

**ATHENS UNIVERSITY  
OF ECONOMICS AND BUSINESS**  
DEPARTMENT OF STATISTICS  
POSTGRADUATE PROGRAM

**STATISTICAL QUALITY CONTROL  
TECHNIQUES FOR AUTOCORRELATED  
PROCESSES**

By

Maria N. Lyra

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**

**STATISTICAL QUALITY CONTROL  
TECHNIQUES FOR AUTOCORRELATED  
PROCESSES**

By

Maria N. Lyra

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  
January 2003





**ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΑΘΗΝΩΝ**

**ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ**

**ΤΕΧΝΙΚΕΣ ΤΟΥ ΣΤΑΤΙΣΤΙΚΟΥ ΕΛΕΓΧΟΥ  
ΠΟΙΟΤΗΤΑΣ ΓΙΑ ΑΥΤΟΣΥΣΧΕΤΙΣΜΕΝΕΣ  
ΔΙΕΡΓΑΣΙΕΣ**

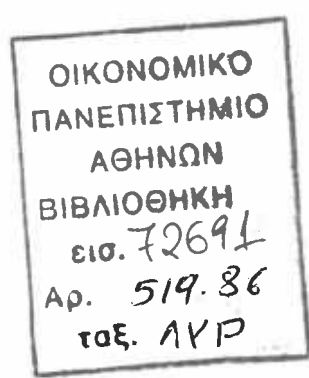
Μαρία Νικολάου Λύρα

**ΔΙΑΤΡΙΒΗ**

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

Αθήνα  
Ιανουάριος 2003





**ATHENS UNIVERSITY  
OF ECONOMICS AND BUSINESS**  
**DEPARTMENT OF STATISTICS**

A Thesis submitted in partial fulfilment of  
the requirements for the degree of  
Master of Science

**STATISTICAL QUALITY CONTROL TECHNIQUES FOR  
AUTOCORRELATED PROCESSES**

**Maria N. Lyra**

**Approved by the Graduate Committee**

S. Psarakis  
Lecturer  
Thesis Supervisor

E. Moustaki  
Assistant Professor  
Member of the Committee

V. Vasdekis  
Lecturer

**Athens, January 2003**

**Michael Zazanis, Associate Professor  
Director of the Graduate Program**



## **ACKNOWLEDGEMENTS**

I thank my mother and father who stood up to me during all this effort. I would also like to thank my professor S. Psarakis with whom I had an excellent collaboration, as well as every single person that provided me with psychological or practical support.





## VITA

I was born in Germany in 1979. I completed my second-degree education at the high school of Pendeli. I studied the faculty of statistics in the Athens University of Economics and Business and I graduated in 2001. A few months later I continued my studies in statistics by following the postgraduate program of the same university.







## **ABSTRACT**

Lyra Maria

### **STATISTICAL QUALITY CONTROL TECHNIQUES FOR AUTOCORRELATED PROCESSES**

January 2003

The Statistical Quality Control (SQC) is a group of techniques which, combined with management, help to achieve continuous improvement in the production process.

A standard tool of SQC is the Statistical Process Control (SPC) applied to processes that generate independent and identically distributed random variables. However, high volume production processes yield process data which are autocorrelated.

To accommodate autocorrelated data, many SPC methodologies have been developed. Independently to the SPC technique, the Engineering Process Control (EPC) strategy has been used in the parts industry. The EPC also aims in quality improvement but via a different path than the one of SQC.

The aim of this dissertation is to analyze both the SPC and EPC strategies for autocorrelated data and to assess the optimization and improvement in the production process when the two techniques are being combined.





## ΠΕΡΙΛΗΨΗ

Μαρία Λύρα

### ΤΕΧΝΙΚΕΣ ΤΟΥ ΣΤΑΤΙΣΤΙΚΟΥ ΕΛΕΓΧΟΥ ΠΟΙΟΤΗΤΑΣ ΓΙΑ ΑΥΤΟΣΥΣΧΕΤΙΣΜΕΝΕΣ ΔΙΕΡΓΑΣΙΕΣ

Ιανουάριος 2003

Ο Στατιστικός Έλεγχος Ποιότητας είναι ένα σύνολο τεχνικών οι οποίες, σε συνδυασμό με τη διαχείριση, συμβάλουν στην επίτευξη διαρκούς βελτίωσης της παραγωγικής διαδικασίας.

Ένα γνωστό βοήθημα του Στατιστικού Ελέγχου Ποιότητας είναι ο Στατιστικός Έλεγχος Διεργασιών, ο οποίος εφαρμόζεται σε διεργασίες που παράγουν ανεξάρτητες και ισόνομα κατανεμημένες τυχαίες μεταβλητές. Εντούτοις, ο μεγάλος παραγωγικός όγκος οδηγεί σε διεργασίες που είναι αυτοσυσχετισμένες.

Προκειμένου να ληφθεί υπόψη η αυτοσυσχέτιση των δεδομένων, έχουν αναπτυχθεί πολλές μεθοδολογίες του Στατιστικού Ελέγχου Διεργασιών. Ανεξάρτητα με αυτή την τεχνική, η στρατηγική του Μηχανικού Ελέγχου Διεργασιών χρησιμοποιήθηκε στη βιομηχανία. Αυτή η μέθοδος επίσης αποβλέπει στη βελτίωση της ποιότητας του προϊόντος αλλά με διαφορετικό τρόπο από το Στατιστικό Έλεγχο Διεργασιών.

Σκοπός της συγκεκριμένης διατριβής είναι να αναλύσει και τις δύο μεθοδολογίες, εφαρμοσμένες σε αυτοσυσχετισμένα δεδομένα, και να διαπιστώσει τη βελτιστοποίηση και τη βελτίωση της παραγωγικής διαδικασίας όταν οι δύο τεχνικές χρησιμοποιούνται σε συνδυασμό.





# TABLE OF CONTENTS

	Page
<b>Chapter 1. Introduction</b>	1
<b>Chapter 2. Overview of the most common control charts</b>	5
2-1. Introduction	5
2-2. Properties of Statistical Process Control (SPC)	5
2-2.1. Basic tools of SPC	5
2-2.2. Out-of-control patterns detected by control charts	7
2-3. Shewhart-type control charts	8
2-3.1. General properties	8
2-3.2. The $(\bar{x}, R)$ control chart	10
2-3.3. The $(\bar{x}, S)$ control chart	12
2-3.4. Selection of the sampling scheme	15
2-4. Control charts for attributes	16
2-4.1. Control charts for fraction/number nonconforming	16
2-4.2. Control charts for nonconformities (defects)	18
2-5. The third-generation control charts: CUSUM and EWMA	19
2-5.1. The CUSUM control chart	19
2-5.2. The EWMA control chart	21
2-6. The Spectral chart	24
<b>Chapter 3. Time series models for autocorrelated data</b>	29
3-1. Introduction	29
3-2. Basic properties of autocorrelated data	29
3-2.1. Autocorrelated data in industry	30
3-2.2. The autocovariance and autocorrelation functions	31
3-3. The ARMA process	32
3-3.1. Definition and properties of the ARMA process	32
3-3.2. Modelling the ARMA processes	33



3-3.3. Goodness-of-fit of the ARMA process	35
3-3.4. The most common ARMA models	37
3-3.4.1. The MA(q) process	37
3-3.4.2. The AR(p) process	38
3-3.4.3. The ARMA(1,1) process	39
3-3.5. Forecasting stationary time series	41
3-4. The ARIMA process	41
3-4.1. Definition of ARIMA models	41
3-4.2. Transformation of the ARIMA models	43
<b>Chapter 4. Control charts for autocorrelated processes</b>	<b>45</b>
4-1. Introduction	45
4-2. Traditional charts modified for autocorrelated processes	45
4-2.1. The modified Shewhart Control Chart	45
4-2.2. The EWMAST chart	47
4-3. Traditional control charts applied to the residuals	50
4-3.1. The Common Cause and Special Cause Charts	50
4-3.1.1. The Common Cause Chart (CCC)	50
4-3.1.2. The Special Cause Chart (SCC)	52
4-3.2. The Weighted and Unweighted Batch Means charts	54
4-3.2.1. The Weighted Batch Means (WBM) chart	54
4-3.2.2. The Unweighted Batch Means (UBM) chart	55
4-4. Control charts applied to the forecast errors	57
4-4.1. The M-M chart	57
4-4.2. The Moving Center-line EWMA control chart	59
4-5. A new chart for correlated data: The ARMAST chart	59
<b>Chapter 5. Performance of charts for autocorrelated data</b>	<b>65</b>
5-1. Introduction	65
5-2. The design of simulation studies	65
5-2.1. The most common performance criteria	66
5-2.2. Guidelines for the simulation procedures	69
5-3. Performance of traditional charts based on simulation studies	71



5-3.1. Performance of the modified Shewhart chart	71
5-3.2. Performance of the EWMAST chart	72
5-4. Performance of the residuals charts	73
5-4.1. Performance of the SCC chart compared to traditional charts	74
5-4.2. Relative performance of the residuals chart	77
5-5. Performance of control charts applied to forecast errors	78
5-6. Performance of the ARMAST chart	81
<b>Chapter 6. The Engineering Process Control (EPC)</b>	<b>83</b>
6-1. Introduction	83
6-2. Differences between the SPC and EPC techniques	83
6-3. Design of EPC	85
6-4. The MMSE and PID controlled processes	91
6-4.1. The I controller	91
6-4.2. The PI controller	93
6-4.3. The PID controller	94
6-4.4. MMSE controllers for disturbance models other than the IMA(1,1)	95
6-5. Applying the feedback control scheme: an example	96
6-6. Minimum cost adjustment: some simple schemes	99
6-6.1. Sampling interval fixed	100
6-6.2. Sampling interval not fixed	101
6-7. Other factors inciting the use of ASPC	103
<b>Chapter 7. Automatic Statistical Process Control (ASPC): more special issues</b>	<b>105</b>
7-1. Introduction	105
7-2. The Run-by-Run controller	105
7-2.1. The Gradual mode in the RbR controller	107
7-2.2. The Rapid mode in the RbR controller	109
7-3. Design maps for the PID controller	112
7-4. MMSE control when measurements are delayed	113



7-5. An economic model for monitoring MMSE-controlled processes	114
7-6. Criticisms concerning the ASPC rule	116
7-7. The Proportional Integral Derivative (PID) chart	118
<b>Chapter 8. Performance of the Automatic Statistical Process Control</b>	<b>121</b>
8-1. Introduction	121
8-2. Performance criteria used for the ASPC technique	121
8-3. Efficiency in SPC monitoring of ASPC controlled processes	124
8-3.1. Comparison of control charts based on the economic model AQC	124
8-3.1.1. Comparisons for AR(1) processes based on the AQC model	124
8-3.1.2. Comparisons for ARMA(1,1) processes based on the AQC model	126
8-3.2. Relative efficiency of charts used in the ASPC scheme	128
8-3.2.1. Sustained shift	128
8-3.2.2. Assignable cause resulting in a trend	128
8-4. Choosing between monitoring the output or the control action	129
8-4.1. The SN ratios under the MMSE controller	129
8-4.2. The SN ratios under the PI controller	130
8-5. Comparisons among the PI and MMSE schemes	132
8-5.1. Relative performance under specific disturbance models	132
8-5.1.1. Performance under ARMA(1,1) disturbance models	132
8-5.1.2. Performance under ARIMA(1,1,1) disturbance models	135
8-5.2. Relative performance under different types of shifts	136
8-5.2.1. Performance when a step change has occurred	137
8-5.2.2. Performance when a drift has occurred	138
8-6. Robustness of the ASPC scheme	138





8-6.1. Controlling an ARMA disturbance with IMA forecasts	139
8-6.2. Controlling an IMA disturbance with ARMA forecasts	139
8-7. Performance of the PID chart	141
<b>Chapter 9. Other types of feedback control</b>	<b>145</b>
9-1. Introduction	145
9-2. Closed-loop controllers	145
9-2.1. Definition of the closed-loop controller	146
9-2.2. Closed-loop output description: the PI controller	146
9-3. Other control schemes	148
9-3.1. Process description of an application	148
9-3.2. Developing different control schemes	149
9-3.3. Evaluation of the control schemes on the application	151
9-3.4. Performance of the control schemes	152
9-3.5. Robustness of the control schemes	153
9-4. The Monitor Wafer Controller (MWC)	153
9-4.1. General remarks	153
9-4.2. Model tuning	154
9-4.3. Stepwise Optimization	158
<b>Chapter 10. Conclusion</b>	<b>159</b>
<b>Appendix</b>	<b>161</b>
<b>References</b>	<b>175</b>





## LIST OF TABLES

Table	Page
9-1. Types of Disturbance included in the model of Eq(9-1)	143
A-1. Factors for constructing Variables Control Charts	162
A-2. ARL Performance of the CUSUM chart with $k=0.5$ and $h=4, 5$	163
A-3. Values of $k$ and the corresponding values of $h$ that give $ARL_0 = 370$ for the two-sided CUSUM chart	163
A-4. Average Run Lengths for several EWMA control schemes	164
A-5. ARL's of EWMAST chart with various $\lambda$ applied to AR(1) processes	165
A-6. Average Run Lengths for CUSUM charts in AR(1) processes using alternative values of $K$	166
A-7. Minimum Batch size required for UBM and WBM charts in AR(1) processes	167
A-8. ARMA charts compared with the corresponding optimal EWMA chart for detecting mean shifts of $1\sigma$ when $(\phi = 0.85)$	167
A-9. Comparisons of ARL's for EWMAST, Residual, Shewhart and M-M charts applied to AR(1) processes	168
A-10. Comparisons of ARL's of the Special-Cause chart (SCC), the Shewhart and the EWMA chart for various ARMA(1,1) parameters	169
A-11. Comparison of Shewhart chart (Residuals, WBM and UBM) ARL's for AR(1) processes	170
A-12. ARL's and CDF's of control charts applied to optimal EWMA forecast residuals for an AR(1) process with parameter $\phi$ and desired in-control ARL of 250	171
A-13. Comparisons of ARL's: ARMAST, EWMAST and SCC on ARMA(1,1) processes	172
A-14. ARL comparisons of control charts on process Output and Control action (MMSE-controlled ARMA(1,1) processes)	173
A-15. ARL comparisons of control charts on process Output and Control action (PI-controlled ARMA(1,1) processes)	173
A-16. Averages of the PM's (and ARL's in parenthesis) for ASPC rules. The assignable cause is a shift in the process mean at observation 251	174



**A-17.** Averages of the PM's (and ARL's in parenthesis) for ASPC 174  
rules. The assignable cause is a trend that starts at observation  
251



## LIST OF FIGURES

Figure	Page
2-1. A typical $(\bar{x}, R)$ control chart.	12
2-2. A typical $(\bar{x}, S)$ control chart.	14
2-3. The p and np charts for $m = 40$ samples of size $n = 5$ each.	17
2-4. The c and u charts for $m = 40$ samples of size $n = 5$ each.	19
2-5. The CUSUM control chart with $n=4$ , $k=0.5$ and $h=4$ .	21
2-6. The EWMA control chart with $n=4$ , $\lambda=0.2$ and $L=3$ .	23
2-7. Example of a Periodogram with five ordinates.	25
2-8. A spectral control chart with frequency $= \omega_k$ and period $= n/k$ .	27
3-1. Independent (a) versus autocorrelated process (b).	31
3-2. The sample/model ACF (a) and the sample/model PACF (b) for the process $X_t = 0.8202X_{t-1} + \epsilon_t - 0.9766\epsilon_{t-1}$ .	36
3-3. The ACF (a) and PACF (b) of the residuals after the model $X_t = 0.8202X_{t-1} + \epsilon_t - 0.9766\epsilon_{t-1}$ has been applied.	36
3-4. An ARIMA model with $d=12$ (a) transformed in an ARMA model (b).	44
4-1. The standard (a) and the modified (b) Shewhart control charts for an ARMA(1,1) model.	47
4-2. The EWMA and the EWMAST charts both applied to autocorrelated data.	50
4-3. The CCC chart for an ARMA(1,1) model with $\phi = 0.74$ and $\theta = 0.32$ .	52
4-4. The residuals chart for a mean shift of $1\sigma$ (a) and of $2\sigma$ (b).	53
4-5. The Shewhart chart of the residuals with $n=3$ for $1\sigma$ (a) and $2\sigma$ (b) shift of the process mean.	57
4-6. Parameter design of ARMA charts.	62
5-1. ARLs of the EWMAST with $\lambda=0.2$ and of the Residuals chart for AR(1) processes with $\phi=0.5, 0.95, -0.5, -0.95$ .	73
5-2. ARLs for a shift of one standard deviation.	76
5-3. ARLs for a shift of 3 standard deviations.	76
5.4. Average Run lengths for control charts applied to forecast	81



residuals used for monitoring various AR(1) processes if their estimated parameter $\phi = 0.5$ .	
<b>5.5. Average Run lengths for control charts applied to forecast residuals used for monitoring various IMA(1,1) processes if their estimated parameter <math>\theta = 0.5</math>.</b>	<b>81</b>
<b>6-1. The control actions for the output of Eq(6-16).</b>	<b>97</b>
<b>6-2. The Individuals and the EWMA charts for the output deviations from target.</b>	<b>98</b>
<b>6-3. The Individuals chart for the adjustments (control actions).</b>	<b>98</b>
<b>6-4. The manual adjustment chart.</b>	<b>99</b>
<b>7-1. The effect of a large shift in the process.</b>	<b>110</b>
<b>9-1. MSE of the output MI under different control strategies.</b>	<b>152</b>
<b>9-2. Schematic of monitor based controller.</b>	<b>155</b>
<b>9-3. Schematic of model tuner.</b>	<b>156</b>



# CHAPTER 1

## Introduction

The development of the statistical field called '*Quality Control*' was due to the need of improving the quality of manufactured goods as well as services, with both being products used by our society. Quality improvement methods can be applied to any area within a company or organization.

The term '*quality*' can be summarized in eight dimensions each one of which specifies quality in a different way. These components are:

- Performance (i.e., if the product does what it is meant to do).
- Conformance (i.e., whether or not the product follows the exact standards specified by the company).
- Reliability (i.e., if the product does or does not fail too often).
- Durability (i.e., the period within which the product is considered as being valid).
- Serviceability (i.e., how easily the product can be repaired).
- Features (i.e., what are the characteristics of the product).
- Appearance (i.e., how attractive is its visual construction).
- Reputation (i.e., how well 'known' the product or the company are).

The quality characteristics cannot be measured in the same way for all cases. There are situations where the characteristic can be measured in a continuous scale as is the length, the weight or the voltage and it is called *variables data*. On the other hand, we confront *attributes data* if the characteristic takes the form of discrete counts, for example when the number of nonconforming products (that is, failing to meet at least one of the specifications) or the number of nonconformities (i.e., specific types of failure) in each unit are measured.



In practical terms, quality is inversely proportional to variability, since more repairs and warranty claims means more rework and, thus, more spent time, effort and money. Consequently, quality improvement is achieved via the reduction of variability in the manufacturing process.

Three major areas where statistics is applied for quality improvement are:

1. **Acceptance sampling.** This is the area of quality control that has been first developed and it is connected with inspection and testing of the incoming raw materials provided by the supplier or of the final product. The inspection is done to a sample of units selected at random from a lot and a decision has to be made about accepting or rejecting the whole lot according to the percentage of the nonconforming products of the sample, that is the ones that fail to meet one or more of the process's specifications.

2. **Statistical Process Control (SPC).** The main tool of this area is the control chart in which the averages of measurements of a quality characteristic in samples taken from the process are plotted versus time or the sample number. The chart consists of three lines: the *central line (CL)*, which shows where this characteristic should fall if there were no exceptional sources of variability, the *upper specification limit (USL)*, which is the largest value allowed for the characteristic, and the *lower specification limit (LSL)*, being the smallest value allowed. As long as the measurements of the samples are within this range, we consider that the product has a satisfactory performance. Statistical process control was an improvement over the acceptance sampling because it detects an eventual problem inside the process instead of just checking the suitability of an already finished product.

3. **Experimental design.** This is the most recent approach in statistical control and it helps to discover the key variables influencing the quality characteristics we are interested in. By systematically varying the controllable input factors of the process, it is possible to determine the effect that these factors have on the output product parameters. This area is a further improvement since it provides a better understanding of the data.





Among these three basic areas, our attention will be focused primarily on the Statistical Process Control (SPC) area because it is the one most broadly used in industry. Control charts were developed under the assumption that process observations are independent. With the development of advanced measurement technology and the increase of sampling frequency, many of today's manufacturing processes display inherently autocorrelated behavior. The presence of autocorrelation in observed data values can profoundly impact the performance of traditional control charts. A solution to this difficulty has been the construction of new control charts or the expansion of the standard ones in order to take into consideration the autocorrelation structure of the data.

After having identified the important variables and their relationship with the process output, then an on-line technique can be employed for monitoring the process. Once the dynamic nature of the relationship between the inputs and the outputs is understood, it may be possible to adjust the process in order to keep future values of the product characteristic close to the target of the process. This adjustment is called **Engineering Process Control (EPC)** or **feedback control adjustment** and it differs from the statistical control charts in which corrective action is taken only after a sample average *has already fallen outside the specified control limits*.

The aim of this thesis is to present the most recent approaches concerning the autocorrelated manufacturing process including both the SPC and EPC techniques for quality improving, which may be combined and integrated to form a more elaborate system termed **Automatic Process Control (APC)**. The performance of the two methods tools is thoroughly studied and the best choice is provided according to the particularity of the data.

Specifically, Chapter 2 discusses the traditional control charts of the SPC technique applied to uncorrelated data, Chapter 3 briefly mentions the most popular time-series models applied to correlated data, Chapter 4 is referred to the control charts used when the process consists of observations that are dependent over time, Chapter 5 concentrates on the performance of the control charts described in Chapter 4, Chapter 6 introduces the most common tools of the EPC/APC system, Chapter 7 is concerned with more



special issues of EPC, Chapter 8 combines the performance of the EPC/APC tools and Chapter 9 describes some other types of EPC techniques less encountered in practice. Finally, in Chapter 10 some general comments and conclusions are applied.



## CHAPTER 2

### Overview of the most common control charts

#### 2-1 Introduction

The Statistical Process Control consists of a variety of tools easy to implement, with the control charts being the ones used extensively in industry because of combining simplicity and effectiveness. A brief description of these tools along with a more elaborate reference to the control charts is the subject of section 2-2. In section 2-3 the Shewhart-type control charts are presented, while section 2-4 is concerned with control charts for attributes, that is, quality characteristics that cannot be conveniently represented numerically. The alternative to Shewhart control charts, i.e., the EWMA and CUSUM charts, are illustrated in section 2-5, while section 2-6 presents the more recent Spectral chart, constructed to detect cyclic behaviors in the process mean.

#### 2-2 Properties of Statistical Process Control (SPC)

The SPC area is known to have seven major tools that help detect the time point at which the process deviates from its normal conditions and that provide ways to identify the cause of this deviation. Many types of control charts are constructed according to the change of the pattern of the process mean that one wishes to detect more effectively.

##### 2-2.1 Basic tools of SPC

It is recommended that before ‘forcing’ the product to meet the requirements needed, the stability of the process must be ensured. In other

words, the process should be centered around a target value for each specific characteristic. A set of diagnostic and problem-solving tools helping to achieve process stability and, thus, to reduce variability is known as the '*magnificent seven*'. These are:

- Histogram or stem-and-leaf plot
- Check sheet
- Pareto chart
- Cause-and-effect diagram
- Scatter diagram
- Control chart

The histogram, as well as the stem-and-leaf plot, displays the frequency distribution, i.e., the arrangement of the data by magnitude. What we look forward to is a histogram (or a stem-and-leaf plot) in which the great mass of the data is centered on the target value and the rest of the data are dispersed little around this nominal value. The check sheet is also used at the early stages of SPC implementation, where the historical or current operating data of the process are collected. The Pareto chart is a frequency distribution of attribute data *arranged by category*, while the cause-and-effect diagram is used to analyze potential causes after a defect or problem has been identified in the process and has been isolated for further study. The scatter diagram plots two variables (these may be the values of an important raw material and the corresponding values of the output characteristic) in order to define the potential relationship between them. The existence of correlation between the two variables does not necessarily imply causality, which must be verified only after having used designed experiments.

The control chart has been introduced as the main tool of SPC because it has been proven to be a technique useful for improving productivity, preventing defects effectively by not resulting in unnecessary process adjustment and, lastly, providing diagnostic information as well as information about process capability. The control chart may be presented by the following general model. If  $h$  is a sample statistic measuring the quality characteristic of interest, with mean  $\mu_h$  and standard deviation  $\sigma_h$ , then the center line and upper and lower control limits are:

$$\begin{aligned} \text{UCL} &= \mu_h + L \sigma_h \\ \text{Center line} &= \mu_h \\ \text{LCL} &= \mu_h - L \sigma_h \end{aligned} \quad (2-1)$$

where  $L$  is the distance of the control limits from the center line expressed in standard deviation units. This general theory of control charts was first proposed by Walter S. Shewhart and the control charts conforming to these principles are called *Shewhart control charts*.

In standard applications of statistical process control, a state of statistical control is identified with a random process, that is, a process generating independent and identically distributed (iid) random variables. Once a state of statistical control is attained, departures typically are reflected in extreme individual observations (outliers) or aberrant sequences of observations (runs above and below a level or runs up and down).

Departures from a state of statistical control are discovered by plotting and viewing data on control charts, such as the Shewhart, the Cumulative Sum (CUSUM), the Exponentially Weighted Moving Average (EWMA), and moving-average charts. Having found departures, we hope to find explanations for them in terms of assignable or special causes. ‘Assignable cause’ is a term introduced by Shewhart, while ‘special cause’ is an alternative term suggested by Deming. We then hope to move from ‘out of control’ to ‘in control’ by correcting or removing the special causes.

### 2-2.2 Out-of-control patterns detected by control charts

The assignable causes that affect a production process may be summarized in groups based on the type of the out-of-control patterns. Some of these types are the following (Beneke et al., 1988):

1. *Sudden shift in level*. This condition is associated with a sudden change in the average of the process. This change could be caused by an alteration of the process setting, a difference in raw materials, or a minor failure of a machine part.

2. *Trend or steady change in level.* This condition is associated with a gradual change in the average of the process. Some of the causes for this condition are tool wear and equipment deterioration.

3. *Several populations.* This condition exists when items come from more than one population. Some of the causes for this condition are items from different suppliers, machines, or workers being plotted on the same chart.

4. *Recurring cycles.* The process has periodic high and low points that might provide cause for concern. Some of the causes for recurring cycles are the seasonal variations in incoming materials, the recurring effects of temperature and humidity, any daily or weekly chemical, mechanical or psychological events, and the periodic rotation of operators.

The Shewhart control charts, as well as the most recently developed charts CUSUM and EWMA, were constructed for detecting special causes of the first three types, while a more recent chart called the '*Spectral chart*' was initiated in an effort to take into account the fourth type. All these charts are further presented in more detail.

## 2-3 Shewhart-type control charts

The Shewhart control charts are charts used to implement variables, that is quality characteristics that are measured on a numerical scale. When dealing with a variable, it is usually necessary to monitor both the mean value of the quality characteristic and its variability. The mean quality level is controlled via the chart for means. Process variability can be monitored with either a control chart for the standard deviation, called the S chart, or a control chart for the range, called an R chart.

### 2-3.1 General properties

In order to construct a control chart for the quality characteristic we are interested in, our first step is to take  $m$  samples each containing  $n$  observations of the characteristic during the process. The  $m$  preliminary

samples help us to construct the trial control limits. If all  $m$  past points are inside the control limits, then we consider that the process is in control and the trial control limits may be used for controlling future production. This analysis of past data is referred to as a phase 1 analysis, in which about 20-25 samples of size 3-5 each should be used. On the other hand, if out-of-control points are found, we should check whether they were due to assignable causes or not. In the first case, the points should be discarded and the trial limits are recalculated using only the remaining points. In the second situation, the points may be eliminated, as previously, or be retained if they do not distort the control limits significantly.

The phase 2 analysis consists of plotting the points of the new collected samples on the control chart with limits calculated from the preliminary samples. If  $\mu$  is the *average of the process*, the best estimator of  $\mu$  is the grand average  $\bar{\bar{x}}$ , where:

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_m}{m} \quad (2-2)$$

and  $\bar{x}_m$  is the average of the  $m$ th sample (consisting of  $n$  observations). Assuming that the *quality characteristic is normally distributed* with mean  $\mu$  and standard deviation  $\sigma$ , the probability is  $1-\alpha$  that any sample mean  $\bar{x}$  will fall between:

$$\mu + Z_{1-\alpha/2} \sigma_{\bar{x}} = \mu + Z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \mu - Z_{1-\alpha/2} \sigma_{\bar{x}} = \mu - Z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (2-3)$$

Therefore, if  $\mu$  and  $\sigma$  are known, they could be used to calculate the upper and lower control limits for sample means. It is customary to replace  $Z_{1-\alpha/2}$  by 3, so that *three-sigma limits* are employed. If a sample mean falls outside of these limits, it is an indication that the process mean is no longer equal to  $\mu$ . The three-sigma limits setting  $Z_{1-\alpha/2} = 3$  correspond to  $\alpha = 0.0027$ , meaning that 27 out of 10 000 observations may fall out of the control limits without being the result of an assignable cause. In other words,  $\alpha$  is the probability of type I error, i.e., the probability that, although the null

hypothesis that the process is in control is true, we have an out-of-control sign. This sign is often called a **false alarm**. On the other hand,  $\beta$  symbolizes the probability that, while the null hypothesis is not true, meaning that the process is out-of-control, we do not have points out of the control limits. Consequently,  $1-\beta$  is the probability that we successfully consider the process as being out-of-control.

The **Average Run Length (ARL)** is the *expected number of samples taken before the shift is detected*, or  $ARL = 1/P(\text{one point plots out of control})$ . Thus, the in-control ARL ( $ARL_0$ ) =  $1/\alpha$ , while the out-of-control ARL ( $ARL_1$ ) =  $1/(1-\beta)$ . Naturally, when the process has been fallen out-of-control, a small value for the ARL is desired, while if there is not any assignable causes disturbing the data, the ARL value is preferred to be large so that a false alarm is avoided.

When the mean  $\mu$  and the standard deviation  $\sigma$  are not known and have to be estimated, there are two approaches available depending on if the standard deviation of the process is estimated by the sample range or by the sample standard deviation. The first approach gives the  $(\bar{x}, R)$  control chart and the second is known as the  $(\bar{x}, S)$  control chart.

### 2-3.2 The $(\bar{x}, R)$ control chart

If  $x_1, x_2, \dots, x_n$  is a sample of size  $n$ , then the *range* of the sample is the difference between the largest and the smallest observation, that is  $R = x_{\max} - x_{\min}$ . Since we have a set of  $m$  samples, we calculate  $m$  sample ranges, i.e.  $R_1, R_2, \dots, R_m$ . The average range  $\bar{R}$  is the average of all of the  $m$  ranges. It has been proven that  $\hat{\sigma} = \frac{\bar{R}}{d_2}$ , where  $d_2$  is the mean of the value  $R/\sigma$ . With  $\bar{x}$  the estimator of  $\mu$  and  $\bar{R}/d_2$  the estimator of  $\sigma$ , the values of the  $(\bar{x}, R)$  chart are:

Control limits of the  $\bar{x}$  chart

$$UCL = \bar{\bar{x}} + 3 \bar{R} / d_2 \sqrt{n} = \bar{\bar{x}} + A_2 \bar{R}$$

$$\text{Center line} = \bar{\bar{x}}$$

$$LCL = \bar{\bar{x}} - 3 \bar{R} / d_2 \sqrt{n} = \bar{\bar{x}} - A_2 \bar{R}$$



**Control limits of the R chart**

$$UCL = \bar{R} + 3 \hat{\sigma}_R = \bar{R} + 3d_3 \bar{R} / d_2 = (1 + 3d_3/d_2) \bar{R} = D_4 \bar{R}$$

$$\text{Center line} = \bar{R}$$

$$LCL = \bar{R} - 3 \hat{\sigma}_R = \bar{R} - 3d_3 \bar{R} / d_2 = (1 - 3d_3/d_2) \bar{R} = D_3 \bar{R} \quad (2-5)$$

where  $A_2 = 3/d_2 \sqrt{n}$ ,  $\hat{\sigma}_R = d_3 \frac{\bar{R}}{d_2}$ ,  $D_3 = 1 - 3d_3/d_2$  and  $D_4 = 1 + 3d_3/d_2$ . If the true values of  $\mu$  and  $\sigma$  are known, then the parameters of the  $(\bar{x}, R)$  chart are modified as follows:

**Control limits of the  $\bar{x}$  chart: standards given**

$$UCL = \mu + 3 \frac{\sigma}{\sqrt{n}} = \mu + A \sigma$$

$$\text{Center line} = \mu$$

$$LCL = \mu - 3 \frac{\sigma}{\sqrt{n}} = \mu - A \sigma \quad (2-6)$$

**Control limits of the R chart: standards given**

$$UCL = d_2 \sigma + 3d_3 \sigma = (d_2 + 3d_3) \sigma = D_2 \sigma$$

$$\text{Center line} = d_2 \sigma$$

$$LCL = d_2 \sigma - 3d_3 \sigma = (d_2 - 3d_3) \sigma = D_1 \sigma$$

Values of  $d_2$ ,  $A$ ,  $A_2$ ,  $D_1$ ,  $D_2$ ,  $D_3$ ,  $D_4$  are tabulated in tables according to the sample size  $n$ . These tables can be easily found in the literature (e.g., see Montgomery, 2001).

An example of an  $(\bar{x}, R)$  chart in which the quality characteristic is the weight of the product and 60 samples of 4 products each have been chosen randomly from the process, is presented in Figure 2-1. The calculations for the limits of the chart have been based on Eq(2-4) and (2-5). The means of all the samples are inside the control limits indicating that these limits may be used in order to check for the compatibility of future samples instead of calculating new limits each time we gather a new random sample. The average ranges of all samples are also inside the control limits of the R chart and, thus, the standard deviations do not imply an out-of-control situation.

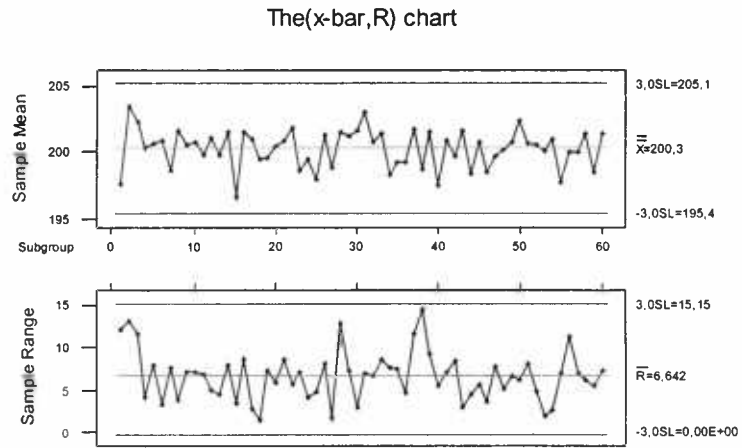


Figure 2-1: A typical  $(\bar{x}, R)$  control chart.

### 2-3.3 The $(\bar{x}, S)$ control chart

The  $\bar{x}$  and S control charts are preferred to the  $\bar{x}$  and R charts when the sample size  $n$  is large (usually greater than 10) or when the sample size is variable from sample to sample, because in this case the range method for estimating  $\sigma$  loses statistical efficiency. An unbiased estimator of  $\sigma^2$  which is widely known is the sample variance:

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (2-7)$$

However, S is an unbiased estimator of  $c_4\sigma$ , where  $c_4$  is a constant depending on the sample size  $n$ . Furthermore, the standard deviation of S is  $\sigma\sqrt{1-c_4^2}$ . Therefore, when  $m$  preliminary samples are available, each of size  $n$ ,  $\bar{S}$  is the average of the standard deviations of all the  $m$  samples and  $\bar{S} / c_4$  is an unbiased estimator of  $\sigma$ . The three-sigma control limits for the  $\bar{x}$  and S charts are:

Control limits of the  $\bar{x}$  chart

$$UCL = \bar{\bar{x}} + 3\bar{S} / c_4\sqrt{n} = \bar{\bar{x}} + A_3\bar{S}$$

$$\text{Center line} = \bar{\bar{x}}$$

$$LCL = \bar{\bar{x}} - 3\bar{S} / c_4\sqrt{n} = \bar{\bar{x}} - A_3\bar{S}$$

**Control limits of the S chart**

$$UCL = \bar{S} + 3 \hat{\sigma}_s = \bar{S} + 3 \frac{\bar{S}}{c_4} \sqrt{1 - c_4^2} = (1 + 3 \sqrt{1 - c_4^2} / c_4) \bar{S} = B_4 \bar{S}$$

$$\text{Center line} = \bar{S}$$

$$LCL = \bar{S} - 3 \hat{\sigma}_s = \bar{S} - 3 \frac{\bar{S}}{c_4} \sqrt{1 - c_4^2} = (1 - 3 \sqrt{1 - c_4^2} / c_4) \bar{S} = B_3 \bar{S} \quad (2-9)$$

If the true values of the mean  $\mu$  and of the standard deviation  $\sigma$  of the process are known, then the limits of the  $\bar{x}$  and S charts become:

**Control limits of the  $\bar{x}$  chart: standards given**

$$UCL = \mu + 3 \frac{\sigma}{\sqrt{n}} = \mu + A \sigma$$

$$\text{Center line} = \mu$$

$$LCL = \mu - 3 \frac{\sigma}{\sqrt{n}} = \mu - A \sigma \quad (2-10)$$

**Control limits of the S chart: standards given**

$$UCL = c_4 \sigma + 3 \sigma \sqrt{1 - c_4^2} = (c_4 + 3 \sqrt{1 - c_4^2}) \sigma = B_6 \sigma$$

$$\text{Center line} = c_4 \sigma$$

$$LCL = c_4 \sigma - 3 \sigma \sqrt{1 - c_4^2} = (c_4 - 3 \sqrt{1 - c_4^2}) \sigma = B_5 \sigma$$

Values of  $c_4$ ,  $A_3$ ,  $B_3$ ,  $B_4$ ,  $B_5$  and  $B_6$  can also be found easily in tables, implemented in the literature, according to the selected sample size  $n$ .

It is generally assumed that when the data arise independently from a common normal distribution, it should not make difference if the range method or the standard deviation method are used as estimates of  $\sigma$ . However, the  $(\bar{x}, S)$  chart seems to be a safer approach in more extreme cases, as when trends and oscillations affect the data (Cryer and Ryan, 1990). With the wide use of relative software, the construction of the  $(\bar{x}, S)$  chart is not time-consuming any more and it has replaced the  $(\bar{x}, R)$  chart that was used for its simplicity only. Figure 2-2 presents a typical  $\bar{x}$  and S chart using Eq(2-8) and (2-9) for the same data as in Figure 2-1.

The ( $\bar{x}$ , S) chart

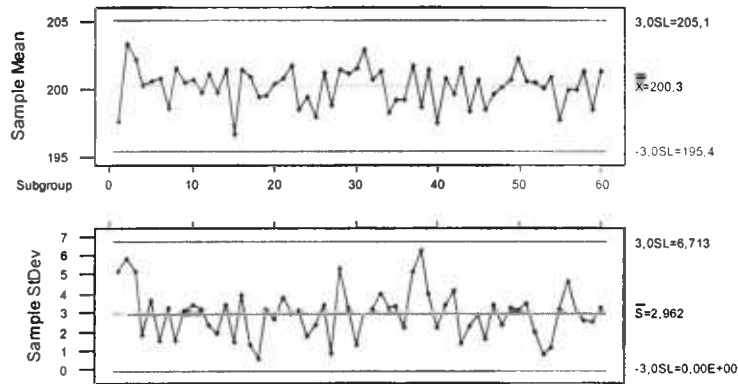


Figure 2-2: A typical ( $\bar{x}$ , S) control chart.

Two subcategories of the  $\bar{x}$  and S charts are:

1. *The  $\bar{x}$  and S charts with variable sample size*

If the samples do not all have the same number of observations, one way to deal with the problem is to use weighted averages in calculating  $\bar{\bar{x}}$  and  $\bar{\bar{S}}$ . This will lead to upper and lower control limits that are not a straight line but vary according to the sample size. An alternative is to calculate the average sample size or to use the sample size which is the most often so as to design a control chart with a single value for the upper and a single for the lower limit.

2. *The  $\bar{x}$  and S charts for Individual measurements*

In many situations, the sample consists of an individual unit because the production rate may be very slow. In many applications of the individuals control chart we use the moving range of two successive observations in order to estimate process variability. The moving range is defined as  $MR_I = |x_I - x_{I-1}|$  and its control limits are:

**Control limits of the MR chart**

$$UCL = \bar{\bar{x}} + 3 \frac{\overline{MR}}{d_2} = \bar{\bar{x}} + 3 \overline{MR} / 1.128$$

$$\text{Center line} = \bar{\bar{x}}$$

$$LCL = \bar{\bar{x}} - 3 \frac{\overline{MR}}{d_2} = \bar{\bar{x}} - 3 \overline{MR} / 1.128 \quad (2-11)$$

**Control limits of the R chart**

$$UCL = D_4 \bar{MR} = 3.267 \bar{MR}$$

$$\text{Center line} = \bar{MR}$$

$$LCL = D_3 \bar{MR} = 0 \quad (2-12)$$

where 1.128 is the value of  $d_2$ , 3.267 is the value for  $D_4$  and 0 is the one for  $D_3$  as they are indicated from the corresponding tables for  $n=2$  (e.g., see Montgomery, 2001). Cryer and Ryan (1990) proposed, however, that the estimation of  $\sigma$  for the Individuals chart should be based on  $S/c_4$  of the  $(\bar{x}, S)$  control chart because  $\bar{MR}/d_2$  tends to inflate the variance considerably.

### 2-3.4 Selection of the sampling scheme

In our analysis about constructing the Shewhart control charts, we considered that the choice of the sample size was evident. In practice, however, it is difficult to decide on a specific value for it. The concept of rational subgroups is of great help but a choice has still to be made between taking samples consecutively or selecting items at a long time interval apart the one from the other. The first approach is more effectively used when the primary scope is to detect shifts in the process while the second when a difference between two samples is of main concern. Consequently, the sampling scheme is defined by both the sample size and the sampling frequency.

A final remark is that a problem may not be indicated exclusively by an out-of-control point, but the existence of a nonrandom pattern may also reveal an abnormality in the process. A set of decision rules for recognizing nonrandom patterns consists of considering the process as being out of control if either:

- One point plots outside the three-sigma control limits
- Two out of three consecutive points plot beyond the two-sigma limits (known as warning limits)
- Four out of five consecutive points plot at a distance of one-sigma or beyond from the center line

- Eight consecutive points plot on one side of the center line.

However, when using the above decision rules simultaneously, the probability of type I error is inflated, resulting in an excessive number of false alarms. That is why, the sensitizing rules should be applied with considerable caution.

## 2-4 Control charts for attributes

When a quality characteristic cannot be conveniently represented numerically, it is often suitable to classify it as **conforming** or **nonconforming** to the specifications. The quality characteristics of this type are called attributes.

### 2-4.1 Control charts for fraction/number nonconforming

The *fraction nonconforming* ( $p$ ) is the ratio of the number of nonconforming items in a population to the total number of items in that population. The sample fraction nonconforming is the ratio of the number of nonconforming units in the sample  $D$  to the sample size  $n$ , that is  $\hat{p}=D/n$ . Because the distribution of the random variable  $\hat{p}$  can be obtained by the binomial, its mean value is  $\mu=p$  and  $\sigma_{\hat{p}}^2=p(1-p)/n$ . If the fraction nonconforming of the process is not known, then it is estimated from  $m$  samples by calculating  $\hat{p}_i=D_i/n$  for each sample and then by averaging the  $m$   $\hat{p}_i$ 's to get:

$$\bar{p} = \frac{\sum_{i=1}^m \hat{p}_i}{m} \quad (2-13)$$

Therefore, by substituting  $p$  with  $\bar{p}$  if the true fraction nonconforming is unknown, the control limits of the chart for fraction nonconforming are specified as:

**Control limits of the p chart**

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

$$\text{Center line} = p \quad (2-14)$$

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

It is, however, possible and more convenient sometimes to base a control chart on the *number nonconforming* (np) rather than the fraction nonconforming (p). The control limits for the np chart are derived by a simple modification of Eq(2-14) as:

**Control limits of the np chart**

$$UCL = np + 3\sqrt{np(1-p)}$$

$$\text{Center line} = np \quad (2-15)$$

$$LCL = np - 3\sqrt{np(1-p)}$$

Figure 2-3 illustrates typical p and np charts in which the number of nonconforming items observed in each sample of size 5 varied between 0 and 2. It seems that only the 13<sup>th</sup> sample has an unusual number of nonconformities (i.e., 3), that is why a signal has been marked at this point. Obviously, the np chart is just a multiplier of the p chart and, thus, the information obtained is similar with both charts.

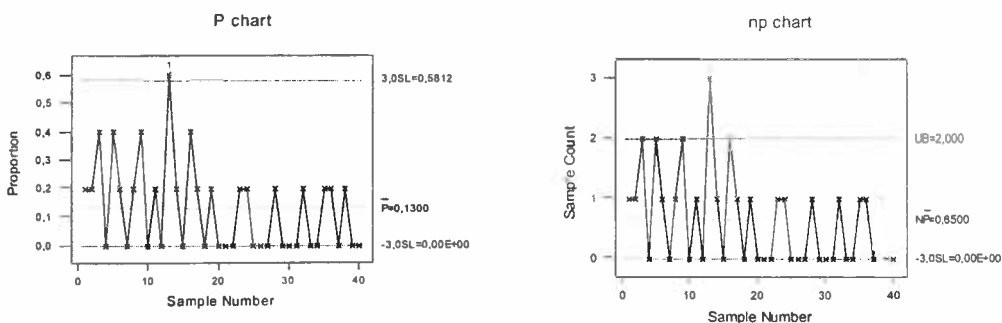


Figure 2-3: The p and np charts for m = 40 samples of size n = 5 each.

## 2-4.2 Control charts for nonconformities (defects)

A **nonconforming** item is a unit of product that does not satisfy one or more of the specifications for that product. Each specific point at which a specification is not satisfied results in a defect or nonconformity. The number of defects ( $c$ ) is considered to follow the Poisson distribution since it defines the number of occurrences at a specific interval. Therefore, the mean and the variance of the random variable  $c$  are both equal to  $c$  itself. If no standard is given for  $c$ , then it is estimated as the observed average number of nonconformities in a preliminary sample of inspection units. The limits of the control chart for nonconformities are:

**Control limits of the  $c$  chart**

$$\begin{aligned} \text{UCL} &= c + 3\sqrt{c} \\ \text{Center line} &= c \\ \text{LCL} &= c - 3\sqrt{c} \end{aligned} \quad (2-16)$$

A more obvious approach would probably be to construct a control chart for the *number of nonconformities per inspection unit* ( $u$ ). The value  $u$  can be considered as the ratio of the total nonconformities in a sample over the  $n$  inspection units. The control chart for nonconformities per unit (the  $u$  chart) has the following limits:

**Control limits of the  $u$  chart**

$$\begin{aligned} \text{UCL} &= u + 3\sqrt{\frac{u}{n}} \\ \text{Center line} &= u \\ \text{LCL} &= u - 3\sqrt{\frac{u}{n}} \end{aligned} \quad (2-17)$$

Figure 2-4 shows both a  $c$  and a  $u$  chart with the number of nonconformities varying between 2 and 7. All samples seem to have a reasonable number of defects under the specific scheme. The lower limit of the chart is set to zero, because there can be no negative value for the defects.



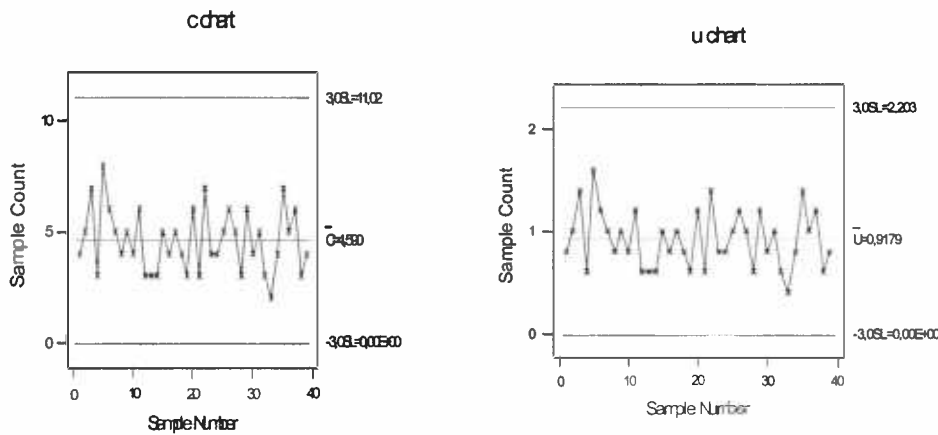


Figure 2-4: The c and u charts for  $m = 40$  samples of size  $n = 5$  each.

## 2-5 The third-generation control charts: CUSUM and EWMA

A great disadvantage of the Shewhart control charts is that they only use information on the last plotted point so as to draw conclusions about the stability of the process. Two effective alternatives to the Shewhart control chart considering the entire sequence of plots are the Cumulative Sum (CUSUM) and the Exponentially Weighted Moving Average (EWMA) control charts. Both the CUSUM and the EWMA charts have been proven (e.g., see Lucas and Saccucci, 1990) to perform better than the Shewhart control chart when we are interested in detecting small shifts. In other words, their ARL value is smaller when the shift of the process mean is between  $0.5\sigma$  to  $2\sigma$  more or less than the initial mean of the process.

### 2-5.1 The CUSUM control chart

If the process is in control, the quality characteristic  $x$  has a normal distribution with mean  $\mu_0$  (this is the target value of the characteristic) and a standard deviation  $\sigma$ . This assumption is the same as the one used in the Shewhart control charts, with the difference that now it is not the averages of the samples that are plotted but these are modified as accumulating derivations from  $\mu_0$ . The derivations that are above target are summarized in

the statistic  $C^+$  (one-sided upper cusum), while the ones below target are symbolized as  $C^-$  (one-sided lower cusum). More precisely,

<p><b>Plotted points on the CUSUM chart</b></p> $C_i^+ = \max \{0, (x_i - \mu_0) - K + C_{i-1}^+\}$ $C_i^- = \min \{0, (\mu_0 - K) - x_i + C_{i-1}^-\},$ <p>where <math>C_0^+ = C_0^- = 0</math> and <math>K = \frac{ \mu_1 - \mu_0 }{2} = \delta\sigma/2</math> when the shift is expressed as</p> $\mu_1 = \mu_0 + \delta\sigma.$ <p><b>Control limits of the CUSUM chart</b> <span style="float: right;">(2-18)</span></p> $UCL = H$ $\text{Center line} = \mu_0$ $LCL = -H$
--

$K$  is often called the '*reference value*' and its value is defined according to the smallest shift in the process mean (measured in standard errors and expressed by  $\delta$ ) that is considered important to be detected quickly. There are many debates concerning the value of  $H$  and  $K$ . However, according to the ARL performance of various values for  $H$  computed using simulation, it has been proven that, by considering  $H$  and  $K$  as multiples of the standard deviation  $\sigma$ , i.e.  $H=h\sigma$  and  $K=k\sigma$ , a value of 4 or 5 for  $h$  and of 0.5 for  $k$  gives smaller values to ARL when a shift has occurred and larger when no shift has occurred than do other choices (see Montgomery, 2001). Thus, these are the recommended values for  $h$  and  $k$ . If an out-of-control point appears, one should search for the assignable cause, take any corrective action required and then reinitialize the cusum at 0.

The cusum chart specified in Eq(2-18) presents the case where the sample size is only one unit. If  $n$  is greater than 1, the value  $x_i$  should be replaced by the mean of the sample  $\bar{x}_i$  and  $\sigma$  by the standard deviation  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ , where  $\sigma$  is usually estimated from the data. Finally, Eq (2-18) could be modified to account for the variability and not only for the derivation from the mean of the data if we set  $(x_i - \mu_0)/\sigma_{\bar{x}}$ , instead of  $x_i - \mu_0$ .

This is often called the '*Standardized Cusum chart*'. Figure 2-5 shows a typical CUSUM chart for the same data used in the previous Figures.

#### The Fast Initial Response (FIR) or Headstart Feature

This procedure was revised by Lucas and Crosier (1982) to improve the sensitivity of a cusum at process start-up. The Fast Initial Response (FIR) or headstart sets the starting values  $C_0^+$  and  $C_0^-$  equally to a nonzero value (a good choice would be  $H/2$ , i.e., a 50% headstart). If the process is in control, the values of  $C_0^+$  and  $C_0^-$  are soon not affected by the headstart because consecutive observations near the target value set the cusums rapidly to zero. On the other hand, if the process is out-of-control, points are plotted out of cusums earlier than when no headstart has been used. Therefore, if the FIR is applied to the CUSUM chart, there is a great possibility of detecting a shift faster.

The CUSUM chart

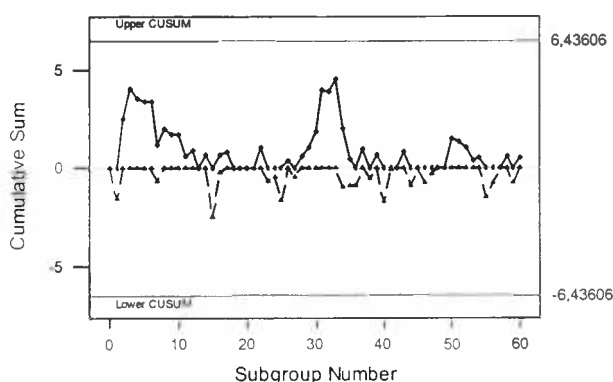


Figure 2-5: The CUSUM control chart with  $n=4$ ,  $k=0.5$  and  $h=4$ .

#### 2-5.2 The EWMA control chart

The EWMA control chart was introduced by Roberts (1959). It is also effective when small shifts should be detected and its performance is approximately equivalent to that of the CUSUM chart. Furthermore, it is somewhat easier to operate and it is model free. The consecutive points plotted on the EWMA chart, as well as the control limits, are calculated as:

Plotted points of the EWMA chart

$$z_i = \lambda x_i + (1-\lambda)z_{i-1} = \lambda \sum_{j=0}^{i-1} (1-\lambda)^j x_{i-j} + (1-\lambda)^i z_0, \text{ with starting value } z_0 = \mu_0 \quad (2-19a)$$

Control limits of the EWMA chart

$$UCL = \mu_0 + L\sigma_z = \mu_0 + L\sigma \sqrt{\frac{\lambda}{(2-\lambda)} [1 - (1-\lambda)^{2i}]} \xrightarrow{i \rightarrow \infty} \mu_0 + L\sigma \sqrt{\frac{\lambda}{(2-\lambda)}} \\ \text{Center line} = \mu_0 \quad (2-19b)$$

$$LCL = \mu_0 - L\sigma_z = \mu_0 - L\sigma \sqrt{\frac{\lambda}{(2-\lambda)} [1 - (1-\lambda)^{2i}]} \xrightarrow{i \rightarrow \infty} \mu_0 - L\sigma \sqrt{\frac{\lambda}{(2-\lambda)}},$$

where  $\sigma$  is the estimated standard deviation of the original data  $X$ .

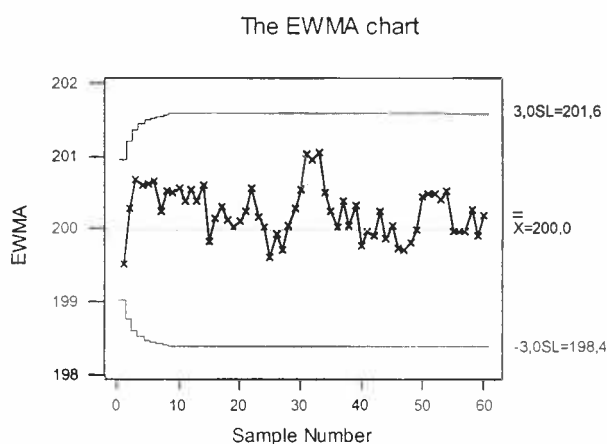
From Eq(2-19a), it is easily seen that the weights  $\lambda(1-\lambda)^j$  decrease geometrically with the age of the sample mean and that they sum to unity. The plotted points are not the observations themselves but they are the sum of the weighted values of all the previous observations. The fact that the observations are accumulated makes the detection of a shift quicker, so that less time is needed to observe an out-of-control signal than when the Shewhart-type control chart is used.

Considering Eq(2-19b), one observes that, since the term  $[1-(1-\lambda)^{2i}]$  approaches unity as  $i$  gets large, after the EWMA chart has run for several time periods, its control limits are stabilized. Simulation procedures for the ARL performance of the EWMA have shown that values of  $\lambda$  in the interval  $0.05 \leq \lambda \leq 0.25$ , and  $L$  around 3 give satisfactory values for the average run length (e.g., see Montgomery, 2001).

A rule of thumb is that if small shifts are to be detected, then a small value for  $\lambda$  is more appropriate and a larger value otherwise. This is to be expected since a smaller smoothing constant gives more weight to older observations and, thus, the effect if the shift is small is accumulated more than when the smoothing constant is large. Hence, smaller shifts are detected sooner when  $\lambda$  is small. In contrast, when the shift is large, it is detected sooner, since more weight is given to more recent observations that have a higher mean. If  $\lambda = 1$ , the EWMA chart is equivalent to the Shewhart.

To design an EWMA control scheme, Lucas and Saccucci (1990) proposed to select a value for the parameter  $\lambda$  that results in the minimum ARL for the specified shift that one wants to detect. They have provided tables with the ARL values for various cases derived by simulation procedures, so as to help someone decide upon the most appropriate value for  $\lambda$ .

If each sample consists of more than one unit ( $n > 1$ ), then  $x_i$  should be replaced by  $\bar{x}_i$  and  $\sigma$  by  $\sigma_{\bar{x}} = \sigma / \sqrt{n}$  in Equation (2-19b). Figure 2-6 represents an EWMA chart for the data that were also used in the previous Figures.



**Figure 2-6: The EWMA control chart with  $n=4$ ,  $\lambda=0.2$  and  $L=3$ .**

### The FIR Feature

As in the case of the CUSUM chart, the FIR feature is also useful in the EWMA schemes, especially if the value of  $\lambda$  is small. This is because, when  $\lambda$  is small, the variance of the control statistic converges slowly to its asymptotic value and, thus, control schemes based only on the asymptotic standard deviation tend to be insensitive at start-up. The FIR feature in the EWMA control chart is obtained by simultaneously implementing two one-sided EWMA's, each with a headstart (HS). One EWMA has a HS above the target value and the other below the target value. If the process is off aim at start-up, the EWMA with the appropriate HS (which usually has a starting value of 50% between the process target and the control limits) will give an out-of-control signal more quickly. On the other hand, if the process is in-control, at least initially, the two EWMA's will tend to converge.

### **Combined Shewhart-EWMA chart**

Although the CUSUM and the EWMA control charts perform well against small shifts, the Shewhart chart reacts better to large shifts. Consequently, the Shewhart control charts should be used along with the third generation charts (EWMA and CUSUM) in order to be protected both against small and large shifts in the mean. This is achieved by adding Shewhart limits to an EWMA (or CUSUM) control scheme, so that an out-of-control signal is given if the EWMA (or CUSUM) statistic is outside the control limits or if the current observation is outside the Shewhart limits.

### **The EWMA control scheme compared to the CUSUM**

The property of the EWMA chart that it is not sensitive to normality has been proven from the fact that applying this chart to nonnormal distributions does not affect the ARL values (e.g, see Montgomery, 2001) something that does not happen in the case of Shewhart or CUSUM charts. In terms of the ARL properties, there is little practical difference between the EWMA and the CUSUM chart. According to the study of Lucas and Saccucci (1990), the ARL values for the EWMA chart are usually somewhat smaller than the ones of the CUSUM up to a value of the shift near the one that the scheme was designed to detect. Beyond this shift, though, the CUSUM chart has smaller ARL values and, thus, it detects the mean shifts more quickly than the EWMA.

## **2-6 The Spectral chart**

Spectral analysis has been used to detect and evaluate periodicities in equally spaced time-ordered data by decomposing the data into its periodic components. This is the technique based on which the spectral control chart has been developed by Beneke et al. (1988) with the purpose of detecting periodic behavior. The spectral control chart is much more recent than the Shewhart, CUSUM and EWMA and it is not very commonly applied in the manufacturing procedure. According to the spectral analysis,  $n$  consecutive measurements of the process average represent a finite realization of a time series and a value computed at time  $t$  can be represented by (Chatfield, 1984):

$$X_t = \alpha_0/2 + \sum_{k=1}^m (\alpha_k \cos \omega_k t + b_k \sin \omega_k t), \quad t = 1, 2, \dots, n$$

where  $\omega_k = 2\pi k/n$ ,  $k = 0, 1, 2, \dots, m$

$$\alpha_k = 2/n \sum_{t=1}^n X_t \cos \omega_k t, \quad k = 0, 1, 2, \dots, m \quad (2-20)$$

$$b_k = 2/n \sum_{t=1}^n X_t \sin \omega_k t, \quad k = 1, 2, \dots, m \quad \text{and } m = n/2 \text{ with } n \text{ even}$$

The frequencies  $\omega_k$  can be expressed in cycles per unit time as  $f_k = \omega_k/2\pi = k/n$ . The period corresponding to the frequency  $f_k$  is then given by  $T = n/k$ . The Fourier series contains periodic components at each of the frequencies  $\omega_1, \omega_2, \dots, \omega_m$ . This type of analysis partitions the variability of the data into components at frequencies  $2\pi/n, 4\pi/n, \dots, \pi$ . The component at frequency  $\omega_k = 2\pi k/n$  is referred to as the  $k$ th harmonic. For  $k \neq n/2$  (i.e.,  $\omega_k \neq \pi$ ), the  $k$ th harmonic is given by:

$$\begin{aligned} \alpha_k \cos \omega_k t + b_k \sin \omega_k t &= R_k \cos(\omega_k t + \phi_k), \quad \text{where} \\ R_k &= \text{amplitude of the } k\text{th harmonic} = (a_k^2 + b_k^2)^{1/2} \quad \text{and} \\ \phi_k &= \text{phase of the } k\text{th harmonic} \end{aligned} \quad (2-21)$$

The periodogram can be visualized as a bar chart consisting of  $k$  cells. Its purpose is to estimate the spectrum of the process, analogous to the way a histogram is used to estimate the probability density function of a distribution. The area of each histogram rectangle is the contribution of each of the frequencies,  $\omega_k$ , to the variance of the data. An example of a periodogram with  $m = 5$  ordinates is given in Figure 2-7.

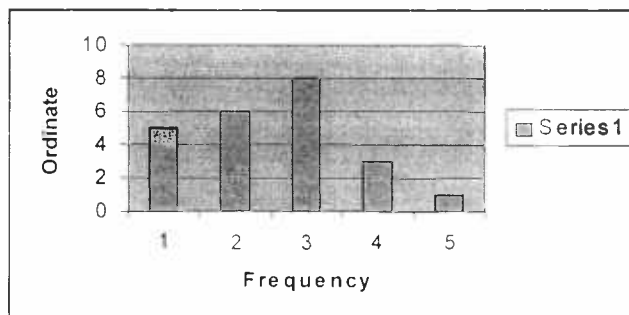


Figure 2-7: Example of a Periodogram with five ordinates.





The periodogram can be estimated from the original observations  $X_1, X_2, \dots, X_n$ . The  $k$ th periodogram ordinate (the height of the  $k$ th histogram rectangle) is calculated as:

$$I(\omega_k) = 1/n\pi \left\{ \left[ \sum_{t=1}^n X_t \cos \omega_k t \right]^2 + \left[ \sum_{t=1}^n X_t \sin \omega_k t \right]^2 \right\}, \quad k = 1, 2, \dots, m \quad (2-22)$$

In order to search for cycles in the original observations, we can test the hypothesis  $H_0: X_t = \mu + \epsilon_t$  versus  $H_1: X_t = \mu + A \cos \omega t + B \sin \omega t + \epsilon_t$ , where  $\mu$ ,  $\omega$ ,  $A$  and  $B$  are unknown constants and  $\epsilon_t$  are independent and identically distributed random variables with mean 0 and standard deviation  $\sigma$ . A statistic that can be used to test the hypothesis is (Fuller, 1976):

$$\xi = \frac{I_L}{\left( \frac{1}{m} \sum_{k=1}^m I(\omega_k) \right)}, \quad \text{where } I_L \text{ is the largest among the } m \text{ periodogram}$$

ordinates, each of which is distributed as a  $X^2$ -distribution with 2 degrees of freedom. (2-23)

The derivation of the distribution of  $\xi$  is provided by Beneke et al. (1988) and a table of the percentage points of the largest ordinate to the average (i.e.,  $\xi$ ) is given by Fuller (1976).

The spectral control chart that detects the presence of cyclic behavior of the process mean is based on a test of the hypothesis mentioned previously. The null hypothesis  $H_0$  (that is, cycles are not present) is rejected if the periodogram ratio,  $\xi$ , is larger than the critical value for the desired significance level.

The spectral control chart consists of an upper limit only, which is the critical value for the desired significance level. The value plotted at each time point is the ratio of the largest periodogram ordinate to the average of all ordinates. An out-of-control signal is given when the value plotted falls above the control limit line. Figure 2-8 illustrates the general form of the spectral control chart. When a value exceeds the upper control limit, the frequency  $\omega_j$  and period  $2\pi/\omega_j$  corresponding to the largest periodogram ordinate are



identified. The first value to be plotted is computed using the first  $n$  observations ( $X_1, X_2, \dots, X_n$ ) and the procedure described previously. To obtain the next plotted value,  $X_1$  is dropped and  $X_{n+1}$  is added, and so forth.

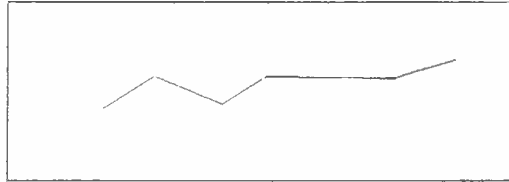


Figure 2-8: A spectral control chart with frequency  $= \omega_k$  and period  $= n/k$ .

Several comments concerning the spectral control chart made by Beneke et al. (1988) are:

(i) Since adjacent values use  $n-1$  of the same process observations in computing the periodogram, there is a significant positive autocorrelation on the spectral control chart. An out-of-control signal does not produce a sudden jump upward as in the Shewhart chart when there is a change in the process average. Rather, the values plotted on the spectral control chart rise slowly as a cyclic tendency becomes more pronounced over time.

(ii) Computations can be expedited in two different ways. Since the process observation  $X_{t-n}$  is dropped and the observation  $X_t$  is added to spectral chart computations, considerable computation time may be saved by storing values from one point in time to the next. If  $n$  is large, a fast Fourier transform may be more efficient.

(iii) The formula for  $I(\omega_k)$  assumes that  $X_t$  are observed at equally spaced time intervals. If the observations are at irregular time intervals or have some consecutive data points missing, this procedure cannot be used. However, if only a few observations are missing and they are not adjacent, interpolation between the known points surrounding the missing ones could solve the problem.

(iv) The spectral control chart works best when detecting cyclic variations that follow a sinusoidal form.

(v) The choices of sampling frequency and the number of observations are related in determining the effectiveness of the spectral chart.

Beneke et al. (1988) showed that the spectral control chart is superior for detecting cyclic variations but it is not effective for detecting shifts in the process mean. On the other hand, the Shewhart and EWMA charts perform poorly when the mean appears to have a cyclic behavior, because their limits are based on the variability of the data, which, apart from the random variability of the process, it also includes the variability caused by the cycle in the process mean. Therefore, the spectral control chart should be used along with the standard control charts, so that both shifts and cycles in the process mean can be detected.

# CHAPTER 3

## Time series models for autocorrelated data

### 3-1 Introduction

In Chapter 2, the standard control charts were presented assuming that the processes of interest are not submitted to a specific pattern and, thus, insinuating that there is no dependence between successive observations over time. However, the existence of correlation among the data is a situation which is very often confronted in practice, since the same machines are used in the fabrication of goods and samples are taken once after a short time interval. The best way to be released from this correlation structure is simply to estimate it and subtract it from the observed data, so as to be left with the uncorrelated structure only.

The estimation of the autocorrelated processes is achieved via the time series models. In section 3-2 the fundamentals of the time series approach are presented, section 3-3 explains the structure of the most widespread stationary time series models, called the ARMA models, and section 3-4 describes the class of nonstationary models, termed as ARIMA processes.

### 3-2 Basic properties of autocorrelated data

Traditional statistical process control (SPC) assumes that consecutive observations from a process are normally and independently distributed with mean  $\mu$  and standard deviation  $\sigma$ . When this assumption is valid, the statistical properties of the control chart (the false alarm rate, the ARL etc.) can be easily determined. If the assumed uncorrelated structure is not valid, then an effort is made to estimate the form of the dependency by trying to find a model that *fits* the data. The first thing to do is calculate the mean, the

variance and the correlation of the observed data, so that an idea is formed about the structure of the process.

### 3-2.1 Autocorrelated data in industry

Often in industrial practice, in continuous as well as discrete production processes, observations are actually not independent but more or less correlated. Autocorrelated behavior means that there are carryover effects from earlier observations. The mechanism of these carryover effects must be sought. Examples include chemical processes where consecutive measurements on product characteristics are interrelated or when the process is organized in batches. Under such conditions, traditional SPC procedures may be ineffective, indeed inappropriate, for monitoring, controlling, and improving process quality. The main difficulty is that when systematic nonrandom patterns are present, casual inspection makes it hard to separate special causes and common causes. A natural solution to this difficulty is to model systematic nonrandom patterns by time-series models that go beyond the simple benchmark of independent and identically distributed (iid) random variables.

One possibility, for example, is a first-order autoregressive model, in which each observation may be regarded as having arisen from a regression model for which the current observation on the process is the dependent variable and the previous observation is the independent variable. If an autoregressive time-series model fits this data set, leaving only residuals that are consistent with randomness, it is futile to search for departures from statistical control since only special causes are now left. Otherwise, these departures will be confounded with the dominant autoregressive behavior of the data. Hence, when the data suggest lack of statistical control, one should attempt to model systematic nonrandom behavior by time-series models- autoregressive or other- before searching for special causes.

Figure 3-1 reveals the difference between an independent (a) and a correlated process (b). In (a) nothing can be said about the next possible value of the process, in (b) however there is strong evidence that the following value will be positive. The existence of autocorrelation may be a good thing



in that it helps predicting the next value. Another aspect with which we could consider the two processes is that a model was fitted adequately to the correlated process of Figure 3-1(b), such that when the model predicted values were subtracted from the data set, what was left was the uncorrelated process of Figure 3-1(a). This procedure of model fitting is similar to the simple regression procedure, where a line is fitted to the data, with the difference that the autocorrelated case is a much more complicated procedure.

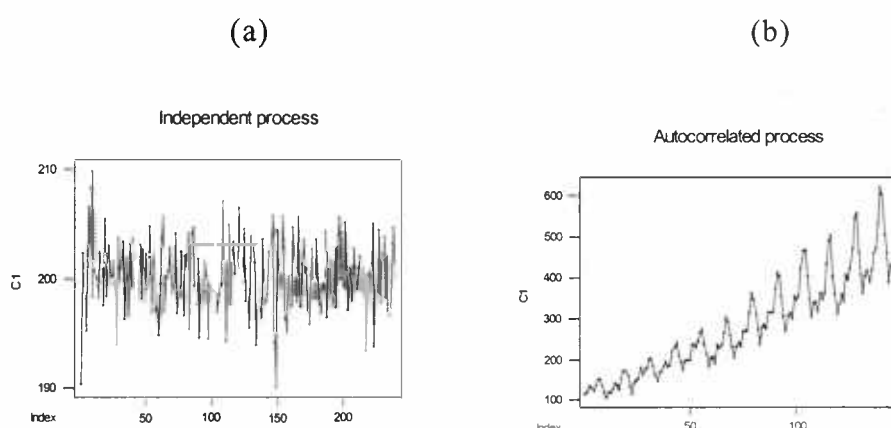


Figure 3-1: Independent (a) versus autocorrelated process (b).

### 3-2.2 The autocovariance and autocorrelation functions

The autocorrelation over a series of time-oriented observations is measured by the **autocorrelation function (ACF)**:

$$\rho_k = \text{Cov}(x_t, x_{t-k}) / V(x_t), \quad k = 0, 1, \dots, \quad \text{where}$$

$\text{Cov}(x_t, x_{t-k})$  is the covariance (ACVF) of observations that are  $k$  time periods apart, i.e.,  $\text{Cov}(x_t, x_{t-k}) = E\{(x_t - \mu_x)(x_{t-k} - \mu_x)\}$

(3-1)

Note that if  $k = 0$ ,  $\text{ACVF} = V(x_t)$

Both  $\mu_x$  and  $V(x_t)$  are assumed to be constant, that is the observations are spread around a fixed value  $\mu_x$ , they have all a constant variance  $V(x_t)$  and the covariance depends only on the lag between the two time periods, i.e.,  $\text{Cov}(x_t, x_{t-k}) = \text{Cov}(x_t, x_{t+k}) = \gamma(k)$ . This is the definition of a **stationary time-series**. The *sample mean* ( $\bar{x}$ ), the *sample autocorrelation function*  $\hat{\rho}(k)$  and

the *sample autocovariance function*  $\hat{\gamma}(k)$  for a set  $x_1, \dots, x_n$  of observations of a *stationary time series* are the sample analogues of those for the mean, autocovariance and autocorrelation functions shown in Eq(3-1). The  $\bar{x}$ ,  $\hat{\gamma}(k)$ ,  $\hat{\rho}(k)$  and the sample *Partial Autocorrelation Function*  $\hat{\alpha}$  equal:

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{t=1}^n x_t \\ \hat{\gamma}(k) &= \frac{\sum_{t=1}^{n-|k|} (x_{t-|k|} - \bar{x})(x_t - \bar{x})}{n}, \quad -n < k < n \\ \hat{\rho}(k) &= \hat{\gamma}(k) / \hat{\gamma}(0), \quad -n < k < n.\end{aligned} \quad (3-2)$$

Usually, we compute values of  $\hat{\rho}(k)$  for  $k \leq n/4$ .

The sample PACF for any set of observations  $x_1, \dots, x_n$  is given by:

$\hat{\alpha}(0)=1$  and  $\hat{\alpha}(k)=\hat{\phi}_{kk}$  for  $k \geq 1$ , where  $\hat{\phi}_{kk}$  is the last component of  $\hat{\phi}_k = \hat{\Gamma}_k^{-1} \hat{\gamma}_k$ ,  $\hat{\Gamma}_k = [\hat{\gamma}(i-j)]_{i,j=1}^k$  is the sample covariance matrix and  $\hat{\gamma}_k = [\hat{\gamma}(1), \dots, \hat{\gamma}(k)]'$ .

### 3-3 The ARMA process

If the autocorrelated data set seems to be stationary, then the most widespread class of models applied to the data is the class of ARMA models.

#### 3-3.1 Definition and properties of the ARMA process

The time series  $\{x_t\}$  is an **ARMA(p,q)** process if it is stationary and if for every  $t$ ,  $X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q}$ , where  $\epsilon_t$  is iid random variable with mean 0 and variance  $\sigma^2$ . (3-3)

Eq(3-3) is valid only when the polynomials  $\phi(z) = (1 - \phi_1 z - \dots - \phi_p z^p)$  and  $\theta(z) = (1 + \theta_1 z + \dots + \theta_q z^q)$  have no common factors and a unique stationary solution exists if and only if  $\phi(z) \neq 0$  for all  $|z|=1$  (see Brockwell and Davis, 1996). Two major properties of the ARMA model are the *causality* and the *invertibility*, which are defined as:

**Causality:**

An ARMA(p,q) process is causal if there exist constants  $\{\psi_j\}$  such that

$$\sum_{j=0}^{\infty} |\psi_j| < \infty \text{ and } \mathbf{X}_t = \sum_{j=0}^{\infty} \psi_j \mathbf{Z}_{t-j} \text{ for all } t. \quad (3-4)$$

Causality is equivalent to the condition  $\phi(z) \neq 0$  for all  $|z| \leq 1$ .

**Invertibility:**

An ARMA(p,q) process is invertible if there exist constants  $\{\pi_j\}$  such that

$$\sum_{j=0}^{\infty} |\pi_j| < \infty \text{ and } \mathbf{Z}_t = \sum_{j=0}^{\infty} \pi_j \mathbf{X}_{t-j} \text{ for all } t. \quad (3-5)$$

Invertibility is equivalent to the condition  $\theta(z) \neq 0$  for all  $|z| \leq 1$ .

**3-3.2 Modeling the ARMA processes**

The determination of an appropriate ARMA(p,q) model to represent an observed stationary time series involves:

1. Estimation of the process mean  $\mu$ .
2. Order selection (the choice of p and q).
3. Estimation of the coefficients ( $\phi_i, i=1, \dots, p$  and  $\theta_j, j=1, \dots, q$ ) and the variance of  $\epsilon_t, \sigma^2$ .

**Estimation of the process mean  $\mu$** 

The mean is estimated by the sample mean given by Eq(3-2), which is not unbiased in this case because of the autocorrelation of the data but which, however, still holds some good properties. It is suggested to subtract the sample mean from the data, so that a zero-mean ARMA model is appropriate to be fitted to the adjusted series.

**Order selection**

As a general rule, choosing p and q arbitrarily large is not advantageous. Fitting a very high order model will generally result in a small  $\sigma^2$ , but when using the fitted model for forecasting, the mean squared error of the forecasts will depend not only on  $\sigma^2$  but also on the errors arising from estimation of the parameters of the model. A criteria helping to decide upon a

good choice for  $p$  and  $q$  is the Akaike (defined below) together with a bias-corrected version of it. After having estimated a number of ARMA models with different values for  $p$  and  $q$ , we select the one with minimum Akaike value.

#### Estimation of the model coefficients and of $\sigma^2$

For fixed values of  $p$  and  $q$ , good estimators for  $\phi$  and  $\theta$  can be found by considering the data to be observations of a stationary Gaussian time series and then maximizing the likelihood with respect to the  $p+q+1$  parameters  $\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q$  and  $\sigma^2$ . Maximization is carried out by searching numerically for the maximum likelihood after specifying initial parameter values with which to begin the search. The closer these are to the maximum likelihood estimates, the faster the search will be.

There exist many methods for preliminary parameter estimation proposed by Brockwell and Davis (1996) as is the 'Yule-Walker Estimation', the 'Burg's algorithm', the 'Innovations Algorithm' and the 'Hannan-Rissanen Algorithm'. After initial values for  $\phi$  and  $\theta$  have been provided by the use of one of the above methods, the maximum likelihood estimation searches for the values of  $\phi$  and  $\theta$  that minimize the reduced likelihood given by:

#### Maximum likelihood estimators

$$\hat{\sigma}^2 = \frac{S(\hat{\phi}, \hat{\theta})}{n}, \text{ where } S(\hat{\phi}, \hat{\theta}) = \sum_{j=1}^n (X_j - \hat{X}_j)^2 / r_{j-1} \text{ and } \hat{\phi}, \hat{\theta} \text{ are the values of } \phi, \theta \text{ that minimize } \ell(\phi, \theta) = \ln(n^{-1}S(\phi, \theta)) + n^{-1} \sum_{j=1}^n \ln r_{j-1} \quad (3-6)$$

Finally, for fixed  $p$  and  $q$ ,  $\phi_p$  and  $\theta_q$  are selected to minimize the bias-corrected Akaike criterion:

#### Akaike criterion

$$AICC = -2\ln L(\phi_p, \theta_q, S(\phi_p, \theta_q)/n) + 2(p+q+1)n/(n-p-q-2) \quad (3-7)$$





By repeating the estimation procedure for various values of  $p$  and  $q$  and calculating the Akaike value using Eq(3-7), we may consider the values for  $p$  and  $q$  that minimize AICC as the optimal ones.

### 3-3.3 Goodness-of-fit of the ARMA process

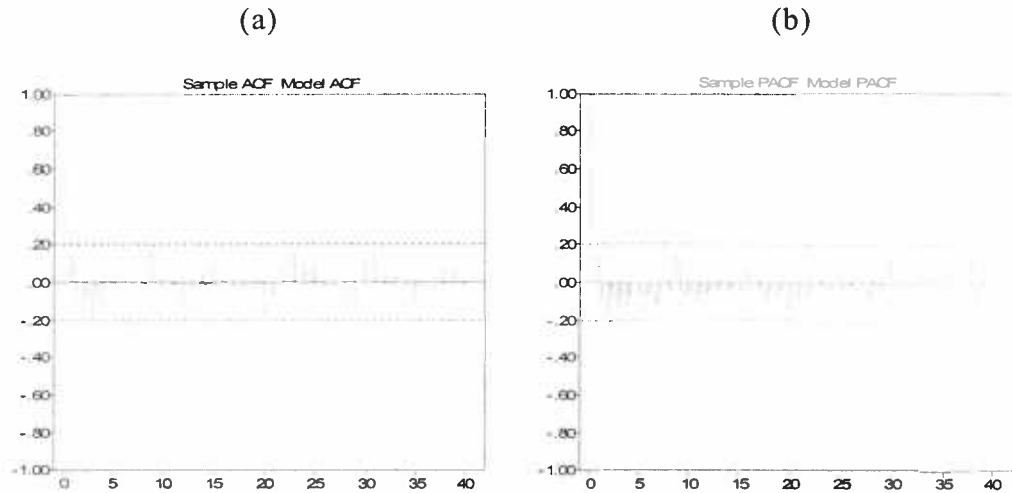
In order to verify the appropriateness of the specified ARMA( $p,q$ ) model, its autocorrelation function is compared with the sample autocorrelation defined in Equation (3-2). Methods for computing the autocorrelation function of causal ARMA processes are given in Brockwell and Davis (1996). Similarly, the sample Partial autocorrelation and the model Partial autocorrelation are compared. If there are not serious deviations between the sample and the model values for all lags, then the particular ARMA( $p,q$ ) model is considered adequate to fit the data. If this is not the case, other models should be tried.

With the purpose of checking the model fit, the most common approach is to calculate the *residuals* (observed values - values estimated by the model). If the model has fitted the data well, the residuals should be left to be *white noise*, that is, a sequence of uncorrelated random variables, each with 0 mean and variance  $\sigma^2$ . The most common way to find if the residuals are white noise is to plot them between the bounds  $\pm 1.96/\sqrt{n}$  [based on the fact that the distribution of an iid sequence is  $N(0,1/n)$ ], in which they will fall with 95% probability. If we compute the autocorrelations of the residuals up to lag 40 and find that more than 2 values (i.e.,  $40 \cdot 0.5$ ) fall outside the bounds, then we should reject the iid hypothesis. Other tests for checking the validity of the hypothesis are the 'Portmanteau test', the 'Turning Point' test, the 'Difference Sign test' and the Rank test' (see Brockwell and Davis, 1996).

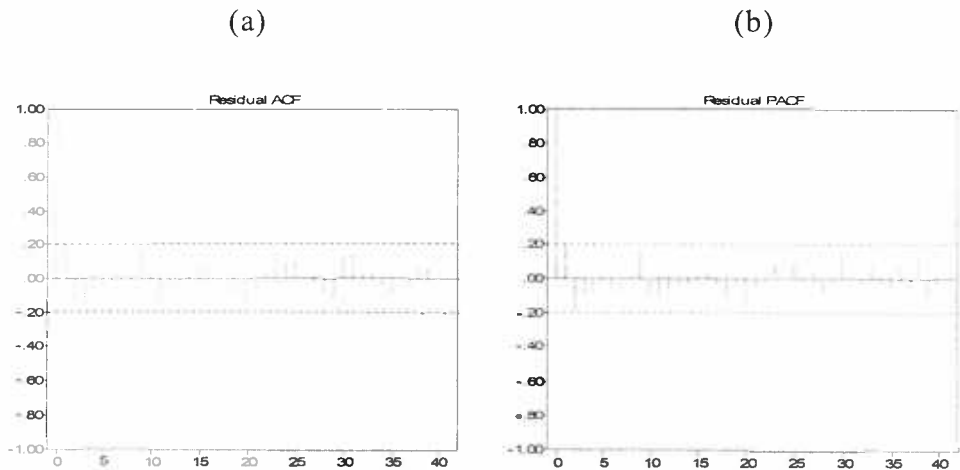
Figure 3-2 presents the ACF/PACF functions where the longer lines indicate the ones estimated by the data and the shorter lines are the ones derived by the Maximum Likelihood estimates for the ARMA(1,1) model:  $X_t = 0.8202X_{t-1} + \epsilon_t - 0.9766\epsilon_{t-1}$ . It is obvious that both the sample ACF and PACF are very close to their model analogues. That is why, both the ACF and PACF of the residuals shown in Figure 3-3 are compatible with the ones coming



from pure white noise, since they are inside the 95% bounds denoted by the dotted lines (apart from when lag = 0 in which case the correlation is always equal to 1).



**Figure 3-2: The sample/model ACF (a) and the sample/model PACF (b) for the process  $X_t = 0.8202X_{t-1} + \epsilon_t - 0.9766\epsilon_{t-1}$ .**



**Figure 3-3: The ACF (a) and PACF (b) of the residuals after the model  $X_t = 0.8202X_{t-1} + \epsilon_t - 0.9766\epsilon_{t-1}$  has been applied.**

### 3-3.4 The most common ARMA models

Usually, the construction of ARMA models with a large number of estimated parameters is avoided due to their complexity. In statistical process control, a few parameters are often sufficient to explain the autocorrelation

structure. The simplest and most popular ARMA charts together with their properties are further presented.

### 3-3.4.1 The MA(q) process

The **MA(q)** is a **Moving Average** process of order  $q$  if the stationary time series  $\{X_t\}$  satisfies the equations:

$$X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q}.$$

where  $\epsilon_t$  is identically and independently distributed random variable with mean 0 and variance  $\sigma^2$  and  $\theta_1, \dots, \theta_q$  are constants with  $\theta_0 = 1$ . (3-8)

In other words, the MA(q) is an ARMA(p,q) process with  $p=0$  and  $q>0$ .

#### a) ACVF and ACF of the MA(q) process

The mean, autocovariance and autocorrelation functions specified by the model MA(q) are:

$$\begin{aligned} E(X_t) &= 0 \\ \gamma(k) &= \begin{cases} \sigma^2 \sum_{j=0}^{q-|k|} \theta_j \theta_{j+|k|}, & \text{if } |k| \leq q, \\ 0, & \text{if } |k| > q \end{cases} \\ \rho(k) &= \gamma(k) / \gamma(0) \end{aligned} \quad (3-9)$$

By setting  $q=1$ , we get the ACVF and ACF of the MA(1) process.

#### b) Order selection

The MA(q) process is said to be  $q$ -correlated because, as seen from Eq (3-9), for  $\text{lag} > q$ , its autocorrelation function becomes 0. Since the inverse is also true, i.e., a stationary  $q$ -correlated time series with mean 0 can be represented as the MA(q) process, a good way to estimate  $q$  is to represent graphically the sample autocorrelation function and specify  $q$  to the value above which the sample autocorrelation becomes 0. More precisely, if the sample ACF of the data is significantly different from 0 for  $k < q$ , i.e.,

$|\hat{\rho}(k)| > 1.96/\sqrt{n}$  for  $k < q$  and it is negligible for  $k > q$  ( $|\hat{\rho}(k)| < 1.96/\sqrt{n}$ ), then a MA(q) model is suggested.

c) *Estimation of the model coefficients*

The preliminary estimation methods used for the MA(q) model are either the “Innovations” or the “Hannan-Rissanen” algorithms. After using the Maximum Likelihood estimation based on the initial values of  $\theta_1, \dots, \theta_q$  estimated by the one of the above preliminary methods, the AICC value should be calculated using Eq(3-7) and the q value resulting in the minimum AICC will be the one retained.

3-3.4.2 The AR(p) process

The time series  $\{X_t\}$  is an **Autoregressive** process of order p, **AR(p)**, if the stationary time series  $\{X_t\}$  satisfies the equations:

$$X_t = \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} + \epsilon_t,$$

where  $\epsilon_t$  is identically and independently distributed random variable with mean 0 and variance  $\sigma^2$ ,  $\epsilon_t$  is uncorrelated with  $X_s$  for  $s < t$  and  $\phi_1, \dots, \phi_p$  are constants. (3-10)

The AR(p) is an ARMA(p,q) process with  $p > 0$  and  $q = 0$ .

a) *ACVF and ACF of the AR(1) process*

There is not a general closed formula for the ACVF and ACF of the AR(p) process. Therefore, only the autocovariance and autocorrelation functions of the AR(1) process will be provided because this is the most common case among all the AR(p) schemes that are used in industry. The mean, ACVF and ACF functions of the AR(1) model are calculated as:

$$E(X_t) = 0$$

$$\gamma(k) = \phi^{|k|} \gamma(0), \text{ where } k = 0, \pm 1, \dots \quad (3-11)$$

$$\rho(k) = \phi^{|k|}, k = 0, \pm 1, \dots$$

*b) Order selection*

The PACF of an AR(p) process has the properties  $\alpha(p)=\phi_p$  and  $\alpha(k)=0$  for  $k>p$ . Therefore, if the sample PACF of a set of  $n$  observations is significantly different from 0 for  $0 \leq k \leq p$  and negligible for  $k>p$ , then an AR(p) model might provide a good representation for the data. To decide what is meant by negligible, we can use the fact that for an AR(p) process the sample PACF values at lags greater than  $p$  are approximately independent  $N(0,1/n)$  random variables, so that if  $|\hat{\alpha}(k)| > 1.96/\sqrt{n}$  for  $0 \leq k \leq p$  and  $|\hat{\alpha}(k)| < 1.96/\sqrt{n}$  for  $k>p$ , an AR(p) model is suggested for the data.

*c) Estimation of the model coefficients*

For an AR(p) model, the preliminary estimation method should be one of the ‘Yule-Walker’ or the ‘Durbin-Levinson’ algorithms. Then, the Maximum Likelihood estimation is applied for various values of  $p$  and the value of  $p$  minimizing the AICC criterion given by Eq(3-7) is the one we finally decide to keep.

### 3-3.4.3 The ARMA(1,1) process

The time series  $\{X_t\}$  is a first-order Autoregressive Moving-Average process, **ARMA(1,1)**, if it is stationary and satisfies for every  $t$  the equation:

$$X_t - \phi X_{t-1} = \epsilon_t + \theta \epsilon_{t-1},$$

where  $\epsilon_t$  is identically and independently distributed random variable with mean 0 and variance  $\sigma^2$  and  $\phi \neq \theta$ . (3-12)

The ARMA(1,1) model contains simultaneously the information provided from both the AR(1) and MA(1) models.

*a) ACVF and ACF of the ARMA(1,1) process*

The autocovariance and autocorrelation functions of the ARMA(1,1) model are given by:

$$\begin{aligned}
 &E(X_t) = 0 \\
 &\gamma(k) = \begin{cases} \sigma^2 \left[ 1 + \frac{(\theta + \varphi)^2}{1 - \varphi^2} \right], & k = 0 \\ \sigma^2 \left[ \theta + \varphi + \frac{(\theta + \varphi)^2 \varphi}{1 - \varphi^2} \right], & k = 1 \\ \varphi^{k-1} \gamma(1), & k \geq 2 \end{cases} \\
 &\rho(k) = \gamma(k) / \gamma(0)
 \end{aligned} \tag{3-13}$$

The variance of an ARMA(1,1) model is (Box and Jenkins, 1976):

$$\begin{aligned}
 &\sigma_x^2 = (1 - 2\phi_1\theta_1 + \theta_1^2) \sigma^2 / (1 - \phi_1^2), \\
 &\text{where } \sigma \text{ is the standard deviation of the random error terms } \epsilon_t.
 \end{aligned} \tag{3-14}$$

*b) Order selection for the general ARMA(p,q) model*

For models with  $p > 0$  and  $q > 0$ , the sample ACF and PACF are difficult to distinguish. However, an approach would be to combine the methods used for the separate MA(q) and AR(p) processes, so that p and q are selected to be the values above which the PACF and the ACF values respectively become 0.

*c) Estimation of the model coefficients*

The preliminary estimation methods used for the ARMA(p,q) model and, consequently, for the ARMA(1,1) model are either the “Innovations” or the “Hannan-Rissanen” algorithms. After using the Maximum Likelihood estimation, the minimization of the AICC value will indicate the values for p and q to use. The parameter estimation and the choice of the ARMA model to be fitted are provided by user friendly packages.

### 3-3.5 Forecasting stationary time series

In time series applications, we often consider the problem of predicting future values of the process,  $X_{n+h}$  when  $h > 0$ , with known mean  $\mu$  and autocovariance  $\gamma$  in terms of the past values  $\{X_n, \dots, X_1\}$  up to time n. The goal is to find the *linear combination* of  $\{X_n, \dots, X_1\}$  which forecasts  $X_{n+h}$  with

minimum squared error. Thus, the linear predictor of  $X_{n+h}$  based on  $\{X_n, \dots, X_1\}$  has the form:  $\hat{X}_{n+h} = c_0 + c_1 X_n + \dots + c_n X_1$ . It is known from the statistical theory that the minimum-squared error criterion finds the optimal value of  $\hat{X}_{n+h}$  by minimizing the quantity:

$$E(X_{n+h} - \hat{X}_{n+h})^2 = E[X_{n+h} - (c_0 + c_1 X_n + \dots + c_n X_1)]^2 \quad (3-15)$$

There are two recursive approaches of determining the best linear predictor. These are the “Durbin-Levinson” and the “Innovations” algorithm. Both approaches are described extensively in Brockwell and Davis (1996).

### 3-4 The ARIMA process

In the previous section, the stationary ARMA models have been studied. However, we should also examine the case in which a set of observations  $\{x_1, \dots, x_n\}$  does not seem to be generated by a stationary time series. If the data do not exhibit apparent deviations from stationarity or, in other words, the autocovariance function of the data is rapidly decreasing, then an ARMA model should be fitted to the mean-corrected data. Otherwise, we confront an ARIMA model and a transformation should be applied to the data in order to make them stationary.

#### 3-4.1 Definition of ARIMA models

A wide range of nonstationary time series may be fitted by Autoregressive Integrated Moving Average (ARIMA) processes, that is, processes which, after differencing finitely many times, reduce to ARMA models. If  $d$  is a nonnegative integer, then  $\{X_t\}$  is an **ARIMA(p,d,q)** model if:

$$Y_t = (1-B)^d X_t \text{ is a causal ARMA}(p,q) \text{ process,}$$

where  $(1-B)^d \equiv \nabla^d$  is the lag- $d$  difference operator and  $B$  is the backward shift operator, such that  $BX_t = X_{t-1}$ . (3-16)



For example, the **lag-one** difference operator  $\nabla$  is:  $\nabla X_t = (1-B)X_t = X_t - X_{t-1}$ , the **lag-two** difference operator  $\nabla^2$  is:  $\nabla(\nabla(X_t)) = (1-B)^2 X_t = (1-2B+B^2)X_t = X_t - 2X_{t-1} + X_{t-2}$ , etc. An equivalent definition of the one given by Eq(3-16) is that  $X_t$  satisfies the equation of the form:

$$\phi^*(B)X_t = \theta(B)\epsilon_t, \text{ with } \phi^*(B)X_t = \phi(B)(1-B)^d X_t, \\ \text{where } \epsilon_t \text{ is white noise with 0 mean and variance } \sigma^2 \text{ and } \phi(z), \theta(z) \text{ are} \\ \text{polynomials of degrees } p \text{ and } q, \text{ respectively, while } \phi(z) \neq 0 \text{ for } |z| \leq 1. \quad (3-17)$$

The ARIMA(p,d,q) process reduces to an ARMA(p,q) process if  $d=0$ .

#### The IMA model

A special case of the ARIMA(p,d,q) model is the Integrated Moving Average model (IMA) which has no autoregressive term (i.e.  $p=0$ ). The first-order IMA (that is, when  $d=1$ ) is specified as:

$$\nabla X_t = \epsilon_t + \theta \sum_{i=0}^{t-1} \epsilon_i \text{ or } X_t = X_{t-1} + \epsilon_t + \theta \sum_{i=0}^{t-1} \epsilon_i \quad (3-18)$$

where  $0 \leq \theta \leq 1$ . An IMA with  $\theta \neq 0$  is a nonstationary process with variance  $\sigma^2(1 + \theta^2 t)$  increasing linearly in  $t$ . Special cases of the IMA family arise when  $\theta = 0$  giving an iid process and when  $\theta=1$  giving a random walk. For  $0 < \theta < 1$ , the IMA is equivalent to a random walk observed with iid measurement error (Box and Jenkins, 1976).

The IMA(1,1) or, equivalently, ARIMA(0,1,1) model has only one  $\theta$  parameter to be estimated, i.e:

$$X_t = X_{t-1} + \epsilon_t - \theta \epsilon_{t-1} \quad (3-19)$$

The model of Eq(3-19) describes nonstationary behavior (that is, the variable  $X_t$  drifts as if there is no fixed value of the process mean). This model often arises in chemical and process plants when  $X_t$  is an uncontrolled process output, so that no control action has been taken to keep the variable to a target value.



### 3-4.2 Transformation of the ARIMA models

Deviations from stationarity are usually suggested by the graph of the time series itself or by the sample autocorrelation function. Three kinds of transformation are commonly applied to the nonstationary time series in order to eliminate the specific cause of nonstationarity. These are:

#### a) *Unstable variability*

If inspection of the graph reveals a strong dependence of variability on the level of the series, then a Box-Cox transformation can be used in order to *stabilize variability*. This transformation proposes a new time series  $\{Y_t\}$  with stabilized variance where:

$$Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, \lambda \neq 0 \\ \ln X_t, \lambda = 0 \end{cases} \quad (3-20)$$

The logarithmic transformation for  $\lambda = 0$  is appropriate when the standard deviation of the series increases linearly with the mean.

#### b) *Trend and seasonality*

If the graph of the time series reveals the existence of trend (the mean of the process is systematically increasing or decreasing) and seasonality (a pattern is repeated every fixed time period), then there are two approaches to the problem:

1. *Classical decomposition* of the series into a trend component, a seasonal component and a random residual component. Then the trend is estimated by applying a smoothing filter (exponential or moving average), the seasonal component is estimated next and, finally, the trend is reestimated by applying a filter or fitting a polynomial (linear or quadratic). All these estimation methods are described in detail in Brockwell and Davis (1996). At last, the estimated trend and seasonal component are subtracted from the model and a stationary time series is left.

2. *Differencing*. If the trend can be expressed in terms of a polynomial of degree  $k$ , then the application of the operator  $\nabla^{k+1}$  to the data eliminates the trend and results in a time series with constant 0 mean. The seasonal component with period  $d$  can be eliminated by applying the  $\nabla_d$  operator defined by  $\nabla_d X_t = (1-B^d)X_t = X_t - X_{t-d}$ . It is preferable to eliminate the seasonal component first and the trend afterwards so as to be left with the residual term which is a stationary time series. Figure 3-4(a) shows a time series with a seasonal component of period 12 and with variability increasing linearly with the mean, while in (b) the logarithmic transformation and the application of the operator  $\nabla_{12}$  has made the series stationary. The fit of an ARMA model can now be suggested as it was done in the previous section.

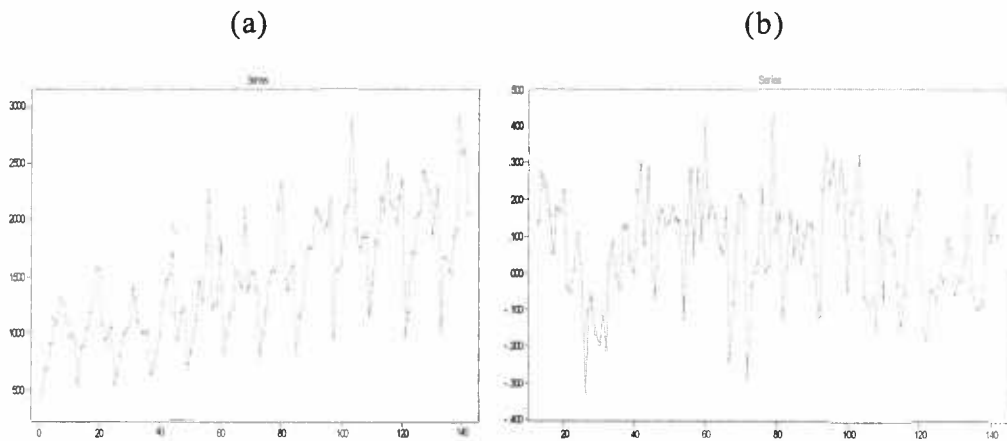


Figure 3-4: An ARIMA model with  $d=12$  (a) transformed in an ARMA model (b).

## CHAPTER 4

### Control charts for autocorrelated processes

#### 4-1 Introduction

In the previous chapter we were concentrated in modeling autocorrelated processes of the form of ARIMA models. By referring to ARIMA models we implicitly assume that the ARMA models are also included, since they form a subcategory of the ARIMA ones. Once an appropriate ARIMA process has been specified for the observed data leaving a pure series of independent stationary residuals with nonrandom pattern, then the control charts are able to test whether the production process is in control, i.e., no shift in the mean or in the standard deviation of the process has been occurred, or not.

A variety of control charts for autocorrelated processes has been proposed. These include traditional charts with modified control limits described in section 4-2, or charts applied to the residuals, as presented in section 4-3. Section 4-4 recommends the use of forecasting tools with traditional control charts applied to the forecast errors. Finally, in section 4-5 a recent chart called the ARMA chart is initiated.

#### 4-2 Traditional charts modified for autocorrelated processes

<sup>1</sup> Since our concern is on processes following a correlated pattern, the lack of independence among the observations should be taken into account when monitoring these processes. A lot of work has been done in an effort to widen the control limits of the standard control charts, so that the number of false alarms due to the inherent patterned structure of the data is reduced.



#### 4-2.1 The modified Shewhart Control Chart

In Chapter 2 we initiated the standard mode of the Shewhart control limits as:

$\begin{aligned} \text{UCL} &= \mu_x + L \sigma_x \\ \text{Center line} &= \mu_x \\ \text{LCL} &= \mu_x - L \sigma_x \end{aligned} \tag{4-1}$
---

We have already discussed that the Shewhart chart is a plot of the observations themselves and it is designed so as to have a small chance of obtaining an out-of-control signal when the process is in control, and a higher chance of an out-of-control signal when the process is out of control. Assuming that only one observation is provided at each time period, that is the Shewhart chart is an Individuals chart, it may account for the autocorrelation of the data if  $\sigma_x$  is not the pure standard deviation of the process any more, but it now depends on the particular correlation structure (i.e., on the parameters of the appropriate ARIMA model). For example, if the fitted model is ARMA(1,1), having the form of Eq(3-12), its variance has been specified from Eq (3-14) as:

$$\sigma_x^2 = (1 - 2\phi_1\theta_1 + \theta_1^2) \sigma_\epsilon^2 / (1 - \phi_1^2),$$

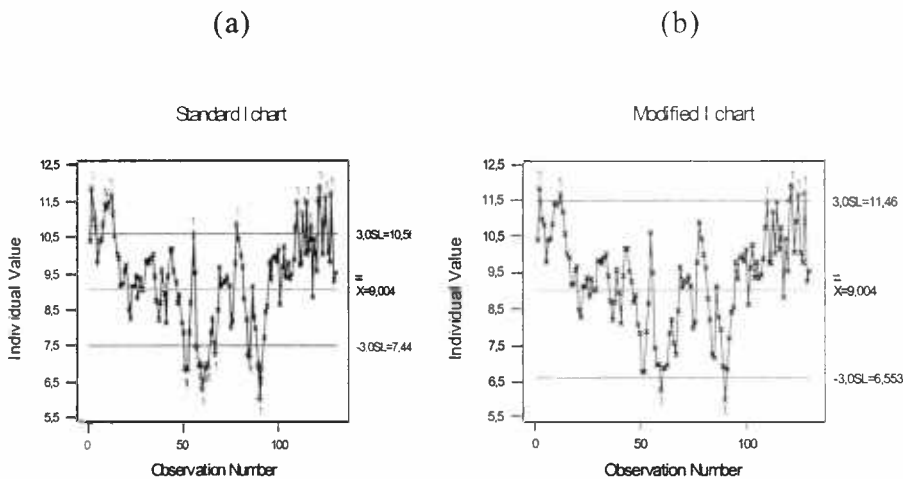
with  $\sigma_\epsilon^2$  being the variance of the random error term (symbolized as  $\sigma^2$  in the previous Chapter). By substituting both the squared error of the variance given from Eq (3-14) and the mean of the process (known or estimated from the data) to Eq (4-1) and by setting  $L$  to a fixed value (usually 3), we have constructed the control limits for the *modified Shewhart Control Chart*. If more than one observation is available at each sample, then the standard deviation  $\sigma_x$  can be substituted by  $\sigma_x/\sqrt{n}$ .

The modified Shewhart Control Chart has been proposed by Wardell, Moskowitz and Plante (1992). It is an extension of the original Shewhart chart since for independent processes, that is, when  $\phi_i$  and  $\theta_j$  are equal to 0,  $\sigma_x \rightarrow \sigma_\epsilon$  and the standard limits of the Shewhart chart are derived.

Figure 4.1(a) presents the original Shewhart chart applied to a set of autocorrelated data with mean 9 for the first 98 points and 10 for the last 32

Because we did not take into account the dependency of the process, run counts suggest that the process is nearly always out of control, sometimes on the high side and sometimes on the low side, so that the shift in the mean is confused with the autocorrelation of the data. Note that the control limits were calculated from the first 98 points because these were considered as the steady state of the process.

Figure 4.1(b) presents the modified Shewhart chart in which we have calculated  $\sigma_x^2 = 0.66$  using Eq(3-14) after fitting adequately an ARMA(1,1) model with  $\phi = 0.74$ ,  $\theta = 0.32$  and  $\sigma_\epsilon^2 = 0.475$ . This chart has much fewer out-of-control points since its limits are wider but it has still confused the common and special causes to some degree by not clearly distinguishing the shift in the mean from the inherent autocorrelation structure.



**Figure 4-1: The standard (a) and the modified (b) Shewhart control charts for an ARMA(1,1) model.**

Obviously, if another ARMA model apart from the ARMA(1,1) is appropriate for the process, then its estimated standard deviation is the one replaced in Eq(4-1). Estimates for the variance of various ARMA models can be found in Box and Jenkins (1976).

#### 4-2.2 The EWMAST chart

The EWMA chart has been proven satisfactory in some cases of autocorrelated processes even by not taking into account the correlation of the

data. However, Zhang (1998) proposed the EWMAST chart to improve the performance of the simple EWMA when the data are related. Its limits are different from the ones of the original EWMA chart being wider when the process is *positively* autocorrelated. The plotted observations are not the original ones,  $x_t$ , but the recursive ones,  $z_t$ , exactly as in the case of the simple EWMA, i.e.,  $z_t = \lambda x_t + (1-\lambda)z_{t-1}$ . What is changed compared to the EWMA chart is just the control limits. That is, the values  $z_t$  are again plotted on a chart with centerline  $\mu$  and  $L\sigma$  limits of the form  $\mu \pm L\sigma_z$  with the only difference being the value of  $\sigma_z$ , which is now calculated as:

$$\sigma_z^2 = \lambda(2-\lambda) \sigma_x^2 \times \left\{ 1 - (1-\lambda)^{2t} + 2 \sum_{k=1}^{t-1} \rho(k) (1-\lambda)^k \times [1 - (1-\lambda)^{2(t-k)}] \right\} \quad (4-2)$$

Assuming no change of autocorrelation in the series  $\{X_t\}$ , the EWMAST chart will signal changes of the process mean. It is not difficult to see that when the data come from an iid process, that is,  $\rho(k) = 0$  when  $k \geq 1$ , Eq (4-2) becomes  $\sigma_z^2 = \lambda(2-\lambda) \sigma_x^2 [1 - (1-\lambda)^{2t}]$ , which is the variance of the original EWMA chart as it was presented in Eq(2-19b). Thus, the ordinary EWMA chart is a special case of the EWMAST when  $\{X_t\}$  forms an iid sequence.

Zhang (1998) also proved that when  $t$  is large, there exists an integer  $M$  so that an approximate variance of  $z_t$  asymptotically is:

$$\sigma_z^2 \approx \lambda(2-\lambda) \sigma_x^2 \times \left\{ 1 + 2 \sum_{k=1}^M \rho(k) (1-\lambda)^k \times [1 - (1-\lambda)^{2(M-k)}] \right\} \quad (4-3)$$

When the process is iid, Eq (4-3) becomes:  $\sigma_z^2 \approx \lambda(2-\lambda) \sigma_x^2$ , being exactly the asymptotic variance of the EWMA chart also shown in Eq(2-19b). Because  $|\rho(k)| < 1$  for  $|k| > 0$ , the approximation of Eq(4-3) is very good even for a fairly strongly autocorrelated process. Zhang suggested, after having conducted simulation studies, to use  $M=25$  when  $\lambda \geq 0.2$ , because  $M$  should be large enough in order to avoid large estimation errors of the autocorrelations.

We remind that in practice,  $\mu$  and  $\sigma_z$  are estimated based on some historical data of  $\{X_t\}$  when the process is under control, so  $\mu$  is replaced by the sample mean and  $\sigma_x^2$  and  $\rho(k)$  in Eq (4-2) by their sample estimates specified by Eq(3-2).

Zhang (1998) summarizes the implementation of the EWMAST chart in the following steps:

1. Determine a period with  $N$  ( $\geq 100$ ) observations when the process is in a stable condition. Calculate the sample process mean,  $\bar{x}$ ,  $\hat{\sigma}_x$ , using the sample process standard deviation and, finally, calculate the sample autocorrelations  $\hat{\rho}(k)$  for  $k=1, \dots, 25$  of all observations to that point.

2. Calculate the approximate EWMAST standard deviation  $\hat{\sigma}_z$  from Eq(4-3) with an appropriate  $\lambda$  (usually equal to 0.2) and  $M=25$ .

3. The EWMAST chart is constructed by charting the values  $z_t$  calculated from Eq(2-19) as in the EWMA chart with centerline at  $\bar{x}$  and limits at  $\bar{x} \pm L \hat{\sigma}_z$ .

4. Once the EWMAST chart gives a signal indicating that the process is out of control, the process mean needs to be reestimated when the process is stable again and the centerline of the chart needs to be adjusted to the new level. The process variance and autocorrelations also need to be checked if they need any adjustment or not. The most practical thing to do is to update the autocorrelations and the process variance at regular intervals and do the same thing for the centerline if it is not fixed to a specified target.

Figure 4-2 illustrates an EWMA and an EWMAST chart for the autocorrelated data also used in Figure 4-1. The variance of the EWMAST chart was calculated by its asymptotic form of Eq(4-3), by setting  $\lambda$  equal to 0.2 and  $M=25$ . Once again, the limits of these charts have been calculated from the first 98 values, which are the ones being in statistical control.

The EWMA chart (with  $\lambda=0.2$ ) has many out-of-control points caused by the autocorrelation of the data so that the shift in the mean does not reveal an unusual situation. On the other hand, the EWMAST has much wider limits, and the mean shift seems to be detected through the observation of an unusual upward trend.





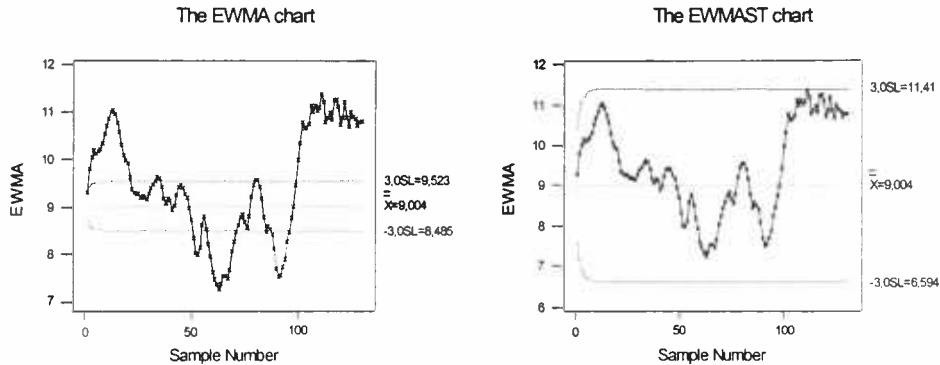


Figure 4-2: The EWMA and the EWMAST charts both applied to autocorrelated data.

### 4-3 Traditional control charts applied to the residuals

Another way to monitor the related observations is to assume that the residuals, left after the estimation of the correlation structure, are uncorrelated. This permits us to apply the standard control charts with no objections any more concerning the process structure.

#### 4-3.1 The Common Cause and Special Cause Charts

The idea of plotting the residuals of the correlated data on a control chart after having fitted the appropriate ARIMA model, instead of charting the original data, has been proposed by Alwan and Roberts (1988). The authors summarized this procedure in two steps by initiating the *Common-cause and Special-cause charts*.

##### 4-3.1.1 The Common Cause Chart (CCC)

The Common Cause Chart (CCC) simply plots the fitted values that are determined by fitting the correlated process by an ARIMA model. This chart assumes that no special causes have occurred. Strictly speaking, it is not a control chart because it has no limits with its intention being of just giving a representation of the current and estimated or predicted state of the process. It provides guidance in seeking better understanding of the process and in achieving real-time process control by giving a view of the level of the



process and of the evolution of that level through time. More precisely, the model of the fitted values can be interpreted as a set of observations which can be thought of as a random disturbance plus a random-walk trend or drift reflecting a certain fraction of the sum of all past random disturbances. Thus, a part of each disturbance continues to affect the process in the future.

The fitted values are estimates of the underlying random-walk trend i.e., they follow a random-walk without drift. The common-cause chart essentially accounts for the systematic variation in the process. In most situations in practice where SPC data are correlated, the systematic variation in the data is much larger, and, thus, more important with respect to influencing product quality, than are special-cause effects. After as much of the systematic variation is removed as possible, a special-cause chart can then be used to establish process capability and to monitor process quality.

Wardell, Moskowitz and Plante (1992) thought that often the forecasts themselves can be used to signal an out-of-control condition before the residuals indicate the change has occurred. For that reason they derived limits for the CCC chart but only for the specific case where the process is described by an ARMA(1,1) model. The plotted values of the CCC chart are the forecasts. The one-step ahead forecast minimizing the mean squared deviation between the forecast and the observed value for the ARMA(1,1) model is given by (Box and Jenkins, 1976) as:

$$F_{t+1} = (1-\phi_1)\mu + (\phi_1 - \theta_1)X_t + \theta_1 F_t,$$

where  $F_{t+1}$  = the forecast made at time  $t$  for period  $t+1$  and  $\mu$  is fixed. The control limits for the CCC chart ( $CL_F$ ) are of the form:

$$CL_F = \mu \pm L\sigma_F,$$

where  $\mu$  is the mean of the process but also of the forecast since the error terms are independent with mean 0, and the variance of the forecast is:

$$\sigma_F^2 = \text{Var}(F_{t+1}) = \frac{(\phi_1 - \theta_1)^2}{(1 - \phi_1^2)} (1 - 2\phi_1^t \theta_1^t + \theta_1^{2t}) \sigma_\epsilon^2,$$

while for large  $t$  the steady-state variance is:

$$\sigma_F^2 = \frac{(\phi_1 - \theta_1)^2}{(1 - \phi_1^2)} \sigma_\epsilon^2, \text{ with } \sigma_\epsilon^2 \text{ being the variance of the error term.}$$



Figure 4-3 illustrates the fitted values of the ARMA(1,1) model used also for the previous figures with  $\phi = 0.74$  and  $\theta = 0.32$ , plotted on an CCC chart. The steady-state variance was easily calculated by Eq(4-4) as equal to 0.19. However, in our case the CCC chart is not of any help since it has many out-of-control points even before the mean shift has occurred.

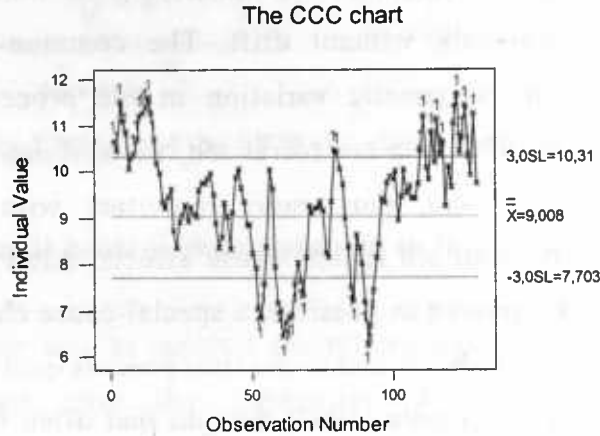


Figure 4-3: The CCC chart for an ARMA(1,1) model with  $\phi = 0.74$  and  $\theta = 0.32$ .

#### 4-3.1.2 The Special Cause Chart (SCC)

This is a traditional Shewhart chart of the residuals (i.e., of the difference between the actual process values and their forecasts) from fitting ARIMA models based on the simple thinking that assignable causes impacting the process should also impact the residuals. The SCC chart can be used in traditional ways to detect any special causes without the danger of confounding special causes with common causes. The residuals are now iid data and, thus, all traditional tools of process control are applicable. Since the mean of the residuals is 0, the centerline of the SCC chart is 0, and the standard deviation used is the standard deviation of the residuals  $\sigma_R^2$ , which must be equal to  $\sigma_\epsilon^2$  if the process is fitted correctly. Thus, the limits of the SCC chart are:

$$UCL_R = L \sigma_\epsilon$$

$$\text{Centerline} = 0$$

$$LCL_R = -L \sigma_\epsilon$$

(4-5)

Some of the major reasons that the Alwan and Roberts (1988) approach is appealing include the following:

- (a) It takes advantage of the fact that the process is correlated to make forecasts of future quality,
- (b) the special-cause chart is based on the assumption that the residuals are random, so all of the assumptions of traditional SPC are met and, hence, any of the traditional tools for SPC can be used, including run rules, cumulative sum (CUSUM) charts, and so forth,
- (c) similarly, the special-cause chart can be used to detect any assignable cause, including changes in the structure of the time series,
- (d) the methodology used to obtain the charts is straightforward and does not require a great deal of sophistication on the part of the user, especially with the availability of user-friendly software packages to fit time series models, and
- (e) unlike other methods for dealing with correlated data that have been limited to AR(1) or MA(1) time series, the method can be applied to any type of time series model.

Figure 4-4(a) shows the SCC chart for the same autocorrelated data as in the previous graphs, in which the ARMA(1,1) model with  $\phi = 0.74$  and  $\theta = 0.32$  has been fitted. The mean shift was not detected, probably because the Shewhart chart does not detect small shifts quickly. On the other hand, Figure 4-4(b), in which the mean for the last 32 observations has shifted to 11 instead of 10 as in Figure 4-4(a), shows that the residuals chart detects the out-of-control state.

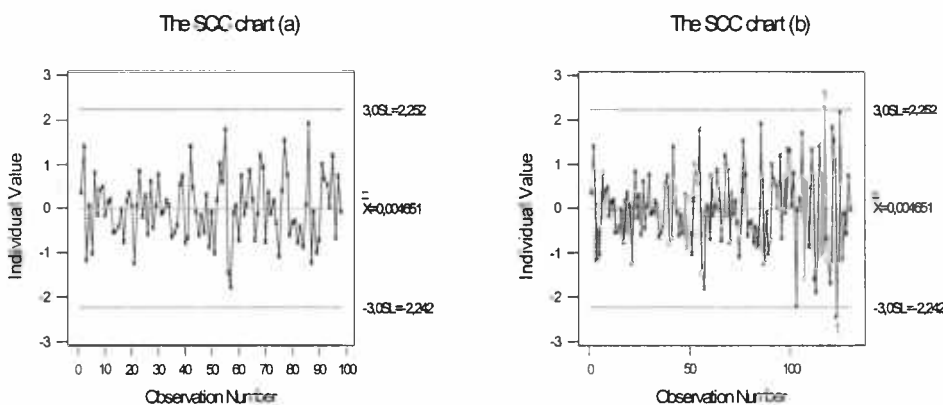


Figure 4-4: The residuals chart for a mean shift of  $1\sigma$  (a) and of  $2\sigma$  (b).

Since any SPC control chart may be used to plot the residuals, apart from the Shewhart chart or the individuals chart, the EWMA and CUSUM may be used as well so as to improve sensitivity to small process shifts. Concerning the CUSUM chart for residuals, Runger, Willemain and Prabhu, (1995) strongly recommend against the standard practice of using  $K = \delta/2$  (where  $\delta$  is the mean shift expressed in terms of the standard deviation) whenever the autocorrelation is reasonably high. They proved that for an AR(1) model with parameter  $\phi$  and if  $\delta = 1$ , a choice of  $K = (1 - \phi)/2$  is to be preferred.

#### 4-3.2 The Weighted and Unweighted Batch Means charts

Runger and Willemain (1995) doubted that the residuals are strictly uncorrelated because the fitted time-series model could possibly be inadequate. They proved that for an AR(1) process, if the mean of the process shifts by  $\delta$  standard deviations from target, then the mean of the first residual after the shift is  $\delta$ , and of the consequent residuals is  $\delta(1 - \phi)$  instead of 0, i.e.:

$$\begin{aligned} R_t &= \delta + \epsilon_t, \text{ if } t=1, \text{ and} \\ &= \delta(1 - \phi) + \epsilon_t, \text{ if } t>1. \end{aligned}$$

Thus, the AR(1) model responds to the change in the mean and partially incorporates the shift in the mean into its forecasts.

##### 4-3.2.1 The Weighted Batch Means (WBM) chart

Bischak, Kelton and Pollock (1993) derived a way to eliminate the possible autocorrelation left in the residuals by using independent subgroups of residuals to monitor the process mean. Starting with an ARMA model, they calculated the weights needed to cancel autocorrelation between batch means as functions of the batch size and the model parameters. If the batch size is  $b$  and the  $j^{\text{th}}$  batch is formed from consecutive data values  $X_{(j-1)b+i}$ , the  $j^{\text{th}}$  weighted batch mean is:

$$Y_j = \sum_{i=1}^b w_i X_{(j-1)b+i}, j=1,2,\dots$$

(4-6)

The batch size  $b$  can be selected to tune performance against a specified shift  $\delta$ . The weights must sum to unity for  $Y_j$  to be an unbiased estimate of the process mean  $\mu$ . For AR(p) processes, the optimal weights are the same for the middle of the batch but differ in sign and magnitude for the first and last values in the batch. For example, for the AR(1) model, Runger and Willemain (1995) proved that the weights are:

$$\begin{aligned} w_1 &= \frac{-\phi}{(b-1)(1-\phi)} \\ w_i &= 1/(b-1), \quad i=2, \dots, b-1 \\ w_b &= \frac{1}{(b-1)(1-\phi)} \end{aligned} \quad (4-7)$$

Given normal data and any batch size  $b>1$ , the optimal weights produce batch means that are iid normal with mean and variance:

$$E(Y_j) = \mu \text{ and } \text{Var}(Y_j) = \frac{1}{(1-\phi)^2(b-1)} \quad (4-8)$$

To construct the WBM chart of the residuals, we form in batches successive residuals derived from the respective observations. Runger and Willemain (1995) proved that the WBM of size  $b$  is the average of  $b-1$  successive residuals divided by  $(1-\phi)$ , i.e., by substituting (4-7) in (4-6), the WBM of the residuals is  $Y_{j+1} = \frac{\bar{r}}{1-\phi}$ , with  $\bar{r}$  being the average of the  $b-1$  residuals  $r_{jb+2}, r_{jb+3}, \dots, r_{jb+b}$ . These  $Y_j$  values of the residuals are plotted on the WBM chart with control limits equal to  $\pm L\sigma_Y$ , where  $\sigma_Y$  is derived from equation (4-8) for an AR(1) model.

#### 4-3.2.2 The Unweighted Batch Means (UBM) chart

The UBM chart, also proposed by Runger and Willemain (1995), differs from the WBM chart in that it gives equal weights to every point in the

batch. In this case all weights are  $w_i = 1/b$  for  $i = 1, \dots, b$  and by substituting these weights to Eq(4-6), we derive the  $j^{\text{th}}$  **unweighted** batch mean as:

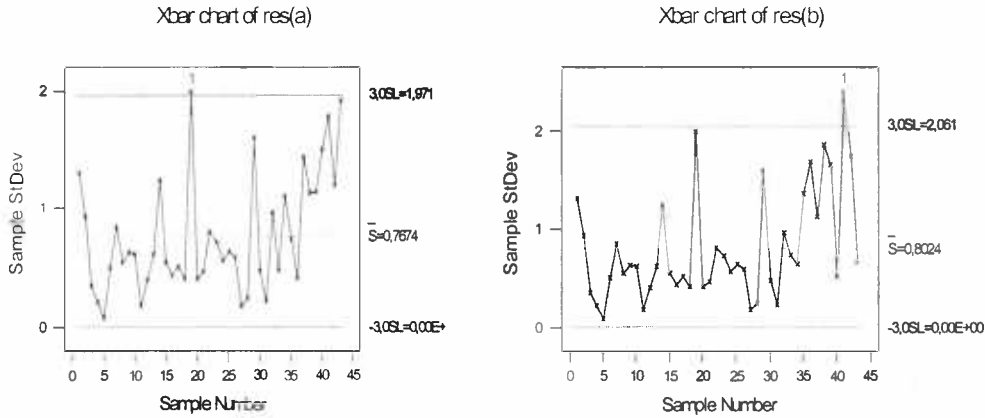
$$V_j = \frac{\sum_{i=1}^b X_{(j-1)b+i}}{b}, j = 1, 2, \dots \quad (4-9)$$

The unweighted batch means define a model-free approach and they can be plotted and approximately analyzed on a standard Individuals control chart with the control limits of traditional individuals charts. As distinct from residual plots, UBM charts retain the simplicity of averaging observations to form a point on a control chart. With UBM, averaging is used to dilute the autocorrelation of the data.

The important implication of UBM is that, though one does not need to make an ARMA model of the data, one has to determine an appropriate batch size  $b$  which is more difficult to do when not being guided by the selection of a time-series model. Runger and Willemain (1995) provided a detailed analysis of batch sizes for AR(1) models and recommended to select the batch size that reduces the lag one autocorrelation of the batch means to approximately 0.1.

Simulation analysis often use Fishman's (1978) procedure: Start with  $b=1$  and double  $b$  until the lag one autocorrelation of the batch means is sufficiently small. The advantages and disadvantages from further increasing the batch size are the same as for a conventional Shewhart chart. Larger batches are more effective for detecting smaller shifts, though smaller batches respond more quickly to larger shifts.

The example of autocorrelated data that we have been used to construct the graphs of this chapter up till now cannot be fitted adequately by a model of AR(1) type, in order to verify the findings of the authors. What we can do, though, is just create subgroups of residuals instead of considering them as individual observations. If for example, we consider every 3 consecutive residuals as one batch, then Figure 4-5 shows that in fact both the  $1\sigma$  shift (a) and the  $2\sigma$  shift (b) in the mean are detected more successfully than in the case of the individual residuals chart shown in Figure 4-4.



**Figure 4-5: The Shewhart chart of the residuals with  $n=3$  for  $1\sigma$  (a) and  $2\sigma$  (b) shift of the process mean.**

#### 4-4 Control charts applied to the forecast errors

Another approach to deal with autocorrelated data has been proposed by Montgomery and Mastrangelo (1991). The proposed procedure is to plot one-step-ahead EWMA prediction errors on a control chart. Two versions of the same procedure are provided: the *M-M chart* and the *Moving Center-line EWMA control chart*.

##### 4-4.1 The M-M chart

The EWMA is not only used to monitor a process but it can also provide a forecast of where the process mean will be at the next time period. The usual EWMA can be written as  $z_t = \lambda x_t + (1-\lambda)z_{t-1} = z_{t-1} + \lambda(x_t - z_{t-1}) = z_{t-1} + \lambda e_t$ , because if we view  $z_{t-1}$  as a forecast of the process mean in period  $t$ , we can think of  $x_t - z_{t-1}$  as the forecast error  $e_t$ .

If the process can be modeled by the first-order integrated moving average model IMA(1,1), being of the form  $x_t = x_{t-1} + \epsilon_t - \theta\epsilon_{t-1}$  shown also in Eq(3-19), it has been proven by Box and Jenkins (1976) that the corresponding EWMA with  $\lambda=1-\theta$  is the optimal *one-step-ahead forecast* for this process. That is, if  $\hat{X}_{t+1}(t)$  is the forecast of the observation in period  $t+1$  made at the end of period  $t$ , then:



$$\hat{X}_{t+1}(t) = z_t.$$

The sequence of one-step-ahead prediction errors  $e_t = x_t - \hat{X}_t(t-1)$  is independently and identically distributed with mean zero. Therefore, control charts could be applied to these one-step-ahead prediction errors. The optimal parameter  $\lambda$  is found by minimizing the sum of squares of the errors  $e_t$ .

In general, if the observations from the process are positively autocorrelated and the process mean does not drift too quickly, the EWMA with an appropriate value for  $\lambda$  will provide an excellent one-step-ahead predictor. Consequently, we would expect many processes obeying first-order dynamics (that is, they follow a slow drift) to be well represented by the EWMA without being exactly modeled by the first-order integrated moving average model. The one-step-ahead prediction errors could be plotted in a Special Cause chart described in section 4-3.1.2 in the place of the residuals.

This scheme, for simplicity called the M-M chart, is like the residual chart except that the IMA(1,1) model is assumed for all the processes and the prediction errors are used. It could be accompanied by a chart of the original observations on which the EWMA forecast is superimposed, as the chart of the original observations allows process dynamics to be visualized while the chart of the residuals does not.

If the process is modeled as an AR(1), then the EWMA forecast may not be the best predictor for this model (as it is for the IMA(1,1)), but it is still an accurate forecast. Cox (1961) has shown that the optimal EWMA forecast (in the sense that the mean squared error is minimized) for an AR(1) process with parameter  $\phi$  is given by:

$$\lambda = 1 - [(1 - \phi)/2\phi], \text{ where } 1/3 < \phi \leq 1 \quad (4-10)$$

By extending the EWMA forecast, if the IMA(1,1) model is not very different from the true process, the one-step ahead prediction could replace the residuals derived from fitting ARIMA models. Thus, a broader idea could be formed by constructing both the residuals and the M-M chart. If the one of the two charts results in out-of-control signals, while the other does not, one should be suspicious about the estimation of the true correlation structure.



#### 4-4.2 The Moving Center-line EWMA control chart

Montgomery and Mastrangelo (1991) initiated another control chart combining both the information about the state of statistical control and the process dynamics. Assuming that the one-step-ahead prediction errors (or alternatively the model residuals)  $e_t$  are normally distributed, then the usual three-sigma control limits on these errors satisfy the probability statement:

$$P[-3\sigma \leq e_t \leq 3\sigma] = 0.9973 \rightarrow P[-3\sigma \leq x_t - \hat{X}_t(t-1) \leq 3\sigma] = 0.9973 \rightarrow$$

$$P[\hat{X}_t(t-1) - 3\sigma \leq x_t \leq \hat{X}_t(t-1) + 3\sigma] = 0.9973,$$

where  $\sigma$  is the standard deviation of the errors or of the residuals  $e_t$ . The Moving Center-line EWMA control chart plots the one-step-ahead predictors instead of the errors themselves and it is constructed as:

$$\begin{aligned} \text{UCL}_{t+1} &= z_t + 3\sigma \\ \text{Center line} &= z_t \\ \text{LCL}_{t+1} &= z_t - 3\sigma \end{aligned} \quad (4-11)$$

where the standard deviation of the one-step-ahead prediction errors or of the model residuals is estimated by dividing the sum of squared prediction errors for the optimal  $\lambda$  by the number of observations  $n$ . The Moving Center-line EWMA control chart would be preferable from an interpretation standpoint to a control chart of residuals and a separate chart of the EWMA's, as it combines information about process dynamics and statistical control in one chart.

#### 4-5 A new chart for correlated data: The ARMAST chart

The Autoregressive Moving Average (ARMA) chart for independent and identically distributed processes and the ARMAST chart for autocorrelated processes were introduced by Jiang, Tsui and Woodall (2000). Assuming that the ARMAST chart is applied to a known stationary process  $\{X_t\}$ , the ARMAST statistic  $Z_t$  with parameters  $\phi$  and  $\theta$  can be represented by:

$$Z_t = \theta_0 X_t + \alpha \sum_{k=1}^{t-1} \varphi^{k-1} X_{t-k}, \text{ where } \alpha = \phi \theta_0 - \theta \text{ and } \theta_0 \text{ is chosen so that the sum of the coefficients is unity when } Z_t \text{ is expressed in terms of } \alpha_t. \quad (4-12)$$

The ARMA chart signals when  $Z_t > L\sigma_z$ , where:

$$\sigma_z^2 = \left\{ \theta_0^2 + 2\theta_0\alpha \sum_{k=1}^{t-1} \varphi^{k-1} \rho(k) + \alpha^2 \sum_{i=1}^{t-1} \sum_{j=1}^{t-1} \varphi^{i+j-2} \rho(j-i) \right\} \sigma_x^2.$$

The asymptotic or steady-state variance is: (4-13)

$$\sigma_z^2 = \left\{ \theta_0^2 + \frac{\alpha^2}{1-\varphi^2} + 2 \left( \theta_0\alpha + \frac{\varphi\alpha^2}{1-\varphi^2} \right) \sum_{k=1}^{\infty} \varphi^{k-1} \rho(k) \right\} \sigma_x^2.$$

When the original process is an ARMA(1,1) process with parameters  $u$  and  $v$ , the application of the ARMA chart to the ARMA(1,1) process is proved by Jiang, Tsui and Woodall (2000) to result in a generalized ARMA(2,2) model, that is:

$$Z_t = (\phi+u)Z_{t-1} - \phi u Z_{t-2} + \theta_0 \alpha_t - (\theta + \theta_0 v) \alpha_{t-1} + \theta v \alpha_{t-2} \quad (4-14)$$

In general, because the autocorrelation structure of the ARMA chart on an ARMA(1,1) process depends on the parameters of the charting process ( $\phi$  and  $\theta$ ) as well as those of the original process ( $u$  and  $v$ ), the performance of the ARMA chart depends on all four parameters. It is, therefore, hard to characterize the performance of the ARMA chart. However, if the parameters of the ARMA chart of Eq (4-14) are chosen as  $\phi=v$  and  $\theta/\theta_0=u$ , then the monitoring process reduces to  $Z_t = \theta_0 \alpha_t$ . Jiang, Tsui and Woodall (2000) proved that this monitoring process as well as the mean shift pattern of the class of ARMA chart are the same with those of the SCC chart of Alwan and Roberts (1988) except for the scaling constant  $\theta_0$ . Therefore, the performance of the ARMA chart with  $\phi=v$  and  $\theta/\theta_0=u$  is identical to the performance of the SCC chart applied to an ARMA(1,1) process with parameters  $u$  and  $v$ .

When the data are uncorrelated, then Eq(4-12) becomes:  $Z_t = \theta_0 \alpha_t - \theta \alpha_{t-1} + \phi Z_{t-1}$  and, since the coefficients must sum to unity when  $Z_t$  is expressed

in terms of  $\alpha_t$ , it is derived that  $\theta_0 = 1 + \theta - \phi$ . Note that the ARMA chart reduces to the EWMA chart if the data are not correlated, and  $\theta=0$  and  $\phi=1-\lambda$ . Thus, the ARMA chart can be considered as an extension of the EWMA chart. The steady-state variance when the data are independent and identically distributed is derived from Eq (4-13) when  $X_t = \alpha_t$  and  $\rho(k)=0$ .

#### Choosing the parameters of the ARMA chart

It is difficult to derive the optimal parameters of the monitoring process, but the authors proposed a heuristic strategy including the calculations of the *transient shift*  $\mu_T = \theta_0\mu$  (that is, the mean value when the process begins, at  $t=0$ ), and of the *steady-state* (or asymptotic, with  $t = \infty$ ) *shift*  $\mu_s = \mu$ .

Consequently, the *transient signal-to-noise ratio* and the *steady-state signal-to-noise ratio* are derived by the formulas:

$$R_T = \mu_T/\sigma_z \text{ and } R_S = \mu_s/\sigma_z, \text{ respectively} \quad (4-15)$$

The transient signal-to-noise ratio measures the capability of a chart to detect a shift in the first few runs, while the steady-state signal-to-noise ratio measures the ability to detect the shift in the later runs and is used if the shift has not been discovered soon after it has been interfered into the process, that is, when the transient signal-to-noise ratio is low.

Jiang, Tsui and Woodall (2000) suggested that if the transient ratio can be tuned to a high enough value of about 5 by choosing appropriate ARMA chart parameters, the corresponding chart will be able to detect the shift quickly. On the other hand, if this ratio is smaller than 3, the shift will likely be missed at the transient state and needs to be detected in the later runs. In this case, the steady-state ratio becomes more important for detecting the shift efficiently at the steady state. The steady state ratio should not be tuned too high, however, because it may result in an extremely small transient ratio and make the transition of the shifts from the transient state to the steady state very slow. To make the chart detect the shift fast in the steady state, a balance is needed to make a trade-off between the transient ratio and the steady-state

ratio when choosing the charting parameters. Based on these guidelines, the design of Figure 4-6 is derived (Jiang, Tsui and Woodall, 2000).

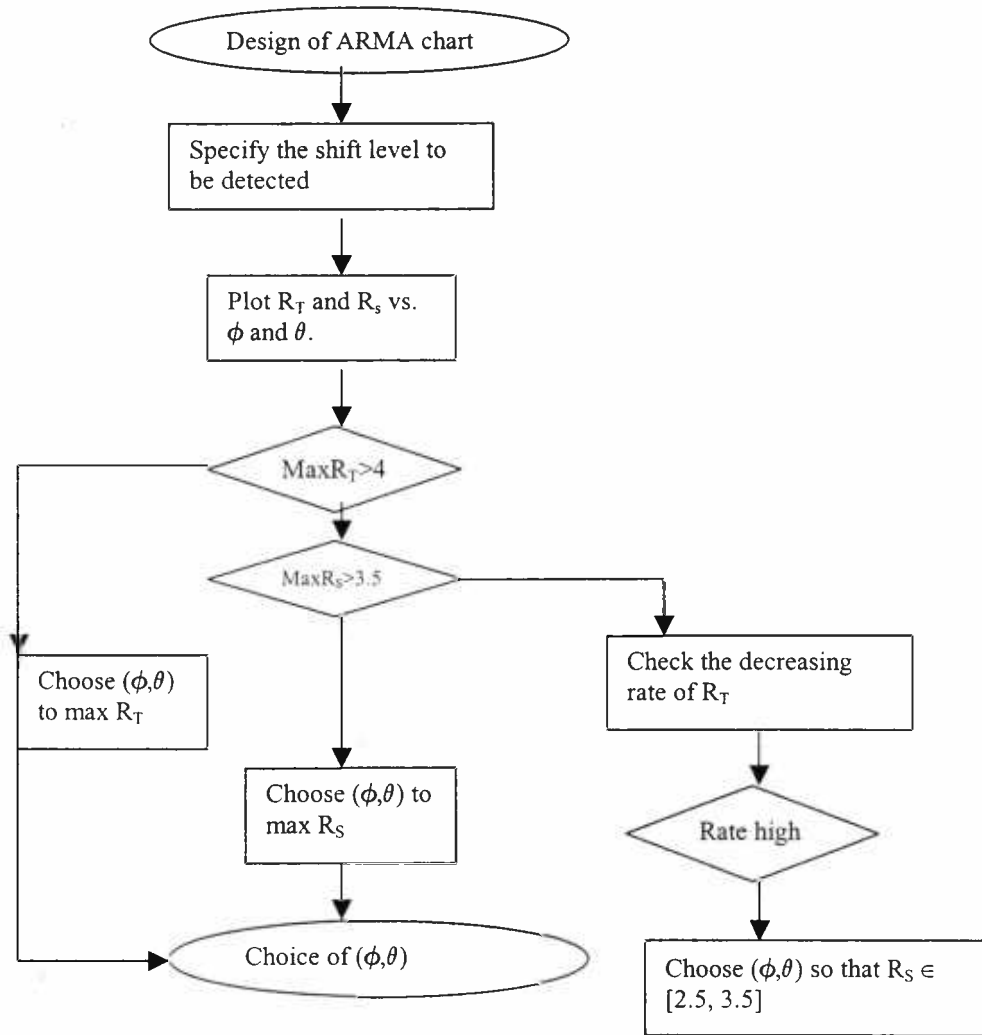


Figure 4-6: Parameter design of ARMA charts.

If one does not know the process model, it is still possible to design an ARMA chart. In this case, the standard deviation of the original process ( $\sigma_x$ ) can be estimated by the sample autocovariance of the data and the standard deviation of the charting process ( $\sigma_z$ ) can be estimated from the sample autocovariance of the ARMA statistics. The ratio  $\sigma_x/\sigma_z$  can then be estimated and the two signal-to-noise ratios can be obtained as:

$$R_T = \theta_0 \sigma_x / \sigma_z \text{ and } R_S = \sigma_x / \sigma_z.$$

Figure 4-2 presenting the EWMA (with parameters  $\theta = 0$  and  $\phi=1-\lambda = 0.8$ ) as well as the EWMAST chart are examples of the ARMA chart. The

performance of the ARMAST chart is studied in detail in the next Chapter. However, even though it may be very sufficient in detecting special causes for some cases, its complexity makes it not at all practical for widespread use.



## CHAPTER 5

### Performance of charts for autocorrelated data

#### 5-1 Introduction

In the previous chapter we were concentrated on the description of the most popular control charts used in the case when the process is of some autocorrelation type. Apart from the design of these charts, one would be interested to know in which case a chart is superior to others, because, naturally, there is no chart outperforming the others for all types of autocorrelation and for all estimation procedures. Section 5-2 describes the criteria used to combine the performance of the charts, as well as the guidelines for conducting simulation studies used for deriving more accurate conclusions on the effectiveness of the charts. In section 5-3 the performance of the modified standard charts is examined. The effectiveness of the residuals charts and of the charts applied to forecast errors is discussed in sections 5-4 and 5-5, respectively. Lastly, section 5-6 concerns the performance of the ARMA chart, presented in the previous chapter.

#### 5-2 The design of simulation studies

The simulation approach is used to create the data needed for combining the effectiveness of the control charts. The reason why simulation is preferred to real data is that, in the second case, one does not know if the statistical control state of the process has been changed and when this change has occurred. On the contrary, if the data is created, one may interpose a shift at a specific time point and, then, he can simply observe by which charts the known shift has been detected. For more accurate results, the above procedure is repeated for many similarly constructed data.

### 5-2.1 The most common performance criteria

The performance of the charts has been studied by many authors having based their results on some performance criteria. The performance criteria most frequently used are:

#### a) *The ARL criterion*

The Average Run Length (ARL) is the most widely used measure of performance that helps one decide about the ability of a chart to detect shifts of the process quickly. As it has been already mentioned, it is the average number of observations until an out-of-control point is observed. For a given control chart, we desire the ARL to be large when no assignable cause has occurred (i.e., when the shift in the process mean is 0), because otherwise we would falsely consider the process as being out-of-control. On the other hand, if there is a shift in the mean, we desire the ARL to be small, that is, to have a signal the quickest possible.

When comparing the performance of different charts, the ARL value is meaningful only if it is initially specified to a fixed number, so that all charts start with the same performance. That is why, the  $\sigma$  multipliers of the control charts (i.e., the  $L$ 's in Eq(2-1)) are manipulated in order to have the same in-control ARL. The value of the in-control ARL used is often 370, because this is the in-control ARL of the  $3\sigma$ -limits Shewhart chart. It is analogous to matching the Type I errors so that the Type II errors can be compared in a more meaningful way.

It is usually difficult to find a close formula for the calculation of the average run length especially for charts designed for autocorrelated processes and, usually, there is no standard calculation method for all ARIMA models. Zhang (1997) has initiated a formula for the ARLs of the residual chart but only for AR(1) and AR(2) processes. Runger and Willemain (1995) suggest a computational method for the ARL of the WBM and UBM charts when the process is of AR(1) type.

Apart from the ARL itself, one could use as performance measurement the probability of signaling within a fixed period. For example, Wiel (1996)





chose the probability of signaling within 10 periods after the shift as an alternative for the ARL value, which he denoted by  $P(10)$ .

*b) The run-length distribution criterion*

A more reliable criterion is to use not only the average of the run length, but, if possible, to have a broad idea about its distribution. However, apart from the run-length distribution of Shewhart control charts on iid data which is widely known to be the geometric, it is much more difficult to calculate the probability distribution of monitoring schemes used for autocorrelated data.

Three good ways to study the run-length distributions of a monitoring scheme are:

- (1) by analytically deriving it,
- (2) by approximating it in a discrete Markov-chain representation, and
- (3) by building it up through Monte Carlo simulation.

Though simple analytical results are available for deriving the run-length distribution of the Shewhart individual charts, this is not the case with the CUSUM scheme for which a computational technique based on Markov chains is, however, available by Brook and Evans (1972), and a similar approach is described in detail for the EWMA charts by Lucas and Saccucci (1990).

By knowing the probability function of the ARL distribution, one can derive the first and second moments, which can be used to find the average run length (ARL) as well as the *standard deviation of the run length* (SRL). This is what was done by Wardell, Moskowitz and Plante (1994) who determined the run-length distribution of the special-cause chart as an extension of the run-length distribution of the standard Shewhart control chart. They found closed-form solutions for the residuals chart if  $q > p$  in the fitted  $ARMA(p,q)$  process and semiclosed-form solutions for  $q \leq p$  for computing the ARL and the SRL recursively. Jiang, Tsui and Woodall (2000) considered a Markov chain approach to approximate the run-length distribution of the ARMA chart.

Knowledge of the run-length distribution enforces the objectivity when comparing different monitoring schemes, because the fact that in some cases the SRL may be larger than the ARL itself shows that in these charts the detection of the shifts is not precise. Thus, these charts may be considered inferior to others with smaller SRL even if their ARL value is smaller, too. However, if only the ARL values were available this conclusion would be impossible.

*c) The Cumulative Distribution Function (CDF) criterion*

The in-control run-length distributions of the Shewhart, the CUSUM and the EWMA charts for uncorrelated observations, as well as of control charts on positively correlated observations, are right-skewed. Furthermore, the EWMA as a forecasting tool, recovers quickly from disturbances in the process and, therefore, the ‘window of opportunity’ available for detecting process shifts may be quite small. This makes the probability of detection within the first few observations after the shift particularly important. That is why Superville and Adams (1994) suggested the use of the CDF of the run length, which is the percentage of signals given by the  $i$ th observation after the shift. The CDF is an alternative criterion to the ARL for detecting disturbances in positively autocorrelated data, because the ARL is often distorted in these cases.

*d) The Dynamic Step Response Function (DSRF)*

The dependence of the EWMA and ARMA forecasts (and, consequently, of the parameters  $\lambda$  and  $(\phi, \theta)$ , respectively) on past data result in a dynamic reaction to the shift and, thus, in the gradual convergence to a new steady value instead of an immediate set up to a new value. Another approach to account for the distortion of the ARL in this case is, apart from the CDF value described above, the proposed by Wardell, Moskowitz and Plante (1992) DSRF criterion. The DSRF describes how the forecast (or the residual in the case of the SCC chart) would dynamically react to a shift in the process mean if there was no noise in the process (i.e., if the process was completely deterministic). The authors specified the DSRF values, *if the*



model of the process is  $ARMA(1,1)$ , for four charts: the Individuals, the EWMA, the CCC and the SCC as follows:

$$\text{Individuals chart: } DSRF_x(j) = \delta / L_x$$

$$\text{EWMA chart: } DSRF_H(j) = \sigma[1-(1-\alpha)^j]/L_H$$

$$\text{CCC chart: } DSRF_F(j) = \frac{\delta}{L_F}(\phi_1 - \theta_1) \left( \frac{1 - \theta_1^j}{1 - \theta_1} \right)$$

$$\text{SCC chart: } DSRF_R(j) = \frac{1}{L_R} \left[ \delta - \delta(\phi_1 - \theta_1) \left( \frac{1 - \theta_1^j}{1 - \theta_1} \right) \right],$$

where  $j$  is the number of observations since the step change occurred,  $\delta$  is the size of the step change and  $L$  is the  $\sigma$  multiplier for each chart.

In all cases, the DRSF has been normalized by the value of the upper control limit, so a normalized response of 1 or larger indicates that the mean of the statistic has exceeded its upper limit (since noise has been removed, the time at which the step response function exceeds 1 or  $-1$  is a rough approximation to the ARLs of the charts, but it is not exact). Knowing the dynamic response of each chart to a shift in the process mean allows us to explain the difference in ARLs when comparing the performance of these four charts.

#### e) The signal-to-noise (SN) ratios

Jiang, Tsui and Woodall (2000) had used the transient and steady-state signal-to-noise ratios in order to choose appropriate parameters of the ARMA chart for autocorrelated processes. Since the transient ratio measures the capability of the chart to detect the shift in the first few runs and the steady-state ratio measures the chart efficiency in later runs, these two ratios can be well used as a performance criterion in the place of the ARL values.

### 5-2.2 Guidelines for the simulation procedures

One of the ways for estimating a performance criterion is known as the Monte- Carlo simulation. This method is said to be the most accurate, since it solves the estimation problems by creating situations close to the real ones. That is why most authors conducted simulation studies to estimate the ARL

value of a control chart according to which they, afterwards, specified an approximate close formula for ARL, or verified the one already formed. The basic idea of the simulation is that in order to have a good idea about the performance of a chart, one should not use one set of data to conclude about the detection or not of a process shift, but this procedure must be repeated many times if one wants to make sure about the results of his study. The general simulation plan is constructed in the following manner (inspired by Adams and Tseng, 1998):

1. For uncorrelated data a series of  $r$  values with  $r$  usually being between 10 000-100 000 is generated from a well-known distribution (usually the normal) with fixed parameters. To create correlated data, a set of  $r$  random values  $[\epsilon_t \text{'s} \sim N(0,1)]$  is generated and the observations  $x_t$ 's are calculated using Eq(3-3), i.e.,  $x_t = \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q}$  of the general ARMA model with fixed values for  $p$  and  $q$ . Several values of  $\phi$  and  $\theta$  are often used ranging from 0 to 1. Apparently, some initial values for  $x$  and  $\epsilon$  should be specified.

2. If the data are autocorrelated, the residuas of the observations or the EWMA forecast residuals are calculated after having estimated the parameter values  $\hat{\phi}$  and  $\hat{\theta}$  by the procedure mentioned in Chapter 3 and described in detail in Box and Jenkins (1976).

3. The first (usually 100) values of the residuals or of the EWMA forecasts (or of the observations themselves if they are iid) are discarded to allow a 'burn-in' period, so that the effect of the initial values is removed.

4. The remained residuals or the EWMA forecasts (or the observations themselves) are monitored using the control charts we are interested in and the chosen performance criterion is calculated for this data set for each chart.

5. Steps (1)-(4) are repeated many times (for example 1000). The performance criterion for each type of control chart is recorded for each simulation repetition (each repetition consisting of 10 000 - 100 000 observations) and the average value of the criterion of interest based on 1000 repetitions is obtained for each chart.



Naturally, this general plan will be different from study to study by often allowing for data sets in which a mean shift occurs at a specific time point or using fixed formulas to calculate some performance criterion, as is the DSRF measurement. Obviously, many criteria may be calculated for each data set to help derive accurate conclusions.

### **5-3 Performance of traditional charts based on simulation studies**

Chapter 4 presented the most popular control charts referring to autocorrelated processes. Among the efforts to find a chart that detects shifts quickly, an idea was to modify the already existing standard charts. The performance of these charts is an important issue, so that one knows which chart is more appropriate according to the manufacturing process of interest.

#### **5-3.1 Performance of the modified Shewhart chart**

Wardell, Moskowitz and Plante (1992) have conducted a simulation approach for an AR(1), a MA(1) and an ARMA(1,1) model and deduced that the modified-Shewhart chart does not perform well (in terms of its ARL values) in the case of correlated data and it rarely achieves a predicted ARL lower than those obtained by other charts. Conventional control charts such as the Shewhart chart (modified or not) are not completely robust to deviations from the assumption of process randomness, namely when observations are correlated.

Increasing the subgroup size of the modified-Shewhart chart for ARMA(1,1) models substantially increased its ability to detect shifts in the mean quickly, exactly as happens when the observations are independent, shown by Wadsworth, Stephens and Godfrey (1986). This would suggest that when it is possible and practical to take observations in subgroups rather than individually, one should do so to improve the performance of the modified-Shewhart control chart. However, if the data are truly autocorrelated, each point on a Shewhart chart will still show runs which are essentially due to correlation resulting from common causes rather than any special cause. Thus,

in autocorrelated processes, care should be taken when using either the simple or the modified Shewhart chart.

### 5-3.2 Performance of the EWMAST chart

Zhang (1998) combined the performance of the EWMAST chart with those of the traditional Shewhart, as well as the residuals chart and the M-M chart, for AR(1), MA(1) and ARMA(1,1) processes in terms of their ARL values. For AR(1) processes, the EWMAST chart (with  $\lambda = 0.1$  and  $\lambda = 0.2$ ) performs better than the residual chart for weak and medium autocorrelations ( $\phi < 0.75$ ), especially for small to medium mean shifts. When the autocorrelations are positive, the EWMAST chart has very large *in-control ARLs*, but the residuals chart has much smaller *out-of-control ARLs* and, thus, performs better than the EWMAST for detecting a mean shift, especially when the shift is large. The ARL values of the EWMAST and of the residuals chart for some of the simulated AR(1) models, studied by Zhang (1998), are shown in Figure 5-1.

For an AR(2) process, when the process is not near nonstationary (i.e.,  $\phi_1 + \phi_2$  is not near 1), the EWMAST performs better than the residuals chart, especially when the mean shift is not large. **Thus, the residuals chart performs better than the EWMAST chart only when the process is near nonstationary with strong positive autocorrelations.**

In most cases of the AR(1) model, the EWMA performs better than the Shewhart chart when the mean shifts are less than or equal to  $2\sigma_x$ . Only when  $\phi$  is positive and large (e.g.,  $\phi = 0.95$ ) and the mean shift is large ( $= 3\sigma_x$ ) does the Shewhart chart perform better than the EWMAST. Similarly to the AR(2) process, **the EWMAST chart performs much better than the Shewhart chart except when the process has strong positive autocorrelations or the mean shift is large.** The conclusions for ARMA(1,1) processes are similar.

Compared to the M-M chart, even when the mean shifts are medium or large, the out-of-control ARLs of the M-M chart are much larger than those of the EWMAST. Only when  $\phi = 0.95$ , the M-M chart performs relatively well but even in this case the residual chart is slightly better than the M-M chart. Zhang (1998) used various  $\lambda$  for the EWMAST chart and showed that for detecting



small shifts, a value of 0.1 or 0.2 for  $\lambda$  is better. Furthermore, the adjustment of the control limits for various in-control ARLs did not effect their conclusions. An obvious advantage of using the EWMAST chart is that there is no need to build a time series model as for the residual chart.

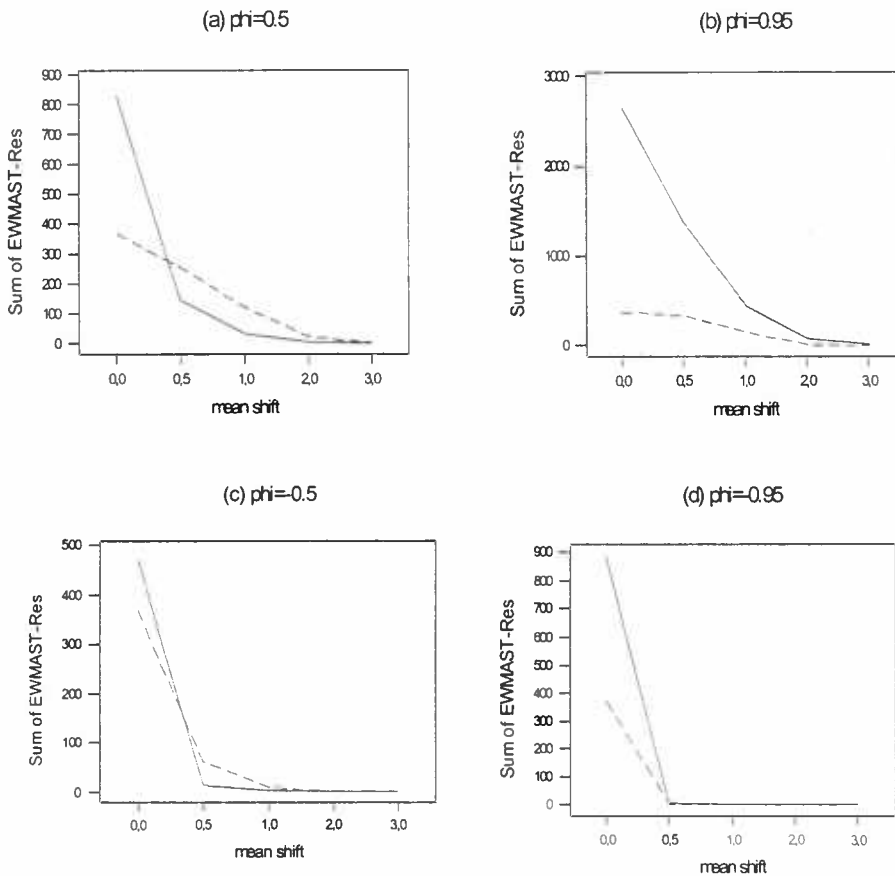


Figure 5-1: ARLs of the EWMAST with  $\lambda=0.2$  and of the Residuals chart for AR(1) processes with  $\phi=0.5, 0.95, -0.5, -0.95$  where: — EWMAST, ---- Residuals chart.

#### 5-4 Performance of the residuals charts

The effectiveness of the residuals chart (i.e., of the SCC chart) should be compared to other types of control charts. Another issue is to assess the relative performance of the residuals chart, that is, to apply different standard charts on the residuals in order to confirm which of the standard charts detects the mean shift more quickly when the residuals and not the original data are the ones plotted.



### 5-4.1 Performance of the SCC chart compared to traditional charts

Many simulation studies have been conducted in order to compare the effectiveness of the residuals charts versus the original ones which have not been modified to take into account the autocorrelation of the data. Wardell, Moskowitz and Plante (1994) determined mathematically the run-length distribution of the special-cause control chart (i.e., the Shewhart chart on the residuals) for a given ARMA(p,q) process and were, thus, able to draw some conclusions based on the relative performance of the SCC chart, the original Shewhart chart and the EWMA applied to correlated data.

To compare the ARLs of the special-cause chart (SCC) to the ARLs of the Shewhart and the EWMA charts for AR(1) processes, Wardell, Moskowitz and Plante (1994) used the simulation procedure and set the parameter  $\lambda$  of the EWMA as 0.1. They proved that in terms of the ARL, the **SCC chart for AR(1) processes is superior to traditional control charts only when the process is highly negatively correlated**. The reason for this is that, when the process is negatively autocorrelated and the mean shifts, the one-step-ahead forecast moves in the opposite direction of the shift. This causes the residual, that is the difference between the observation and the forecast, to be very large, and, hence, the shift is detected earlier.

Wardell, Moskowitz and Plante (1992) showed that when the process is AR(1), the modified-Shewhart and the Common Cause (CCC) charts have the same ARL, though when the process is MA(1), the SCC chart and the CCC have the same ARL. They also proved that when the process is ARMA(1,1) with various parameters for  $\phi$  and  $\theta$  for a shift of *1 standard deviation*, the EWMA chart has a smaller ARL than the modified Shewhart, the SCC and the CCC charts over most of the stationary region. This conclusion is shown graphically in Figure 5-2.

As the shift of the mean is increased to *3 standard deviations*, the CCC chart dominates most of the region with the EWMA still being superior when the autoregressive parameter  $\phi$  is negative and the moving average parameter  $\theta$  is positive, as shown in Figure 5-3. By setting the in-control ARL to 110 instead of 370, the EWMA chart again dominates most of the stationary region when the shift in the mean is small. However, as the shift increases





the other charts perform much better and again the CCC chart has the smallest predicted ARL among the 4 charts when the shift in the mean is 3 standard deviations.

Thus, according to Wardell, Moskowitz and Plante (1992) for an ARMA(1,1) process, the EWMA chart is very good at detecting small shifts and performs well for large shifts when the autoregressive parameter is negative and the moving average parameter is positive. However, since in practice we are usually more interested in detecting the larger, more costly shifts quickly, the SCC becomes more attractive. **As the shift increases, the SCC and CCC charts perform better over a wide range of the ARMA(1,1) parameters.** This is especially true if these two are used conjointly as they should.

It is also advantageous to draw limits to the CCC chart so as to help detect shifts in the process mean, since, when the shift is 3 standard deviations, the out-of-control condition is predicted by the forecasts in the CCC chart before the other charts indicate a process change. As no chart is obviously dominant under every condition, it would be worthwhile to measure the degree of correlation in the process data to decide which control chart would best suit the particular needs.

The most interesting part, though, were the deductions of Wardell, Moskowitz and Plante (1994) concerning the probability mass function of the run-length distribution for an SCC chart which has a very long tail and, thus, inflates both ARL and SRL. Therefore, although the ARL of the SCC for an AR(1) model is often longer than other charts (since a positive autocorrelation of the process is more possible than a negative one), the SCC chart has a higher probability of detecting a shift immediately during the dynamic response to the shift. The original Shewhart chart had a smaller probability of detecting the shift immediately for an AR(1) model, but the probability increased with higher run lengths. This results from the fact that when the shift first occurs, there is a large discrepancy between the observation and its forecasted value in the SCC chart, giving a large residual. In the next instance, however, the forecasted values begins to reflect the fact that the observations have shifted upward, and the forecasts shift up also. Hence, the residuals become smaller again.



The SRL of the SCC chart is smaller than the ARL when the process is negatively autocorrelated, while in the case of positive autocorrelation the SRL is greater than the ARL, so the time at which we actually detect the signal is not at all precise, except in the case where the ARL is equal to 1.

Wardell, Moskowitz and Plante (1992) used the DSRF measurement to determine how the EWMA forecasts, as well as, how the residuals from the ARMA model, will react dynamically to a shift in the process mean. They used the formulas described in section 5-2.1d) and found that for an ARMA(1,1) model when the data are negatively correlated, the CCC DSRF moves in the opposite direction of the shift in the mean. This causes the time until the forecast exceeds its control limits to be very long. The times when the CCC chart does well when the autocorrelation is negative are when the forecast falls below the lower control limit. This occurs when the autocorrelation is close to  $-1$  and when the shift is relatively large.

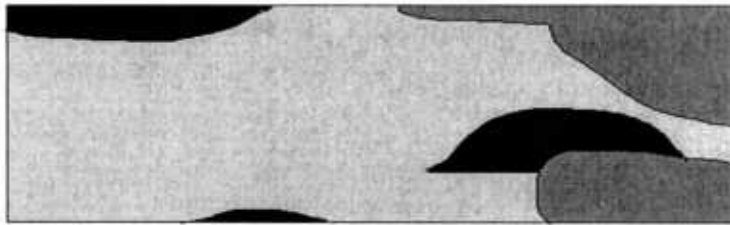






Figure 5-2: ARLs for a shift of one standard deviation:

EWMA , CCC , SCC 



Figure 5-3: ARLs for a shift of 3 standard deviations:

EWMA , CCC , SCC , Shewhart 

In both charts the horizontal axis is the values of  $\rho$  and the vertical the  $\delta$  values so that the dominance region for different combinations of  $\phi$  and  $\theta$  is seen for all the four charts.

#### 5-4.2 Relative performance of the residuals chart

We have already discussed that the residuals, after an ARIMA model has been fitted to the data, may be plotted on every traditional chart with no exception. Up to now we were concentrated only on the SCC chart, which is the Shewhart chart of the residuals, but the other control charts could be equally used. It is very obvious that since the residuals are uncorrelated, all the conclusions concerning the traditional control charts for iid data are also valid for the residuals charts. Thus, as Lucas and Saccucci (1990) have deduced, the EWMA and CUSUM control charts are effective when small shifts in the mean should be detected and the Shewhart control schemes are superior for detecting large shifts. Furthermore, the properties of the EWMA are very close to those of the CUSUM schemes.

However, both Wardell, Moskowitz and Plante (1994) and Zhang (1998) showed that the residuals charts may not have after all the expected performance since in many cases they have been outperformed by traditional charts and even by the simple Shewhart chart. This may be due to the fact that if the model of the process and its parameters have not been well estimated, then some autocorrelation in the data still remains and, thus, the residuals are not as uncorrelated as they should be. To solve this problem, Runger and Willemain (1995) proposed the weighted batch means which use the same time series model to determine the weights that render batch means uncorrelated.

They made comparisons of Shewhart ARLs for AR(1) data and showed that both batch means residual charts (WBM and UBM) outperform the simple residuals charts in almost all cases studied with the UBM chart performing best of all, although that in general model-based inference parametric models are more powerful than nonparametric.

**The WBM is superior to the residuals chart (i.e. the SCC) in the case of small mean shifts**, because for independent data, it is well known that larger subgroups provide greater sensitivity to small shifts. On the other hand, **the residuals chart is more effective than any batching strategy for very large shifts** because, for a large enough  $\delta$  even one observation is

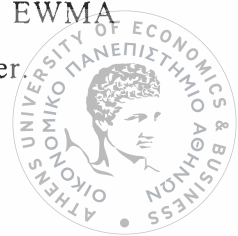
unequivocal evidence of a shift in the mean. In this case, there is no need to wait to collect an entire batch of data to declare an out-of-control condition.

### 5-5 Performance of control charts applied to forecast errors

Another way of estimating the residuals is to use the EWMA predictions proposed by Montgomery (1991) and, thus, calculate the forecast residuals. Superville and Adams (1994) performed a simulation study to quantify the effect of forecast recovery on the ARL of forecast-based schemes (EWMA predictors) and the relative performance of the Individuals, CUSUM and EWMA control charts applied to forecast errors for the AR(1) model. Control limits for each control chart on the forecast errors were determined to yield in-control ARLs of 250 or 500, the parameter  $\lambda$  for the EWMA chart on forecast residuals was set equal to 0.1 (its optimal value minimizing the sum of squares of the errors  $e_t$ ) as denoted from Eq(4-10) for the AR(1) model. The values of ARL and CDF were recorded for different shift sizes.

Based on the ARL values, Superville and Adams (1994) showed that the **CUSUM and EWMA charts detect more quickly shifts in the process mean of size up to  $2\sigma$** . On the other hand, for larger shifts, the Individuals chart signals more quickly. These results are consistent with those of Lucas and Saccucci (1990) for uncorrelated data. However, the magnitude of the ARLs for the independent case ( $\phi=0$ ) are significantly smaller than those for the non-independent case ( $\phi>0$ ) due in a large part to the recovery property of the forecasting tool.

The percentage of signals detected by the  $i$ th observation after the shift, that is, the CDF criterion presented in section 5-2.1c), also reveals a dominance of the CUSUM and EWMA control charts for small shifts, while the Individuals control chart provides a higher probability of signaling quickly for larger shifts. A substantial remark based on the CDF criterion, though not detected by the ARLs, is that the Individuals chart produces a signal on the observation immediately following the shift for approximately 50% of the simulated data sets with relatively few signals for subsequent observations, while the probability of either the CUSUM or the EWMA control chart of detecting a shift just after it has occurred is much lower.



We remind that the same conclusion was drawn by Wardell, Moskowitz and Plante (1994) for the residuals chart. This phenomenon results from the tendency of the EWMA forecast to recover quickly from process shifts suggesting that **in forecasting schemes, the superiority of the EWMA and CUSUM charts for detecting step shifts is doubtful**, at least in the case of AR(1) models. That is why, for correlated data, the monitoring of forecast residuals should be accompanied by the monitoring of the raw data.

A relative study was conducted by Adams and Tseng (1998), who simulated an AR(1) process and an IMA(1,1) process, obtained one-step-ahead forecasts and constructed EWMA, CUSUM and Individuals charts (with and without the sensitivity rules described in Chapter 1) on the forecast residuals. Because of the estimation error in  $\phi$ , the resulting one-step-ahead forecast residuals are not iid but are positively correlated, hence, the ARL of each control chart differs from the expected value of 225 which is the ARL when  $\phi$  has been estimated correctly.

For AR(1) models, **if  $\phi$  is overestimated, then the EWMA control chart and the CUSUM provide larger ARLs than the Individuals control chart with and without run rules**. The difference among the ARLs of the four control charts increases with the magnitude of overestimation of  $\phi$ . **When  $\phi$  is underestimated, the EWMA and CUSUM control charts provide shorter ARLs than the other two control charts (with and without run rules)** with the difference among the ARLs of the four control charts being less substantial as the magnitude of underestimation of  $\phi$  increases.

Figure 5-4 displays the ARLs for the Individuals control chart, the Individuals control chart with runs rules, the EWMA control chart and the CUSUM control chart applied to forecast residuals for an AR(1) process for which the true  $\phi$  is estimated as being equal to 0.5. This figure shows graphically the above discussion about the robustness of the charts of interest.

In general, the results for an IMA(1,1) process are the opposite of those for an AR(1) process. **When  $\theta$  is overestimated, the EWMA and CUSUM control charts provide smaller ARLs than anticipated**, with the differences among ARLs for the four control charts decreasing as the magnitude of the overestimation of  $\theta$  increases. **The Individuals control chart with or without run rules have smaller ARLs than anticipated when  $\theta$  is**

**underestimated**, though the EWMA and CUSUM have very large ARLs in this case.

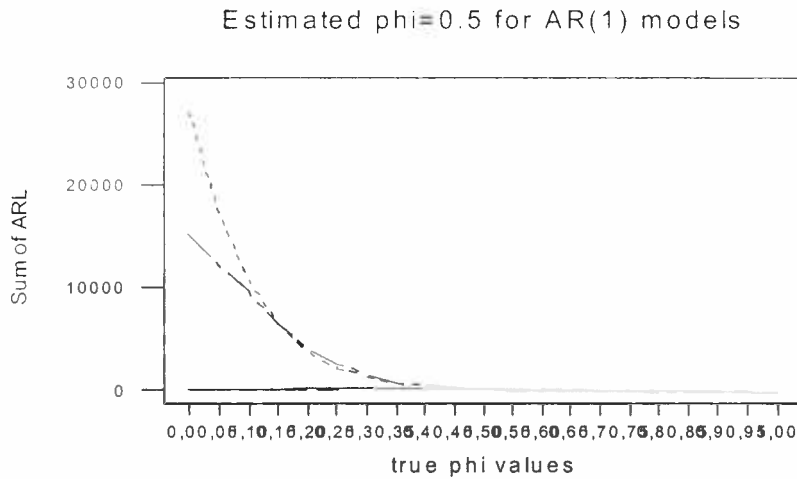
The performance of the two Individuals control charts are superior to the EWMA and CUSUM in the sense that they are less sensitive to estimation error in the process parameters. The explanation for the nonrobustness of the EWMA and CUSUM charts on the misspecification of the model parameters is the following: when the ARMA parameters are estimated with error, the forecast based monitoring scheme does not remove all the autocorrelation structure from the raw data, and the forecast residuals are correlated. As a result, the basic assumption of independence for the traditional control charts applied to the forecast residuals is still violated. Unlike the CUSUM and EWMA control charts, the Individuals control chart does not incorporate all past data into the plotted value, thus, its performance is not as seriously affected by the autocorrelation of forecast residuals caused by the estimation error. In general, the Individuals control chart with run rules is more affected by autocorrelation of the forecast residuals than is the Individuals control chart without run rules. The case of IMA(1,1) models with estimated  $\theta = 0.5$  is illustrated in Figure 5-5.

To summarize the conclusions of Adams and Tseng (1998) on the robustness of forecast-based control charts, we deduce that the performance of the Individuals control charts is superior to the performances of the EWMA and CUSUM charts in the sense that the in-control ARLs are better maintained in the presence of estimation error. When  $\phi$  is overestimated or  $\theta$  is underestimated, the forecast residuals are negatively correlated, thus, the EWMA and CUSUM charts provide much larger than anticipated ARLs, though the opposite is true when the forecast residuals are positively correlated. Finally, Adams and Tseng (1998) showed that substantial sample sizes are required for estimating process parameters and that updating and validating parameter estimates would be prudent.

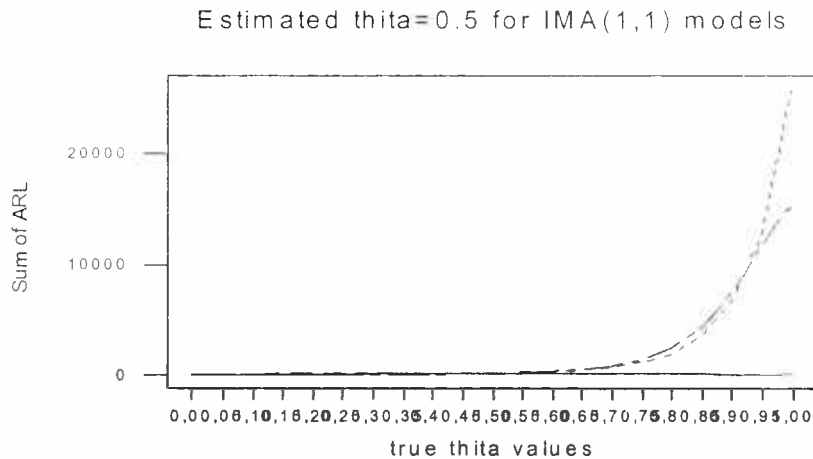
Unlike the fact that Montgomery (2001) suggests to use the EWMA predictor even when the process is not of IMA(1,1) type, Tseng and Adams (1994) demonstrate that the use of the EWMA forecast for models other than the IMA(1,1) can lead to unexpected performances and should be avoided.







**Figure 5-4: Average Run lengths for control charts applied to forecast residuals used for monitoring various AR(1) processes if their estimated parameter  $\phi = 0.5$ : — Ind with runs, ---- Ind, ..... EWMA, . — . CUSUM.**



**Figure 5-5: Average Run lengths for control charts applied to forecast residuals used for monitoring various IMA(1,1) processes if their estimated parameter  $\theta = 0.5$  with the annotation of Figure 5-4.**

## 5-6 Performance of the ARMAST chart

Jiang, Tsui and Woodall (2000) assumed that because both the SCC and EWMAST charts are special cases of the ARMAST chart, it is possible to derive an ARMA chart with appropriate parameter values that outperforms

both the SCC and EWMA charts. They proved that the ARMAST with appropriately chosen parameters either outperforms or performs competitively with the best of the EWMAST and SCC charts at least for AR(1) and ARMA(1,1) processes. Since both the transient and steady-state ratios are higher than those of the EWMAST chart, the ARMAST chart significantly improves the efficiency of detecting small mean shifts. When the data are uncorrelated, though, the optimal ARMA chart can be slightly better than the EWMA chart but the difference is not as significant as for the autocorrelated processes.

Jiang, Tsui and Woodall (2000) also showed that the WBM with optimal batch size is not as competitive as the ARMAST chart in all AR(1) processes. For ARMA(1,1) processes it is difficult to derive the optimal batch size and weights for the WBM chart. For an AR(1) model, the ARMAST chart is better than the CUSUM chart of residuals for detecting small shifts but it is worse than the SCC for detecting large shifts because the ARMA chart is designed to detect small shifts. It can be shown that the *EWMA chart applied to the residuals* can be modeled as a special case of the higher-order ARMA charts. This design, however, may involve too many parameters making it too complicated for implementation.





# CHAPTER 6

## The Engineering Process Control (EPC)

### 6-1 Introduction

Adams and Tseng (1998) showed that when the parameter estimates of the ARIMA model are not very appropriate, serious distortion may be done to the forecast-based schemes. Given the difficulties of model specification and a lack of robustness to parameter estimation, one might conclude that forecast-based monitoring schemes are of little value in practical settings.

There is a second approach, though, for monitoring a process and reducing its variability. This alternative method is based on adjusting the process using information about its current value or the deviation of the current value from a desired target, and it is often called **feedback adjustment** or **Engineering Process Control (EPC)**. This approach obviously differentiates from the SPC technique that reduces variability by detecting and removing causes of variation. Section 6-2 concentrates on the skeptic of the new approach compared to the widespread SPC methodology, section 6-3 describes the background and design of EPC, section 6-4 presents the most popular tools of EPC, section 6-5 gives an application that helps to better understand the use of EPC, section 6-6 refers to some other EPC schemes having the property of minimum cost instead of minimum variance, and section 6-7 discusses cases in which the EPC method is appropriate.

### 6-2 Differences between the SPC and EPC techniques

SPC aims to identify and act on causes of process changes (special/assignable causes) and on important contributors to variation that can be regarded as chronic noise (common causes). Control charts are very popular

as an on-line SPC tool and they involve two phases: the first is to achieve a state of statistical control and the second is to monitor the process so as to signal statistically significant deviations from the previously established statistical control. However, many commonly encountered circumstances are not amenable to the SPC approach. Apart from the case in which the process is autocorrelated, so the wrongly estimated model makes the SPC approach almost useless, other examples include:

- The process is subject to occasional shifts. Even when the causes of the shifts are known, it may be impossible to remove them, as is, for example, the raw material variability.
- The process is undergoing slow drift. In this case, SPC is not very effective because the drift must drift a certain distance before control action is taken, in response to an alarm. But if an inexpensive control action is available, then there is no reason to wait until the process has drifted 'far enough'. In addition, SPC does not specify what the control action should be.

In cases such as these, the EPC approach may be a good alternative since it does not remove the root or assignable causes but it uses continuous adjustments to keep process variables on target by transferring variability in the output variable to an input control variable with which the relationship is known.

In general, both techniques have the objective of reducing variability but they accomplish it in different ways. SPC looks for signals indicating assignable causes assuming that the process data can be described in terms of statistically independent observations that fluctuate around a constant mean. On the other hand, EPC is based on process compensation and regulation (instead of process monitoring), in which some manipulable process variables are adjusted to keep the process output on target. In other words, EPC assumes that there is a specific dynamic model that links the process input (manipulated variable) to the process output (quality characteristic) and, thus, a series of regular control actions to the input variable will keep the process output close to the desired target.

Montgomery (2001) gives a successful example helping someone to comprehend the superiority of the feedback adjustment to the SPC method in some cases: consider the process of driving a car, with the objective of keeping it in the center of the right-hand lane (or equivalently, minimizing variation around the center of the right-hand lane). The driver can easily see the road ahead, and process adjustments (corrections to the steering wheel position) can be made at any time at negligible cost. Consequently, if the driver knew the relationship between the output variable (car position) and the manipulated variable (steering wheel adjustment), he would obviously prefer using the feedback-control scheme to control the car position, rather than wait until the car is off the road to take some action.

### 6-3 Design of EPC

Provided that the future process performance is predictable (that is, there is a correlation in the data able to be identified) and that there exist readily adjustable variables for which the impact of performance is known, then the data is a good candidate for EPC. In order to have a better understanding of the procedure, it is preferable to follow the successive steps proposed by Faltin et al. (1993), summarized below.

#### Step 1: Initial process assessment

Successful use of EPC requires past information to be representative of future performance. The autocorrelation of the observed process performance might be empirically assessed from a plot of the process against time or it may involve a formal autocorrelation analysis. For such an evaluation to provide useful results, the ‘right data’ must be available, that is compensatory control actions that wipe out the underlying relationships should be taken into account so as not to effect the statistical evaluations.

#### Step 2: Model formulation

Model formulation involves building upon the previously established correlations to develop a mathematical model that will be useful for future

prediction and, thus, for control. Typical models to represent changes in process performance over time involve three components:

$$Y_t = X_{t-1} + D_t + e_t,$$

where  $Y_t$  is the output of the process at time period  $t$ ,  $X_{t-1}$  is the effect of any control action taken after the  $(t-1)^{st}$  observation,  $D_t$  is the effect of the underlying disturbances on the true process mean at time  $t$ , and  $e_t$  is an independent random variable with mean 0 and variance  $\sigma_e^2$  indicating the random effect of all the remaining negligible components.

1) *Developing a time series model for the disturbance term  $D_t$*

The disturbance term  $D_t$  includes all of the correlation of the data through time. Autocorrelation is not necessarily bad. It, does, however, mean that the process is somewhat predictable and this suggests the possibility of compensation. The disturbances may be changes in the raw materials tending to have immediate as well as lasting effects to the output variable.

The disturbance term, that is, the autocorrelation structure or common causes of the model in terms of the SPC jargon, is usually represented by an appropriate autoregressive integrated moving average (ARIMA) model. The time series model provides a framework to determine how well past process performance predicts future performance and it usually works well provided that the major causes that impacted process performance in the past continue to apply. A simple ARIMA model frequently used is the first-order integrated moving average [IMA(1,1)] for which we have already discussed that the best predictor for the next measurement is an exponentially weighted moving average (EWMA) of the current and past measurements, that is,

$$\hat{z}_{t+1} = \theta \hat{z}_t + (1-\theta)z_t = (1-\lambda) \hat{z}_t + \lambda z_t,$$

because  $\theta = 1-\lambda$  (6-1)

The simplicity of the IMA(1,1) model and especially of the EWMA predictor is a major reason for its attractiveness. Higher order ARIMA models can also be applied if they provide richer and more accurate presentations.

Sometimes, one can improve predictions of process performance by basing these not only on feedback of past performance, but also on

measurements of other important process variables. If there is sufficient understanding of the process and timely measurements on appropriate impacting variables, they can be used as a feedforward mechanism for predicting future process performance. An example of feedforward adjustment is measurements of relevant properties of the raw materials that feed into the process.

Thus, one could add to the time series model, appropriate regression-type terms involving measurements of variables that impact process performance. Adding such terms will make it necessary to modify the part of the time series model that previously accommodated the effect of variables that are now explicitly included in the model. Use of a feedforward scheme, when applicable, has the potential of providing important improvements in a timely manner.

## 2) *Including control variables ( $X_t$ ) in the model*

Our previous goal was to consider predicting future process performance as a function of past process performance and of impacting process variables that can be measured, which, at least in the short run, cannot be changed. For EPC to be viable, there must, in addition, exist control variables whose adjustment will have a predictable effect on process performance and will provide the desired short-term reduction in variability. Thus, we need to add into the model the impact on process performance of such control variables and, therefore, represent the process dynamics appropriately. In the chemical industry the amount of catalyst is a popular control variable.

The relationship between the control and the output variables may be of many different types. The simplest and most often used relation, however, is of linear form, leading to the following model for the effect of the manipulate variable to the quality characteristic of interest:

$$Y_t = X_{t-1} + D_t = bu_{t-1} + D_t,$$

where  $X_{t-1}$  has the linear form of  $bu_{t-1}$ , with  $u_{t-1}$  being the measurement of the control variable at the end of period  $t-1$ , and  $D_t$  is the time series model already specified.

3) *The error-term  $e_t$  of the model*

The error-term  $e_t$  encompasses error of observation (that is, sampling and measurement error), as well as sources of process variability between observations. The error term is assumed to have 0 mean because the partial errors annul the one the other, and variance equal to  $\sigma_e^2$  which is estimated by the data. The variance  $\sigma_e^2$  is the fundamental variability of the complete process or, in other words, is the common-cause variability. Since the definition of common-cause variability is the one that is common to all material produced or the one that can be affected only by a change in the system, the common-cause variability is associated with  $\sigma_e^2$ .

**Step 3: Procuring the needed data**

Given a potentially appropriate process model, the next step is to estimate the parameters of the model and assess its usefulness. This requires appropriate data from the process. Often, EPC is applied on an existing process on which some sort of action has been taken in the past. Carefully maintained records of process performance over a long period of time might provide a basis for fitting and evaluating a process model.

However, sometimes, recording of process performance that impacts variables and control actions is not sufficient because there have been no past adjustments or the control variables are confounded with the impacting process variables and an initial analysis is required. When observed data are insufficient, one must introduce deliberate variation into the process by means of a dither signal. If these perturbations are large enough so that their impact is detected early, statistical estimation may be enhanced and that is why they should not be so large as to distort normal process behavior.

Gustavsson et al. (1977) gave a quite general theory showing that prediction-error variance is increased when estimation is performed within an overparameterized model class. On the other hand, unbiased prediction may not even be possible if model orders are underspecified. The purpose of the identification and estimation step of the EPC procedure is to come up with a model that captures the effects of the control and disturbance portions of the innovations form of a process model.

#### Step 4: Model fitting and evaluation

Given a model and appropriate data, one should estimate the parameters in the model and to evaluate the adequacy of the fit. The estimation of the time series model is done by the procedure described in Chapter 2, while the effect of the control variable may be shown visually from its scatter plot with the output variable and can then be estimated using the regression technique.

Apart from the goodness-of-fit revealed by the residuals, a good property of the model is its ability to predict the process adequately. Because the fit of the data used to construct the model often provides a very optimistic evaluation, a verification of the model applied to data not used in the model fit but coming from a subsequent period of time, is essential. This step is finished if we are convinced that the future performance of the process is adequately predicted from past data on process performance which impacts variables and control actions.

#### Step 5: Developing a control rule

Apart from the model specified for the manufacturing process, one should also come up with a model for the control variable. The optimal control rule usually is obtained after changing the control variable (whenever this is practically feasible) and then determining the magnitude of the change to be made by minimizing the variance of the true process performance. This leads to a simple rule when process performance for the most recent observation is known at the time the next one is prepared. An example of a control rule could be the following:

$$u_t = c_1 u_{t-1} + c_2 y_t,$$

where  $c_1$  and  $c_2$  are determined directly from the previously fitted model.

If the cost of the control action is also considered, then a control rule minimizing the total cost could be preferred over the minimization of the process performance variance.

#### Step 6: Developing a monitoring procedure

Although that the feedback control scheme is often preferred over statistical monitoring if the data is appropriate, it does not make any attempt



to identify an assignable cause possibly effecting the process. All EPC schemes do is react to the process upsets, but they do not make any effort to remove the assignable causes. Consequently, in processes where feedback control is used there may be substantial improvement if also control charts are applied for statistical process monitoring. The systems where both an EPC and an SPC tool for process monitoring have been implemented are often referred to as **Algorithmic Statistical Process Control (ASPC)**.

The control chart should be applied to either the difference between the control variable and its target (that is, the control error) or to the sequence of adjustments to the manipulated variable (that is, the control action). The monitoring of the control action is often useful because process performance already includes the impact of the control variables and, thus, fundamental changes that may be compensated for by ever-increasing control action may remain undetected for a long time. Points that lie outside the control limits on these charts would identify periods when the control errors are large or when large changes to the manipulated variable are being made. These periods would possibly be good opportunities to search for assignable causes.

The monitoring scheme of ASPC can serve a variety of purposes (e.g., see Tucker et al., 1993) including those of:

- (1) verifying identifications
- (2) determining whether the values of process or the model parameter values are varying
- (3) assisting the search for root causes.

Once the appropriate monitoring scheme has been agreed upon (which could be any of the SPC charts already discussed as the Shewhart type, EWMA or CUSUM control charts), one would use past data to determine the control limits. The ASPC implementation yields quality improvement both by removing sources of variability and by compensating for predictable process deviations from target. It revises the SPC dictum 'don't act without statistical evidence' to 'act on the statistical evidence'.

#### **Step 7: On-line implementation and assessment**

Implementation of the ASPC scheme requires much planning and should be led by the responsible process engineer. It is prudent to introduce





ASPC in stages, so as to build confidence in the results, assure that there will be no disasters and identify possible problems.

#### 6-4 The MMSE and PID controlled processes

The **Minimum-Mean-Squared-Error (MMSE)** controller is based on the idea of minimizing the mean square error of the process output deviations from the target. It is equivalent to the minimum squared error procedure used in order to specify the best forecast for a future value of the time series data discussed in Chapter 3. According to the control action and the model applied to the disturbance term, it takes different formulas. On the other hand, the **Proportional-Integral-Derivative (PID)** controller has a specific formula no matter what the process model is. In some cases the Minimum-Mean-Squared-Error is given by the PID controller, so that the two techniques coincide. Both controllers have interesting advantages as well as drawbacks and the choice between them should be made with caution.

##### 6-4.1 The I controller

The skeptic behind the PID controller is to find a formula that cancels out the disturbance term of the model. This formula is fixed and the only thing to be done is to estimate its parameters in order to find the optimal PID scheme. The I controller is one of the simplest forms of the PID controller. The analytical derivation of its formula is explicitly described by Montgomery (2001) and it is the following:

Suppose that the manipulated variable,  $u_t$ , has linear relationship with the deviation of the output characteristic from its target ( $Y_t$ ), so that:

$$Y_t = gu_{t-1} \quad (6-2)$$

where  $g$  is a constant usually called the process gain. If the disturbance term is modeled as an IMA(1,1) model of the form given by Eq(3-19), it can then be predicted adequately using the EWMA prediction as  $\hat{D}_t = \hat{D}_{t-1} + \lambda(D_{t-1} -$

$\hat{D}_{t-1}) = \hat{D}_{t-1} + \lambda e_{t-1}$ , with  $e_t = D_{t-1} - \hat{D}_{t-1}$  being the prediction error at time period  $t$  and  $0 < \lambda \leq 1$  is the weighting factor of the EWMA, where  $\lambda = 1 - \theta$  and  $\theta$  is the moving average parameter of the IMA(1,1) model. The adjusted process then taking into account both the disturbance term and the control action becomes:

$$Y_t = D_t + gu_{t-1} = e_t + \hat{D}_t + gu_{t-1}, \text{ since } e_t = D_t - \hat{D}_t \rightarrow D_t = e_t + \hat{D}_t \quad (6-3)$$

Eq (6-3) makes it obvious that by setting  $gu_{t-1} = -\hat{D}_t$ , which is equivalent to setting  $u_{t-1} = -(1/g)\hat{D}_t$ , the disturbance is cancelled out. Thus, the adjustment made at the time point  $t+1$  compared with the one made at the previous time point is:

$$u_{t+1} - u_t = -(\hat{D}_t - \hat{D}_{t-1})/g \quad (6-4)$$

The difference in the two EWMA predictions is written as:

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

Therefore, Eq (6-4) becomes:

$$u_{t+1} - u_t = -\lambda e_{t-1}/g = (-\lambda/g)e_{t-1} \quad (6-5)$$

The actual set point for the manipulated variable at the end of the period  $t+1$  is the sum of all the adjustments through time  $t+1$ , so that:

$$u_{t+1} = \sum_{j=1}^{t+1} (u_j - u_{j-1}) = (-\lambda/g) \sum_{j=1}^{t+1} e_j = k_I \sum_{j=1}^{t+1} e_j \quad (6-6)$$

Note that if the target value is 0, then the output can also be viewed as the deviation from target (i.e., the output error or control error). In fact the actual error at time  $t$  is the difference between the output and the target, i.e.,  $e_t = Y_t - \text{Target}$ , because by subtracting the measurement error  $e_t$  from the output variable  $Y_t$  in Eq(6-3), what is left is the prediction of the process.

Thus, from now on, by referring to the output we will implicitly refer to  $e_t$  and the control error will numerically equal the prediction error. The I controller is a pure feedback control scheme that sets the level of the manipulatable variable equal to a weighted sum of all current and previous process deviations from target as shown in Eq(6-6).

Another way to express the control equations that define this adjusting mechanism is given by Del Castillo (2002) as:

$$u_t = -\alpha_t / g,$$

$$\text{where } \alpha_t = \lambda(e_t - bu_{t-1}) + (1 - \lambda)\alpha_{t-1}, \quad 0 \leq \lambda \leq 1 \quad (6-7)$$

and  $b$  is an off-line estimate of the input-output gain  $g$

Because of this recursive form of the value  $\alpha_t$ , the I controller is often called the EWMA controller.

It can be shown that if the deterministic part of the process model  $Y_t = gu_{t-1}$  is correct and if the disturbance  $D_t$  is predicted perfectly apart from random error by an EWMA, this is the optimal control rule because it minimizes the mean-squared error of the process output deviations from target. In other words, if the dynamic model is  $Y_t = gu_{t-1}$ , the MMSE controller is given by the I controller.

#### 6-4.2 The PI controller

The PI controller, which is the most popular among the PID controllers, is derived if we consider that the adjustments to the process should take into account the two last errors instead of the last one as previously. That is,  $u_{t+1} - u_t = (c_1/g)e_{t+1} + [(c_1+c_2)/g]e_t = k_p e_{t+1} + k_I e_t$ . By summing this expression up, we get:

$$u_{t+1} = k_p e_{t+1} + k_I \sum_{j=1}^{t+1} e_j \quad (6-8)$$

In the case of the I controller,  $k_I$  was determined by the gain  $g$  (which is a known constant) and the EWMA parameter  $\lambda$ . For the PI controller, the



constants  $c$ 's (or  $k$ 's) should be chosen so as to minimize the mean-squared error around the target value.

In the above case, the dynamic model  $Y_t = gu_{t-1}$  has been considered, in which it is assumed that all of the change induced by a step change in  $u$  will occur in a single time interval. A more complicated but reasonable approach is to consider that a unit step change in  $u$  will have an effect on the output variable for more than one time period. Therefore, knowing that  $t$  time periods after a unit step change is made in  $u$ , the change in  $Y$  will be  $g(1-\delta^t)$ , where  $0 \leq \delta < 1$ . For this dynamic model the output change asymptotically approaches  $g$  units. The value of  $\delta$  measures the inertia in the process dynamics with  $\delta$  close to 0 corresponding to little or no inertia. For example, the first-order dynamic model that can approximate the behavior of a number of processes is characterized by the difference equation:

$$Y_t = \delta Y_{t-1} + g(1-\delta)u_{t-1} \quad (6-9)$$

The simplified dynamic model of Eq(6-2) corresponds to setting  $\delta=0$ .

Supposing now that the disturbance is represented by the nonstationary IMA(1,1) model and that the process dynamics is represented by Eq(6-9), it has been proven by Box and Kramer (1992) that PI adjustments of the form of Eq(6-8) produce MMSE about the target value provided that the proportional and integral constants  $k_p$  and  $k_i$  are set to the values:

$$k_p = \lambda\delta/[g(1-\delta)] \text{ and } k_i = \lambda/g \quad (6-10)$$

The control action specified from Eq(6-10) is of practical use only if  $\delta$  is fairly small. As  $\delta$  becomes larger and, in particular, as it approaches unity, the MMSE scheme requires excessive control action.

#### 6-4.3 The PID controller

The PID controller is the general form of this specific type of controllers and it is derived in the same way as the I controller with the

difference that the adjustments now depend on the last three random errors. Eq(6-11) gives the general formula of the PID controller.

$$u_t = k_p e_t + k_I \sum_{i=1}^t e_i + k_D (e_t - e_{t-1}) \quad (6-11)$$

$\downarrow$   
Proportional

$\downarrow$   
Integral

$\downarrow$   
Derivative term

where  $e_i$  s are the deviations of the output variable from its target value, and  $k_p$ ,  $k_I$ ,  $k_D$  are constants manipulated to minimize the resulting process variation. By omitting one or more of these terms, Eq (6-11) results in several special cases including those of the I and PI controllers described previously. That is, if  $k_p = k_D = 0$ ,  $k_I = (-\lambda/g)$  and  $t+1=t$ , then the I (integral ) controller is formed, though  $k_D = 0$  results in the PI controller, and if  $k_I = 0$  the PD controller is derived. Choosing the constants  $k$ 's or equivalently  $c$ 's is usually called *tuning the controller*.

When the disturbance term follows the IMA(1,1) model and the control action is linearly correlated with the output variable, then the MMSE is achieved with the PID controller, that is the MMSE controller reduces to the PID one.

#### 6-4.4 MMSE controllers for disturbance models other than the IMA(1,1)

1) If the disturbance term  $D_t$  is defined by an ARMA(1,1) process i.e.,

$$D_t = \phi D_{t-1} + \epsilon_t - \theta \epsilon_{t-1},$$

and the control action is defined as  $Y_t = -u_{t-1}$  (which is as setting  $g = -1$  to Eq(6-2), the model for both the dynamic behavior of the process and the disturbance effects becomes:  $Y_t = D_t - u_{t-1}$ . The MMSE controller in this case is defined by Box, Jenkins and Reinsel (1994) as:

$$u_t = \phi u_{t-1} + (\phi - \theta) Y_t \quad (6-12)$$

where  $Y_t$  is the output variable and  $u_t$  is the control action taken at the time period  $t$ . If the disturbance model is known, then  $Y_t = \epsilon_t$  and  $\sigma_Y = \sigma_\epsilon$ . Under



the MMSE control policy, it is not difficult to show, as it was done by Box, Jenkins and Reinsel (1994), that the standard deviation of the control action in this case equals:

$$\sigma_u = \frac{|\theta - \phi|}{\sqrt{1 - \phi^2}} \sigma_a \quad (6-13)$$

2) If the disturbance term is an autoregressive process of order 1, that is:

$$D_t = \phi D_{t-1} + \epsilon_t,$$

then it has been proven by Box and Jenkins (1976) that the sum of the AR(1) model of the disturbance and of the white noise term  $e_t$  in the general model:  $Y_t = u_{t-1} + D_t + e_t$  results in the ARMA(1,1) model. MacGregor (1990) proved that the control action of the MMSE controller in this case has the following form:

$$u_t = \phi u_{t-1} - (\phi - \theta) Y_t \quad (6-14)$$

## 6-5 Applying the feedback control scheme: an example

An application based on the EPC and APC schemes would make clear the need for their use. Suppose that the dynamic model of the process is given by Eq(6-9) with  $\delta=0.1$ , and that the disturbance model is the IMA(1,1) model. Then, if Eq(6-10) is used for this simulated data, then the MMSE controller in this case is the PI controller.

Montgomery (2001) showed that in order for  $\lambda$  to minimize the sum of the squared forecast errors for the process disturbance if the true optimum value for  $\lambda$  is  $\lambda_0$ , a value for  $\lambda$  in the 0.2-0.4 range does not inflate the variance of the output much. In contrast, if  $\lambda = 1$  (or equivalently  $\theta = 0$ ) in Eq(6-1), this implies that the adjustment made is exactly equal to the current deviation from target and this choice of  $\lambda$  doubles the output variance. On the other hand, when  $\lambda = 0$  (or  $\theta = 1$ ), this means that the process is in statistical

control and it will not drift off target, so that no adjustment is being done. Thus, a value of 0.2 for  $\lambda$  is a good approximation that works well in practice.

By setting  $\delta = 0.1$ ,  $\lambda = 0.2$  and  $g = 1.5$ , the values for  $k_p$  and  $k_I$  are calculated from Eq(6-10) as 0.015 and 0.13, respectively. The control actions are then calculated from Eq(6-8) as:

$$u_t = 0.015e_t + 0.13 \sum_{j=1}^t e_j \quad (6-15)$$

The model of the process under the dynamic model calculated by Eq(6-9) becomes:

$$Y_t = D_t + 0.1Y_{t-1} + 1.35u_{t-1} \quad (6-16)$$

where the disturbance model is the IMA(1,1) of Eq(3-19) with  $\theta = 0.4$ . The output for 493 observations of the process was derived under the model given by Eq(6-16) and the control actions were specified by Eq(6-15). At time  $t = 260$  a disturbance consisting of a sustained shift of magnitude 1 unit was introduced into the process. Figure 6-1 shows the control actions made using Eq(6-15) in which it is apparent the MMSE controller (or the PI more precisely) compensates for this assignable cause to a large degree, since the adjustment is larger after observation 260.

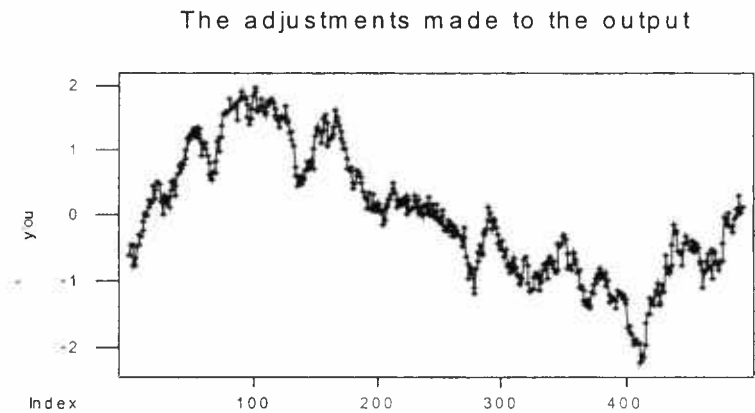


Figure 6-1: The control actions for the output of Eq(6-16).



Figure 6-2 shows the Shewhart and the EWMA control charts of the output deviations from target after the adjustments have been made. In both charts the shift in the mean is detected at observations 412 and 414 and there is also a downward trend after observation 260. The Shewhart control chart for the sequence of the adjustments seems to be in statistical control, as is shown in Figure 6-3.

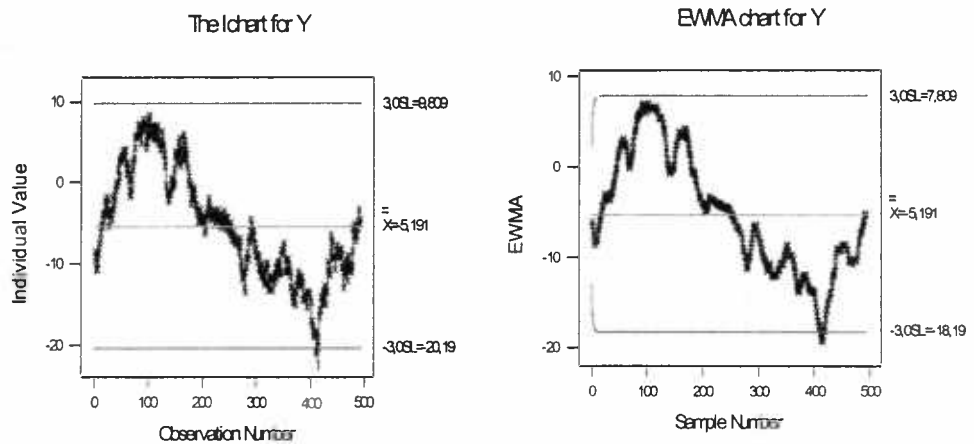


Figure 6-2: The Individuals and the EWMA charts for the output deviations from target.

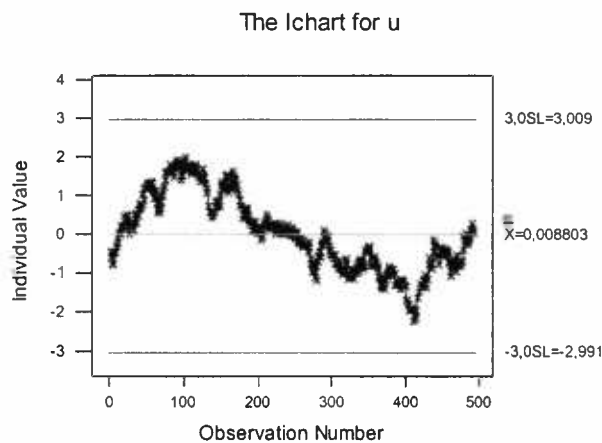


Figure 6-3: The Individuals chart for the adjustments (control actions).

The adjustments made to the output achieved to keep it close to target, though the application of simple SPC charts would reveal an out-of-control state with no indications of improvement. However, the application of SPC to the output deviations from target, after the adjustments have been made, helped to find the shift in the mean, something which would not be detected



with the EPC alone. This is how the combination of feedback control and control charts outperforms the use of one of these techniques alone.

## 6-6 Minimum cost adjustment: some simple schemes

The process adjustment schemes so far considered had the property of minimizing the mean squared error of the output quality characteristic about the target value  $T$ . Some forms of the MMSE controller, as is for example the control action of the I controller given by Eq (6-5), minimizes the output variance and at the same time the mean overall cost of adjustment *if it could be assumed that the cost of being off target was proportional to the square of the deviation from target and that other variable costs were negligible*.

In a case like this, the manual adjustment chart described by Montgomery (2001) could be used as in Figure 6-4. That is, an adjustment scale is added to the plot of the output deviations, so that by observing the current output deviation from target, the amount of adjustment to apply is written on a vertical scale. In the application of the previous section,  $k_I$  was  $\lambda/g=1/7.5$ , meaning that the divisions on the adjustment scale would be arranged so that one unit of adjustment equals 7.5 units on the output scale. Furthermore, the units on the adjustment scale that correspond to the output values above 0 (since the target value of 5 has been subtracted) are negative, whereas the units that correspond to output deviations from target below 0 are positive. This is naturally happening because when the output is above its target the manipulated variable should be reduced to drive the output toward the target and the opposite happens when the output tends to be directed below the target.

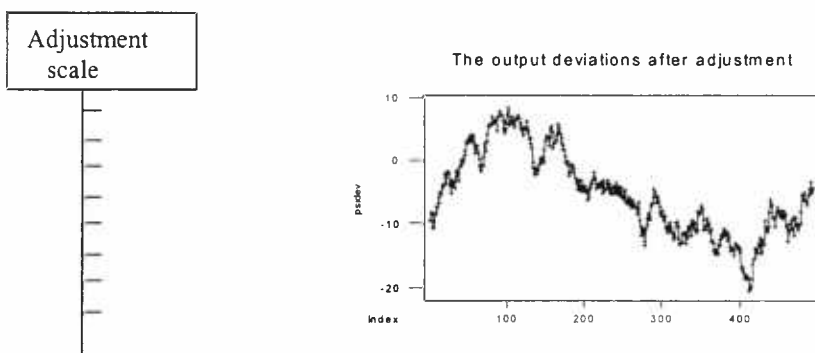


Figure 6-4: The manual adjustment chart.

If, however, other costs, such as that of adjusting the process or of taking an observation have also to be considered, then the minimum-cost feedback schemes become more complicated. If this is the case, Box and Kramer (1992) studied two possible situations: considering the sampling interval fixed or not fixed.

### 6-6.1 Sampling interval fixed

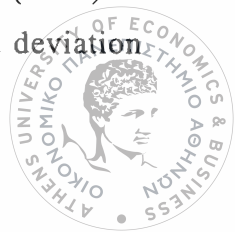
When the cost of observing the process is negligible, then the only existing costs are the cost of being off target for one time interval ( $C_T$ ), which is assumed by Box and Kramer (1992) as proportional to the square of the deviations from target, and the cost incurred by adjusting the process ( $C_A$ ), considered to be fixed. Therefore, it is of interest to make some modifications to the feedback adjustment procedure so that less frequent adjustments are made. The two most popular modifications are the bounded and the rounded adjustment charts.

#### 1) *The bounded adjustment chart*

This type of adjustment chart resembles superficially to an EWMA monitoring chart. An adjustment is made to the process only when an EWMA forecasted value of the output variable falls outside of the control limit lines  $T \pm L$  and not at each time point. After an adjustment is made, the forecasted value of the next observation is set to 0 and the following values are computed recursively. This type of adjustment is far more economic than is the case of continuous adjustments.

The boundary value is not determined, however, by questions of statistical significance but rather depend on the ratio  $C_A / C_T$  of the adjustment cost to the cost of being off target. Simple tables that allow the value of  $L$  to be calculated for given  $C_A / C_T$ ,  $\lambda$  and  $\sigma_\alpha$  (which is the amount of being off target) and that provide corresponding average run lengths between adjustments are given in Kramer (1989).

The appropriate choice of  $C_T$  that determines the off-target cost is sometimes difficult. An argument used, for example, by Taguchi (1981) is that, assuming that the cost is a quadratic function of the off-target deviation



with a minimum at  $T$ , to determine the whole cost curve we only need to know an additional point. To obtain such a point it is argued that as the deviation from target increases it will reach a point  $T \pm \Delta$ , say, at which the manufactured material must be discarded or reprocessed at a cost  $c_0$ . Then,  $C_T$  is calculated as:  $C_T = c_0 \sigma_\alpha^2 / \Delta^2$ .

## 2) The rounded adjustment chart

Box and Luceno (1997) discussed the rounded adjustment chart used to assist operating personnel in making simple adjustments. The adjustment scale is now rounded to a fixed number of zones (usually four or five) on either side of the target. Each zone corresponds to a specific adjustment (for example reduce or increase the manipulatable variable by 1 unit, 2 units etc). Usually the central zone corresponds to no adjustment at all.

## 6-6.2 Sampling interval not fixed

Kramer (1989) showed that when the cost of taking an observation is included, the minimal cost scheme is still of the form of the EWMA bounded adjustment chart but with a sampling interval of  $m$  units. To obtain minimum cost schemes when the cost incurred each time the process is observed ( $C_M$ ) is not negligible, it is often necessary to lengthen the sampling interval in comparison with schemes when the sampling cost is very low. Kramer (1989) gives a formula for calculating the overall cost when the cost of adjustment, the monitoring cost and the cost of being off target are included. By minimizing this overall cost, it is possible to determine:

- (a) when the process should be adjusted
- (b) what size of adjustment should be made
- (c) how often the process should be sampled and the data should be collected
- (d) the average interval between adjustments (based on the ARLs) associated with each scheme.

Box and Cramer (1992) presented a graph from which one can find immediately the value for  $m$  to use when the disturbance term is modeled as an IMA(1,1) process but this chart requires values for three quantities: the

nonstationarity measure of the IMA disturbance process  $\lambda$  [given by Eq(3-19) if  $\theta$  is replaced by  $1 - \lambda$ ] and the two ratios  $R_A$  and  $R_M$ , where:

$$R_A = (C_A / C_T) / \lambda^2 \text{ and } R_M = (C_M / C_T) / \lambda^2 \quad (6-17)$$

If the limit lines yielding minimum overall cost are set at  $T \pm L$  with  $L = l\lambda\sigma_\alpha$  where  $\lambda$  and  $\sigma_\alpha$  are the parameters of the original disturbance process monitored at unit intervals, Box and Cramer (1992) provided also a chart for finding the appropriate value for  $l$  in conjunction with  $\lambda$ ,  $R_A$  and  $R_M$ . The two charts for finding appropriate values for  $l$  and  $m$  have resulted in some special cases when the sampling interval is not fixed. Some of these cases are:

- $\lambda = 0$ . When this is the case, it is derived from Eq(3-19) that the disturbance is a white-noise stationary process. The standardized limit and the monitoring interval  $m$  are then both infinite according to the relevant chart of Box and Cramer (1992) for finding  $m$ . Thus, the control action to be taken for a process known to be in a perfect state of control is no action at all.
- $C_A$  negligible. In this case, for any fixed  $m$ ,  $l$  tends to 0 and the limits  $T \pm L$  converge on the target value. Adjustments must, therefore, be made as each new value becomes available. Each adjustment is made to cancel the deviation of the exponentially smoothed value from the target value and the total adjustment at time  $t$  is then proved to be the I controller. Notice that the I controller provides minimum cost of regulation only when the cost of adjustment  $C_A$  is negligible.
- $C_A$  is negligible,  $C_M$  is not and  $m$  is not fixed. In this case, a feedback scheme is obtained in which limit lines are on the target ( $l=0$ ), so that adjustments are made after each observation, but these observations may be made less frequently.
- $C_A$  is not negligible,  $m$  is fixed. If this is the case, the action limits  $T \pm l\lambda\sigma_\alpha$  adjustments are determined directly by  $R_A$ .
- $\lambda = 1$ . This random-walk case could theoretically occur and only in this case could adjustment action based on  $z_t$  rather than  $\hat{Z}_t$  be justified. However, this degree of nonstationarity is so extreme, that it can hardly be met in practice.

## 6-7 Other factors inciting the use of ASPC

Considering the advantages of the ASPC technique discussed extensively, Faltin et al. (1993) remarked that the use of this method has better results if:

1) Measurement variability is modest relative to process variability, that is ASPC works best when the 'signal-to-noise' ratio is high, meaning that the magnitude or true process variability is large relative to measurement error.

2) The impact of the control variables is understood because there is a good physical understanding of the control variables and their impact on performance.

3) Performance measurements are obtained in a timely manner.

Although the measurements of some process parameters are obtained continuously, there is often a lengthy delay in the time required to obtain measurements of quality. The effectiveness of process adjustments to compensate for performance variability is comprised as the delay time increases. To reduce such delays surrogate measurements are sometimes used. Measurement delay is discussed in more detail in the following chapter.





## CHAPTER 7

# Automatic Statistical Process Control (ASPC): more special issues

### 7-1 Introduction

In the previous chapter, the design of the EPC/ASPC technique was discussed and the most popular controllers were described. In this chapter, our attention is focussed on some more specialized aspects of the feedback control adjustment, as is the on-line process control used not only to bring the process close to target so as to differentiate between the common and special causes, but to adjust the process between the collection of the data set. This type of controller is called the “Run-by-Run” controller and it is discussed in section 7-2. A design map evaluated to derive easily the unknown constants of the PID controller is presented in section 7-3. Another issue is the fact that often the previous measurement(s) are not available for estimating the process output at the next time point. This is confronted in section 7-4, while the use of an economic model applied to MMSE controllers is the subject of section 7-5. Section 7-6 is referred to some drawbacks of the ASPC approach, while section 7-7 presents a control chart based on the PID scheme and, therefore, called the PID chart.

### 7-2 The Run-by-Run controller

**The Run-by-Run (RbR)** controller is a group of algorithms designed to be used for on-line process control, that is, control of a process during production. It responds to post-process by updating models of the process between runs (instead of during a run) and providing a new recipe for use in

the next (or a subsequent) run of the process. The recipe itself may include changes in set points during a run. The RbR controller does not, however, modify the recipe during a run based on measurements made while the process is running, as is done by the *Real Time controller*. Its objective is to reject various disturbances frequently found in RbR processes, such as shifts and trends, as well as autocorrelated disturbances.

The RbR controller has two modes of operation: optimization and control. The distinction between the two modes is that in the optimization mode, it is expected that the process can be significantly improved, while in the control mode, the concern is to maintain the performance of the process in the face of disturbances. Thus, although the optimization mode temporarily increases the variability of the process output by exploring the process space in order to improve it, the goal of the control mode is to reduce this variability and keep the process on a target value.

Sachs, Hu and Ingolfsson (1995) studied the RbR controller thoroughly, assuming that the dynamics of the process is captured by the relation:

$$y_t = \alpha_t + \beta_t x_t + \epsilon_t,$$

where  $x_t$  is the controllable input variable,  $y_t$  is the output variable,  $\epsilon_t$  is the random error term with variance  $\sigma^2$  and the parameters  $\alpha_t$  and  $\beta_t$  may be random variables. The intercept term  $\alpha_t$  and the process sensitivity  $\beta_t$  may change with time. The appropriate prediction equation is then:

$$\hat{y}_t = \alpha_{t-1} + b_{t-1} x_t \quad (7-1)$$

which is used to select a recipe for the next run at which the process output is likely to be close to a target  $T$ , i.e., a recipe  $x_t$  satisfying  $\alpha_{t-1} + b_{t-1} x_t = T$ . The values  $\alpha_{t-1}$  and  $b_{t-1} x_t$  in the prediction Eq(7-1) are estimates of the parameters  $\alpha_t$  and  $\beta_t$ . Sachs, Hu and Ingolfsson (1995) assumed that the process outputs are measured for every run and that the process sensitivities  $b_0$  stay constant over time, so only the intercept term  $\alpha_{t-1}$  is updated each time an output measurement  $y_t$  becomes available.

Concerning the disturbance term, two generic situations are studied by the authors:



(a) *Gradual mode*: The process is drifting slowly (on the order of about  $1\sigma$ ). For this slow drifting process, with no radical departure from the predicted process behavior having been signaled, the disturbance is smaller than the noise in order of magnitude in the process and the process model is updated gradually.

(b) *Rapid mode*: The process is subject to occasional, large shifts on the order of  $2\sigma$  or larger. If this rapid mode influences the process, then the output measurements will be in significant disagreement with their predicted values and will signal an alarm. The process mode has then to be updated rapidly to allow the process to quickly adapt to the disturbance and return the output to target.

The guidelines used is to preliminarily apply a control chart to the difference between the prediction of the model given by Eq(7-1) and the measured value. Since the RbR controller functions by making small changes to the input parameters, so as to keep the process closer to target, the adaptation of the model that takes place in the gradual mode accomodates this slow shift. Since the gradual model adaptation does not remove the effect of a rapid shift, a control chart can be used to distinguish between slow drifts and rapid shifts. In the RbR controller it is expected that the gradual mode will remove the effect of small changes and, therefore, the suggested control chart is the Shewhart, since our interest is now on detecting large shifts in the process.

### 7-2.1 The Gradual mode in the RbR controller

As has been mentioned previously, the purpose of the gradual mode is to compensate for drift in the process by gradually updating a model for the process and prescribing a corrective action based on that model. In order to control the process, we use our current model of the process to predict what the output for the next run will be as a function of the input for that run and select an input value for which the predicted output is on target. Since the estimates are not perfect, the recipe chosen by the controller may not be the ideal one and the actual output value may be different from the predicted one. But if the slope estimate, assumed to be constant, is ‘good enough’ and the

intercept estimate is updated in an intelligent way, the input settings will converge to the ideal recipe (i.e., control action) and the output values will converge to target.

#### The algorithm

When the process is subject to noise and data from  $t$  runs are available, it is reasonable to compute the estimate  $\alpha_t$  as some weighted average of the numbers  $(y_1 - bx_1)$ ,  $(y_2 - bx_2)$ , ...,  $(y_t - bx_t)$ . The best choice is to use the Exponentially Weighted Moving Average (EWMA) so that the weight decays gradually with age in the geometric fashion. If  $\lambda$  is the weight assigned to the most recent data point, then the EWMA may be expressed in a recursive form. The gradual mode algorithm for a single-input-single-output process is then defined by the following two recursive relations given by Sachs, Hu and Ingolfsson (1995):

$$\begin{aligned} x_t &= (T - \alpha_{t-1}) / b & \text{and} \\ \alpha_t &= \lambda(y_t - bx_t) + (1 - \lambda) \alpha_{t-1} \end{aligned} \quad (7-2)$$

The first equation in (7-2) specifies how the recipe for the  $t^{\text{th}}$  run should be selected and the second one describes how the intercept estimate should be estimated using the EWMA property after the output from run  $t$  has been measured. In the case where the controller is applied after every run and the full prescribed control action is taken, the intercept estimate and the control action are given by:

$$\begin{aligned} \alpha_t &= \lambda(y_t - T) + \alpha_{t-1} \\ x_t &= -(\lambda/b) \sum_{i=1}^{t-1} (y_i - T) + x_1 \end{aligned} \quad (7-3)$$

It can be easily seen that the EWMA controller of Eq(7-3) is the Integral (I) controller given by Eq(6-6) with a measurement delay of one run. In order to ensure that the sequence of recipes suggested by the algorithm will eventually converge to the new ideal recipe when the process is single-input-single-output, that it is subject to uncorrelated noise and drifts an amount  $\delta$

between successive runs, Sachs, Hu and Ingolfsson (1995) came up with the condition  $0 < \lambda\beta/b < 2$  that guarantees the controlled process is stable. This condition places a restriction on how poor the estimate of the process sensitivity can be. The first inequality implies that the estimated sensitivity  $b$  must have the same sign as the true sensitivity  $\beta$ . The second inequality ensures that the magnitude of  $b$  is at least  $\lambda\beta/2$ .

Another matter we have to check is the performance of the controller when the process is operating in control. It is known that if the output from a process is not correlated, then any control action will increase the process variability and the best action is to leave the process alone. Sachs et al. (1995) checked that with an EWMA weight of 0.1, the application of the EWMA controller to a non-autocorrelated process will result in a small increase of the standard deviation of the process. However, if a slow drift is present, the uncontrolled process will drift off target, while the controlled process will stay close to target. As a practical matter, there is a wide range of EWMA weights where the controller effectively compensates for drift with minimal negative impact for the case when no drift is present.

The authors have also considered the case of multiple inputs which is an extension of the one with a single input just described.

### 7-2.2 The Rapid mode in the RbR controller

After the SPC mode has signaled a large shift, or a shift has occurred after specification changes or after maintenance operations, the rapid mode must be able to prescribe proper control action for the process. The most significant feature of this kind of disturbance is that it changes the process level by a large amount in a small number of runs since gradual mode was not able to compensate for it. According to Sachs et al. (1995) the compensation done for large, occasional shifts is more appropriate when the RbR controller:

- 1) estimates the magnitude and location of the disturbance
- 2) assesses sequentially the probability that a step of the magnitude and location estimated in 1) actually took place
- 3) uses the estimations from steps 1) and 2) to prescribe control actions.

1) *Estimating the magnitude and location of the disturbance*

The procedure of the estimation is illustrated in Figure 7-1. The data points are the output measurements, adjusted for the effect of the input variable. The two horizontal lines are fitted to the data so as to minimize the sum of the squared deviations of the data from the lines. The position of the breakpoint provides an estimate of the location of the shift disturbance and the distance between the lines is an estimate of its magnitude.

Let  $z_t = y_t - bx_t$  be the output measurements, adjusted for the effect of the process parameters. In the case of approximating the change by adapting the intercept only, if the process drift is slow enough to make the estimated intercept term  $\alpha_t$  change little during the  $k$  runs, then  $z_t$  can be approximated by:

$$z_t \approx \alpha + e_t, \text{ where } e_t = y_t - T \quad (7-4)$$

If the gradual mode of the RbR controller performs its job adequately, then, in the absence of shifts, the deviations from target  $e_t$  will be iid with mean 0 and variance  $\sigma^2$ . If a step of magnitude  $d$  occurred between runs  $t-m$  and  $t-m+1$ , then:

$$\begin{aligned} z_t &\approx \alpha + e_i, & \text{for } i = t-k+1, \dots, t-m \\ &\approx \alpha + d + e_i, & \text{for } i = t-m+1, \dots, t. \end{aligned}$$

Based on these observations, the procedure to estimate the magnitude and location of the shift is to minimize the sum of squared deviations of the  $z_t$ 's from estimates of the process intercept  $\hat{z}_t$ , i.e.,  $\sum_{i=t-k+1}^t (z_i - \hat{z}_t)^2$ . This rapid mode algorithm will give both an estimate of the shift magnitude and location.

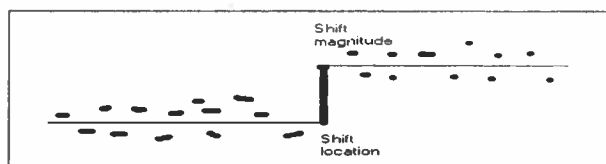


Figure 7-1: The effect of a large shift in the process

## 2) *Assessing the probability of a shift*

The general framework described by Sachs, Hu and Ingolfsson (1995) for deciding whether the disturbance actually did occur is the use of a sequential Bayesian statistics approach if the estimates from the previous step for the shift magnitude and location are  $d$  and  $m$ , respectively.

If there was indeed a change in the level of the process, then with the accumulation of supporting data the shift probability  $f_t$  will tend to increase towards 1. On the other hand, if the alarm was due to a random fluctuation, the subsequent data would discount the probability of the shift, so that the value of  $f_t$  decreases towards 0. The authors symbolize as  $Z_{m+}$  the data derived after the possible shift from runs  $t-m+1, \dots, t$ . According to Baye's rule, the probability  $f_t$  at run  $t$  that a shift with magnitude  $d$  occurred between run  $t-m$  and run  $t-m+1$  is:

$$f_t = P\{\text{shift with magnitude } d \text{ occurred } m \text{ runs ago} \mid Z_{m+}\} = \\ = P\{Z_{m+} \mid \text{shift}\} f_{t-m} / [P\{Z_{m+} \mid \text{shift}\} f_{t-m} + P\{Z_{m+} \mid \text{no shift}\} (1 - f_{t-m})],$$

where  $f_{t-m}$  is the shift probability before the data set  $Z_{m+}$  is available.

We can assume a prior probability distribution for the shift probability,  $p(f \mid Z_{m-})$ , where  $Z_{m-}$  is the data available before the shift. Again using Baye's rule, we can obtain an expression for the sequential nature of updating the probability distribution for  $F_t$ :

$$p(f_t \mid Z_t) = h P(Z_{m+} \mid F_{t-m} = f) p(f_{t-m} \mid Z_{m-}),$$

where  $Z_t$  is the complete set of data,  $p(f_{t-m} \mid Z_{m-})$  is the prior probability distribution for  $f_{t-m}$ ,  $p(f_t \mid Z_t)$  is the posterior probability distribution for  $f_t$ ,  $P(Z_{m+} \mid F_{t-m} = f)$  is the likelihood function incorporating the information on shift probability from the data and  $h$  is the normalization constant.

One advantage of the Bayesian approach is that it elicits from the user explicitly what are the assumptions for the parameter of interest by having the user specify the prior distribution. Another is that since the parameter of interest is updated with each data point, the newly obtained posterior distribution always reflects the information from the latest data. A discrete and a non-discrete case for the prior distribution of the shift are described in detail by Sachs, Hu and Ingolfsson (1995).

### 3) *Compensation strategy*

The next and final task of the rapid mode is to decide how to compensate for the change in the output. The criterion used is to minimize the sum of expected squared deviations of the output from target on all the runs after the application of the rapid mode algorithm. Currently, the algorithm minimizes the expected squared deviations from the target for the next run. The minimization of all runs is performed in order to:

$$\text{minimize } E[(y_{t+1} - T)^2 | Z_t.$$

Assuming that only the intercept term changes, the minimization is to be performed with respect to the amounts by which the estimated intercept term has changed. The expected magnitude of a shift is:

$E[F_t]d + (1 - E[F_t])0 = E[F_t]d$ . This is the amount by which the intercept term is adjusted in the initial version of the RbR controller. The adjustments in the process parameters (inputs) are determined by solving a model where the constant term has changed by an amount  $E[F_t]d$ . However, when process noise is large in comparison to the changes in the slopes, the compensation factor will become independent of the shift probability.

### 7-3 Design maps for the PID controller

Tsung and Shi (1999) worked on the estimated parameters of the PID controller, used in an RbR control process, and they ended up with some maps that direct us in the choice of the constants  $k_p$ ,  $k_I$  and  $k_D$  when the disturbance model is an ARMA(1,1) and the control action has its full effect on the process output in one run (that is, the pure dynamics model is valid). The optimal PID controller is the minimum process variability under the constrain that the PID control strategy has been used.

The authors derived the relationship between the PID control parameters and the process minimum-squared-error of the ARMA(1,1) model. Then, based on the optimization index, the design of the PID parameters  $k_p$ ,  $k_I$  and  $k_D$  was obtained by minimizing the squared-error within the stability region:



$$(k_p, k_I, k_D) = \begin{cases} k_I \geq 0, \\ k_p + k_I / 2 + 2k_D < 1, \\ -1 < k_D < 1, \\ -k_D(1 + k_p + k_I) - k_p < 1 \end{cases} \quad (7-5)$$

Tsung and Shi (1999) represented the choices of  $k_p$ ,  $k_I$  and  $k_D$  in a two-dimensional parameter space, using contour plots. The contour plots are graphic representations with lines (as on a map) connecting the points on a response surface that have the same elevation or response value. In this case, the vertical axis is the disturbance parameter  $\phi$ , the horizontal axis is the disturbance parameter  $\theta$  and the response value is the corresponding control parameter, which is labeled on each contour line. The PID parameters and their relationships, when the disturbance model is of ARMA(1,1) type and the pure dynamics model is appropriate, are summarized in PID design maps (Tsung and Shi, 1999).

As an example, if  $\phi=0.86$  and  $\theta=0.16$  are the parameters of a disturbance model, then the PID parameters are obtained directly from the design maps as  $k_p = 0.24$ ,  $k_I = 0.58$  and  $k_D = -0.1$ . Some obvious remarks concerning the design maps are:

(a) If  $\phi \leq 0.5$  or if  $\phi \leq \theta$ , the value of  $k_I$  is zero, that is Integral (I) control action is not needed. Thus, in this region PID control coincides with PD control.

(b) Near the region  $\phi = \theta$ , the choices of  $k_p$ ,  $k_I$  and  $k_D$  are close to zero. This is because in this case the ARMA(1,1) model becomes a white noise process, so the best control strategy is to not adjust the process.

#### 7-4 MMSE control when measurements are delayed

If a measurement is delayed, that is the preceding measurement is not available, then the formulas for the control action cannot be applied but it is possible to minimize the output MSE with respect to the available data. An example is given by Vander Wiel et al. (1992) who studied that if the model of the process is specified by  $Y_t = u_{t-1} + D_t$ , with  $D_t$  following the AR(1)



model, and, thus, the control action used is the one of Eq(6-14), then because of the measurement delay, the model is formatted as:

$$\hat{y}_{t|t-i} = u_{t-1} + \hat{D}_{t|t-i},$$

where  $\hat{y}_{t|t-i}$  is the  $i$ -step MMSE forecast of  $Y_t$  and  $\hat{D}_{t|t-i}$  is the  $i$ -step MMSE forecast of the model noise term of AR(1). By certainty equivalence the  $i$ -step MMSE feedback rule is found by setting  $\hat{y}_{t|t-i} = 0$  resulting in  $u_{t-1} = -\hat{D}_{t|t-i}$ .

When  $Y_{t-1}$  becomes available, the update equation for  $\hat{D}_{t|t-i}$  is:

$$\hat{D}_{t|t-1} = \theta \hat{D}_{t-1|t-2} + (\phi - \theta)(Y_{t-1} - u_{t-2}).$$

When needed for two-step adjustments,  $\hat{D}_{t|t-2}$  can be computed from:

$$\hat{D}_{t|t-2} = \phi \hat{D}_{t-1|t-2}.$$

In the absence of laboratory delays, this procedure is equivalent to Eq(6-14). Importantly, adjusting to negate the forecast of  $D_t$  minimizes the output MSE in period  $d$  regardless of what control policy was used during the previous periods. For example, if the process had been adjusted using a two-step MMSE rule through period  $t-1$ , but beginning in period  $t$  laboratory measurements are no longer delayed, then simply switching to the one-step rule of Eq(6-14) will not minimize the output MSE in period  $t$ . The derivation of Eq(6-14) assumes that the rule will be used in each period. Using the intermediate quantity  $\hat{D}_{t|t-i}$  avoids this difficulty.

### 7-5 An economic model for monitoring MMSE-controlled processes

Assuming that the output of the system (with the system inertia being one-step delay) can be expressed as  $e_t = D_t - u_{t-1}$ , where  $D_t$  is the disturbance and  $u_{t-1}$  the control action, Jiang and Tsui (2000) derived an economic model when the feedback control is a Minimum Mean Square Error (MMSE) control scheme. The authors extended the formulation of economic models for monitoring continuous time production processes described in Montgomery (2001), by defining a production cycle for discrete time processes.

Supposing that the production process travels between only two states, the in-control and out-of-control states, the time interval from the beginning



of the in-control state to the adjustment of the out-of-control state is called a **production cycle**. It is assumed that the shift occurs only once in each cycle with no other shift occurring before the previous shift is detected and removed and that the process is not self-correcting, that is, once a transition to an out-of-control state has occurred, the process may return to the in-control condition only by external adjustment.

Jiang and Tsui (2000) subgrouped the costs associated with a production cycle into four categories:

- (i) the sampling and testing cost associated with the subgroup size
- (ii) the diagnosis cost  $C_D$  associated with identifying special causes from out-of-control signals
- (iii) the adjustment cost  $C_A$  associated with correction of special causes, and
- (iv) the cost associated with production of nonconforming items.

If the sampling size and frequency are fixed, the first category of cost is not considered and this path was followed by the authors. For the cost of nonconforming items, the quadratic quality loss function was used, i.e.,  $L(y,t) = A_0(\sigma^2 + \mu_t^2)$ , where  $A_0$  is the cost coefficient and  $\sigma^2$  is the variance of the white noise. It follows that the total cost of a production cycle (denoted as the total quality cost,  $L_T$ ) consists of two parts: the in-control ( $L_{in}$ ) and the out-of-control cost ( $L_{out}$ ), i.e.,

$L_T = L_{in} + L_{out},$	$\text{where}$	
$L_{in} = (1/p)A_0\sigma^2 + (1/p)aC_D =$		
$\Downarrow$	$\Downarrow$	
$\text{quadratic loss of}$	$+$	$\text{expected diagnosis cost}$
$\text{the in-control items}$		$\text{for false-alarms,}$
		$\text{and}$
$L_{out} = \sum_{T=1}^{\infty} \left[ \sum_{t=0}^{T-1} A_0(\sigma^2 + \mu_t^2) \right] P_T(\mu) + C_D + C_A \quad (7-6)$		



where  $a = 1/ARL_0$ ,  $T$  is the out-of-control run length and  $P_T(\mu)$  is its probability mass function (pmf). Obviously,  $\sum_{T=1}^{\infty} P_T(\mu) = 1$  and the in-control and out-of-control run lengths, respectively, are:  $ARL_0 = 1/a$ , and  $ARL_1 = \sum_{T=1}^{\infty} TP_T(\mu)$ .

Averaging the total quality cost over the entire production cycle, Jiang and Tsui (2000) came up with the Average Quality Cost (AQC) as:

$$L_A \equiv L_T / [(1/p) + ARL_1] = (\sigma^2 + aC_D) + \frac{\sum_{T=1}^{\infty} P_T(\mu) \sum_{i=0}^{T-1} \mu_i^2 - [ARL_1 \times a - 1]C_D}{(1/p) + ARL_1} \quad (7-7)$$

It has been mentioned that different SPC charts can be applied to monitor the MMSE-controlled processes, and criteria are needed to choose the appropriate charts. Traditionally, the Average Run Length (ARL) has been commonly used for such purposes. However, when the process mean is dynamic, the ARL may not be complete as it does not take into account the run length variation. On the other hand, the Average Quality Cost (AQC) of Eq(7-7) can serve as a good alternative since it considers the run length variation together with the dynamic nature of the output mean shift pattern. Comparisons of the AQC and the ARL values for MMSE-controlled processes with AR(1) and ARMA(1,1) disturbances are provided in Chapter 8.

## 7-6 Criticisms concerning the ASPC rule

SPC practitioners have sometimes criticized feedback controllers for:

- (a) overcompensating disturbances
- (b) compensating disturbances rather than removing them
- (c) concealing information that might be used for quality improvement.

Box and Kramer (1992) responded adequately to these criticisms by insisting that about issue (a) it is the occasional misapplication of ASPC that

has resulted in overcompensation and not inherent problems with the technique itself. If the controller is of the right design but is mistuned or if the design is not appropriate, then the adaptation will not be very successful but this is the case with every method in the statistical field.

In terms of (b), it is well known that disturbances may not be eliminated in all cases and, thus, the adjustment is the only remaining solution. With reference to (c), the authors support that the feedback adjustment does not have to conceal important features of the process if one does not want to. The disturbance model and the dynamic model, that together define the common-cause system, are taken into account when designing the controller and could be changed by management.

The superiority of the ASPC technique over robust and adaptive approaches has been discussed by Tucker, Faltin and Vander Wiel (1993). Naturally, from time to time, many authors initiate new methods by trying to outperform the ASPC mode or at least reduce some of its drawbacks. Tucker, Faltin and Vander Wiel (1993) referred to the results of some studies about the relative performance of the ASPC compared to other elaborated rules.

Harris and MacGregor (1987) built a set of control equations in their effort to compensate adequately for process/model mismatch, i.e., to form a robust method. ASPC can, however, result in better process performance than the approach of Harris-MacGregor because the last one tends to average over the uncertainties for which ASPC can properly correct.

The application of *adaptive Bayesian methods* has the double advantage of both probing the system for information and driving the outputs to their target values. As a matter of fact, though, when reidentification or reestimation was needed, the periods of probing were similar for the Bayesian methods and the ASPC.

The *non-adaptive rules* may be either self-optimizing (that is, their control rule achieves asymptotically the same optimum performance as when parameters are known) or self-tuning (if their parameter estimates converge to values that result in an optimal control rule). Even in this case, Tucker, Faltin and Vander Wiel (1993) support that the ASPC performs better than non-adaptive controllers in the long run.

### 7-7 The Proportional Integral Derivative (PID) chart

In the same way that the MMSE predictor is closely tied to the corresponding MMSE scheme in feedback control problems, Jiang et al. (2002) used an analogous relationship between PID control and the corresponding PID predictor to propose a new class of procedures for process monitoring. As in SCC charts, they transformed the autocorrelated data to a set of residuals by subtracting the PID predictor and monitoring the residuals. Jiang et al. (2002) proved that the PID predictor corresponding to the PID control is given by:

$$\hat{D}_{t+1} = \hat{D}_t + k_I e_t + k_p(1-B)e_t + k_D(1-B)^2 e_t \quad (7-8)$$

where  $B$  is the backward shift operator defined in Chapter 3 as:  $Be_t = e_{t-1}$ . The PID predictor, as the PID controller specified in Eq(6-11), has only three terms since the prediction update is based on the three most recent terms. When  $k_D = 0$  in Eq(7-8), the PI predictor is derived, which corresponds to the Proportional Integral control scheme commonly used in industry.

The PID charts are obtained by subtracting the PID predictor specified from Eq(7-8) from the original data to yield the PID-based residuals and monitoring the residuals. Because the residuals are somewhat correlated, we must take into account the correlation structure when computing the control limits and then any of the traditional approaches such as the Shewhart, CUSUM or EWMA can be used to monitor the residuals. When the monitored disturbance process  $D_t$  is stationary, it is reasonable to require  $e_t$  to be stationary too, because otherwise the charting process  $e_t$  would drift even when the monitored process has no shift. This requirement holds if the conditions for the PID parameters ( $k_p$ ,  $k_I$ ,  $k_D$ ) proposed by Tsung and Tsui (1999) and given by Eq(7-5) are satisfied.

Because  $e_t = D_t - \hat{D}_t$ , we have  $e_t - e_{t-1} = (D_t - D_{t-1}) - (\hat{D}_t - \hat{D}_{t-1})$  and Eq(7-8) can also be written as:

$$e_t = (1-k_I)e_{t-1} - k_p(1-B)e_{t-1} - k_D(1-B)^2 e_{t-1} + (D_t - D_{t-1}) \quad (7-9)$$

Special cases of the PID charts arise by setting appropriate chart parameters to 0. These coincide with some well-known control charts in the literature:

1. The I chart has  $k_p = k_D = 0$ . By Eq(7-9),  $\hat{D}_t = D_t - e_t = D_{t-1} - (1 - k_I)e_{t-1} = k_I D_{t-1} + (1 - k_I) \hat{D}_{t-1}$ , that is,  $\hat{D}_t$  is an EWMA predictor of  $D_t$ . Since the I predictor is the well-known EWMA predictor, consequently, the I chart is the same as the M-M chart.

2. The P chart has  $k_I = k_D = 0$ , so Eq(7-9) becomes  $e_t = -k_p e_{t-1} + D_t$ . Hence,  $e_t = (1 + k_p B)^{-1} D_t = D_t + (-k_p) D_{t-1} + (-k_p)^2 D_{t-2} + \dots = \tilde{D}_t / \lambda$ , where  $\lambda = 1 + k_p$  and  $\tilde{D}_t = [D_t + (1 - \lambda) D_{t-1} + (1 - \lambda)^2 D_{t-2} + \dots]$  is an EWMA of  $D_t$ . In other words, when the Shewhart chart is applied to the P chart with  $-1 < k_p \leq 0$ , this is equivalent to the EWMAST chart provided that  $\lambda = 1 + k_p$ .

3. When  $D_t$  is an iid process, the EWMAST chart becomes the EWMA chart, so the P chart is then equivalent to the EWMA chart.

Note that there is no connection between the EWMA chart and the EWMA predictor. The I control leads to the EWMA predictor, though the EWMA prediction-based chart is the I chart (i.e., the M-M chart).

#### Choosing the parameters of the PID chart

Following the signal-to-noise (SN) ratios introduced by Jiang et al. (2000) so as to choose the appropriate parameters of their ARMA chart, as explained in detail in section 4-5, Jiang et al. (2002) also proposed to choose the parameters of the PID chart using the transient capability  $R_T = \mu_T / \sigma_e$  and the steady-state capability  $R_S = \mu_S / \sigma_e$ , where now  $\sigma_e$  is the variance of the charting process  $e_t$  when the monitored process  $D_t$  is in control. For the PID chart with a mean shift of  $\mu$  at time  $t_0$  (that is, the process is  $D_t$  for  $t < t_0$  and becomes  $\mu + D_t$  for  $t \geq t_0$ ), the ratios take the following form:

$$\begin{aligned} R_T &= \mu / \sigma_e \text{ and} \\ R_S &= \mu / [\sigma_e (1 + k_p)] = R_T / (1 + k_p), \text{ if } k_I = 0, \\ &= 0, \text{ if } k_I > 0 \end{aligned} \quad (7-10)$$



We note that the D term ( $k_D$ ) does not affect the values of  $\mu_T$  and  $\mu_S$ . The transient capability  $R_T$  measures the chart's capability to detect the shift in the first few runs and is more appropriate for large shifts. If the chart fails to signal early, then the steady-state capability,  $R_S$ , becomes important for detecting the shift efficiently in later runs. Because  $R_S = 0$  for  $k_I > 0$ , the PID chart with  $k_I > 0$  (for example, the I or M-M chart) is generally not good for detecting small shifts. According to the heuristic rules proposed by Jiang et al. (2000), the same algorithm may be adapted to guide us towards the appropriate parameters for the PID chart. This is:

1. Specify the shift level  $\mu = k\sigma_D$  to be detected.
2. Compute  $\{\max R_T\}$ , the maximum value of  $R_T$  for the PID chart, by varying its parameters ( $k_p, k_I, k_D$ ).
3. If  $\{\max R_T\} > 5$ , then choose the PID chart with the transient capability equal to  $\{\max R_T\}$  and stop, otherwise go to step 4.
4. Compute  $\{\max R_S\}$ , the maximum value of  $R_S$  for the PD chart, by varying its parameters ( $k_p, 0, k_D$ ).
5. If  $\{\max R_S\} \leq 3.5$  or if  $R_T \geq 1$  when  $R_S$  is maximized, then choose the PD chart with the steady-state capability equal to  $\{\max R_S\}$ , otherwise choose a PD chart with  $R_S \in [2.5, 3.5]$  to balance the values of  $R_T$  and  $R_S$ .

Practically,  $\{\max R_T\}$  in step 2 is often obtained from a PI chart. Furthermore, there do not exist closed-form expressions for  $\{\max R_T\}$  and  $\{\max R_S\}$  and the maximization is done with numerical methods.



## CHAPTER 8

# Performance of the Automatic Statistical Process Control

### 8-1 Introduction

After having studied the advantages of the EPC/ASPC technique and provided analytical discussion about this method, it would be interesting to investigate its performance among the choice of different control charts used to monitor the process after the adjustment has been made. Section 8-2 investigates some performance criteria additional to the ones for comparing the control charts of Chapter 5, section 8-3 assesses the effectiveness of the charts applied together with the EPC method and draws some conclusions about their appropriateness. In section 8-4 the choice between monitoring the output or the control action is investigated, while the differences between the MMSE and the PI controllers are elucidated in section 8-5. Section 8-6 is concerned with the robustness properties of the MMSE and the PI controllers and section 8-7 discusses the performance of the PID chart.

### 8.2 Performance criteria used for the ASPC technique

Apart from the Average Run Length and some other performance criteria discussed extensively in Chapter 5, some supplementary criteria used exclusively in the field of feedback adjustment are:

#### a) *The PM criterion*

The PM (Performance Measurement) criterion was the performance measurement used in Montgomery et al. (1994) to compare the SPC schemes

applied after the ASPC method and it was named in this way because it was the principal criterion in the authors's study. The PM is simply the average squared deviation of the output from the target T, that is:

$$PM = \frac{1}{n} \sum_{t=1}^n (Y_t - T)^2 \quad (8-1)$$

Obviously, the smaller the PM value is, the more effective the adjustment scheme has been.

*b) The criterion of Absolute Efficiency of variation reduction (AE)*

This criterion is the ratio of the variance of the disturbance model  $\sigma_D^2$  over the variance of the output error  $\sigma_e^2$  (that is, the deviation of the quality characteristic from target), i.e.,

$$AE = \sigma_D^2 / \sigma_e^2 \quad (8-2)$$

The larger the value of AE or the smaller the standard deviation of the output, the better is the performance.

If the MMSE control is used, then, assuming that the model as well as its parameters are known,  $AE = 1$  since the variance of the output error  $\sigma_e^2$  is minimized only if it is equal to the disturbance variance  $\sigma_D^2$ . On the other hand, the PID schemes have usually  $AE \leq 1$  because their goal is not the minimization of the output variance.

Since the AE criterion compares the performance of a given control scheme with the MMSE scheme (because  $AE_{MMSE} = 1$ ), an equivalent formula for computing the AE when the PID scheme is compared to the MMSE is:

$$AE \equiv MSE_{MMSE} / MSE_{PID} \quad (8-3)$$

*c) The Relative Efficiency (RE) criterion*

This is defined, per analogy to the previous definition as:

$$RE \equiv MSE_{No-control} / MSE_{PID} \quad (8-4)$$



which compares the variability of the PID control scheme with the variability of the no-control strategy. Because the no-control strategy is a special case of the PID scheme with  $k_p = k_I = k_D = 0$ , it is clear that  $RE \geq 1$ . The values of  $RE$  give us a measure of the improvement over the no-control strategy, i.e., the lower bound of the control performance. It is obvious that, apart from the PID controller, any control scheme of interest can be compared with the no-control action.

*d) The signal-to noise (SN) ratio*

This performance measure introduced by Jiang, Tsui and Woodall (2000) in order to study the average run length ARL of their ARMA chart when a process shift occurs, can also be used to predict the performance of any SPC monitoring charts. We remind that the transient and steady-state signal to noise ratios for a statistic  $Z_t$  are given by the formulas:

$$R_T = \mu_T / \sigma_z \text{ and } R_S = \mu_S / \sigma_z,$$

with  $\sigma_z$  being the standard deviation of the charted statistic and  $\mu_T$  (or  $\mu_S$ ) the transient (or steady-state) mean shift level of the charted process, i.e., the process mean level when  $t=0$  (or  $t = \infty$ ).

The SN ratio works as follows: the transient ratio measures the capability of a chart to detect shifts in the first few runs and this makes it an important indicator in detecting large shifts. If the chart fails to signal a shift in the early runs, then the steady-state ratio becomes critical in indicating how efficient the monitoring chart is for detecting the shifts in the later runs. The above general rule used from Jiang, Tsui and Woodall (2000) can also be applied under APC controlled processes.

Generally, the chart with a higher  $R_T$  is often preferred if its  $R_T$  value is high enough (say more than 4). It is then expected that the preferred chart will quickly signal in the first runs after the shift. However, if all candidate charts have a relatively small  $R_T$  value (say less than 3), then the chart with a larger  $R_S$  is preferred even if its  $R_T$  value is somewhat smaller. It is then expected that the preferred chart will efficiently signal in the later runs. When all charts have moderate values for both ratios, then their performance is considered similar.

Note, though, that the above SN rule is somewhat ad hoc. It may not be reliable if  $R_T$  is high but  $R_S$  is extremely small (about 0), since the shift on the chart could drop down to zero too quickly to let the chart signal the shift in the first few runs. In this case, a long ARL is to be expected.

*e) The Average Quality Cost (AQC) criterion*

The general form of the Average Quality Cost (AQC) value was given by Eq(7-7). In order to assess the effectiveness of this measurement, it was used by Jiang and Tsui (2000) along with the corresponding ARL values to compare the performance of traditional SPC charts after the MMSE controller has been applied.

### **8-3 Efficiency in SPC monitoring of ASPC controlled processes**

The most evident confirmation about the efficiency of the ASPC approach would be to check the ability of the standard control charts to detect shifts in the process mean after the control action has been applied. The charts that achieve this more quickly than the others are considered as the most effective.

#### **8-3.1 Comparison of control charts based on the economic model AQC**

Jiang and Tsui (2000) were based on the fact that monitoring the process output of an MMSE-controlled autocorrelated process is equivalent to monitoring the forecast error of the same process. Various disturbance models of AR(1) and ARMA(1,1) type in which the MMSE scheme is applied were used in their simulation studies.

##### **8-3.1.1 Comparisons for AR(1) processes based on the AQC model**

Assuming that a step shift  $\mu$  occurs at time 0 and that the MMSE control scheme is applied to an AR(1) disturbance process, the mean of the process before and after the shift occurrence is:

$$\mu_t = \begin{cases} 0, & t < 0 \\ \mu, & t = 0 \\ (1-\phi)\mu, & t > 0 \end{cases} \quad (8-5)$$

The formula of the Average Quality Cost (AQC) shown in Eq(7-7) is formatted after being applied under the AR(1) model as (Jiang and Tsui, 2000):

$$L_A = (\sigma^2 + (1-\phi)^2 \mu^2) + \frac{\mu^2 - (1-\phi)^2 \mu^2 - (1/p)[(1-\phi)^2 \mu^2 - (a+p)C_D]}{(1/p) + ARL_1} \quad (8-6)$$

Comparing different SPC charts with the same  $\alpha$  and diagnosis cost ( $C_D$ ), Jiang and Tsui found that  $L_A$  is a monotonically increasing function of the out-of-control ARL ( $ARL_1$ ) if the following condition holds:

$$\mu^2 - (1-\phi)^2 \mu^2 - (1/p)[(1-\phi)^2 \mu^2 - (\alpha+p)C_D] < 0 \quad (8-7)$$

This is consistent with the traditional argument that the smaller the out-of-control ARL, the smaller the cost (specified by the AQC value), and, thus, the better the performance of the monitoring chart. However, when the condition of Eq(8-7) is reversed ( $>0$ ), the monitoring chart with the larger  $ARL_1$  will have a smaller AQC. This condition is not trivial and it occurs when:

(1)  $\phi$  is very close to one.

In this case, the shift can be significantly recovered by the MMSE control action and the process experiences an approximately zero shift, so that  $(1-\phi)\mu \approx 0$ . According to the assumption that no other shift would occur during the out-of-control period before an adjustment is made, the Average Quality Cost (AQC) of a chart with a longer  $ARL_1$  will be smaller than that with a shorter  $ARL_1$ . It is possible, however, that the longer the ARL is, the higher chance there is that another shift might happen. Therefore, the overall cost will be increased.

(2) *the diagnosis cost is very high.*

If this is the case, then the quality cost after the shift may be lower than the quality cost before the shift. Thus, a monitoring chart with a longer ARL may result in a smaller AQC, though, again another shift may occur before the shift is removed and this might increase the cost.

Jiang and Tsui (2000) compared three control charts, the Individuals Shewhart, the EWMA and the combined EWMA- Shewhart charts to find how the AQC criterion is different from the ARL. By keeping  $C_D$  and  $\alpha$  constant, they proved that when the diagnosis cost is negligible, the AQC is proportional to ARL through the proportional constant  $P = (1/p+1)(1-\phi)^2 - 1$ . When the diagnosis cost cannot be neglected, the AQC is no longer proportional to ARL and the performance of the chart depends on the magnitude of the cost. In this situation, it is optimal not to do SPC monitoring because the ASPC scheme is able to compensate for the special cause so that the control charts are not cost effective.

The simulation study of the authors showed that when diagnosis costs are small, the best chart in terms of AQC is consistent with that in terms of ARL and this is the combined EWMA- Shewhart. This chart dominates for large and median shifts, while the EWMA alone is the best chart when the shift is small. For the cases with large diagnosis cost, the Individuals chart has the smallest AQC, although it often has the longest ARL among the three charts.

### 8-3.1.2 Comparisons for ARMA(1,1) processes based on the AQC model

The mean shift pattern of the output for an ARMA(1,1) model with a step shift at time  $t_0$  was specified by Jiang and Tsui, 2000 as:

$$\mu_t = \begin{cases} 0, & t < t_0 \\ \mu, & t = t_0 \\ \left(\frac{1-\phi}{1-\theta} - \frac{\theta-\phi}{1-\theta}\theta^{t-t_0}\right)\mu, & t > t_0 \end{cases} \quad (8-8)$$

The AQC derived after substituting the mean shift pattern of Eq(8-8) into the formula of Eq(7-7), is proven to be:

$$L_A = (\sigma^2 + \alpha C_D) + \frac{\sum_{T=1}^{\infty} P_T(\mu) \sum_{t=0}^{T-1} (B^2 \theta^{2t} - 2AB\theta^t) + A^2 - (ARL_1 \times a - 1)C_D}{(1/\lambda) + ARL_1} \quad (8-9)$$

Jiang and Tsui (2000) used the Markov chain method to calculate the AQC and ARL values of the ARMA(1,1) model and they found that the AQC values of the three charts of interest for the same mean shift are not proportional to the ARL values when the diagnosis costs are zero, small or large. For non-zero diagnosis costs, this happens in the same way as for the AR(1) process explained earlier. For zero diagnosis cost, the phenomenon can be explained by the fact that the mean shift function in Eq(8-8) is not constant over time. The ARL does not take into account the mean shift change and assigns an equal weight to each run length probability. On the contrary, the AQC assigns a different weight proportional to the mean shift change to each run length probability. As a result, the two criteria will be different when the mean shift changes significantly over time.

Among the three charts, when the diagnosis cost is negligible the optimal result under the AQC criterion is consistent with the ARL indicating the EWMA as, usually, the best for detecting small shifts and the Individuals control chart as the best for large shifts. When the integrated AQC criterion is used, the combined EWMA-Shewhart and the Individuals charts are uniformly the best. In general, when the two parameters  $\phi$  and  $\theta$  approach to the boundary of opposite directions, so that the variance of the underlying process becomes large, the Individuals chart outperforms the combined EWMA- Shewhart in terms of the integrated AQC value.

The integrated AQC measurement when the shift magnitude follows a distribution function  $F(\mu)$ , is defined by Jiang and Tsui (2000) as:

$$IAQC = \int_0^{\infty} AQC(\mu) dF(\mu) \quad (8-10)$$

### 8-3.2 Relative efficiency of charts used in the ASPC scheme

Montgomery, Keats, Runger and Messina (1994) used two types of assignable causes: a sustained shift and a trend, with the purpose of detecting the most powerful SPC tools on the basis of these two kinds of causes. The performance criteria in which their simulation studies were based are the ARL value and the PM criterion specified by Eq(8-1).

#### 8-3.2.1 Sustained shift

Montgomery, Keats, Runger and Messina (1994) used a simulation study to investigate the success of four control charts, being the Shewhart chart for Individuals with  $3\sigma$  limits, the EWMA with  $\lambda = 0.1$  and  $0.4$  and  $3\sigma$  limits and the CUSUM chart with parameters  $k = 0.5$  and  $h = 5$  in detecting several sustained shifts of magnitudes ranging from 1 to 10 units. The model used is the one leading to the MMSE controller of Eq(6-12).

The first thing worth mentioning is that the combined EPC/SPC scheme had a smaller PM value than the EPC rule alone, leading to the conclusion that, integrating an SPC rule with EPC by applying control charts to the output deviation from target, results in reducing overall variability if assignable causes in the form of sustained shifts occur. Some indication that the Individuals chart performs better than the other charts for large shifts of magnitude 7.5 and more was perceived.

By comparing the ARLs it was shown that the small shifts were difficult to detect under all schemes since the effect of an assignable cause is converted from a step change in a correlated process to a patterned change in an autocorrelated change, but with the application of EPC, active control is compensated for it.

#### 8-3.2.2 Assignable cause resulting in a trend

A similar study was conducted with the shift interfering in the process being continuous, so as to create a trend ranging from 0.05 to 1 units per period. Once again, the PM criterion revealed the superiority of the EPC/SPC rule to the EPC rule alone, but now the indices is that the three non-Shewhart



procedures provide more reduction in the variability than does the Individuals chart, with the EWMA ( $\lambda = 0.1$ ) and CUSUM being particularly effective.

In terms of the ARL values, again the EWMA and CUSUM seem to perform best though the choice between them is not critical because they behave similarly when being adjusted.

#### 8-4 Choosing between monitoring the output or the control action

We have already referred to the fact that a mean shift in the process output (which is the same as the forecast error) changes over time due to the ASPC compensation and, thus, it affects the performance of the SPC monitoring chart. Therefore, in some cases it may be more efficient to monitor the control action instead of the process output. Jiang and Tsui (2002) performed simulation studies with the purpose of finding if monitoring the output or the control action is more effective in detecting a mean shift. By assuming a one unit delay, after a step mean shift has taken place, the output takes the following form:

$$e_t = D_t - u_{t-1} + h_t, \text{ where } h_t=0 \text{ for } t<0 \text{ and } h_t=h \text{ for } t \geq 0 \quad (8-11)$$

Jiang and Tsui (2002) based their conclusions by using the SN ratio as the only performance criterion and by applying both the MMSE and the PI control schemes.

##### 8-4.1 The SN ratios under the MMSE controller

If the disturbance model is an ARMA(1,1) model, then the MMSE controller has the form of Eq(6-12). For simplicity, the mean shift magnitude can be defined in terms of the standard deviation of the process output as  $h=\mu\sigma_e$ . Then, after mathematical calculations, the signal to noise ratios described in section 8-2d) reduce to (Jiang and Tsui, 2002):

$$\begin{aligned} R_T^e &= \mu, \quad R_S^e = \mu \left| (1-\phi)/(1-\theta) \right| \text{ for the output error, and,} \\ R_T^u &= \mu \sqrt{1-\varphi^2}, \quad R_S^u = \mu \sqrt{1-\varphi^2} / (1-\theta) \text{ for the control action} \end{aligned} \quad (8-12)$$



It is interesting to note that  $R_T^e$  is always at least  $R_T^u$ , since  $\sqrt{1-\phi^2} \leq 1$  and, according to the heuristic rule, the value of  $R_T$  is critical for detecting large mean shifts in the early runs. As a result, if the mean shift magnitude is large (4 or  $5\sigma_e$ ), it is expected that monitoring the output is more efficient than monitoring the control action. On the other hand, if the shift magnitude is small, so that both transient ratios are small, the efficiency of the charts will depend on  $R_S$ . When  $\phi > 0$ ,  $R_S^u > R_S^e$ , so it is expected that monitoring the control action will be more efficient. This is quite logical because when the correlation is positive, the MMSE automatically reduces the shift level in the output and this makes detection difficult. When  $\phi < 0$ , then monitoring the output will be the best choice, since the shift will be amplified and, thus, easier to detect.

The simulation results confirmed these theoretical considerations by showing that in terms of the ARL values, **monitoring the output was quicker than monitoring the control action for detecting a shift of  $5\sigma_e$  or of  $3\sigma_e$  along with  $\phi > 0$ , though a control chart of the control action performs best when the shift is  $3\sigma_e$  or less and at the same time  $\phi$  has a negative value.** The Shewhart chart was the only SPC tool used by the authors.

#### 8-4.2 The SN ratios under the PI controller

For the pure P controller ( $k_I = k_D = 0$ ) the control action is the output scaled by  $k_p$ , since  $u_t = k_p e_t$ . **Therefore, the performance of the output chart and the control action for the P controller is the same.** Concerning the PI controller, when a step mean shift  $h$  occurs at time 0, the two SN ratios were calculated by Jiang and Tsui (2002) as:

$R_T^e = h/\sigma_e$ ,  $R_S^e = 0$  for the output and

$R_T^u = (k_p + k_I)h/\sigma_u$ ,  $R_S^u = h$  for the control action, where:

$$\sigma_e = \sqrt{\frac{1 - \kappa_p \phi + \theta \kappa_p + \theta \phi \kappa_p^2 - \phi \theta - \theta \phi^2 k_p}{(1 + \phi \kappa_p)(1 - \kappa_p^2)(1 - \phi^2)}} \sigma_e \quad \text{and} \quad \sigma_u = k_p \sigma_e \quad (8-13)$$



It is important to note that the mean shift of the output converges to 0 (i.e.,  $\mu_s^e = 0$ ) due to the I component, that is the mean shift is completely compensated for in the steady state. However, this robustness property turns to be a disadvantage for SPC monitoring since the mean shift often cannot be detected, especially when it is small.

Note that the mean shift of the pure P controller does not converge to 0 because the two transient and steady-state mean shift patterns of the P scheme are:

$$\begin{aligned}\mu_T^e &= \eta, \mu_S^e = \eta/(1+\kappa_p), \text{ and} \\ \mu_T^u &= \kappa_p h, \mu_S^u = \kappa_p h/(1+\kappa_p)\end{aligned}\tag{8-14}$$

This further explains that the robustness property of the PI control is coming from the I-component rather than the P-component of the controller.

The simulation approach, applied for the same ARMA(1,1) disturbance models as in the case of the MMSE controller of section 8-4.1, revealed that the ARL was smaller when monitoring the control action than the output under the PI controller. This result can also be explained in terms of the SN ratios: since the steady-state ratio of the output chart is always 0 due to the I-component, the performance of the output chart is mainly determined by the transient ratio. In other words, if the chart misses the shift in its transient state, it is very hard to detect it in the steady-state and a longer ARL is to be expected. On the other hand, the non-zero steady state ratio of the control action chart can help the chart to efficiently signal the shift in the later runs, although that its transient ratio is smaller than the one of the output.

For the P controller the ARLs were, as expected, the same no matter if the monitoring data was the output or the control action because the latter is a multiplier of the first one. As a result, **monitoring the control action is always more efficient than monitoring the output when the PI controller is used.**

The authors noted that it was not meaningful to compare the SPC efficiency of the MMSE and PI controlled processes as the shift levels used were not the same due to the difference of the output variance.

## 8-5 Comparisons among the PI and MMSE schemes

Section 8-4 gave an idea on the difference between the PI and the MMSE schemes by trying to find cases in which monitoring the control action is superior to monitoring the output. The performance of the two controllers when they are applied to the same set of data has been more analytically studied.

### 8-5.1 Relative performance under specific disturbance models

Tsung, Wu and Nair (1998) used the stationary disturbance model ARMA(1,1), as well as the nonstationary model ARIMA(1,1,1), in order to draw conclusions about the efficiency of the two most common adjustment schemes.

#### 8-5.1.1 Performance under ARMA(1,1) disturbance models

Tsung, Wu and Nair (1998) first proved that, under any stationary disturbance model  $D_t$ , the stability region of PI schemes is:

$$\{(k_p, k_I) : k_p > -1, k_I \geq 0, k_p + k_I/2 < (1+\delta)/(1-\delta)\} \quad (8-15)$$

where  $\delta$  measures the inertia for the process dynamics given by Eq(6-9). That is, the PI schemes should be restricted to be in the region of Eq(8-15), which as a matter of fact, gets larger as  $\delta$  increases.

Considering the pure gain model [i.e., when  $\delta=0$  in Eq(6-9)] and the ARMA(1,1) as the disturbance model, Tsung, Wu and Nair (1998) conducted a simulation study using the optimal PI schemes (that is, the ones that minimize the output variance within the class of PI schemes). They showed that when  $\phi$  was close to  $\theta$ , all the PI schemes (P, I and PI) had very high efficiency ( $AE > 0.99$ ). This is because in the limiting case where  $\phi = \theta$ , the ARMA(1,1) model reduces to white noise for which the no-control strategy is optimal. It can be proved that under the ARMA(1,1) disturbance model, the MMSE scheme is a P scheme when  $\phi = 0$  (that is for MA(1) models). That is why

the optimal P schemes had high efficiency when  $\phi$  was close to 0. In general, P schemes get less efficient as the difference  $|\phi - \theta|$  gets large.

The only case in which the PI controller was not very efficient was when  $\phi \approx -1$  and  $\theta \approx 1$ , but the process disturbance has then large negative correlation and this is not a situation of practical interest in industry. Thus, there is little loss of efficiency in using the PI controller over the MMSE.

The authors presented some contour plots about the AE values [described in section 8-2b)] of the PI controller for several values of the parameter  $\delta$  in the process dynamics. These contour plots showed that the PI controller has high efficiency along the diagonal line  $\phi = \theta$  because, as has been already said, the process disturbance reduces to white noise in this case.

An important issue concerning the inertia parameter is that when  $\delta = 0$ , the PI controller has high efficiency if  $\phi = 0$  because the P scheme is optimal for MA(1) models. On the other hand, the region of high efficiency when  $\delta > 0$  is centered at the line  $\phi = \delta$ . Tsung, Wu and Nair (1998) proved also mathematically that for the general first-order dynamic model with process inertia  $\delta$ , the P scheme is optimal for an ARMA(1,1) process when  $\phi = \delta$ . As a consequence, the region where PI schemes perform well shifts up as  $\delta$  increases.

Tsung and Shi (1999) were concentrated on the fact that if there is an integral control involved in the PID controlled process, a steady-state shift of the process outputs will be eliminated immediately after a process change, leaving a limited 'window of opportunity' during which the process change must be detected. To overcome this problem they thought of jointly monitoring both the output and input variable using bivariate SPC to improve the efficiency of detection than plot them separately. Bonferroni's inequality was used to control the overall error probabilities when monitoring multiple characteristics simultaneously. The proposed SPC chart design was directly based on the PID controlled process model, since the standard deviations of the output and of the manipulated input after process control were functions of the PID control parameters.

The control limits of the joint charts ( $CL_e$  and  $CL_u$ ) were written by Tsung and Shi (1999) as:

$$\begin{aligned} CL_e &= \pm L_e \sigma_e \\ CL_u &= \pm L_u \sigma_u \end{aligned} \quad (8-16)$$

The joint decision rule suggests that the controlled process is out of control when either the controlled output or the manipulated input is outside the limits.

A simulation approach was then applied by Tsung and Shi (1999) by using the PID controller of Eq(6-11) as the control action and an output of the form  $e_t = u_{t-1} + D_t + \mu_t$ , with  $\mu_t$  being a step shift. The Bonferroni's approach was the one preferred for the SPC monitoring. The parameters of the PID controller were derived by the design maps explained in section 7-3, proposed by the same authors. For each disturbance model parameter set, the process was simulated to obtain the ARL value and the simulation data showed that the geometric assumption of the run-length was acceptable.

The authors presented the contour plots of the ARLs of joint monitoring, which led them to the following conclusions:

(a) For large mean shifts (e.g.  $\mu_t = 2\sigma_D$ ), with  $\sigma_D$  being the standard deviation of the disturbance, Bonferroni's approach for the PID controlled processes performs quite well with ARLs < 10. Especially when  $\phi < \theta$ , most of the ARLs were less than 3.

(b) For medium mean shifts (e.g.  $\mu_t = 1\sigma_D$ ), Bonferroni's approach performs well when  $\phi < \theta$ , but it does not perform well (ARL > 50) in the remaining cases.

(c) For small mean shifts (e.g.  $\mu_t = 0.5\sigma_D$ ), the performance is satisfactory only near the region of  $\phi$  close to  $-1$  and  $\theta$  close to  $1$ .

The only SPC chart used was the Shewhart but it is well known that this chart is not sensitive to small mean shifts. The power of small-shift detection could be improved by replacing the Bonferroni's approach with a more advanced multivariate SPC scheme, such as the multivariate CUSUM or EWMA.

The use of Bonferroni's approach is not recommended for pure P (Proportional) processes because then, the manipulated inputs are

proportional to the controlled outputs, so the simultaneous monitoring of both inputs and outputs is redundant and the corresponding type I error is wrongly designed by Bonferroni's inequality. Thus, **when  $\phi$  is close to zero and where the best choices of  $k_I$  and  $k_D$  are close to 0, the joint monitoring approach should be avoided.**

Furthermore, Tsung and Shi (1999) investigated the AE values of the PID controller compared to the ones of the MMSE controller as described by Eq(8-3) and the RE values using Eq(8-4). **They showed that the PID schemes perform well with AE more than 90% for most of the ARMA(1,1) parameter space.** Only for  $\phi$  close to  $-1$  was the AE not as good, but the RE was still much greater than 1.

#### 8-5.1.2 Performance under ARIMA(1,1,1) disturbance models

In order to investigate the efficiency of the PI controller under the ARIMA(1,1,1) disturbance model, Tsung, Wu and Nair (1998) proved initially that, for a first-order nonstationary disturbance model, the stability region for PI schemes is:

$$\{(k_p, k_I) : k_p > -1, k_I > 0, k_p + k_I/2 < (1+\delta)/(1-\delta)\} \quad (8-17)$$

Unlike the stationary case, now  $k_I$  cannot be 0, so the I mode of action is necessary when nonstationarity is present.

It has been analytically proved that the MMSE schemes coincide with I control schemes under the ARIMA(1,1,1) disturbance model with  $\phi=0$ , that is, the IMA(1,1). According to Tsung, Wu and Nair (1998), when  $\delta=0$ , there exists the symmetry property  $AE(\phi, \theta) = AE(-\phi, -\theta)$ . Furthermore, there is again high efficiency of the PI controller along the line  $\phi = \theta$ , because this time  $D_t$  reduces to a random walk for which the I control is optimal. The I control is also optimal for IMA(1,1) models, as we have seen, and that is why the performance near the line  $\phi=0$  is very good.

Comparing the I with the PI control, it was shown that the AE values increase substantially under the second scheme and, thus, there is considerable gaining using the PI over the I control schemes. An awkward

conclusion was that optimal I control with ARIMA(1,1,1) disturbance models behaves the same way as optimal P control under ARMA(1,1) models.

When  $\delta > 0$ , the pure I scheme is no longer optimal for an IMA(1,1) disturbance. It was proven by Tsung, Wu and Nair (1998) that the MMSE scheme in this case is a PI scheme with  $k_p = (1 - \theta)\delta/(1 - \delta)$  and  $k_I = 1 - \theta$ . Thus, the P mode of action is necessary even for IMA(1,1) models when there is process inertia, and the magnitude of the P mode increases with  $\delta$ . It has been also proven by the same authors that under the ARIMA(1,1,1) disturbance with  $\phi = \delta$ , the MMSE scheme is a PI scheme with  $k_p = \delta/(1 - \delta)$  and  $k_I = (1 - \theta)/(1 - \delta)$ . Consequently, **the PI schemes have high efficiency in the regions where  $\phi$  is close to 0,  $\delta$ , or  $\theta$** . This explains why the region of high efficiency gets bigger as  $\delta$  gets larger.

As a conclusion, the authors showed that PI controllers can have approximately the same control performance as the MMSE controllers for reducing process variation.

### 8-5.2 Relative performance under different types of shifts

In section 8-5.1 our concern was to find the controller that best attracts the data towards their target value. The efficiency of a controller is, however, also assessed according to its ability to detect the special causes, that is, to not fully compensate for the deviations from target, so that if a sudden shift takes place, there are chances to find it. Jiang and Tsui (2002) examined the effectiveness of the MMSE and the PI controller when the special cause that results in a mean shift in the process takes the form of two types: a step shift and a drift. If the mean shifts affect the process, starting from example from the 101<sup>st</sup> observation onwards, then the step shift is a constant shift of the form:

$$\mu_t = \begin{cases} 0, & t < 101 \\ \mu, & t \geq 101 \end{cases} \quad (8-18)$$

while the drift is a linear trend shift of the form:

$$\mu_t = \begin{cases} 0, t < 101 \\ (t - 100)\mu, t \geq 101 \end{cases} \quad (8-19)$$

### 8-5.2.1 Performance when a step change has occurred

Jiang and Tsui (2002) used an example in which a step shift was interfered into the process, and showed that the MMSE controller was able to compensate for most of the shift and adjusted the mean shift pattern to a small constant both when the output and the control action were monitored. The control chart, though, detected the shift less quickly than the output chart. Since the controller has quickly compensated for the shift, there was a spike at the beginning of the shift but the data quickly returned to their original pattern in both the output and control action cases. The performance can be explained by the SN ratios: since the transient ratio of the output chart is higher than that of the control chart, the output chart has a higher probability to signal the shift in the first few runs than the control action chart.

Although the MMSE and PI controllers have approximately the same performance in terms of reducing the variation of the autocorrelated processes, as it was proven by Tsung, Wu and Nair (1998), they behave quite differently when a mean shift occurs. The PI controller completely compensated for the step shift and adjusted the mean shift pattern to zero for the output chart and to a positive constant for the control action chart. This is the result of the robustness property of the PI controller in the process output, which makes it difficult to detect the unanticipated shift. More explicitly, since the steady-state ratio of the control action is higher under the PI control than under the MMSE control, the control action chart under the PI scheme is expected to signal earlier than under the MMSE scheme.

A more thorough study of the same authors showed that **monitoring the output gives a smaller ARL than monitoring the control action for the MMSE controller, and vice versa for the PI controller**. Furthermore, because under the PI controller, the steady-state ratio is 0, the step shift is harder to detect from the PI controlled output than from the MMSE output, unless it is detected at once.



### 8-5.2.2 Performance when a drift has occurred

When Jiang and Tsui (2002) introduced a drift in the process, they found that the MMSE controller was unable to compensate for the upward linear trend and it resulted in a linear shift pattern both in the output and the control action with a slower upward trend in the first than in the second. This trend consequently led to out-of-control signals in the Shewhart chart which were triggered more quickly in the control action chart than in the output chart. This fact is explained as follows: because the steady-state ratio of both charts is infinity, obviously, the mean shift is easily detected in both charts, with a small advantage for the control chart.

On the contrary, the PI controller significantly compensated for the linear upward trend and adjusted the mean shift to a small constant above zero for the output chart, confirming the robustness property of the PI controller. For the control action, the mean shift pattern behaved similarly to the original linear upward shift. In terms of the SN ratio, the small value of the steady-state ratio for the output chart makes it difficult to detect the mean, though the value of infinity for the same ratio of the control action implies that the mean shift can be detected. Therefore, **control actions are a better data stream to be monitored when the PI controller is used.**

## 8-6 Robustness of the ASPC scheme

Luceno (1998) studied the performance of the ASPC mode when the process model is stationary (more specifically, when the disturbance follows an ARMA(1,1) model) but it is misspecified as being the nonstationary process IMA(1,1), as well as the inverse case. He proved that, in the context of feedback control, if the process is affected by stationary disturbances, then by assuming that these disturbances are nonstationary, one loses little efficiency. In contrast, if the process is inflated by nonstationary disturbances, then assuming stationary disturbances, the adjustments made will be quite inefficient. That is why, it is better to apply a nonstationary model than a stationary one whenever their ability to model the disturbance term is not well distinguished.





### 8-6.1 Controlling an ARMA disturbance with IMA forecasts

Assuming that the system is responsive, that is, the full effect of an adjustment  $u_t - u_{t-1}$  made at time  $t$  is realized at the output within the next unit interval, suppose that the true disturbance is generated by the ARMA(1,1) model given by Eq(3-12), that is,  $D_t = \phi D_{t-1} + \epsilon_t - \theta \epsilon_{t-1}$ , but compensation for this disturbance is made using the IMA model shown in Eq(3-18), i.e., so that the forecast  $\hat{D}_t$  is  $\hat{D}_t = \lambda D_{t-1} + \tilde{\theta} \hat{D}_{t-1}$ , where  $\tilde{\theta} = 1 - \lambda$  and the MMSE control consistent with the IMA model is applied. If we substitute these  $D_t$  and  $\hat{D}_t$  values to the equation  $e_t = D_t - \hat{D}_t$  that gives the errors of the output which coincide with the disturbance errors, Luceno (1998) showed that we are led to:

$$(1 - \phi B)(1 - \tilde{\theta} B)e_t = (1 - B)(1 - \theta B)\epsilon_t \quad (8-20)$$

The errors at the output are not a white-sequence as they should be if exact MMSE control had been used, but they are generated by a stationary and noninvertible ARMA(2,2) model, as shown by Eq(8-20). The resulting variance of  $e_t$  is then larger than  $\sigma_\epsilon^2$  but still finite. Thus, the output error may not be independent and identically distributed, but it continues at least to be stationary.

### 8-6.2 Controlling an IMA disturbance with ARMA forecasts

If the disturbance is generated by an IMA model of the form  $D_t = D_{t-1} + \epsilon_t - \theta \epsilon_{t-1}$ , but compensation is made using an ARMA(1,1) model, so that the forecast  $\hat{D}_t$  is computed as  $\hat{D}_t = \tilde{\phi} D_{t-1} + \epsilon_t - \tilde{\theta} \epsilon_{t-1}$  with  $\lambda = \tilde{\phi} - \tilde{\theta}$ , and the MMSE policy is used, then the errors of the output  $e_t = D_t - \hat{D}_t$  are derived as:

$$(1 - B)(1 - \tilde{\theta} B)e_t = (1 - \tilde{\phi} B)(1 - \theta B)\epsilon_t \quad (8-21)$$

which is a nonstationary ARIMA(1,1,2) model. Therefore, the variance of  $e_t$  tends to infinity when  $t$  increases and this policy is, hence, untenable.

The inadequacy of the policy based on the ARMA(1,1) model is not immediately apparent in practice due to the starting conditions  $D_0 = 0$  and  $\hat{D}_0 = 0$ . Thus, the output errors of the process will be normally close to 0 at first, but, after a number of observations has passed, they will start to be positive or negative and this trend will be more and more evident. The trouble in this case is not that the stationary disturbance model is wrong because no model can be expected to be perfectly true in practice, but the trouble is that it is not robust. Thus, it is dangerous to use a stationary disturbance model to deal with a process that may be nonstationary, though the reverse is rather safe.

The explanation for this is that, even though  $\hat{D}_t = \tilde{\varphi} D_{t-1} + \epsilon_t - \tilde{\theta} \epsilon_{t-1}$  provides the MMSE forecast of  $D_t$ , under both the ARMA(1,1) and the IMA(1,1) models, in the first case  $\lambda + \tilde{\theta} = \tilde{\varphi} < 1$ , but in the second  $\lambda + \tilde{\theta} = \tilde{\varphi} = 1$ . Therefore,  $\hat{D}_t$  is an exponentially weighted moving average (EWMA) of past data under the IMA model, but it is not an average of past data under the ARMA model. This small difference between the IMA and ARMA models yields essential differences in their performance as disturbance models. Note that, despite of the similarity between the two models expressed by Eq(3-12) and Eq(3-18), according to which the model of Eq(3-12) could be seen as a generalization of Eq(3-18), there are profound differences in the two models, since the first is stationary but the second is not.

A solution to the problem of the inappropriateness of the MMSE controller when a nonstationary disturbance is considered as stationary is given by Tsung, Wu and Nair (1998) by simply using the PI controller in the place of the MMSE. More precisely, they proved mathematically that in the case where  $D_t$  is *any first-order nonstationary* disturbance model, any PI control scheme applied to  $D_t$  leads to stationary output  $e_t$  provided that  $k_I > 0$ , that is, the I mode of action is active. Thus, a PI scheme that is optimal for a (wrong) stationary ARMA(1,1) model will still yield stationary output when the true disturbance is an ARIMA(1,1,1) or an IMA(1,1) as long as  $k_I > 0$ . The MMSE schemes, on the other hand do not have this robustness property because we saw that they resulted in a nonstationary model for  $e_t$ . A



simulation study conducted by the authors confirmed their theoretical deductions.

In the opposite case in which an ARMA(1,1) disturbance model is misidentified as an ARIMA(1,1,1) disturbance, Tsung, Wu and Nair (1998) showed that the PI scheme, as well as the MMSE scheme discussed previously, also leads to a stationary output  $e_t$ . Although that both schemes had comparable performances under this type of model misidentification, the simulation approach of the authors showed that the variances of the output  $e_t$  were often smaller under the PI scheme than under the MMSE scheme.

**Therefore, when one takes into account the model uncertainty, the PI schemes can be more efficient than the corresponding MMSE schemes.**

### 8-7 Performance of the PID chart

In the previous chapter, the PID chart was proposed by Jiang et al. (2002). The authors did not limit themselves to the presentation of this chart, but they also used the simulation approach in order to prove its efficiency compared to other traditional charts. They first compared the PID chart to the SCC chart for various step shifts in terms of their ARL values under the ARMA(2,1) disturbance model. They showed that the SCC chart is very good for detecting large shifts but performs poorly for small to moderate shifts. The EWMAST (or otherwise the P chart) is better for these cases but its ARLs are still quite large. The solution is given by the PD chart, which detects small shifts more quickly than the EWMAST chart, though, at the same time, the PI chart has comparable performance to the SCC chart for detecting large shifts. Obviously, the class of PID charts is flexible because the chart parameters can be tuned to achieve good performance for small or large shifts. This flexibility is absent from the EWMAST and SCC charts.

Jiang et al. (2002) also conducted a simulation study for uncorrelated processes, that is, when the disturbance is white noise (i.e.,  $D_t = \epsilon_t$ ). In this case, it was shown that the P chart with  $-1 < k_p \leq 0$  performs better than the PI chart (with  $k_i > 0$ ). In particular, the Shewhart chart (corresponding to  $k_p = k_i = k_D = 0$ ) is the best for detecting large mean shifts and the P chart with a value of  $k_p = -0.8$  is the best for detecting small and moderate shifts. This latter



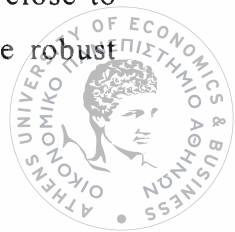
observation is consistent with the results of Lucas and Saccucci (1990), because for the iid processes, the P chart is equivalent to the EWMA chart with  $\lambda = 1 + k_p = 1 - 0.8 = 0.2$ .

Jiang et al. (2002) showed that for an iid process, both capability indices ( $R_T$  and  $R_S$ ) are higher for the Shewhart chart than for the PID with  $k_I > 0$ , because then  $R_S = 0$ . Therefore, **for iid data, we must focus only to PD charts**, and since the P charts (with  $k_p > 0$ ) do not perform well, we consider only the PD charts with  $k_p \leq 0$ . Finally, it was observed by the authors that the D chart with  $k_D = -0.1$  is the best (and better than the Shewhart) for detecting large mean shifts, while the PD chart with  $k_p = -0.8$  and  $k_D = 0.2$  is the best (and better than the EWMA with  $\lambda = 0.2$ ) for detecting small shifts. For detecting both small and large shifts, the ARL of the PD chart depends mainly on  $k_p$  ( $-1 < k_p \leq 0$ ), though for fixed  $k_p$ , the ARL varies slightly as  $k_D$  changes. **Therefore, if the data are uncorrelated, the P-component of the PID chart is the most important, the D-component gives a small improvement and the I-component should not be used.**

Concerning autocorrelated processes with the disturbance model being an ARMA(1,1), the authors showed that by finding the appropriate parameters of the PID chart according to their heuristic algorithm, one may achieve to design both a PID chart that detects small shifts and one that detects large shifts, with both outperforming the traditional SPC charts. The two PID charts should be used on the data so that small and large mean shifts are detected. The PID charts outperform EWMAST and M-M charts, indicating that **the D component is useful in the autocorrelated case.**

#### **Robustness of the PID chart to parameter misspecification**

Jiang et al. (2002) tried to compare the robustness of the SCC chart with that of the PID chart by using an ARMA(1,1) disturbance model with various parameter estimates different from the true ones. We had presented in Chapter 5 the simulation study of Adams and Tseng (1998), who found that the SCC chart is not robust when the model parameters are incorrectly estimated. The PID chart adjusted by Jiang et al. (2002) to detect large shifts had similar performance with the SCC chart. When the process was close to being nonstationary, however, the PID chart was proven much more robust



and this was due to the non-zero I term, which is appropriate for processes near nonstationarity.

On the other hand, the PID chart designed by the authors to detect small shifts was proved to be less robust than the SCC chart. This happens because a control chart that is sensitive to small shifts will also be sensitive to the estimation errors and, thus, is less robust. In general, it is difficult to design a control chart that is both sensitive to small shifts and robust to model misspecification.





## CHAPTER 9

### Other types of feedback control

#### 9-1 Introduction

In the previous three chapters a detailed description has been made concerning the most widely used tools of the feedback control scheme: the MMSE and the PID controllers. A different identification technique called the ‘closed-loop identification’ is evaluated in this chapter. This technique is based on directly modeling the closed-loop deviation from target, rather than attempting to model the controller or the open-loop process, and it is described in section 9-2. Apart from different methods applied to the two popular control schemes, a big range of other similar schemes has been developed, but their complexity or high cost have reduced their application. Some of them, however, are effective for particular sets of data and that is why they are mentioned in section 9-3. The final section, 9-4, gives some information about another feedback scheme mainly used in semiconductor processes.

#### 9-2 Closed-loop controllers

In the discrete-part manufacturing processes, we saw that the quality characteristic consists mainly of two parts: the process dynamics usually of the pure gain type and a noise dynamics, which is an additive disturbance exhibiting dynamic behavior. The identification techniques for determining the structure of the underlying stochastic process were relied until now on open-loop experimentation. However, open-loop identification experiments may be too costly, particularly if the process exhibits severe drift when left uncontrolled.



### 9-2.1 Definition of the closed-loop controller

It is often desirable to have available an identification technique that selects a disturbance model while a controller is operating, because in doing so, cost is reduced even if the controller in use is not optimal in any sense. With knowledge of the disturbance model affecting a process and with estimates of model parameters, the controller can be tuned to achieve more desirable performance. Del Castillo (2002) pursued an identification approach based on the autocorrelation structure of the closed-loop output provided that the algebraic form of the controller is known.

The types of disturbances considered by Del Castillo (2002) were all the particular cases included in a possibly non-invertible IMA(1,1) with drift process. The drift process is included if in the formula of Eq(3-19), expressing the IMA(1,1) model with no drift, the value  $\delta$  of the drift is added. Therefore, the formula of the *IMA(1,1) disturbance with a drift* is:

$$D_t = \delta + D_{t-1} + \epsilon_t - \theta\epsilon_{t-1}, \quad |\theta| \leq 1 \quad (9-1)$$

Obviously, when  $\delta = 0$ , Eq(3-19) is derived. The special cases of the model of Eq(9-1) are summarized in Table 9-1.

**Table 9-1: Types of Disturbance included in the model of Eq(9-1)**

	$\delta = 0$	$\delta \neq 0$
$\theta = 0$	Random walk (RW)	RW with drift (RWD)
$0 <  \theta  < 1$	IMA(1,1)	IMA(1,1) with drift
$\theta = 1$	White noise	Deterministic trend+noise (DT)

### 9-2.2 Closed-loop output description: the PI controller

For a PI controller with the form of Eq(6-8), Del Castillo (2002) showed that the controlled deviations from target for a process with any of the disturbances originating from Eq(9-1) follow an ARMA(2,1) process of the form:



$$(1 - \phi_1 B - \phi_2 B^2)(e_t - \mu_e) = (1 - \theta B)\epsilon_t, \quad \text{where}$$

$\phi_1 = 1 + gk_p$  and  $\phi_2 = gk_i$  with  $g$ ,  $k_p$  and  $k_i$  being the parameters denoted in Chapter 6.

The mean of the deviations from target is  $\mu_e = \delta / (1 - \phi_1 - \phi_2)$  (9-2)

Stability is achieved if  $|\phi_2| < 1$ ,  $g(k_i - k_p) < 2$  and  $g(k_i + k_p) < 2$ . Del Castillo showed that the model of Eq(9-2) reduces to an ARMA(1,1) process if a pure integral (I) controller, such as the EWMA controller is used. He also proved that the mean square deviation (MSD) of the output is given by:

$$\text{MSD}(e_t) = \text{Var}(e_t) + \mu_e^2,$$

where  $\gamma_0 = \text{Var}(e_t)$  is obtained from the variance of an ARMA(2,1) process (Box et al. 1994):

$$\gamma_0 / \sigma_\epsilon^2 = \{-\phi_1 \theta - \phi_1 \phi_2 \theta + [1 - \theta(\phi_1 - \theta)](1 - \phi_2)\} / \{1 - \phi_2 - \phi_1^2 - \phi_1^2 \phi_2 - \phi_2^2(1 - \phi_2)\},$$

while the variance of the adjustments is:

$$\text{Var}(u_t - u_{t-1}) = (k_p^2 + k_i^2)\gamma_0 + 2k_p k_i \gamma_1, \quad \text{with}$$

$$\gamma_1 = (\phi_1 \gamma_0 - \sigma_\epsilon^2 \theta) / (1 - \phi_2) \quad (9-3)$$

Given the derived expressions for  $\text{Var}(u_t - u_{t-1})$  and  $\text{MSD}(e_t)$ , Del Castillo followed the Box and Luceno (1997) approach and solved :

$$\min J = \text{MSD}(e_t) / \sigma_\epsilon^2 + \rho \text{Var}(u_t - u_{t-1}) / \sigma_\epsilon^2 \quad \text{subject to}$$

$$|gk_i| < 1, \quad g(k_i - k_p) < 2 \quad \text{and} \quad g(k_i + k_p) < 2 \quad (9-4)$$

Del Castillo provided several solutions to the model for the cases  $\rho = 0$  and  $\rho = 1$ . When  $\rho = 0$ , he found that the sign of the drift parameter ( $\delta$ ) does not affect the solutions. If  $\delta = 0$  and  $\theta = 1$ , the model of Eq(9-1) becomes the Shewhart's model for which no adjustment is necessary. In general, larger values of  $(k_p, k_i)$  are required to minimize MSD for larger drifts. In the case of  $\rho = 1$ , he showed that larger values of  $g$  result in better control. In addition, the larger the drift parameter, the larger the coefficients  $(k_p, k_i)$  should be, with highest values for the case of a random walk disturbance ( $\theta=0$ ).

The study of Del Castillo (2002) proved that neglecting the drift to determine the PI controller parameters can result in large deviations from

optimality, with respect to both  $\text{MSD}(e_t)$  and  $\text{Var}(u_t - u_{t-1})$ , particularly when a highly constrained solution is sought (i.e., when a large value of  $\rho$  is used).

In conclusion, PI controllers are quite robust with respect to many different disturbances and dynamics, as it has been discussed in Chapter 8, but there are limits to such robustness: a drift in the disturbance will require a different treatment from the case when there is no drift. If there is drift, the PI controller settings of Box and Luceno (1997) provide minimum variance control but not minimum MSE control, due to an offset for which the PI controller is unable to compensate. It is true, however, that if a good estimate of the drift is available from previous open-loop experience with the process, then feedforward control can compensate for the remaining variability and then the Box and Luceno settings will be optimal.

Similarly to the PI controller, a closed-loop identification may also be used in the case where the MMSE controller is judged as the appropriate scheme.

### 9-3 Other control schemes

The brief presentation of other control methods is done via a polymerization example provided by Capilla et al. (1999). The example helps to better comprehend the need for using other control methods if the standard ones do not give satisfactory results.

#### 9-3.1 Process description of an application

According to the polymerization process initiated by Capilla et al. (1999), large volumes of a polymer of a certain grade is produced and the key quality characteristic is polymer viscosity, measured by melt index (MI). The objective is to minimize MI variation around a target level of 0.8 viscosity units. Adjustments to viscosity can be made by varying the temperature of the reactor (T), which is a ready compensatory variable and whose changes represent negligible cost when compared to off-target viscosity cost.

The overview of the polymerization process and the consistent cross correlation function (CCF) of the series, which shows the dynamics of the



relationship between MI and T, suggest a tentative model for the measured viscosity variation  $\nabla MI_t$  at time t:

$$\nabla MI_t = w_1 \nabla T_{t-1} + w_2 \nabla T_{t-2} + \epsilon_t - \theta \epsilon_{t-1} \quad (9-5)$$

where  $\nabla T_{t-1}$  is the temperature adjustment at t-1 and  $\epsilon_t \sim \text{independent } N(0, \sigma_\epsilon^2)$ . In other words, the model dynamics is a second-order autoregressive moving average discrete transfer function model with IMA(1,1) disturbance. After estimating the parameters of the model in Eq(9-5), it was deduced that the parameter  $\theta$  is not significant and, thus, the term  $\theta \epsilon_{t-1}$  was dropped out from the model.

### 9-3.2 Developing different control schemes

#### 1) The MMSE controller

In the petrochemical process example, the cost of being off target is the overriding concern and it was supposed by Capilla et al. that it is a quadratic function of the viscosity deviations from the target. Therefore, control algorithms can be designed to minimize the mean squared deviation of viscosity from its target value. The MMSE controller optimizes the performance index as:

$$\min \{E[MI_{t+1} - \text{Target}]^2\} = \min \{[\sigma^2(MI_{t+1}) + (\text{bias})^2]\} \quad (9-6)$$

Although such a strategy seems to be appealing, the MMSE controller may not be ideal in practice. It may have undesirable properties, such as requiring excessive control action or having performance and stability characteristics that are sensitive to the accuracy of the process model.

#### 2) The CMV controller

Modified control schemes can sometimes be employed in which reduced control action can be achieved at a cost of small increases in the mean squared error (MSE) at the output. This can be accomplished by

optimizing a quadratic performance index involving the deviation of the output from its target and the deviation of the input from its steady state.

When the disturbance model exhibits non-stationary behavior, however, as is our case, it is impossible to stabilize the variance of the output from target when the manipulated variable is constrained to its steady-state value. The temperature must be allowed to float with no steady-state value. This is accomplished by constraining the change in control action  $\nabla T_t = T_t - T_{t-1}$ . In other words, in this modified scheme, the MSE of the output will be minimized subject to a constraint on the variance of the temperature adjustments  $\nabla T_t$ .

This **Constrained Minimum Variance (CMV)** controller optimizes the following quadratic objective function:

$$\min E \{ (MI_{t+1} - \text{Target})^2 + r(\nabla T_t)^2 \} \quad (9-7)$$

where the constraint parameter  $r$  is like a Lagrangian multiplier.

### 3) The CC controller

An alternative and much simpler approach to constraining the variations in the manipulated variable was proposed by Clarke and Hasting-James (1971). Instead of minimizing Eq(9-7), they treated the simpler problem of minimizing an instantaneous performance index:

$$\min \{ (MI_{t+1/t}^* - \text{Target})^2 + r(\nabla T_t)^2 \} \quad (9-8)$$

where  $MI_{t+1/t}^*$  is the minimum variance forecast of  $MI_{t+1}$  made at time  $t$ . This criterion usually results in controllers that, for the same constraint on the variance of  $\nabla T_t$ , have only slightly larger variance for the output than the CMV controllers. Added to it, this **Clarke's constrained (CC)** controller is much more easy to derive than the CMV one. It might be called a 'short-sighted' or 'one-step optimal controller' because it does not take into consideration the effect that the present adjustment ( $\nabla T_t$ ) will have on future outputs at lead times greater than the process time-delay.



### 9-3.3 Evaluation of the control schemes on the application

Capilla et al. (1999) used closed-loop operation, so they were concentrated to the output error after applying the control action. Under the CC controller, which was judged as the most appropriate to use in this example, the resulting output error ( $e_t$ ) of the adjusted process has been proven by Capilla et al. to follow an autoregressive moving average ARMA(1,1) process:

$$\begin{aligned} (1-\phi B)e_t &= (1-\theta B)\epsilon_t, \quad \text{where} \\ \phi &= [-w_2 - (r/w_1)] / [w_1 + (r/w_1)] \\ \theta &= -w_2 / [w_1 + (r/w_1)] \end{aligned} \quad (9-9)$$

This ARMA(1,1) process is both stationary and invertible if  $r > w_1(w_2 - w_1)$ . The authors showed that, as the constraint parameter  $r$  increases, the output-error variance increases and the variance of the temperature adjustment decreases.

The differences between the CMV and the CC controllers were found by the authors to be:

(1) The CMV controller was stable for all values of the constraining parameter  $r$ . Even if  $r = 0$ , in spite of the system being nonminimum phase, one obtains the controller that has minimum variance among all controllers with finite variance for  $\nabla T_t$ .

(2) For increasing values of  $r$  ( $r \geq 0$ ), the variance of the output increased monotonically and the variance of the adjustments decreased monotonically in the CMV controller.

If  $r = 0$ , minimizing Eq(9-8) is equivalent to minimizing the mean squared deviation of viscosity from its target as shown in Eq(9-6). The control action at time  $t$  that produces the MMSE of viscosity around its target is obtained by setting  $T_t$ , so the one-step-ahead minimum-variance forecast of MI at time  $t$ ,  $MI_{t+1/t}^*$ , equals the target value. Thus, the algorithm of the MMSE controller is obtained as a special case of the CC controller for  $r = 0$ . Accordingly, the two-step-ahead minimum variance forecast of MI taken at



time  $t$ ,  $MI_{t+2/t}^*$ , is also a special case of the CC controller for  $r = w_1 w_2$ , and so forth.

### 9-3.4 Performance of the control schemes

Capilla et al. (1999) used the MMSE controller, the CC with  $r = 0.2$  and the CC with  $r = 0.5$  for the polymerization data, and compared their performance with the actual control done by process operators (Manual) and the simulated situation in which no EPC would have resulted from setting  $T$  fixed (NO EPC). The performance measurement was the MSE of the MI under every control strategy and it is shown in Figure 9-1. Figure 9-1 implies that operators managed to reduce viscosity deviations from target by 71% from what would have resulted had temperature been fixed (NO EPC).

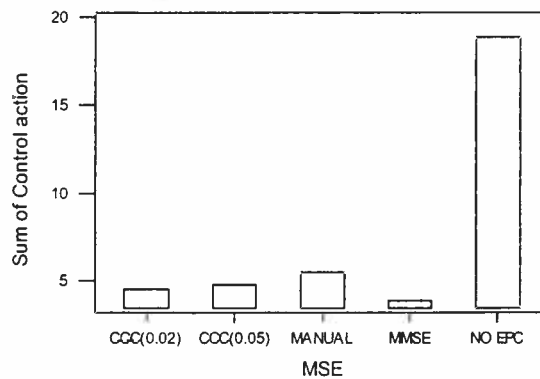


Figure 9-1: MSE of the output MI under different control strategies.

The performance of the controllers was also studied under different types of assignable causes having interfered into the process. These were: (i) MI measurement error in the laboratory, (ii) temperature sensor calibration failure, (iii) a pulse shift (outlier) in the sequence  $\{\epsilon_t\}$ , and (iv) a sustained shift (permanent change). It was proven that the EPC rules gave superior performance to no-control for every assignable cause, even when control actions were based on wrong information. However, the controllers were affected by the assignable causes. The combination of EPC with SPC tools improved the ability to detect the assignable causes in all cases.

### 9-3.5 Robustness of the control schemes

Capilla et al. (1999) also showed that, as the constraining parameter  $r$  of the CC controller increases, the stability region of the controller becomes wider and, hence, the stability robustness also increases. Even if moderate mismatch errors in the transfer function model parameters exist, the closed-loop systems studied were noticeable better strategies than the no-control systems.

In conclusion, the study of Capilla et al. (1999) detected the superiority of the CC controller compared to the MMSE for the polymerization process both in performance and robustness. Nevertheless, although this controller compensates for a moderate change in the transfer-function parameters to a large degree, the MSE of the output may increase. If the change still remains undetected, the amount of product out of specifications, and, therefore, the associated cost, can be considerably high.

## 9-4 The Monitor Wafer Controller (MWC)

A **Monitor Wafer Controller (MWC)** is a Statistical Quality Control based controller (that is, it takes corrective actions only after obtaining a statistically significant indication that the current model no longer represents the process behavior). This type of controller differentiates from the continuous controllers that assume a drift in the process outputs between consecutive runs and correct for such drifts using a drift model.

### 9-4.1 General remarks

The MWC is important in semiconductor processes. It uses periodic measurements made on selected product wafers to control the process. The monitor wafer measurements do not require in-situ sensors and can be used on existing equipment without any equipment modifications. Furthermore, a MWC does not require frequent measurements of the quality characteristic being controlled. The goal of the monitor wafer based control strategy is to determine whether the process state has been changed from its previous





estimate based on the monitor wafers, and if so, determine a set of changes to the equipment settings to generate a new recipe and bring the product quality characteristics on target. The process state is represented by the composite process models.

The MWC has been developed by Mozumder et al. (1994) and its model comprises an *intristic* and an *extrinsic* part. Together they form the composite model that represents the process state. The intristic model represents the initial state of the process. The extrinsic model transforms the inputs and outputs of the intrinsic model. Based on the monitor wafer measurements and statistical quality control, it is determined whether the state of the process is significantly different from that represented by the composite models. If the process state is different, only the extrinsic models are adapted to capture the new process state.

The adapted composite models are used for adjusting the process recipe. If the adjusted recipe results in acceptable product, it is used for future wafers, otherwise the process state is reestimated. The schematic of the controller is shown in Figure 9-2 as it has been presented by Mozumder et al. (1994). A more explicit description of the controller is followed, though more details can be found in Mozumder et al. (1994).

#### 9-4.2 Model tuning

Estimating the state of the process and updating the composite models to adapt to the new state, based on the measurements of the product parameters from monitor wafers, is termed *model tuning*. A MWC employs a layered model and a *multivariate* tuning methodology [i.e., it is not a single-input single-output (SISO) controller, but a multiple-input multiple-output (MIMO) system] that enables independence of the intristic model form.

To achieve independence from the form of the intristic model, the differences in the composite model predictions and the observations are attributed to a change in the extrinsic model. The transformations of the extrinsic model can either take the form of a gain (multiplicative factor) or offset (additive factor) applied to each of the controllable settings.





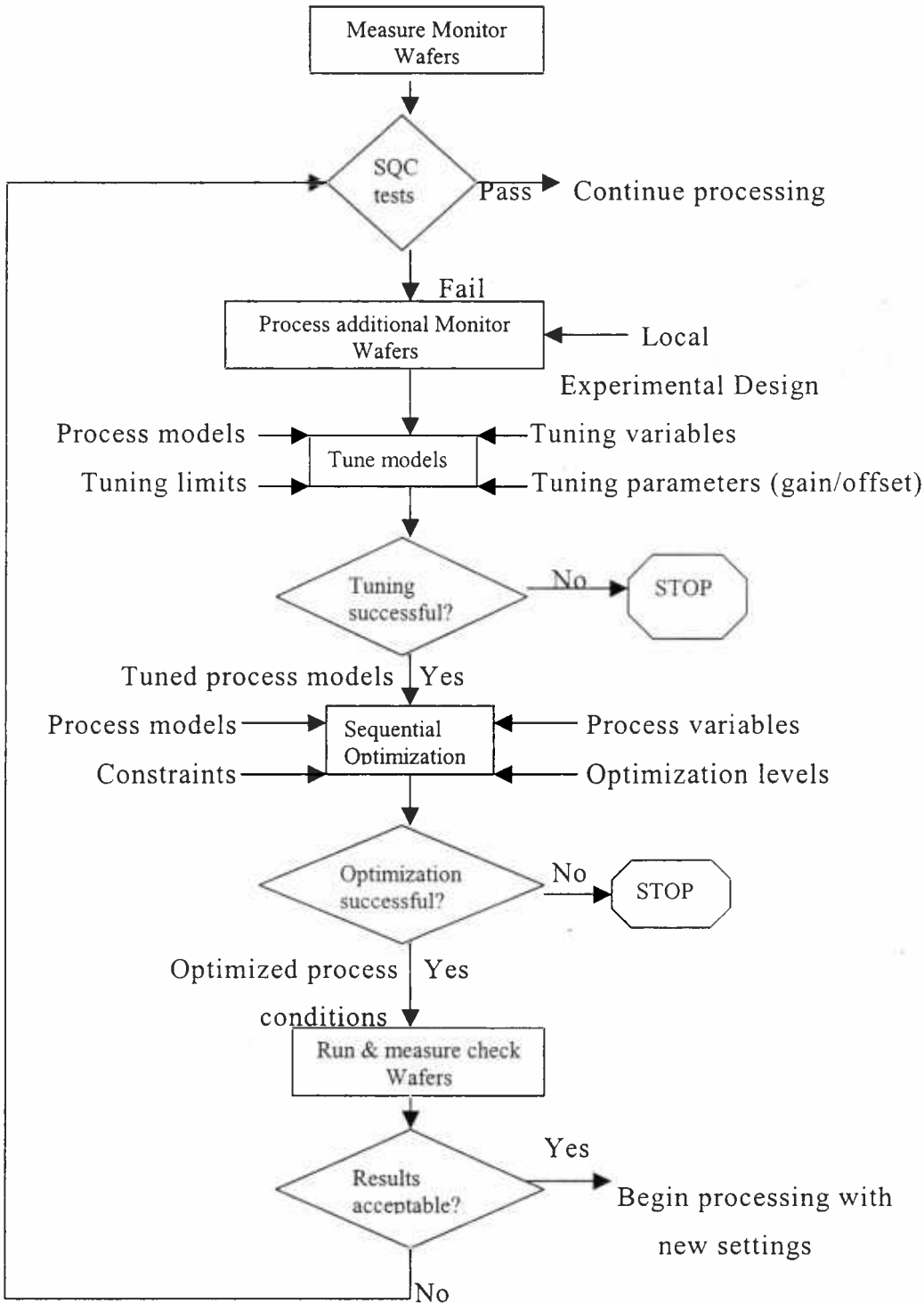


Figure 9-2: Schematic of monitor based controller.

1) *Model tuning algorithm:* Let  $\mathbf{y} = [y_1, \dots, y_m]$  represent the outputs and  $\mathbf{x} = [x_1, \dots, x_n]$  represent the inputs to the process. If  $\hat{\mathbf{y}} = [\hat{y}_1, \dots, \hat{y}_m]$  are the values predicted by the composite process models, the intrinsic

process models are represented as  $\hat{y} = F(x, \phi)$ , where  $F$  is the functional form of the intrinsic model relating  $x$  to  $y$ , and  $\phi$  represents the coefficients operating on  $x$ .

Let  $G$  and  $o$  be a diagonal matrix of gains and a vector of offsets and inputs, respectively, and  $\delta$  represent a constant 'bias' to the model (that is, the offset associated with  $\hat{y}$ ). The tuning procedure aims to determine the values  $G$ ,  $o$  and  $\delta$  such that the aggregate difference between the actual data from the monitor wafers and the corresponding predictions is minimized. The resulting tuned model can then be estimated as:

$$\hat{y} = F((Gx + o), \phi) + \delta \quad (9-10)$$

Figure 9-3 shows a schematic of the composite process and equipment model (Mozumder et al. ,1994). The model consists of 3 layers: the center block ( $F$ ) representing the intrinsic component of the model, and the two outer blocks being the extrinsic components at the input ( $Gx + o$ ), and output ( $\delta$ ), respectively.

The tuning methodology aims at tuning the three-layered model by only calibrating the extrinsic models ( $Gx + o$ ,  $\delta$ ) in Figure 9-3, leaving the intrinsic model unchanged. Since the tuning does not alter the functional form ( $F$ ) or the coefficients  $\phi$ , the tuning procedure is independent of the intrinsic model form. The effectiveness of the controller will depend on the accuracy of the intrinsic model itself and not on the intrinsic model form.

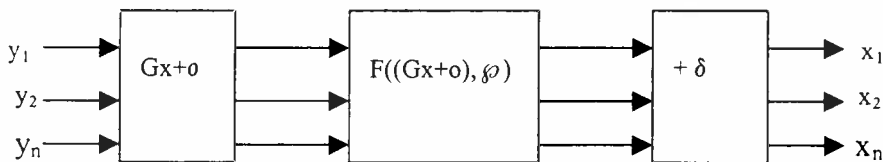


Figure 9-3: Schematic of model tuner.

If the measurements made on  $s$  monitor wafers are represented as  $y^1, \dots, y^s$  and they correspond to the input conditions  $x^1, \dots, x^s$ , then the tuning problem can be transformed into the following weighted least square minimization problem:

$$\min_{G, o} \sum_{i=1}^s \frac{1}{s_i^2} \sum_{j=1}^s (y_i^j - \hat{y}_i^j)^2, \text{ where } \hat{y}^j = \mathbf{F}((\mathbf{G}\mathbf{x}^j + \mathbf{o}), \boldsymbol{\varphi}), \text{ i.e., } \delta = 0, \text{ and} \quad (9-11a)$$

$$\min_{\delta} \frac{1}{s_i^2} \sum_{j=1}^s (y_i^j - \hat{y}_i^j)^2, \text{ where } \hat{y}^j = \mathbf{F}((\mathbf{G}^* \mathbf{x}^j + \mathbf{o}^*), \boldsymbol{\varphi}) + \delta \quad (9-11b)$$

with  $\mathbf{G}^*$  and  $\mathbf{o}^*$  being the optimal values of  $\mathbf{G}$  and  $\mathbf{o}$ , derived from Eq(9-11a). The variables  $s_i^2$  represent the estimates of variances corresponding to the prediction errors associated with  $\hat{y}_i$ . They serve as normalizing factors for the optimization so that the influence of errors for different parameters are weighted by their model prediction error variances.

The new state of the process can be presented by the tuned composite models:

$$\hat{y} = \mathbf{F}((\mathbf{G}^* \mathbf{x} + \mathbf{o}^*), \boldsymbol{\varphi}) + \delta^* \quad (9-12)$$

where  $\delta^*$  is the optimal value of  $\delta$  derived from Eq((9-11b).

2) *Selection of tuning data: Local Experiment Design.* The number of parameters measured on the monitor wafers (which usually correspond to the parameters to be controlled in the process) is often smaller than the number of input parameters to be tuned. Moreover, output parameters are often correlated, making the effective number of independent variables even smaller. Consequently, the degrees of freedom in the output may be less than the number of parameters to be tuned.

To avoid this problem local designed experiments are conducted to get sufficient observations for tuning the extrinsic models. In addition to using the measurements from the current settings, additional wafers are processed at settings which are different from the current settings. A rule of thumb is that the degrees of freedom from the output measurements should be twice or more than the number of input parameters to be fitted.

More details concerning the experiment design for tuning the controller can be found in Mozumder et al. (1994).

### 9-4.3 Stepwise Optimization

Once the extrinsic models are tuned, new equipment settings need to be determined to bring the product quality parameters to target. The new recipe is found by minimizing the difference between the model prediction from the tuned composite model and the target output values. Both input and output constraints are used in the optimization: the input constraints being the region of acceptability for the equipment settings, and the output constraints being the specification limits on the output parameters. The optimization can be formulated as:

$$\begin{aligned} \min_x \sum_{i=1}^m \frac{1}{w_i^2} [y_i^* - \hat{y}_i]^2, \text{ such that} \\ \mathbf{y}^L \leq \mathbf{F}((\mathbf{G}^* \mathbf{x} + \mathbf{o}^*), \boldsymbol{\phi}) + \boldsymbol{\delta}^* \leq \mathbf{y}^H \\ \mathbf{z}^L \leq \mathbf{H}(\mathbf{x}) \leq \mathbf{z}^H \\ \mathbf{x}^L \leq \mathbf{x} \leq \mathbf{x}^H \end{aligned} \quad (9-13)$$

where  $\mathbf{y}^L$ ,  $\mathbf{y}^H$  are the low and high specification limits on the outputs  $\hat{\mathbf{y}}$ , respectively,  $\mathbf{H}$  is any transformation on the inputs,  $\mathbf{z}^L$  and  $\mathbf{z}^H$  are the low and high specification limits on the variables defined by the function  $\mathbf{H}$ , respectively,  $\mathbf{x}^L$  and  $\mathbf{x}^H$  are the low and high limits on the inputs  $\mathbf{x}$ , respectively,  $y_i^*$  is the target output value corresponding to  $y_i$  and  $w_i$  is the weighting corresponding to the  $i$ th output parameter  $y_i$ .

It is usually desirable that the new optimal settings be close to the current settings, since the models are tuned using local data and it may be possible that all parameters will not have to be changed to get the process back on target. Therefore, the starting point for the optimization can be set to the current settings or to a predefined nominal setting (usually the optimal point derived from the untuned process).

After the controller has been optimized, the SPC charts should be applied to the adjusted process in order to detect special causes in the process.



# CHAPTER 10

## Conclusion

The manipulation of production processes having a correlation structure of some type was the question of interest in this dissertation. The most common area of the Statistical Quality Control (SQC) field, called the Statistical Process Control (SPC) area, uses control charts to assess if the process is in statistical control or if external causes operating in the process result in extreme values for the quality characteristic of interest.

If the process is completely random a variety of control charts is used in practice. One type is the Shewhart chart in which the actual observations of the characteristic are plotted and, thus, it is used to detect eventual large shifts in the mean of the process. On the other hand, in the EWMA and CUSUM charts, the accumulated observations are the ones plotted and that is why these charts are more effective whenever small shifts in the process should be found quickly. A more recent chart trying to reveal possible cycles in the process is the Spectral chart. These three types of charts could be all applied separately to the process, if the most common process shifts are to be detected.

However, the above charts are not valid in the case of autocorrelated data, because then the distinction between inherent and exterior causes is difficult to be made. Many attempts have been tried to handle this situation.

One approach is to use traditional control charts with modified control limits that take into account the autocorrelation structure of the data, as is the modified-Shewhart and the EWMAST control charts. A second approach is to apply a time series model and subtract the predicted values from the observed ones. The error terms derived in this way are almost uncorrelated (the best the prediction is, the more uncorrelated the errors are) and the application of the standard control charts is, therefore, possible. A third possibility is to use the

EWMA prediction, so that forecasted values are derived without having to apply a model to the data.

At the same time with the SPC tools, the EPC method for autocorrelated processes has been developed in the parts manufacturing industry. Though the aim of the EPC method is, similarly, to reduce variability, this is achieved by keeping the process close to a target value, so that its inherent structure does not result in excessive upward or downward trends.

More specifically, the EPC model consists of three parts: the disturbance error term that accounts for the correlation structure of the process and, therefore, takes the form of a time series model, a manipulated variable having a well-known relationship with the measured characteristic, so that adjustments of this variable bring the process close to desired values, and, finally, the random error term.

Once the two first terms have been quantified, the deviation of the random error term from the observed data becomes the output variable of the process. Because the correlation structure and any input variables effecting the process have been considered in the model, the output variable does not fluctuate uncontrolled, but stays close to a specified value. Two popular EPC techniques are the Proportional Integral Derivative (PID) and the Minimum Mean Square Error (MMSE) controllers. The first ones try to cancel out the disturbance term by adjusting the manipulated variable in terms of the output variable, while the MMSE control scheme has the objective of minimizing the output deviation from its target value.

After the process has been centered to a target, then standard SPC control charts could be applied to detect any special causes interfering into the data. This combination of the EPC/SPC techniques, known as the Automatic Statistical Process Control (ASPC) approach, is highly recommended for even more reliable results. The drawback of the ASPC method is its complexity from estimating many parameters. Estimation is the only way, though, to get an idea about the inherent structure of the process.

The best solution would obviously be to avoid correlation by sampling the observations less frequently. However, the nature of the process itself makes sometimes the dependency inevitable.



# APPENDIX



**Table A-1: Factors for constructing Variables Control Charts**

	Chart for Averages			Chart for Standard Deviations					Chart for Ranges					
	Factors for Control Limits			Factors for CenterLine	Factors for Control Limits				Factors for CenterLine	Factors for Control Limits				
n	A	A <sub>2</sub>	A <sub>3</sub>	c <sub>4</sub>	B <sub>3</sub>	B <sub>4</sub>	B <sub>5</sub>	B <sub>6</sub>	d <sub>2</sub>	d <sub>3</sub>	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>
2	2.121	1.880	2.659	0.7979	0	3.267	0	2.606	1.128	0.853	0	3.686	0	3.267
3	1.732	1.023	1.954	0.8862	0	2.568	0	2.276	1.693	0.888	0	4.358	0	2.575
4	1.500	0.729	1.628	0.9213	0	2.266	0	2.088	2.059	0.880	0	4.698	0	2.282
5	1.342	0.577	1.427	0.9400	0	2.089	0	1.964	2.326	0.864	0	4.918	0	2.115
6	1.225	0.483	1.287	0.9515	0.030	1.970	0.029	1.874	2.534	0.848	0	5.078	0	2.004
7	1.134	0.419	1.182	0.9594	0.118	1.882	0.113	1.806	2.704	0.833	0.204	5.204	0.076	1.924
8	1.061	0.373	1.099	0.9650	0.185	1.815	0.179	1.751	2.847	0.820	0.388	5.306	0.136	1.864
9	1.000	0.337	1.032	0.9693	0.239	1.761	0.232	1.707	2.970	0.808	0.547	5.393	0.184	1.816
10	0.949	0.308	0.975	0.9727	0.284	1.716	0.276	1.669	3.078	0.797	0.687	5.469	0.223	1.777
11	0.905	0.285	0.927	0.9754	0.321	1.679	0.313	1.637	3.173	0.787	0.811	5.535	0.256	1.744
12	0.866	0.266	0.886	0.9776	0.354	1.646	0.346	1.610	3.258	0.778	0.922	5.594	0.283	1.717
13	0.832	0.249	0.850	0.9794	0.382	1.618	0.374	1.585	3.336	0.770	1.025	5.647	0.307	1.693
14	0.802	0.235	0.817	0.9810	0.406	1.594	0.399	1.563	3.407	0.763	1.118	5.696	0.328	1.672
15	0.775	0.223	0.789	0.9823	0.428	1.572	0.421	1.544	3.472	0.756	1.203	5.741	0.347	1.653
16	0.750	0.212	0.763	0.9835	0.448	1.552	0.440	1.526	3.532	0.750	1.282	5.782	0.363	1.637
17	0.728	0.203	0.739	0.9845	0.466	1.534	0.458	1.511	3.588	0.744	1.356	5.820	0.378	1.622
18	0.707	0.194	0.718	0.9854	0.482	1.518	0.475	1.496	3.640	0.739	1.424	5.856	0.391	1.608
19	0.688	0.187	0.698	0.9862	0.497	1.503	0.490	1.483	3.689	0.734	1.487	5.891	0.403	1.597
20	0.671	0.180	0.680	0.9869	0.510	1.490	0.504	1.470	3.735	0.729	1.549	5.921	0.415	1.585
21	0.655	0.173	0.663	0.9876	0.523	1.477	0.516	1.459	3.778	0.724	1.605	5.951	0.425	1.575
22	0.640	0.167	0.647	0.9882	0.534	1.466	0.528	1.448	3.819	0.720	1.659	5.979	0.434	1.566
23	0.626	0.162	0.633	0.9887	0.545	1.455	0.539	1.438	3.858	0.716	1.710	6.006	0.443	1.557
24	0.612	0.157	0.619	0.9892	0.555	1.445	0.549	1.429	3.895	0.712	1.759	6.031	0.451	1.548
25	0.600	0.153	0.606	0.9896	0.565	1.435	0.559	1.420	3.931	0.708	1.806	6.056	0.459	1.541

For n > 25:

$$A = 3/\sqrt{n}, A_3 = 3/[c_4\sqrt{n}], c_4 \approx 4(n-1)/(4n-3), B_3 = 1 - [3/c_4\sqrt{2(n-1)}], B_4 = 1 + [3/c_4\sqrt{2(n-1)}], B_5 = c_4 - [3/\sqrt{2(n-1)}], B_6 = c_4 + [3/\sqrt{2(n-1)}]$$





**Table A-2: ARL Performance of the CUSUM chart with  $k=0.5$  and  $h=4, 5$**

Shift in the Mean (multiple of $\sigma$ )	$h=4$	$h=5$
0	168	465
0.25	74.2	139
0.50	26.6	38.0
0.75	13.3	17.0
1.00	8.38	10.4
1.50	4.75	5.75
2.00	3.34	4.01
2.50	2.62	3.11
3.00	2.19	2.57
4.00	1.71	2.01

**Table A-3: Values of  $k$  and the corresponding values of  $h$  that give  $ARL_0 = 370$  for the two-sided CUSUM chart**

$k$	0.25	0.5	0.75	1.0	1.25	1.5
$h$	8.01	4.77	3.34	2.52	1.99	1.61



**Table A-4: Average Run Lengths for several EWMA control schemes**

Shift in Mean (multiple of $\sigma$ )	L = 3.054 $\lambda = 0.40$	L = 2.998 $\lambda = 0.25$	L = 2.962 $\lambda = 0.20$	L = 2.814 $\lambda = 0.10$	L = 2.615 $\lambda = 0.05$
0	500	500	500	500	500
0.25	224	170	150	106	84.1
0.50	71.2	48.2	41.8	31.3	28.8
0.75	28.4	20.1	18.2	15.9	16.4
1.00	14.3	11.1	10.5	10.3	11.4
1.50	5.9	5.5	5.5	6.1	7.1
2.00	3.5	3.6	3.7	4.4	5.2
2.5	2.5	2.7	2.9	3.4	4.2
3.00	2.0	2.3	2.4	2.9	3.5
4.00	1.4	1.7	1.9	2.2	2.7



**Table A-5: ARL's of EWMAST chart with various  $\lambda$  applied to AR(1) processes**

$\phi$	Shift	$\lambda$				
		0.05	0.1	0.2	0.3	0.4
0.25	0	1,567.42	994.81	664.58	579.76	477.42
	0.5	56.94	61.47	74.15	88.99	103.00
	1	19.37	16.78	17.47	19.05	21.85
	2	7.69	6.13	5.02	4.59	4.57
	3	4.96	3.79	2.93	2.56	2.35
0.50	0	1,902.41	1,171.41	829.46	703.96	586.63
	0.5	110.60	123.63	147.31	160.60	164.94
	1	30.74	28.78	31.94	26.23	30.32
	2	10.90	8.85	7.49	10.90	9.70
	3	6.65	5.13	3.93	3.36	3.06
0.75	0	2,454.73	1,467.07	1,135.18	957.05	845.77
	0.5	296.53	330.18	333.84	321.46	308.40
	1	67.39	72.88	81.97	87.25	87.42
	2	19.15	16.12	14.92	15.05	14.17
	3	10.41	8.19	6.32	5.48	4.97
0.95	0	4,014.82	3,357.63	2,653.05	2,295.66	2,060.27
	0.5	1,846.89	1,664.52	1,376.44	1,215.05	1,076.98
	1	487.12	497.92	446.26	396.32	374.82
	2	83.47	81.59	74.14	66.03	62.01
	3	29.20	22.99	18.82	16.29	14.57
-0.25	0	1,178.36	764.98	497.10	422.01	388.95
	0.5	24.26	22.57	25.17	31.36	40.60
	1	10.02	7.83	7.06	7.56	8.44
	2	4.52	3.57	2.90	2.66	2.59
	3	3.08	2.45	1.97	1.73	1.61
-0.50	0	1,031.92	641.37	471.58	389.41	380.35
	0.5	16.00	13.86	14.85	19.19	25.44
	1	7.20	5.60	4.93	5.09	5.68
	2	3.49	2.78	2.31	2.10	2.07
	3	2.44	1.95	1.62	1.45	1.38
-0.75	0	849.48	550.23	438.21	403.16	407.28
	0.5	9.68	8.16	8.31	11.19	16.51
	1	4.79	3.75	3.35	3.49	3.86
	2	2.448	2.06	1.77	1.72	1.74
	3	1.81	1.53	1.28	1.20	1.18
-0.95	0	1,284.28	727.05	760.63	834.69	900.04
	0.5	5.08	4.43	5.02	7.60	18.32
	1	2.81	2.42	2.39	2.54	2.85
	2	1.65	1.44	1.40	1.42	1.48
	3	1.15	1.04	1.24	1.03	1.05



**Table A-6: Average Run Lengths for CUSUM charts in AR(1) processes using alternative values of K**

$\phi$	K	H	Shift $\delta$			
			0	0.5	1	2
0	0.050	17.75	739.99	39.62	19.44	9.73
	0.125	12.11	739.97	31.88	14.59	7.09
	0.250	8.02	740.02	28.83	11.42	5.22
	0.375	6.00	739.96	30.66	10.24	4.35
	0.500	4.78	740.03	35.29	9.93	3.86
	0.750	3.34	740.14	49.97	10.88	3.39
0.25	0.050	17.75	739.99	53.36	25.82	12.60
	0.125	12.11	739.97	44.96	19.75	9.16
	0.250	8.02	740.02	44.50	16.12	6.76
	0.375	6.00	739.96	51.13	15.29	5.67
	0.500	4.78	740.03	61.52	15.86	5.09
	0.750	3.34	740.14	88.13	19.89	4.68
0.5	0.050	17.75	739.99	81.76	38.99	18.51
	0.125	12.11	739.97	74.95	31.21	13.61
	0.250	8.02	740.02	83.83	28.09	10.33
	0.375	6.00	739.96	102.02	29.85	9.03
	0.500	4.78	740.03	123.50	34.38	8.58
	0.750	3.34	740.14	167.99	48.85	9.16
0.75	0.050	17.75	739.99	167.69	80.86	37.31
	0.125	12.11	739.97	175.15	73.98	29.36
	0.250	8.02	740.02	211.05	82.71	25.91
	0.375	6.00	739.96	249.59	100.73	27.23
	0.500	4.78	740.03	284.62	122.01	31.20
	0.750	3.34	740.14	342.61	165.96	43.89
0.9	0.050	17.75	739.99	355.22	206.18	99.66
	0.125	12.11	739.97	382.36	219.94	95.34
	0.250	8.02	740.02	430.93	262.70	111.25
	0.375	6.00	739.96	468.84	303.91	135.56
	0.500	4.78	740.03	498.07	339.30	160.97
	0.750	3.34	740.14	539.93	394.73	206.07



**Table A-7: Minimum Batch size required for UBM and WBM charts in AR(1) processes**

$\phi$	b	$\sigma_{UBM}/\sigma_{\epsilon}$	$\sigma_{WBM}/\sigma_{\epsilon}$
0.00	1	1.0000	NA
0.10	2	0.7454	1.1111
0.20	3	0.6701	0.8839
0.30	4	0.6533	0.8248
0.40	6	0.6243	0.7454
0.50	8	0.6457	0.7559
0.60	12	0.6630	0.7538
0.70	17	0.7405	0.8333
0.80	27	0.8797	0.9806
0.90	58	1.2013	1.3245
0.95	118	1.6827	1.8490
0.99	596	3.7396	4.0996

Note: Batch size chosen to make lag-1 autocorrelation of batch means  $\leq 0.10$ .

**Table A-8: ARMA charts compared with the corresponding optimal EWMA chart for detecting mean shifts of  $1\sigma$  when  $\phi = 0.85$**

	EWMA	ARMA chart					
	$\lambda = 0.15$	$\theta = -0.075$	$\theta = -0.05$	$\theta = -0.03$	$\theta = 0.03$	$\theta = 0.10$	$\theta = 0.30$
$\mu$	L= 2.913	L=2.832	L=2.843	L=2.868	L=2.952	L=3.023	L=3.080
0	499	503	498	496	501	503	508
0.5	36.2	35.8	35.5	35.9	36.7	40.6	62.0
1	10.3	10.3	10.2	10.1	10.8	11.0	15.6
2	3.97	4.25	4.11	4.01	3.92	3.85	4.16
3	2.56	2.94	2.78	2.69	2.47	2.25	2.00
4	2.01	2.31	2.22	2.11	1.86	1.58	1.25



**Table A-9: Comparisons of ARL's for EWMAST, Residual, Shewhart and M-M charts applied to AR(1) processes**

$\phi$	Shift	EWMAST chart		Residual	Shewhart	M-M
		$\lambda = 0.1$	$\lambda = 0.2$			
0.25	0	994.81	664.58	370.40	382.65	-
	0.5	61.47	74.15	206.04	156.65	-
	1	16.78	17.47	75.42	47.53	-
	2	6.13	5.02	12.24	7.33	-
	3	3.79	2.93	2.85	2.21	-
0.50	0	1,171.74	829.46	370.40	389.71	390.04
	0.5	123.63	147.31	258.42	170.32	378.06
	1	28.78	31.94	123.82	53.48	368.20
	2	8.85	7.49	24.22	8.94	298.37
	3	5.13	3.93	4.14	2.53	162.14
0.75	0	1,467.07	1,135.18	370.40	516.58	375.18
	0.5	330.18	333.84	311.23	235.02	374.15
	1	72.88	81.97	197.74	76.89	361.34
	2	16.12	14.92	40.24	13.69	211.68
	3	8.19	6.32	3.01	3.65	33.64
0.95	0	3,357.63	2,653.05	370.40	1,382.22	375.91
	0.5	1,664.52	1,376.44	330.96	753.16	359.23
	1	497.92	446.26	138.84	286.05	170.64
	2	81.59	74.14	1.08	46.80	1.54
	3	22.99	18.82	1.00	9.13	1.00
-0.25	0	764.98	497.10	370.40	368.82	-
	0.5	22.57	25.17	106.59	156.69	-
	1	7.83	7.06	23.30	42.06	-
	2	3.57	2.90	3.44	6.15	-
	3	2.45	1.97	1.57	1.87	-
-0.50	0	641.37	471.58	370.40	413.32	-
	0.5	13.86	14.85	61.21	165.18	-
	1	5.60	4.93	10.45	45.49	-
	2	2.78	2.31	2.11	6.10	-
	3	1.95	1.62	1.33	1.74	-
-0.75	0	550.23	438.21	370.40	483.87	-
	0.5	8.16	8.31	22.12	184.92	-
	1	3.75	3.35	3.58	51.70	-
	2	2.06	1.77	1.50	6.83	-
	3	1.53	1.28	1.00	1.65	-
-0.95	0	765.28	887.92	370.40	1,213.91	-
	0.5	4.46	4.96	2.67	440.88	-
	1	2.39	2.34	1.42	129.54	-
	2	1.50	1.42	1.00	13.80	-
	3	1.04	1.03	1.00	1.06	-

**Table A-10: Comparisons of ARL's of the Special-Cause chart (SCC), the Shewhart and the EWMA chart for various ARMA(1,1) parameters**

$(\phi, \theta)$	Shift	SCC	Shewhart	EWMA
(0.95, 0.9)	0	370.38	370.82	366.37
	0.5	272.90	163.31	62.81
	1	135.35	48.53	15.94
	2	18.53	6.91	5.54
	3	2.38	2.02	3.49
(0.95, -0.9)	0	370.38	385.13	366.44
	0.5	42.75	267.86	248.74
	1	1.00	123.08	111.62
	2	1.00	25.68	29.13
	3	1.00	1.00	10.28
(0.475, 0.9)	0	370.38	394.29	377.50
	0.5	10.53	166.97	7.33
	1	4.74	47.31	3.79
	2	2.18	5.83	2.06
	3	1.39	1.80	1.55
(0.475, 0)	0	370.38	365.34	376.53
	0.5	253.13	166.77	70.05
	1	117.96	51.05	20.69
	2	22.64	8.69	7.16
	3	4.02	2.50	4.28
(0.475, -0.9)	0	370.38	382.60	362.78
	0.5	265.34	190.65	85.90
	1	108.52	60.64	25.49
	2	2.79	11.26	8.56
	3	1.01	3.01	4.97
(0, 0.450)	0	370.38	381.31	383.02
	0.5	210.64	170.85	45.67
	1	78.83	49.01	14.53
	2	12.74	8.12	5.59
	3	2.77	2.26	3.58
(-0.475, 0.9)	0	370.38	378.60	378.17
	0.5	3.06	144.60	4.11
	1	1.94	41.85	2.36
	2	1.24	6.63	1.39
	3	1.01	1.55	1.01
(-0.475, -0.9)	0	370.38	373.96	392.93
	0.5	184.67	164.29	36.54
	1	60.11	45.92	12.036
	2	8.37	7.00	4.83
	3	2.10	2.13	3.13
(-0.95, 0.9)	0	370.38	366.86	369.74
	0.5	1.50	138.59	3.36
	1	1.00	53.16	1.98
	2	1.00	6.58	1.00
	3	1.00	1.00	1.00
(-0.95, -0.9)	0	370.38	382.56	364.04
	0.5	147.52	158.10	25.90
	1	40.04	47.00	9.15
	2	5.64	6.50	4.03
	3	1.87	2.02	2.66



**Table A-11: Comparison of Shewhart chart (Residuals, WBM and UBM) ARL's for AR(1) processes**

$\phi$	Method	b	Shift: $\delta/\sigma$			
			0.5	1	2	4
0	RES	1	2823	520	34	2
0.25	RES	1	4360	1183	116	3
	WBM	4	2066	320	23	4
	UBM	4	1279	149	11	4
	WBM	23	233	34	23	23
	UBM	23	210	32	23	23
0.50	RES	1	6521	2818	506	17
	WBM	8	2230	378	33	8
	UBM	8	1607	225	20	8
	WBM	43	397	66	43	43
	UBM	43	367	63	43	43
0.90	RES	1	9801	9234	7279	1828
	WBM	58	6119	2548	548	96
	UBM	58	5619	2133	423	81
	WBM	472	2547	823	476	472
	UBM	472	2504	809	476	472
0.99	RES	1	9995	9974	9677	4508
	WBM	596	9691	8868	6605	3238
	UBM	596	9631	8670	6178	2847
	WBM	2750	9440	8129	6605	3238
	UBM	2750	9420	8074	5434	3225





**Table A-12: ARL's and CDF's of control charts applied to optimal EWMA forecast residuals for an AR(1) process with parameter  $\phi$  and desired in-control ARL of 250**

$\phi$	$\delta$	Control chart	ARL	Number of time periods after the shift : CDF					
				1	2	3	4	5	6
0	0	Individuals	250	0.3	0.7	1.1	1.4	1.7	2.3
		CUSUM	252	1.0	1.6	2.7	3.4	3.9	4.1
		EWMA	250	0.9	1.5	1.7	2.4	2.8	3.3
	1	Individuals	33.8	3.5	5.3	7.4	10.6	13.1	14.9
		CUSUM	8.4	2.8	6.2	13.6	23.5	33.1	43.3
		EWMA	8.7	1.7	5.0	8.9	17.2	26.1	35.4
	2	Individuals	5.1	21.6	35.3	49.0	58.8	66.7	72.7
		CUSUM	3.3	5.8	27.1	60.7	83.5	94.3	97.8
		EWMA	3.9	5.1	18.8	40.4	68.5	85.4	94.4
	3	Individuals	1.8	53.4	79.5	90.4	96.2	98.4	99.1
		CUSUM	2.1	14.2	76.5	97.9	99.8	100	100
		EWMA	2.6	10.2	47.5	83.0	98.0	99.0	100
0.5	1	Individuals	242.4	2.2	3.0	3.4	3.8	3.9	4.0
		CUSUM	235.9	1.5	3.4	4.4	4.9	6.1	6.6
		EWMA	234.7	1.7	3.8	6.0	7.5	9.1	10.2
	2	Individuals	218.1	15.1	16.4	17.0	17.5	17.9	18.3
		CUSUM	184.6	4.8	14.1	20.6	25.5	27.4	28.7
		EWMA	134.4	6.7	17.0	25.1	30.2	35.7	39.0
	3	Individuals	133.2	45.5	48.5	49.8	50.1	50.4	50.6
		CUSUM	85.1	17.6	42.1	53.8	61.7	65.5	67.2
		EWMA	29.7	17.7	47.6	63.0	71.7	76.5	79.8
0.7	1	Individuals	241.3	2.6	3.2	3.6	3.8	4.0	4.5
		CUSUM	237.6	1.4	2.5	3.3	3.7	4.9	5.4
		EWMA	220.2	1.9	3.7	5.1	6.6	7.6	8.0
	2	Individuals	213.2	15.3	15.8	16.1	16.4	16.7	16.9
		CUSUM	195.4	6.2	12.1	15.8	17.2	18.1	18.4
		EWMA	194.0	5.7	12.1	15.6	18.1	20.5	22.4
	3	Individuals	129.7	47.4	48.2	48.2	48.4	48.6	48.8
		CUSUM	142.7	19.5	33.6	39.3	42.5	43.6	44.5
		EWMA	106.0	21.5	33.7	39.9	44.0	46.3	49.2
0.9	1	Individuals	240.9	3.0	3.7	4.2	4.7	5.4	5.7
		CUSUM	239.5	1.5	2.6	3.5	4.4	5.0	5.2
		EWMA	236.7	1.1	2.2	3.1	3.7	4.3	4.4
	2	Individuals	200.5	17.9	18.0	18.3	18.7	18.9	19.2
		CUSUM	216.3	5.8	9.1	10.0	10.7	11.7	12.7
		EWMA	193.5	4.4	6.1	7.6	8.5	9.5	10.2
	3	Individuals	123.7	50.0	50.2	50.4	50.6	50.8	50.9
		CUSUM	171.1	19.0	24.7	26.7	27.6	28.7	29.1
		EWMA	147.1	13.4	17.3	19.2	20.6	22.8	23.4



**Table A-13: Comparisons of ARL's: ARMAST, EWMAST and SCC on ARMA(1,1) processes**

	Process parameters		Charting parameters					
Shift	u	v	$\phi$	$\theta$	ARMAST	EWMAST	SCC	WBM
0	-0.95	0	0	-0.49	370	370	370	370
0.5					2.65	4.31	2.67	3.16
1					1.42	2.20	1.42	2.00
2					1.00	1.29	1.00	2.00
3					1.00	1.01	1.00	2.00
0	-0.475	0	0.9	0.1	370	370	370	370
0.5					13.2	14.7	65.5	17.1
1					4.78	4.97	11.4	6.27
2					2.31	2.32	2.20	2.79
3					1.64	1.63	1.35	2.02
0	0.475	0	0.9	0.1	370	370	370	370
0.5					65.6	83.3	253	65.6
1					20.3	22.4	118	25.5
2					6.61	6.17	22.6	9.78
3					3.67	3.40	4.20	5.50
0	0.95	0	0.92	0.4	370	370	370	370
0.5					226	237	331	247
1					102	108	139	136
2					25.8	25.7	1.08	60.7
3					8.65	8.30	1.00	36.7
0	0.475	-0.9	0.9	0.1	380	370	370	-
0.5					84.7	105	109	-
1					25.4	29.8	22.8	-
2					7.94	7.68	2.79	-
3					4.29	4.02	1.01	-
0	0.95	0.45	-0.9	0.1	378	370	370	-
0.5					224	226	350	-
1					95.4	97.5	275	-
2					23.6	21.9	43.5	-
3					5.14	7.15	1.30	-
0	0.95	-0.9	-0.9	36.1	370	370	370	-
0.5					42.8	240	42.8	-
1					1.00	110	1.00	-
2					1.00	26.4	1.00	-
3					1.00	8.48	1.00	-



**Table A-14: ARL comparisons of control charts on process Output and Control action (MMSE-controlled ARMA(1,1) processes)**

$(\phi, \theta)$	Shift	Output		Control	
		$(R_T, R_S)$	ARL	$(R_T, R_S)$	ARL
(0.8, -0.3)	0	(0, 0)	370	(0, 0)	370
	1	(1, 0.15)	325	(0.6, 0.46)	208
	3	(3, 0.45)	86.1	(1.8, 1.38)	33.4
	5	(5, 0.75)	2.69	(3, 2.3)	6.49
(0.7, 0.2)	0	(0, 0)	370	(0, 0)	371
	1	(1, 0.37)	208	(0.71, 0.89)	77.4
	3	(3, 1.11)	16.7	(2.13, 2.67)	5.13
	5	(5, 1.85)	1.15	(3.55, 4.45)	1.40
(-0.7, -0.2)	0	(0, 0)	370	(0, 0)	370
	1	(1, 1.42)	18.1	(0.71, 0.60)	113
	3	(3, 4.26)	1.54	(2.13, 1.80)	7.68
	5	(5, 7.10)	1.02	(3.55, 3)	1.32
(-0.8, 0.3)	0	(0, 0)	370	(0, 0)	370
	1	(1, 2.60)	4.52	(0.60, 0.86)	61.4
	3	(3, 7.80)	1.50	(1.80, 2.58)	3.15
	5	(5, 13)	1.02	(3, 4.3)	1.45

**Table A-15: ARL comparisons of control charts on process Output and Control action (PI-controlled ARMA(1,1) processes)**

$(\phi, \theta)$	$(k_p, k_I)$	Shift	Output		Control	
			$(R_T, R_S)$	ARL	$(R_T, R_S)$	ARL
(0.8, -0.3)	(0.3, 0.7)	0	(0, 0)	370	(0, 0)	370
		1	(1, 0)	356	(0.55, 0.55)	189
		3	(3, 0)	179	(1.65, 1.65)	26.7
		5	(5, 0)	9.33	(2.75, 2.75)	5.93
(0.7, 0.2)	(0.4, 0.1)	0	(0, 0)	370	(0, 0)	370
		1	(1, 0)	341	(0.8, 1.60)	42.6
		3	(3, 0)	102	(2.40, 4.80)	5.38
		5	(5, 0)	1.63	(4, 8)	1.43
(-0.7, -0.2)	(-0.4, 0)	0	(0, 0)	370	(0, 0)	370
		1	(1, 1.67)	13.2	(1, 1.67)	13.2
		3	(3, 5)	1.55	(3, 5)	1.55
		5	(5, 8.33)	1.02	(5, 8.33)	1.02
(-0.8, 0.3)	(-0.8, 0)	0	(0, 0)	370	(0, 0)	370
		1	(1, 5)	3.78	(1, 5)	3.78
		3	(3, 15)	1.48	(3, 15)	1.48
		5	(5, 25)	1.02	(5, 25)	1.02



**Table A-16: Averages of the PM's (and ARL's in parenthesis) for ASPC rules. The assignable cause is a shift in the process mean at observation 251**

Shift	Prior to shift	EPC	EPC and Shewhart	EPC and EWMA ( $\lambda = 0.1$ )	EPC and EWMA ( $\lambda = 0.4$ )	EPC+CUSUM (h=5, k=0.5)
1	2.538	2.638	2.552	2.552	2.552	2.552
			(102.1)	(112.9)	(114.7)	(105.4)
2	2.538	2.679	2.594	2.594	2.594	2.593
			(93.3)	100.7)	(101.8)	(94.9)
5	2.552	2.929	2.754	2.811	2.793	2.785
			(31.1)	(61.6)	(48.8)	(39.7)
7.5	2.544	3.298	2.962	3.033	2.929	2.943
			(3.0)	(24.4)	(12.2)	(5.7)
10	2.544	3.838	3.094	3.273	3.111	3.311
			(1.0)	(5.2)	(1.8)	(1.3)

**Table A-17: Averages of the PM's (and ARL's in parenthesis) for ASPC rules. The assignable cause is a trend that starts at observation 251**

Trend magnitude	Prior to trend	EPC	EPC and Shewhart	EPC and EWMA ( $\lambda = 0.1$ )	EPC and EWMA ( $\lambda = 0.4$ )	EPC+CUSUM (h=5, k=0.5)
0.5	2.555	3.064	2.872	2.807	2.963	2.778
			(119.9)	(119.8)	(149.2)	(111.8)
0.10	2.534	4.398	3.519	2.594	2.594	2.593
			(109.7)	(73.5)	(109.2)	(68.3)
0.25	2.543	13.963	4.085	2.811	3.467	2.291
			(60.8)	(33.9)	(49.5)	(31.6)
0.50	2.557	46.842	4.165	2.976	3.413	2.910
			(32.1)	(17.4)	(24.2)	(15.8)
1.00	2.551	179.57	3.814	2.989	3.122	2.885
			(14.3)	(9.1)	(10.0)	(7.8)



## REFERENCES

- Adams, B.M. and Tseng, I.-T. (1998).** Robustness of Forecast-Based Monitoring Schemes, *Journal of Quality Technology*, 30, 328-339
- Alwan, L.C. and Roberts, H.V. (1988).** Time-Series Modeling for Statistical Process Control, *Journal of Business & Economic Statistics*, 6, 87-95
- Beneke, M., Leemis, L.M., Schlegel, R.E. and Foote, B.L. (1988).** Spectral Analysis in Quality Control: A Control Chart Based on the Periodogram, *Technometrics*, 30, 63-70
- Bischak, D.P., Kelton, W.D. and Pollock, S.M. (1993).** Weighted Batch Means for Confidence Intervals in Steady-State Simulations, *Management Science*, 39, 1002-1019
- Box, G.E.P. and Jenkins, G.M. (1976).** *Time Series Analysis: Forecasting and Control*, 2nd 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, NJ
- Box, G.E.P. and Kramer, T. (1992).** Statistical Process Monitoring and Feedback Adjustment-A Discussion, *Technometrics*, 34, 251-267
- Box, G.E.P. and Luceno (1997).** *Statistical Control by Monitoring and Feedback Adjustment*, John Wiley, New York
- Brockwell, P.J., Davis, R.A., (1996).** *Introduction to Time Series and Forecasting*, Springer
- Brook, D. and Evans, D.A. (1972).** An Approach to the Probability Distribution of CUSUM Run Length, *Biometrika*, 59, 539-549
- Capilla, C., Ferrer, A. and Romero, R. (1999).** Integration of Statistical and Engineering Process Control in a Continuous Polymerization Process, *Technometrics*, 41, 14-28
- Chatfield, C. (1984).** *The Analysis of Time Series: An introduction*, 3rd ed., Chapman & Hall, New York
- Clarke, D.W. and Hasting-James, R. (1971).** Design of Digital Controllers for Randomly Disturbed Systems, *Proceedings of the Institute of Electrical Engineers*, 118, 1503-1506
- Cox, D.R. (1961).** Prediction by Exponentially Moving Average and Related Methods, *Journal of the Royal Statistical Society*, B23, 414-422



- Cryer, J.D. and Ryan, T.P. (1990).** The Estimation of Sigma for an X Chart:  $\overline{MR}/d_2$  or  $S/c_4$ ?, *Journal of Quality Technology*, 22, 187-192
- Del Castillo, E. (2002).** Closed-Loop Disturbance Identification and Controller Tuning for Discrete Manufacturing Processes, *Technometrics*, 44, 134-141
- Faltin, F.W., Hahn, G.J., Tucker, W.T. and Vander Wiel, S.A. (1993).** Algorithmic Statistical Process Control: Some Practical Observations, *International Statistical Review*, 61, 67-80
- Fishman, G.S. (1978).** Grouping observations in Digital Simulation, *Management Science*, 24, 510-521
- Fuller, W.A. (1976).** *Introduction to Statistical Time Series*, John Wiley, New York
- Gustavsson, I., Ljung, L. and Soderstrom, T (1977).** Identification of Processes in Closed Loop- Identifiability and Accuracy Aspects, *Automatica*, 13, 59-75
- Harris, T.J. and MacGregor, J.F. (1987).** Design of Multivariate Linear-quadratic Controllers Using Transfer Functions, *American Institute of Chemical Engineers Journal*, 33, 1481-1495
- Jiang, W. and Tsui, K.-L. (2000).** An economic model for integrated APC and SPC control charts, *IIE Transactions*, 32, 505-513
- Jiang, W. and Tsui, K.-L. (2002).** SPC Monitoring of MMSE- and PI-Controlled Processes, *Journal of Quality Technology*, 34, 384-398
- Jiang, W., Tsui, K.-L. and Woodall, W.H. (2000).** A New SPC Monitoring Method: The ARMA Chart, *Technometrics*, 42, 399-410
- Jiang, W., Wu, H., Tsung, F., Nair, V.N., Tsui, K.-L. (2002).** Proportional Integral Derivative Charts for Process Monitoring, *Technometrics*, 44, 205-214
- Kramer, T. (1989).** *Process Control From an Economic Point of View*, Ph.D. dissertation, University of Wisconsin-Madison
- Lucas, J.M. and Crosier, R.B. (1982).** Fast Initial Response for CUSUM Quality-Control Schemes: Give Your CUSUM a Head Start, *Technometrics*, 24, 199-205



- Lucas, J.M. and Saccucci, M.S. (1990).** Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements, *Technometrics*, 32, 1-12
- Luceno, A. (1998).** Performance of Discrete Feedback Adjustment Schemes With Dead Band, Under Stationary Versus Nonstationary Stochastic Disturbance, *Technometrics*, 40, 223-233
- MacGregor, J.F. (1990).** A Different View of the Funnel Experiment, *Journal of Quality Technology*, 22, 255-259
- Montgomery, D.C. (2001).** *Introduction to Statistical Quality Control*, 4th ed. John Wiley & Sons, 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. and Mastrangelo, C.M. (1991).** Some Statistical Process Control Methods for Autocorrelated Data, *Journal of Quality Technology*, 23, 179-204
- Mozumder, P.K., Saxena, S. and Collins, D.J. (1994).** A Monitor Wafer Based Controller for Semiconductor Processes, *IEEE Transactions on Semiconductor Manufacturing*, 7, 400-411
- Roberts, S.W. (1959).** Control Chart 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
- Runger, G.C., Willemain, T.R. and Prabhu, S. (1995).** Average Run Lengths for Cusum Control Charts applied to Residuals, *Communication in Statistics, Part A-Theory and Methods*, 24, 273-282
- Sachs, E., Hu, A. and Ingolfsson, A. (1995).** Run by Run Process Control: Combining SPC and Feedback Control, *IEEE Transactions on Semiconductor Manufacturing*, 8, 26-43
- Superville, C.R., Adams, B.M. (1994).** An Evaluation of Forecast-based Quality Control Schemes, *Communication in Statistics, Part B-Simulation*, 23, 645-661





- Tagushi, G. (1881).** *On-Line Quality Control During Production*, Japanese Standard Association, Tokyo
- Tseng, S. and Adams, B.M. (1994).** Monitoring Autocorrelated Data with an Exponentially Weighted Moving Average Forecast, *Journal of Statistical Computation and Simulation*, 50, 187-195
- Tsung, F. and Shi, J. (1999).** Integrated design of run-to-run PID controller and SPC monitoring for process disturbance rejection, *IIE Transactions*, 31, 517-527
- Tsung, F., Wu, H. and Nair, V.N. (1998).** On the Efficiency and Robustness of Discrete Proportional-Integral Control Schemes, *Technometrics*, 40, 214-221
- Tucker, W.T., Faltin, F.W. and Vander Wiel, S.C. (1993).** Algorithmic Statistical Process Control: An Elaboration, *Technometrics*, 35, 363-374
- 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: Concepts and an Application, *Technometrics*, 34, 286-297
- Wadsworth, H.M., Stephens, K.S. and Godfrey, A.B. (1986).** *Modern Methods for Quality Control and Improvement*, John Wiley & Sons, New York
- Wardell, D.G., Moskowitz, H. and Plante, R.D. (1992).** Control Charts in the Presence of Data Correlation, *Management Science*, 8, 1084-1105
- 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
- Zhang, N.F. (1997).** Detection Capability of Residual Chart for Autocorrelated Data, *Journal of Applied Statistics*, 24, 475-492
- Zhang, N.F. (1998).** A Statistical Control Chart for Stationary Process Data, *Technometrics*, 40, 24-38





