pm L$. This boundary value is determined from process operators after comparison of the off target cost and the adjustment cost.

As an example in figure 4.9 we have a bounded adjustment chart of 50 observations where the relationship between the input and output in this process is given by

$$Y_t - T = 0.9X_t$$

where the process gain is $g = 0.9$, $\lambda = 0.3$, $T = 0$ and $L = \pm 10$. That means that adjustments are made when EWMA exceeds $L = \pm 10$. Note that adjustments are made at points 3, 5, 8, 14, 29.

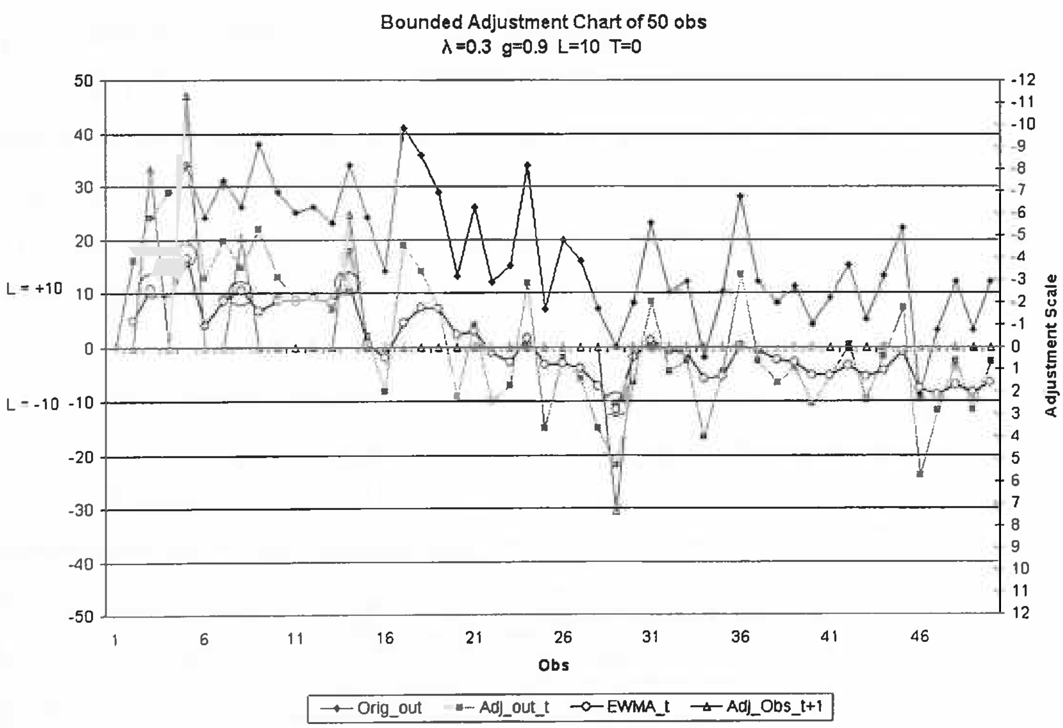


Figure 4.9 Bounded adjustment chart showing the original and adjusted output the EWMA and the real adjustments. The circled EWMA indicate where adjustments are made

4.3.7 Other types of controllers

Until now, the adjustments schemes that presented in this chapter are feedback adjustment schemes. These controllers are most widely used in many industrial areas. However there are situations where is preferable to reduce variation in the input of a process with or without conjunction of a feedback adjustment. This form of controller is called feedforward controller.

4.3.7.1 Feedforward controllers

EPC schemes depend on the information of the process dynamics in order to determine the adjustments. In feedback control, one must know (or assume) a dynamic model linking the manipulated variable to the output. In feedforward control one must additionally know the relationship between the input variable and the output. So, if we can measure fluctuations in an input variable that can be observed but not changed, it may be possible to compensate some other variables. This is a disadvantage of feedforward control with comparison to feedback control where we don't need to know accurately the source or the magnitude of the disturbance. However, there are processes where the feedforward control is applicable, such as in lithography.

A simple feedforward control scheme is illustrated in figure 4.10

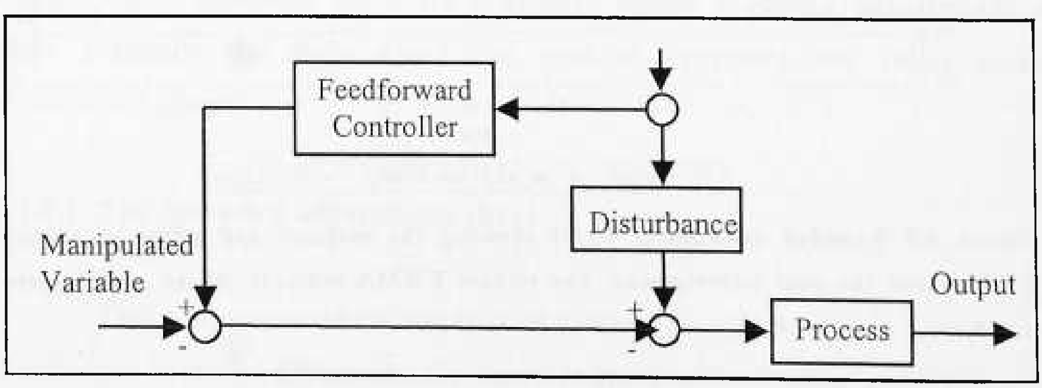


Figure 4.10 A feedforward scheme

4.3.7.2 Feedforward as a supplement to feedback controllers

Box and Luceno (2002) proposed the complementary use of feedforward and feedback adjustments to compensate for expected level shifts in the process mean where this is necessary in an industrial process when a new batch of feedstock material is set up. Authors are studied and compared five alternative control schemes:

(i) Only feedforward adjustment (FF).

(ii) Only feedback adjustment (FB).

(iii) Feedback – Feedforward adjustment (FB+FF).

(iv) Feedback in conjunction with indirect forward to increase the sensitivity of feedback (FB+FFS).

(v) Feedback with both direct and indirect feedforward (FB+FF+FFS).

Box and Luceno assumed, for all the above five control schemes, that the deviation from target Y_t at time t is represented by an IMA time series model as,

$$Y_t - Y_{t-1} = a_t - \theta a_{t-1} \tag{4-21}$$

where a_t is a white noise sequence. They also assumed that a batch of feedstock lasts for T intervals and that an event occurring in the i^{th} interval of the h^{th} batch is indexed by $t = i + hT$.

Comparison of the various schemes results in

(a) FD adjustments are not suitable to eliminate non stationary disturbances and any FB scheme is possible to do much better.

(b) Some improvement can be feasible by the use of feedback scheme with direct and/or indirect feedforward supplement.

4.4 Grubb's adjustment rules

In section 4.2.1 Deming's funnel experiment was presented under the assumption that the funnel was initially on target. However, when the process is initially off target Grubbs (1954, 1983) proposed some process adjustment rules. These rules could be applied in any manufacturing process where an incorrect setup operation can result in drastic consequences in the quality of the output. According to Grubbs, assume initially that the setup offset d (units) is an unknown constant not necessarily equal to zero. For the first part the mean deviation from target is

$$\mu_1 = d + U_0$$

where U_0 is the initial setting of the machine setpoint (controllable factor). Grubbs assumed that $U_0 = 0$ that is, before start up the machine was on target. If d is known we could set $U_t = -d$ for all t and completely eliminate the offset; this trivial case is seldom possible.

The first observed deviation from target is given by

$$Y_1 = \mu_1 + u_1 = d + U_0 + u_1$$

where $u \sim N\left(0, \sigma_u^2\right)$ and models part to part variability and the measurement error. In contrast with some previous EPC adjustment methods where the adjustments affect the deviations from target, now it is assumed that the adjustments affect the mean of the process (DelCastillo, 1998). That is, the adjustment ∇U_1 will result in a new process mean of

$$\mu_2 = \mu_1 + \nabla U_1$$

and the second deviation from target will be

$$Y_1 = \mu_1 + u_2$$

Continuing in this form, a general expression for the mean and the deviations from the target is,

$$\mu_t = \mu_{t-1} + \nabla U_{t-1}$$

$$Y_t = \mu_t + u_t \qquad (4\text{-}22)$$

respectively. Grubbs(1954,1983) proposed to find the adjustment weights $\{K_i\}_{i=1}^{n}$ that is, $\nabla U_{t-1} = -K_t Y_t$ solving the following problem:

$$\boxed{\begin{array}{l} \min Var\left(\mu_{n+1}\right) \\ subject\ to:\ E\left[\mu_{n+1}\right] = 0 \end{array}} \qquad (4\text{-}23)$$

That means that we want to have a process that on average is on target after n parts have been processed with a minimum variance around the target. We also point out that μ_t becomes a random variable since it's a function of adjustments based on random observations. Therefore Grubbs showed that the weights that solve the problem satisfy the following equation:

$$K_t = 1/t \quad t = 1,2,... \tag{4-24}$$

so the adjustments are

$$\nabla U_t = U_t - U_{t-1} = -\frac{Y_t}{t}$$

That means that the adjustment weights follow the harmonic series $\{1, 1/2, 1/3, ...\}$. For this reason some authors calls this adjustment scheme as "harmonic rule".

4.4.1 Grubbs extended rule

A second adjustment rule was also proposed by Grubbs (1954). Now the setup offset d is described as a random variable. Grubbs assumed that the distribution of d has a known mean equal to zero and a known variance σ_d^2. He claimed that the variance of setup (σ_d^2) is due to changes in the machine. Furthermore, he showed that the optimal weights to solve the problem (4-23) are given by,

$$K_t = \frac{1}{t + \frac{\sigma_u^2}{\sigma_d^2}} \tag{4-25}$$

This rule is called "Grubbs extended rule".

4.5 Self tuning (adaptive) controllers

Self tuning controllers are based on the recursive estimation of the controller's parameters as if they were the true parameters. That is, the uncertainty associated with the estimates is ignored. Adaptive controllers assume, in general, that the parameters of a process change with time so that the controller must vary (adapt) its own parameters accordingly. This type of controllers are particularly important in manufacturing since they can avoid the production of scrap (nonconforming products) associated with industrial estimation experiments (DelCastillo, 2002). Furthermore, the estimation procedure can be performed online. Thus, each observation that is obtained should be utilized by the controller allowing controlling the process after startup (Astrom and Wittenmark, 1997).

4.5.1 Direct and Indirect controllers

In general two types of self tuning controllers exist; the indirect and the direct self tuning controllers. Indirect self tuning (ST) controllers have a recursive estimator that estimates the parameters of the process. Then these parameters are used by the controller. On the contrary, in the direct (ST) controllers the parameters of the controller are estimated directly by the recursive function. The objective is to find an equivalent model for the process so that, its parameters are directly correspond with the controllers parameters. A practical recommendation, before someone utilizes any type of a self tuning controller, is the extensive simulation of a wide variety of possible model processes and disturbances before implementation.

4.5.2 An example of indirect ST controller.

Assuming that a process is described as a first order transfer function process with ARMA (1,1) noise,

$$Y_t = \phi Y_{t-1} + gX_{t-1} + (1-\theta B)\varepsilon_t$$

and suppose that we seek an MMSE controller

$$X_t = -\frac{\phi - \theta}{g} Y_t$$

and the corresponding MMSE ST controller is:

$$X_t = -\frac{\hat{\phi}_t - \hat{\theta}_t}{\hat{g}_t} Y_t$$

where $\hat{\phi}$, $\hat{\theta}$ are the parameters estimates of the ARMA(1,1) model and $\hat{g}$ is the process gain at time t. The figures 4.11, 4.12 and 4.13 show the parameters estimates outputs and inputs respectively of a simulation of an indirect ST controller were

$g = 1.2$, $\phi = 0.3$, $\theta = -0.6$ and $\varepsilon_t \sim N(0,3)$.

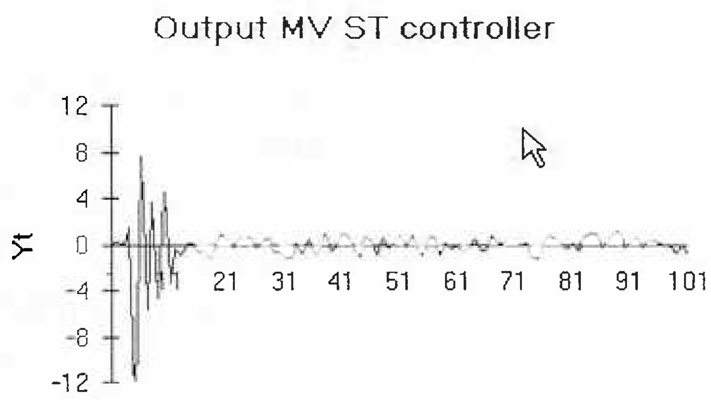


Figure 4.11 Output values of the quality characteristic using indirect ST controller

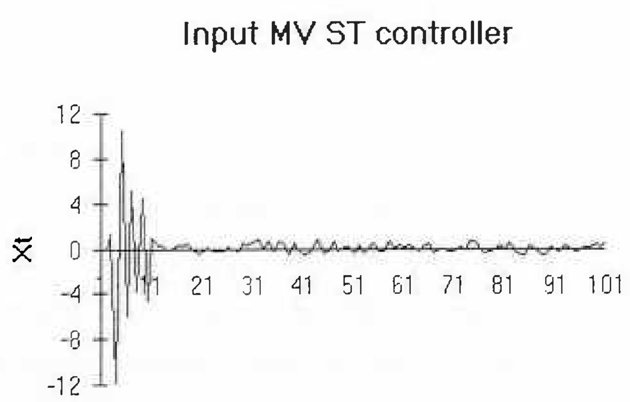


Figure 4.12 Values of the manipulated variable suggested by the indirect ST controller

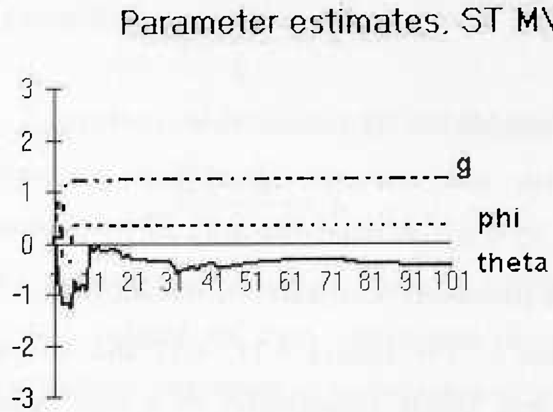


Figure 4.13 Parameter estimates used by the indirect ST controller

It is worthwhile to point out that:

- The estimates of g and ϕ converge rapidly to their true values but the estimate of θ is varying much more than the others.
- There is short transient in all three graphs. This occurs because the controller needs to probe the process making drastic changes in the manipulated variable. That is, the controller tries to probe the process in order to "learn" (e.g., get better parameter estimates) and then bring the process back to the target. These two features of the behavior of a ST controller, namely, probing and control are referred to the control literature as the "dual control effect" (Feld'baum, 1965).

Chapter 5

Optimal Feedback Controllers

5.1 Introduction

In chapter 4 our major concern was to present the most widely used feedback controllers in the modern industry. In this chapter, we are interested in finding how to adjust a process so that our adjustments to be "optimal", in the sense that the output of a process is close to the desired target value. In section 5.2 the MMSE controllers are presented and compared with the PI ones. In section 5.3 the optimal designs of the most widely used EWMA controllers are discussed and some examples are illustrated. Furthermore, in section 5.4 a variance constrained self tuning controller is presented.

5.2 Optimal feedback controllers

In this section certain types of controllers like MMSE controllers are discussed in the sense of optimization. These controllers are rather impractical in industrial processes (Box and Luceno 1995, DelCastillo 2002). Furthermore, PID controllers are by far the most common controllers and many process control devices in modern industries are only equipped with PID controllers rather than MMSE controllers (Tsung and Shi, 1999). However, these types of controllers are optimal under certain criteria and in this rationale these controllers are presented in the following sections. In addition, the MMSE schemes are still useful as benchmarks in evaluating the performance of other schemes so that, minimum variance control strategies provides a great deal of insight into the type of adjustments problems we are concerned with.

5.2.1 Minimum Variance (MMSE) Controllers

Minimum variance or minimum squared error controllers (MMSE) are seek to minimize the variability of the quality characteristic that is, the mean squared error of the process output deviation from the target. These controllers ignore any cost in doing so, that is, there are situations where a MMSE control requires inappropriately large manipulations of the compensating variable and from this side are impractical. However, it is easy to be modified to achieve the variability of the adjustments below a certain level.

(i) Process with no Dynamics

Assume that a process can be described as

$$Q_t = \mu + bX_t + \varepsilon_t \qquad (5\text{-}1)$$

that means that there is not a dynamic relationship between Y_t and X_t, where Q_t is the quality characteristic. Y_t is the deviation from target value ($Y_t = T - Q$) where T is the target value. The value μ is the mean quality characteristic assumed known. Our concern, is to minimize the MSE of the quality characteristic adjusting the manipulated variable X_t. If we set $E(Q_t) = T$ that is, the MSE is equal to the minimum possible variance of σ_ε^2, the controller is

$$X_t = \frac{T - \mu}{b} \qquad (5\text{-}2)$$

The equation (5-2) denotes that all the time the adjustments $(1-B)X_t = 0$. If $X_0 \neq \dfrac{T-\mu}{b}$ only one adjustment is necessary$[(1-B)X_1]$. If $\mu = T$, there is no need for adjustments. This is the same like in Deming's funnel experiment (rule 2).

(ii) Process with Dynamics

Assuming now that a process can be described as

$$Y_t = -aY_{t-1} + bX_{t-1} + c\varepsilon_{t-1} + \varepsilon_t \qquad (5\text{-}3)$$

where $|c|<1$ (invertibility) and Y_t denotes the deviation from target. If Y_{t+1} is the next deviation from target depends on X_t the minimum squared error is given by

$$MSE(Y_{t+1})=Var(Y_{t+1})+E(Y_{t+1})^2 \tag{5-4}$$

This quantity is minimized when $E(Y_{t+1})^2=0$. Thus, the MMSE equals the minimum possible variance where

$$MSE(Y_{t+1})=Var(Y_{t+1})\geq E(\varepsilon_{t+1})^2=\sigma_\varepsilon^2 \tag{5-5}$$

The one step ahead forecast error is

$$\hat{Y}_{t+1}=-aY_t+bX_t+c\varepsilon_t \tag{5-6}$$

and

$$X_t=\frac{a-c}{b}Y_t \tag{5-7}$$

since we choose $E(Y_{t+1})=\hat{Y}_{t+1}=0$ in order minimize the MSE and e_t is an estimate of ε_t where $e_t=Y_t-\hat{Y}_{t-1}=Y_t$. If we substitute (5-7) into (5-3) we get $Y_{t+1}=-cY_t+c\varepsilon_t+\varepsilon_{t+1}$ and using backshift operators can be written as

$$(1+cB)Y_{t+1}=(1+cB)\varepsilon_{t+1}\Rightarrow Y_t=\varepsilon_t \tag{5-8}$$

The (5-8) equation implies that the minimum variance of σ_ε^2 is realized.

(iii) Impractical use of a MMSE controller

In the previous chapter the Proportional Integral (PI) feedback adjustment scheme was presented in the form of

$X_t=K_P e_t+K_I\sum_{j=1}^{t}e_j$ where X_t is the manipulated variable of a process.

If we use this feedback scheme and we want to model the relationship between the input and the output of a process allowing inertia characteristics, a useful way to do this is through the following first order dynamic model:

$$Y_t=\text{constant}+\delta Y_{t-1}+g(1-\delta)X_{t-1} \qquad 0\leq\delta\leq 1 \tag{5-9}$$

where the inertial properties can be understood as at t time periods after a unit step is made in X the change in Y is $g(1-\delta^t)$ where δ and g are parameters of the model and g is the system gain. It can be shown (Box and

Kramer(1992)) that if the disturbance of the process say N_t can be predicted perfectly apart from random error by an IMA (0,1,1) model that is, $N_t - N_{t-1} = \varepsilon_t - \theta\varepsilon_{t-1}$ then the PI adjustment scheme can produce MMSE about the target value,

$$k_P = \frac{\lambda\delta}{g(1-\delta)} \text{ and } k_I = \frac{\lambda}{g} \text{ where } \lambda = 1-\theta \tag{5-10}$$

That means that this is an optimal control rule in the sense that it minimizes the MSE of the process output deviation from the target value. This control equation is practical only if δ is fairly small. As δ becomes larger and in particular as it approaches unity the MMSE scheme requires excessive control action (Box et al., 1994).

5.2.2 The Generalized Minimum Variance (GMV) Controller

There are circumstances where the manipulated variable exhibits a large variability under MMSE control and the control action is not practical or feasible. This occurs because the MMSE controller transfers the variability from the quality characteristic to the manipulated variable.

Consider the process which is described as

$$(1-\phi B)Y_t = gX_{t-1} + (1-\theta B)\varepsilon_t \tag{5-11}$$

where the MMSE controller is (5 -7)

$$X_t = \frac{\theta-\phi}{g} Y_t \tag{5-12}$$

In Figure 5.1 we simulated from this process where $\phi = 0.2$, $\theta = -0.6$, $g = 1$. In the first case an open loop simulation X_t is limited to a range of values $(-1,1)$ due to scaling. In the second simulation the closed loop MMSE is achieved using the controller $X_t = -0.8Y_t$. We can see that the MMSE controller compared to the open loop performance transfers the variability from the output to the input (manipulated variable).

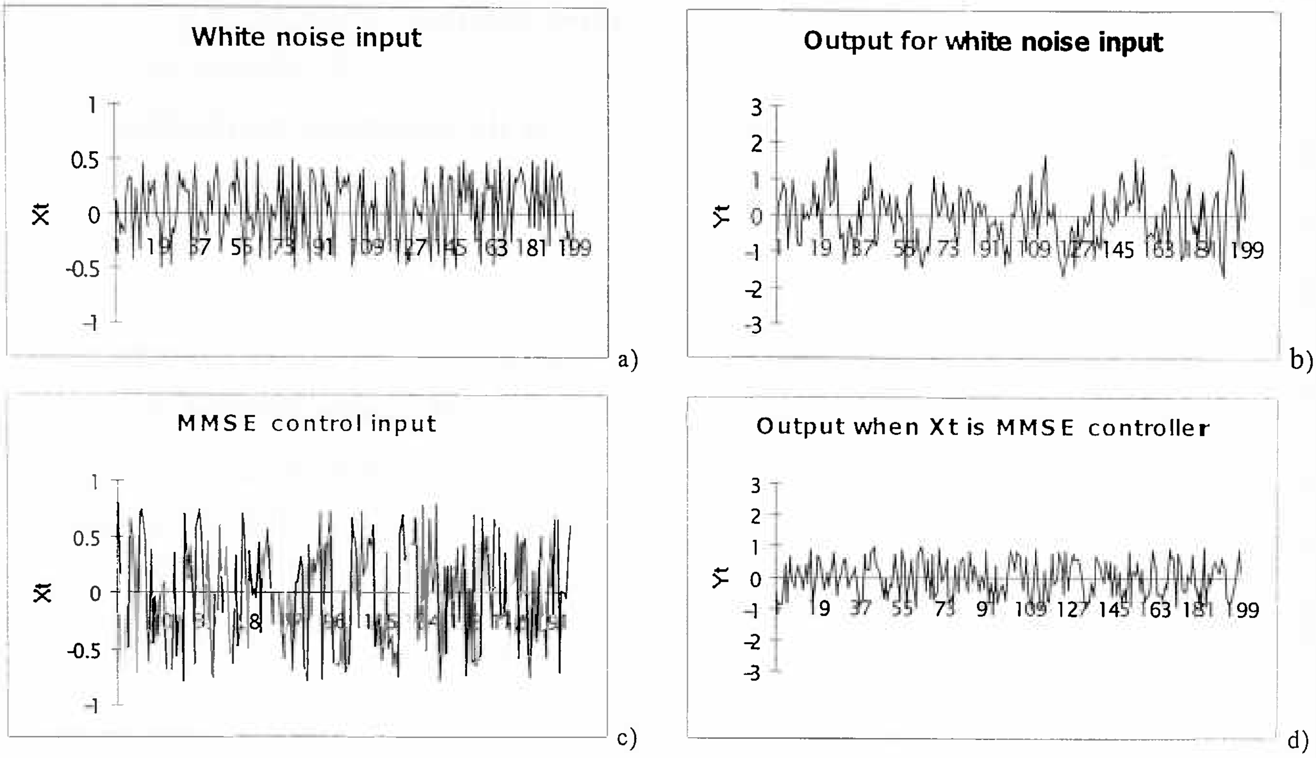


Figure 5.1 Effects on input and output a) White noise input b) Output of white noise input c) MMSE control input d) output when X_t is an MMSE control

The problem is that an extreme adjustment of the manipulated variable is not practical in many processes due to a high cost or limitations in the range of values of the manipulated variable. A method that proposed by Clarke and Gawthrop (1975) is the generalized minimum variance controller (GMV) where the GMV control minimises the squared weighted difference between the desired value and the predicted output while penalising excessive control effort. Assuming that we have a Box – Jenkins transfer function (3.7.4) with IMA (1,1) noise:

$$Y_t = \frac{g}{1-\delta B} X_{t-1} + \frac{1-\theta B}{1-B} \varepsilon_t$$

It can be shown that the GMV controller can be given by the equation:

$$\boxed{X_t = -\frac{(1-\theta B)(1-\delta B)}{g(1-B)+(\lambda/g)(1-\delta B)(1-\theta B)} Y_t \qquad (5\text{-}13)}$$

The effect of increasing λ is that we take large reductions in $Var(X_t)$ for small increments of $Var(Y_t)$.

5.2.3 The Constrained MMSE (CMMSE) control

Box et al., (1994) proposed that when the adjustments are described from a stationary time series (a constrained scheme), we can find an unconstrained minimum scheme such as,

$$X_t = \sigma_\varepsilon^2 + a\sigma_X^2 \tag{5-14}$$

where a is an undetermined multiplier that assigns the relative quadratic costs of variations of ε_t and X_t. This scheme is called constrained MMSE scheme (CMMSE). The same authors also proved, that a process with first order dynamics and IMA (0,1,1) is an unconstrained MMSE scheme which can be formulated as:

$$X_t = -\frac{\lambda(1-\delta B)}{g(1-\delta)}\varepsilon_t$$

For this unconstrained scheme the corresponding CMMSE is given by the function:

$$X_t = \left[k_1 + (1-\lambda)k_0\right]X_{t-1} - (1-\lambda)k_1 X_{t-2} - \frac{\lambda(1-k_0 B)(1-\delta B)}{g(1-\delta)}\varepsilon_t \tag{5-15}$$

where k_0 and k_1 are complicated functions of g, λ, δ and are given by a table of optimal values for the constrained scheme provided by them.

We can use a PI controller to balance the input and output variances of a process with a technique similar with the generalized minimum variance controller (5.2.2). Box and Luceno (1995) introduced a PI controller as

$$(1-B)X_t = -G(Y_t + P(1-B)Y_{t-1}) = c_1 Y_t + c_2 Y_{t-1} \tag{5-16}$$

where $-G = k_I$, $P = \frac{k_P}{k_I}$ and c_1 , c_2 are constants which control a first order system with IMA (1,1) noise. The authors showed that the variance of the adjustments is given by

$$Var\left[(1-B)X_t\right] = \left(c_1^2 + c_2^2\right)Var(Y_t) + 2c_1c_2Cov(Y_t, Y_{t-1})$$

Box and Luceno (1995) made available tables, where someone could choose appropriate values of G and P so that to minimize the expression:

$$\min_{G,P}\left\{\frac{Var(Y_t)}{\sigma_\varepsilon^2} + a\frac{Var((1-B)X_t)}{\sigma^2}\right\} \tag{5-17}$$

More recently, DelCastillo (2001b) proved that someone can solve (5 -17) for different values of a and for a more complicated disturbance instead of IMA(0,1,1). This disturbance can be modelled as

$$N_t = \omega + N_{t-1} - \theta\varepsilon_{t-1} + \varepsilon_t \qquad |\theta| \leq 1 \tag{5-18}$$

where ω is a drift parameter, very common in industrial processes, were tool - wearing phenomena exist in discrete part manufacturing. DelCastillo, deduced that for constrained PI controllers the results varying and depending on whether the IMA(0,1,1) process has or not a drift. Therefore, in the case were the process has a drift but a no-drift IMA(0,1,1) is implemented the resulting controller cannot compensate for an offset. If the drift is known (i.e previous open loop occurrence) then Box and Luceno (1995) settings (tables of G and P) are optimal.

5.2.4 Optimal PI controllers

Optimal PI schemes are schemes that minimize the output variance. In section 4.3.4.3 we presented the proportional integral (PI) controller as a member of a more general category of PID controllers. The PI controllers are very popular primarily, due to their simple structure and ease of implementation. Furthermore, are quite efficient when process disturbance is described as an ARMA (1,1) or ARIMA (1,1,1) model and process dynamics is a simple first order model (Tsung et al., 1998). Under this framework PI controllers are discussed and their efficiency is compared with MMSE controllers. In the next paragraphs Tsung et al., (1998) assumed that:

a) The process output deviation from target under a discrete feedback control is given by

$$e_t = Y_t + D_t$$

where Y_t is the process output from the process dynamics and D_t is the process disturbance. They also assumed that the target value is equal to zero that is, the output is viewed as the deviation from target

b) The first order dynamic scheme is described as in (5-9)

$$Y_t = c + \delta Y_{t-1} + g(1-\delta)X_{t-1} \quad 0 \leq \delta \leq 1$$

where c, g are constants and δ measures the process inertia.

c) The proportional integral (PI) scheme is formulated using the incremental form (4.3.3.1) as

$$X_t = k_0 - k_P - k_I \frac{1}{1-B} e_t$$

d) The efficiency of the PI controllers is measured using the absolute efficiency criterion (AE) where $AE = \frac{\sigma_a^2}{\sigma_e^2}$ and σ_a^2 and σ_e^2 denote the variances of the white noise a_t and output e_t respectively.

e) The efficiency results are based on the assumption that the disturbance model parameters are known. However, the authors proved that when we substitute these parameters with estimates, the conclusions about the efficiency of PI schemes are still valid.

When a process disturbance D_t is described as an ARMA (1,1) model that is,

$$D_t = \frac{1-\theta B}{1-\phi B} a_t \tag{5-19}$$

where $a_t \sim WN(0,\sigma_a^2)$ and $|\phi|<1$, $|\theta|<1$ and and $\delta=0$ the MMSE is given by

$$X_t = \frac{\theta - \phi}{1-\phi B} e_t \tag{5-20}$$

Furthermore, $k_I \geq 0$ that is, P and PI schemes are considered. In this special case not only PI controllers but P and I controllers, as these described in section 4.3.3, are very high efficient when ϕ is close to θ where $AE>.99$. This is explained because the ARMA(1,1) model reduces to white noise so, no control strategy is optimal. On the other hand when ϕ is close to -1 and θ is close to 1 the PI controllers are loosing their efficiency. In the case where $\phi=0$ the MMSE scheme coincides with a P scheme which is very efficient. When the absolute difference between ϕ and θ ($|\phi-\theta|$) gets large the P schemes are loosing their efficiency.

In the general case, where $\delta>0$, the PI schemes have high efficiency when $\phi=\theta$. This follows from the earlier observation, that process disturbance reduces to white noise for which no control strategy is optimal. When δ was equally to zero PI schemes had high efficiency when $\phi=0$. Now the same result holds when $\phi=\delta$.

There is the case where the MMSE scheme for the ARIMA (1,1,1) model (Box et al.,1994) is given by

$$X_t = \frac{\theta-\phi-1+\phi B}{(1-\phi B)(1-B)} e_t \tag{5-21}$$

Furthermore, always $k_I > 0$ so we cannot consider pure P schemes anymore. The authors studied the case in which one could misidentify a nonstationary ARIMA(1,1,1) with stationary ARMA(1,1) model. They concluded, that PI schemes were optimal, although anyone could misidentify the disturbance model. On the other hand they proved that MMSE schemes have not this robustness property.

In this paragraph PI and MMSE schemes are discussed under the common situation where the disturbance model is described as ARMA (1,1) or as ARIMA(1,1,1) model. Tsung et al., (1998) investigated the efficiency and robustness property of PI controllers in comparison with MMSE controllers. Moreover they also investigated that PI controllers are more robust in the presence of model misspecification than MMSE controllers.

5.3 Tuning for optimization the EWMA controllers

Process adjustment schemes based on EWMA statistic are very popular in the manufacturing of semiconductors. The analysis of this adjustment concentrated on a single EWMA controller and a double EWMA controller applied to a system that exhibits a deterministic trend disturbance. Other types of drift models can provide a better description of this disturbance such as, drift disturbances including a random walk with drift. In these cases the value of smoothing parameters (λ_i) is arbitrarily chosen by the operators in the manufacturing processes. In the next paragraphs optimal values of these parameters that is, optimal weights for single EWMA and double EWMA controllers are derived from the case of particular drift disturbance.

5.3.1 Tuning a single EWMA controller

Recall that the single EWMA controller (4.3.4.2) is given by $X_n = \dfrac{T-a}{b}$ and $\hat{a}_n = \lambda\left(Y_n - \hat{b}X_{n-1}\right) + (1-\lambda)\hat{a}_{n-1}$ where X_n is the manipulated variable and $\hat{a}_n$ is an estimate of a at run n. Ingolfsson and Sachs (1993) showed that if the disturbance is a deterministic trend the asymptotic mean square deviation from target ($AMSD_{DT}$) is,

$$AMSD_{DT} = \lim_{t\to\infty} E\left[(Y - T_t)^2\right] = \frac{2\sigma^2}{2-\lambda\xi} + \frac{\delta^2}{\xi^2\lambda^2} \qquad (5\text{-}22)$$

where $\xi = \beta/b$ is a measure of the bias in the estimate of the gain. Furthermore, they showed that as long as $|1-\lambda\xi| < 1$ the quality characteristic is asymptotically stable. Lately, Smith and Boning (1997) investigated the value of λ that minimizes $AMSD_{DT}$ and satisfies the stability condition which is given by the following equation:

$$\sigma^2\xi^3\lambda^3 - \delta^2\xi^2\lambda^2 + 4\delta^2\xi\lambda - 4\delta^2 = 0 \qquad (5\text{-}23)$$

Assuming now that the drift disturbance is described as random walk with drift (RWD), that is, $D_t = D_{t-1} + \delta + \varepsilon_t$, the $AMSD_{RWD}$ is,

$$AMSD_{RWD} = \frac{\sigma^2}{\lambda\xi(2-\lambda\xi)} + \frac{\delta^2}{\xi^2\lambda^2} \qquad (5\text{-}24)$$

DelCastillo (2001a) proved that the minimization of $AMSD_{RWD}$ with respect to λ is given by:

$$\lambda = \frac{4\delta^2 - \sigma^2 - \sigma\sqrt{8\delta^2 + \sigma^2}}{2\left(\delta^2 - \sigma^2\right)\xi} \qquad (5\text{-}25)$$

DelCastillo compared the quantities $AMSD_{DT}/\sigma^2$ and $AMSD_{RWD}/\sigma^2$ as a function of λ for various combinations of the relative drift $|\delta/\sigma^2|$ and various values of the gain bias $\xi = \beta/\delta$. The author concluded, that for large relative drift both AMSDs behave almost identically indicating that a large value of λ is appropriate. If the relative drift is small, then a process with a deterministic trend (DT) requires a considerably smaller λ than a process with RWD disturbance. The reason for this behavior is that when $\delta \to 0$ the DT model tends to Shewart's model, which implies that there is no drift and a little or no control effort is necessary. On the contrary as $\delta \to 0$ the RWD model tends to a random walk which drifts considerably and a tight control is required to reduce AMSD. In the latter case the optimal weights using the AMSD criterion are such that $\lambda_{RWD} > \lambda_{DT}$.

In addition, DelCastillo (2001), studied the performance of a single EWMA in the presence of a variety of disturbances considering a more general model that is,

$$D_t = \delta + D_{t-1} - \theta\varepsilon_{t-1} + \varepsilon_t, \quad |\theta| \le 1 \tag{5-26}$$

where θ , δ are parameters. This model includes the following particular cases:

- Random walk (RW) process when $\theta = 0$ and $\delta = 0$.
- Random walk with drift (RWD) when $\theta = 0$ and $\delta \neq 0$.
- IMA(1,1) when $\theta \neq 0$ and $\delta = 0$.
- IMA(1,1) with drift when $\theta \neq 0$ and $\delta \neq 0$.
- White noise when $\theta = 1$ and $\delta = 0$.

For this process model he proved that the variance of the adjustments is,

$$Var(\nabla X_t) = \frac{\lambda\left[1+\theta^2 - 2(1-\lambda\xi)\theta\right]}{b^2\xi(2-\lambda\xi)} \tag{5-27}$$

If the cost of adjustments is not negligible the combination of the equations (5 -24), (5 -27) leads to the following optimization model as proposed by Box and Luceno (1997). The objective function can be solved trading off the variability of the adjustments with the output mean square error:

$$\boxed{\begin{aligned} &\min_{\lambda} J = \frac{AMSD(e_t)}{\sigma_\varepsilon^2} + \rho\frac{Var(\nabla X_t)}{\sigma_\varepsilon^2} \\ &\textit{subject to}: |1-\lambda\xi| < 1 \end{aligned}} \tag{5-28}$$

where ρ is the relative cost that needs to be defined by the process engineer.

As an example we assume that the disturbance is described by (5 -26). Suppose that the estimated process gain is $b = 1.3$, the estimated drift is $\delta = 2\sigma$ and the IMA(1,1) parameter is $\theta = .3$. The following table shows the consequent EWMA designs for various values of ρ from 0 to 100.

ρ	λ	AMSD	$Var(\nabla X_t)$
0	1	5.09	0.644
5	0.98	5.22	0.616
10	0.865	6.36	0.456
25	0.718	8.75	0.305
50	0.615	11.55	0.226
100	0.523	15.63	0.168

Table 5.1 EWMA Designs

It is obvious that if a design ignores the variability of the adjustments large values of λ are necessary.

5.3.2 Tuning a double EWMA (DEWMA) controller

In the case where a DEWMA controller is more appropriate than a single EWMA controller and the disturbance is modeled as a deterministic trend DelCastillo (1999) proved that the quality characteristic is stable if and only if the following two conditions are satisfied:

$$\begin{aligned} &\left|1-.5\xi(\lambda_1+\lambda_2)+.5z\right|<1 \\ &\left|1-.5\xi(\lambda_1+\lambda_2)-.5z\right|<1 \end{aligned} \quad \text{where } z=\sqrt{\xi^2(\lambda_1+\lambda_2)^2-4\lambda_1\lambda_2\xi} \qquad (5\text{-}29)$$

In addition the same author proved that equations (5 -29) are also required for a more general system where the disturbance D_t follows a possibly nonstationary ARMA(1,q) model with drift,

$$D_t = \phi D_{t-1} + \delta + \Theta(B)\varepsilon_t$$

where $|\phi| < 1$ is the autoregressive parameter and $\Theta(B)$ is a q-order moving average polynomial.

If the variability of the adjustments $\left(Var(\nabla X_t)\right)$ can be ignored an optimization approach for tuning the DEWMA controllers proposed by DelCastillo (1999, 2001a). This approach is based on the concept that there is a "trade off" between the magnitude of the transient effect and the asymptotic (long run) variance. The term asymptotic or "long-run" is used because in this type of controller there are transient effects due to initialization of the EWMA equations (5 -18) that are accounted separately. In the long run the DEWMA controller eliminates the offset and the process approximately will be on target. A measure of the expected transient up to a specified run with number n is given by the average mean square deviation ($\overline{MSD}$):

$$\overline{MSD}=\frac{1}{n}\sum_{i=1}^{n}E\left(Y_t\right)^2 \tag{5-30}$$

Small values of weights (λ_1 , λ_2), close to zero, make $\overline{MSD}$ very large but they tend to minimize the asymptotic mean square deviation (5 -24). On the other hand large values of the weights, close to one, make AMSD very large but minimize the $\overline{MSD}$. For the deterministic trend drift some complicated expressions are provided by DelCastillo (1999). Using these expressions someone can solve the following equations:

$$\min_{\lambda_1,\lambda_2}\left(w_1 AMSD\left(Y_t\right)+w_2\overline{MSD}\right)$$

$$\text{subject to:}\ 0<\lambda_1 , \lambda_2 \leq 1 \tag{5 -31}$$

The parameters $\left(w_1 , w_2\right)$ are set by the process engineer according to the nature of the process. Specifically if $w_1=0$ and $w_2=1$, an "all bias" solution is obtained that is, a solution that gives all the weight to transient effect; if $w_1=1$ and $w_2=0$ we get the "all variance" solution and if $w_1=w_2=1$, the "trade-off" solution is obtained which provides adequate performance in a variety of cases. As an example suppose that in a deterministic trend case the $a=2$, the drift rate is $\delta=0.1$ and $\sigma^2=1$. Using appropriate software (DelCastillo, 1999) the following table illustrates the "all-bias", "all-variance" and "trade-off" controller designs for $n = 20, 50, 100$

n	w_1, w_2	$AMSD$	$\overline{MSD}$	λ_1	λ_2
	1 1	2.629	0.255	0.001	0.668
20	0 1	3.164	0.2	0.05	1
	1 0	2.622	0.271	0.001	0.619
	1 1	2.625	0.117	0.001	0.647
50	0 1	3.164	0.08	0.05	1
	1 0	2.622	0.122	0.001	0.619
	1 1	2.624	0.069	0.001	0.639
100	0 1	3.164	0.04	0.05	1
	1 0	2.622	0.072	0.001	0.619

Table 5.2 DEWMA Designs

We can see that for all three cases ($n = 20, 50, 100$) the "all variance" solution is the same (AMSD) because it does not depend on n. Furthermore, the "all bias" solution is also giving the same AMSD, because always $\sigma^2 = 1$. It is interesting to observe that the "trade off" solution is giving small values (close to zero) to λ_1 and large values to λ_2. This is an indication that unless the drift rate (δ) is very strong a single EWMA controller is enough.

5.4 The variance constrained PI self tuning controller

Box and Luceno (1995) proposed an optimal PI controller when the variance of the adjustments is constrained below a certain upper limit (section 5.2.3). The same authors noted that for certain values of the PI controller parameters (c_1, c_2), the variance of the adjustments of a quality characteristic has drastically decreased at the expense of a small increase in output variance. Box and Luceno approach requires that the process engineer must specify the relative weight, taking into consideration the adjustment and output variances or the selection of a particular design from a table of input and output variances (obtained by varying the relative weight). In some cases it is desirable for the process engineer to achieve particular maximum adjustment variability (e.g., safety considerations). In these cases a variance constrained self tuning (ST) controller is more appropriate.

Variance constrained self tuning (ST) PID controllers have been proposed by many authors such as Cameron and Seborg (1983), Katende and Jutan (1993). However, these controllers require that the user have to specify the value of the Lagrance multiplier of the adjustment variance constraint. In addition, their controllers are based on recursive least squares estimation and the convergence analysis of these types of controllers is complex. DelCastillo (2000) proposed a variance constrained ST-PI controller where the engineer needs only to specify the maximum desired variance for the adjustments, based on the machine's limitations.

5.4.1 Derivation of the variance constrained ST PI controller

Assuming that a process is described by a first order dynamic system as,

$$Y_t = S_t + N_t = \frac{g(1-\delta)}{1-\delta B} X_{t-k} + \frac{(1-\theta B)}{1-B}\varepsilon_t$$

where X_t denotes the level of a manipulated variable and Y_t the deviation from the target of the quality characteristic at time t, S_t is a dynamic signal and N_t an added noise that follows an IMA(1,1) noise. The sequence of random errors $\{\varepsilon_t\}$, is a white noise process and g,δ are constants.

Furthermore, the corresponding PI controller is derived by the following equation (Box and Luceno, 1995):

$$\nabla X_t = c_1 Y_t + c_2 Y_{t-1} = -G(Y_t + P\nabla Y_t)$$

where the relationship between the parameters (G,P) and (c_1,c_2) is

$$c_1 = -(1+P)G \ and \ c_1 = GP$$

and (G,P) are tuning constants (5 -16).

The proposed controller is based on the Clarke and Gawthrop (1975) approach that were among the first to notice that a minimum output variance requires a large input adjustments. The rationale behind Clarke and Gawthrop approach is to find the adjustment ∇X_t which attempts at each instant time t to bring the k-step ahead forecast of the quality characteristic to target, subject to a constraint on the magnitude of the present control action (MacGregor and Tidwell (1977)).

The proposed ST PI controller is given by the equation:

$$\boxed{\nabla X_t = v_t \left(c_1 Y_t + c_2 Y_{t-1} \right) \qquad (5\text{-}27)}$$

where v_t is an additional tuning parameter and c_1, c_2 are the recommended Box and Luceno parameters that can be taken as initial tuning parameters. Thus, what is proposed is to tune the parameters c_1, c_2 by multiplying them by v_t at each sample. The values of v_t will provide the minimum output variance that a PI controller can provide. This method requires minimum process information with the exception of the system input output delay. The ST controller can be used continuously or when the parameters have converged to their new settings $\left(v^{*}c_1 , v^{*}c_2\right)$ can be used in a conventional controller and the self tuning method can be terminated.

Chapter 6

Integration of SPC - EPC

6.1 Introduction

So far, we presented SPC and EPC control separately. This is not strange as SPC and EPC control were developed in a relatively separate way each other. Both of them have scored significant successes in the context of quality improvement. Several authors have attempted to delineate both approaches into the environment in which each one is best suited. However, the experience suggests that considerable improvement to product quality is often best attainable through an integration of techniques of both methods. This strategy is called integrated process control (IPC). In section 6.2 is discussed the comparison of SPC and EPC control. In section 6.3 is presented the algorithmic statistical process control (ASPC) while in section 6.4 is discussed the SPC monitoring of MMSE and PI controllers. In section 6.5 different approaches of SPC/EPC schemes are presented when a transient disturbance exists. Finally, in section 6.6 a SPC/EPC scheme is presented where the EPC component is a feedforward control and in 6.7 section is discussed an integrated adaptive controller.

6.2 Comparison of SPC and EPC control

Although SPC and EPC have as common goal the reduction of variability, they achieve their aim in different ways. EPC is based on control theory and concerns the yields in products, attempting to keep the process' output close to the target making compensatory adjustments to the process' input. SPC is trying to produce items with the smallest possible variability looking for assignable causes in the process data. EPC has the following drawbacks comparing with SPC (Kramer, 1989):

- Overcompensating disturbances resulting in increased variation.
- Does not remove the root of assignable causes but it uses continuous adjustments to keep process output on target.
- Concealing assignable causes rather than revealing them

On the other hand, SPC alone has limitations. It does not control the system and has a limited success in process industries, due to the nature of the data which are often autocorrelated. There have been disagreements between control engineers and statisticians regarding the effectiveness of EPC versus SPC. This is primarily due to the lack of knowledge about control systems on the part of statisticians and the lack of knowledge of SPC on the part of engineers (Janakiram and Keats, 1998). However, lately there has been a major interest integrating the SPC and EPC so as to improve the quality through further reduction of variability.

Messina (1992) has presented the most known differences between SPC and EPC in a tabular form as follows

	SPC	EPC
Philosophy	Minimize variability by detection and removal of process upsets	Minimize variability by adjustment of process to counteract process upsets
Application	Expectation of process stationarity	Expectation of continuous drift
Deployment		
1. Level	Strategic	Tactical
2. Target	Quality characteristics	Process parameters
3. Function	Detecting disturbances	Monitoring set points
4. Cost	Large	Negligible
5. Focus	People and methods	Equipment
Correlation	None	Low to high
Results	Process improvement	Process optimization

Table 6.1 SPC and EPC comparison (Messina (1992))

It is obvious that Table 6.1 indicates that SPC and EPC have nothing in common. However, MacGregor (1988) suggested that stochastic control theory connects these two fields.

6.3 The Algorithmic Statistical Process Control

Algorithmic statistical process control (ASPC) is an integrated approach proposed by Vander Wiel et al., (1992). With this approach we improve the quality through the appropriate process adjustments (i.e., using a controller) and through the elimination of assignable causes of variability.

6.3.1 The concept of ASPC

Vander Wiel et al., (1992) mentioned, that autocorrelation which is common in continuous process industries is not necessarily bad. They meant that the process is predictable and so there is the possibility of compensation. For example, in a chemical process the raw material mixing is used to make the incoming stock homogeneous. In this case the autocorrelation cannot be eliminated. In this frame, ASPC reduces predictable quality variation using feedback and feedforward procedures and then monitors the complete system to remove inevitable assignable causes of variability. MacGregor (1988) was the first who suggested that SPC charts can be used to monitor the performance of a controlled system. Another researcher (Barnard, 1959) that was among the first to suggest that is preferable to make process adjustment at the signals of control charts.

6.3.2 The batch polymerization example

Vander Wiel et al., (1992) advocated their proposal through an application of a typical chemical process. This chemical process produces a polymer resin where the polymerization occurs in five batch reaction lines. The quality characteristic is the intrinsic viscosity of the polymer which is measured after completing each batch. The objective is to minimize the viscosity variation about a target level T. The level of viscosity depends on the amount of catalyst added in any batch. Process measurements from previous batches are used by experienced operators to adjust the amount of catalyst in future batches based on general guidelines and Shewart's chart approach. The authors suggested modeling the measurement viscosity of batch t using a simple autoregressive moving average transfer function (ARMAX) model as

$$Y_t = bX_{t-1} + \frac{(1-\theta B)}{(1-\rho B)} e_t \tag{6-1}$$

where Y_t is the observed viscosity deviation from target, X_{t-1} is the catalyst deviation from nominal and θ, ρ are the parameters of the ARMAX model. Furthermore, they used alternatively to Shewart's chart, the MMSE criterion in order to specify the amount of catalyst minimizing the $MSE(Y_t)$. The authors studied the following two cases:

In the situation where no delays exist a MMSE controller for the model (6-1) will be optimal producing Y_t with the smallest MSE. This controller is described as

$$\boxed{X_{t-1} = \rho X_{t-2} - \left[\frac{\rho - \theta}{b}\right] Y_{t-1}} \tag{6-2}$$

If we substitute (6-2) in (6-1) reveals that the control action results in the process is $Y_t = e_t$. There are two advantages of this control action compared with Shewart's chart. Firstly, it is simple and secondly it can reduce viscosity variation even further.

If a measurement is delayed (6-2) is not appropriate. However, it is possible to minimize the MSE as follows

$$\begin{gathered} \hat{Y}_{t-i} = bX_{t-1} + \hat{N}_{t-i} \\ N_t = \frac{(1-\theta B)}{(1-\rho B)} e_t = Y_{t-i} = bX_{t-1} \end{gathered} \tag{6-3}$$

where $\hat{Y}_{t-i}$ is the i-step minimum MSE forecast of Y_t, $\hat{N}_{t-i}$ is i-step minimum MSE forecast of the noise term. By setting $\hat{Y}_{t-i} = 0$ results in $X_{t-1} = -\hat{N}_{t-i}/b$.

Up to this point, our major concern was the algorithmic part of ASPC as applied to the process using transfer function models and stochastic control. Nevertheless, in ASPC the statistical monitoring plays a significant role with the purpose of detection and signaling when the closed loop process is not consistent with the estimated model and the control algorithm. In the polymerization application an abrupt change of the nominal amount of catalyst required to produce the target viscosity has changed (i.e. due to resuppling in the tanks of raw material feeding the reactors). In this situation where after a "shift" there is short transient period until the output mean stabilizes in a new level, the authors applied a CUSUM chart. The main purpose of CUSUM chart was to detect such shifts as quickly as possible when they occurred in order to determine if any of the devices was really responsible.

The authors proposed a general four step procedure as follows:

- Develop a time series transfer function model for the process output including the effect of past performance.
- Design a control rule for the estimated model taking into account the cost.
- Use SPC charts to monitor the closed loop process.
- When a monitoring signal occurs and an assignable cause really exists remove it. If there is not an assignable cause, estimate again system parameters or even identify again the process and return to second step.

6.4 SPC monitoring of MMSE and PI controllers

The performance of SPC monitoring an APC controlled process depends on the monitoring data stream (the output of the control action), the APC control scheme and the underlying autocorrelated process. Jiang and Tsui (2002) studied and compared the performance of MMSE and PI controlled process, monitoring the output and the control action of them. Their study is based on the signal to noise (SN) ratio.

This ratio was introduced by Jiang et al., (2000) in order to study the average run length (ARL) of a control chart when a process shift occurs. The authors introduced the transient and the steady state signal to noise (SN) ratio as follows

$$R_T^z = \frac{\mu_T^z}{\sigma_z} \quad and \quad R_S^z = \frac{\mu_S^z}{\sigma_z} \tag{6-4}$$

where σ_Z is the standard deviation of the charted statistic and μ_T^Z *and* μ_S^Z are the transient (when the process begins, $t=0$) and steady (asymptotic $t=\infty$) mean shift level of the charted process respectively. The transient SN ratio measures the ability of the chart to detect a shift in the first few runs (transient state) while the steady SN ratio measures the efficiency of the chart to detect the shift in the later runs.

The appropriate monitoring chart is selected among several candidates using the two SN ratios with an ad hoc manner. The chart with the higher R_T value is often preferable if its R_T value is high enough (usually more than 4). When all the candidates have a relatively small chart (say less than 3) the chart with the larger R_S value is preferred even if its R_T may be smaller. When all the charts have moderate values of the two ratios their levels are comparable.

6.4.1 SPC monitoring of MMSE controlled process

Suppose that the model which describes the dynamic behavior of the quality characteristic and noise disturbance effects is

$$Y_t = D_t + c(B)X_{t-1} \tag{6-5}$$

where Y_t is the output, X_t is the manipulated variable employed with a transfer function $c(B)$ and the disturbance term D_t is a stationary ARMA(1,1) process which formulated as

$$D_t = \phi D_{t-1} + \alpha_t - \theta\alpha_{t-1} \tag{6-6}$$

where $a_t \sim WN\left(0, \sigma_a^2\right)$.

For an ARMA(1,1) disturbance model the MMSE controller is defined as $X_t = \dfrac{\theta-\phi}{1-\phi B} e_t$ (4-20) and the standard deviation of the control action is

$$\sigma_X = \frac{|\theta-\phi|}{\sqrt{1-\phi^2}}\sigma_a \tag{6-7}$$

Consider the case where the step mean shift η takes place at time $t=0$ that is,

$$Y_t = D_t - X_{t-1} + \eta_t \text{ where } \eta_t = \begin{cases} 0 & t<0 \\ \eta & t \geq 0 \end{cases}$$

It can be proved that the transient and steady state mean shifts are,

$$\mu_T^Y = \eta \ , \ \mu_S^Y = \frac{1-\phi}{1-\theta}\eta \text{ and } \mu_T^X = (\phi-\theta)\eta \ , \ \mu_S^X = \frac{\phi-\theta}{1-\theta}\eta$$

respectively.

Applying a Shewart's chart to study the performance of these two data streams and assuming for simplicity that $\eta = \mu\sigma_Y$ the SN ratios (6-4) reduce to

$$R_T^Y = \mu \ \ R_S^Y = \mu\left|\frac{1-\phi}{1-\theta}\right| \text{ and } R_T^X = \mu\sqrt{1-\phi^2} \ \ R_S^Y = \mu\frac{\sqrt{1-\phi^2}}{1-\theta} \qquad (6\text{-}8)$$

It is obvious that $R_T^Y \geq R_T^X$ since $\sqrt{1-\phi^2} \leq 1$. According to the rule which described above, the value of a large value of R_T is important to detect large shifts; otherwise the efficiency of the charts will depend on the values of R_S. In this case the values of R_S are appropriate. When $\phi > 0$ the steady state ratio becomes larger than that of the output chart and it is expected that monitoring the control action will be more efficient than the process output. That is, a process with a positive value of ϕ will drive the MMSE controller to reduce the output shift level which makes detection difficult. In the same way a negative value of ϕ will increase the output shift level.

6.4.2 SPC monitoring of PI controlled process

Assuming that a PI controller (4.3.3.3) is described as

$$X_t = K_P e_t + K_I \sum_{j=1}^{t} e_j .$$

For this type of controller when a mean shift occurs, monitoring the output and monitoring the control action results in very different outcomes at time $t=0$. The transient and steady state mean shifts of the process output and the control action are given by

$$\mu_T^Y = \eta \ , \ \mu_S^Y = 0 \text{ and } \mu_T^X = (k_P + k_I)\eta \ , \ \mu_S^X = \eta \qquad (6\text{-}9)$$

respectively. It is essential to see that the output converges to zero due to the I component. That implies that the mean shift is completely compensated for the steady state. This is a disadvantage for SPC monitoring since the mean cannot be detected. Using simulation the authors showed that when P and I components are used, monitoring the control action is more efficient than monitoring the output. On the other hand, if the I component equals to zero the output chart is preferable than the control action chart.

6.5 Combining SPC and EPC control when a transient disturbance exists

It is common in many industrial processes to detect a disturbance due to the unforeseen incidents and the intentional adjustments. Changes in product design (short production runs), fluctuations in material properties (such as chemical impurities) and power or equipment failure results in a large portion to out of specification product. More generally transient disturbances typically fall into two categories:

- Step change disturbances (e.g., from adjustments in the process),
- Linear change disturbances or ramps (e.g., from a tool wearing) Shao (1998).

In a combined SPC/EPC control, the SPC approach is more than a process monitor. The SPC method is incorporated into the EPC control algorithm so that it can trigger the controller to make adjustments in the input (Nembhard and Mastrangelo, 1998).

6.5.1 Effective integrated process control in a transient period using simulation and an ARMA (p,q) disturbance model

Nembhard (1998) used simulation models in order to study the performance of a system when a combination of SPC and EPC control is employed. She called this combination integrated process control (IPC). The author used two simulation models, where the first one represents a noisy dynamic system and the second one represents the designs that integrate EPC and SPC control. She also used three different policies; Policy A and B use an integrated process control while Policy C involves only EPC control. In more specific words these policies are described as:

- Policy A: PI/Shewart-AE(activate and evaluate). When an observation exceeds the limits the PI controller is activated to take over and make adjustments to the process.
- Policy B: PI/Shewart-AR(activate and regulate). In addition to the policy A the Shewart policy ceases to allow PI adjustments after s observations are within the control limits of the new target level.
- Policy C: PI only.

The control policies are evaluated, based on four performance measures: Average squared error of the output from target, number of adjustments, average magnitude of adjustments and number of alarms.

Furthermore, the author supposed that the noise process can be modeled by a stationary and invertible ARMA(p,q) process

$$Y_n = \xi + \sum_{i=1}^{p} \phi_i Y_{n-i} + X_n - \sum_{j=1}^{q} \theta_j X_{n-j} \qquad (6\text{-}10)$$

where ξ is constant, Y_n , X_n is the output and the manipulated variable respectively and ϕ , θ are appropriate AR and MA coefficients. The dynamic process is described by a first or second order linear ordinary differential equations as

$$\frac{Y(s)}{X(s)} = \frac{K/\tau}{s+1/\tau} \quad and \quad \frac{Y(s)}{X(s)} = \frac{K_1 K_2 / \tau_1 \tau_2}{s^2 + \left(\tau_1 + \tau_2 / \tau_1 \tau_2\right)s + \left(1/\tau_1 \tau_2\right)} \qquad (6\text{-}11)$$

respectively where K is the steady state process gain and τ is the process time constant (the speed of the response).

Some conclusions of the author's proposal are the following:

- When a first order process models the noise process the PI controller is sufficient to maintain the process output at the target level minimizing the MSE.
- When the adjustment cost is of a primary concern and a first order process is used policy B gives the best performance.
- When a second order process is appropriate to model the noise process neither PI controller nor Shewart chart produces minimum variance. However,

an integrated policy either A or B is recommended, where the minimum variance and/or the number of adjustments is of our concern.

- Since a PI controller is not optimal for a second order process it is possible that the Shewart limits will be exceeded in this system. Therefore policy A is preferable.

6.5.2 IPC design for startup operations with ARIMA (0,1,1) disturbance model.

In the previous paragraph an effective IPC policy using simulation was discussed, assuming that the noise process is described by a stationary ARMA (p,q) model. Nembhard et al., (2001) studied the case where the noise process is represented by an ARIMA model and in particular an IMA (1,1) model. Furthermore, they proposed an IPC mechanism which combines a PID controller and a moving centerline exponentially weighted moving average (MCEWMA) chart for individual observations (section 2.4.1.1). The MCEWMA chart adapts the EWMA for the autocorrelated data given by an IMA(1,1) disturbance model (Montgomery and Mastrangelo, 1991). The main measure of improvement is to reduce the sum of squared errors (SSE) from target, where the target changes when the transition begins.

Consider that the system is described by a discrete first order process as

$$Y_n = aY_{n-1} + bX_{n-1}$$

where Y_n, X_n is the process output and input respectively a measures the degree of inertia for the process dynamics and b is a constant. Additionally, the noise process is described as

$$(1-B)Y_t = (1-\theta B)\varepsilon_t$$

where θ is the moving average operator and $\varepsilon_t \sim N\left(0,\sigma_\varepsilon^2\right)$.

The EPC policy is a PID controller (4.3.3.1) which is formulated as follows

$$X_t = K_P e_t + K_I \sum_{j=1}^{t} e_j + K_D \nabla e_t$$

and the SPC policy involves the MCEWMA (2.4.1.1) chart.

The author studied three control policies; the IPC, the EPC and the SPC policies and compared them using three metrics:

- The sum of squared errors (SSE)
- The number of adjustments
- The magnitude of adjustments

According to the SSE criterion the sum of squared errors is smaller when EPC policy is used compared with IPC and SPC policies. This is an expected result because EPC is designed to minimize the sum of squared errors. If SSE is the only criterion the EPC policy always gives the best performance.

According to the second criterion, the SPC policy does not make any adjustment, EPC policy makes an adjustment in every period and IPC policy makes an adjustment only when the observation is outside the limits of the MCEWMA. Therefore, when the cost is not negligible the IPC policy is preferable.

Finally, in relation with the third criterion when the number of adjustments is small for the IPC policy, the size of adjustments is small as well.

Generally, the EPC approach minimizes the error but the number and the magnitude of adjustments can be reduced using the integrated monitoring and control approach.

6.5.3 Transient disturbances with Cuscore control charts

So far we discussed the integrated process control schemes when a transient disturbance is present, where the SPC component is a usual Shewart chart. Nevertheless, Shewart's charts are not sensitive in minor shifts and CUSUM charts respond slowly to large process shifts (Montgomery 2001). Shao (1993) showed that the Cuscore control chart (2.4.1.3) is effective not only in detecting process shifts but also in identifying transient disturbances. In the next subparagraphs the effective integration of SPC and EPC, when a transient disturbance exists, is discussed through the use of Cuscore control as the SPC component (Shao, 1998).

The integral and PI controllers are typical in industry. Furthermore, the MMSE controller is equivalent to the integral and PI controllers for a zero and first order noise system which follow an IMA(1,1) model. Shao (1998) proposed in his study the MMSE controller instead of integral and PI controller for feedback control action (EPC) in a closed loop system. When a step change disturbance enters the process the author proved that the output deviation is

$$Y_{t+i} = D\theta^{t-1} + a_t$$

where Y_t is the output of the process, D is the magnitude of the level of the step change and θ is the moving average operator. It can be shown that the Cuscore statistic (2.4.1.3) can be written as

$$Q_0 = \sum_{t=1}^{n} Y_t D\theta^{t-1} \tag{6-12}$$

In the case where a linear change transient disturbance exists the author proved that the output deviation from the target is derived by

$$Y_{t+i} = b\left(\frac{\theta^i - 1}{\theta - 1}\right) + a_{t+i}$$

where a is the magnitude of the level and b is the slope when the linear disturbance is reformed as $D_t = a + bi$ while i stands for the time when the transient disturbance started affecting the process. Thus, the cuscore statistic would be

$$Q_0 = \sum_{t=1}^{n} Y_t b\left(\frac{\theta^t - 1}{\theta - 1}\right) \tag{6-13}$$

Shao (1998) used as a performance metric (PM) the average squared deviation from target (T), that is, $PM = 1/n \sum_{t=1}^{n} (Y_t - T)^2$. Additionally, using simulation, he compared the cuscore control chart with three different control charts:

- Shewart chart for individuals
- Cusum chart
- EPC only (no chart)

The simulated results recommended that the combined EPC/SPC scheme has smaller PM than EPC alone. In addition, the PM for the EPC/Cuscore scheme seems to be smaller than EPC/Shewart and EPC/CUSUM when a negligible trend magnitude 0.1 exists. However, the results of the simulation indicated that the cuscore control chart had the shortest average time for detecting the linear change disturbances.

6.5.4 An EWMA Bounded adjustment scheme with adaptive noise variation

One of the basic assumptions, when an integrated EPC/SPC is applied in a process, is that the standard deviation of the $WN\left(0,\sigma_a^2\right)$ component of the measurement is known and remains constant. However, some processes can operate on a short run basis and previous measurements are not available to adequately determine the variance of the disturbances. To facilitate the control of this kind of process, Nembhard and Mastrangelo (1998) and Nembhard et al., (2001) were based in a MCEWMA chart in which control limits of varying magnitude are established about a moving centerline. EPC was added in the form of a PID controller in order to adjust the process when a measurement exceeds one of the control limits. In that case, the control limits were based on the magnitude of the standard deviation of the EWMA values as it was previously estimated using an estimate of σ_a. O'Shaugenessy and Haugh (2002) proposed a recursive estimation method of the process standard deviation where the ERC/SPC scheme is combined by an MCEWMA chart (Montgomery and Mastrangelo, 1991) as SPC component and an integral control action as EPC component. The advantage of this method is based on little or no even previous data in order to estimate σ_a.

Consider as usual that a process measurement Y can be affected by a stochastic component D and a component driven by a manipulated variable X. Thus, assuming that the disturbance term is described by an IMA(1,1) model,

$$D_t = D_{t-1} + a_t - \theta a_{t-1}$$

where a_t is white noise series and θ is the MA operator and the feedback controller is a pure integral controller as

$$X_t = X_0 + k_t \sum_{i=1}^{t} (Y_i - T)$$

where X is determined to offset the error between the Y and the target T. Furthermore the optimal predictor $\hat{Y}$ of Y is given by

$$\hat{Y} = \lambda Y_t + (1-\lambda)\hat{Y}_{t-1}$$

where λ is the "smoothing constant" of the EWMA statistic. Additionally the standard deviation σ_e of the prediction error $e_t = Y_t - \hat{Y}_{t-1}$ will be equivalent to σ_a.

O'Shaugenessy and Haugh (2002) proved that the upper and lower control limits of the MCEWMA are:

$$\begin{aligned} UCL_t &= \hat{Y}_t + (1.25K)\hat{M}_t \\ LCL_t &= \hat{Y}_t - (1.25K)\hat{M}_t \end{aligned} \tag{6-14}$$

respectively, where K defines the number of standard deviations used to set the width of control limits and $\hat{M}_t$ the estimated mean absolute deviation (M) of the forecast error from the mean of errors.

The iterative procedure is based on the following steps

- Starting values of $\hat{Y}$ and $\hat{M}$
- Estimation of σ_e using $1.25\hat{M}_t$
- Evaluation of UCL and LCL
- If the process measurement exceeds the limits a control action takes over.

Generally, there are two conditions that require an adjustment of the process. The first, when the process drifts and is inevitable if the process is nonstationary; and the second, when a large disturbance or assignable cause perturb the system.

The authors used simulation and showed that their proposal not only is best suited when a noise term is represented as a nonstationary time series but it can effectively reduce the variance of measurements associated with stationary disturbances. Furthermore, as the algorithm is based on recursively estimating the process standard deviation, the process control approach is appropriate during startup conditions as well as during processes that exhibit noise variances that change over time. However, appropriate values of a, λ, and K can dramatically influence the performance of the MCEWMA chart. Thus, a value of $\lambda = 0.5$ will minimize the possibility of overestimating the noise standard deviation. Any value of a greater than 0.1 will exhibit increasingly larger variation. Small increases in the K value will result in a large reduction of the number of adjustments without a corresponding large increase in the output variance.

6.6 An Integrated SPC/EPC scheme with feedforward control

Feedforward controller (4.3.7.1) is a particular type of controller in EPC. If a known input variable, say Z_t, can be measured and proper relationships are made among the input variable, the compensating (manipulated) variable X_t and the desired output Y_t then a feedforward control can be developed. For example, consider again the case of the polymerization process (6.3.2) where the input feed stream (Z_t) can be measured and its effect to polymerization resin (Y_t) at some future time is predicted. After that, an adjustment in catalyst feed rate (X_t) results in offsetting the potential effect to the polymerization output. Montgomery et al., (2000) studied this case in a chemical process, where the output was an intermediate product used for the production of a synthetic resin, through two different scenarios. A chemical mixture of varying concentration was the input of the system and the compensating variable was the pressure, that is, the feeforward control. Furthermore, they applied as an SPC scheme a CUSUM and a EWMA chart without disqualifying other monitoring schemes such as Cuscore control charts. In addition, they supposed that the input feed concentration was described by an IMA (1,1) model. Therefore, it could be reasonably forecasted by a EWMA statistic. Both of these scenarios were studied through simulation and the average run length (ARL) performance of the combined procedures observed.

6.6.1 The shift scenario

The input feed concentration varies over time. A shift in the concentration is characterized by a sudden change (i.e., a freezing valve) that is beyond the process usual variability at a specific period of time t. Thus, an unshifted and shifted disturbance process can be described as,

$$\begin{aligned} Z_t &= \left(Z_{t-1} + e_t - \theta e_{t-1}\right) \\ Z_{tshift} &= \left(Z_{t-1} + e_t - \theta e_{t-1}\right) + k \end{aligned} \qquad (6\text{-}15)$$

respectively, where θ is the moving average operator and k a constant term which represents the shift. At time $t{=}t_{shift}$ the shift of k units is introduced into the process and after several runs the disturbance variable Z_t was recorded. The controller takes over and the process comes back in its previous state but at the expense of increased compensations actions.

6.6.2 The trend scenario

A trend in the concentration is represented by a gradual every - time - unit change in concentration (i.e., a constant valve wearing) that is well beyond the usual variability of the process. The trend can be introduced into the process in changing e_t of the equation (6-15) to $N(\mu,1)$ at time $t{=}t_{trend}$. The larger the value of μ, the greater the slope of the trend. Unlike with the shift scenario, the controller cannot completely compensate the trend and the controlled output comes at the expense of increasingly larger and probably more costly compensations actions.

6.6.3 Integrating SPC and feedforward control

As we saw, in the presence of an assignable cause in the first case (shift scenario) the feedforward controller achieved to come back the process in its previous state but at the expense of increased compensation actions and in the second case (trend scenario) the controller never fully compensates the disturbance. The authors used EWMA and CUSUM charts, because of their effectiveness in detecting small shifts, as an SPC scheme and the performance of one was evaluated. They concluded that generally CUSUM performs better than EWMA in the sense that (ARLs) are smaller for a wide range of shift/trend magnitudes.

6.7 Integration of SPC/EPC using adaptive controllers

Considerable research has been devoted to the advantages of integrated SPC and EPC schemes. However, little research focused on ways to identify and compensate for the source of process disturbance. The majority of the research assumed that assignable causes may be eliminated immediately after an out of control signal is triggered by the SPC scheme (Montgomery et al., 1994). In fact, the immediate elimination of assignable causes is impossible. Shao et. al., (1999) proposed an adaptive EPC/SPC scheme to counteract the effects of the disturbance during the interim period between the initial appearance of a disturbance and the eventual elimination of its assignable causes.

Shao's proposal extends in the case where the underlying disturbance is a step –change or a linear change disturbance. Furthermore, the author assumed that the process is described by a usual first order process with IMA (1,1) noise process that is,

$$
\begin{aligned}
Y_{t+1} &= \frac{q}{(1-pB)} X_t + d_{t+1} \\
d_{t+1} &= \frac{(1-\theta B)}{(1-B)} a_{t+1}
\end{aligned}
\tag{6-16}
$$

where Y_{t+1} is the output deviation from target at time $t+1$, X_t is the manipulated variable at time t, p,q are fixed parameters where $p+q=1$ (Box and Luceno, 1997), d_{t+1} is an IMA(1,1) process, θ is the parameter of the IMA(1,1) process and a_{t+1} is a white noise sequence. For the model in equation (6-16) minimizing the variance of the output deviations can be achieved by the MMSE controller where,

$$X_t = -\frac{(1-\theta)p}{q}\left[Y_t + \frac{(1-p)}{p}\sum_{j=\infty}^{t} Y_j \right] \qquad (6\text{-}17)$$

In this case the MMSE control is an optimal PI control because the process output deviations can be shown to follow a white noise sequence.

Shao (1993) proved that the results when a step change or a linear change disturbance occurs are:

$$Y_{t+1} = L\theta^{i-1} + a_{t+1}$$
$$Y_{t+1} = \delta\left(\frac{-1+\theta^{i}}{\theta-1}\right) + a_{t+1} \qquad (6\text{-}18)$$

respectively where L is the level of the step change disturbance, δ is the trend rate of the linear disturbance and i stands for the elapsed time since the introduction of the disturbance into the process.

6.7.1 Guidelines of detection and identification

In addition to EPC component (MMSE controller) the author introduced a control chart for individuals as SPC component. When the process exhibits the control limits and the out of control alarm is triggered the identification mechanism is activated to examine the pattern of the output deviations.

(i) Step change disturbance case

The first equation of (6-15) suggests that when a step change disturbance occurs the output deviation values would return to a_t at rate θ. The return rate θ has a negligible effect on the subsequent output deviations at time $i+5$ and later. Once the out of control signal is triggered by SPC we start to collect six observations. If 4 or more of 6 observations are within the values of $0 \pm \sigma$ then we conclude that the underlying disturbance is a step change disturbance; otherwise is a white noise sequence. If it is a step change disturbance Shao (1993) showed that the optimal controller is,

$$X_t = X_{t-1} - \frac{(Y_t - pY_{t-1})}{q} \quad (6\text{-}19)$$

(ii) Linear change disturbance case

The second equation of (6-18) indicates that the occurrence of a linear disturbance results in a jump of the output deviation and the subsequent output values approach an approximately constant value $\delta/(1-\theta)$. Consequently, the differences between $\delta/(1-\theta)$ and the output deviations at time $i+5$ and later should behave as a white noise sequence. The same technique as above could be used to detect a linear disturbance. If a linear disturbance really exists, Shao (1993) proved that the optimal controller is,

$$X_t = 2X_{t-1} - X_{t-2} - \frac{\left[2Y_t - (1+2p)Y_{t-1} + pY_{t-2}\right]}{q} \quad (6\text{-}20)$$

6.7.2 Conclusions

What Shao proposed in his study could be considered as a sequence of the following steps:

- The process is tuned by a MMSE controller and monitored by an individual control chart.
- The underlying disturbances are detected and are identified accordingly to the above guidelines.
- The adaptive control is activated to compensate for the effects of the disturbance during the short-term period between the initial detection of the disturbance and its elimination.
- The same controller (MMSE controller) is used to compensate for noise after the assignable causes have been eradicated.

Chapter 7

Multivariate SPC - EPC

7.1 Introduction

In the previous chapters we analyzed the univariate process in which there is only one process output variable or quality characteristic. In reality, however, many industrial processes involve several related variables. In this chapter we present SPC and EPC techniques for multivariate processes. In case where SPC is employed, we assume that two or more quality characteristics are of our concern and when EPC is applied we assume that exist multiple controllable factors (inputs). The general case deals with a process where multiple inputs and multiple outputs (MIMO) define the quality of a product. In section 7.2 are presented the most usual multivariate control charts. In section 7.3 the EPC for multivariate process is discussed, while in section 7.3 the integration of EPC using joint monitoring SPC techniques is presented.

7.2 The multivariate SPC

Most industrial processes are described by several quality characteristics that require monitoring. These quality characteristics are obviously correlated and control charts for monitoring individual quality characteristics may be misleading for detecting changes in the overall quality of the product (Mastrangelo et al., 1996; Liu, 1995). Furthermore, the overall "Type I error" decreases, as the number of variables increases (Montgomery, 2001). Thus, it is desirable to have control charts that can monitor multivariate measurements simultaneously (Lowry and Montgomery, 1995).

7.2.1 The Hotelling T^2 control chart

The most widely known multivariate control chart is the Hotelling T^2 control chart for monitoring the mean vector of a process. This type of chart was introduced by Hotelling (1947) who applied his procedures to bombsight data during World War II. This chart could be applied in processes with individual or subgrouped data and it is analogous to t-statistic in the univariate case.

We assume that p related quality characteristics are jointly monitored. Furthermore, it is assumed that the joint probability of these quality characteristics is $p-$ variate normal distribution and a set of the quality characteristic means is represented by the $p\times 1$ vector $\overline{\mathbf{x}}^{\mathrm{T}}=\left[\overline{x}_1,\overline{x}_2,\ldots,\overline{x}_p\right]$.
The Hotelling T^2 control chart is given by the formula:

$$T^2=n\left(\overline{\mathbf{x}}-\overline{\overline{\mathbf{x}}}\right)^{\mathrm{T}}\mathbf{S}^{-1}\left(\overline{\mathbf{x}}-\overline{\overline{\mathbf{x}}}\right) \tag{7-1}$$

where $\overline{\overline{\mathbf{x}}},\mathbf{S}$ are the estimates of the vector of in-control means $\boldsymbol{\mu}^{\mathrm{T}}=\left[\mu_1,\mu_2,\ldots,\mu_p\right]$ and the covariance matrix $\boldsymbol{\Sigma}$ respectively and n is the sample size.

Alt (1985) proposed two different phases to select the control limits of a Hotelling T^2 statistic, known as Phase I and Phase II. The first phase utilizes a large sample of data so that the parameters and control limits are well estimated for Phase II. Therefore, in Phase II the control chart is used for monitoring the process (Mason et al., 1997; Montgomery, 2001). The control limits of Phase I and II are given by the following equations.

Phase I	Phase II
$UCL=\dfrac{p(m-1)(n-1)}{mn-m-p+1}F_{a,p,mn-m-p+1}$ $UCL=0$	$UCL=\dfrac{p(m+1)(n-1)}{mn-m-p+1}F_{a,p,mn-m-p+1}$ $UCL=0$

where p is the number of variables, m is the number of preliminary samples and n is the sample size.

When the subgroup size is, $n=1$ the Hotelling T^2 statistic becomes

$$T^2 = n(\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\overline{\mathbf{x}}) \tag{7-2}$$

The T^2 chart is obtained by plotting successive T^2 values versus time. On this chart there is only an upper control limit (UCL). In figure 7.1, as an example, a T^2 chart is presented with upper control limit 13.207 and 14.579 for Phase I and II respectively and the significance level is, $\alpha = 0.001$. Data (Mitra, 1998) correspond to a process with two quality characteristics ($p=2$), 20 sample means ($m=20$) with sample size $n=4$. A point outside the control region exists in the 9th sample.

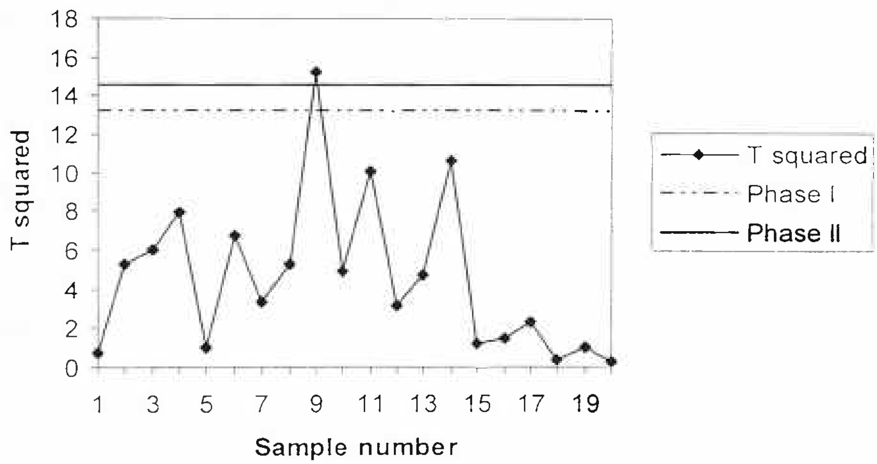


Figure 7.1: A T^2 control chart with Phase I and II control limits

By using the Hotelling T^2 statistic we have a single control limit which produces a Type I error and uses the available information with respect to means, variances and correlation between the variables.

7.2.2 The Multivariate CUSUM and EWMA control charts

The Hotelling T^2 control chart, which was described above, uses information only from the current sample and so it is insensitive to small and moderate shifts in the mean vector. Therefore, multivariate CUSUM (MCUSUM) and EWMA (MEWMA) charts are developed when the detection of small changes in the mean of the process is important. The multivariate schemes of CUSUM and EWMA are either directionally invariant or direction specific (Lowry and Montgomery, (1995)). The average run length (ARL) of a directionally invariant chart is established by the distance of the off target mean and the on target mean, while the direction specific chart's ARL is derived by this distance and the direction that the off target mean is relative to on target mean. Generally, MCUSUM and MEWMA charts are direction specific charts. The most widely used ones, are presented in the following paragraphs.

7.2.2.1 The Multiple Univariate Cusum chart

Woodall and Ncube (1985) proposed the simultaneous use of several CUSUM procedures so that a single multivariate CUSUM procedure is considered. Assuming that independent p-variate random variables $\mathbf{X}_n = \left(X_{1n}, X_{2n}, \ldots, X_{pn}\right)^T$, $n=1,2,\ldots$ observed successively, with known variance - covariance matrix Σ, a one- sided or two – sided CUSUM procedure can be applied to each sequence of random variables $\{\mathbf{X}_{in}\}$, $i=1,2,\ldots,p$. The i^{th} two-sided CUSUM procedure is derived by the following cumulative sums:

$$\begin{aligned} S_{i,n} &= \max\left(0, S_{i,n-1} + X_{i,n} - K_i\right),\ 0 \le S_{i,0} < h_i \\ T_{i,n} &= \min\left(0, T_{i,n-1} + X_{i,n} + K_i\right),\ -h_i \le T_{i,0} < h_i \end{aligned} \tag{7-3}$$

where K_i is the reference value of the i^{th} variable and h is the decision interval. This chart signals when either of the p CUSUM procedures signals, that is, determining the minimum run length $N(i)$ where for a two - sided procedure is,

$$N(i) = \min\left\{n : S_{i,n} \ge h_i \ or\ T_{i,n} \le -h_i\right\} \tag{7-4}$$

Healy (1987) extended the results of Woodall and Ncube by introducing an MCUSUM chart based on a linear combination of the univariate variables. This is a CUSUM of T^2 chart that may have a better ARL performance when the goal is to detect a shift in the covariance matrix.

7.2.2.2 The Grosier's MCUSUM charts

The ARL of Woodall and Ncube(1985) procedure depends on the direction that the mean vector shifts. In contrast, the ARL of the multivariate T^2 chart depends on the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ only through the non centrality parameter $\mathbf{d}$. Based on this result, Grosier (1988) proposed two multivariate CUSUM charts. The first one, is called multivariate CUSUM of T (COT) and is based on the squared root of T^2 statistic, while the second can be derived by replacing the scalar quantities of the univariate CUSUM chart with vectors (CUSUM vector scheme).

Assuming that $\mathbf{x_n}$ is a p-variate observation with known variance – covariance matrix $\boldsymbol{\Sigma}$ and mean vector $\boldsymbol{\mu}$. The CUSUM (COT), which is proposed by Grosier is given by:

$$S_n = \max\left(0, S_{n-1} + T_n - K\right) \tag{7-5}$$

where T_n, $n = 1,2,...$ the scalars of CUSUM and S_0 , $K \geq 0$

The COT chart signals when $S_n \geq h$. The decision interval h is computed using the Markov chain approach. Furthermore the COT chart is a directionless chart.

The second procedure, the CUSUM vector chart, is produced directly from the observations, that is,

$$Y_n = \sqrt{\mathbf{S_n^T \Sigma^{-1} S_n}} \tag{7-6}$$

where,

$$\mathbf{S_n} = 0 \qquad\qquad if\, \mathbf{C_n} \leq K$$
$$\mathbf{S_n} = \left(\mathbf{S_{n-1}} + \mathbf{x_n} - \boldsymbol{\mu}_0\right)^T \left(1 - K/\mathbf{C_n}\right) \; if\, \mathbf{C_n} \geq K$$

and

$$\mathbf{C_n} = \left[\left(\mathbf{S_{n-1}} + \mathbf{x_n} - \boldsymbol{\mu}_0\right)^T \Sigma_0^{-1} \left(\mathrm{S_{n-1}} + \mathrm{x_n} - \mu_0\right)\right]^{1/2}$$

This chart signals when $Y_n > h$. Both of these charts give faster detection of small shifts in the mean vector than multivariate T^2 charts. The CUSUM vector chart is more preferable than COT chart, because of faster detection in shifts of mean vector its direction specific property.

7.2.2.3 The MC1 and MC2 charts

Pigniatello and Runger (1990) introduced two multivariate CUSUM charting procedures. Both of these procedures are based on quadratic forms of the mean vector. Their difference, focuses on the point at which the accumulation is made. That is, MCUSUM#1(MC1) accumulates the **X** vectors before producing the quadratic forms, while MCUSUM#2(MC2) calculates a quadratic form for each **X** and then accumulates those quadratic forms.

The MC1 chart can be constructed as follows:

$$MC1_t = \max\left\{\left\|C_t\right\| - Kn_i, 0\right\} \qquad (7\text{-}7)$$

where a usual choice of $K = 0.5 \times \lambda^2 \boldsymbol{\mu}_1$, $\boldsymbol{\mu}_1$ is the out of control mean vector , n_t is the number of subgroups since the most recent renewal and

$$\left\|\mathbf{C}_t\right\| = \sqrt{\mathbf{C}_t^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{C}_t}$$

where

$$\mathbf{C}_t = \sum_{i=t-n_t+1}^{t} (\mathbf{X}_i - \boldsymbol{\mu}_0)$$

This chart signals when $MC1_t > h$. The value of h is determined using the Markov chain approach.

The MC2 chart is based on the one - sided univariate CUSUM and can be formed as

$$MC2_t = \max\left\{0, MC2_{t-1} + X_t^2 - K\right\} \qquad (7\text{-}8)$$

where $MC2_t = 0$, $K = 0.5 \times \lambda^2 \boldsymbol{\mu}_1 + p$,and X_t^2 is formulated as

$$X_t^2 = (\mathbf{X}_t - \boldsymbol{\mu}_0)^{\mathbf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{X}_t - \boldsymbol{\mu}_0)$$

As before, the process is out of control if $MC2_t > h$.

7.2.2.4 An MCUSUM which is based on sequential probability ratio (SPRT)

CUSUM procedures are a set of sequential procedures based on likelihood ratios for detecting a shift in a process. Healy (1987) based on this fact he developed a multivariate CUSUM in order to detect a shift in a process. Assuming that $\mathbf{x}_n$ may come from a p - variate normal distribution with in control mean vector $\boldsymbol{\mu}_0$ and an out of control mean vector $\boldsymbol{\mu}_1$ and known common covariance matrix $\boldsymbol{\Sigma}_0$. Healy, proved that a CUSUM for detecting a shift in the mean for the p - variate normal distribution is

$$\mathbf{S}_n = \max\left(\mathbf{S}_{n-1} + \mathbf{a}^T\mathbf{x}_n - K, 0\right) \tag{7-9}$$

where

$$\mathbf{a}^T = \frac{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \Sigma^{-1}}{\left[(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)\right]^{-1/2}}$$

and

$$K = .5 \times \frac{(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)^T \Sigma^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)}{\left[(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)\right]^{-1/2}}$$

This CUSUM can be written as

$$\mathbf{S}_n = \max\left(\mathbf{S}_{n-1} + \mathbf{a}^T(\mathbf{x}_n - \boldsymbol{\mu}_0) - .5D, 0\right)$$

where D is the non centrality parameter which is given by

$$D = \left[(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)\right]^{-1/2}$$

This CUSUM signals when $S_n > h$. Furthermore, the quantity $\mathbf{a}^T(\mathrm{x}_n - \mu_0)$ follows a univariate $N(0,1)$ if $\mathbf{x}_n = \boldsymbol{\mu}_0$ and a univariate $N(D,1)$ if $\mathrm{x}_n = \mu_1$. Therefore, the multivariate CUSUM procedure reduces to a univariate normal CUSUM procedure. The main advantage of this CUSUM procedure is that depends on $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}$ only through D; that is, the ARL's do not change as the number of variables increases.

Analyzing the same set of data as in paragraph 7.2.1 a multivariate CUSUM which is based on Healy's approach is presented in figure 7.2. The control limits in this case, just as happened in the univariate case, are expected to lie three times the standard deviation of the variable that we are analyzing, that is, the upper and lower control limits are 16.034 and -16.034 respectively.

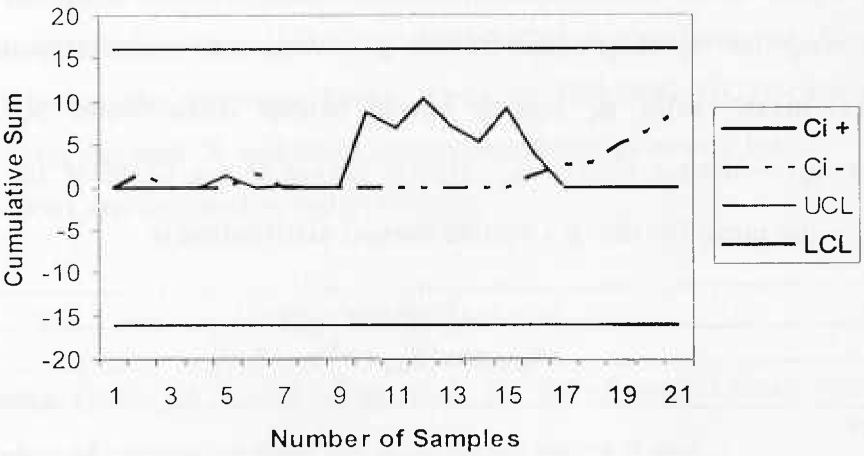


Figure 7.2: An MCUSUM based on SPRT

7.2.2.5 Multivariate EWMA charts

Lowry et al., (1992) developed a multivariate version of EWMA control chart. The MEWMA equations are:

$$\mathbf{Z}_i = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{Z}_{i-1}$$
$$\mathbf{T}_i^2 = \mathbf{Z}_i^{\mathrm{T}} \Sigma_{\mathbf{Z}_i}^{-1} \mathbf{Z}_i \qquad (7\text{-}10)$$

where $\mathbf{x}_i$ is a p - variate observation that follows a p - variate Normal distribution with known mean vector $\boldsymbol{\mu}_0$ and a known covariance matrix $\Sigma_0, 0<\lambda\le 1$, $\mathbf{Z}_0 = \mathbf{0}$

and

$$\Sigma_{Z_i} = \frac{2}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]\Sigma$$

which is analogous to the variance of the univariate EWMA. The decision of selecting appropriate values of λ is not an easy one. Prabhu and Runger (1997) showed that in - control run length is not strongly affected by the value of λ. On the other hand, the optimal value of λ depends on the number of variables and the size of the shift. The same authors deduced that values of λ between 0.1 and 0.5 illustrate very similar shift detection properties. Furthermore, Lowry et al., (1992) proved that the MEWMA is more sensitive in detecting mean shifts from Hotelling's T^2 chart.

Analyzing the same set of data as before a MEWMA control chart which is based on Lowry's et al. approach is illustrated in figure 7.3. The value of λ=0.1 and the correspond upper control limit according to Runger and Prabhu (1997) recommendations is $UCL = 8.64$.

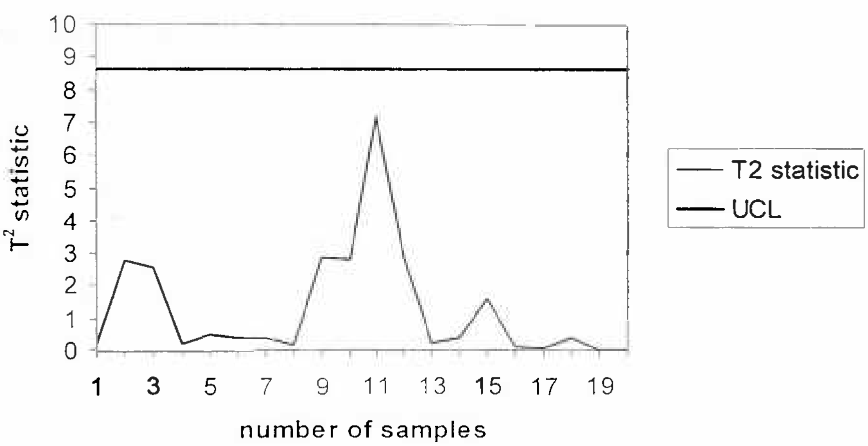


Figure 7.3: An MEWMA control

7.2.3 Monitoring multivariate processes using regression adjustment

Various researchers have developed methods to monitor multivariate processes that do not depend on the Hotelling T^2 statistic. Hawkins (1991) proposed a regression - based control technique, (regression adjustment) which is based on plotting univariate control charts of the residuals from each variable when that variable is regressed on all the others. Furthermore, if the proper set of variables is included in the regression model, the residuals are uncorrelated even though the original variable exhibits correlation. Hawkins showed that the ARL performance of this scheme is very competitive with other methods, but depends on the types of control charts applied to residuals. This approach is very important when the process has a distinct hierarchy of variables, that is, a set of input variables and a set of output variables. Sometimes, this process is called cascade process (Hawkins, 1993b). Cascade process has many possible applications in chemical manufacturing when exist several inputs and outputs or/and there are autocorrelated variables.

7.2.4 Interpretation of Out – of – Control Signals

The objective of performing a multivariate statistical process control is to monitor the process in order to detect any assignable causes. It is crucial to find which of the variables or a subset of them is responsible for the out of control signal. However, it is difficult to analyze the cause of an out of control signal because of the complexity of multivariate control charts and the cross correlation between variables.

Alt (1985) proposed the use of a univariate $\overline{x}$ chart on the individual variables with Bonferroni - type control limits. This approach reduces the number of false alarms associated with simultaneous univariate control charts.

Jackson (1980) recommended the use of control charts of principal components instead of the original variables. The disadvantage of this approach is the difficulty of interpreting principal components with respect to the original variables.

Mason et al., (1995, 1997) proposed to decompose the T^2 statistic into orthogonal components when correlation exists. This approach is called MYT decomposition and can be very beneficial, in interpreting control chart signals, whether the T^2 statistic is chosen as the primary charting statistic or another multivariate charting procedure is used (like MCUSUM or MEWMA).
Furthermore, graphical techniques were developed for interpretation of out of control signals such as the multivariate profile (MP) charts (Fuchs and Benjamini, 1994) and Dynamic Biplots (Sparks et al., 1997).

7.3 The multivariate EPC

Many manufacturing processes have by nature, multiple input and multiple output (MIMO) variables. However, MIMO process feedback control has not been fully investigated in the literature. In the next subparagraphs, are presented a multivariate EWMA and a multivariate double EWMA controller that can be applied in semiconductor industries.

7.3.1 A multivariate EWMA controller

The MIMO process feedback control problem is often encountered in the semiconductor industry. Tseng et al., (2002) proposed a multivariate EWMA (MEWMA) controller, using a linear MIMO model, which could be applied in the above industries under the following two processes:

- A chemical mechanical polishing process.
- A silicon epitaxy process.

A linear $(m \times n)$ MIMO process can be described as:

$$\mathbf{y}_t = \boldsymbol{\alpha} + \boldsymbol{\kappa}(t-1) + \mathbf{B}\mathbf{u}_{t-1} + \boldsymbol{\varepsilon}_t$$

where, $\mathbf{u}_{t-1}$ denotes the vector of m input variables, $\mathbf{y}_t$ denotes the vector of n output variables, $\boldsymbol{\varepsilon}_t$ denotes the process disturbance and $\boldsymbol{\alpha}, \boldsymbol{\kappa}$ and $\mathbf{B}$ are unknown vectors and a matrix to be estimated. For simplicity the authors assumed the case where $\boldsymbol{\kappa} = 0$.

Similar to the single EWMA controller (Ingolfsson and Sachs, 1993) the proposed MEWMA controller is,

$$\hat{\boldsymbol{\alpha}}_t = \omega\left(\mathbf{y}_t - \hat{\mathbf{B}}\mathbf{u}_{t-1}\right) + (1-\omega)\hat{\boldsymbol{\alpha}}_{t-1} \tag{7-11}$$

where, $\hat{\boldsymbol{\alpha}}_t, \hat{\mathbf{B}}$ denote the least squares estimate (LSE) of $\boldsymbol{\alpha}, \mathbf{B}$ respectively, and ω is a discount factor ($0 < \omega < 1$).

The authors advocated that an optimal discount factor for the above controller is given minimizing the mean square error (MSE). The MSE at run t is:

$$E(\mathbf{y}_t - \boldsymbol{\tau})(\mathbf{y}_t - \boldsymbol{\tau})^T$$

where $\boldsymbol{\tau}$ denotes the desired target.

Moreover, they proved that within a finite number of production runs, say N, the optimal discount factor satisfies the following equation:

$$S(\omega)=trace\left(\sum_{i=1}^{N}E(\mathbf{y}_t-\tau)(\mathbf{y}_t-\tau)^T\right) \tag{7-12}$$

7.3.2 A multivariate double EWMA controller

Recent work in the area of semiconductor manufacturing, concentrated in the application of the EWMA statistic for process adjustment purposes (Sachs et. al., 1995; Del Castillo and Hurwitz, 1997). DelCastillo (1999) presented a controller based on two coupled EWMA equations, the double EWMA (DEWMA) feedback controller. The robustness and stability conditions of DEWMA controllers as well as the case of random walk with drift and IMA(1,1) disturbances, were analyzed by DelCastillo (1999, 2001). However, all of these analyses were restricted to single input single output (SISO) processes, although, some of these processes fall into the category of multivariate case. Del Castillo and Rajagopal (2002) extended the DEWMA feedback controller to the multivariate (MIMO) case.

The authors assumed that there are p outputs and m inputs (manipulated variables) and the process is modeled as:

$$\mathbf{y}_t=\boldsymbol{\alpha}+\boldsymbol{\beta}\mathbf{u}_{t-1}+\mathbf{N}_t$$

where,

$$\mathbf{N}_t=\boldsymbol{\delta}\mathbf{t}+\boldsymbol{\varepsilon}_t$$

In the above equations, $\mathbf{N}_t$ is assumed to be a multivariate deterministic trend (DT) noise model and the MIMO DEWMA controller is derived for this type of disturbance. Furthermore, $\boldsymbol{\delta}$ is a $p\times p$ diagonal matrix of the average drift rate per time unit, $\mathbf{t}$ is a $p\times 1$ vector and $\boldsymbol{\varepsilon}_t$ is a multivariate white noise sequence. Furthermore, the $p\times 1$ vector $\mathbf{y}_t$ contains the quality characteristics (outputs), $\boldsymbol{\alpha}$ is a $p\times 1$ vector containing the offset parameter of each of the outputs, $\boldsymbol{\beta}$ is a $p\times m$ process gain matrix and $\mathbf{u}_{t-1}$ is the $m\times 1$ vector of the inputs. Such a model is applicable not only in semiconductor manufacturing but in many other batch processes (i.e. a drift in the process is due to wear out phenomena).

At the end of each run t, a corrective action $\mathbf{u}_t$ is chosen, which gives a prescribed set of inputs (recipe) to the process engineer for use in the next run.

This action minimizes the one step a head predicted deviation from target, that is,

$$\min Z = \|\mathbf{y}_{t+1} - \mathbf{T}\|^2$$

where,

$$\hat{\mathbf{y}}_{t+1} = \mathbf{A}_t + \mathbf{D}_t + \mathbf{B}\mathbf{u}_t$$

where $\mathbf{T}$ is the desired target $p \times 1$ vector and $(\mathbf{A}_t + \mathbf{D}_t)$ estimates the $(\boldsymbol{\alpha} + \boldsymbol{\delta}(\mathbf{t}+1))$. Furthermore, $\mathbf{A}_t$, $\mathbf{D}_t$ can be expressed as functions of EWMA weight matrices $(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2)$, similarly to the univariate case (λ_1, λ_2). It can be proved that the corrective action is described as,

$$\mathbf{u}_t = \left(\mathbf{B}^T\mathbf{B}\right)^{-1} \mathbf{B}^T (\mathbf{T} - \mathbf{A}_t - \mathbf{D}_t) \tag{7-13}$$

In the particular case where $m = p$ the last equation reduces to:

$$\mathbf{u}_t = \mathbf{B}^{-1}(\mathbf{T} - \mathbf{A}_t - \mathbf{D}_t)$$

Usually, the industrial practice with MIMO processes consists in applying several SISO feedback controllers acting in parallel. Therefore, the authors compared their proposed controller with a "parallel SISO" policy using a simulation study.

Their conclusions can be summarized as follows:

- For both the parallel SISO and MIMO DEWMA schemes, there is an optimal weight combination (λ_1, λ_2). This combination of weights performed well, irrespectively of the correlation between the outputs.
- For both the parallel SISO and MIMO DEWMA schemes, high values of weights result in large values of the MSE. For the parallel SISO case, the average MSE can diverge when there is a strong correlation between the outputs. On the contrary, the MIMO DEWMA controller provided a stable performance.
- In the parallel SISO case, when very small values of the EWMA weights are chosen, the behavior is very strange while such behavior was not observed for the MIMO case.

7.4 Integrating the EPC processes with joint monitoring SPC schemes

In the previous chapter were presented and discussed various techniques of combined EPC and SPC schemes in order to improve the product quality. The usual practice is the monitoring of the controlled output using various SPC charts. However, monitoring the controlled outputs alone is usually not effective because feedback control action causes the input of the process to adapt to process changes. Alternatively, someone could apply SPC techniques to the control action of the controlled process (Faltin and Tucker, 1991; Messina et al., 1996). Since the control action compensates for process changes, monitoring the control action it is possible to be more effective to detect these changes (MacGregor, 1991). Nevertheless, joint monitoring of both the control action and the process output using bivariate charts may outperform than conventional SPC approaches.

7.4.1 Joint monitoring of a PID controlled process

Generally, data from a PID controlled process are autocorrelated. Furthermore, various researchers studied the topic of autocorrelated SPC (Alwan and Roberts, 1988; Montgomery and Mastrangelo, 1991; Wardell et al., 1994). It is worth noting that monitoring the controlled output focuses on the "transient" period, while monitoring the manipulated input focuses on the "static" period. Once a mean shift in the input is added to the system, the output will experience a time period with larger dynamic output ("transient" period). After the transient period, the system's output will remain in a small static range ("static" period). However, it is known from the control theory, that when the integral (I) control mode takes action the mean shift of the process output will be eliminated immediately after the transient time period. Thus, there is only a limited "window of opportunity" (Van der Wiel, 1996) during which the process change must be detected. All conventional SPC techniques suffer from this problem. Therefore, a possible solution is a joint monitoring SPC strategy that involves both the input and the output of a process using bivariate SPC schemes.

Tsung et al., (1999), proposed a bivariate SPC scheme using the Bonferroni's approach. This approach was recommended by Alt (1985) because of its simplicity in determining the variables which are responsible for an out of control condition. The most important advantage of this approach is that there is no need to assume that the covariance matrix among the process variables is known. Moreover, the EPC scheme is a PID controlled process where the process disturbance follows an ARMA (1,1) model. As a criterion for optimizing this process the authors used the mean squared error (MSE) of the process outputs. Tsung (1998) derived the relationship between the PID control parameters and the process MSE for an ARMA (1,1) model which is described as:

$$MSE\left(k_P, k_D, k_I, \phi, \theta\right) \equiv \left(I + 2\rho_1 \left(II + III\right)\sigma_D^2\right) \tag{7-14}$$

where, k_P, k_D, k_I are the parameters of the PID controlled process, ϕ, θ are the parameters of the ARMA(1,1) model, I, II, III are functions of the PID and the disturbance parameters and ρ_1, σ_D^2 are the first order autocorrelation and the variance of the disturbance $\{D_t\}$ of an ARMA(1,1) process respectively.

Furthermore, the standard deviation of the process output after the PID control is,

$$\sigma_e = \sqrt{I + 2\rho_1 \left(II + III\right)\sigma_D} \tag{7-15}$$

and the standard deviation of the manipulated input is,

$$\sigma_X = \sqrt{I' + 2\rho_1 \left(II' + III'\right)\sigma_D} \tag{7-16}$$

where, I', II', III' are also functions of the PID and the disturbance parameters.

Based on (7-15), (7-16) and assuming that the process data are normally distributed the control limits of the joint charts are given by the following equations:

$$\begin{aligned} CL_e &= \pm z_{(1-a/4)}\sigma_e \\ CL_X &= \pm z_{(1-a/4)}\sigma_X \end{aligned} \tag{7-17}$$

The joint decision rule suggests that the controlled process is out of control when either the manipulated input or the controlled output is outside the control limits. As a conclusion, the authors proposed that their approach can be used using other disturbance models instead of ARMA (1,1) model.

Another joint monitoring scheme, which is based on Hotelling's T^2 statistic, was proposed by Tsung et al., (1999). For this scheme the EPC component is a PID controlled process, as before, where the process disturbance follows an ARMA (1,1) model. Now, the SPC scheme is the Hotelling's multivariate control chart that gives an out of control signal when

$$\chi_t^2 = \mathbf{X}_t^{\mathrm{T}}\Sigma^{-1}\mathbf{X}_t > CL_T$$

where, the monitored characteristics are the vector of the controlled output and the manipulated input $\left(\mathbf{X}_t = (y_t, u_t)^T\right)$, and Σ is the covariance matrix of $\mathbf{X}_t$. Thus, the control limit for Hotelling's approach is:

$$CL_T = \chi^2_{\alpha,2} \tag{7-18}$$

where α is the desired false alarm rate. The joint decision rule suggests that the controlled process is out of control when χ_t^2 exceeds the control limit.

Although Hotelling's approach requires the knowledge of the covariance matrix, it is superior to Bonferoni's approach in the sense that it gives exact type I errors, and not conservative ones. On the other hand, if the estimated covariance matrix from the model is "bad" the Hotelling's approach may perform worse than the Bonferroni's approach. In general, for small correlations Bonferroni's approach performs better than Hotelling's approach, while for large correlations the opposite holds (Tsung et al., 1999).

7.4.2 The dynamic T^2 chart for monitoring feedback controlled processes

Tsung et al., (2002) they proposed to improve the detect - ability of joint monitoring schemes using dynamic T^2 statistics that take into consideration the effects of dynamics and autocorrelation due to feedback control. Thus, the EPC scheme is based on a PID controller and the SPC scheme on a similar to the usual T^2 control chart but the data vector is composed of time shifted observations of both the input and the output. Therefore, the dynamic T^2 statistic is described as,

$$DT_t^L = \mathbf{X}_t^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{X}_t$$

where, $\mathbf{X}_t$ is the vector of the controlled output and input ($\mathbf{X}_t = (y_t, u_t, y_{t-1}, u_{t-1}, \ldots, y_{t-L}, u_{t-L})^T$) respectively, $\boldsymbol{\Sigma}$ is the covariance matrix and L is a user specified time shift factor.

It is crucial to select an appropriate time lag parameter L for the dynamic monitoring scheme. A smaller value of L, than necessary, may lead to ineffective monitoring, while a larger one may lead to redundant computing. The authors recommended selecting a L value by fitting an $AR(\infty)$ model to the output (y_t). Thus, the "best" value of L would be such that $\left|(\theta - \phi)\theta^{j-1}\right| < \xi$ for $j > L$. Using extensive simulations the authors showed that a satisfactory value of $\xi = 0.1$.

The authors compared their proposed scheme (L=1,2) with the "static" (L=0) T^2 scheme. They found that the overall performance of the dynamic T^2 monitoring scheme is better than the "static" joint monitoring scheme. However, for a process with large correlation the two schemes have similar performance. More specifically, DT^1 is consistently better than the other schemes for large mean shifts, while DT^2 is consistently better than the other schemes for small mean shifts.

Chapter 8

Conclusions and Further Research

8.1 Introduction

This chapter presents some comments on this dissertation as well as possible extensions for further research. In the next section some conclusions and further research topics of SPC and EPC approaches are presented in the univariate and multivariate case respectively.

8.2 Conclusions and Further Research

Usually, two different technical groups have been concerned for developing process controls in modern industries. Quality engineers and applied statisticians focus on monitoring quality characteristics through statistical control charts, while control (process) engineers mostly concentrate on on-line process adjustments using engineering process control (EPC) techniques. Several research ideas about SPC arose from the Journal of Quality Technology (JQT) panel discussion edited by Montgomery and Woodall (1997). However, this dissertation focus on engineering process control methods as well as the combination of SPC and EPC for the same process in an integrated form and with this rationale some conclusions and research ideas are discussed.

(i) Some of the conclusions are:

- When a process exhibits autocorrelation various remedial actions can be taken in order to monitor it. The effect of autocorrelation on a control chart scheme depends primarily on (1) The actual model that the process really follows, (2) the model that we consider to describe the in-control operation, (3) the way we estimate the in-control model and (4) the type of disturbance that the chart is supposed to detect. The first three considerations are very important because the estimated model will affect the chart scheme.

• Quality engineers very often need to adjust a process. Feedback adjustments using PI or PID control action are no more difficult than the usual control charts. Moreover, when frequent adjustments are not feasible or making an adjustment is expensive, "deadband policies" can be shown to be optimal if there is a fixed adjustment cost.

• "Optimal" controllers minimize the variability of the quality characteristic ignoring any cost associated with this action. Furthermore, proportional integral derivative (PID) controllers are by far the common adjustment policy in industrial processes. Hence, minimum mean squared error (MMSE) controllers are rather impractical in industrial processes. On the other hand, this type of controller is still useful as benchmark in evaluating the performance of other adjustment schemes.

• The common goal of SPC and EPC approaches is the reduction of variability using different ways. A significant amount of work has appeared in the 1990s about methods that combine EPC and SPC schemes for the same process in an integrated form. Simulation as well as case studies showed that:

o SPC/EPC schemes result in significant improvement in adjustment variance at the expense of slight increase in output variance,

o SPC/EPC schemes reduce the frequency and magnitude of adjustment when compared to EPC schemes.

(ii) Some topics for further research are:

• The most SPC methods have been developed and customized over the years to maximize the information obtained from relatively limited data. Now, in the most industrial process there is often a large amount of data. Thus, SPC methods are needed for massive datasets (Runger et al., 1997)

• Most of the research on SPC methods has been focused on monitoring the process mean. There has been a trend on monitoring process variability, but more work is needed on this topic.

• There are many difficulties associated with interpreting signals from multivariate control charts. Therefore, more work is needed on graphical methods for data visualization. Furthermore, data reduction methods are also important.

- There are industrial processes were multiple input and/or multiple output (MIMO) are more appropriate to define the quality of the product. In this case many authors are developed multivariate schemes which are generalizations of the corresponding schemes in the univariate case. However, the complexity of these schemes discourages the quality practitioners to apply them in real life situations. Hence, simple tools of sophisticated methods are topics of major priority.

REFERENCES

Alt, F.B. (1985). *Multivariate Quality Control*, in Encyclopedia of Statistical Sciences, Kotz, Johnson and Read, 110 – 122, Wiley, New – York

Alwan, L.C. (1992). Effects of Autocorrelation on Control Charts, *Communications in Statistics – Theory and Methods*, 21 (4), 1025 - 1049

Alwan, L.C. and Radson, D.(1992). Time- Series Investigation of Subsample mean Charts, *IIE Transactions*, 24, 66 – 80

Alwan, L.C. and Roberts, H.V. (1998). Time Series Modelling for Statistical Process Control, *Journal of Business and Economic Statistics*, 6, 87 – 95

Aström, K.J. and Häglund , T. (1995). PID controllers: Theory, Design and Tuning, 2nd ed. Research, Triangle Park, N. C

Aström, K.J. and Wittenmark, B. (1997). *Computer Controlled Systems: Theory and Design*, 3rd ed. Prentice Hall, Upper Saddle River, New Jersey

Barnard, G.A. (1959). Control Charts and Stochastic Processes, *Journal of the Royal Statistical Society*, 21(2), 239 – 271

Box, **G.E.P. and Cox, D.R., (1964).** An analysis of transformations (with discussion), *Journal of the Royal Statistical Society*, B, 26, 211 - 252

Box, G.E.P. and Jenkins, G.M.,(1976). *Time Series Analysis: Forecasting and Control*, revised ed. Holden – Day, San Francisco

Box, **G.E.P., Jenkins, G.M. and Reinsel G.C. (1994).** *Time Series Analysis, Forecasting and Control*, 3rd ed. Prentice Hall, Englewood Cliffs, New Jersey

Box, G.E.P. and Kramer, T. (1992). Statistical Process Monitoring and Feedback Adjustment: A Discussion, *Technometrics*, 34(3), 251 – 267

Box, G.E.P. and Luceno, A. (1995). Discrete Proportional – Integral Control with Constrained Adjustment, *The Statistician*, 44(4), 479 – 495

Box, G.E.P. and Luceno, A. (1997). *Statistical Control by Monitoring and Feedback Adjustment.* Wiley, New York

Box, G.E.P. and Luceno, A. (2002). Feedforward as a Supplement to Feedback Adjustment in Allowing for Feedstock Changes, *Journal of Applied Statistics*, 29(8), 1241 – 1254

Brockwell, P.J. and Davis, R.A., (1996). *Introduction to Time Series and Forecasting.* Springer - Verlag, New York

Butler, and Stefani, J.A. (1994). Supervisory Run to Run Control of a Polysilicon Gate Etch Using in situ Ellipsometry, *IEEE Transactions on Semiconductor Manufacturing*, 7(2), 193 – 201

Cameron, F. and Seborg, D.E. (1983). A Self Tuning Controller with PID structure, *International Journal of Control*, 38(2), 401 – 417

Chen, A. and Guo, R.-S. (2001). Age – Based Double EWMA Controller and Its Application to CMP Processes, *IEEE Transactions on Semiconductor Manufacturing*, 14(1), 11 – 19

Chen, A. and Elsayed, E.A. (2002). Design and performance Analysis of the Exponentially Weighted Moving Average Mean Estimate for Processes Subject to Random Step Changes, *Technometrics*, 44(4), 379 – 389

Clarke, D.W. and Gawthrop, P.J. (1975) Self – Tuning Controller, *Proceedings of the Institution of Electrical Engineers*,122(9), 929 – 934

Crowder, S.V. (1992). An SPC model for short productions runs: minimizing expected cost, *Technometrics*, 34, 64 – 73

DelCastillo, E. (1998). A note on two Process Adjustment Models, *Quality and Reliability Engineering International*, 14, 23 – 28

DelCastillo, E. (1999). Long Run and Transient Analysis of a Double EWMA Feedback Controller, *IIE Transactions*, 31(12), 1157 – 1169

DelCastillo, E. (2000). A Variance Constrained PI Controller that Tunes Itself, *IIE Transactions*, 32(6), 479 – 491

DelCastillo, E. (2001) Identification and Fine Tuning of Closed – Loop Processes under Discrete EWMA and Pi Adjustments, *Quality and Reliability Engineering International*, 17, 419 – 427

DelCastillo, E. (2001a) Some Properties of EWMA feedback Quality Adjustment Schemes for Drifting Disturbances, *Journal of Quality Technology*, 33(2), 153 – 166

DelCastillo, E. (2001b) Closed - Loop Disturbance Identification and Controller Tuning for Discrete manufacturing Processes, *Technometrics*, 44(2), 134 – 141

DelCastillo, E. (2002) *Statistical Process Adjustment for Quality Control.* Wiley, New York

DelCastillo, E. and Hurwitz, A. (1997). Run to Run Process Control: A Review and Some Extensions, *Journal of Quality Technology*, 29(2), 184 – 196

DelCastillo, E. and Rajagopal, R. (2002). A Multivariate Double EWMA Process Adjustment Scheme for Drifting Processes, *IIE Transactions*, 34, 1055 – 1068

Deming, W.E. (1986). *Out of the Crisis.* Mass: Center for Advanced Engineering Study, Cambridge

Dodge, H.F. and Romig, H.G. (1959). *Sampling Inspection Tables, Single and Double Sampling.* Wiley, New York

Faltin, F.W. and Tucker, W.T. (1991). *Online Quality Control for the Factory of the 1990s and Beyond,* in *Statistical Process Control in Manufacturing,* Keats, J.B. and Montgomery, D.C. (eds.,). Marcel – Dekker, 331 – 354, New York

Feld'baum, A. A. (1965). *Optimal Control Theory.* Academic Press, San Diego

Fuchs, C. and Benjamini, Y. (1994). Multivariate Profile Charts for Statistical Process Control, *Technometrics*, 36(2), 182 – 195

Grosier, R.B. (1988). Multivariate Generalizations of Cumulative Sum Quality Control Schemes, *Technometrics*, 30(3), 291 – 303

Grubbs, F.E. (1954). An Optimum Procedure for Setting Machines or Adjusting Processes, *Industrial Quality Control*, July; reprinted in *Journal of Quality Technology* 1983,15(4), 186-189

Hawkins, D.M. (1991). Multivariate Quality Control based on Regression – Adjusted variables, *Technometrics*, 33(1), 61 – 75

Hawkins, D.M. (1993b). Regression Adjustment for Variables in Multivariate Quality Control, *Journal of Quality Technology*, 25(3), 170 – 182

Healy, J.D. (1987). A Note on Multivariate CUSUM Procedures, *Technometrics*, 29(4), 409 – 412

Hunter, J.S. (1994). *Beyond the Shewart SPC Paradigm.* Presentation given at the Department of Statistics, Southern Methodist University, Dallas, Texas

Hotelling, H. (1947). *Multivariate Quality Control, Techniques of Statistical Analysis*, edited by Eisenhart, Hastay and Wallis. MacGraw – Hill, New York

Ingolfsson, A. and Sachs, E. (1993). Stability and Sensitivity of an EWMA controller, *Journal of Quality Technology*, 25, 271 – 287

Jackson, J.E. (1980). Principal Components and Factor Analysis: Part I-Principal Components, *Journal of Quality Technology*, 12

Janakiram, M. and Keats, J.B.(1998). Combining SPC and EPC in a Hybrid Industry, *Journal of Quality Technology*, 30, 189 – 200

Jensen, K.L. and Vanderman, S.B. (1993). Optimal adjustment in the presence of deterministic process drift and random adjustment error, *Technometrics*, 35, 376 – 389

Jiang, W., Tsui, K.L. and Woodall, W. H.(2000). A new SPC Monitoring Method: The ARMA Chart, *Technometrics*,42, 399 – 410

Jiang, W. and Tsui, K.L. (2002). SPC Monitoring of MMSE and PI – Controlled Processes, *Journal of Quality Technology*, 34(4), 384 – 398

Katende, E. and Jutan, A. (1993). A new Constrained Self – Tuning PID Controller, *Canadian Journal of Chemical Engineering*, 82, 35 – 45

Kramer, T. (1989). Process Control from an Economic Point of View, unpublished Ph.D. dissertation, University of Wisconsin – Madison, Dept., of Statistics

Liu, R.Y. (1995). Control Charts for Multivariate Processes, *Journal of the American Statistical Association*, 90(432), 1380 -1387

Lowry, C.A., Woodall, W.H, Champ, C.W. and Rigdon, S.E. (1992). A Multivariate EWMA Control Chart, *Technometrics*, 34(1), 46 – 53

Lowry, C.A. and Montgomery, D.C. (1995). A review of Multivariate Control Charts: A Unified Approach, *IIE Transactions*, 26 800 – 810

Lu, C.W., and Reynolds, M.R., (2001). Cusum Charts for Monitoring an Autocorrelated Process, *Journal of Quality Technology*, 33, 316 – 334

MacGregor, J.F. (1990). A different view of the funnel experiment, *Journal of Quality Technology*, 22, 255 – 259

MacGregor, J.F. (1988). On- line Statistical Process Control, *Chemical Engineering Process*, October, 21 – 31

Mason, R.L., Tracy, N.D. and Young, J.C. (1995). Decomposition of T^2 for Multivariate Control Chart Interpretation, *Journal of Quality Technology*, 27(2), 99 – 108

Mason, R.L., Tracy, N.D. and Young, J.C. (1997). A Practical Approach for Interpreting Multivariate T^2 Control Chart Signals, *Journal of Quality Technology*, 31(2), 155 – 165

Mastrangelo, C.M., Runger, G.C. and Montgomery, D.C. (1996) Statistical Process Monitoring with Principal Components, *Quality and Reliability Engineering International*, 12, 203 – 210

Messina, W.S. (1992). Strategies for the Integration of Statistical and Engineering Process Control". Ph.D. Dissertation, Arizona State University, Tempe, AZ

Messina, W.S., Montgomery, D.C., Keats, J.B. and Runger, G.C. (1996). *Strategies for Statistical Monitoring of Integral Control for the Continuous Process Industries, in* Statistical Applications in Process Control, Keats, J.B. and Montgomery, D.C. (eds.,). Marcel – Dekker, 193 – 215, New York

Mitra, A. (1998). *Fundamentals of Quality Control and Improvement.* Prentice Hall, New Jersey

Montgomery, D.C. (2001). *Introduction to Statistical Quality Control* 4th ed. Wiley, New York

Montgomery, D.C., Keats, J.B., Runger, G.C. and Messina, W.S. (1994). Integrating Statistical Process Control and Engineering Process Control, *Journal of Quality Technology*, 26, 79 – 87

Montgomery, D.C., Keats, J.B., Yatskievitch, M. and Messina, W.S. (2000). Integrating Statistical Process Monitoring with Feedforward Control, *Quality and Reliability Engineering International*, 16, 515 – 525

Montgomery, D.C. and Mastrangelo, C.M. (1991). Some Statistical Process Control Methods for Autocorrelated Data, *Journal of Quality Technology*, 23, 179 - 193

Moyne, J., DelCastillo, E. and Hurwitz, A., (2000). *Run to Run Control in Semiconductor Manufacturing*. CRC Press, Boca Raton

Nembhard, H.B. (1998). Simulation Using the State - Space Representation of Noisy Dynamic Systems to Determine Effective Integrated Process Control Designs, *IIE Transactions*, 30, 247 – 256

Nembhard, H.B. and Mastrangelo, C.M. (1998). Integrated Process Control for Startup Operations, *Journal of Quality Technology*, 30(3), 201 – 211

Nembhard, H.B. Mastrangelo, C.M. and Kao, M.S. (2001). Statistical Monitoring Performance for Startup Operations in a Feedback Control System, *Quality and Reliability Engineering International*, 17, 379 – 390

O'Shaugenessy, P. and Haugh, L.D. (2002). An EWMA – based Bounded Adjustment Scheme with Adaptive Noise Variance Estimation, *Journal of Quality Technology*, 34(3), 327 – 338

Pigniatello, J.J. and Runger, G.C. (1990). Comparison of Multivariate CUSUM Charts, *Journal of Quality Technology*, 22(1), 173 – 186

Prabhu, S.S. and Runger, G.C. (1997). Designing a Multivariate EWMA chart, *Journal of Quality Technology*, 29(1), 8 – 15

Ramirez, G. (1998). Monitoring Clean Room Air Using Cuscore Charts, *Quality and Reliability*, 14, 281 – 289

Roberts, S.W. (1959). Control Charts Tests Based on Geometric Moving Averages, *Technometrics*, 1,239 - 250

Runger, G.C. and Willemain, T.R. (1995). Model – Based and Model – Free Control of Autocorrelated Processes, *Journal of Quality Technology*, 27, 283 – 292

Sachs, E., Hu, A. and Ingolfsson, A. (1995). Run by Run Process Control: Combining SPC and Feedback Control, *IEEE Transactions on Semiconductor Manufacturing*, 8(1), 26 – 43

Shao, Y.E. (1993). A Real Time Stochastic Process Control System for Process Manufacturing, Ph.D., Dissertation, Rensselaer Polytechnic Institute

Shao, Y.E. (1998). Integrated Application of the Cumulative Score Control Chart and Engineering Process Control, *Statistica Sinica*, 8, 239 – 252

Shao, Y.E., Runger, G.C., Haddock, J. and Wallace, W.A. (1999). Adaptive Controllers to Integrate SPC and EPC, *Communications in Statistics – Simulation*, 28(1), 13 – 36

Shewart, W.A. (1931). *Economic Control of Quality of Manufacturing Product.* Van Nostrand Reinhold, New Jersey

Smith, T.H. and Boning, D.S. (1997). Artificial Neural Network Exponentially Weighted Moving Average Controller for Semiconductor Processes, *Journal of vacuum Science and Technology* 15(3)

Sparks, R.S., Adolphson, A. and Phatak, A. (1997). Multivariate Process Monitoring Using Dynamic Biplot, *International Statistical Review*, 65(3), 325 – 349

Tseng, S.T., Chou, R.J and Lee, S.P (2002). A study on a Multivariate EWMA Controller, *IIE Transactions*, 34, 541 - 549

Tsung, F. and Apley, D. W. (2002). The Dynamic T^2 Chart for Monitoring Feedback – Controlled Processes, *IIE Transactions*, 34, 1043 – 1053

Tsung, F. and Shi, J.J. (1999). Integrated Design of Run to Run PID Controllers and SPC Monitoring for Process Disturbance Rejection, *IIE Transactions*, 31(6), 517 – 527

Tsung, F., Shi, J.J. and Wu, C.F.J. (1999). Joint Monitoring of PID controlled Processes, *Journal of Quality Technology*, 31(3), 275 – 285

Tsung, F., Wu, H. and Nair, V. (1998). On the Efficiency and Robustness of Discrete Proportional – Integral Schemes, *Technometrics*, 40(3), 214 – 222

Vander Wiel, S.A. (1996) Monitoring Processes that Wander Using Integrated Moving Average Models, *Technometrics*, 38, 139 – 151

Vander Wiel, S.A., Tucker, W.T., Faltin, F.W. and Doganaksoy, N. (1992). Algorithmic Statistical Process Control: An Elaboration, *Technometrics*, 35, 363 – 374

Wardell, D.G., Moskowitz, H. and Plante, R.D. (1994). Run – Length Distributions of Special – Cause Control Charts for Correlated Processes, *Technometrics*, 36, 3 – 17

West, D., Dellana, S. and Jarrett, J. (2002). Transfer Function Modeling of Processes with Dynamic Inputs, *Journal of Quality Technology*, 34, 315 – 326

Woodall, W.H. and Ncube, M.M. (1985). Multivariate CUSUM Quality Control Procedures, *Technometrics*, 27(3), 285 – 292