

**ENHANCING THE EFFICIENCY OF MULTIVARIATE
CONTROL CHARTS IN THE PRESENCE OF MISSING VALUES
USING IMPUTATION BY MACHINE LEARNING
TECHNIQUES AND DEVELOPING A NEW SELF-STARTING
CONTROL CHART FOR HIGH DIMENSIONAL DATA**

BY

SIKA JAWARA

A Thesis Presented to the
COLLEGE OF GRADUATE AND INTERDISCIPLINARY STUDIES
KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In
APPLIED STATISTICS

MAY 2024

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN 31261, SAUDI ARABIA

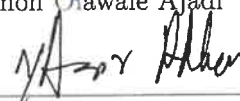
DEANSHIP OF GRADUATE STUDIES

This thesis, written by **SIKA JAWARA** under the direction of his thesis adviser and approved by his thesis committee, has been presented to and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN APPLIED STATISTICS**.

Thesis Committee



Dr. Jimoh Olawale Ajadi (Adviser)



Dr. Nasir Abbas (Member)



Dr. Mohammad Hafidz Omar (Member)


Prof. Monther R. Alfuraidan
Department Chairman
Prof. Suliman Al-Homidan
Dean of Graduate Studies

Date



©Siaka Jawara
2024

Dedication

To my dear family, whose steadfast support and unwavering faith in my capabilities have been my greatest pillars of strength. Your constant encouragement, boundless love, and tireless prayers, day, and night, have been instrumental in helping me reach this milestone. With deep gratitude and affection, I dedicate this thesis to you.

ACKNOWLEDGMENTS

In embarking on this academic endeavor, I've been fortunate to be under the exceptional guidance of Dr. Jimoh Ajadi, my supervisor. His unwavering support, profound insights, and mentorship have indelibly shaped the trajectory of this thesis. Acknowledgment is also due to the Department of Mathematics for their steadfast support and the scholarship that facilitated my studies. Their continuous assistance has been integral to my academic pursuits. I am equally indebted to the esteemed members of my thesis committee for their valuable contribution and guidance. Also, a special recognition goes to Prof. Riaz for his mentorship. Furthermore, I extend my appreciation to my professors and colleagues for fostering an enriching learning environment and for their readiness to offer guidance and support whenever needed. Lastly, my deepest gratitude is reserved for my family—my parents, siblings, whose unwavering support, encouragement, and belief in my capabilities have been the bedrock of my resilience and determination throughout this journey.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	vi
LIST OF TABLES	x
LIST OF FIGURES	xi
ABSTRACT (ENGLISH)	xii
ABSTRACT (ARABIC)	xiv
CHAPTER 1 INTRODUCTION	1
1.1 General Control Chart Basics	1
1.2 The Hotelling T^2 Control Chart	2
1.3 Motivation	3
1.4 Missing Data in Quality Control	5
1.5 Multivariate Control Chart	6
1.6 Self-Starting Control Chart	7
1.7 Thesis Objective and Significance	8
1.8 Thesis Layout	10
CHAPTER 2 THEORETICAL BACKGROUND AND LITERATURE REVIEW	11
2.1 Overview of Control Chart	11
2.2 Control Chart Literature with Missing Values	14
2.3 Overview of Self-Starting Control Chart and Related Literature	15

CHAPTER 3	ENHANCING THE EFFICIENCY OF MULTIVARIATE CONTROL CHARTS IN THE PRESENCE OF MISSING VALUES USING MACHINE LEARNING TECHNIQUES	18
3.1	Missing Value Imputation Methods	19
3.1.1	Traditional Imputation Method	19
3.1.2	Machine Learning-Based Imputation Methods	20
3.2	Evaluating the Effect of Missing Data Handling Methods on Hotelling T^2 Control Chart Performance in Phase I	27
3.3	In-control Performance	28
3.3.1	Results of the In-control Performance when Correlation Coefficient Between Quality Characteristics is Zero	29
3.3.2	Results of the In-control Performance when Correlation Coefficient Between Quality Characteristics is 0.7	30
3.4	Application	31
3.5	Summary	35
CHAPTER 4	A SUPERVISED MULTIVARIATE SELF-STARTING EXPONENTIALLY WEIGHTED MOVING AVERAGE CHART BASED ON INTERPOINTS DISSIMILARITIES	36
4.1	Introduction	36
4.2	Change-point model	37
4.3	Existing self-starting statistics based on sum of squared of interpoints dissimilarities	39
4.4	Multivariate Self-starting exponentially weighted moving average chart based on interpoints dissimilarities	42
4.4.1	A self-starting statistics based on sum of squared of interpoints dissimilarities	42
4.4.2	An exponentially weighted moving average structure based on interpoints dissimilarities	44
4.4.3	Implementation of the proposed scheme	45

4.5	Simulation Study	47
4.5.1	Methods for simulating data	48
4.5.2	Summary of simulation results	49
CHAPTER 5 CONCLUSION AND FUTURE WORK		63
REFERENCES		66
VITAE		74

LIST OF TABLES

3.1	The values of h that result in an ICARL when no missing value is present.	28
3.2	Values of ICARL when $\rho = 0$.	30
3.3	Values of ICARL when $\rho = 0.7$.	31
4.1	<i>ARL</i> profile values when multivariate dataset simulated from MNORM and MGAMMA distribution for $P = 2$.	52
4.2	<i>ARL</i> profile values when multivariate dataset simulated from MPOIS and MNB distribution for $P = 2$.	53
4.3	<i>ARL</i> profile values when multivariate dataset simulated from MGAMMA and MNORM distribution for $P = 10$.	54
4.4	<i>ARL</i> profile values when multivariate dataset simulated from MPOIS-NORM and MNORMMGAMMA distribution for $P = 2$.	55
4.5	<i>ARL</i> profile values when multivariate dataset simulated from MNORM and MGAMMA distribution for $P = 10$.	56
4.6	<i>ARL</i> profile values when multivariate dataset simulated from MPOIS and MNB distribution for $P = 10$.	57
4.7	<i>ARL</i> values of the chart under ID covariance structure and Euclidean distance when multivariate of 10-dimensional ($P=10$) dataset simulated from MGAMMA, MNORM, MPOIS, MNB, MNORMMGAMMA, and MPOISNORM distributions.	61

LIST OF FIGURES

3.1	Scatterplot of wholesale distributor	32
3.2	Monitoring plot of a wholesale distributor without missing values	34
3.3	short	34
4.1	Results show (a) average run length (ARL), (b) median run length (MDRL), (c) 75th percentile of run length (75th PRL), and (d) standard deviation of run length of the charts from MNORM distribution with auto-regress (AR), equi-covariance (EQ), independent (ID) and non-sparse (NSP) covariance structures $p = 2$ or 10	58
4.2	Results show (a) average run length (ARL), (b) median run length (MDRL), (c) 75th percentile of run length (75th PRL), and (d) standard deviation of run length of the charts from MGAMMA, MNORM, MPOIS, MNBINORM, MNORMGAMMA, or MPOISNORM distribution when $p = 2, 3, 4, 5, 10$ and 20	59
4.3	Results show average run length (ARL) based on Euclidean, Manhattan and Bray-Curtis distance measures under independent (ID) covariance structure for $p = 10$	60

THESIS ABSTRACT

NAME: Siaka Jawara

TITLE OF STUDY: Enhancing the Efficiency of Multivariate Control Charts in the Presence of Missing Values Using Imputation by Machine Learning Techniques and Developing a New Self-Starting Control Chart for High Dimensional Data

MAJOR FIELD: Applied Statistics

DATE OF DEGREE: May, 2024

This thesis sets out to accomplish two primary objectives. First, we employ imputation techniques based on machine learning approaches to investigate the impact of missing data in Phase I on the performance of Hotelling T^2 control chart. Six distinct imputation methods were evaluated: k -nearest neighbor (kNN), multiple imputation by chained equation ($MICE$), weighted nearest neighbor ($wkNN$), weighted nearest neighbor selection ($wkNNSel$), missforest (which are machine learning-based methods), along with mean imputation (a traditional method). The study utilize estimates of the mean vector and variance-covariance matrix after handling missing value to construct the Hotelling T^2 chart statistics. Through a comprehensive simulation

study involving various parameters such as the number of quality characteristics p , correlation between them ρ , percentage of missing values m , and a sample size of 200, we assess the effectiveness of these methods in estimating ICARL. Our simulation results show that the MICE-based imputation method outperforms other methods when the correlation coefficient is 0 or 0.7 and the number of quality characteristics is 15 or fewer. However, for larger numbers of quality characteristics, $wkNN$ and kNN outperform other methods when correlation coefficient is 0 and 0.7, respectively. Our second objective focuses on the introduction of a novel self-starting control chart. This innovative strategy will represents a significant advancement in the realm of statistical process control, offering a potent solution for monitoring and managing processes characterized by a multitude of correlated quality attributes. Through a simulation studies, the efficacy and practical utility of this innovative approach are demonstrated, highlighting its potential to revolutionize quality management practices across diverse industries. In essence, this thesis not only contributes to the theoretical advancement of statistical process control but also offers tangible solutions to address the practical challenges posed by contemporary datasets in real-world quality management scenarios.

ملخص الرسالة

الاسم: سياكا جاوارا

عنوان الدراسة: تحسين كفاءة مخططات التحكم متعددة المتغيرات في وجود قيم مفقودة باستخدام الإسناد بواسطة تقنيات التعلم

الآلي وتطوير مخطط تحكم جديد ذاتي البدء للبيانات عالية الأبعاد

التخصص: الإحصاء التطبيقي

تاريخ الدرجة العلمية: مايو 2024

تهدف هذه الرسالة إلى تحقيق هدفين أساسيين. أولاً، نستخدم تقنيات التعويضاً طرحة هذه تسعى لتحقيق هدفين رئيسيين. أولاً، نستخدم تقنيات التعويض بناءً على مناهج التعلم الآلي للتحقيق في تأثير البيانات المفقودة في المرحلة الأولى على أداء مخطط التحكم T^2 Hotelling. تم تقييم ستة أساليب تعويض متميزة: الجار الأقرب، (kNN) التعويض المتعدد بواسطة معادلات السلسلة، (MICE) الجار الأقرب الموزون، (wkNN) اختيار الجار الأقرب الموزون، (wkNNSel) والغابة العشوائية (وهي طرق تعتمد على التعلم الآلي)، بالإضافة إلى التعويض بالمتوسط (وهو طريقة تقليدية). الدراسة تستخدم تقديرات متجه المتوسط ومصفوفة التباين والتغاير بعد معالجة القيم المفقودة لإنشاء إحصائيات مخطط T^2 Hotelling. من خلال دراسة محاكاة شاملة تشمل معايير مختلفة مثل عدد خصائص الجودة p ، الارتباط بينها ρ ، نسبة القيم المفقودة m ، وحجم العينة 200، نقيم فعالية هذه الأساليب في تقدير متوسط طول التشغيل في الوضع الطبيعي. (ICARL) نتائج المحاكاة لدينا تظهر أن طريقة التعويض المستندة إلى MICE تتفوق على الطرق الأخرى عندما يكون معامل الارتباط 0 أو 7.0 وعدد خصائص الجودة 15 أو أقل. ومع ذلك، بالنسبة لعدد أكبر من خصائص الجودة، فإن أساليب wkNN و kNN تتفوق على الأساليب الأخرى عندما يكون معامل الارتباط 0 و 7.0، على التوالي. يركز هدفنا الثاني على تقديم مخطط تحكم ذاتي البدء جديد. هذه الاستراتيجية المبتكرة تمثل تقدماً كبيراً في مجال التحكم الإحصائي في العمليات، حيث تقدم حلاً قوياً لمراقبة وإدارة العمليات التي تتميز بعدد كبير من خصائص الجودة المرتبطة. من خلال دراسات المحاكاة، يتم إظهار فعالية وفائدة هذا النهج

المبتكر عملياً، مما يبرز إمكاناته في إحداث ثورة في ممارسات إدارة الجودة عبر الصناعات المختلفة. باختصار، هذه الأطروحة لا تساهم فقط في التقدم النظري للتحكم الإحصائي في العمليات، بل تقدم أيضاً حلولاً ملموسة لمواجهة التحديات العملية التي تفرضها مجموعات البيانات المعاصرة في سيناريوهات إدارة الجودة في العالم الحقيقي.

CHAPTER 1

INTRODUCTION

1.1 General Control Chart Basics

A control chart visually displays the performance of specific quality characteristic(s) in a process over time. It typically includes a central line representing the process mean, along with upper control limits (UCL) and lower control limits (LCL). These control limits can either be set by the practitioner or computed from the mean of the plotted data, with the latter being more commonly used in practice. [1] indicates that control limits are typically set to plus or minus three standard errors from a centerline that represents the historical average of the plotted statistic. To effectively implement a control chart, it is essential to observe the process data over a period of time [2]. Control charts are instrumental in determining whether a process is "in control" or "out of control" by distinguishing between two types of variation: chance causes and assignable causes. Chance causes of variation, which are inherent to the process, indicate that the process is in control [3]. Conversely, assignable causes, which signify that the process is out of control, necessitate further investigation to identify

the source of the variation and to restore the process to its controlled state.

The application of control charts is divided into two main stages: control (Phase I) and monitoring (Phase II). Phase I involves a retrospective analysis of historical system data to identify periods when the system was under statistical control. In contrast, Phase II focuses on the ongoing monitoring of the process to detect any shifts from an in-control state to an out-of-control state. It is essential to confirm that the process is in control before implementing Phase II control charts.

1.2 The Hotelling T^2 Control Chart

A multivariate control chart is a statistical tool utilized in quality and statistical process control to track process parameters defined by many quality characteristics. In the statistical process control (SPC) literature, several multivariate control chart types, including Hotelling T^2 [4], multivariate cumulative sum (MCUSUM) [5], and multivariate exponentially weighted moving average (MEWMA) [6] have been proposed. However, in our research, we employed the Hotelling T^2 control chart since it is the commonly used multivariate control chart on phase I analysis and thus takes into account the correlated structure among variable.

To utilize the T^2 statistic for multivariate process control, simultaneous monitoring of p -correlated quality characteristics is essential. We assume these p -quality attributes follow a p -variate normal distribution denoted as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ represents the mean vector and $\boldsymbol{\Sigma}$ denotes the variance-covariance matrix. Estimation of both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is typically required in practical scenarios. During the phase I procedure, a

historical dataset is used, comprising m observation vectors where each vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ is derived from a sample size of $p = 1$. The T^2 statistic is computed based on a single observation vector $\mathbf{x} = (x_1, x_2, \dots, x_p)$, typically obtained from in-control process data. The primary objective is to detect deviations in the process mean vector $\boldsymbol{\mu}$ from the hypothesized mean vector $\boldsymbol{\mu}_0$. The test statistic on the T^2 control chart for the i^{th} observation is expressed as:

$$T_i^2 = (\mathbf{x}_i - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_0) \quad (1.1)$$

Here, $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ represent the estimators of the variance-covariance matrix and the process mean vector, respectively.

The sample mean vector $\bar{\mathbf{X}}$, which serves as an estimator for $\boldsymbol{\mu}_0$, is given by:

$$\bar{\mathbf{X}} = \frac{1}{m} \sum_{i=1}^m x_i \quad (1.2)$$

And the sample covariance matrix $\mathbf{S}$, an estimator for the process variance-covariance matrix $\boldsymbol{\Sigma}$, is given by:

$$\mathbf{S} = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})(x_i - \bar{x})^T. \quad (1.3)$$

1.3 Motivation

In quality control, the validity, precision, and generalizability of engineers' conclusions can be severely impacted by the presence of missing values. Dealing with missing data

appropriately depends on various factors such as the context, goal, and underlying assumptions of the data. While the traditional method involves deleting samples with missing observations, this approach is simple and quick but can lead to reduced sample sizes, biased results, and loss of important information. An alternate method is to replace the missing values with reasonable estimates, which helps mitigate bias and ensures that results reflect a more accurate representation of underlying patterns and relationships. This approach also helps maintain statistical power, resulting in more robust and trustworthy analyses. Moreover, most statistical software programs require handling missing values before conducting any analysis [7]. Thus, this thesis aims to improve the overall quality and reliability of control chart implementation in Phase I datasets with missing values by offering practitioners a robust solution for addressing missing values through imputation using machine learning techniques. The performance of each methods used in imputing missing values were assessed through estimating the In-control average run length (ICARL). The missing values imputation method that results in an ICARL value closest to the specified ICARL value is considered to have the best performance.

Additionally, in modern manufacturing, with its emphasis on just-in-time production and reduced inventories, historical data may be insufficient to precisely estimate parameters for in-control processes. Increased automation and instrumentation have facilitated data collection, leading to a larger number of variables under observation. Consequently, achieving desired control chart performance may necessitate a larger sample size. In response, a self-starting approach can serve as an alternative to conventional control charts with estimated parameters. Hence, our objective is to devise a

novel supervised self-starting control chart tailored for multivariate data. The chart's performance was assessed through an analysis of different run length metrics, including Average Run Length (ARL) and Standard Deviation of the Run Length (SDRL).

1.4 Missing Data in Quality Control

The existence of missing value within multivariate control charts poses a unique set of challenges to the field of quality control. These missing values can manifest in both the phase II online data and the phase I historical dataset, stemming from various underlying factors. Addressing this issue is of paramount importance for engineers, as the integrity of data at either stage is critical for accurate process evaluation and control.

Notably, missing values within the phase I dataset bear a heightened significance. Mishandling of these missing values during the analysis of historical data can lead to a cascade of inaccuracies in the calculation of vital process parameters. Consequently, this can result in a compromised phase I performance, as it hinges on the precise evaluation of prior samples, whether they were in or out of control. The ripple effect of such inaccuracies extends to the phase II multivariate control chart, influencing its performance and effectiveness [8]. The consequences of missing data on the efficacy of multivariate control charts has garnered attention from researchers aiming to devise robust strategies for handling this challenge. Approaches such as imputation techniques utilizing machine learning algorithms have emerged as promising avenues to mitigate the adverse effects of missing values in quality control processes [9], [10],

[11]. These advancements aim not only to preserve the integrity of the data but also to fortify the reliability and accuracy of multivariate control chart analyses. By exploring these avenues, we endeavor to enhance the field of quality engineering in contemporary industries.

1.5 Multivariate Control Chart

Multivariate control charts are essential for tracking and preserving process quality in the context of SPC. These charts, designed to evaluate multiple correlated quality characteristics, provide a comprehensive view of process performance. Despite this, missing data significantly impairs these charts' accuracy and usefulness in practical applications. The impact of missing data on multivariate control charts is profound. It introduces uncertainty and can lead to erroneous conclusions about process stability. Missing values can significantly impact the accuracy and reliability of control charts, potentially leading to false alarms or failing to detect actual process shifts. Hence, the imputation of missing values emerges as a crucial enhancement. Addressing these missing values not only prevents the loss of valuable information but also preserves sample size integrity, ensuring the robustness and effectiveness of the control chart. Traditional approaches to address missing data often fall short, necessitating a more sophisticated strategy. This is where the power of machine learning techniques comes to play. By leveraging advanced imputation techniques, such as machine learning algorithms, we employed several machine learning imputation methods to fortify the capabilities of multivariate control charts and bolster their utility in real-world quality

control applications. This innovative approach not only overcomes the limitations associated with traditional methods but also harnesses the power of computational techniques and elevate the accuracy and efficiency of control chart analyses.

1.6 Self-Starting Control Chart

Modern manufacturing places a strong emphasis on just-in-time production, reduced inventories, and rapid innovation. However, these developments may leave us with insufficient historical data to calculate the in-control process's parameters with accuracy. A higher sample size may be necessary to achieve the desired control chart performance as a result of increased automation and instrumentation, which facilitate data collection and increase the number of variables under observation. In such scenarios, relying solely on conventional control charts with estimated parameters may not suffice. This is where a self-starting approach can be particularly advantageous. A self-starting control chart can begin monitoring a process with limited initial data and adapt as more data becomes available. This approach uses sequential updates to improve parameter estimates and control limits dynamically, making it a valuable tool in highly automated and data-rich environments. The self-starting control chart methodology involves updating control limits and parameter estimates continuously as new data points are collected. This allows for real-time adjustments and ensures that the control chart remains sensitive to any changes in the process. It provides a more accurate reflection of the current state of the process, leading to more reliable detection of out-of-control conditions. Furthermore, self-starting control charts can

help maintain process stability and quality in environments with frequent changes and updates. For example, in industries where product designs or manufacturing processes are regularly updated, having a flexible and adaptive control chart can prevent the delays and inaccuracies associated with recalculating parameters from scratch each time a change occurs. Thus, the integration of self-starting control charts into modern manufacturing processes addresses the challenges of limited historical data and the need for rapid adaptation. By continuously updating control limits and parameter estimates, self-starting control charts enhance the accuracy and reliability of process monitoring, ensuring higher quality and efficiency in production processes.

1.7 Thesis Objective and Significance

The primary aim of this thesis is to explore how various approaches of handling missing values, integrated with advanced machine learning imputation techniques, affect the performance of the Hotelling T^2 control chart. Specifically, we aim to evaluate how these approaches influence the estimated ICARL value, a key metric in assessing the efficacy of control charting in SPC applications. In addition to evaluating the impact of missing value imputation methods on the performance of Hotelling T^2 control charts, we seek to introduce a novel self-starting control chart. This innovative approach is designed to address the challenges associated with overseeing and managing processes characterized by multiple correlated quality characteristics. Thus, this self-starting control chart will offer practitioners a robust framework for real-time monitoring and intervention in high-dimensional datasets. Our dedication throughout this thesis is to

maintain the highest standards of precision, accuracy, and reliability in our analysis, thereby contributing to the advancement of statistical process control. In summary, our objectives in this thesis are to:

- Enhance multivariate control chart performance in the presence of missing data using advanced machine learning imputation methods. To achieve this, we will focus on improving the accuracy and reliability of control charts when there exist missing values in historical dataset. Specifically, our goals are to:
 - analyze the effectiveness of different imputation techniques in imputing missing values in multivariate control charts.
 - implement the most effective imputation methods in real-world scenarios and assess their practical benefits and challenges in maintaining process control.
- Introduce and validate a novel self-starting control chart for high-dimensional datasets. This objective aims to develop and test a control chart that can adapt to new data points and provide real-time updates. Specifically, our goals are to:
 - develop a self-starting control chart framework that dynamically adapts to new data points and updates control limits in real-time.
 - assess the chart’s performance through an analysis of different run length metrics, including ARL and SDRL.

1.8 Thesis Layout

The subsequent chapters of this thesis follow a structure outline: Chapter 2 provides an extensive review of the theoretical framework and prior research pertinent to the study. In Chapter 3, we utilize machine learning imputation methods to improve the effectiveness of multivariate control charts in scenarios with missing data. Chapter 4 introduces a novel approach: a supervised self-starting exponentially weighted moving average control chart based on interpoints dissimilarities. Lastly, Chapter 5 presents the conclusions drawn from the study and outlines potential directions for future research.

CHAPTER 2

THEORETICAL BACKGROUND AND LITERATURE REVIEW

The theoretical background section serves as the foundation for this thesis, providing a comprehensive exploration of the key concepts necessary for understanding the methodologies employed.

2.1 Overview of Control Chart

In chapter One (Subsection 1.0.1), we discussed the fundamental principles of control charts, detailing their role in monitoring process stability and detecting variations. Control charts come in various forms, each tailored to different scenarios. Univariate control charts, for instance, focus on a single quality characteristic, offering a clear and concise view of the process's stability. On the other hand, multivariate control charts extend this concept to scenarios where multiple correlated quality characteristics are interdependent. They are especially vital in industries where product quality depends

on a combination of factors.

Building on that foundation, we now turn our attention to the specific challenges and approaches within Phase I multivariate control charting. While Phase II charting strategies have received significant attention in the literature, Phase I multivariate control charting approaches are equally crucial. For instance, [12] introduced a Phase I control chart useful when only a single preliminary sample is available, assuming a multivariate normal model. This approach provides precise control limits for each statistic plotted. However, obtaining exact simultaneous control limits can be challenging, as [13] highlighted through their use of simulation to approximate these limits. Additionally, [14] examined the Phase I scenario, demonstrating that a mean shift cannot be effectively detected using the standard estimate of the covariance matrix. They explored various estimators that provide more accurate estimates under control conditions.

Nonetheless, since Shewhart initially presented control charts, technological developments have made it possible to collect more data at more precise granularities about processes. Thus, due to this information overload, the quality engineer is no longer able to effectively use single or several univariate charts for a process [15]. [16] indicated that rather than relying solely on the values of these attributes, the evaluation of a product's quality relies, among other things, on their combined effects. Since the multivariate control chart is design for tracking the process parameters of several correlated quality characteristics[17], it is a better tool to utilize than several charts when assessing distinct attributes [18]. In situations where a multivariate normal distribution appropriately models quality measurements and the objective is to monitor

the means of these measurements during Phase II, the utilization of a multivariate control chart presents itself as a practical alternative to employing separate charts. Noteworthy options in this category encompass the MEWMA chart [6], the MCUSUM chart [5], and Hotelling’s T^2 chart [19]. These charts provide a means to detect alterations in one or more of the variables, consequently signaling an out-of-control state. Consequently, an out-of-control signal indicated by these charts signifies that changes have occurred in one or more of the variables [6].

Multivariate control charts, originally introduced by Harold Hotelling in 1947 [20], offer a powerful tool for aggregating information from multiple process or product variables onto a single control chart. This is achieved through the use of the T^2 statistic, which quantifies the distance between the values (or their sample statistics) of these variables and the vector of their means, while also considering their inter-variable correlations as expressed in the covariance matrix [21]. Among the primary advantages of multivariate control charts lies in their capacity to detect certain signs of process deviation more effectively compared to traditional univariate charts. This heightened sensitivity arises from their consideration of the correlations between the variables of interest. Additionally, these charts offer the capability to set and control the type I error rate during their design, providing a reliable means to detect out-of-control processes. Another notable advantage is their capacity to consolidate data from multiple variables onto a single chart, streamlining the monitoring process for operators tasked with ensuring process quality [21].

2.2 Control Chart Literature with Missing Values

The issue of missing data is pervasive in statistical analyses, presenting a significant challenge in various fields. While extensive research has addressed methods to handle missing data, there has been a notable gap in exploring how control charting techniques specifically accommodate this issue. This literature review aims to explore key studies that have delved into the intersection of control charting and missing data.

[9] devised a multivariate control chart tailored for clinical laboratories. However, their approach, which entailed removing observations with missing values during the chart development process, may result in the loss of vital information and a reduction in sample size, thereby affecting the chart's sensitivity and effectiveness. [10] introduced an effective variance control chart designed to track the variance-covariance matrix of a process, even when confronted with missing data. In a similar vein, [11] put forth an innovative solution by advocating for the adjustment of weights in MEWMA control charts to accommodate missing data in Phase II. [22] stated that clinical records often contain incomplete data, requiring imputation of missing values before applying multivariate control charts. Their simulation study investigated the impact of imputing missing values using methods such as mean imputation, median imputation, and regression imputation on the run length performance of several multivariate charts, including the MEWMA, MCUSUM, and T^2 control charts. They concluded that regression imputation performs better than competing methods for imputing missing values. [23] examined the impact of five distinct approaches to handling missing values namely complete case analysis, unconditional mean substitution,

conditional mean substitution, stochastic regression imputation, and maximum likelihood estimation on the phase II MEWMA control chart's run-length performance. They came to the conclusion that, when the entire phase I dataset is not taken into account, some approaches for handling missing values have a negative impact on the control chart's performance in comparison to its correspondence. When estimating the Phase I historical dataset in control charts, [8] examined the effects of four imputation techniques: mean imputation, stochastic regression, regression, and the expectation-maximization algorithm for estimating Phase I historical dataset in control chart and estimated the unknown parameters in the Hotelling's T^2 chart statistic. They reported that the stochastic regression method has the best overall performance among all the competing methods. Also, in a research study focusing on estimating parameters from a "treated Phase I dataset" to assess Phase II control chart performance, various methods for handling missing data were examined. These methods includes complete case analysis, mean substitution, median imputation, regression, and stochastic regression [23]. Simulation results highlighted the superiority of imputation methods over case deletion approaches. Specifically, stochastic regression imputation demonstrated the best overall performance in improving the accuracy of control charting.

2.3 Overview of Self-Starting Control Chart and Related Literature

Self-starting control chart is an alternate method to parameter estimation in Phase I. The technique begins by monitoring the process without relying on prior knowledge

of the process. The self-starting employs the difference of each observation vector from the mean of previous processes. For instance, [24] introduced a self-starting multivariate control chart for monitoring process mean vectors. This technique was applied to monitor the Hotelling T^2 and MEWMA chart, as initially proposed by [4] and [6], respectively. [25], [26] developed a change-point model based on likelihood ratio statistics, which also employs self-starting techniques. [27] introduced a self-starting chart based on recursive residuals for monitoring linear profile, integrating two EWMA charts to track regression coefficients and process variability. [28] proposed another self-starting chart that combines the likelihood ratio test and EWMA statistic to monitor both process mean and variance simultaneously. Monitoring with Hotelling's T^2 control chart can only begin once the number of observations exceeds the dimensions of the process. [29] developed a multivariate self-starting chart that relies on the quadratic form of the observation vectors from a multivariate process, recommending a univariate EWMA chart to track a statistic based on this quadratic form. Moreover, [30] provide a comprehensive review of the impact of parameter estimation on control chart performance. Their findings indicate that when the reference sample size is small, control charts with estimated parameters exhibit a significant bias in the in-control (IC) average run length (ARL) and decreased sensitivity to the process, as indicated by the out-of-control (OC) ARL. Additionally, after short runs, false alarm rates increase substantially [31]. A straightforward solution to this issue is to increase the Phase I sample size to reduce the variability of estimates, particularly the process variance. However, this is often impractical as users typically need to monitor processes during the initial stages, such as in job-shop environments. Another

solution is to employ self-starting methods that update parameter estimates with new observations while concurrently monitoring the process to determine if it remains in control [32], [33]. To enhance the performance of self-starting control charts, [34], [35] introduced an adaptive Cusum of the Q chart, which can detect various shifts adaptively. Specifically, [25], [26], [36] proposed a change-point model based on the likelihood ratio for online monitoring, which also functions as a self-starting method.

The existing nonparametric control charts, as well as self-starting control charts, have limitations when it comes to managing processes that involve correlated counts datasets. These charts are primarily intended for processes that involve continuous quality characteristics which are not normally distributed [37]. Consequently, they are not well-suited for efficiently handling multivariate discrete observation data, such as those commonly encountered in health-related monitoring.

CHAPTER 3

ENHANCING THE EFFICIENCY OF MULTIVARIATE CONTROL CHARTS IN THE PRESENCE OF MISSING VALUES USING MACHINE LEARNING TECHNIQUES

This chapter focuses on using imputation techniques to enhance the usefulness of multivariate control charts when there are missing values. Imputation involves estimating missing values based on observed data, enabling the completion of datasets, and facilitating more accurate analysis. By exploring various imputation methods, we aim to

develop strategies that improve the reliability and effectiveness of multivariate control charting.

3.1 Missing Value Imputation Methods

In the realm of SPC, ensuring data integrity is paramount for accurate analysis and decision-making. However, datasets with missing values pose a challenge, potentially hindering the effectiveness of control charting and process monitoring. To overcome this challenge, researchers and practitioners have developed various missing value imputation methods tailored to the unique requirements of control chart analyses. In this section, we explore both traditional and machine learning-based methods for imputing missing values. By comparing their effectiveness in terms of estimating the ICARL, we aim to provide insights that would assist practitioners in making informed decisions when faced with missing data in control chart applications.

3.1.1 Traditional Imputation Method

Several imputation methods have been described in the literature. The most common methods for imputing missing values in realm of control chart include:

Mean Imputation: mean imputation is a commonly used technique for handling missing data in statistical analyses. This method involves replacing missing values with the arithmetic mean of the observed values within each variable. The simplicity and ease of implementation make mean imputation an attractive option. However, while it can provide a quick solution, it comes with significant drawbacks. Mean

imputation can introduce bias, especially if the missing data are not random, and tends to underestimate the variability in the dataset. These limitations can affect the accuracy and reliability of subsequent analyses, making it crucial for researchers to carefully consider the assumptions and potential impacts when employing this method.

3.1.2 Machine Learning-Based Imputation Methods

As the landscape of data analysis advances, the use of machine learning techniques becomes more common in numerous sectors. The necessity for sophisticated imputation methods in control chart analysis has prompted researchers to investigate machine learning algorithms. In this subsection, we present four state-of-the-art machine learning techniques:

Multiple Imputation by Chained Equation (MICE): This imputation method entails the iterative utilization of multiple regression models to predict missing values based on other variables in the dataset [38]. It takes into consideration the data type of each variable during the modeling process. Specifically, continuous variables are imputed using predictive mean matching, whilst variables that are binary are handled by logistic regression [38]. The approach involves four primary steps, as delineated by Azur et al. [38], which are iteratively executed until satisfactory imputations are achieved. Missing values are initially substituted by the variable's mean observed values, acting as placeholders. These mean imputations are then reset to "missing" in the subsequent step. The third stage comprises regressing the values that were observed of a variable, denoted as "x" on the remaining variables, with "x" serving as the dependent variable and the others are independent. Predictions obtained

from this regression model are subsequently utilized to substitute the values missing in the fourth step. The imputed value, along with observed values for other variables, is integrated into the set of independent variables. This iterative process is applied for each variable with missing values, with one iteration comprising steps "2" through "4" [39]. For further insights into Multiple Imputation by Chained Equation (MICE), readers are directed to Azur et al. [38].

k-Nearest Neighbor Imputation: The kNN approach uses a distance metric between instances to categorize the closest neighbors of missing values, which are then used for imputation [40]. kNN imputation can be done using a variety of distance measures. The Euclidean distance as shown in **Eq.3.1** is the most used distance measure as it is proven to provide a plausible estimate for missing values [41].

$$Dist_{xy} = \sqrt{\sum_{k=1}^m (X_{ik} - X_{jk})^2} \quad (3.1)$$

$Dist_{xy}$ denote Euclidean distance, k denotes the data feature, and $j = 1, 2, 3, \dots$ k represents the dimension of the data. The values of the j attributes with values missing are represented by X_{ik} and those with complete data by X_{jk} .

After determining which k points have the smallest distance, the weight mean estimate is computed as shown in **Eq.3.2**.

$$X_k = \frac{\sum_{j=1}^j w_j v_j}{\sum_{j=1}^j w_j} \quad (3.2)$$

where j denote the variables used, and $j=1,2,3,\dots,k$; X_k is the estimated mean. w_j

denotes the closest neighbors observed, v_j represents complete values for attributes with missing data.

Next, the weighted value is obtained using **Eq.3.3**

$$w_j = \frac{1}{Dist_{(x,y)^2}} \quad (3.3)$$

However, kNN imputation possesses limitations, such as lower accuracy in variable imputation and higher computation time since it examines all datasets to locate the most similar cases [42]. A modified version of the kNN approach to missing data imputation has been proposed in shahla [43], [44]. This method uses a weighted average of nearest neighbors (NNs) based on L_q distances.

Weighted Neighbors: Imputation results are not always improved and using data from all nearest neighbors might be intensive, particularly when dealing with huge datasets. An alternative approach proposed by shahla [43] is to use weighted nearest neighbor method. Two crucial steps are needed in using NNs to impute missing values:

1. define how the NNs are computed and which distance measure to used
2. how the calculated NNs are used to obtain imputed values

In the computation of NNs, consider an $n \times p$ data matrix with n observations on p covariates, let $X = (x_{is})$ be the i -th observation of the s th variable. Let $O = (o_{is})$ be the appropriate $n \times p$ -dummies matrix with the following entries:

$$o_{is} = \begin{cases} 1 & \text{if } x_i \text{ is not missing} \\ 0 & \text{if } x_i \text{ is missing} \end{cases}$$

For the observed data, the L_q metric can be used to calculate the distance between two observations x_i and x_j , represented by rows in the data matrix as:

$$d_q(x_i, x_j) = \left(\frac{1}{m_{ij}} \sum_{s=1}^p |x_{is} - x_{js}|^q \cdot I(o_{is} = 1) \cdot I(o_{js} = 1) \right)^q \quad (3.4)$$

where $I(a)$ is the indicator variable employed in the fomulation; It takes the value 1 if a is true, otherwise 0. The number of valid components in calculating distances is denoted by $m_{ij} = \sum_{s=1}^p I(o_{is} = 1) \cdot I(o_{js} = 1)$. Imputation of distances only uses vector components for which observations are available in both vectors, not all vector components. $C_{ij} = (s : I(o_{is} = 1) \cdot I(o_{js} = 1) = 1)$ is the components that are included in the calculation of neighbors. The distances define the NNs when assuming a certain value. It's crucial to keep in mind that the quantity of components varies within the sample, as C_{ij} is dependent on the particular observation necessitating imputation.

After obtaining the distances that define the NNs, we used these distances to impute missing values through weighted method.

Imputation by Weighting

One drawback of the k-NN imputation method is that it treats the k th nearest neighbor value as equally important as the first nearest neighbor value. A preferable alternative involves incorporating weights that consider the distances between observations. We advocate for computing a weighted average of

neighbors by utilizing distances derived from kernel functions. The weighted imputation estimate considered has the form shown in **Eq.3.5**

$$\hat{x}_{is} = \sum_{j=1}^k w(x_i, x_{(j)}) x_{js} \quad (3.5)$$

with weights

$$w(x_i, x_{(j)}) = \frac{f\left(d\left(\frac{(x_i, x_{(j)})}{\lambda}\right)\right)}{\sum_{l=1}^k f\left(d\left(\frac{(x_i, x_{(l)})}{\lambda}\right)\right)} \quad (3.6)$$

where $f(\cdot)$ serves as the kernel function and the rotation parameter is λ . As λ diminishes, the weights experience a substantial decrease in correlation with distance, whereas as $\lambda \rightarrow \infty$, all neighbors are attributed equal weight. To emphasize the significance of λ as the pivotal changing parameter, many of the potential neighbors should be employed.

Shahla [43] proposed an improved version of the observed data metric based on L_q to measure the distance between two observations, x_i and x_j .

Examine the imputation of x_i in the component s ($o_{is} = 0$). When calculating distances from the reduced dataset, an additional weight is included in the distance calculation ($x_i, o_{is} = 1$). In particular, the component-specific distance is calculated for a given distance L_q .

$$d_{q,c}(x_i, x_j) = \left(\frac{1}{m_{ij}} \sum_{s=1}^p |x_{is} - x_{js}|^q I(o_{is} = 1) I(o_{js} = 1) C(r_{sl}) \right)^{\frac{1}{q}} \quad (3.7)$$

The actual relationship between covariates s and l is indicated by r_{sl} , and $C(\cdot)$

translates the correlations into weights using a convex function given on the interval $[-1,1]$. The transformation is designed in a way that only the covariates that have a strong correlation with component s strongly influence the distance calculation. Let $C(\cdot)$ be a convex function given by:

$$C(r) = \begin{cases} \frac{|r|}{1-c} - \frac{c}{1-c} & \text{if } |r| > c \\ 0 & \text{if } |r| \leq c \end{cases} \quad (3.8)$$

The component does not contribute to the distance if $|r_{sj}| \leq c$. In the event that $|r_{sj}| > c$, the weight increases linearly with increasing correlation and is equal to 1 when $|r| = 1$.

Using the distances calculated according to equation 3.10, the observations are sorted in ascending order as $x_{(1)}, \dots, x_{(k)}$, where $X_j^T = (x_{(j)1}, \dots, x_{(j)p})$ represents the j th nearest neighboring observation with $d(x_i, x_{(1)}) \leq \dots \leq d(x_i, x_{(k)})$.

The estimated value of x_{is} is determined by:

$$\hat{x}_{is} = \sum_{j=1}^k w(X_i, X_{(j)}) x_{(j)s} \quad (3.9)$$

with a weighting factor given as:

$$w(x_i, x_j) = \frac{f\left(\frac{d(X_i, X_j)}{\lambda}\right)}{\sum_{h=1}^k f\left(\frac{d(X_i, X_{(h)})}{\lambda}\right)} \quad (3.10)$$

where $f(\cdot)$ represents a kernel function and λ serves as a tuning parameter.

Random Forest: Random Forest expands upon the Classification and Regression

Tree (CART) methods proposed by Breiman et al. [45]. It can handle both regression-type problems (predicting a continuous dependent variable) and classification-type problems (predicting a categorical dependent variable). In regression problems, Random Forest consists of an ensemble of simple trees, each trained on a subset of independent variables. These trees collectively contribute their predictions, which are aggregated to produce an estimate of the dependent variable. To construct a regression forest, variables and data are randomly selected using an iterative bootstrap sampling technique. Additionally, Random Forest implementation involves tuning stopping parameters and decision tree regularization. These parameters include the number of predictor variables randomly selected in each node (denoted as k) and the maximum number of trees in the forest. It is essential to note that limiting the number of predictor variables may reduce predictive effectiveness by excluding variables crucial for capturing data variability and trends. The stopping parameters, also known as control parameters, dictate when the algorithm should halt once sufficient results are achieved [46]. The imputation algorithm use for filling in missing data in random forest is called **MissForest**.

3.2 Evaluating the Effect of Missing Data Handling Methods on Hotelling T^2 Control Chart Performance in Phase I

We conducted a simulation study to assess the effectiveness of six imputation methods, namely kNN, wNNSel, wNN, MICE, and MissForest (machine learning-based methods), as well as Mean imputation (a conventional approach), in addressing missing values' impact in Phase I on the performance of Hotelling T^2 control chart, particularly concerning the estimation of an ICARL. The estimation of ICARL entails four sequential steps.

Initially, a Phase I historical dataset is generated using a multivariate normal distribution with a mean vector $\mu_0 = (\mathbf{0}, \mathbf{0}, \dots, \mathbf{0})_{1 \times p}$ and a variance-covariance matrix $\Sigma_0 = \mathbf{I}_p$, tailored to specific parameter values of n and p .

The second step involves intentionally introducing missing values at rates of 5%, 10%, 15%, 20%, and 30% into the initially generated in-control dataset. The missing data pattern being considered is the general pattern of missingness, where any variables within the dataset may be missing for any observations. Specifically, a random deletion of precisely $m\%$ from the total number of values in the dataset ($n \times p$ values) occurs the phase I dataset.

Moving on to the third step, various missing value imputation approaches are applied to estimate the values that are missing from the dataset. Next, the mean vector μ is estimated using the sample mean vector $\bar{\mathbf{X}}$, whereas the variance-covariance

matrix Σ is estimated using the sample variance-covariance matrix $\mathbf{S}$.

The fourth step involves generating Phase II data employing T^2 chart with the estimated parameters $\bar{\mathbf{X}}$ and $\mathbf{S}$. Subsequently, these four steps are repeated 10,000 times in an iterative manner to assess the in-control average run length (ICARL).

3.3 In-control Performance

The Phase I in-control performance of T^2 control chart was assessed after estimating the missing values using various imputation methods such as kNN, wkNN, wNNSel, MICE and MissForest (machine learning-based approaches) and mean imputation (a traditional methods). We account for two values of the correlation coefficient ρ , three values of the quality attributes p , six values of the percentage of missingness m , and a historical sample of size 200. The effectiveness of the imputation method is measured by comparing the ICARL value obtained from various missing value imputation methods with 200. Missing value imputation methods with an ICARL value close to 200 are believed to have the highest performance. Additionally, the values of h that yield an ICARL 200 when there is no missing value were calculated using 10,000 simulation runs. The values of h that correspond to the number of quality characteristics (p) and sample size (m) are shown in Table 3.1.

Table 3.1: The values of h that result in an ICARL when no missing value is present.

Quality Characteristics	No. of Observation	h
5	200	17.53
6	400	19.6
10	200	27.43
30	200	67.4

3.3.1 Results of the In-control Performance when Correlation Coefficient Between Quality Characteristics is Zero

The estimated ICARL values for the six approaches for addressing missing values used in this thesis when $\rho = 0$ are provided in Table 3.2. These ICARL values were obtained using the simulation procedure described earlier (in section 3.2), and the chart statistics were constructed using the control limits specified in Table 1. In general, the results of Table 2 illustrate a pronounce decrease in ICARL values as the percentage of missing values increases across all values of p . Specifically, for all methods that handle missing values, the ICARL values exhibit a decline with rising percentage of missing values. As an example, with $n = 200$, $p = 5$, and $m = 5\%$, the ICARL values generated by kNN is 160.89. When the percentage of missing values rises to 30%, this ICARL falls to 59.65. In instances where $\rho = 0$, the MICE method for imputing missing values consistently outperforms the other methods. Across all scenario examined in our simulations with $\rho = 0$, the ICARL values estimated by the MICE method tend to be closer to 200, except for $p = 30$, where the wkNN method outperforms the MICE method. Furthermore, the MissForest, kNN, wNNSel, machine learning-base methods, and mean, a traditional method, for imputing missing values demonstrate better performance when considering small values of quality characteristics p ; however, they performs poorly for all cases compared to other machine learning-based imputation methods examined in this research when p assumes larger values.

Table 3.2: Values of ICARL when $\rho = 0$.

		Imputation Methods						
Quality	Charact	m	kNN	MICE	MissForest	Mean	wNNSel	wkNN
$p = 5$		0.00	203.29	199.41	202.36	202.81	200.97	202.85
		0.05	153.14	193.18	140.99	143.13	140.58	153.78
		0.10	116.66	180.02	101.98	102.48	100.38	116.62
		0.15	88.06	170.21	71.03	70.78	72.86	89.39
		0.20	68.38	159.16	50.84	50.46	50.58	66.61
		0.30	39.49	133.99	26.34	25.77	26.31	36.31
$p = 10$		0.00	197.95	200.47	199.98	199.98	199.98	199.85
		0.05	139.62	175.70	123.80	128.95	131.41	152.43
		0.10	96.25	155.12	76.37	84.83	86.76	107.12
		0.15	70.68	133.33	47.85	55.02	55.17	81.21
		0.20	49.16	108.78	30.77	37.07	36.81	58.45
		0.30	26.26	72.66	12.74	16.77	16.38	30.76
$p = 30$		0.00	199.08	198.83	196.32	198.15	195.85	199.72
		0.05	111.89	117.67	94.51	105.82	105.47	114.96
		0.10	69.38	68.27	46.98	60.13	58.34	68.17
		0.15	43.44	36.81	25.12	34.19	33.11	44.64
		0.20	28.85	17.16	13.22	20.10	19.34	28.54
		0.30	12.83	3.51	4.64	7.44	7.10	13.72

3.3.2 Results of the In-control Performance when Correlation

Coefficient Between Quality Characteristics is 0.7

This subsection presents the estimated ICARL values for the cases where $\rho = 0.7$, estimating missing values through various methods, including kNN, wkNN, wNNSel, MICE, MissForest (machine learning-based approaches) and Mean (a traditional method). Table 3.3 provides the estimated ICARL values obtained using the specified methods when $p = 5, 10$, and 30 . Across all cases considered in this research, a consistent finding shows that the ICARL values corresponding to the methods increase as the value of correlation coefficient between quality characteristics increases from 0 to 0.7. For instance, consider the estimated ICARL values using the MICE method,

which is 180.02 when $\rho = 0$, $p = 5$, $n = 200$, and $m = 10\%$. As the correlation coefficient increases to 0.7, the ICARL value rises to 187.84. However, in cases involving both a high correlation coefficient and a substantial number of missing values, this increase might be very large. For the kNN method, the estimated ICARL value is 155.45 when $p = 30$, $m = 20\%$, and $\rho = 0.7$, whereas its value equals 28.85 when $\rho = 0$.

Table 3.3: Values of ICARL when $\rho = 0.7$.

		Imputation Methods						
Quality	Charact	m	kNN	MICE	MissForest	Mean	wNNSel	wkNN
$p = 5$		0.00	204.09	202.48	203.50	202.45	207.57	202.88
		0.05	160.89	193.33	147.37	220.32	151.29	160.92
		0.10	133.12	187.84	106.80	199.86	116.65	133.36
		0.15	108.41	174.94	76.57	161.42	98.54	114.59
		0.20	91.31	165.23	54.10	125.55	86.14	95.01
		0.30	59.65	137.08	28.56	62.16	58.10	62.80
$p = 10$		0.00	199.98	199.98	199.98	199.98	200.52	199.98
		0.05	167.62	175.42	130.60	243.20	153.99	163.39
		0.10	146.11	154.00	88.18	217.96	112.05	129.73
		0.15	128.58	131.29	58.34	175.66	84.23	106.44
		0.20	109.62	105.50	39.77	130.67	62.66	87.84
		0.30	71.46	63.35	18.16	57.22	38.25	67.25
$p = 30$		0.00	198.11	200.79	193.60	194.13	196.66	196.88
		0.05	195.96	116.85	118.24	265.66	176.09	180.99
		0.10	199.09	59.61	69.76	269.88	139.73	140.55
		0.15	186.27	27.33	43.11	214.28	101.96	112.93
		0.20	155.45	11.09	26.11	148.08	69.59	83.76
		0.30	84.83	2.19	10.01	51.27	27.91	46.79

3.4 Application

This section aims to assess various methods for missing value imputation using real-world data. The dataset utilized, sourced from the UCI machine learning repository

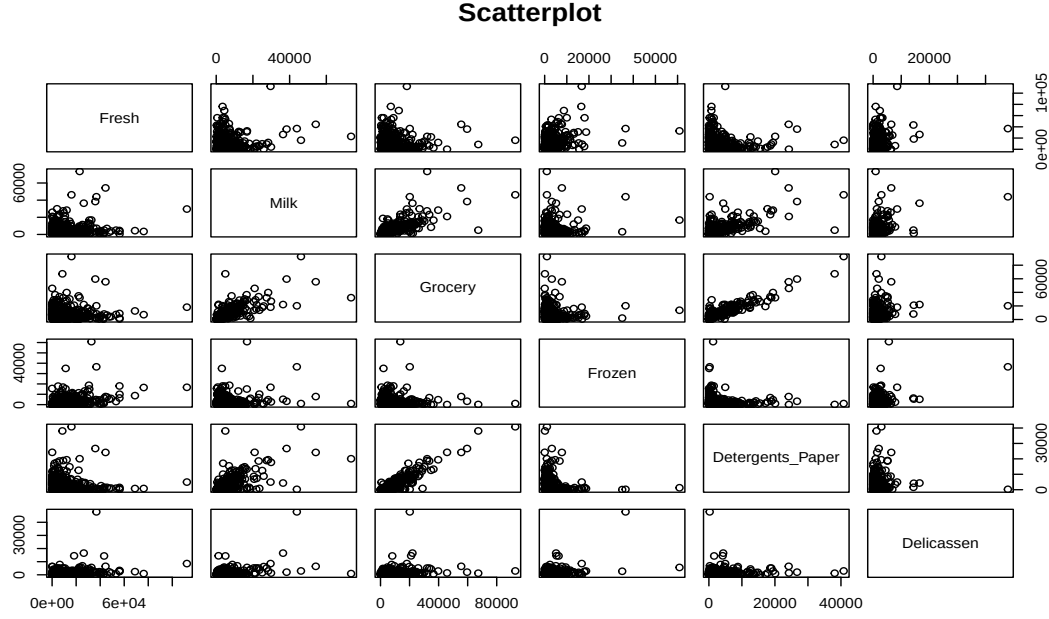


Figure 3.1: Scatterplot of wholesale distributor

database, pertains to clients of a wholesale distributor. It comprises annual spending (in monetary units) on different product categories, comprising eight features: channel, region, fresh, milk, grocery, frozen, detergent (paper), and delicatessen. The dataset consists of 440 observations. We remove channel and region in our analysis since they are categorical data.

To begin the analysis, we constructed a scatter plot matrix by plotting each feature against every other feature, as demonstrated in figure 1. This visualization provides insight into the relationships between different product categories and helps identify patterns or correlations within the data. Upon examining the scatter plot matrix, we observed that while some features exhibit positive relationships, there is no evident relationship among other variables.

To evaluate the performance of different methods of missing value imputation, we initially plotted the monitoring graph for the real dataset using Hotelling’s T^2 control chart. We removed all points that fall outside the upper control limit to refine the dataset. The graph of monitoring plot before and after refining the data is shown in figure 3.2. Subsequently, we introduced 30% missingness into this refined dataset to evaluate the performance of various missing value imputation methods: kNN, MICE, MissForest, mean, wNNSel, and wkNN. For each imputation method, we plotted a corresponding monitoring graph to analyze their effectiveness in imputing missing values as shown in Figure 3.3. Our findings revealed that the MICE imputation method closely aligned with the monitoring graph of the refined dataset, outperforming the other methods. Therefore, we recommend employing the MICE imputation method for practical applications, particularly for imputing missing values in phase I with quality characteristics of 15 or fewer. This recommendation is further supported by simulation results discussed in sections 3.2 and 3.3. The monitoring graphs for wkNN and wNNSel were not included, as their graphs were analogous to that of kNN.

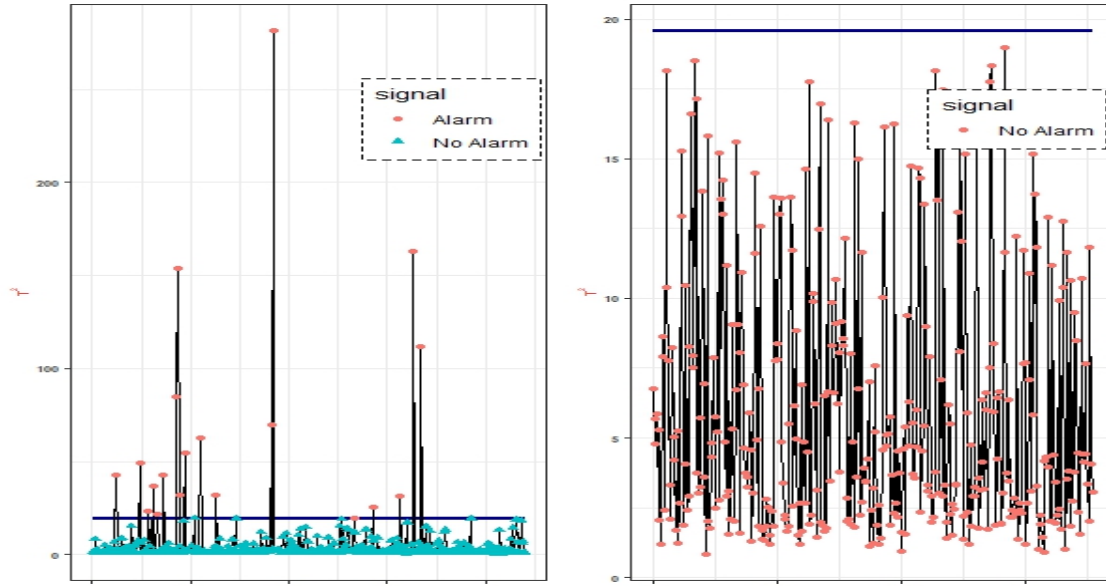


Figure 3.2: Monitoring plot of a wholesale distributor without missing values

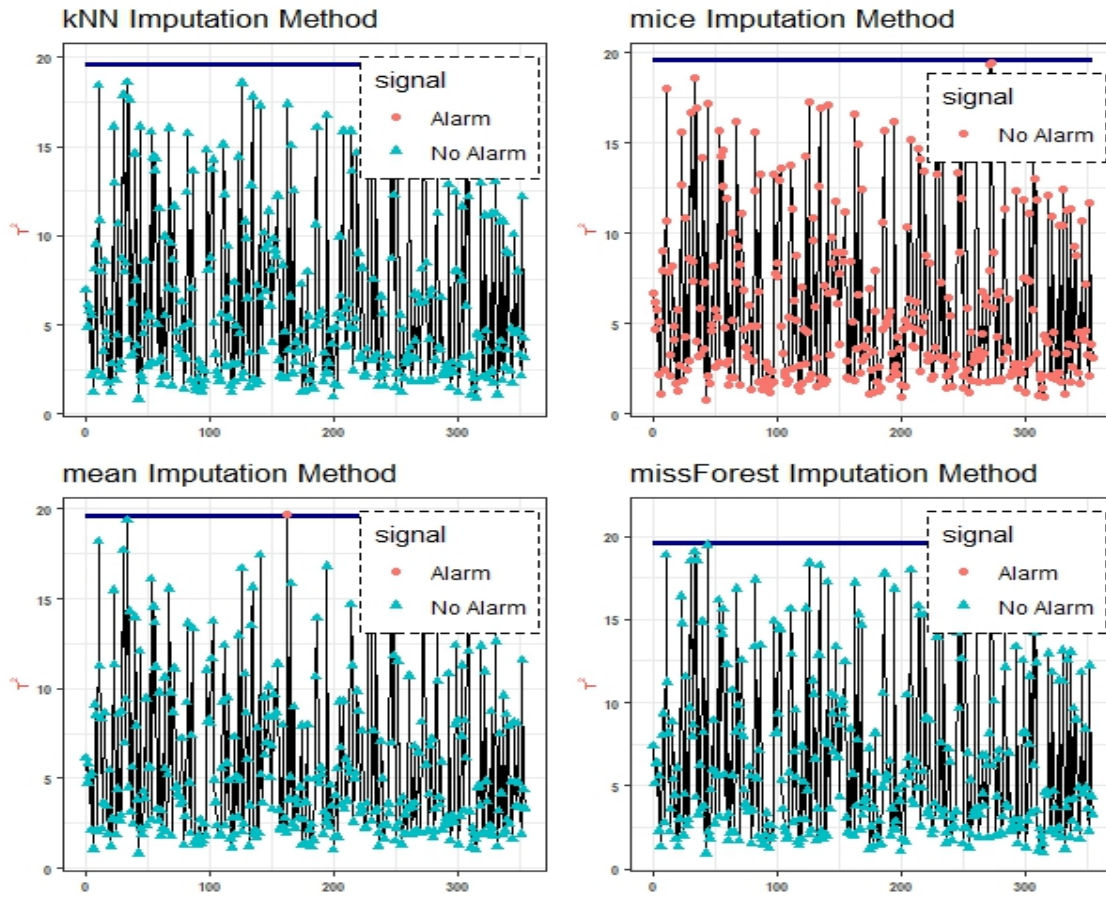


Figure 3.3: Monitoring plots of a wholesale distributor through four comparative charts.

3.5 Summary

In this study, we investigate the performance of various imputation methods for handling missing values in Phase I and their impact on the Hotelling T^2 control chart. Among the methods evaluated, Multiple Imputation by Chained Equations (MICE) demonstrated the highest performance for correlation coefficients (ρ) of 0 and 0.7 when the number of quality characteristics (p) was 5 or 10. For larger datasets with $p = 30$, the weighted k-nearest neighbors (wkNN) method performed best for $\rho = 0$, while the k-nearest neighbors (kNN) method excelled for $\rho = 0.7$. Conversely, the MissForest method showed the lowest performance across all scenarios. Mean imputation, MissForest, and weighted nearest neighbor selection (wNNSel) yielded nearly identical ICARL values in most cases.

CHAPTER 4

A SUPERVISED MULTIVARIATE SELF-STARTING EXPONENTIALLY WEIGHTED MOVING AVERAGE CHART BASED ON INTERPOINTS DISSIMILARITIES

4.1 Introduction

In today's dynamic and complex industrial landscape, the need for effective SPC methodologies is more pronounced than ever. Traditional univariate control charts

have been extensively utilized for monitoring single quality characteristics. However, modern manufacturing processes often involve multiple correlated quality variables, necessitating devising a multivariate control charts. In this chapter, we introduce a novel approach to multivariate statistical process control, termed the Supervised Multivariate Self-Starting Exponentially Weighted Moving Average (SSEWMA) chart. This chart is specifically designed to address the challenges associated with monitoring processes characterized by multiple correlated quality attributes. At the core of our proposed methodology lies the concept of interpoints dissimilarities, which enables the characterization of relationships among multivariate observations. Moreover, the self-starting nature of our proposed chart eliminates the need for historical data for initialization, making it particularly suitable for monitoring newly established processes. This novel approach aims to contribute to the advancement of multivariate SPC methodologies, offering practitioners a valuable tool for maintaining process stability and enhancing product quality.

4.2 Change-point model

We examine a series of independent observations $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})'$ ($i = 1, \dots, n$) from a p -dimensional process characterized by a mean vector $\boldsymbol{\mu}_0 = (\mu_{01}, \dots, \mu_{0p})$ and covariance matrices $\boldsymbol{\Sigma}$. The operational model

$$\begin{cases} F_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), & \text{if } i \leq \tau \\ F_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), & \text{if } i > \tau \end{cases}$$

establishes a p-variate representation of change-point dynamics within distribution F_p . The parameter τ , known as the change point, remains unknown. The remaining parameters, $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\Sigma}_1$, delineate the mean vectors and covariance matrices associated to the in-control and out-of-control states. Although some change-point models assume certain parameters to be known, we concentrate on scenarios where all parameters remain unknown. Numerous variations of the general model exist; we focus on a scenario where $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0$, resulting in a mean-only change-point model. Additionally, we assume that the unidentified parameters remain constant after the shift, signifying the presence of a single-step change point awaiting identification.

In the classical setting, the random variables $\mathbf{X}_i$ are drawn from multivariate normal distribution, and the change-point model becomes:

$$\begin{cases} N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), & \text{if } i \leq \tau \\ N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), & \text{if } i > \tau \end{cases}$$

Several models for detecting mean changes in univariate/multivariate normal processes have been developed, as documented in works such as [14], [25], [26], [36], [47], [48], [?], [49], among others. These models have been extended to monitor linear profiles [50], [27], [51] and multistage processes [52]. An overview of these methodologies is available in [53].

Although classical self-starting charts have demonstrated utility for datasets with multivariate independent normal or "nearly" normal distributions, many real-world systems exhibit non-normal distributions. Furthermore, the monitoring of processes

characterized by discrete, count, or mixed discrete and continuous features is increasingly important. In such scenarios, the statistical assumptions of classical charts may not hold, impacting their performance. Additionally, routine tests for normality may be impractical.

In this work, we assume $\mathbf{X}_i$ are random samples from a process having an unknown continuous or a mixture of continuous distribution with an unknown location parameter $\boldsymbol{\mu}$. In some cases, such distributions may also include standard deviation, scale, or rate parameter, etc. For example, the multivariate normal distribution has both mean vector and covariance parameters; the multivariate gamma distribution has shape and scale parameters. Our proposed self-starting method based on sum of squared of interpoints dissimilarities does not make any assumption regarding the process distribution. The only assumption we make is that the true process distribution is characterized by parameters that include location or mean parameter (μ).

4.3 Existing self-starting statistics based on sum of squared of interpoints dissimilarities

We initiate by presenting a succinct overview of univariate self-starting statistics, which rely on the summation of squared interpoints dissimilarities, before introducing their multivariate analog. Here, we denote the in-control value of μ as μ_0 . At any given time t , our aim is to assess the degree to which the observation x deviates from the norm, taking into account observations up to time $(t - 1)$. This assessment is

quantified by

$$d_t = x_t - \bar{x}_{(t-1)} \quad (4.1)$$

where $\bar{x}_{(t-1)} = \frac{1}{(t-1)} \sum_{i=1}^{(t-1)} x_i$ [54], [55]. Thus, d_t denotes the difference between an initial observation at time t from the mean of previous observations. We investigate that, under the null hypothesis assuming the process is "in control," the x_t follow a randomized pattern with a uniform expected value μ , determined by $\bar{x}_{(t-1)}$. Consequently, under the null hypothesis, d_t constitute stochastic deviations with an expected value of zero, as expressed by

$$H_0 : \mu = \mu_0 \quad \text{implies} \quad H_0 : d_t = 0$$

abnormally large value of d_t signify the process at time t is "out of control." In essence, we anticipate that a shift in μ_0 would manifest as significantly large absolute values of d_t .

In monitoring a p -dimensional quality characteristic $\mathbf{X}_{n \times p}$, the nominal value of the location parameter is $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ when the process is under control. Thus, detecting deviations in the mean vector $\boldsymbol{\mu}$ equates to testing the hypothesis:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

Under the null hypothesis, where the process is "in control," individual data points stochastically fluctuate around a central value $\boldsymbol{\mu}_0$ in the multivariate p -space over time. Here, we assume $\boldsymbol{\mu}_0$ to be unknown and expect natural variations if the process

remains in-control. In this multivariate setting, the difference between the present observation and the average of all previous observations is expressed as:

$$\mathbf{d}_t = \mathbf{x}_t - \bar{\mathbf{x}}_{t-1} \quad (4.2)$$

where $\bar{\mathbf{x}}_{(t-1)}$ represents a vector of means $\bar{x}_{(t-1)k} = \frac{1}{(t-1)} \sum_{i=1}^{(t-1)} x_{ik}$, $k = 1, 2, \dots, p$, with $\mathbf{x}_0 = \mathbf{0}$ [54], [55].

In the case where $\mathbf{x}$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, [54], [55] introduced scaled statistics as $\mathbf{u}_t = \mathbf{d}_t \sqrt{\frac{t-1}{t}}$, with $\mathbf{u}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. It was demonstrated that the scaled statistics $\mathbf{u}_t$ are independent and identically distributed under the "in-control" scenario. A transformation was proposed as follows:

$$Q'(\mathbf{z}_t) = \Phi^{-1} \left[F_{p, t-p-1} \left(\frac{(t-1)(t-p-1)}{tp(t-2)} \right) \times (\mathbf{u}_t^T \mathbf{S}_{\mathbf{u}_{t-1}}^{-1} \mathbf{u}_t) \right] \quad (4.3)$$

where $t = p+2, p+3, \dots$, $F_{p, t-p-1}$ denotes the cumulative distribution function (CDF) of the F distribution with p and $t-p-1$ degrees of freedom, and Φ^{-1} represents the inverse of the CDF of the standard normal distribution [54], [55]. By applying the transformation in Equation (4.3), the in-control transformed chart statistic $Q'(\mathbf{z}_t)$ conforms to a standard normal distribution (i.e., $Q'(\mathbf{z}_t) \sim N(\mathbf{0}, \mathbf{I})$). The sample (or empirical) covariance matrix $\mathbf{S}_{\mathbf{z}_t}$ of $\mathbf{z}_t$ at time t is computed as $\mathbf{S}_{\mathbf{u}_t} = \frac{1}{t-1} \mathbf{H}_t$, where $\mathbf{H}_t$ is recursively defined as $\mathbf{H}_t = \mathbf{H}_{t-1} + \mathbf{u}_t \mathbf{u}_t^T$, $t = 3, 4, \dots$, with $\mathbf{H}_0 = \mathbf{0}_{k \times k}$ [54], [55]. For more insights into Equations (4.2 - 4.3) and their extensions in classical self-starting control charts for multivariate normally distributed datasets, readers are

referred to [24].

4.4 Multivariate Self-starting exponentially weighted moving average chart based on interpoints dissimilarities

4.4.1 A self-starting statistics based on sum of squared of interpoints dissimilarities

A notable limitation of Equation (4.3) is its heavy reliance on the assumption of multivariate normality in the process dataset, an assumption that may not always hold in real-world monitoring scenarios. To address this, we explore a straightforward extension of Equation (4.2) that does not depend on the distributional assumptions of the process. We investigate the utilization of the sum of squared interpoint dissimilarities.

[56] proposed a multivariate criteria, similar to Equation (4.2), as the dissimilarity between the observation at time t and the centroid in the multivariate space formed by the preceding $(t - 1)$ observations. Consequently, the criterion based on Euclidean distances is expressed as:

$$d_t = ||\mathbf{x}_t - \bar{\mathbf{x}}_{(t-1)}|| = \sqrt{\sum_{k=1}^p (x_{tk} - \bar{x}_{(t-1)k})^2} \quad (4.4)$$

This criterion contrasts with Equation (4.2) in that, as a distance metric, it exclusively takes positive values and lacks an expected value of zero. However, even under the

null hypothesis, it remains a stochastic quantity with a consistent expected value that varies stochastically over time.

A generalization of Equation (4.4) is necessary to develop a multivariate monitoring criterion based on any selected distance or dissimilarity metric. In such instances, the "centroid" no longer refers to the arithmetic mean, particularly in non-Euclidean spaces. However, the criterion is based on the fact that the sum of squared interpoint distances divided by the number of points equals the sum of squared departures from the centroid. For a set of m points in multivariate space, this relationship holds:

$$\sum_{i=1}^m \|\mathbf{x}_i - \bar{\mathbf{x}}_n\|^2 = \frac{1}{m} \sum_{i=1}^{m-1} \sum_{j=i+1}^m \|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad (4.5)$$

known to hold for Euclidean distances and extended by [57] to other dissimilarity measures using principal coordinates. This result was leveraged by [58] for partitioning multivariate sums of squares in ANOVA designs directly through interpoint dissimilarities. In a similar study, [56] used interpoint dissimilarities to determine the monitoring criterion, eliminating the necessity of calculating the "centroid" for non-Euclidean spaces. Specifically, at any given time t , [56] defined the following two quantities:

$$SS_t = \frac{1}{t} \sum_{i=1}^{t-1} \sum_{j=i+1}^t D_{ij}^2,$$

$$SS_{t-1} = \frac{1}{(t-1)} \sum_{i=1}^{t-2} \sum_{j=i+1}^{t-1} D_{ij}^2$$

where D denotes dissimilarities or distances between every pair of points in $\mathbf{X}$, and SS_t and SS_{t-1} denote the accumulated sum of squared interpoint dissimilarities among all observations until time t , divided by t , and until time $(t-1)$, divided by $(t-1)$,

respectively, based on the findings from step 1. They demonstrated that the criterion of interest (Equation 4.5) is simply derived as:

$$d_t = \sqrt{\frac{t}{(t-1)}(SS_t - SS_{t-1})} \quad (4.6)$$

4.4.2 An exponentially weighted moving average structure based on interpoints dissimilarities

Utilizing d_t (as in Equation 4.6), we construct Exponentially Weighted Moving Average (EWMA) statistics [59] for monitoring shifts in the process location parameter:

$$W_t = \lambda d_t^z + (1 - \lambda)W_{t-1}, \quad (4.7)$$

where $d_t^z = \frac{d_t - \bar{d}_{t-1}}{S_{d_{t-1}}}$ represents the standardized version of d_t . Here, $\bar{d}_{t-1}$ and $S_{d_{t-1}}$ denote the mean and standard deviation of all preceding d_t values. The parameter λ denotes the smoothing factor, with $0 < \lambda \leq 1$. To identify small shifts, smaller λ values are preferred. An EWMA chart with $\lambda = 1.0$ is equivalent to a memoryless control chart. The initial value W_0 is set to $W_0 = 0$ when $i = 1$. The expected value and variance of W_i are $E(W_i) = 0$ and $Var(W_i) = \left\{ \frac{\lambda}{2-\lambda} (1 - (1-\lambda)^{2i}) \right\}$, respectively. The varying LCL and UCL of W_i are given by:

$$LCL_i = -L\sqrt{Var(W_i)} \quad (4.8)$$

and

$$UCL_i = L\sqrt{Var(W_i)}, \quad (4.9)$$

respectively. These parameters are selected to achieve the desired ICARL performance for the chart. The term $(1 - (1 - \lambda)^{2i})$ in the LCL and UCL approaches unity as i increases. Utilizing time-varying limits enhances the robustness of the chart in detecting initial out-of-control conditions, thus, we adopt them in our study.

4.4.3 Implementation of the proposed scheme

The proposed chart can be implemented step-by-step as follows:

Step 1: calculations of distance between pairs of points

Given a $n \times p$ dimensional process sample, where $n = n_1 + n_2$, n_1 denotes the number of samples used to obtain the charts limits and n_2 is the number of samples used for monitoring, the proposed method begin by calculating the distance (denoted by D_{ij}) between every pair of points. Different distance measures including Euclidean distance, Manhattan distance, Bray-Curtis distance have been proposed in the literature, the choice of distance measure depends on the forms of the process data.

Step 2: sum of squares of distances

utilizing the outcome from step 1, compute the distance squared matrix (denoted by D_{ij}^2) in which the following axioms must be satisfied; (i) all the main-diagonal entries are zero i.e $D_{ij}^2 = 0$, if $i = j$ (ii) all the off-diagonal entries are positive i.e $D_{ij}^2 > 0$, if

$i \neq j$ (iii) the matrix is symmetric. Once the distance squared matrix satisfies these axioms, the lower diagonal entries of the distance squared matrix are used to compute SS_t and SS_{t-1} which are the sum of inter-point distances among all point up to time t divided by t and time $(t - 1)$ divided by $(t - 1)$, respectively. From the results of SS_t and SS_{t-1} , the statistic d_t , is computed, as given in equation (6).

Step 3: calculation of EWMA statistics

Use the desired quantity λ smoothing parameter of the EWMA chart and the standardized version of statistic d_t computed in step 2, calculate the EWMA statistic W_t , using equation (7) for each sample n_1 with $W_0 = 0$ as the initial value.

Step 4: Calculation of parameter L and control limit

Use the first n values of W_t to obtain the chart control varying limits i.e LCL_i (equation 4.8) and UCL_i (equation 4.9). The value L of the LCL_i and UCL_i are obtained using a binary search algorithm. This algorithm operates on principle of dividing the data in the list to half on every iteration until the the required value is sited in the search result. Suppose we defined $d_{t[0:n-1]}$ as a sequence order or an array of available in-control sample of d_t and L be the target value. Let $d_{t[L]}$, $d_{t[U]}$, $d_{t[(0+n-1)/2]}$ and $d_{t[0]}$ denote the lower-bound (left side of the array), upper-bound (right side of the array), mid-value and initial value of d_t , respectively. To search the target value from $d_{t[0:n-1]}$, the process is initiates by locating the middle element of the array $d_{t[0:n-1]}$. Thereafter, the target value is compare with the middle element $d_{t[(0+n-1)/2]}$. If the target value is smaller than $d_{t[(0+n-1)/2]}$ then the next searching will be between $d_{t[(0+n-1)/2]}$ and

$d_{t[U]}$, otherwise it will be between $d_{t[L]}$ and $d_{t[(0+n-1)/2]}$ for comparison and matching. This process is continue on the sub-array until the target value is match or the array reduce to zero. Once the target value L is matched, the varying control limit can be determined from the in-control data using equation (8) and (9).

Step 5: Calculation of EWMA statistics for monitoring remaining n_2 samples

Repeat step 1 to 3 and calculate the EWMA statistics W_t for the remaining n_2 samples. Any of the EWMA statistics W_t that exceeds the control limit (i.e $W_t > UCL$ or $W_t < UCL$), indicate out-of-control point. Investigate the cause(s) for individual out-of-control point and take the necessary corrective action, otherwise, the process is said to be in-control and the monitoring continues.

4.5 Simulation Study

We assess the effectiveness of the proposed chart in identifying shifts in the location parameters for datasets generated from: (i) two multivariate discrete distributions (MDD), (ii) two multivariate continuous distributions (MCD), and (iii) two mixtures of discrete and continuous distributions (MDCD). These distributions were employed to assess the chart's performance across various scenarios. For each case, we vary the dimensionality $p = 2, 3, 4, 5, 10$ or 20 . Initially, we investigate the chart's performance under a Gaussian copula model with an independent covariance structure and Euclidean distance matrix. We introduce shifts (δ) to the (in-control) location param-

eters after collecting 200 (in-control) samples. We explore the impact of covariance structure, distance measures, and the number of in-control samples before applying shifts on the chart's performance.

4.5.1 Methods for simulating data

We create simulated datasets for Multivariate Data Distribution (MDD) from two distinct distributions: (i) the Multivariate Poisson distribution (MPOIS), and (ii) the Multivariate Negative Binomial distribution (MNBINORM). These datasets are generated utilizing a Gaussian Copula model, with Poisson and Negative Binomial marginal distributions, respectively. The marginal parameters for MPOIS and MNBINORM are established as $\mu \times \mathbf{1}_p$, where $\mathbf{1}_p$ represents a column vector of ones, with $\mu = 5$.

Furthermore, we explore two Multivariate Continuous Data Distributions (MCDD): (i) the Mixture of Multivariate Poisson and Normal distributions (MPOIS-NORM), and (ii) the Mixture of Multivariate Gamma and Normal distributions (MGAMMANORM). To generate multivariate mixture datasets with dimensions $p = \{2, 3, 4, 5, 10, \text{ or } 20\}$, we initially simulate data of dimension $p_1 = \{1, 1, 2, 2, 5, \text{ or } 10\}$, respectively. These datasets are generated from (i) MPOIS with marginal parameters $\mu \times \mathbf{1}_{p_1}$, or (ii) MGAMMA with marginal parameters $\mu \times \mathbf{1}_{p_1}$. Subsequently, the remaining dimension of the data $p_2 = p - p_1$ are simulated from MNORM with marginal parameters $\mu \times \mathbf{1}_{p_2}$. The columns of the simulated MPOIS dataset of dimension (or MGAMMA) p_1 and the simulated MNORM dataset with dimension p_2 are then put together to produce the multivariate mixture dataset with dimension p .

The proposed chart is applied to a p -dimensional process observation of size n . Initially, a random sample of size n_1 is drawn from an in-control normally distributed multivariate process. The criterion of interest d_t and its standardized version for the n_1 sample are computed to design the EWMA statistic (as given in Equation (7)) and the control limits (as given in Equation (8)) of the chart. Additionally, with location parameters set to different shifts, we generate a random vector comprising n_2 observations from several multivariate distributions (previously discussed). The efficacy of the charts is assessed by considering various covariance structures, distance measures, and the number of in-control samples.

4.5.2 Summary of simulation results

The chart's performance is assessed through an analysis of different run length metrics, including ARL and SDRL. The in-control ARL (ARL_0) reflects the average run length when the process is under control, indicating no shift in the process mean. Conversely, the out-of-control ARL (ARL_1) represents the average run length when the process is out of control, indicating absence of shift in process mean.

We set $ARL_0 = 200$ for the in-control process and compare the chart's performance in various simulated conditions. Next, we evaluate ARL_1 values for different out-of-control situations. We employ a binary search approach, as described in step 3 of the implementation, to ascertain the control limits that yield the desired ARL_0 . A chart with a smaller ARL_1 value is considered more efficient in detecting mean shifts compared to others.

Based on the average of 50,000 run lengths, we provide the results from the Gaus-

sian copula model with an independent covariance structure and Euclidean distance matrix when shift sizes are applied to the (in-control) location parameter after 200 (in-control) samples from $p = 2, 3, 4, 5, 10$, or 20 simulated from each of the six distributions in Table 4.1. Figure 4.1 present comparisons of different distributions and dimensions of the sample process. Additionally, Figure 4.3 illustrates the run length properties of the chart from the MNORM distribution with auto-regress (AR), equi-covariance (EQ), independent (ID), and non-sparse (NSP) covariance structures when $p = 10$.

The auto-regressive structure is defined as:

$$\sigma_{ij} = \begin{cases} \frac{1}{(1-\theta^2)} & \text{if } i = j \\ \frac{\theta^{|j-i|}}{1-\theta^2} & \text{if } i \neq j \end{cases}$$

where $\theta \in [-1, 1]$ is an auto-regressive parameter from the AR(1) model, and $i, j = 1, \dots, p$. We use $\theta = 0.3$.

Furthermore, we examine EQ covariance structure as $\Sigma_{ij} = 1$, if $i = j$, else ϵ where $i, j = 1, \dots, p$ and $\epsilon \in [0, 1]$. We use $\epsilon = 0.6$. The ID covariance structure is represented as $\Sigma_{ij} = \sigma_{i,j}^2$, if $i = j$, else 0 where $i, j = 1, \dots, p$.

The NSP covariance structure considered has the form:

$$\sigma_{ij} = \begin{cases} 1 & \text{if } i = j \\ \frac{\beta-1}{\beta\sqrt{p}} & \text{if } i \neq j \end{cases}$$

where β varies with p for $p = \{2, 3, 4, 5, 10 \text{ or } 20\}$, and $i, j = 1, \dots, p$.

Lastly, the sparse covariance structure is defined as: $\sigma_{i,i+1} = \sigma_{i,i-1} = 0.4, \sigma_{ii} = \phi$, and zero elsewhere.

Tables 4.2 and 4.3 present the results under Euclidean, Manhattan and Bray-Curtis distance measures for $p = 2$ or 10, examining the effects of the distance measures on the chart's performance.

Finally, Tables 4.7 and 4.8 provides ARL_1 values under the independent covariance structure and Euclidean distance matrix for different numbers of in-control samples, with sizes $n = 100, 150, 200$, and 300, to explore the effect of the number of in-control samples on the chart's performance.

Table 4.1: *ARL* profile values when multivariate dataset simulated from MNORM and MGAMMA distribution for $P = 2$.

Distribution	Shift	Euclidean		Manhattan		Bray-Curtis	
		ARL	SDRL	ARL	SDRL	ARL	SDRL
MNORM	0	207.3	229.4	205.3	227	201.9	230.4
	0.4	184.1	229.7	177.2	205	261.5	296.7
	0.6	146.5	201.2	142.2	196	265.2	280.2
	0.8	95.9	158	95.9	159	272	269.8
	1	51.1	103	47.8	96.5	232	240.5
	1.2	22.9	40.8	21.9	35.1	169.7	214
	1.4	13.3	17.6	12.7	16.7	94.9	158.9
	1.8	6.7	3.9	6.6	3.9	16.3	36.5
	2	5.4	2.8	5.3	2.8	9.7	7.9
	2.5	3.6	1.4	3.5	1.4	5.5	2.3
	3	2.8	1	2.6	0.9	3.9	1.3
	4	1.9	0.6	1.8	0.6	2.6	0.7
	5	1.5	0.5	1.4	0.5	2.1	0.3
MGAMMA	0	206.7	250	245	255	199.1	214.9
	0.4	155.9	207.1	191	202	209.8	226.2
	0.6	128.8	164.5	170	164	212	218.3
	0.8	106.3	137.3	145	135.2	218.6	238
	1	92.1	122.8	121	104	210.9	229.3
	1.2	70.9	98.3	97.6	39	209	216.3
	1.4	58.5	79.3	72.2	30	207.6	208.8
	1.8	35.4	49.6	47.3	20	199.7	190
	2	28.3	39.4	30.5	17	192	183.4
	2.5	17	17	15.8	11	165	159.6
	3	11.2	9.6	8.6	8	124.4	141.4
	4	6.5	4.3	3.8	5	38	73.3
	5	4.3	2.4	2.2	4	10.9	11.4

Table 4.2: *ARL* profile values when multivariate dataset simulated from MPOIS and MNB distribution for $P = 2$.

Distribution	Shift	Euclidean		Manhattan		Bray-Curtis	
		ARL	SDRL	ARL	SDRL	ARL	SDRL
MPOIS	0	201.2	221.5	201.1	221	201.3	221.7
	0.4	153.5	174.5	157.4	187	187.4	209.2
	0.6	132.1	163	127.9	149	180	204
	0.8	108.2	139.6	103.1	127	164.9	185.8
	1	88.9	117.1	83.1	111	159.1	173.3
	1.2	68.4	89.4	68.2	92	148.8	160.4
	1.4	51.5	67.5	51.6	67.8	134.4	137.8
	1.8	31.1	41.6	29.4	35	124.4	122.3
	2	24.2	26.7	23.6	26.8	120.2	111.2
	2.5	14.8	14.2	13.7	11.9	116.4	98.8
	3	9.6	7.1	9.4	6.9	113.6	83.5
	4	5.9	3.6	5.6	3.3	120.8	68.2
	5	4.1	2.1	3.9	2	123.7	66.2
MNB	0	211.4	367.4	210.2	329	200.1	228.4
	0.4	130.7	185.3	132.1	188	193.7	217.3
	0.6	110.2	157.5	104.6	140	180.3	208.6
	0.8	93.2	123	87.6	111	194.5	210.9
	1	75.6	97.4	75.4	113	190.9	206.1
	1.2	66.8	87.4	60.7	76.2	190.3	213.2
	1.4	55.4	67.8	51.7	64.1	187.4	211.9
	1.8	41.5	49.2	38.1	46.5	183.2	206.4
	2	36.4	41.3	33.9	36.5	184.2	207.4
	2.5	27.7	32.7	26.2	26.9	172.9	196.2
	3	21.3	21.6	21	21	182.9	196.2
	4	14.7	13.7	14.1	12.5	182.9	208.1
	5	11.1	9.4	10.4	8.6	180.3	181.8

Table 4.3: *ARL* profile values when multivariate dataset simulated from MGAMMA and MNORM distribution for $P = 10$

Distribution	Shift	Euclidean		Manhattan		Bray-Curtis	
		ARL	SDRL	ARL	SDRL	ARL	SDRL
MGAMMA	0	203.7	226	200.4	216	203.2	220.8
	0.4	184.5	201.7	187.2	209	189.5	210.7
	0.6	171.2	197.8	171.9	196	181.6	206
	0.8	149.7	182.8	148.9	179	168.7	187.8
	1	132.9	161.6	129.2	157	160.1	175.3
	1.2	110.1	143.6	104.8	127	150.8	162.4
	1.4	87.6	112.6	84.7	105	136.5	139.9
	1.8	56.3	75	53.7	65.5	128.7	124.4
	2	43.5	53.5	43.6	54.7	122.2	113.2
	2.5	26.5	29.4	25.5	16.1	118.4	100.8
	3	16.6	14.4	16.1	14.1	117.9	86.6
	4	9.1	5.8	8.8	5.5	123.8	72.2
	5	6	3.2	5.8	3	124.7	71.2
MNORM	0	200.1	236.2	197.1	229	203.1	230.3
	0.4	164.3	210.8	172.6	200	195.9	219.5
	0.6	145.3	205.3	143.6	167	186.3	210.6
	0.8	124	160.8	128.6	145	195.5	222.9
	1	102.4	131.7	107.5	128	193.9	215.1
	1.2	92.3	118.3	92.8	110	193.2	221.2
	1.4	75.1	93.2	75.9	87.8	191.4	212.9
	1.8	56.8	68.1	52.4	56.1	188.4	206.7
	2	48.1	55.6	46	52.9	187.8	203.2
	2.5	36	37.4	33.8	34.6	188.2	208.6
	3	27	27.2	24.5	22.2	179.9	194.4
	4	17.7	15.8	15.9	12.5	182.9	210.6
	5	12.7	10.1	7.9	9	181.3	183.8

Table 4.4: *ARL* profile values when multivariate dataset simulated from MPOIS-NORM and MNORMGAMMA distribution for $P = 2$

Distribution	Shift	Euclidean		Manhattan		Bray-Curtis	
		ARL	SDRL	ARL	SDRL	ARL	SDRL
MPOISNORM	0	201.8	244.4	207.2	243	201.1	237.2
	0.4	162.8	203.5	163.3	191	237.3	270.9
	0.6	141	185	137	171	244.4	265
	0.8	111.4	147.7	105.7	135	259.7	284.6
	1	85.7	119.6	79.8	119	253.5	258.5
	1.2	64.4	94.2	55.3	84.9	257.4	271.5
	1.4	49.3	75.4	37.1	58.2	246.2	252
	1.8	25	37.5	17.5	25.2	191.9	212.5
	2	18.1	19.2	13.2	12.3	147.7	185.6
	2.5	10.4	7.8	8	5	58.2	112.4
	3	7.2	4.3	5.5	2.9	16.3	35
	4	4.2	1.9	3.4	1.5	6.5	2.8
	5	3	1.2	2.5	0.9	4.4	1.4
MNORMGAMMA	0	203.3	253.8	198.5	244	202.5	230.4
	0.4	156	206	155.2	194	223.4	237.4
	0.6	133.4	181.4	134.2	181	217.9	231.7
	0.8	117.4	166.9	170.2	152	219.9	230.1
	1	90.2	126.5	82.6	130	220	229
	1.2	69	101	62.3	102	211	220.4
	1.4	53.8	93.8	43.1	71	192.8	194.5
	1.8	29.1	45.4	20.4	27.1	138.3	165
	2	21.6	28.7	14.4	14.8	110.1	151
	2.5	11.8	10.1	8.5	5.8	38.1	79
	3	7.6	4.9	5.9	5.8	12.9	18.9
	4	4.5	2.2	3.5	1.5	5.8	2.6
	5	3.1	1.2	2.6	1	4	1.3

Table 4.5: *ARL* profile values when multivariate dataset simulated from MNORM and MGAMMA distribution for $P = 10$

Distribution	Shift	Euclidean		Manhattan		Bray-Curtis	
		ARL	SDRL	ARL	SDRL	ARL	SDRL
MNORM	0	200.1	222.4	195.7	213	193.1	205.9
	0.4	200.9	212.8	201.8	226	167.2	181.8
	0.6	194.8	216.5	197	215	149.7	149.9
	0.8	191.4	220.7	182.8	205	142.3	134.5
	1	170.6	209.3	164.8	208	132.3	111.3
	1.2	129.5	192	130.2	185	130.7	97.9
	1.4	83.4	141.3	82.9	149	134.9	90.8
	1.8	25.8	48.8	25.5	51.1	137.6	85.1
	2	16.5	27.2	15.4	22.5	130.2	85.4
	2.5	7.8	4.4	7.3	4	69.2	81
	3	5	2.2	4.9	2.1	16.6	29.1
	4	3	0.9	2.8	0.9	4.8	1.8
	5	2.1	0.6	2	0.6	3	0.8
MGAMMA	0	204.1	231.4	201.2	227	201.6	225.4
	0.4	182.3	213.3	179.8	198	177.7	201.1
	0.6	161.3	190.6	165.4	201	163.6	189.1
	0.8	150.9	181.6	149.6	178	147.7	166.3
	1	129.5	174.4	129	152	135.3	144.5
	1.2	113.9	151.2	110.5	152	124.9	129.3
	1.4	91.1	120.5	90.4	112	116	122.2
	1.8	65.3	87	58.7	76.6	101	94.6
	2	53.4	74.4	46.3	57.2	97.6	85.7
	2.5	33.2	44.1	28.8	34.3	95.2	78.2
	3	20.8	20.8	18.1	16.8	93.6	70.8
	4	11.1	8.2	9.7	6.3	97.8	63.9
	5	7.1	4.2	6.2	3.5	92.7	65

Table 4.6: *ARL* profile values when multivariate dataset simulated from MPOIS and MNB distribution for $P = 10$

Distribution	Shift	Euclidean		Manhattan		Bray-Curtis	
		ARL	SDRL	ARL	SDRL	ARL	SDRL
MPOIS	0	203.7	226	200.4	216	203.2	220.8
	0.4	184.5	201.7	187.2	209	189.5	210.7
	0.6	171.2	197.8	171.9	196	181.6	206
	0.8	149.7	182.8	148.9	179	168.7	187.8
	1	132.9	161.6	129.2	157	160.1	175.3
	1.2	110.1	143.6	104.8	127	150.8	162.4
	1.4	87.6	112.6	84.7	105	136.5	139.9
	1.8	56.3	75	53.7	65.5	128.7	124.4
	2	43.5	53.5	43.6	54.7	122.2	113.2
	2.5	26.5	29.4	25.5	16.1	118.4	100.8
	3	16.6	14.4	16.1	14.1	117.9	86.6
	4	9.1	5.8	8.8	5.5	123.8	72.2
	5	6	3.2	5.8	3	124.7	71.2
MNB	0	200.1	236.2	197.1	229	203.1	230.3
	0.4	164.3	210.8	172.6	200	195.9	219.5
	0.6	145.3	205.3	143.6	167	186.3	210.6
	0.8	124	160.8	128.6	145	195.5	222.9
	1	102.4	131.7	107.5	128	193.9	215.1
	1.2	92.3	118.3	92.8	110	193.2	221.2
	1.4	75.1	93.2	75.9	87.8	191.4	212.9
	1.8	56.8	68.1	52.4	56.1	188.4	206.7
	2	48.1	55.6	46	52.9	187.8	203.2
	2.5	36	37.4	33.8	34.6	188.2	208.6
	3	27	27.2	24.5	22.2	179.9	194.4
	4	17.7	15.8	15.9	12.5	182.9	210.6
	5	12.7	10.1	7.9	9	181.3	183.8

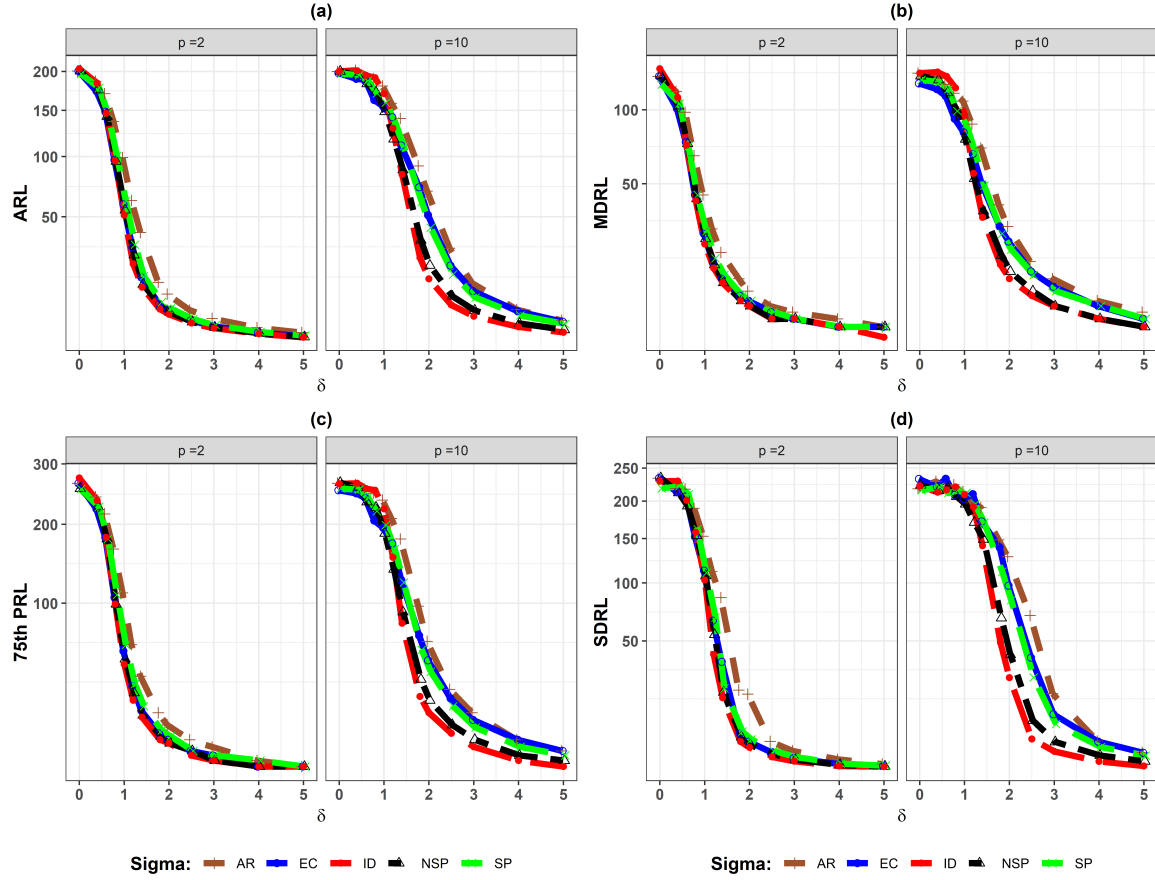


Figure 4.1: Results show (a) average run length (ARL), (b) median run length (MDRL), (c) 75th percentile of run length (75th PRL), and (d) standard deviation of run length of the charts from MNORM distribution with auto-regress (AR), equi-covariance (EQ), independent (ID) and non-sparse (NSP) covariance structures $p = 2$ or 10

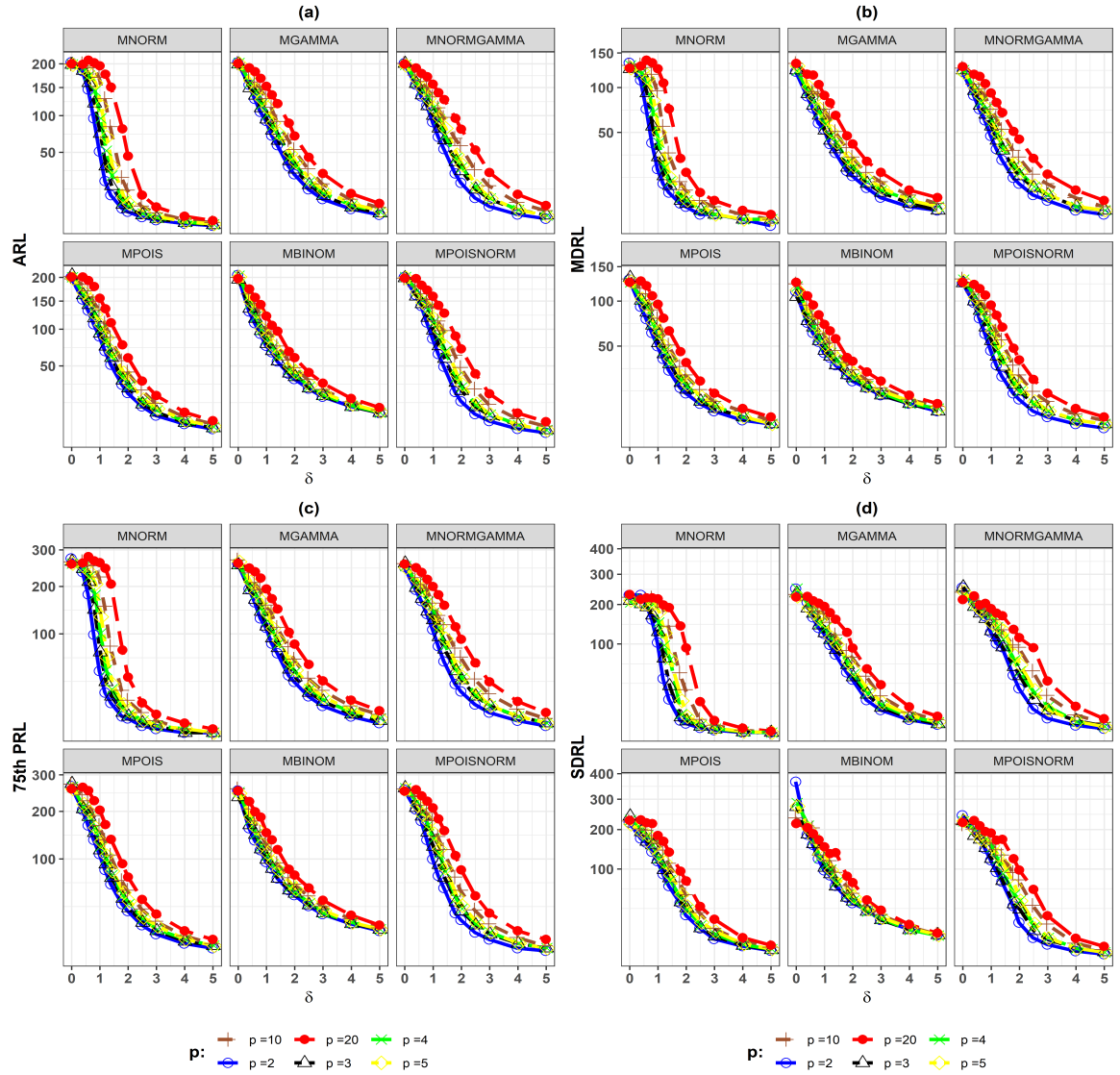


Figure 4.2: Results show (a) average run length (ARL), (b) median run length (MDRL), (c) 75th percentile of run length (75th PRL), and (d) standard deviation of run length of the charts from MGAMMA, MNORM, MPOIS, MNBINORM, MNORMGAMMA, or MPOISNORM distribution when $p = 2, 3, 4, 5, 10$ and 20 .

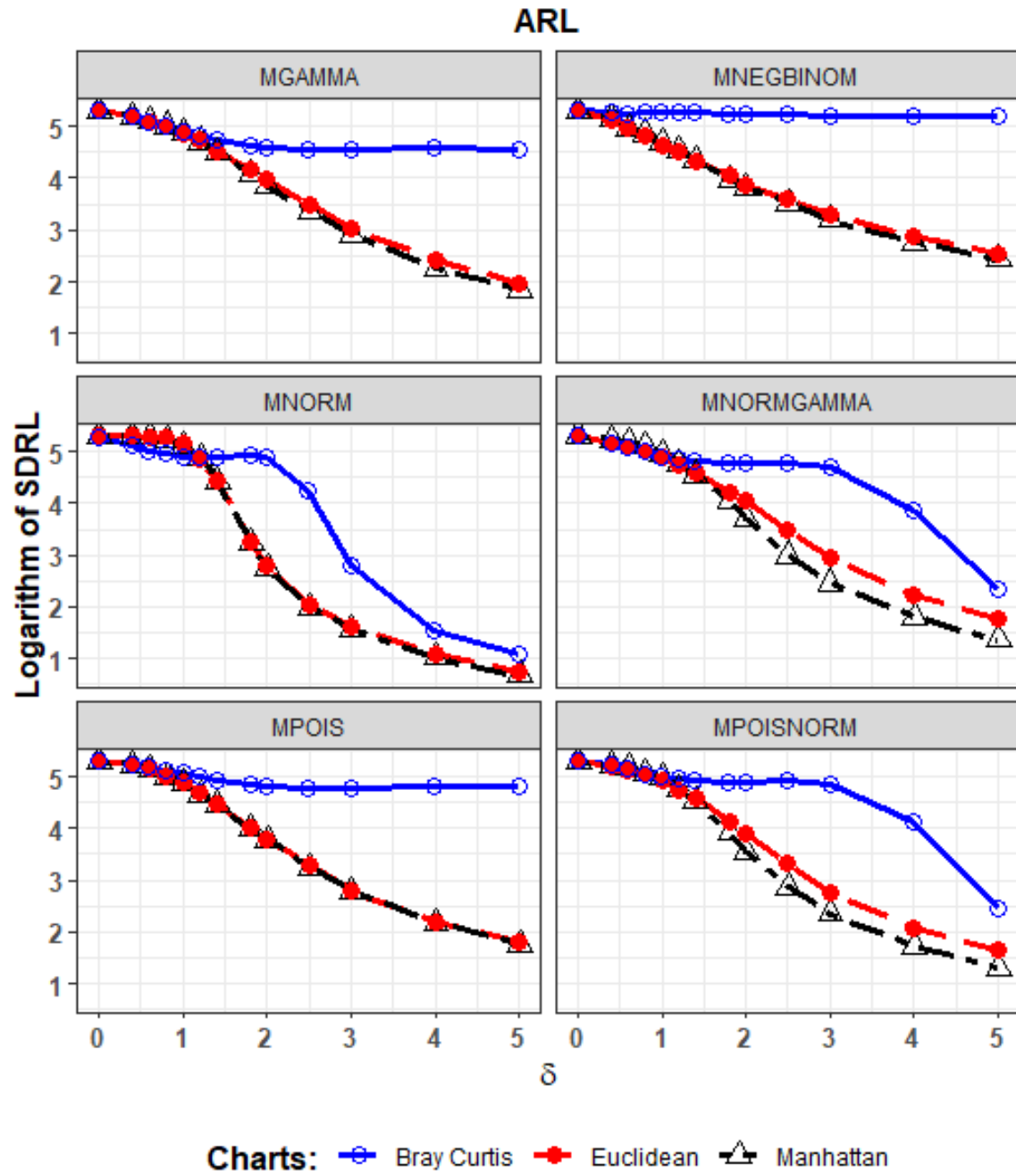


Figure 4.3: Results show average run length (ARL) based on Euclidean, Manhattan and Bray-Curtis distance measures under independent (ID) covariance structure for $p = 10$

Table 4.7: ARL values of the chart under ID covariance structure and Euclidean distance when multivariate of 10-dimensional (P=10) dataset simulated from MGAMMA, MNORM, MPOIS, MNB, MNORMGAMMA, and MPOISNORM distributions

Distribution	IC Sample	δ					
		0	0.4	0.6	0.8	1	1.2
MGAMMA	100	199.71	201.72	188.2	174.38	168.83	147.87
	150	199.26	185.33	171.6	163.12	142.45	123.69
	200	204.12	182.32	161.3	150.91	129.5	113.94
	300	200.61	172.81	153.3	131	113.78	96.55
MNORM	100	202.91	211.85	205.3	205.56	199.02	179.77
	150	204.26	204.96	206.1	201.66	183.47	153.51
	200	200.06	200.88	194.8	191.43	170.61	129.53
	300	197.72	199.17	188.4	179.03	142.94	96.06
MPOIS	100	197.81	201.54	198.3	186.31	169.67	152.71
	150	205.11	189.55	173.9	162.16	145.04	118.2
	200	203.74	184.54	171.2	149.71	132.92	110.13
	300	200.42	177.24	156	136.05	115.04	90.2
MNB	100	210.35	189.45	171.6	150.19	129.65	117.73
	150	199.71	170.29	154.6	134.2	114.98	100.93
	200	200.08	164.35	145.3	124.04	102.43	92.32
	300	202.32	156.62	133.6	112.69	95.63	81.67
MNORMGAMMA	100	196.31	191.66	178.7	172.6	160.38	151.39
	150	196.7	176.93	171.6	159.7	145.52	128.93
	200	202.13	175.66	162.4	151.35	132.55	113.28
	300	202.51	167.67	154.8	138.53	119.73	104.71

Distribution	IC Sample	δ						
		1.4	1.8	2	2.5	3	4	5
MGAMMA	100	129.53	98.84	84.72	54.06	32.13	14.33	8.02
	150	107.94	73.41	63.13	38.77	23.91	11.82	7.37
	200	91.09	65.27	53.38	33.16	20.78	11.11	7.13
	300	81.27	53.07	44.42	27.56	18.64	10.48	6.89
MNORM	100	149.25	75.74	39.4	10.3	5.59	3.1	2.2
	150	108.24	40.79	20.95	8.19	5.19	3.01	2.14
	200	83.38	25.83	16.46	7.76	5	2.98	2.12
	300	58.26	19.05	13.3	7.36	4.79	2.88	2.08
MPOIS	100	130.76	92.17	73.7	41.99	24.49	10.88	6.59
	150	101.41	65.25	52.71	30.08	18.69	9.44	6.19
	200	87.6	56.32	43.47	26.52	16.64	9.07	6.01
	300	73.22	46.43	36.93	22.86	15.39	8.71	5.76
MNB	100	100.33	76.42	65.64	46.59	33.76	20.55	14.25
	150	83.32	63.47	54.82	38.38	29.18	18.48	13.31
	200	75.11	56.75	48.09	35.99	27.02	17.72	12.68
	300	66.62	53.66	45.73	33.4	25.69	17.14	12.13
MNORMGAMMA	100	132.52	107.75	92.06	59.9	32.53	12.02	6.52
	150	118.84	82.39	68.78	39.81	22.14	9.97	6.14
	200	97.37	67.28	58.01	32.39	19.37	9.39	5.89
	300	83.37	55.87	45.83	27.75	17.28	8.95	5.7

CHAPTER 5

CONCLUSION AND FUTURE WORK

The general motivation of this thesis is to employ machine learning-based imputation methods to improve the performance of multivariate control chart in phase I analysis in the presence of missing values and to develop a novel self-starting control chart aim overcoming the limitation of traditional control charts.

In chapter 3, we investigate the effects of different methods for handling missing values in Phase I on the performance of Hotelling T^2 control chart, focusing on the in-control average run length (ICARL). Six imputation methods were considered: kNN, wNN, wNNSel, MICE, missForest (machine learning-based approaches) and Mean (a traditional method). Through a simulation study involving various parameters such as the number of quality characteristics, correlation between them, percentage of missing values, and a sample size of 200, we assess the effectiveness of these methods. After conducting a thorough analysis, we recommend employing an imputation-based

method to create a complete phase I dataset. Our evaluation focused on ICARL performance under different scenarios. When the number of quality characteristics is 5 or 10, the MICE-based imputation method consistently outperforms others, regardless of levels of correlation between them and percentage of missing values. However, as the number of quality characteristics increases to 30, the wkNN and kNN method demonstrates better performance when the correlation was 0 and 0.7, respectively. Also, the application of competing methods is demonstrated using real-world data from a wholesale distributor. In conclusion, our study highlights the superiority of both MICE, wkNN and kNN in enhancing the overall quality and reliability of control chart implementation in phase I datasets with missing values. The MICE method is recommended when the number of quality characteristics is 15 or less, while wkNN and kNN is preferred for larger number of quality characteristics, offering practitioners robust solutions for effectively addressing missing values.

In chapter 4, we devised a multivariate self-starting exponentially weighted moving average chart based on interpoints dissimilarities. We assess the performance of the control chart across various scenarios. We fixed $ARL_0 = 200$ during in-control processes and evaluated ARL_1 for different out-of-control situations. A smaller ARL_1 indicates higher efficiency in detecting mean shifts. Our analysis, based on an average of 50,000 run lengths, employed the Gaussian Copula model with independent covariance structure and Euclidean distance matrix. We introduced shifts to the location parameter after 200 samples from datasets with dimensions ranging from $P = 2, 3, 4, 5, 10$ or 20, sampled from six distributions. For comparison, we utilized Euclidean, Manhattan and Bray-Curtis distance measures, particularly for $P = 2$ or 10, to assess

their impact on chart performance. The results showed that the chart under Euclidean and Manhattan distance are always almost similar and more efficient than the chart under Bray-Curtis distance across the six distributions. In addition, the increase in the dimensions from 2 to 10 change the scenario as shown in figure 4.3 such that the chart under the three distance measures have almost similar ability in detecting small shifts in location process across all the distribution except the chart under Bray-Curtis that outperformed others under MNORM. However, for large shifts size, it is notices that the chart under Bray-Curtis distance performed worst while Manhattan distance gives best performance from MNORMGAMMA and MPOISNORM.

Future research entails investigating the out-of-control average run length (OOCARL) performance of the Hotelling T^2 control chart while comparing various strategies for handling missing values in Phase I analysis. Furthermore, exploring the impact of missing values on the MEWMA and MCUSUM) control charts using machine learning-based imputation methods is strongly recommended. Additionally, extending the proposed self-starting control chart to include other multivariate and discrete distributions through the use of alternative distance measures is highly encouraged.

REFERENCES

- [1] J. R. Dever, “The applicability of statistical process control to systems involving people processes and business rhythms,” Ph.D. dissertation, The George Washington University, 2009.
- [2] A. P. Sage, *System of systems engineering: innovations for the 21st century*. John Wiley & Sons, 2011.
- [3] D. C. Montgomery, *Statistical quality control*. Wiley New York, 2009, vol. 7.
- [4] H. Hotelling, “Multivariate quality control-illustrated by the air testing of sample bombsights,” *Techniques of statistical analysis*, 1947.
- [5] G. C. Runger and M. C. Testik, “Multivariate extensions to cumulative sum control charts,” *Quality and Reliability Engineering International*, vol. 20, no. 6, pp. 587–606, 2004.
- [6] C. A. Lowry, W. H. Woodall, C. W. Champ, and S. E. Rigdon, “A multivariate exponentially weighted moving average control chart,” *Technometrics*, vol. 34, no. 1, pp. 46–53, 1992.

- [7] A. G. Di Nuovo, “Missing data analysis with fuzzy c-means: A study of its application in a psychological scenario,” *Expert Systems with Applications*, vol. 38, no. 6, pp. 6793–6797, 2011.
- [8] M. A. Mahmoud, N. A. Saleh, and D. F. Madbully, “Phase i analysis of individual observations with missing data,” *Quality and Reliability Engineering International*, vol. 30, no. 4, pp. 559–569, 2014.
- [9] H. Furutani, K. Yamamoto, H. Ogura, and Y. Kitazoe, “New control chart for multivariate data with missing values,” *Computers and biomedical research*, vol. 21, no. 1, pp. 1–8, 1988.
- [10] J. C. García-Díaz, “The ‘effective variance’ control chart for monitoring the dispersion process with missing data,” *European Journal of Industrial Engineering*, vol. 1, no. 1, pp. 40–55, 2007.
- [11] S. R. Wilson, “Control charts with missing observations,” Ph.D. dissertation, Virginia Tech, 2009.
- [12] N. D. Tracy, J. C. Young, and R. L. Mason, “Multivariate control charts for individual observations,” *Journal of quality technology*, vol. 24, no. 2, pp. 88–95, 1992.
- [13] A. J. Hayter and K.-L. Tsui, “Identification and quantification in multivariate quality control problems,” *Journal of quality technology*, vol. 26, no. 3, pp. 197–208, 1994.

- [14] J. H. Sullivan and W. H. Woodall, “Change-point detection of mean vector or covariance matrix shifts using multivariate individual observations,” *IIE transactions*, vol. 32, no. 6, pp. 537–549, 2000.
- [15] M. J. Mihalcin, T. A. Mazzuchi, S. Sarkani, and J. R. Dever, “Quality control—an approach applying multivariate control charts during the operation of systems involving human processes,” *Systems Engineering*, vol. 17, no. 2, pp. 204–212, 2014.
- [16] S. Kotz and N. L. Johnson, *Process capability indices*. CRC press, 1993.
- [17] J. O. Ajadi, Z. Wang, and I. M. Zwetsloot, “A review of dispersion control charts for multivariate individual observations,” *Quality Engineering*, vol. 33, no. 1, pp. 60–75, 2021.
- [18] P. Lewicki, T. Hill, and C. Qazaz, “Multivariate quality control,” *Quality*, vol. 46, no. 4, p. 38, 2007.
- [19] B. Murphy, “Selecting out of control variables with the t^2 multivariate quality control procedure,” *Journal of the Royal Statistical Society Series D: The Statistician*, vol. 36, no. 5, pp. 571–581, 1987.
- [20] H. Scheffe, “Selected techniques of statistical analysis: For scientific and industrial research and production and management engineering.” 1948.
- [21] M. Rogalewicz, “Some notes on multivariate statistical process control,” *Management and Production Engineering Review*, 2012.

- [22] M. Waterhouse, I. Smith, H. Assareh, and K. Mengersen, "Implementation of multivariate control charts in a clinical setting," *International Journal for Quality in Health Care*, vol. 22, no. 5, pp. 408–414, 2010.
- [23] D. F. Madbuly, P. E. Maravelakis, and M. A. Mahmoud, "The effect of methods for handling missing values on the performance of the mewma control chart," *Communications in Statistics-Simulation and Computation*, vol. 42, no. 6, pp. 1437–1454, 2013.
- [24] J. H. Sullivan and L. A. Jones, "A self-starting control chart for multivariate individual observations," *Technometrics*, vol. 44, no. 1, pp. 24–33, 2002.
- [25] D. M. Hawkins, P. Qiu, and C. W. Kang, "The changepoint model for statistical process control," *Journal of quality technology*, vol. 35, no. 4, pp. 355–366, 2003.
- [26] D. M. Hawkins and K. Zamba, "A change-point model for a shift in variance," *Journal of Quality Technology*, vol. 37, no. 1, pp. 21–31, 2005.
- [27] C. Zou, C. Zhou, Z. Wang, and F. Tsung, "A self-starting control chart for linear profiles," *Journal of Quality Technology*, vol. 39, no. 4, pp. 364–375, 2007.
- [28] Z. Li, J. Zhang, and Z. Wang, "Self-starting control chart for simultaneously monitoring process mean and variance," *International Journal of Production Research*, vol. 48, no. 15, pp. 4537–4553, 2010.
- [29] J. Schaffer, "A multivariate application of the q chart." 1998.

- [30] W. A. Jensen, L. A. Jones-Farmer, C. W. Champ, and W. H. Woodall, “Effects of Parameter Estimation on Control Chart Properties: a Literature Review,” *Journal of Quality Technology*, vol. 38, no. 4, pp. 349–364, 2006.
- [31] D. P. Bischak and D. Trietsch, “The rate of false signals in $\bar{u}$ control charts with estimated limits,” *Journal of Quality Technology*, vol. 39, no. 1, pp. 54–65, 2007.
- [32] D. M. Hawkins, “Self-starting cusum charts for location and scale,” *Journal of the Royal Statistical Society Series D: The Statistician*, vol. 36, no. 4, pp. 299–316, 1987.
- [33] D. M. Hawkins and Q. Wu, “The CUSUM and the EWMA Head-to-Head,” *Quality Engineering*, vol. 26, no. 2, pp. 215–222, 2014.
- [34] C. P. Quesenberry, “On properties of q charts for variables,” *Journal of Quality technology*, vol. 27, no. 3, pp. 184–203, 1995.
- [35] Z. Li and Z. Wang, “Adaptive cusum of the q chart,” *International Journal of Production Research*, vol. 48, no. 5, pp. 1287–1301, 2010.
- [36] D. M. Hawkins and K. Zamba, “Statistical process control for shifts in mean or variance using a changepoint formulation,” *Technometrics*, vol. 47, no. 2, pp. 164–173, 2005.
- [37] G. Tuerhong and S. Bum Kim, “Gower distance-based multivariate control charts for a mixture of continuous and categorical variables,” *Expert Systems With Applications*, vol. 41, pp. 1701–1707, 2014.

- [38] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple imputation by chained equations: what is it and how does it work?” *International journal of methods in psychiatric research*, vol. 20, no. 1, pp. 40–49, 2011.
- [39] H. Hegde, N. Shimpi, A. Panny, I. Glurich, P. Christie, and A. Acharya, “Mice vs ppca: Missing data imputation in healthcare,” *Informatics in Medicine Unlocked*, vol. 17, p. 100275, 2019.
- [40] J. Maillo, S. Ramírez, I. Triguero, and F. Herrera, “knn-is: An iterative spark-based design of the k-nearest neighbors classifier for big data,” *Knowledge-Based Systems*, vol. 117, pp. 3–15, 2017.
- [41] A. Amirteimoori and S. Kordrostami, “A euclidean distance-based measure of efficiency in data envelopment analysis,” *Optimization*, vol. 59, no. 7, pp. 985–996, 2010.
- [42] X. Wu, H. Akbarzadeh Khorshidi, U. Aickelin, Z. Edib, and M. Peate, “Imputation techniques on missing values in breast cancer treatment and fertility data,” *Health Information Science and Systems*, vol. 7, pp. 1–8, 2019.
- [43] G. Tutz and S. Ramzan, “Improved methods for the imputation of missing data by nearest neighbor methods,” *Computational Statistics & Data Analysis*, vol. 90, pp. 84–99, 2015.
- [44] S. Faisal, “wnnsel: Weighted nearest neighbor imputation of missing values using selected variables,” 2017.
- [45] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.

- [46] K. M. Ramachandran and C. P. Tsokos, *Mathematical statistics with applications in R*. Academic Press, 2020.
- [47] K. Zamba and D. M. Hawkins, “A multivariate change-point model for statistical process control,” *Technometrics*, vol. 48, no. 4, pp. 539–549, 2006.
- [48] —, “A multivariate change-point model for change in mean vector and/or covariance structure,” *Journal of Quality Technology*, vol. 41, no. 3, pp. 285–303, 2009.
- [49] Y. Li, Y. Liu, C. Zou, and W. Jiang, “A self-starting control chart for high-dimensional short-run processes,” *International Journal of Production Research*, vol. 52, no. 2, pp. 445–461, 2014.
- [50] C. Zou, Y. Zhang, and Z. Wang, “A control chart based on a change-point model for monitoring linear profiles,” *IIE transactions*, vol. 38, no. 12, pp. 1093–1103, 2006.
- [51] M. A. Mahmoud, P. A. Parker, W. H. Woodall, and D. M. Hawkins, “A change point method for linear profile data,” *Quality and Reliability Engineering International*, vol. 23, no. 2, pp. 247–268, 2007.
- [52] C. Zou, F. Tsung, and Y. Liu, “A change point approach for phase i analysis in multistage processes,” *Technometrics*, vol. 50, no. 3, pp. 344–356, 2008.
- [53] A. Amiri and A. Allahyari, “Change Point Estimation Methods for Control Chart Postsignal Diagnostics: A Literature Overview,” *Quality and Reliability Engineering International*, vol. 28, no. 7, pp. 673–685, 2012.

- [54] C. P. Quesenberry, “Spc q charts for start-up processes and short or long runs,” *Journal of quality technology*, vol. 23, no. 3, pp. 213–224, 1991.
- [55] ———, *SPC methods for quality improvement*, 1997.
- [56] M. J. Anderson and A. A. Thompson, “Multivariate control charts for ecological and environmental monitoring,” *Ecological Applications*, vol. 14, no. 6, pp. 1921–1935, 2004.
- [57] B. H. McArdle and M. J. Anderson, “Fitting multivariate models to community data: a comment on distance-based redundancy analysis,” *Ecology*, vol. 82, no. 1, pp. 290–297, 2001.
- [58] M. J. Anderson, “A new method for non-parametric multivariate analysis of variance,” *Austral ecology*, vol. 26, no. 1, pp. 32–46, 2001.
- [59] S. W. Roberts, “Control Chart Tests Based on Geometric Moving Averages,” *Technometrics*, vol. 1, no. 3, pp. 239–250, 1959.

VITAE

CONTACT INFORMATION

Address: Brikama Jamisa, West Coast Region

Cell phone: (+220) 3203232 / 9564777

Email: siakajawara00@gmail.com

Address: MDI Road, Kanifing, P.O Box 35350, Serrekunda, The Gambia.

EDUCATION BACKGROUND

King Fahd University of Petroleum and Minerals

Aug 2022 - Present

- MSc. Applied Statistics
-

University of The Gambia

Jul 2015 - Aug 2019

- BSc. Economics (major) and Mathematics (minor).

West African Senior School Certificate Examination (WASSCE) Sept 2012 - May 2015

- John Pickering Comprehensive Senior Secondary School, Lamin.
- 9 credits including English, Mathematics, Economics and General Science.
- Highest Grades in John Pickering Comprehensive Senior Secondary School.
- Awarded as the overall best outstanding student in John Pickering Comprehensive Senior Secondary School, 2015.

Institute of Professional Financial Managers, Brikama Sept 2014 - Dec 2014

- Certificate in Accounting and Finance.

West African Examination Council (WAEC) Sept 2009 - Jun 2012

- Aggregate: 16
 - Standard Aggregate in Kerewan Upper Basic School.
-

RELEVANT EXPERIENCE

Graduate Teaching Assistant

King Fahd University of Petroleum and Minerals

Aug 2022 - Present

- Meeting with students during office hours.
- Grading, moderating, and returning scripts in a timely manner.
- Uploading lecture slides, assigning reading materials, and solutions to quizzes and exams to the virtual learning sites of the university.
- Proctor examinations.

Planner

Ministry of Fisheries and Water Resources and National Assembly Matters, Banjul,
the Gambia

Nov 2020 – Aug 2022

- **Departmental Assistance:** Assisted department heads in the management and development of departments.
- **Strategic Planning and Appraisal:** contributed to the identification, formulation, and appraisal of strategic plans, policies, projects, and programs. Prepared groundwork for review.
- **Strategic Planning and Appraisal:** contributed to the identification, formulation, and appraisal of strategic plans, policies, projects, and programs. Prepared groundwork for review.

- **Policy Advisory Role:** responsible for advising on policy formulation and facilitating its implementation.
- **Data Collection and Analysis:** actively involved in data collection and analysis.

Economics, Commerce and Mathematics Teacher

Maahad Senior Secondary School, Brikama

Sept 2018 - Oct 2020

- Prepare lesson plans with a variety of subjects.
- Help students who are physically and emotionally challenged in a curriculum that emphasizes comprehensive inclusion.
- Participate in and observe efficient methods of classroom management.

ADDITIONAL EXPERIENCE

Secretary

Islamic Club of John Pickering Comprehensive Senior Secondary School Oct 2013 -
Apr 2014

- maintaining efficient organization and minute-taking of meetings.
- establishing efficient administrative and record-keeping systems.
- Communication and correspondence.

LANGUAGES AND IT KNOWLEDGE

- English: Fluent
- Mandinka: Fluent
- Wollof: intermediate

COMPUTER SKILL

I have a very good knowledge of:

- R Language.
- Latex (Overleaf).
- Stata.
- Microsoft words and excel.

AWARDS AND INTEREST

- Awarded the overall best outstanding student in John Pickering Comprehensive Senior Secondary School, 2015.
- Football: member of star academy 2009-2011, captain of my home team 2010.

ACHIEVEMENTS

I have completed writing an economics book for senior secondary schools. The book, titled Understanding Basic Economics for Senior Secondary Schools, greatly centres on the major topic that comprise the economics syllabus, as per the WASSCE, the international examination grade 12 students usually sit.

References

Available upon request.