INTEGRATION OF DATA MINING ALGORITHMS AND CONTROL CHARTS FOR MULTIVARIATE AND AUTOCORRELATED PROCESSES

by

WEERAWAT JITPITAKLERT

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
December 2009

To my father, Wichai Jitpitaklert, my mother, Paweenrat Ampawanon,
and my sister, Srita Jitpitaklert

ACKNOWLEDGEMENTS

This dissertation would not be possible without many great people that have provided me with numerous supports. I am extremely fortunate to be so blessed to be surrounded with love and care. I would like to thank the following people who have supported my doctoral study. I would like to express my sincerest gratitude to my supervising professor, Dr. Seoung Bum Kim, for invaluable advice and knowledge for both study and personal life. He taught me discipline, consistently encouraged me to never ever give up, and patiently guided me towards my goals. I am highly grateful to my supervising committee members, Dr. Victoria C.P. Chen, Dr. Jamie Rogers, Dr. Sung Seek Moon, and Dr. Suk Young Kang, for their interest and helpful advice on this dissertation. I am grateful to Dr. Jamie Rogers and Dr. Donald H. Liles for providing me a chance to be a part of the honorable society at The University of Texas at Arlington (UTA). My appreciation is extended to my academic advisors Dr. Sheik Imrhan for his guidance and support through out my doctoral study. My thanks also go to the Industrial and Manufacturing Systems Engineering staffs: Christie Murphy, Kimetha Williams, Julie Estill, Rose, and Joyce for their assistants with the necessary administrative tasks during my graduate years.

I am grateful to all the teachers and professors who taught me during the years I spent in schools and university, first in Vajiravudh College, then in Suankularb Wittayalai School and then in Thammasat University.

I would like to thank my COSMOS colleagues: Bancha Ariyajunya, Chatabush Roongrat, Ching-feng Lin, Dr. Chivalai Temiyasathit, Dr. Huiyuan Fan, Dr. Panaya Rattakorn, Dr. Panitarn Chongfuangprinya, Passakorn Phananiramai, Poovich Phaladiganon, Dr. Prattana Punnakitikashem, Dr. Siriwat Visoldilokpun, Surachai Charoensri, Dr. Thuntee Sukchotrat, and Wei-Che Hsu, for helpful discussions and friendships. Many thanks also go to all of my friends: Dr. Ake Tonanont, Ajjaree

Limpamont, Borrom Akarajanthachot, Kotcharaht Nilrach, Kenny Maykin, Panita Suebvisai, Dr. Mathupayas Thongmak, Nitcha Suwongtham, Pisit Thanapattum, Pituck Kijpalakorn, Popon Singhapan, Punnapob Punnakitikashem, Dr. Sanya Yimsiri, Sarayuth Supakawanit, Sopis Komolchokthavee, Suparat Srisontisuk, Surbpong Trihattakarn, Thanat Thanapattum, Dr. Temyos Pandejpong, Wansamorn Chiravajr, Dr. Yodchanan Wongsawat, for always giving me useful advice and taking care of me. I also thank several of my friends who I have not stated the names here.

I would also like to extend my appreciation to Royal Thai Government for choosing me as a government scholar, for giving me a great honor to serve my country, Thailand, where I grew up in warm environment, received education from very notable academic institutions, meet good people, and made lifetime friends.

Finally, I would like to express my deepest gratitude to my father, mother, sister, niece, and nephews for their sacrifice, encouragement and patience. Thank you all so much for giving me endless love.

July 21, 2009

ABSTRACT

INTEGRATION OF DATA MINING ALGORITHMS AND CONTROL CHARTS FOR MULTIVARIATE AND AUTOCORRELATED PROCESSES

WEERAWAT JITPITAKLERT, Ph.D.

The University of Texas at Arlington, 2009

Supervising Professor: Seoung Bum Kim

The objective of this dissertation is to integrate state-of-the-art data mining algorithms with statistical process control (SPC) tools to achieve efficient monitoring in multivariate and autocorrelated process. Process monitoring and diagnosis have been widely recognized as important and critical tools in system monitoring for detection of abnormal behavior and quality improvement. Although traditional SPC tools are effective in simple manufacturing processes that generate a small volume of independent data, these tools are not capable of handling the large streams of multivariate and autocorrelated data found in modern manufacturing/service systems. As the limitations of SPC methodology become increasingly obvious in the face of ever more complex processes, data mining algorithms, because of their proven capabilities to effectively analyze and manage large amounts of data, have the potential to resolve the challenging problems that are stretching SPC to its limits. This dissertation consists of two main components; data mining model-based control charts and one-class classification-based control charts.

First, we propose a new control chart technique that integrates state-of-the-art data mining algorithms with SPC techniques to achieve efficient monitoring in multivariate and autocorrelated processes. The data mining algorithms include arti-

ficial neural networks, support vector regression, and multivariate adaptive regression splines. The residuals of data mining models were utilized to construct multivariate cumulative sum control charts to monitor the process mean. Simulation results from various scenarios indicated that data mining model-based control charts performs better than traditional model-based control charts.

Second, we examine the feasibility of using one-class classification-based control charts to handle autocorrelated multivariate processes. In recent years, statistical process control (SPC) of multivariate and autocorrelated processes has received a great deal of attention. Modern manufacturing/service systems with more advanced technology and higher production rates can generate complex processes in which consecutive observations are dependent and each variable is correlated. These processes obviously violate the assumption of the independence of each observation that underlies traditional SPC and thus deteriorate the performance of its traditional tools. The popular way to address this issue is to monitor the residuals with the traditional SPC approach. However, this residuals-based approach requires an accurate prediction model necessary to obtain the uncorrelated residuals. Furthermore, these residuals are not the original values of the observations and consequently may have lost some useful information about the targeted process. We use simulated data to present an analysis and comparison of one-class classification-based control charts and the traditional Hotelling's T^2 chart.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Statistical Process Control

Quality control and process improvement usually play an important role in strategic planning of organizations. With an appropriate quality control system, the businesses/manufacturers can maintain and continually improve the quality of products and processes. Statistical Process Control (SPC) contains a collection of problem-solving tools and has been frequently used for detecting and reducing process variability. One of the primary tools in SPC is control charts.

Control chart techniques were originally developed by Shewhart [1]. Control charts are popular because of the simplicity of use as well as graphical display capability. There are two basic components in typical control charts: monitoring statistics and control limits. Monitoring statistics could be any measurable value or function of interested process characteristics. Control limits, generally estimated from the underlying distribution of the process characteristic when the process is in control, are thresholds used to specify the in-control or out-of-control status of the process. Control charts will generate an alarm when the monitoring statistics exceed (or fall below) the control limits, then the appropriate action can be taken to correct and maintain the process quality. Figure 1.1 shows an example of Shewhart control chart along with center line (CL), upper and lower control limit (UCL and LCL). The observations were generated from a normal distribution with mean equal to ten, and standard deviation equals to one. The control limits are calculated based on three-sigma control limits. The center line is the average value of the observations. All observations lie within the control limit representing that the process is in-control.

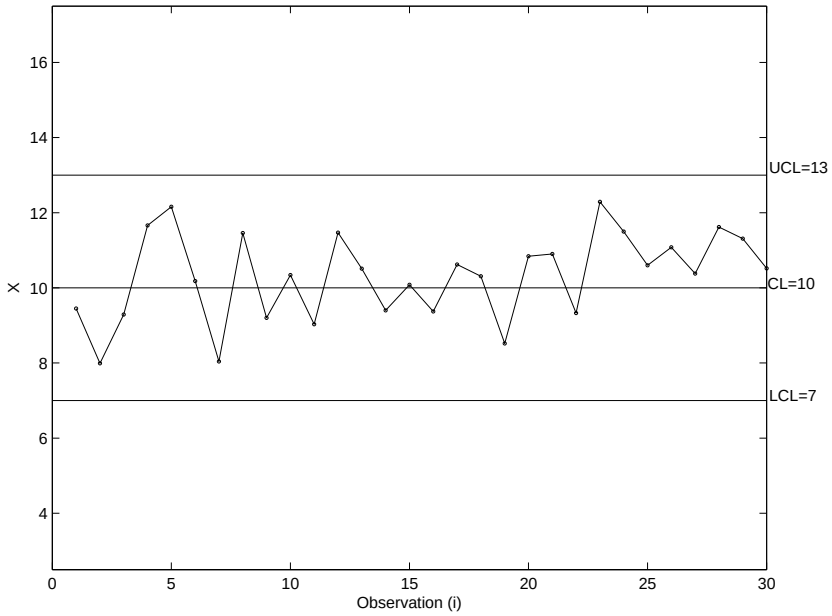


Figure 1.1. Shewhart control chart.

1.2 Multivariate and Autocorrelated Process

In this dissertation, the focus is on the monitoring of multivariate and autocorrelated process. This type of process comprises two structures; autocorrelated structure and multivariate structure. The details of each are discussed, respectively.

1.2.1 Autocorrelated Process

An autocorrelated process is usually referred to as a time-series process. In autocorrelated process, the consecutive observations are monitored at different points in time. The most distinguishing property is that the consecutive observations are unlikely to be independent. That means, the current observation is the function of the past observation and the future observation is the function of the current observation. There are many examples of this type of process such as daily stock market value, monthly unemployment figures or number of influenza cases observed over some time period [2]. The traditional SPC methodology has assumptions that the in-control data is normally and independently distributed. Thus, the autocorrelated process obviously violates this assumption. This could result in the deterioration of tradi-

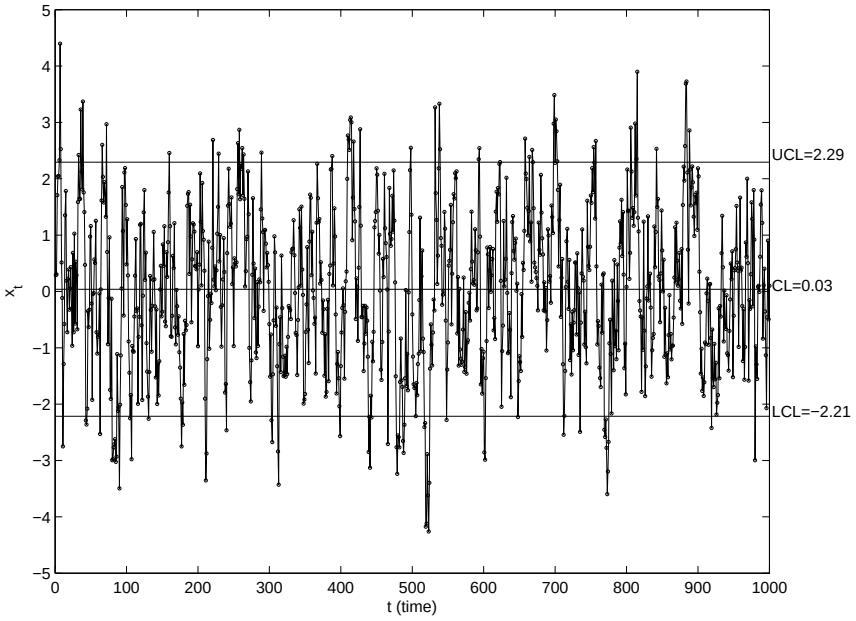


Figure 1.2. Autocorrelated process.

tional control chart performance such as decreasing of the in-control run length, and increasing of the false alarm rate [3] [4] [5] [6] [7]. Figure 1.2 shows an autocorrelated process containing 1000 observations. This process is an in-control process; however there are many points lying out of control limits representing many false alarms. In Figure 1.3, further study of the process shows a scatter plot of x_{t-1} versus x_t. The dots in the scatter plot lie from bottom left to top right of the graph revealing the positive relation between x_{t-1} and x_t. That means, as the value of x_{t-1} increase, the value x_t of would increase as well.

1.2.2 Multivariate Process

There could be more than one process variable which need to be monitored. Occasionally, these process variables can be highly correlated to each other. Monitoring each process variable individually not only takes resource, might be impractical but also could give misleading result. Therefore, there is a need to monitor multiple process variables simultaneously. Some multivariate control charts have been proposed to address with high correlation among variables are such as Hotelling's T^2 charts

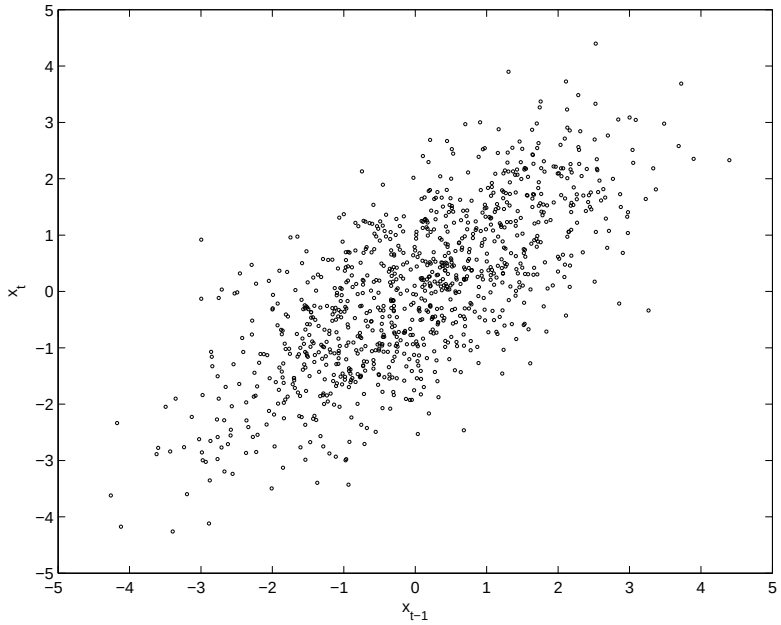


Figure 1.3. Scatter plot of x_t versus x_{t-1}.

[8], multivariate cumulative sum control chart [9] [10] [11] [12], multivariate exponentially weighted moving average control chart [13]. Multivariate control charts give a single graphical chart that simultaneously monitors all process variables instead of using multiple univariate control charts. Figure 1.4 shows a Hotelling's T^2 chart for a multivariate normal process of five dimensions. The crosscorrelation degree of the process is 0.5. The control limit is calculated with alpha equal to 0.01.

1.3 Data Mining

As more dataset gathered over time have grown in size and complexity, the need for technique to extract data into information has been increasing. Data mining is a collection of useful techniques to handle large amount of data and to provide solutions for complex situations. Data mining commonly involves four classes of tasks [14] including classification; clustering; regression; and association discovery. Classification involves arranging the data into groups based on quantitative information on one or more characteristics inherent in the items and based on a training set of previously labeled items. Famous classification algorithms include nearest neighbor, neural

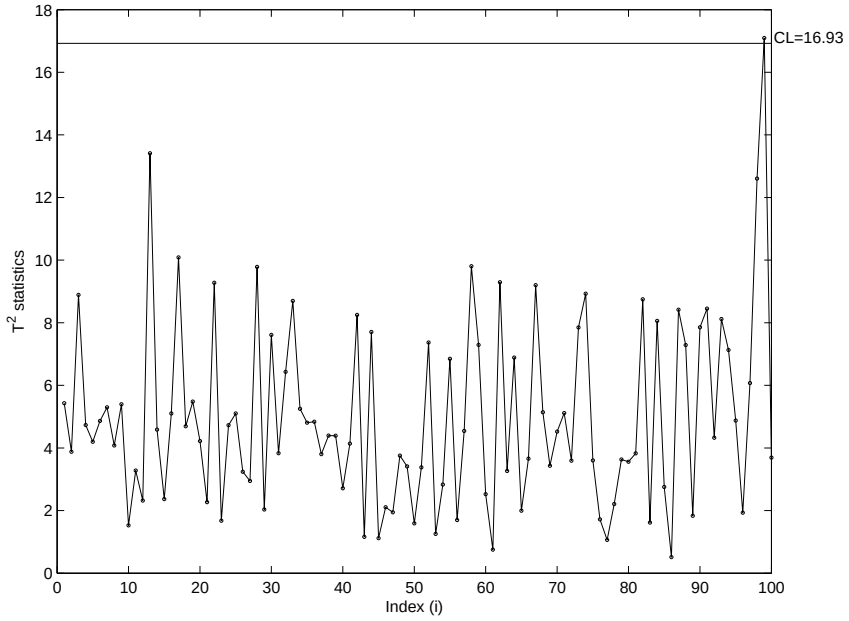


Figure 1.4. Hotelling's T^2 chart.

network, classification tree, and support vector machines. Cluster analysis involves assigning objects into groups by minimizing within-group variation and maximizing between-group variation [15]. These variations can be evaluated based on distance metrics between observations in the dataset. Regression refers to techniques for the modeling and analysis of numerical data consisting of values of a dependent variable and of one or more independent variables. Regression models often provide a description of how the independent variables affect the dependent variable. Moreover, regression models are frequently used for prediction problems. Association rule learning involves discovering interesting relations between variables and use this information for decision making. This sometimes referred to as "market basket analysis", usually employed in marketing activity planning [16].

1.4 Motivation and Contribution

Multivariate and autocorrelated process is a complex process containing multiple process variables; each variable has degree of autocorrelation, and each variable is correlated with the other variables. There have been studies regarding the effect of

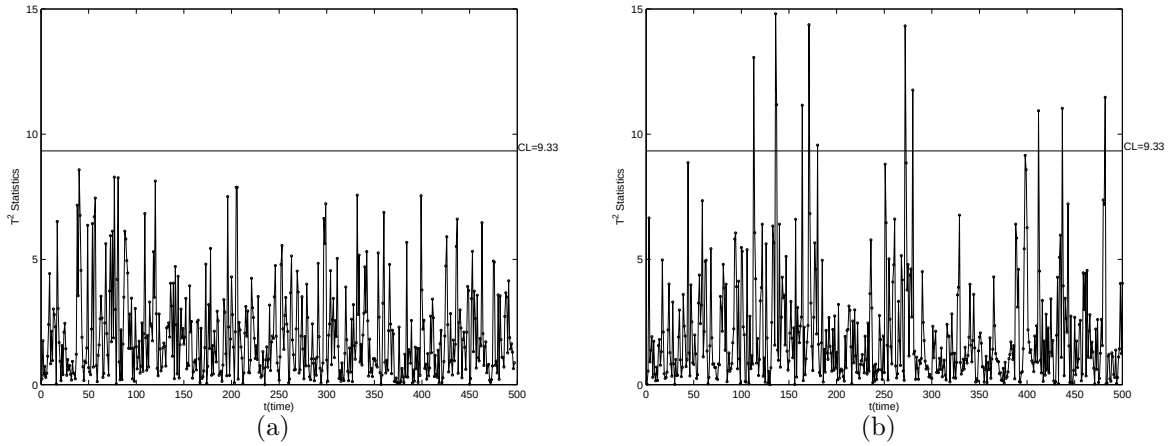


Figure 1.5. Hotelling's T^2 chart for (a) Phase I SPC (b) Phase II SPC.

autocorrelation on the performance of multivariate control charts. Many researchers conclude that the performance of control charts will be deteriorated [17] [18] [19] [20] [21] [22] [23]. To illustrate, Figure 1.5(a) and Figure 1.5(b) represent phase I and phase II SPC, respectively. In brief, SPC can be divided into two phases. The objective of phase I SPC is to identify the in-control process and construct the control limit. The objective of phase II SPC is to monitor ongoing process with control limit obtained from phase I SPC. Figure 1.5(a) shows a Hotelling's T chart in phase I SPC. The original process contains two dimensions; the autocorrelation degree equals to 0.25, and the cross correlation degree equals to 0.7. The control limit shown is calculated based on alpha equal to 0.01. All the points lie within control limit showing in-control process status. Figure 1.5(b) shows a Hotelling's T chart in phase II SPC. The process in this figure is also from in-control process as same as the process in Figure 1.5(a); consequently one would expect all the points lying within the control limit. However, an amount of points exceed the control limit, representing process false alarms. This is an example of deterioration in monitoring multivariate and autocorrelated data with traditional SPC technique.

Data mining algorithms are known as efficient approaches in dealing with various types of processes such as nonnormal process and autocorrelated process. By

the integration of data mining algorithms with SPC, the aim of this dissertation is to propose some data mining SPC approaches for monitoring multivariate and autocorrelated process.

1.5 Outline of the Dissertation

Chapter 2 introduces data mining model-based SPC control charts for multivariate and autocorrelated process. Data mining model-based methods and traditional SPC model-based methods are used to obtain the residuals. The residuals are monitored by multivariate cumulative sum control chart (MCUSUM) and the performance is compared based on average run length measures. Chapter 3 presents the feasibility of monitoring multivariate and autocorrelated process with one-class classification-based control charts. The one-class classification-based control chart which developed from the integration of k-nearest neighbors data description (kNNDD) and SPC will be used to monitor the process observations without using residuals. The performance comparison between traditional multivariate control charts and one-class classification-based control charts are presented under various simulation scenarios. Chapter 4 summarizes this dissertation and presents some ideas about future research.

CHAPTER 2

DATA MINING MODEL-BASED CONTROL CHARTS FOR MULTIVARIATE AND AUTOCORRELATED PROCESSES

2.1 Introduction

One of the key management systems in organizations is planning for quality. Organizations consider planning for quality as a part of their strategic planning. Without careful strategic planning for quality, organizations could lose large amounts of money, market share, time, and effort [24]. Therefore, business/manufacturers should focus on planning for quality as a way to develop a competitive edge in the market. Quality control and improvement include a set of activities implemented to achieve product and service specifications. SPC methodologies have frequently been used to avoid poor quality. A control chart is an important tool used in SPC to monitor the performance of a process over time to keep the process within control limits. Control charts are based on solid statistical theory and provide a comprehensive graphical display that can be readily configured by the users with minimal assistance. A typical control chart comprises the monitoring statistics and the control limits. When the monitoring statistics exceed (or fall below) the control limits, an alarm is generated so that the process can be investigated before defective units are produced.

Univariate control charts were devised to monitor the quality of a single process characteristic. However, modern processes often involve a large number of highly correlated process characteristics. Although univariate control charts can be applied to each individual characteristic, this technique may lead to unsatisfactory results when multivariate problems are involved. Moreover, high-throughput technologies in modern industries are capable of generating data for short intervals that in their brevity leads to an autocorrelation problem. Traditional multivariate control charts

were developed and came into use to solve these problems. However, they have become less and less capable of handling the large streams of complex and auto/cross-correlated data found in modern manufacturing and service systems.

Hotelling's T^2 chart is the most widely used multivariate control chart because it can simultaneously and efficiently monitor multiple correlated process characteristics. The main assumptions of T^2 control charts are the normality and independency of observed process data. That is, successive multivariate observations are assumed to be independent, identically, and normally distributed over time. Some other types of multivariate control charts include the multivariate cumulative sum (MCUSUM) control chart [9] [10] [11] [12] and the multivariate exponentially weighted moving average (MEWMA) control chart [13]. Both were devised for increased sensitivity to detect small process shifts. Although the MCUSUM and MEWMA charts are known to be relatively robust, compared with Hotelling's T^2 control chart, for non-normal and autocorrelated data, failure to use multivariate control charts carefully with autocorrelated data may result in deterioration of monitoring performance [4] [6]. Increased rates of false alarms are one possible result of such deterioration.

Model-based control charts that yield residuals - the difference between the actual values and the fitted values from the models used - have been the traditional way to address autocorrelation problems in process monitoring. Model-based control charts have been effectively used in monitoring multistage processes in which the output process variable(s) of interest are related to the input process variables from the previous and current stages [17]. A regression adjustment control chart, developed by Hawkins [25], monitors the residuals from the process variable of interest when that variable is regressed on all the others. A regression adjustment control chart is especially useful when a process variable of interest exhibits autocorrelation because the residuals from the regression model are typically uncorrelated. However, its parametric assumption of an error term in linear regression analysis limits its applicability

for handling nonnormal process data. A number of other model-based control charts are available [26] [4] [27] [28] [29].

Alwan [26] proposed a two-step approach containing two control charts, one called a common-cause chart and the other, a special-cause chart. The approach works well in detecting large process mean shifts. Montgomery and Mastrangelo [4] proposed the EWMA center line control chart. Their approach works well if the observations are positively autocorrelated and if the process mean does not drift too rapidly. Runger and Willemain [27] proposed the unweighted batch means (UBM) chart. This approach monitors the average value of observations and does not use a residual-based control chart. Zhang [28] proposed an exponentially weighted moving average for stationary process (EWMAST) chart to deal with a stationary autocorrelated process. The chart works well when the process has low positive autocorrelation and small mean shifts. Jiang et al. [29] proposed a charting technique based on autoregressive moving average statistics, the ARMA chart. All of the methods discussed above, however, deal with the occurrence of autocorrelation in univariate processes. They do not address autocorrelation in multivariate processes.

As the limitations of SPC methodology become increasingly obvious in the face of evermore complex manufacturing processes, data mining algorithms, because of their proven capabilities to effectively analyze and manage large amounts of data, have the potential to resolve the problems that are stretching SPC to its limits. Despite their great potential, however, few efforts have been made to integrate data mining algorithms with SPC. Arkat et al. [22] used artificial neural networks (ANNs) to build a model and construct a MCUSUM chart using the residuals for multivariate and autoregressive processes. They compared the average run length (ARL) performance of the three methods: autocorrelated charts, time-series-based residuals charts, and ANN-based residuals charts and concluded that ANN-based residuals charts outperformed the other two charts for small mean shifts in processes. ARL is the average number of observation required for the chart to detect a change [30].

In-control ARL (ARL_0) and out-of-control ARL (ARL_1) were, respectively, calculated under in-control and out-of-control processes. Issam and Mohamed [23] used support vector regression (SVR) to construct the residuals-based MCUSUM control chart. They calculated the residuals from one-step-ahead prediction. That is, current observations are used as input to forecast future observations. They concluded that SVR-based residuals charts performed better than time-series-based residuals charts and ANN-based residuals charts when small mean shifts were involved. This idea is interesting, but their main conclusion was derived based on limited simulation scenarios. Their studied did not investigate the different degrees of autocorrelation. Thus, their methods need to be justified much more thoroughly via simulation under various scenarios.

Our proposed approach differs from Issam and Mohamed [23] in how it finds residuals. To illustrate, for a process with three variables; x_1, x_2, and x_3, we use x_1 and x_2 as inputs to create a model that predicts x_3. The residuals of this model are obtained for monitoring x_3. We apply the same procedure to the other variables until we get the residuals from all variables. The assumption behind our proposed approach to obtain the residuals is that degrees of autocorrelation of individual process variables are not significantly different. This is a reasonable assumption because the process variables from an equipment may have similar degrees of autocorrelation. In the present study, we conducted a simulation study under various scenarios including multiple dimensions and different degree of autocorrelation.

The focus of the present study is the development of the new process monitoring methodology that can effectively deals with complex multivariate autocorrelated processes. Specifically, we use such state-of-the-art data mining models as multivariate adaptive regression splines (MARS), ANNs, and SVR. Multivariate control charts will then be used to monitor the residuals of the output variables from these data mining models.

The rest of this paper is organized as follows. In Section 2.2, we briefly explain the data mining models used for the model-based control charts. Section 2.3 illustrates the simulation study and performance comparisons among various data mining model-based control charts based on ARL measures. Section 2.4 presents our concluding remarks.

2.2 Data Mining Algorithms and MCUSUM Chart

2.2.1 Multiple linear regression

Multiple linear regression (MLR) is a parametric approach that renders a linear equation to examine the relation of the mean response to multiple predictor variables. The coefficient of each predictor variable in the linear equation is estimated by a least squares estimation technique that minimizes the summation of the squared deviation between the actual and fitted values. MLR models have been widely used for prediction problems because of their simplicity. However, MLR models may lead to inefficient and unsatisfactory conclusions when the relationship between the response and predictor variables is nonlinear. Moreover, a parametric assumption of error term in MLR often restricts its applicability to many complicated multivariate data.

2.2.2 Time-series regression

Although linear regression models are easy to implement, they do not account for the autocorrelation structure of the process. The time-series regression procedure consists of two steps. In the first step, an ordinary least square regression procedure is implemented to fit the model. Next, the autocorrelation function and the partial autocorrelation function of the residuals are employed to determine the appropriate autoregressive and moving average time-series model [31]. In the second step, the generalized least squares with a maximum likelihood estimation technique are applied to estimate the parameters of a time-series regression model.

2.2.3 Artificial neural networks

ANNs, inspired by the way biological nervous systems learn, are widely used for prediction modeling in many applications [32]. ANN models are typically represented by a network diagram containing several layers (e.g., input, hidden, and output layers) that consist of nodes. These nodes are interconnected with weighted connection lines in which these weights are adjusted as training data are presented to the ANN. The neural network training process is an iterative adjustment of the internal weights to bring the network's output closer to the desired values through minimizing the mean squared error.

In the present study we used the same parameter setup of an ANN as was done in a previous study by Issam and Mohamed [23]. To be specific, We used three layers consisting of input, hidden, and output layers. We used mean squared error (MSE) as a stopping criterion. The neural network model stops training if no significant change in the value of MSE occurs in two consecutive epochs. The activation function we used was Purelin, which is a linear transfer function. Learning rate and momentum rates were, respectively, 0.05 and 0.1.

2.2.4 Support vector regression

SVR is a regression version of the support vector machines algorithm. The basic idea of SVR is to find a function $f(x)$ that predicts the response variable based on the predictors with the maximum acceptance error. Another requirement for the function $f(x)$ is that it should be as flat as possible [33]. Thus, the parameters are estimated by solving a convex optimization problem. SVR is capable of handling nonlinearity by using kernel functions that map input space into a new feature space. We used the Gaussian radial kernel function and its related parameters in the SVR models as the one done by Issam and Mohamed [23].

2.2.5 Multivariate adaptive regression splines

Multivariate adaptive regression splines (MARS) is a method for estimating a completely unknown relationship between a response variable (performance measurement) and several predictor variables [34]. It is one of the few tractable methods for high-dimensional problems with interactions. MARS is a data-driven statistical linear model with a forward stepwise algorithm to select model terms, which is followed by a backward procedure to prune the model. The approximation bends at "knot" locations into a model curvature, and one of the objectives of the forward stepwise algorithm is to select appropriate knots. Smoothing at the knots is an option that may be used if derivatives are desired.

2.2.6 Multivariate cumulative sum control chart (MCUSUM)

Crosier [11] extended the univariate CUSUM scheme into vector-valued CUSUM. The univariate CUSUM scheme can be represented as follows:

$$S_n = max(0, S_{n-1} + (X_n - a) - k\sigma), \tag{2.1}$$

where a is the target value for the mean, σ is the standard deviation of the X's, k is the reference value, which is often chosen about halfway between the target mean and the out-of-control mean, and S_0 is the starting value and is set equal to zero [24]. By replacing the scalars in (2.1) with vectors, we can extend univariate CUSUM into multivariate CUSUM.

H is the decision interval or threshold to decide if the process is in control. Crosier [11] explained that the desired ARL_0 has to be specified first, then we can manually adjust the H value to yield the desired ARL_0. The ARL_0 value is user-defined. In his work, Crosier [11] set the target ARL_0 at 200, then adjusted the H value for each simulation scheme. Each scheme will have a different H value. The problems are how to find k, and how to interpret taking the maximum of a vector and the null vector [11]. Crosier [11] has addressed this issue clearly in his work in which

he also recommended 0.5 for the k value in a MCUSUM control chart. According to Crosier [11], the calculation of a MCUSUM control chart can be demonstrated as follows:

$$C_n = [(S_{n-1} + X_n - a)^T \Sigma^{-1} (S_{n-1} + X_n - a)]^{1/2}, \tag{2.2}$$

where a is the target value of the process mean vector, X_n is the vector value of the process, Σ is the covariance matrix of the process, and n is the number of observations.

$$S_n = 0 \qquad if C_n \leq k,$$
$$S_n = (S_n + X_n - a)(1 - \frac{k}{C_n}) \qquad if C_n \geq k,$$

where $S_0 = 0$ and $k > 0$ Let

$$Y_n = [S_n^T \Sigma^{-1} S_n]^{1/2}, \tag{2.3}$$

where Y_n is the monitoring statistic in the MCUSUM control chart and H is the decision interval. MCUSUM control charts would generate an alarm when Y_n exceeds the threshold H. In the present study we set the k value equal to 0.5, which is the value that typically has been used [11] [13] [22] [23].

2.3 Simulation

2.3.1 Simulating multivariate and autocorrelated data

A simulation study was conducted to examine the performance of the proposed data mining model-based control charts under various scenarios. Multivariate autoregressive datasets were generated by a stationary vector autoregressive model [20]. A vector autoregressive (VAR) model consists of the following three components: a process mean vector (μ) , an autoregressive coefficient matrix (Φ) , and a covariance matrix of the residuals (Σ_r). If the multivariate autoregressive processes of m dimensions contain autocorrelation of an order p, we can express the VAR model as follows:

$$X_t = \mu + \Phi_1 (X_{t-1} - \mu) + ... + \Phi_p (X_{t-p} - \mu) + \epsilon_t, \tag{2.4}$$

where X_t is the m-dimensional process vector, μ is the $m \times 1$ process mean vector, (Φ) is the $m \times m$ autoregressive coefficient matrix, and ϵ_t is the m-dimensional white noise process vector with zero mean and covariance matrix (Σ_r). To ensure that the process is stationary, the autoregressive coefficient matrix needs to be a positive definite matrix. Equivalently, all eigenvalues of the autoregressive coefficient matrix need to be less than one [35] [36].

2.3.2 Simulation scenarios

Table 2.1 shows a summary of simulation scenarios. The simulations start from low-positive autocorrelation to medium- and high-positive autocorrelation, and from processes with two dimensions to processes with five and ten dimensions. The detailed information of the parameter values can be found in Appendix A.

As a way of quantifying the magnitude of the shift in the out-of-control data in a multivariate setting, we define the noncentrality parameter. Let μ_0 and Σ_X be, respectively, the mean vector and the covariance matrix of a multivariate process when there is no shift in the mean process. Let $\mu_1 = \mu_0 + \delta$ be the mean vector shift.

The noncentrality parameter λ is defined by

$$\lambda = \sqrt{\delta^T \mathbf{\Sigma}_{\mathrm{X}}^{-1} \delta}, \tag{2.5}$$

where δ is the magnitude of the shift. In this study, the process mean is shifted for ten cases and is shifted equally in all dimensions. Each process contains 1000 observations. Table 2.2 shows the noncentrality parameter values and the amount of mean shift in each dimension. The noncentrality parameter value is set from very small value (0.05λ) to large value (3λ). Appendix C provides a detailed description of the noncentrality parameter along with a numerical example.

2.3.3 Simulation results

We compared six different model-based control charts (none, multiple linear regression model, time-series regression model, ANN, MARS, and SVR) under the

nine simulation scenarios shown in Table 2.1. We considered the scenarios with different numbers of dimensions and different degrees of autocorrelation.

For comparison, we used ARL_1 from the model-based MCUSUM charts. In general, we prefer the procedure that produces lower ARL_1 given the similar values of ARL_0. The threshold H in the model-based MCUSUM charts is manually adjusted so that the values of ARL_0 in the different control charts are approximately the same at 200. This threshold value was used to monitor the out-of-control process when the process mean is shifted, and we can calculate ARL_1 for the different model-based MCUSUM charts. In this simulation, the average value of ARL was calculated from 1000 replications. Figure 2.1, 2.2, and 2.3 show the values of ARL_1 obtained from six different model-based MCUSUM charts against different mean shifts. ARL_1, obtained from data mining model-based MCUSUM charts (ANN, MARS, SVR), is shown by a solid line; ARL_1, obtained from traditional model-based MCUSUM methods (none, multiple linear regression, time-series regression), is shown by a dashed line. The standard error of 1000 replications is less than 0.01.

All simulation scenarios returned similar results in that data mining model-based MCUSUM charts yielded a smaller ARL_1 than traditional model-based MCUSUM charts. To put it simply, on average, data mining model-based methods can detect an out-of-control status quicker than the traditional methods. The difference can be seen clearly in small mean shifts. For large mean shifts, because the process mean shifts are large, all methods can readily detect the shifts.

Of the three data mining model-based MCUSUM control charts, SVR performed the best in Scenarios 2.1(a), 2.1(c), 2.2(c), and 2.3(a). ANN performed the best in Scenarios 2.1(b), 2.2(b), 2.3(b), and 2.3(c). Both performed comparably in Scenario 2.2(a). Compared with SVR and ANN, MARS performed the worst in all nine simulation scenarios. Among the traditional model-based MCUSUM control charts, the MCUSUM chart without using any models performed the worst; the time-

series regression model-based and multiple linear regression-based MCUSUM charts were comparable performers.

The results also revealed that as the degree of autocorrelation increases, the control charts would have higher values of ARL_1. This can be seen by comparing three panels in each of Figures 2.1, 2.2, and 2.3. This may indicate that higher autocorrelation deteriorates the ability of control charts for rapid detection of the shift.

Further, it is interesting to observe in Figures 2.1(c), 2.2(c), 2.3(a), 2.3(b), and 2.3(c) that the performance of data mining model-based MCUSUM charts is clearly superior to traditional model-based MCUSUM charts. This demonstrates that data-mining model-based MCUSUM charts performed better especially in higher positive autocorrelation and higher process dimensions.

2.4 Concluding Remarks

This study proposes model-based control charts based on data mining algorithms. The proposed charts address a growing need in process control for a way to deal with correlation among variables and autocorrelation within variables without introducing unreliability that would be marked by increasing rates of false alarms. Three data mining model-based techniques and three traditional techniques were compared in this study based on a measurement of ARL performance. Given similar ARL_0, the preferred techniques are those that yield the smaller ARL_1. The simulation results, based on 1000 replications, indicated that data mining model-based techniques, especially ANN and SVR, performed better than traditional model-based techniques and much better than direct monitoring of a cross/auto-correlated process. The difference in performance is obvious in smaller mean shifts. In addition, data mining model-based control charts also performed better in higher positive autocorrelation processes and in high-dimensional processes. Therefore, these results show

that data mining can provide a sound and promising solution for multivariate and autocorrelated process control.

Table 2.1. Simulation scenarios

Scenarios	Number of Dimensions	Autocorrelation Degree (Coefficient in Autoregressive Coefficient Matrix (Φ))	Crosscorrelation Degree (Coefficient in Correlation Matrix)
1	2	Low positive (0.25)	0.7
2	2	Medium positive (0.50)	0.7
3	2	High positive (0.75)	0.7
4	5	Low positive (0.25)	0.6
5	5	Medium positive (0.50)	0.6
6	5	High positive (0.75)	0.6
7	10	Low positive (0.25)	0.5
8	10	Medium positive (0.50)	0.5
9	10	High positive (0.75)	0.5

Table 2.2. Noncentrality parameter values and individual mean shifts for 2-dimension, 5-dimension and 10-dimension scenarios (All dimensions are shifted equally as the values shown in bracket)

No.	λ	2 dimensions Shift	5 dimensions Shift	10 dimensions Shift
1	0.05	[0.46]	[0.25]	[0.13]
2	0.10	[0.95]	[0.50]	[0.26]
3	0.15	[1.45]	[0.70]	[0.39]
4	0.20	[1.90]	[0.90]	[0.52]
5	0.25	[2.30]	[1.20]	[0.64]
6	0.50	[4.60]	[2.50]	[1.29]
7	1.00	[9.00]	[5.00]	[2.58]
8	1.50	[14.00]	[7.00]	[3.86]
9	2.00	[18.00]	[9.50]	[5.15]
10	3.00	[27.00]	[13.50]	[7.73]

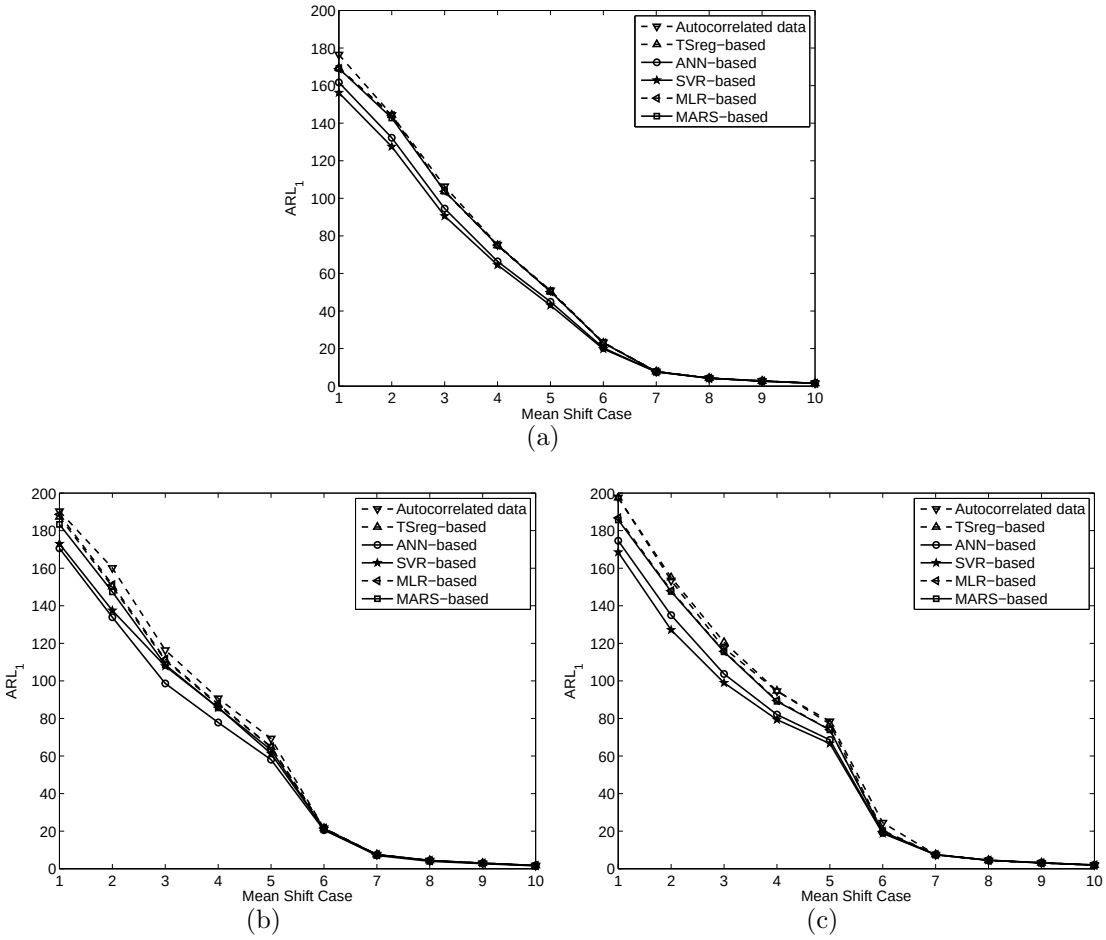


Figure 2.1. Out-of-control ARL (ARL_1) for six different control charts with two dimensions and (a) low (b) medium, and (c) high positive autocorrelation. The maximum standard error from 1000 replications is 0.01.

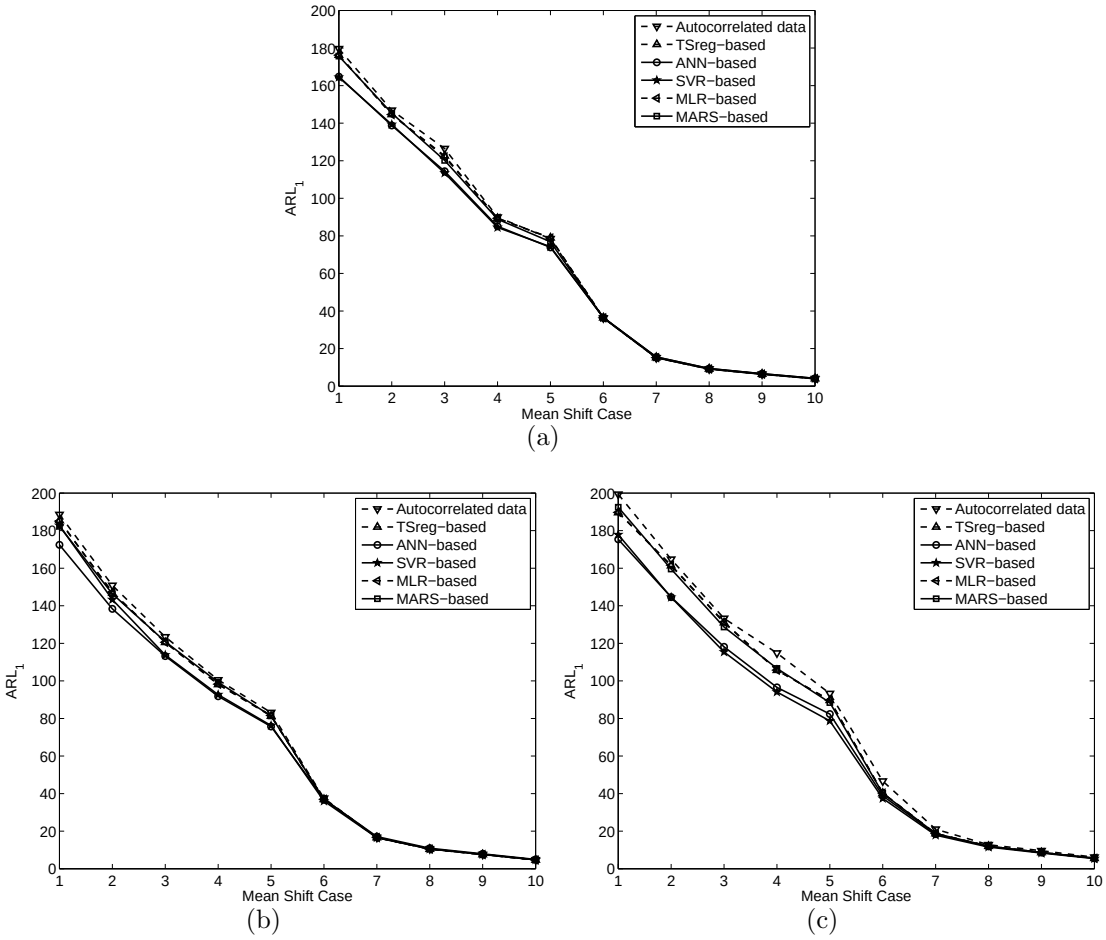


Figure 2.2. Out-of-control ARL (ARL_1) for six different control charts with five dimensions and (a) low (b) medium, and (c) high positive autocorrelation. The maximum standard error from 1000 replications is 0.01.

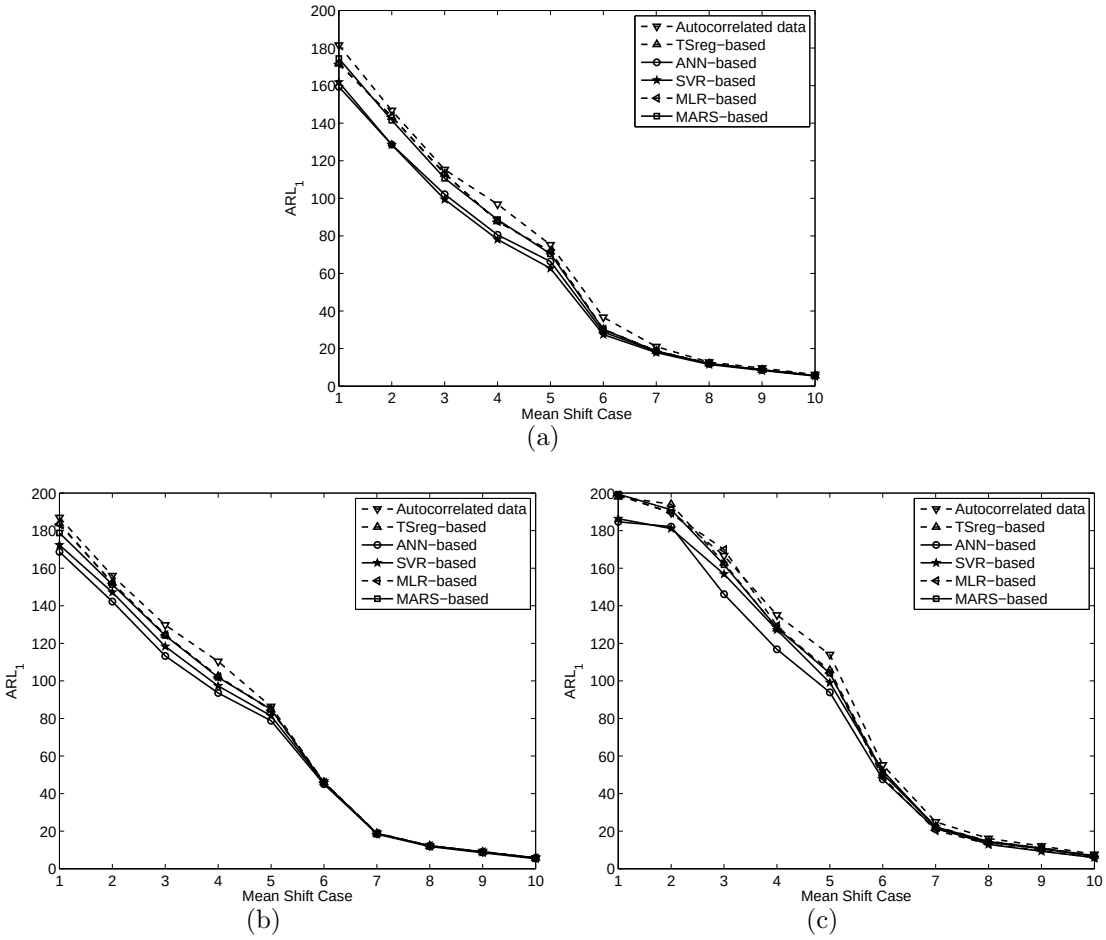


Figure 2.3. Out-of-control ARL (ARL_1) for six different control charts with ten dimensions and (a) low (b) medium, and (c) high positive autocorrelation. The maximum standard error from 1000 replications is 0.01.

CHAPTER 3

ONE-CLASS CLASSIFICATION-BASED CONTROL CHARTS FOR MONITORING MULTIVARIATE AND AUTOCORRELATED PROCESSES

3.1 Introduction

Statistical process control (SPC) is one of the most widely used techniques for quality control. One of the important tools in SPC is a control chart that monitors the performance of a process over time to keep the process in control. Control charts have been widely used because of their excellent capability to generate graphical output so that users can readily interpret the outcomes of control charts.

Although traditional control charts are effective in simple manufacturing processes that generate a small volume of independent data, these charts falter when confronted by the large streams of complex and correlated data encountered in modern manufacturing systems. Most traditional control charts assume that the process is independent and identically distributed. However, high-throughput technologies in modern industries are capable of generating short-interval data that leads to an autocorrelation problem. The unguarded use of traditional control charts in an autocorrelated process results in deterioration of their monitoring performance in such ways as a decrease in the length of the in-control run length and an increase in the false alarm rate [3] [4] [5] [6] [7].

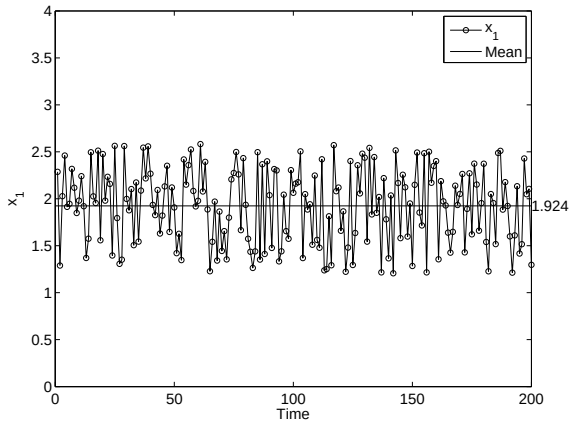
In addition to autocorrelation, the processes used in modern manufacturing systems involve a number of process variables that are correlated with each other. Many multivariate control charts have been developed to handle multivariate processes. These include Hotelling's T^2 charts [8], multivariate exponentially weighted moving average control charts [13], and multivariate cumulative sum control charts [9] [10] [11] [12]. Despite their effectiveness in multivariate processes because they

take into account the correlation of the process variables, most of the existing multivariate control charts require their observations to be independent of each other (uncorrelated). Some studies have investigated the effect of autocorrelation on the performance of multivariate control charts and have concluded that autocorrelation hampers their performance [17] [37] [18] [19] [20] [21] [22] [23].

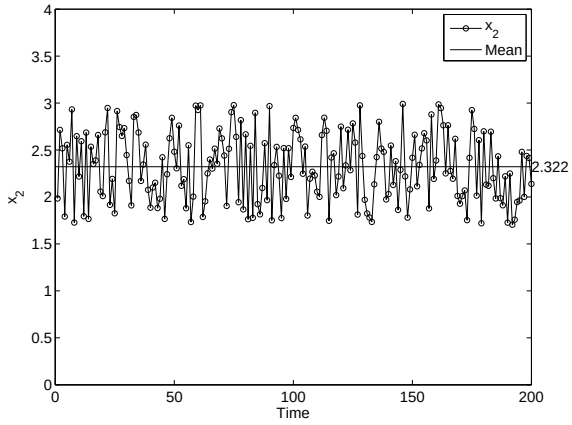
Figure 3.1 and Figure 3.2 give examples of Hotelling's T^2 charts for an unautocorrelated multivariate process and an autocorrelated multivariate process, respectively. Figure 3.1 displays the time-sequence plot of two process variables and the plot of the T^2 chart for this process. The time-sequence plots of each variable show random, nonpattern fluctuation, and no upward or downward trend exists. As expected from the unautocorrelated observations of process variables, the T^2 chart shows no systematic patterns. Moreover, the T^2 values are near zero, indicating that the process observations fluctuate around the mean value of the process.

Figure 3.2 illustrates the effect of autocorrelation on the multivariate process of the T^2 chart. The plot of x_1 shows the upward trend, and the plot of x_2 shows the fluctuation pattern and upward movement. The observations of each process variable obviously are not maintained at the mean value. The T^2 chart for this autocorrelated multivariate process looks entirely different from the previous T^2 chart shown in Figure 3.1. This T^2 chart shows a systematic pattern, and most of the T^2 values are not near zero. This is evidence that a variation in process variables has caused the variation in the T^2 chart. Because the T^2 is a squared statistic, such trends and variations in the process variables would produce large T^2 values [37]. The above simple example argues strongly for the need to develop efficient control charts to monitor autocorrelated multivariate processes.

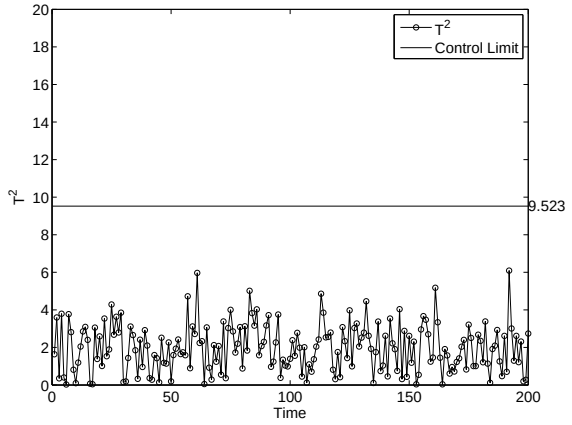
Although autocorrelation and crosscorrelation issues often occur concurrently in modern process systems, to this point research on these issues in control charts has been conducted separately. Model-based control charts that use residuals have been the way to monitor autocorrelated multivariate processes. The residual is the



(a)

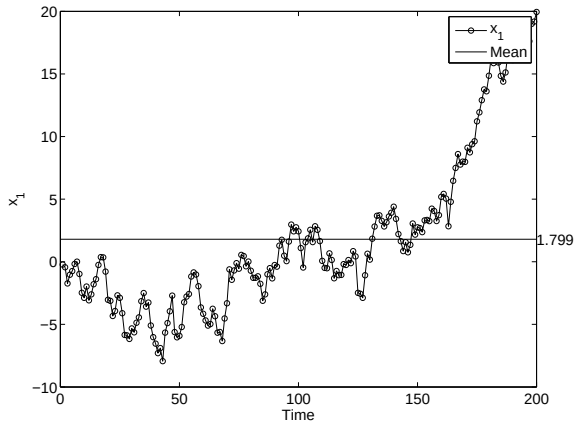


(b)

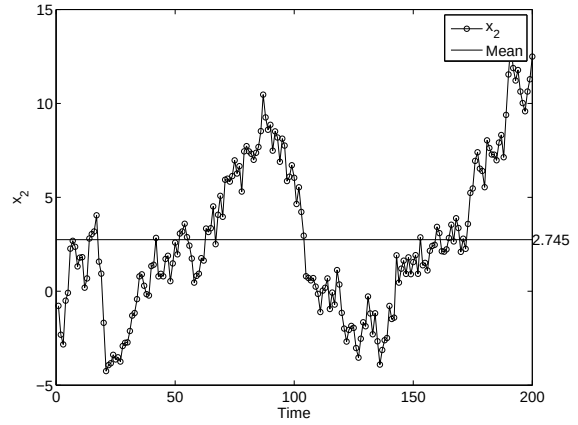


(c)

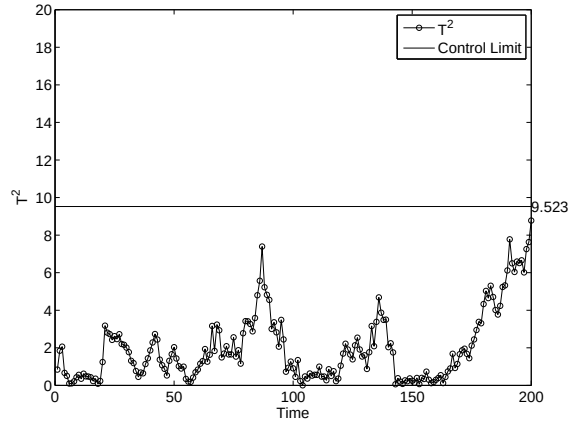
Figure 3.1. (a) Plot of variable x_1 without time dependency (b) Plot of variable x_2 without time dependency (c) Plot of T^2 chart for unautocorrelated multivariate process.



(a)



(b)



(c)

Figure 3.2. (a) Plot of variable x_1 without time dependency (b) Plot of variable x_2 without time dependency (c) Plot of T^2 chart for autocorrelated multivariate process.

difference between the actual values and the fitted values from the model. These model-based control charts usually have two steps. The first step is an effort to obtain the residuals. The second is the construction of control charts based on these residuals, which typically are uncorrelated if the prediction model is properly developed.

In model-based control charts, an accurate prediction model is necessary to obtain the uncorrelated residuals. The prediction models that have been used include time series models [38] [22] [23], regression models [25], and data mining models [22] [23]. Although all of these methods perform reasonably well within the experimental settings for which they have been designed, no consensus exists about which of them best satisfies all conditions. Moreover, because the residuals are not the original values of the observations, they cannot be readily interpreted, and the extraction of meaningful information is cumbersome.

The main objective of the present study is to examine the feasibility of one-class classification (OCC)-based control charts as a way to efficiently monitor autocorrelated multivariate processes. To be specific, we implemented an OCC control chart based on a k nearest-neighbors data description (kNNDD) algorithm [39]. The OCC control charts overcome a limitation posed by the model-based control charts and can be constructed without losing any information from the original process data. To the best of our knowledge, the present study is the first attempt to propose actual data-based control charts for monitoring autocorrelated multivariate processes.

The rest of this paper is organized as follows. In Section 3.2, we briefly explain kNNDD-based OCC control charts . Section 3.3 presents the simulation study used to explore the performance of OCC control charts and compare them with the traditional Hotelling's T^2 charts in terms of Type I and Type II error rates. Section 3.4 presents our concluding remarks.

3.2 kNNDD-Based OCC Control Chart (K^2 Chart)

Recently, Sukchotrat el at. [40] developed an K^2 chart that integrates a traditional control chart technique with a kNNDD algorithm, one of the one-class classification algorithms. A kNNDD algorithm solves one-class classification problems by estimating the local density of the data [41] [39]. Let z be a data point from a training dataset, k be the number of nearest-neighbor data points of point z. A cell or a hypersphere in p dimensions will contain a data point z from the training dataset. The volume of this cell will expand until it contains k (a user-specified value) nearest-neighbor data points from the training dataset [39]. The local density of point z is then estimated by:

$$d(\mathbf{z}) = \frac{i/N}{V\|\mathbf{z} - \mathrm{NN}_i(\mathbf{z})\|}, \tag{3.1}$$

where V is the volume of the cell, $\mathrm{NN}_i(\mathbf{z})$ be the i^{th} nearest neighbor training observation of a data point $\mathbf{z}$.

Likewise, the local density of $\mathrm{NN}_i(\mathbf{z})$ will be:

$$d(\mathrm{NN}_i(\mathbf{z})) = \frac{i/N}{V\|\mathrm{NN}_i(\mathbf{z}) - \mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))\|}, \tag{3.2}$$

where $\mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))$ is the i^{th} nearest neighbor of $\mathrm{NN}_i(\mathbf{z})$ in the same training dataset. For decision criteria, the kNNDD method will classify data point z as the target class if the ratio of the local density of z to the local density of $\mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))$ is greater than or equal to 1, as shown below:

$$\frac{d(\mathbf{z})}{d(\mathrm{NN}_i(\mathbf{z}))} = \frac{\|\mathrm{NN}_i(\mathbf{z}) - \mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))\|}{\|\mathbf{z} - \mathrm{NN}_i(\mathbf{z})\|} \geq 1. \tag{3.3}$$

To calculate the average of k nearest neighbors, the equation (3.3) will become:

$$\frac{\sum_{i=1}^{k} \|\mathrm{NN}_i(\mathbf{z}) - \mathrm{NN}_i(\mathrm{NN}_i(\mathbf{z}))\|}{\sum_{i=1}^{k} \|\mathbf{z} - \mathrm{NN}_i(\mathbf{z})\|} \geq 1. \tag{3.4}$$

The K^2 monitoring statistics is calculated as follows:

$$K^2 = \frac{\sum_{i=1}^{k} \|\mathbf{z} - \mathrm{NN}_i(\mathbf{z})\|}{k}. \tag{3.5}$$

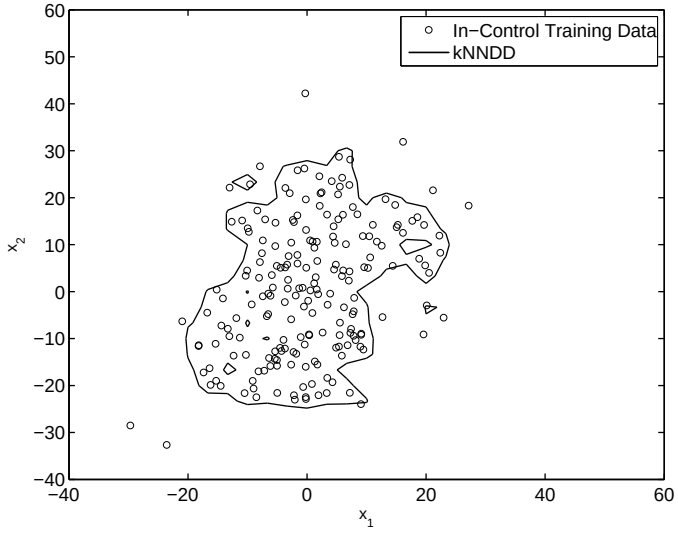


Figure 3.3. Control boundary of kNNDD constructed from an autocorrelated multivariate process.

The control limits of a K^2 chart are calculated by a bootstrap-based percentile procedure. A detailed description for designing a K^2 chart can be found in Sukchotrat el at. [40]. Figure 3.3 shows an example of a kNNDD control boundary constructed from an autocorrelated multivariate process. Figure 3.4 shows an example of a K^2 chart for an autocorrelated multivariate process.

3.3 Simulation

3.3.1 Simulation setup

A simulation study was conducted to examine the performance of the K^2 charts in autocorrelated multivariate processes and compare them with Hotelling's T^2 charts under various scenarios. To provide the simulated data, the vector autoregressive processes (VAR) models with two dimensions were used with different degrees of autocorrelation. Table 3.1 summarizes the simulation process configuration.

For the K^2chart, the parameter k should be determined before its construction. In general, different values of k are examined to find the best one that produces the smallest error rate. Here, we used $k = 2$. A training dataset contains 500 observations,

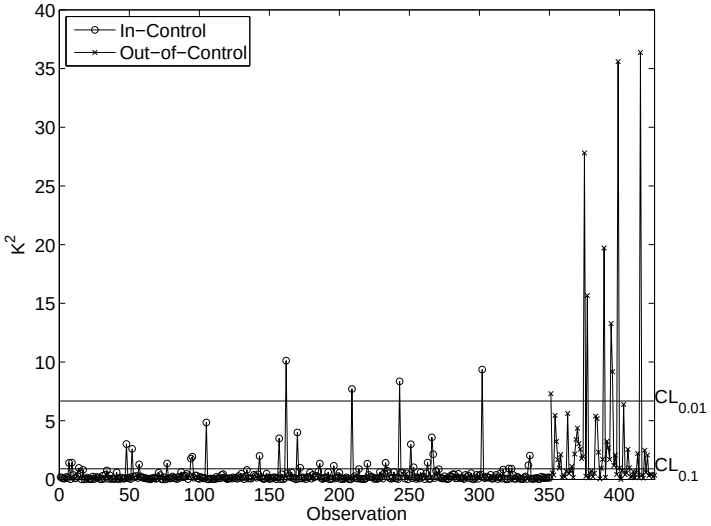


Figure 3.4. K^2 chart for autocorrelated multivariate process.

Table 3.1. Simulation scenarios

Scenarios	Autocorrelation Degree (Coefficient in Autoregressive Coefficient Matrix (Φ))	Crosscorrelation Degree (Coefficient in Correlation Matrix)
1	[0.25 0.75] (Mixed positive)	0.5
2	[0.25 0.25] (Low positive)	0.5
3	[0.50 0.50] (Medium positive)	0.5
4	[0.75 0.75] (High positive)	0.5

and the testing dataset contains $1,000$ observations (500 in-control and 500 out of control). In this simulation, the averages of Type I and Type II error rates were calculated from $1,000$ replications. A Type I error rate is defined as the ratio of the number of in-control observations that are incorrectly identified as out of control to the total number of in-control observations. A Type II error rate is defined as the ratio of the number of out-of-control observations that are not identified as out of control to the total number of the out-of-control observations.

3.3.2 Simulating autocorrelated multivariate data

In the present study, a vector autoregressive (VAR) model of order one is used to generate autocorrelated multivariate processes. A VAR model has three components. The three are a process mean vector (μ) , an autoregressive coefficient matrix (Φ) ,

and a covariance matrix of the residuals (Σ_r) . The m dimensional VAR model with p degrees of autocorrelation can be expressed as follows:

$$X_t = \mu + \Phi_1(X_{t-1} - \mu) + ... + \Phi_p(X_{t-p} - \mu) + \epsilon_t, \tag{3.6}$$

where X_t is the m-dimensional process vector, μ is the m by 1 process mean vector, Φ is the m by m autoregressive coefficient matrix, and ϵ_t is the m-dimensional white noise process vector with a zero mean and a covariance matrix Σ_r.

To generate the out-of-control data, three different degrees of shift were considered. Unlike univariate cases in which the shifts can be expressed in terms of standard deviation, multivariate cases involve more than one process variable. Thus, in multivariate cases, shifts usually can be expressed in terms of the following noncentrality parameter λ, which is a function of the magnitude of the shift δ and the estimated covariance matrix Σ_{X}:

$$\lambda = \sqrt{\delta^T \mathbf{\Sigma}_{\mathrm{X}}^{-1} \delta}. \tag{3.7}$$

In the present study, we considered three different magnitudes of the mean shift, which is shifted equally in all dimensions (λ=0.5(small), λ=1(medium), λ=2(large)). Further, we assumed a constant and unchanged covariance matrix.

3.3.3 Simulation results

Two control charts were compared (Hotelling's T^2 chart and K^2 chart) under the four simulation scenarios. Each scenario has a different degree of autocorrelation as shown in Table 3.1. The simulation results of all four scenarios are shown in Figures 3.5 − 3.8. For comparison, we used Type I and Type II error rates as the performance measurement. In general, we prefer a chart that yields a lower Type II error rate, given the similar values of Type I error rates. In this simulation, the average values of Type I and Type II errors were calculated from $1,000$ replications. In Figures 3.5 − 3.8, each figure consists of a three noncentrality parameter mean shift size (0.5λ, 1λ, and 2λ) as shown, respectively, in subfigures (a), (b), and (c).

The performance of Hotelling's T^2 charts and the K^2 charts are shown, respectively, by lines with triangles and lines with circles. The average standard errors from $1,000$ simulation runs of Type I and Type II error rates are approximately 0.0001.

All simulation scenarios provided similar results in that the K^2 charts yielded smaller Type II error rates than the T^2 charts, given similar Type I error rates. To put it simply, on average, K^2 charts are superior to T^2 charts in detecting out-of-control observations. The difference is clearly noticed in situations of small mean shifts. For situations with large shifts in mean, all charts performed comparably well.

To facilitate discussion, we grouped the simulation scenarios into two categories: scenarios with mixed degrees of positive autocorrelation and scenarios with equally positive degrees of autocorrelation. One of these with a mixed degree of positive autocorrelation is Scenario 1 (shown in Figure 3.5). Those scenarios with equal degrees of positive autocorrelation are further divided into three levels (low, medium, and high). The scenario with an equally low degree of positive autocorrelation is Scenario 2 (shown in Figure 3.6). Scenario 3 (shown in Figure 3.7) is the equally medium positive autocorrelation degree. The scenario with an equally high degree of positive autocorrelation is scenario 4 (shown in Figure 3.8).

As the degree of autocorrelation increased, the performance of all charts declined. The negative effect that autocorrelation processes have on control charts is well-known and likely accounts for the deterioration documented here. The performance comparisons of the T^2 charts and K^2 charts at different degrees of autocorrelation degree are shown in Figures 3.9 (a) and (b). Figure 3.9 (a) shows the performance of the T^2 charts across different degrees of positive autocorrelation. The lines with squares represent the performance of the T^2 charts with zero autocorrelation processes. The process with zero autocorrelation is a process that has unautocorrelated observations. We can clearly see that the T^2 charts perform worse as autocorrelation increases.

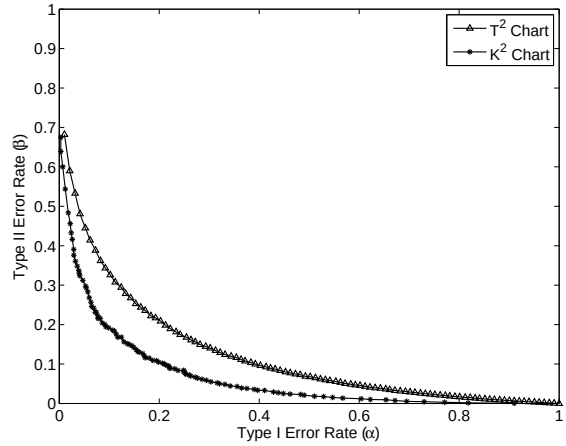
Similarly, Figure 3.9 (b) shows the performance of the K^2 charts across different degrees of positive autocorrelation. Note that the lines with squares represent the performance of K^2 charts with zero autocorrelation processes. Although the performance of the K^2 charts declines as autocorrelation increases, it is interesting to observe that the performance of the K^2 charts deteriorates less than the performance of the T^2 charts, implying the the K^2 charts are relatively robust to the degrees of autocorrelation in a process.

3.4 Concluding Remarks

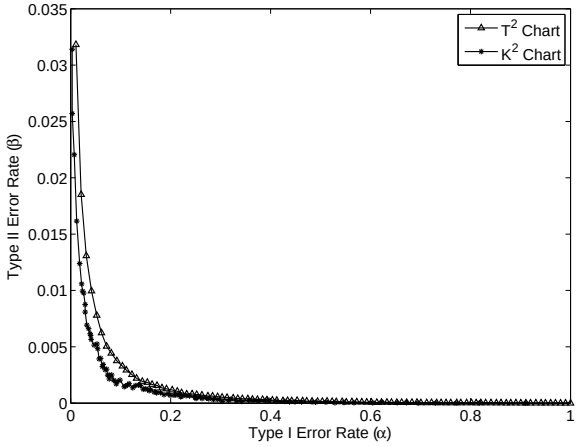
Autocorrelation observations are common in many industrial processes. Because failure to use multivariate control charts carefully with autocorrelated data may causes inaccurate monitoring result, it is important to develop control charts that can effectively handle autocorrelated observations. This study presents a K^2 chart that integrates the OCC algorithm and control chart techniques as a method to monitor autocorrelated multivariate processes. The K^2 chart is derived from a kNNDD algorithm, which is a modified version of the k nearest-neighbor algorithm that has proven its capability to effectively analyze and manage large amounts of data with only a minimal set of modeling assumptions. Moreover, unlike model-based control charts that use residuals, K^2 charts use original observations to monitor autocorrelated multivariate processes.

To demonstrate the effectiveness of the K^2 control charts, we conducted simulation studies under various autocorrelated scenarios, thus demonstrating that the K^2 charts outperformed the T^2 control charts. In particular, K^2 charts performed notably better than T^2 control charts in situations involving small mean shifts. Moreover, the performance of K^2 charts is not significantly affected by the degrees of autocorrelation.

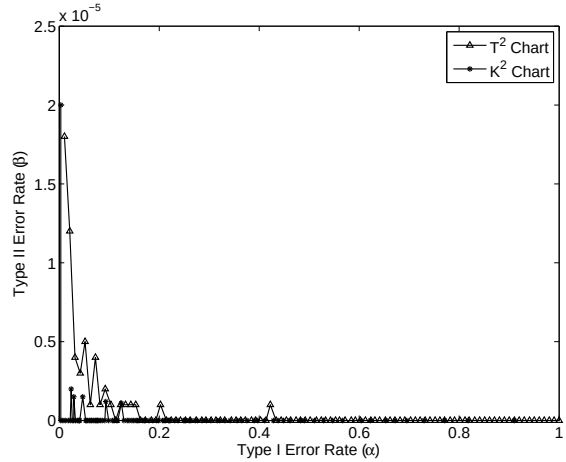
Our study extends the application scope of both the control chart method and the OCC algorithm. We hope that the procedure presented here stimulates further



(a)

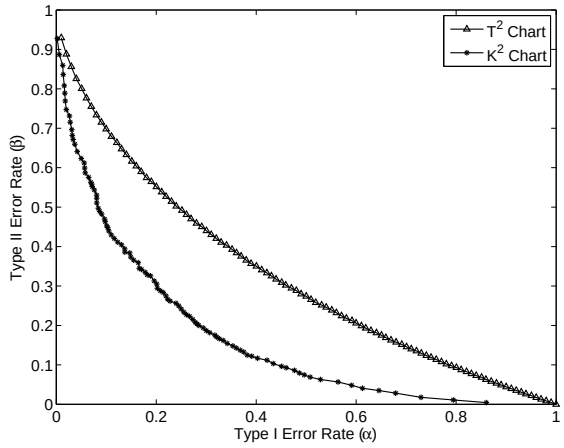


(b)

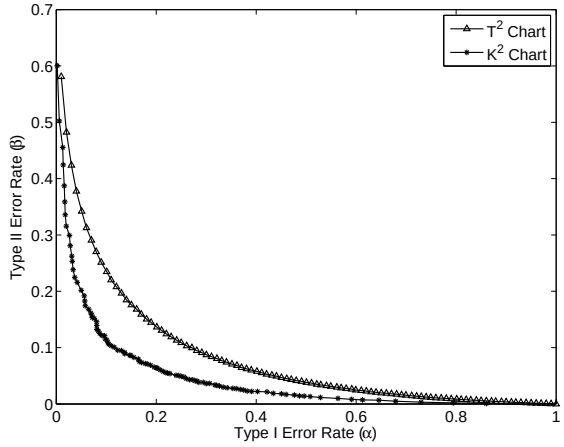


(c)

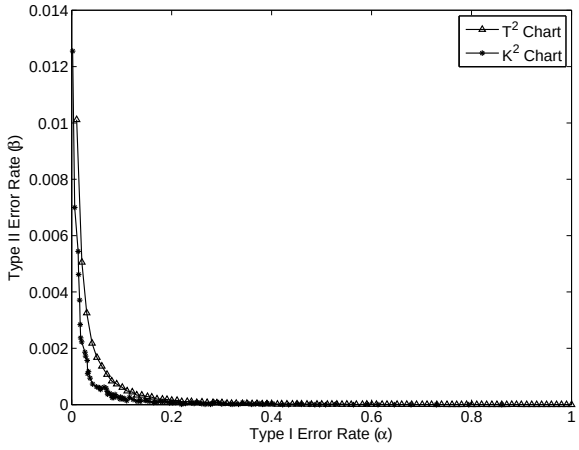
Figure 3.5. Type I and Type II error rates for two different control charts for three different mean shift sizes ((a) 0.5 λ, (b) 1 λ, and (c) 2 λ) with mixed degrees of positive autocorrelation and 0.5 degree of crosscorrelation (Scenario 1).



(a)

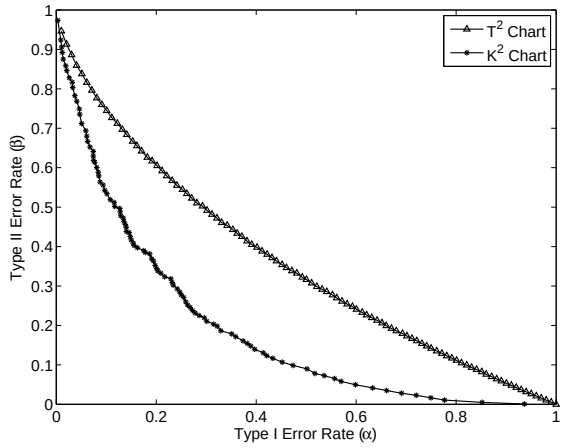


(b)

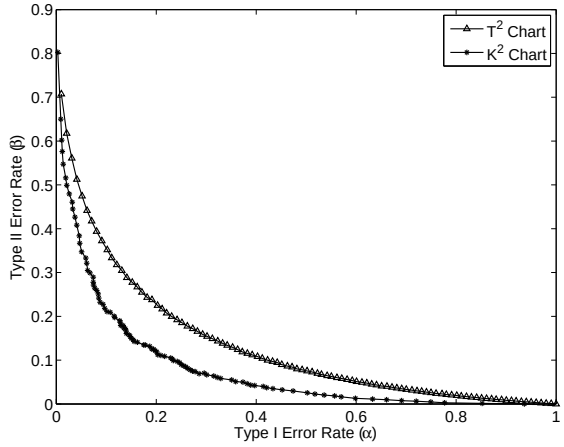


(c)

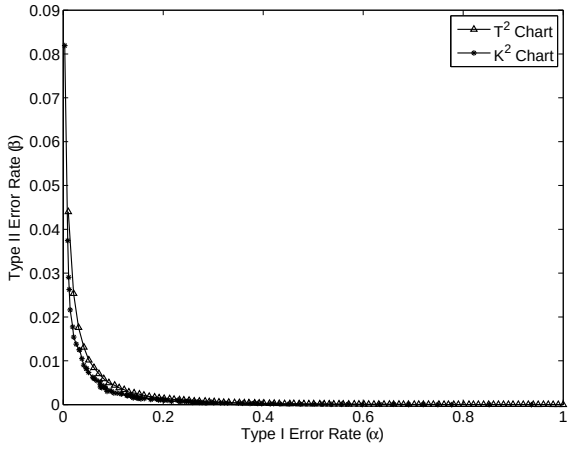
Figure 3.6. Type I and Type II error rates for two different control charts for three different mean shift sizes ((a) 0.5 λ, (b) 1 λ, and (c) 2 λ) with low degrees of positive autocorrelation and 0.5 degree of crosscorrelation (Scenario 2).



(a)

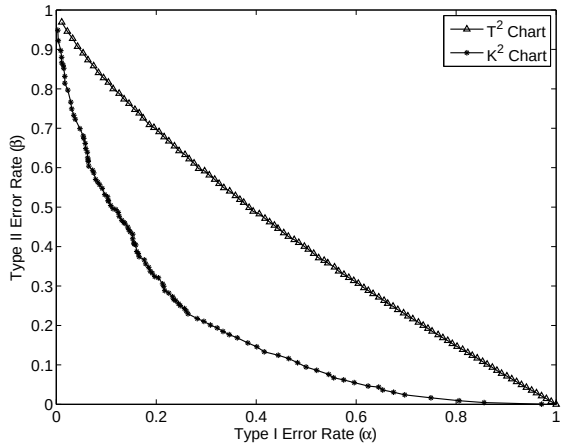


(b)

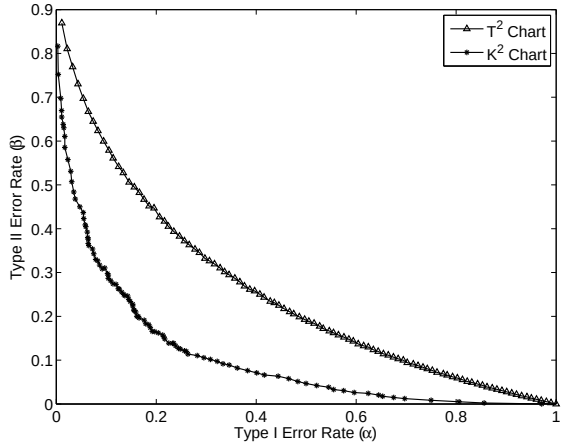


(c)

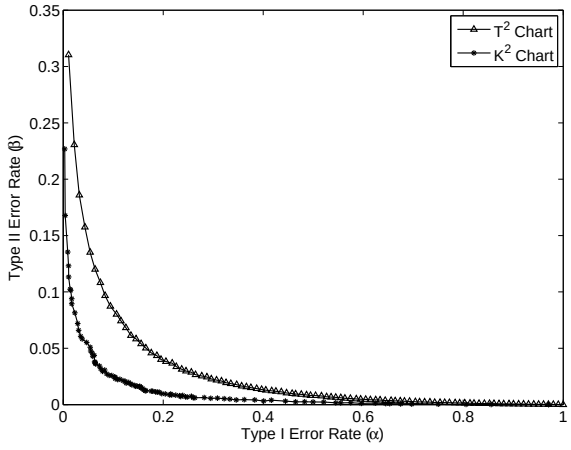
Figure 3.7. Type I and Type II error rates for two different control charts for three different mean shift sizes ((a) 0.5 λ, (b) 1 λ, and (c) 2 λ) with medium degrees of positive autocorrelation and 0.5 degree of crosscorrelation (Scenario 3).



(a)



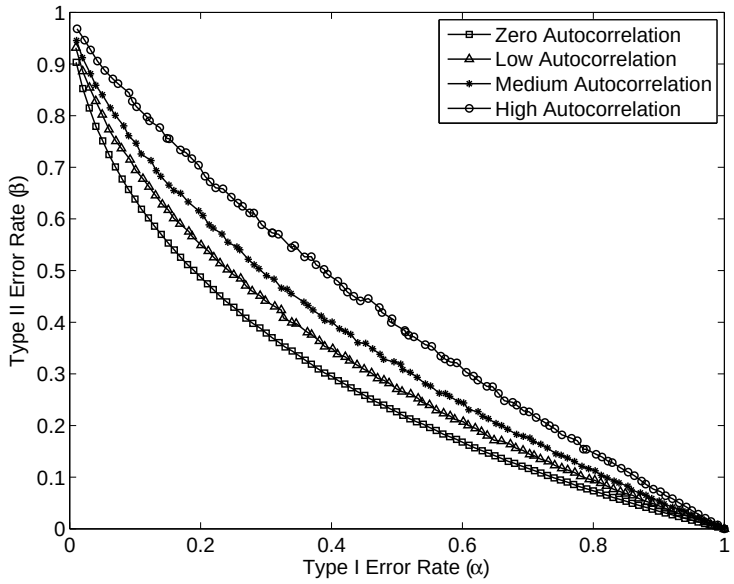
(b)



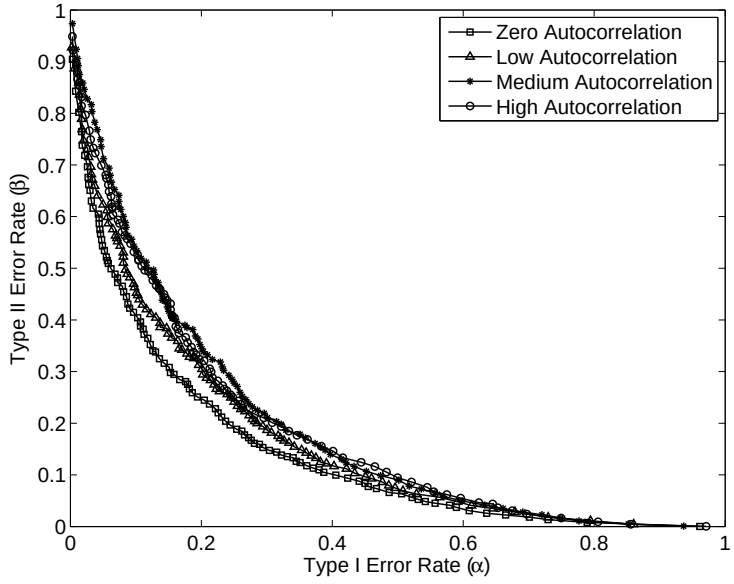
(c)

Figure 3.8. Type I and Type II error rates for two different control charts for three different mean shift sizes ((a) 0.5 λ, (b) 1 λ, and (c) 2 λ) with high degrees of positive autocorrelation and 0.5 degree of crosscorrelation (Scenario 4).

investigation into development of better procedures for OCC modeling in monitoring autocorrelated multivariate processes.



(a)



(b)

Figure 3.9. (a) Performance comparison of T^2 chart for different autocorrelation degrees (b) Performance comparison of K^2 chart for different autocorrelation degrees.

CHAPTER 4

SUMMARY AND FUTURE DIRECTIONS

In this dissertation, some data mining algorithms have been integrated with statistical process control for multivariate and autocorrelated process monitoring. In Chapter 2, we proposed data mining model-based control charts that utilize the residuals from the process. First, the data mining algorithms were used to develop prediction models. Second, we obtain the residuals from the difference between the actual values and the predicted values. Then, the residuals, assumed to be uncorrelated, would be monitored by the traditional multivariate control charts for process monitoring. Based on the simulation data with different degree of autocorrelation, different degree of crosscorrelation, and different numbers of process dimensions, the data mining model-based control charts have performed better than the traditional methods. In Chapter 3, we examined the possibility of using one-class classification-based control charts for multivariate and autocorrelated process monitoring. We proposed to monitor multivariate and autocorrelated process with one-class classification-based control charts without using residuals. Using simulated data, comparisons between traditional multivariate control charts and one-class classification-based control charts have been made. The results revealed that one-class classification-based control charts are superior to traditional multivariate control charts in all scenarios.

To monitor such a complex process as multivariate and autocorrelated process is a very challenging task. For future research in this direction, we can integrate two procedures proposed in this dissertation by develop data mining model-based control charts and monitor the residuals with one-class classification-based control charts. Moreover, we can integrate exponentially weighted moving average procedure

with one-class classification-based control charts for multivariate and autocorrelated process monitoring.

APPENDIX A

PARAMETERS UTILIZED IN GENERATING VECTOR AUTOREGRESSIVE PROCESSES IN CHAPTER 2

This section includes parameters used in generating vector regressive process for each scheme. μ is the in-control process mean vector, Φ is autoregressive coefficient matrix, and Σ_r is the covariance matrix of the residuals. Table A.1 displays parameters for scenarios 1, 2, and 3. All three scenarios are bivariate processes. Clearly, all have the same process mean and covariance matrix of the residual. The only difference is the autoregressive coefficient matrix. Tables A.2 to A.4 include parameters for five dimensions scenarios. Tables A.5 to A.7 include parameters for ten dimensions scenarios.

Table A.1. Parameters used in generating vector autoregressive processes for three different degrees of positive autocorrelation under 2 dimensions.

Parameters	Low Positive (Scenario1)		Medium Positive (Scenario2)		High Positive (Scenario3)	
μ	0	0	0	0	0	0
Φ	0.25	0.0177	0.50	0.0177	0.75	0.0177
	0.6493	0.25	0.6493	0.50	0.6493	0.75
Σ_r	99.91	63.99	99.91	63.99	99.91	63.99
	63.99	69.52	63.99	69.52	63.99	69.52

Table A.2. Parameters used in generating vector autoregressive processes for 5 dimensions and low positive autocorrelation (Scenario 4)

	Parameters				
μ	0	0	0	0	0
Φ	0.25	-0.02	-0.08	0.08	-0.02
	-0.02	0.25	-0.03	0.03	-0.05
	-0.08	-0.03	0.25	0.08	-0.05
	0.08	0.03	0.08	0.25	0.04
	-0.02	-0.05	-0.05	0.04	0.25
Σ_r	100.00	45.54	-18.09	74.66	-26.43
	45.54	100.00	-44.31	42.69	-17.64
	-18.09	-44.31	100.00	-5.95	30.82
	74.66	42.69	-5.95	100.00	-23.26
	-26.43	-17.64	30.82	-23.26	100.00

Table A.3. Parameters used in generating vector autoregressive processes for 5 dimensions and medium positive autocorrelation (Scenario 5)

	Parameters				
μ	0	0	0	0	0
Φ	0.50	-0.02	-0.08	0.08	-0.02
	-0.02	0.50	-0.03	0.03	-0.05
	-0.08	-0.03	0.50	0.08	-0.05
	0.08	0.03	0.08	0.50	0.04
	-0.02	-0.05	-0.05	0.04	0.50
Σ_r	100.00	45.54	-18.09	74.66	-26.43
	45.54	100.00	-44.31	42.69	-17.64
	-18.09	-44.31	100.00	-5.95	30.82
	74.66	42.69	-5.95	100.00	-23.26
	-26.43	-17.64	30.82	-23.26	100.00

Table A.4. Parameters used in generating vector autoregressive processes for 5 dimensions and high positive autocorrelation (Scenario 6)

	Parameters				
μ	0	0	0	0	0
Φ	0.75	-0.02	-0.08	0.08	-0.02
	-0.02	0.75	-0.03	0.03	-0.05
	-0.08	-0.03	0.75	0.08	-0.05
	0.08	0.03	0.08	0.75	0.04
	-0.02	-0.05	-0.05	0.04	0.75
Σ_r	100.00	45.54	-18.09	74.66	-26.43
	45.54	100.00	-44.31	42.69	-17.64
	-18.09	-44.31	100.00	-5.95	30.82
	74.66	42.69	-5.95	100.00	-23.26
	-26.43	-17.64	30.82	-23.26	100.00

Table A.5. Parameters used in generating vector autoregressive processes for 10 dimensions and low positive autocorrelation (Scenario 7)

	Parameters									
μ	0	0	0	0	0	0	0	0	0	0
Φ	0.25	-0.01	0.05	-0.02	0.00	0.00	-0.05	0.04	0.11	-0.04
	-0.01	0.25	-0.02	-0.06	0.07	-0.03	0.03	-0.03	0.06	0.00
	0.05	-0.02	0.25	-0.03	0.07	0.02	-0.03	0.00	-0.04	-0.04
	-0.02	-0.06	-0.03	0.25	0.02	0.04	-0.03	0.07	0.01	-0.03
	0.00	0.07	0.07	0.02	0.25	0.03	0.03	0.06	-0.05	-0.01
	0.00	-0.03	0.02	0.04	0.03	0.25	0.08	-0.04	0.07	0.01
	-0.05	0.03	-0.03	-0.03	0.03	0.08	0.25	0.03	0.00	-0.04
	0.04	-0.03	0.00	0.07	0.06	-0.04	0.03	0.25	0.02	0.05
	0.11	0.06	-0.04	0.01	-0.05	0.07	0.00	0.02	0.25	0.03
	-0.04	0.00	-0.04	-0.03	-0.01	0.01	-0.04	0.05	0.03	0.25
Σ_r	100.00	18.37	57.18	22.68	18.06	17.13	-6.44	16.64	-6.21	-13.77
	18.37	100.00	-7.52	20.70	14.55	-31.37	-1.50	55.29	5.55	11.23
	57.18	-7.52	100.00	-7.49	8.37	28.78	27.84	-1.52	8.59	-23.70
	22.68	20.70	-7.49	100.00	35.34	4.06	-36.92	10.44	-30.70	39.02
	18.06	14.55	8.37	35.34	100.00	-35.38	-12.99	-33.43	30.90	8.35
	17.13	-31.37	28.78	4.06	-35.38	100.00	-20.84	-3.35	-46.60	-12.73
	-6.44	-1.50	27.84	-36.92	-12.99	-20.84	100.00	15.87	12.98	4.12
	16.64	55.29	-1.52	10.44	-33.43	-3.35	15.87	100.00	-19.25	4.57
	-6.21	5.55	8.59	-30.70	30.90	-46.60	12.98	-19.25	100.00	-29.83
	-13.77	11.23	-23.70	39.02	8.35	-12.73	4.12	4.57	-29.83	100.00

Table A.6. Parameters used in generating vector autoregressive processes for 10 dimensions and medium positive autocorrelation (Scenario 8)

	Parameters									
μ	0	0	0	0	0	0	0	0	0	0
Φ	0.50	-0.02	0.10	-0.05	0.00	-0.01	-0.11	0.08	0.22	-0.09
	-0.02	0.50	-0.04	-0.13	0.15	-0.07	0.06	-0.06	0.11	0.00
	0.10	-0.04	0.50	-0.06	0.15	0.05	-0.07	0.01	-0.07	-0.07
	-0.05	-0.13	-0.06	0.50	0.05	0.09	-0.05	0.15	0.02	-0.05
	0.00	0.15	0.15	0.05	0.50	0.07	0.06	0.13	-0.10	-0.03
	-0.01	-0.07	0.05	0.09	0.07	0.50	0.15	-0.09	0.14	0.01
	-0.11	0.06	-0.07	-0.05	0.06	0.15	0.50	0.06	0.01	-0.08
	0.08	-0.06	0.01	0.15	0.13	-0.09	0.06	0.50	0.04	0.10
	0.22	0.11	-0.07	0.02	-0.10	0.14	0.01	0.04	0.50	0.06
	-0.09	0.00	-0.07	-0.05	-0.03	0.01	-0.08	0.10	0.06	0.50
Σ_r	100.00	18.37	57.18	22.68	18.06	17.13	-6.44	16.64	-6.21	-13.77
	18.37	100.00	-7.52	20.70	14.55	-31.37	-1.50	55.29	5.55	11.23
	57.18	-7.52	100.00	-7.49	8.37	28.78	27.84	-1.52	8.59	-23.70
	22.68	20.70	-7.49	100.00	35.34	4.06	-36.92	10.44	-30.70	39.02
	18.06	14.55	8.37	35.34	100.00	-35.38	-12.99	-33.43	30.90	8.35
	17.13	-31.37	28.78	4.06	-35.38	100.00	-20.84	-3.35	-46.60	-12.73
	-6.44	-1.50	27.84	-36.92	-12.99	-20.84	100.00	15.87	12.98	4.12
	16.64	55.29	-1.52	10.44	-33.43	-3.35	15.87	100.00	-19.25	4.57
	-6.21	5.55	8.59	-30.70	30.90	-46.60	12.98	-19.25	100.00	-29.83
	-13.77	11.23	-23.70	39.02	8.35	-12.73	4.12	4.57	-29.83	100.00

Table A.7. Parameters used in generating vector autoregressive processes for 10 dimensions and high positive autocorrelation (Scenario 9)

	Parameters									
μ	0	0	0	0	0	0	0	0	0	0
Φ	0.75	-0.01	0.05	-0.02	0.00	0.00	-0.05	0.04	0.11	-0.04
	-0.01	0.75	-0.02	-0.06	0.07	-0.03	0.03	-0.03	0.06	0.00
	0.05	-0.02	0.75	-0.03	0.07	0.02	-0.03	0.00	-0.04	-0.04
	-0.02	-0.06	-0.03	0.75	0.02	0.04	-0.03	0.07	0.01	-0.03
	0.00	0.07	0.07	0.02	0.75	0.03	0.03	0.06	-0.05	-0.01
	0.00	-0.03	0.02	0.04	0.03	0.75	0.08	-0.04	0.07	0.01
	-0.05	0.03	-0.03	-0.03	0.03	0.08	0.75	0.03	0.00	-0.04
	0.04	-0.03	0.00	0.07	0.06	-0.04	0.03	0.75	0.02	0.05
	0.11	0.06	-0.04	0.01	-0.05	0.07	0.00	0.02	0.75	0.03
	-0.04	0.00	-0.04	-0.03	-0.01	0.01	-0.04	0.05	0.03	0.75
Σ_r	100.00	18.37	57.18	22.68	18.06	17.13	-6.44	16.64	-6.21	-13.77
	18.37	100.00	-7.52	20.70	14.55	-31.37	-1.50	55.29	5.55	11.23
	57.18	-7.52	100.00	-7.49	8.37	28.78	27.84	-1.52	8.59	-23.70
	22.68	20.70	-7.49	100.00	35.34	4.06	-36.92	10.44	-30.70	39.02
	18.06	14.55	8.37	35.34	100.00	-35.38	-12.99	-33.43	30.90	8.35
	17.13	-31.37	28.78	4.06	-35.38	100.00	-20.84	-3.35	-46.60	-12.73
	-6.44	-1.50	27.84	-36.92	-12.99	-20.84	100.00	15.87	12.98	4.12
	16.64	55.29	-1.52	10.44	-33.43	-3.35	15.87	100.00	-19.25	4.57
	-6.21	5.55	8.59	-30.70	30.90	-46.60	12.98	-19.25	100.00	-29.83
	-13.77	11.23	-23.70	39.02	8.35	-12.73	4.12	4.57	-29.83	100.00

APPENDIX B

PARAMETERS UTILIZED IN GENERATING VECTOR AUTOREGRESSIVE PROCESSES IN CHAPTER 3

This section includes parameters used in generating vector regressive process for each scheme. μ is the in-control process mean vector, Φ is autoregressive coefficient matrix, and Σ_r is the covariance matrix of the residuals. Table B.1 displays parameters for scenarios 1, 2, 3, and 4. All four scenarios are bivariate processes. Clearly, all scenarios have the same process mean and covariance matrix of the residual. The only difference is the autoregressive coefficient matrix.

Table B.1. Parameters used in generating vector autoregressive processes.

Parameters	Mixed Positive (Scenario1)		Low Positive (Scenario2)		Medium Positive (Scenario3)		High Positive (Scenario4)	
μ	0	0	0	0	0	0	0	0
Φ	0.25	0.0177	0.25	0.0177	0.50	0.0177	0.75	0.0177
	0.0177	0.75	0.0177	0.25	0.0177	0.50	0.0177	0.75
Σ_r	100.00	55.53	100.00	55.53	100.00	55.53	100.00	55.53
	55.53	100.00	55.53	100.00	55.53	100.00	55.53	100.00

APPENDIX C

NONCENTRALITY PARAMETER

A noncentrality parameter is used in generating out-of-control data. Let μ_0 and Σ_X be the mean vector and the covariance matrix of the in-control multivariate process. Let $\mu_1 = \mu_0 + \delta$ be the mean vector of the out-of-control process. So, δ is the difference between μ_0 and μ_1. The following numerical example will show how to calculate a noncentrality parameter.

First, we find the covariance matrix of the in-control data, Σ_X. Then we find the inverse matrix of Σ_X.

$$Let \quad \mu_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}., \mu_1 = \begin{bmatrix} 1 & 1 \end{bmatrix}., and \quad \Sigma_X = \begin{bmatrix} 1.00 & 0.50 \\ 0.50 & 0.33 \end{bmatrix}.$$

, therefore the inverse matrix will be

$$\Sigma^{-1}{}_X = \begin{bmatrix} 4 & -6 \\ -6 & 12 \end{bmatrix}.$$

and δ will equal to [1 1]. Plug these values into the noncentrality parameter equation (equation 3.7). The noncentrality parameter will be 2. However, if in this case, δ equal to [2 2], with the same Σ_X, the noncentrality parameter will equal to 4.

REFERENCES

[1] W. A. Shewhart, *Economic Control of Quality of Manufactured Product.* Princeton, NJ: Van Nostrand Press, 1931.

[2] S. D. S. Shumway, R. H., *Time Series Analysis and Its Applications: With R Examples, Springer.* New York, NY.: Springer, 2006.

[3] H. W. G. Berthouex, P. M. and L. Pallesen, "Monitoring sewage treatment plants: Some quality control aspects," *Journal of Quality Technology*, vol. 10, pp. 139–149, 1978.

[4] D. Montgomery and C. Mastrangelo, "Some statistical process control methods for autocorrelated data," *Journal of Quality Technology*, vol. 23, no. 3, pp. 179–204, 1991.

[5] R. W. H. Harris, T.J., "Statistical process control procedures for correlated observations," *Canadian journal of chemical engineering*, vol. 69, no. 1, pp. 48–57, 1991.

[6] A. L.C., "Effects of autocorrelation on control chart performance," *Communication in Statistics-Theory and Methods*, vol. 21, pp. 1025–1049, 1992.

[7] W. H. Woodall and F. W. Faltin, "Autocorrelated data and spc," *ASQC Statistics Division Newsletter*, vol. 13, pp. 18–21, 1993.

[8] H. Hotelling, *Multivariate Quality Control*, ser. Techniques of Statistical Analysis, C. Eisenhart, M. W. Hastay, and W. A. Wallis, Eds. New York, NY: McGraw-Hill, 1947.

[9] W. Woodall and M. Ncube, "Multivariate cusum quality control procedures," *Technometrics*, vol. 27, pp. 285–292, 1985.

[10] J. Healy, "A note on multivariate cusum procedures," *Technometrics*, vol. 29, no. 4, pp. 409–412, 1987.

[11] R. Crosier, "Multivariate generalization of cumulative sum quality-control schemes," *Technometrics*, vol. 30, no. 3, pp. 291–303, 1988.

[12] J. J. Pignatiello and G. Runger, "Comparisons of multivariate cusum charts," *Journal of Quality Technology*, vol. 22, pp. 173–186, 1990.

[13] W. W. C. C. Lowry, C.A. and S. Rigdon, "A multivariate exponentially weighted moving average control chart," *Technometrics*, vol. 34, no. 1, pp. 46–53, 1992.

[14] P.-S. G. Fayyad, Usama and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[15] Z.-A. Gaber, M.M. and S. Krishnaswamy, "Mining data streams: a review," *SIGMOD Record*, vol. 34, no. 2, pp. 18–26, 2005.

[16] I.-T. Agrawal, R. and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference*, 1993.

[17] J.-D. Loredo, E.N. and C. Borror, "Model-based control chart for autoregressive and correlated data," *Quality and reliability engineering international*, vol. 18, pp. 489–496, 2002.

[18] A. A. Kalgonda and S. R. Kulkarni, "Multivariate quality control chart for autocorrelated processes," *Journal of Applied Statistics*, vol. 31, pp. 317–327, 2004.

[19] R. Noorossana and S. Vaghefi, "Effect of autocorrelation on performance of the mcusum control chart," *Quality and Reliability Engineering International*, vol. 22, no. 2, pp. 191–197, 2006.

[20] P.-X. Jarrett, J.E., "The quality control chart for monitoring multivariate autocorrelated processes," *Computational Statistics & Data Analysis*, vol. 51, pp. 3862–3870, 2007.

[21] X. Pan and J. Jarrett, "Using vector autoregressive residuals to monitor multivariate processes in the presence of serial correlation," *International Journal of Production Economics*, vol. 106, pp. 204–216, 2007.

[22] N.-S. T. A. A. B. Arkat, J., "Artificial neural networks in applying mcusum residuals charts for ar(1) processes," *Applied Mathematics and Computation*, vol. 189, pp. 1889–1901, 2007.

[23] L. Issam, B.K. amd Mohamed, "Support vector regression based residual mcusum control chart for autocorrelated process," *Applied Mathematics and Computation*, vol. 201, pp. 565–574, 2008.

[24] D. C. Montgomery, *Introduction to Statistical Quality Control*, 5th ed. New York, NY: Wiley, 2005.

[25] D. M. Hawkins, "Multivariate quality control based on regression-adjusted variables," *Technometrics*, vol. 31, no. 1, pp. 61–75, 1991.

[26] A. L.C. and H. Roberts, "Time-series modeling for statistical process control," *Journal of Business & Economic Statistics*, vol. 6, no. 1, pp. 87–95, 1988.

[27] G. Runger and T. Willemain, "Model-based and model-free control of autocorrelated processes," *Journal of Quality Technology*, vol. 27, no. 4, pp. 283–292, 1995.

[28] N. Zhang, "A statistical control chart for stationary process data," *Technometrics*, vol. 40, no. 1, pp. 24–38, 1998.

[29] T. K. Jiang, W.J. and W. Woodall, "A new spc monitoring method: The arma chart," *Technometrics*, vol. 42, no. 4, pp. 399–410, 2000.

[30] W. Woodall and D. Montgomery, "Research issues and ideas in statistical process control," *Journal of Quality Technology*, vol. 31, pp. 376–385, 1999.

[31] . J. G. M. Box, G. E. P., *Times series analysis: Forecasting and control (Revised)*. Oakland, CA: Holden-Day, Inc., 1976.

[32] B. Ripley, *Pattern Recognition and Artificial Neural Networks*. NY: Cambridge University Press, 1996.

[33] C. N. and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. UK: Cambridge University Press, 2000.

[34] J. Friedman, "Multivariate adaptive regression splines (with discussion)," *ANNals of Statistics*, vol. 19, pp. 1–141, 1991.

[35] B. Biller and B. Nelson, "Modeling and generating multivariate time-series input processes using a vector autoregressive technique," *ACM Transactions on Modeling and Computer Simulation*, vol. 13, no. 3, pp. 211–237, 2003.

[36] B. Pfaff, "Var, svar and svec models: Implementation within r package vars," *Journal of Statistical Software*, vol. 27, no. 4, 2008.

[37] R. L. Mason and J. C. Young, *Multivariate Statistical Process Control with Industrial Applications*. Philadelphia, PA: American Statistical Association and Society for Industrial and Applied Mathematics, 2002.

[38] J. Fox, *An R and S-PLUS companion to Applied Regression*. Thousand Oaks, CA: Sage, 2002.

[39] D. M. J. Tax, "One-class classification: Concept-learning in the absence of counter-examples," Ph.D. dissertation, Delf University of Technology, 2001.

[40] K. S. Sukchotrat, T. and F. Tsung., "One-class classification-based control charts for multivariate process monitoring," *IIE Transactions (In Press)*.

[41] K. H. P. N. R. T. Breunig, M. M. and J. Sander, "Lof: identifying density-based local outliers," in *Preceedings of the ACM SIGMOD 2000 international conference on management of data*, vol. 29, 2000, pp. 93–104.

BIOGRAPHICAL STATEMENT

Weerawat Jitpitaklert received his B.B.A. degree from Thammasat University, Thailand, in 2001. In 2002, he worked for Muang Luang Transport Company Limited, Thailand. In 2004, he completed a M.S. degree in Logistics from the University of Texas at Arlington (UTA). In 2005, Thailand government offered him a comprehensive scholarship for pursuing a Ph.D. degree. Then, he has continued pursuing a Ph.D. degree and working as a graduate research assistant at UTA. His research interest is applications of data mining, multivariate statistical process control, time-series analysis, operation management, and logistics/supply chain management. He is a member of the Center of Stochastic Modeling, Optimization & Statistics (COSMOS) at UTA, Institute for Industrial Engineers (IIE), Institute for Operation Research and Management Science (INFORMS). His web address is http://www.weerawat.net.