A big data computational framework for enterprise level statistical process monitoring

by

Siim Koppel

B.S., Tallinn University of Technology, 2000
M.S., Tallinn University of Technology, 2003

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Industrial and Manufacturing Systems Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2020

Abstract

The emergence of big data storage together with the evolution of sensor technologies has expanded the amount of data that complex manufacturing facilities can produce. Almost all process variables in the factory can be measured and the data can be stored in data lakes in cloud servers. This big data phenomenon has presented challenges and opportunities for quality improvement teams. While the traditional control charts are still widely used, they are often isolated tools for monitoring product quality characteristics scattered in a manufacturing system. The need to monitor full systems becomes even more pressing with the emergence of smart factories in the next industrial revolution called Industry 4.0.

The goal of this research is to develop a big data computational framework for enterprise-level process monitoring that tracks different variables simultaneously and provides near-time system status updates. To achieve this goal, a novel methodology called Technique of Uniformally Formatted Frequencies (TUFF) is developed that standardizes continuous, discrete and profile variables into comparable statistics, classifies these statistics into four colors using ideas from pre-control charts and summarizes these colors to a single frequency table. This table is used to compare the current situation to historic data and to decide if the performance of the system has changed. A higher resolution of the results identifies the temporal and spatial location of possible change. The comprehensive monitoring method uses all the available data and monitors both quality characteristics as well as process parameters near-time. Additionally, the method is easily scalable to handle big data level datasets. Extensive simulation studies identify the sensitivity and other characteristics of the TUFF method.

This dissertation also redefines one of the more popular Six Sigma continuous improvement methods of DMAIC (Define, Measure, Analyze, Improve, and Control) for the manufacturing environment. The redefined method is Measure, Define, Analyze, Improve and Control (MDAIC), where the unit in need of improvement is identified automatically by the data. The research integrates the TUFF statistical system monitoring method to the MDAIC framework and provides a solution for the implementation of the method in a big data environment based on the MapReduce algorithm.

A big data computational framework for enterprise level statistical process monitoring

by

Siim Koppel

B.S., Tallinn University of Technology, 2000
M.S., Tallinn University of Technology, 2003

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Industrial and Manufacturing Systems Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2020

Approved by:

Major Professor
Shing I Chang

Copyright

Abstract

The emergence of big data storage together with the evolution of sensor technologies has expanded the amount of data that complex manufacturing facilities can produce. Almost all process variables in the factory can be measured and the data can be stored in data lakes in cloud servers. This big data phenomenon has presented challenges and opportunities for quality improvement teams. While the traditional control charts are still widely used, they are often isolated tools for monitoring product quality characteristics scattered in a manufacturing system. The need to monitor full systems becomes even more pressing with the emergence of smart factories in the next industrial revolution called Industry 4.0.

The goal of this research is to develop a big data computational framework for enterprise-level process monitoring that tracks different variables simultaneously and provides near-time system status updates. To achieve this goal, a novel methodology called Technique of Uniformally Formatted Frequencies (TUFF) is developed that standardizes continuous, discrete and profile variables into comparable statistics, classifies these statistics into four colors using ideas from pre-control charts and summarizes these colors to a single frequency table. This table is used to compare the current situation to historic data and to decide if the performance of the system has changed. A higher resolution of the results identifies the temporal and spatial location of possible change. The comprehensive monitoring method uses all the available data and monitors both quality characteristics as well as process parameters near-time. Additionally, the method is easily scalable to handle big data level datasets. Extensive simulation studies identify the sensitivity and other characteristics of the TUFF method.

This dissertation also redefines one of the more popular Six Sigma continuous improvement methods of DMAIC (Define, Measure, Analyze, Improve, and Control) for the manufacturing environment. The redefined method is Measure, Define, Analyze, Improve and Control (MDAIC), where the unit in need of improvement is identified automatically by the data. The research integrates the TUFF statistical system monitoring method to the MDAIC framework and provides a solution for the implementation of the method in a big data environment based on the MapReduce algorithm.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First, I would like to thank my major advisor Dr. Shing I Chang for the time and energy he invested in growing and molding me as a scholar. I am as hard-headed as they come, so it was not an easy task. Additionally, I would like to thank the members of my committee for their time and effort.

Secondly, I would like to thank Dr. Todd Easton for having a really nice open door policy and the patience to listen to my attempts to make a joke. These moments of getting my mind off one topic actually helped me find solutions.

I would also like to thank Dr. Luke Muggy, Dr. Mohammadhossein Amini, Dr. Fabio Vitor-Torres, Palamandadige Kasun Fernando and other students that I met over these years. I enjoyed communicating with all of you and sharing the occasional meal and soccer games.

The most important thank you goes out to my patient wife Dr. Kadri Koppel, who has supported, pushed, dragged, motivated (and so on) me through this journey. I hope there will be a time when I can really repay her for all the misery I have caused.

Dedication

This dissertation is dedicated to the memory of my father Tiit Koppel, who taught me to search for answers and solutions regardless of the size of the problem and handed down an inquisitive mind.

Chapter 1 - Introduction

1.1 Introduction and research motivations

Over the course of human history, three industrial revolutions have taken place. The first was during the 18th and 19th centuries when technological advances started the trend of moving from hand production to machine production. The second industrial revolution is widely considered to be the introduction of the conveyor method as a mean of mass production at the beginning of the 20th century (Schwab, 2016). The mass production dictated a need for parts with constant measurements for quality control. That need was satisfied with the introduction of control charts by Shewhart in 1924 (Montgomery, 2012). The third revolution was the introduction of the microchip to automate the production in the 1970s (Schwab, 2016).

Now humankind is on the brink of the fourth industrial revolution that is based on the digitalization of production. It is characterized by the fusion of technologies that is blurring the lines between the physical and digital world (Schwab, 2016). Today's production processes are becoming more complex with every passing day. The amount of data received from a production line is large and each production company usually owns several facilities consisting of a number of production lines. All the data generated can be conveniently saved to data warehouses and managed there with the help of big data tools. Additionally, the advances in cyber-physical systems, industrial internet, and digital manufacturing have enabled the connection between physical production equipment to computational cyberspace or cloud (Chang, 2017).

All the stages of product life from design to production to quality assurance to customer service can all be delivered through a cloud platform. Every client of the platform has access to a large amount of data. At the same time, the methods that are used to assess the quality and health of the process are largely the same as the ones used after the second industrial revolution. This research is transformative in that it introduces a method that will allow the quality assurance and process monitoring to evolve from the second industrial revolution to cater to the needs of the fourth revolution.

One of the more common tools to assess the quality of the products is statistical process monitoring (SPM) or as previously known, statistical process control (SPC) (Montgomery, 2012). Unfortunately, most existing SPM methods are not capable of analyzing large datasets generated by the production equipment and real-time quality characteristics that are collected from all over the factory floor. The most traditional SPC tool X-bar and R control chart often deal with only one quality characteristic at a time. Multivariate control charts, for example, Hotelling's T^2 charts are rarely capable of using more than 10 characteristics (Montgomery, 2012). More than 1000 sensors may need to be monitored in a modern production line, but control charts are still widely used despite their shortcomings. Other solutions, some of which will be discussed in the next section, have been proposed over the years to monitor the processes, but most of them are either computationally difficult or are hard to fathom for the end-users, so their use in industry is marginal. The current SPM methods are also largely ignoring different spatial and temporal spaces.

The list below provides a few key motivations for this research:

1. Can an algorithm be created for enterprise-level or system-wide SPM that is easily scalable over a large number of variables, is computationally easy to execute in a big data environment and helps define the time and location of the change?
2. Can an algorithm be created for SPM that is capable of merging different types of data, specifically continuous, profile and binomial data?
3. How to implement this type of algorithm in big data environments?

1.2 State of the art in SPM

Weese *et al* (2016) did a comprehensive review of the statistical learning methods up to 2015. They claimed to view these methods from the scalability to big data problems but did not offer any solutions on how to do it. They looked at unsupervised learning approaches such as Principal Component Analysis (PCA), Partial Least Squares (PLS), Factor Analysis (FA), Least Absolute Shrinkage and Selection Operator (LASSO), multivariate exponential weighted moving average (EWMA) charts, Support Vector Machines (SVM) and other clustering methods; and supervised learning methods such as control chart pattern recognition, regression-based methods, and neural networks.

In another review, Yin *et al* (2014) looked at basic data-driven approaches for SPM. They concluded that due to uncertainties and complexities of modern industrial systems it is virtually impossible to construct process models based on the first principles or linear regression. They reviewed the usage of PCA and PLS within the multivariate statistical process framework.

Ge *et al* (2013) reviewed and classified most of the data-based process monitoring methods based on which processes these are best applicable to, such as non-Gaussian processes, non-linear processes, time-varying, and multimode processes and dynamic processes.

A review on the usage of support vector machines in statistical process monitoring can be found by Cuentas *et al* (2016). These authors aimed to provide researchers with a starting point to potentiate the performance of the support vector machine classifier to achieve the best possible performance and improve detection efficiency. The problem of scalability to big data application was not one of the goals of the review.

Maleki *et al* (2017) did a review of research conducted on measurement errors in SPM. The conclusion of their review was that the effect of contaminated measurements must be explored more in detail in the event of big data applications. Also, they brought out that very little attention has been given to multivariate processes and profile monitoring applications.

A systematic comparison of PCA-based SPM methods was carried out by Rato *et al* (2016). The methods were used on different datasets such as autocorrelated data, nonstationary data, etc. and the pros and cons of each method were discussed.

A comparative study on monitoring schemes for non-Gaussian distributed processes can be found by Li & Qin (2016). The focus was on support vector data description, kernel density estimation, Independent Component Analysis (ICA) and statistical pattern analysis.

De Ketelaere *et al* (2015, 2016) reviewed SPM on time-dependent data, focusing mainly on PCA-based methods and how can they handle different types of data such as auto-correlation and non-stationarity.

More recently, Zhu *et al* (2018) studied robust data mining approaches in industrial process monitoring in cases when the data was considered to be polluted because of outliers or missing data. The methods were divided into two main groups: robust data mining for data preprocessing and statistical modeling. Different methods of outlier detection, missing data imputation, normalization, principal component analysis, Bayesian principal component analysis, and non-linear/non-Gaussian/dynamic modeling were discussed. The authors also discussed big data modeling, data selection, data dependence analysis, and information fusion, and multi-view monitoring.

Jiang *et al* (2019) reviewed different multi-block PCA, PLS and canonical correlation analysis (CCA) methods and how these methods fit into their own proposed plant-wide data-driven distributed monitoring method. The paper examined basic multivariate statistical process monitoring methods (Hotelling T^2, PCA, PLS, CCA), then established motivations for distributed monitoring, where the basic methods were used in smaller subsets within the system and based on these statistics, decisions were made on the full system. The decision rules were Bayesian interference statistic or Bayesian fault diagnosis system.

Abundant research has been conducted in this field as can be seen from the reviews. Three main methods are used: PCA, PLS and control charts. The following is an overview of some of the methods that have been published in recent years.

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. Tong & Yan (2015) looked at a modified multiblock PCA algorithm for decentralized SPM. The measured variables were divided into several overlapping blocks, which then were analyzed with multiblock PCA. Since the method relies on Hotelling T^2, the use of it in big data applications is doubtful. The numeric example given used 8 variables. Jiang *et al* (2016) explored the distributed PCA process with the help of fault-relevant variable selection and Bayesian interference. The optimal subset of variables was identified for each fault using the optimization algorithm. A sub-PCA model was established for each subset and finally monitoring results were combined through the Bayesian process. The examples given used up to 32 variables.

Gajjar *et al* (2018) investigated the use of sparse principal component analysis to detect faults. The reason for using SPCA was that interpreting the principal components from a large dataset was deemed challenging. The SPCA would use sparse loadings to some principal components and therefore make the detection of faulty variables easier. The proposed method was used on the Tennesee-Eastman process with 32 variables. There are many other studies published, but they all rely on similar or same simulated datasets of up to 32 variables, for example, see Jiang & Yan (2014) and Zheng *et al* (2015). Menafoglio *et al* (2018) studied profile monitoring of probability density functions using simplicial functional PCA. Their work was using image data to detect faults. The proposed approach summarized the random occurrences of faults via their probability density functions and then monitors these functions using PCA. The example provided was on metal foam porosity faults.

Zhang *et al* (2018) developed fault detection and diagnosis method based on weighted and combined indexes of the residual subspace associated with PCA. First, the residual subspace was divided into two subspaces according to the residual percent variance. The subspaces were active subspace and stable subspace based on the idea that the residual variance should reach a steady state when the factors begin to account for random errors. Next, the authors established a weighted index by combining Hotelling T^2 statistics and the Euclidean distance statistics which was then monitored to detect faults.

Partial Least Squares (PLS) regression is a statistical method that finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Wang *et al* (2015) used kernel PLS based prediction model construction to monitor non-linear processes. Relations between input and output variables were considered in a high-dimensional space, latent vector matrices were calculated through reformed nonlinear iterative partial least squares method. The examples used showed only univariate datasets. Wang and Yin (2015) used orthogonal signal correction and modified PLS to detect quality-related faults. OSC was used to remove undesired systematic variation from the input space. Then the input was decomposed into two orthogonal subspaces of fully responsible for predicting the output and uncorrelated to output. The maximum number of variables used was 32. Zhou *et al* (2018) introduced a statistical model based on a global plus local projection to the latent structures that focused more attention on the relevance of extracted principal components. The linear correlation information between the process and quality variables extracted was maximized and the local nonlinear structural correlation information was extracted as much as possible. The Tennessee-Eastman process was once again used as an example with 32 variables.

Tong *et al* (2019) proposed the use of distributed partial least squares based residual generation for SPM. Their proposed method developed a soft sensing model for each specific variable and the generated residuals reflected the variations in the input-output relationship. The example was the Tennessee-Eastman process. The biggest issue with PLS is that like any regression modeling, the larger the number of variables, the more difficult it is to estimate variance-covariance matrices.

As stated previously, the control charts are still used and also studied. Exponentially Weighted Moving Average (EWMA) and cumulative sum (CUSUM) charts are two variations of control charts. EWMA chart tracks the exponentially-weighted moving average of all prior sample means. EWMA weights samples in geometrically decreasing order so that the most recent samples are weighted most highly while the most distant samples contribute very little. Haq *et al* (2015) proposed an EWMA chart that was using the best linear unbiased estimator paired ranked set sampling method. They claim that the method performs better than other similar methods in detection and also is more sensitive. However, the method is used for univariate data. Multivariate EWMA chart based on the variable selection using the Akaike information criterion. The number of variables in examples was five. Huwang *et al* (2018) introduced a charting methodology for monitoring linear profiles. The method was developed based on the spatial rank of the vector of the estimators of the regression coefficients and error variance. These coefficients and error variance was then monitored using EWMA charts.

Suman and Das (2019) used EWMA on latent variables scores to detect developing faults early. First, the raw data was analyzed using PLS and then the scores were monitored using EWMA. The CUSUM chart accumulates information from the process over time and therefore is more sensitive to small and moderate changes in the process mean. Zhang and Woodall (2016) explored dynamic probability control limits for lower and two-sided risk-adjusted CUSUM charts. They discovered that the in-control performance of their charts varies significantly less than with the chart with fixed limits. Saleh *et al* (2016) examined the CUSUM control chart with estimated parameters by considering between-practitioner variability. They recommended designing the charts using a bootstrap approach.

Daryabaryi *et al* (2019) explored the overall performance of the Bernoulli CUSUM chart on the presence of measurement errors and considering learning effect using the average number to signal measure as the performance metric. They showed that measurement errors deteriorated the performance of the chart and that the effect of Type I measurement errors was larger than the Type II error. The biggest shortcoming of the EWMA and CUSUM charts is that most of them are used to monitor changes in mean and do not consider variance shift at all. The other issue is also scalability of the methods. Usually, these charts are given for a univariate or small number of multivariate datasets.

In conclusion, the largest shortcomings of current SPM methods are that they are focused on quality characteristics and not on other processes, they are not applied for the whole system, they are not capable of functioning with different data types, they ignore spatial and temporal spaces, they do not consider data streaming and they do not scale up beyond 50 dimensions.

1.3 Challenges and Goals

Most of the data generated by the production companies usually fulfill the requirement of the definition of big data by large volume and velocity. The use of video cameras and photo lenses add to the variety of data. The research proposed in this document focuses on the volume and velocity dimension of the big data in manufacturing. Comparisons between classical SPM implementations versus the opportunities that big data paradigm brings are listed in Table 1 (Chang, 2017)

Table 1. Classic SPC implementation vs the big data opportunities

SPM Functions	**Classical SPM**	**Big Data Opportunities**
Data Availability	Scarce	Abundant
Data Type	Restricted to numbers	Numbers, text, image, voice
Data Collection speed	Slow	Fast
Data Collection frequency	Infrequent	Real- or near-time
Data analysis	Using samples	Using all observations
Data dimensions	Small (less than 10)	Large (hundreds or thousands)
Data Archive	Limited past	All historical observations
Data visualization	Control charts	Spatial, temporal, multiscale
Data subject	On product	On both process parameters and products

The first hurdle is the volume of the data. The ever-expanding frontiers in data recording and storing have made it possible for enterprises to gather large amounts of data about every detail in their processes. The amount of new data generated every second can be difficult to fathom.

Sensors are capable of reading 400 to 1000 values per second, which adds the velocity dimension. There can be thousands of sensors on the production line, several lines in a facility and several facilities in a corporation. Enterprises have large quantities of historic raw data, unfortunately, little is used to generate knowledge about processes. The traditional SPM assumes that you can monitor local quality characteristics and therefore the system-level monitoring is ignored. The system-level monitoring would allow predicting the output. If settings change due to wear of the machine or the tool, the quality of the product can change as well.

The second issue is that the quality characteristics can be in different formats. Some processes are continuous, some can be described best by profiles and some use batches. To complicate matters further, the use of cameras is steadily rising and more information is captured by pictures or even movies.

Most of the methods described in the previous section are capable of dealing with either one type of data or with a small number of variables. When the requirements are much larger, the methods are not capable of handling such situations. There is certainly a need for a method that can accommodate a large amount of historical data. It should be scalable, simple to implement and understand, and computationally efficient. The method is useless if analyzing 1 minute of data from one machine takes 1 hour.

After all the analysis is completed, the knowledge must be summarized and presented to the decision-makers. Visualization has been proven to be an effective way of conveying messages. Also, it is important that analysis can be summarized depending on the level of interest. The factory manager is interested mainly in the status of one’s factory; the area manager is interested in the summarization of the situation of the factories in their area etc. The method proposed could be used to analyze all the data at the same time and have the flexibility to cater to the needs of most clients.

Thus, the goals of this research are to

- create an algorithm for system-wide Statistical Process Monitoring that is
    - usable with continuous and profile variables
    - capable of merging different types of data into a uniform format
    - able to identify timeframes and location of possible changes
- establish a solution for big data implementation on the Statistical System Monitoring method

1.4 Research Contributions and Organization of Dissertation

This research was inspired by the USA presidential elections in 2016. While there are 50 states that contribute to the election of the president, most of the attention is focused on the few "swing" states that have historically elected both Republican and Democrat candidates. A large number of states almost always vote for Republican (Texas) or Democratic (California). Similar tendencies can be credited to the manufacturing industry. The data from a large number of variables are gathered, but only a few are monitored and analyzed thoroughly because they are the most important or most prone to change (such as "swing" states). However, similarly to presidential elections, all the variables must be accounted, just in case there are changes in those variables.

Traditional quality monitoring methods that are still widely used in industry have limited capabilities when it comes to system-wide monitoring. They are usually geared towards monitoring locally without taking into account the events from upstream of the production line.

They are also focused on sampling and monitoring quality characteristics assuming that if those are in-control, the process parameters are also in-control. Big data allows much more flexibility in storing and analyzing each variable regardless if it is the current or historic measurement. In this dissertation, a novel system monitoring method called Technique of Uniformally Formatted Frequencies (TUFF) is proposed. A classification based on pre-control charts is utilized to monitor continuous variable, profile variable and attribute variable based data in a big data environment. All the readings are analyzed and summarized on different levels of interest to detect changes in overall performance and to pinpoint possible culprits of change. A novel approach to the Six Sigma continuous improvement process is also proposed where the signal for possible improvement projects is initiated by data rather than a person. The research is expected to help owners of the smart factories monitor the whole factory at once and identify the location and timeframe of possible issues in quality characteristics as well as process parameters for more in-depth analysis.

1.4.1 Profile Monitoring using Modified Sample Entropy

First, a short literature review is presented on the current profile monitoring methods and Adjusted Modified Sample Entropy. A method to monitor a single profile using a modified Sample Entropy is presented next. A simulation study shows the sensitivity of the method to the changes in the underlying model of the profile as well as the variation changes of the profile. Specifically, the chapter makes the following contributions:

- A novel method for profile monitoring is presented that uses modified sample entropy
- The profile is characterized by a single value or set of values when the profile is segmented

- A simulation study shows that the method is capable of detecting changes in underlying models as well as variation along the profile
- The segmentation of the profile helps to identify the faulty part of the process

1.4.2 Statistical System Monitoring (SSM) for Enterprise-Level Quality Control

Statistical system monitoring called TUFF is proposed where all the process parameters and quality characteristics are considered simultaneously for change detection. Background into traditional multivariate statistical process control is presented with a literature review on Hotelling T^2, Principal Component Analysis, Group Control Charts, and Pre-control charts are discussed. The desired properties of enterprise-level system monitoring are presented and a method for monitoring continuous variables in system-level is proposed. A continuous variable is monitored using classification based on the pre-control chart and group control chart ideas. The classifications are summarized and presented in "traffic-light" visualization. Results from simulation studies are presented that show the thresholds of changes for decision-makers, the maximum number of variables that can be used in the group control phase and the sensitivity of the method. Recommendations for implementation of the method in a big data environment are provided. The chapter makes the following contributions:

- The requirements for the system-wide monitoring are identified
- A method is proposed on how to monitor continuous variable in a system-wide environment
- A step-by-step example is provided on how to implement the TUFF method in an enterprise as well as in big data environment

- Simulation studies are conducted to identify the characteristics of the method, including thresholds for decision-makers, the maximum number of variables that can be grouped together and what is the sensitivity of the method

1.4.3 Monitoring Profile Data in a System-Wide Monitoring Framework in the Big Data Era

A literature review on single and multiple profile monitoring is presented. The background on pre-control charts, group control charts and continuous variable monitoring in the system-wide framework is explained. The monitoring of a profile type of data with TUFF is proposed. The method uses classification based on pre-control charts and group control charts with the "traffic-light" type of visualization as the statistics to monitor. The raw profile is monitored with a method that can characterize the profile, for example, one from Chapter 2, and these characteristics are then classified. Results from simulation studies show the recommendation for machine-level statistics and sensitivity of the method in the profile domain. The main contributions are

- A method is proposed that standardizes profile monitoring into similar representation as to the continuous variable tracking in the previous chapter
- The monitoring has shifted from raw data to characterizing values
- Simulation studies are carried out to identify which summarization method would be capable of identifying changes best out of average over all control points, a worst-case out of all control points and single characterizing value (Adjusted modified Sample entropy). The most sensitive method is to use worst-case out of all control points

- The second simulation study establishes the sensitivity of the method. Small changes in a small number of profiles (1% of profiles) are not detected, however, medium changes are detectable in 1%, 5%, 10% and 20% of changed profiles cases

1.4.4 MADIC- a Six Sigma Implementation strategy in Big Data Environments

A literature review is presented on the current state of Six Sigma in big data environments. One of the more popular techniques of Six Sigma, DMAIC (Define, Measure, Analyze, Improve, Control) is proposed to be redefined in the realm of big data possibilities in the manufacturing environment. An application of the redefined process is presented. The use of system-wide monitoring is discussed in terms of continuous improvement. The indexing method in a big data environment is proposed and shown in an example of the implementation of the system-wide monitoring. Specifically, the chapter's contributions are:

- Refining DMAIC in a manufacturing environment, where the signal for potential improvement candidate comes from data rather than from a person. The new sequence of steps would be: Measure, Define, Analyze, Improve, Contol
- The merging of the continuous variable, profile variable and attribute variable data into system-wide monitoring TUFF framework is proposed using pre-control charts, group control charts and classifications on characterizing values of different variable types
- A big data implementation solution is proposed for the Six Sigma process and for the TUFF method using a MapReduce algorithm

1.4.5 A Visualization tool for Multivariate Process Monitoring in Data-abundant environment using Adjusted Modified Sample Entropy

A background on modified sample entropy, adjusted modified sample entropy, trellis displays, and star glyphs are presented. A method for visualization is presented that uses elements from the background. A visualization tool is presented that can be used to monitor a large number of variables. The method is demonstrated in different settings. The main contributions:

- A novel tool is presented to visualize a large number of variables simultaneously
- The tool is capable of helping to determine changes in variables

Chapter 2 - Profile Monitoring using Modified Sample Entropy

2.1 Introduction

A profile is defined as a relationship between a response variable and the explanatory variable(s) (Woodall, 2007). The explanatory variable is usually either time or space. Since a lot of processes in manufacturing are using profiles to characterize their performance, profile monitoring has drawn attention over the last 10-15 years. Various methods have been proposed, some of which have been described in the following section. The problem with most of those is that the algorithms are usually computationally demanding and lack the capability of easy comparison. Since there is a trend to generate larger quantities of data, there is a need for a method that has the capability to detect and quantify changes and could be used to analyze all the incoming data.

2.2 Background

2.1.1 Current profile monitoring methods

Studies on profile monitoring methods in quality control before 2007 have been reviewed by Woodall (2007). Studies after 2007 include for example Chang and Yamada (2010), who studied monitoring of non-linear profiles using wavelet filtering and B-Spline approximation. In another study, Chou *et al* (2014) researched simultaneous process monitoring for multiple linear or non-linear profiles. Zeng *et al* (2014) studied Phase I monitoring of profile data in non-normality assumption cases. These authors used independent component analysis to transform multivariate coefficient estimates to independent univariate data and then used univariate nonparametric control charts to detect changes. Chang *et al* (2014) tried to detect changes in wave profiles in real-time [5]. In another study, Shang *et al* (2016) looked at change point detection with binary data profiles and random predictors using a logistic model. Paynabar *et al* (2016) looked at multivariate profile monitoring using multidimensional functional principal component analysis.

2.1.2 Adjusted Modified Sample Entropy

This study used Adjusted Modified Sample Entropy as the tool to detect changes in profiles. It is a version of Sample Entropy that has been upgraded by Xie *et al* (2010) to detect smaller changes and by Kong *et al* (2015) and Koppel *et al* (2016) to help detect changes in both mean shift and variance change. In essence, the Sample Entropy calculates the negative natural logarithm of the conditional probability that two similar sequences for m points remain similar within a tolerance of r at the next (m+1) point. The Sample Entropy has a Heaviside step function where a value of 1 is assigned if the distance between two random points in the time series is less than the set threshold of r and 0 if the distance is greater. Xie *et al* (2010) introduced a smoothing concept to the method, where the step function is replaced with a fuzzy membership function that assigns weights as values. The closer the value of the member is to the set goal, the higher the weight is assigned.

Kong *et al* (2015) introduced a transformation of the original data to Sample Entropy. The original data is segmented, the means of each segment are calculated and then the transformation is calculated using the following formula

$$\boldsymbol{y_{ij}} = \boldsymbol{x_{ij}}\left(\left|\frac{\boldsymbol{\bar{x}_{\imath}} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}\right| + \mathbf{1}\right) \tag{1}$$

where $\bar{x}_{\imath}$ is the estimated mean of i^{th} segment set; μ is the desired mean of the variable of interest, and σ is the desired standard deviation of the variable of interest. After the transformation, the Sample Entropy is used on the new data. The method is capable of detecting mean shift and variance changes. Koppel *et al* (2016) used the same transformation to transform the original data for Modified Sample Entropy.

2.3. The proposed method for detecting changes in profiles using Modified Sample Entropy

The proposed method for change detection is presented in this section. As is the case with most quality control methods, this includes Phase I and Phase II procedures. The parameters are set for the upper control limit in Phase I and the monitoring is conducted in Phase II.

Phase I steps include the following. Import training set of profiles that have been proven to be good. The suggested number of profiles is between 10 and 40. Next, calculate the baseline by averaging each point in time over all the selected profiles. After that, remove the baseline from all the profiles. Then determine where the profile has direction changes by either visual inspection or using any method of change point detection for segmentation purposes. Segmentation can be also done by dividing the profile into equal chunks with the same number of observations in each of the segments. Next, divide the profile into segments and calculate the mean for each segment. Input the transformation of Adjusted Modified Sample Entropy according to equation (1) on all the segments on all the profiles of the training set. Then calculate Modified Sample Entropy for each segment of each profile in the training set. Finally, calculate the upper control level for each segment using mean plus *k* times the standard deviation of the entropy values. The *k* is selected to be 4 to lessen the Type 1 error. The lower level is not necessary, because that implies that the mean has not changed and the variance of the profile under investigation is less than the variance of the training set or the approved profiles.

Phase II steps include the following: import the new profile (this step can be also done on each segment as it is generated to monitor the process near-time), remove the baseline determined in the Phase I, calculate the mean for each segment. Then input the transformation of Adjusted Modified Sample Entropy according to equation (1) on all the segments of the profile, or the segment generated in the near-time application. Next, calculate the Modified Sample Entropy for each segment of the profile or for the segment under investigation. If any value of the Modified Sample Entropy in any segment exceeds the upper control level, the profile is marked and needs closer attention to determine if the output of the process is acceptable or not and the cause of the change.

2.4. Simulation Study on the use of the proposed method

The following simulation study was conducted in order to demonstrate the capability of the proposed method to detect changes. A simple profile was generated and analyzed (Figure 1). The profile had three segments: an upward slope, a stable segment and then a downward slope mimicking for an example temperature reading of a vulcanizing process.

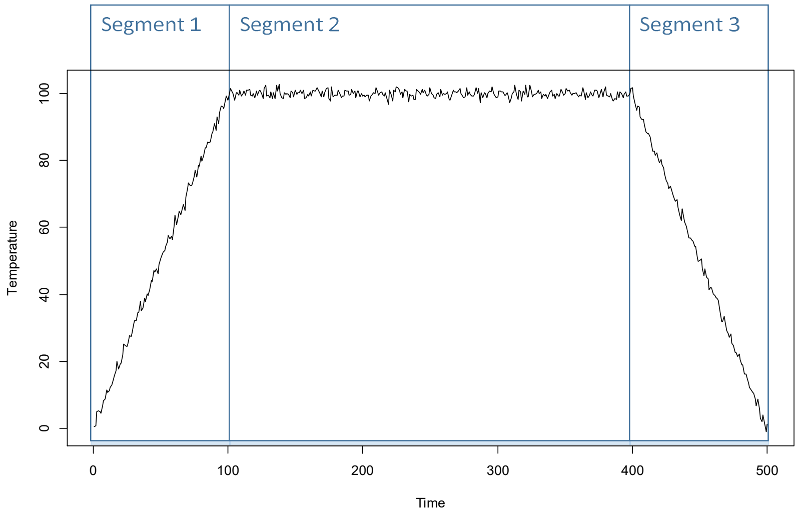


Figure 1. A sample of the temperature profile in vulcanizing

The profile was simulated using the following models:

	Segment 1:100	Segment 101:400	Segment 401:500
Model	*y=x+e*	*y=100+e*	*y=500-x+e*

where *y* is the output, *x* is the time value and *e* is error term. In this case, the error term was generated based on a normal distribution with a mean of 0 and a standard deviation of 1.

A total of twenty such profiles were generated. The profile was segmented based on the changes of directions in profile according to the procedure introduced in the previous chapter. The segmentation was done visually because the model was simple. The mean values for each segment were calculated and the transformation was conducted. Then the modified Sample Entropies for each segment of each profile were calculated. Based on these numbers, upper control limits were calculated for each segment using the mean values plus four times the standard deviation. The control limits were as follows:

	Segment 1	Segment 2	Segment 3
Upper Control Limit	0.3277	0.2602	0.2796

In Phase II, five different scenarios were used: 1) no change, 2) variance change for all of the profile, 3) the first slope was steeper, 4) the second slope was steeper and 5) the holding segment had a higher temperature. A total of 100 profiles was simulated in each scenario while changing the needed parameter. The count of profile segments with an alarm was generated out of each scenario.

In scenario 1, the slope models and the holding segment model were the same; the error was generated using a normal distribution with a mean of 0 and a standard deviation of 1. All the profiles were segmented into three segments as the Phase I profiles. The baseline was removed from each generated profile. The means of each segment of each profile were calculated and used to transform the original data with the Adjusted Modified Sample Entropy transformation formula. Then the Modified Sample Entropy was calculated to each of the segments of each profile and compared to the upper control limit identified in Phase I.

The results showed that in segment 1 there were four profiles with a value that was higher than the upper control limit for the first segment. There were no alarms on the profiles of the second segment and four alarms on the third segment, but those were not on the same profiles as the alarms for the first segments.

In scenario 2, the variance changed for the whole profile, but the models for slopes and holding segment stayed the same (Figure 2). The variance change was achieved by generating the error term using a normal distribution with a mean of 0 and a standard deviation of 2.

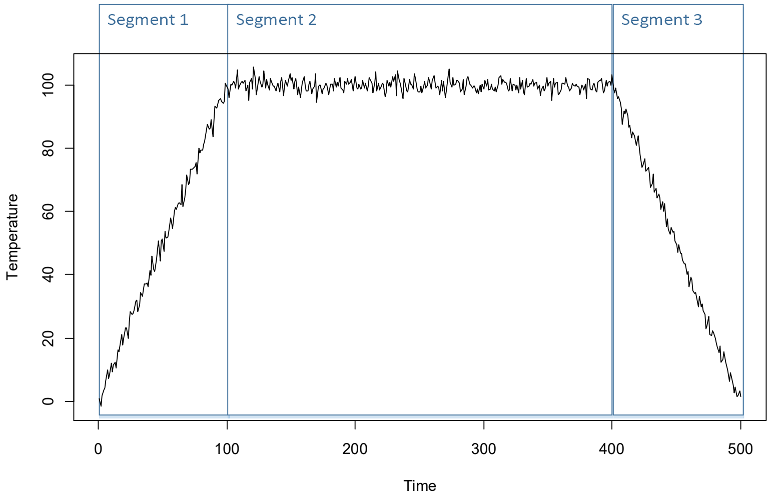


Figure 2. Example of the profile with larger variance

All the steps of Phase II were carried out on all the 100 profiles. The results indicated that the alarm was generated in all the slope segments and the holding segments. Therefore, the change was detected in all the segments.

In scenario 3, the first segment had a steeper slope and therefore the holding time was longer than the holding segment in Scenario 1 and 2 (Figure 3). The third segment did not change and the error was generated using a normal distribution with a mean of 0 and a standard deviation of 1.

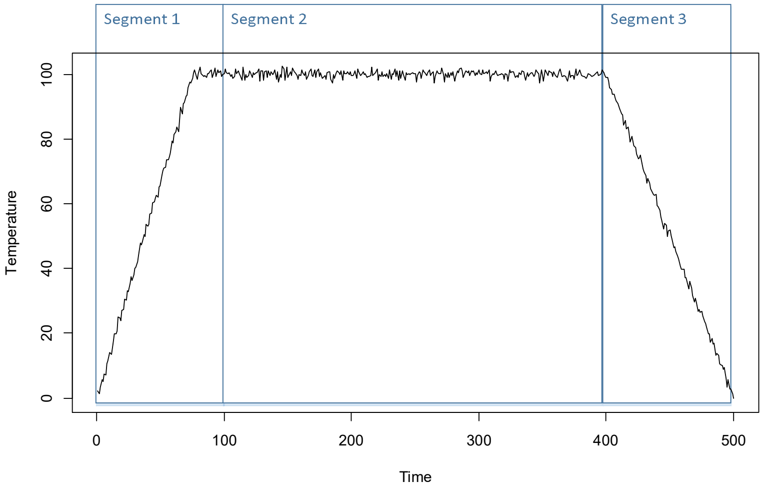


Figure 3. Example of a profile with a steeper first slope

The profiles were segmented, the baseline was removed, means were calculated, original data was transformed and Modified Sample Entropies were calculated on all the segments of all the profiles. The results showed that there were 100 alarms in segment 1, none in segment 2 and 6 in segment 3. The algorithm was capable of detecting the change in all the profiles and accurately identifying the segment where the change had happened.

Scenario 4. The third segment had a less steep slope and therefore the holding time is shorter than the holding segment (Figure 4). The first segment did not change from Phase I profiles and the error was still normal distribution with a mean of 0 and standard deviation of 1.

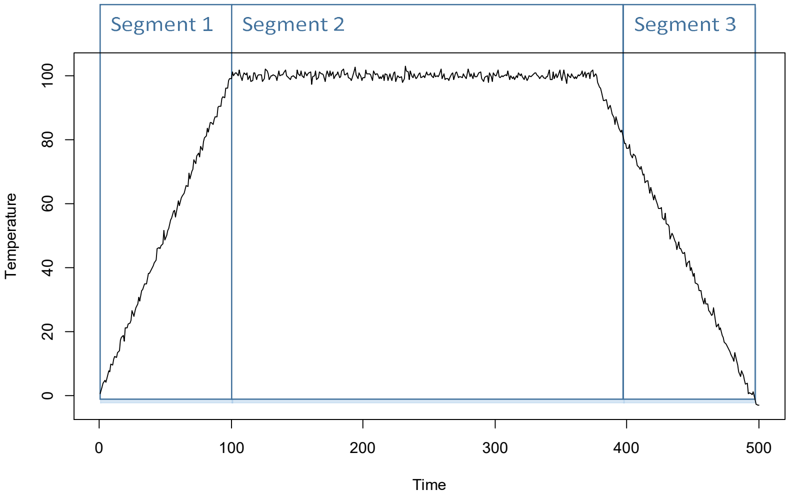


Figure 4. An example of a profile with a less steep second slope

The steps of the method were completed and the results showed 1 alarm in Segment 1, 100 alarms in Segment 2 and 100 in Segment 3. Therefore, the method was capable of detecting the change in all the segments and also identifying the segments where the change had happened.

Scenario 5. The error term was the same normally distributed with a mean of 0 and standard deviation of 1. The first slope had the same angle as in Scenario 1, but it continued further to 120 (Figure 5). Therefore, some of the increase happened in the holding segment. Holding time was the same, but in higher temperature, affecting the third segment as well. The cool-down was with the same slope as the Scenario I profiles, but since it started later, the minimum temperature was not achieved.

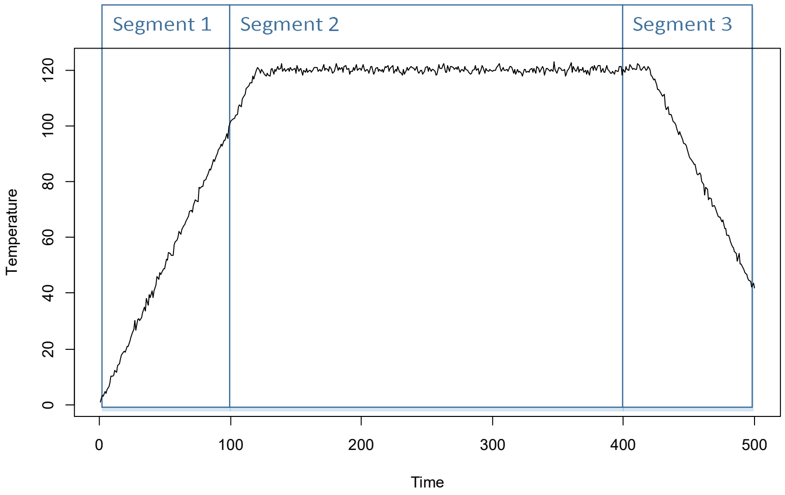


Figure 5. An example of a profile with higher temperature in the holding segment

The results indicate that no alarms were detected in the first segment, 100 alarms were in the second and third segment. The method was capable of detecting all the changes and also identifying all the segments where the change happened.

2.5. Conclusions and Future Studies

A novel method for change detection in profile monitoring has been presented in this study. It is based on the Adjusted Modified Sample Entropy. The tool is capable of identifying changes in variation and in the shape of the profile. It is also capable of identifying which segment of the profile is changing.

For future research, while the base of this method has been studied in big data applications, there is a need to investigate the capabilities and adaptability of this method to function in the big data environment. It shows a lot of promise in providing input in the indexing of massive datasets. The idea is that all the raw data is analyzed with simple methods, given identification and stored into a data warehouse. It can then be retrieved for further, more accurate analysis if needed. The outcomes of simple methods are then used for monitoring the health of systems and cumulative reporting with different reporting horizons. There is also a need to study more complex cases of profiles like second-order and third-order models and different error distributions.

Chapter 3 - Statistical System Monitoring (SSM) for Enterprise-Level Quality Control

3.1 Introduction

Statistical process control (SPC) approaches were first introduced by Walter Shewhart (1930). The use of statistical methods such as hypothesis testing in graphic forms coupled with the Central Limit Theorem has served numerous applications well especially in manufacturing to achieve desired product quality. The core concept is the use of a set of historical data deemed in control to set up a pair of control charts. This process is usually called Phase I SPC. Then during Phase II SPC, statistics of future observations of a quality characteristic are plotted for continuous monitoring. If any point plots outside control limits, then the process under monitoring is deemed out of control. Process engineers are then informed of fault diagnoses.

This basic framework remains unchanged to this day although multiple revisions such as CUSUM and EWMA charts have been proposed to improve the sensitivity of detecting small process shifts (Runger *et al*, 2004; Lowry *et al*, 1992; Hu *et al*, 2007). Hotelling T^2 (Hotelling, 1947) was proposed to extend univariate quality characteristics to multivariate quality characteristics. However, the combinational development of omnipresence of sensors and cloud computing often-called Industrial 4.0 or cyber-physical systems (Lee *et al*, 2015) has opened up opportunities to rethink implementation strategies of SPC for manufacturing. Traditional SPC methods, often restricted to product quality characteristics, cannot take advantage of big data generated from a production process equipped with thousands of process parameters and hundreds of product characteristics scattered throughout a production system.

We propose a system-wise process monitoring framework to answer this challenge. The proposed framework is called Technique of Uniformally Formatted Frequencies (TUFF) and it is used for **statistical system monitoring** or SSM in that all process parameters and QCs are considered simultaneously for change detection. A system is composed of multiple processes that may be hierarchical. This chapter provides the SSM framework that adopts parts of a collection of monitoring methods such as pre-control and group control chart algorithms. The core concept of TUFF is to, first, quantify the performance of a system or subsystem composed of process parameters and product quality characteristics over time into three zones (green, yellow, and red zone in Figure 6). Then, only those segments of time series that exhibits changes in terms of the statistics reflecting green, yellow, and red zones are to be analyzed. A small manufacturing example with three departments is used to demonstrate the use of the proposed method. The goal is to monitor the full system, not just the individual parts. Simulated data sets are generated to demonstrate the properties of the proposed method.

Unlike the traditional methods where measurement is restricted to physical products or work in progress, the TUFF framework integrates process parameters associated with products or work in process for process monitoring and defect prevention. Since the number of parameters is usually very large, a high dimensional problem often confronts traditional control charts. For example, the production of semiconductor wafers includes hundreds of processes and thousands of process parameters. The TUFF framework contains multiple techniques for dimension reduction and feature selection. The following section briefly outlines some of these methods in the content of statistical process control or monitoring.

3.2 Background

3.2.1 Traditional Multivariate SPC

Much research has been generated on the topic of statistical process monitoring over the last decades. Most work focuses on a univariate quality characteristic.

3.2.1.1 Hotelling T^2

Hotelling T^2 is one of the oldest SPC methods for monitoring multivariate processes. This method is monitoring the mean vector of the process. The monitoring is done by plotting a chi-squared control chart (Hotelling, 1947).

The statistic plotted is calculated based on the vector of variable means over some time. Then a covariance matrix is used to calculate the statistic. There also lies the biggest issue of the method – the estimation of the elements in a covariance matrix. It is possible to calculate the covariance matrix when there are about 10 variables, but anything over that, the task becomes very difficult or next to impossible to implement. Also, Hotelling T^2 is often applied to several QCs. Usually, process parameters associated with the QCs are not considered in the same vector. Multivariate versions of Exponentially Weighted Moving Average (EWMA) and cumulative sum (CUSUM) charts (Lowry *et al*, 1992; Pignatello & Runger, 1990; Crosier, 1988) suffer the same drawbacks. These control charts were merely used to enhance the chart performance of catching small shifts.

3.2.1.2 PCA

When a large number of multiple quality characteristics are encountered, Principal Component Analysis (PCA) is often used for dimension reduction. PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (Montogomery, 2012). Jiang and Yan (2014) proposed a tool to monitor multi-mode plant-wide processes by using mutual information-based multi-block PCA, joint probability and Bayesian inference. Tong and Yan (2017) studied a modified multiblock PCA algorithm for decentralized Statistical Process Monitoring. In another study, Liu *et al* (2015) explored statistical process monitoring with the integration of data projection and one-class classification using PCA. In a different study, Zheng *et al* (2015) studied a time series model coefficients monitoring approach for controlled processes. Jiang *et al* (2016) proposed a distributed PCA process with the help of fault-relevant variable selection and Bayesian interference. In another study, Gajjar *et al* (2016) proposed to detect faults with the help of Sparse Principal Component Analysis. Gajjar and Palazoglu (2016) also proposed a data-driven multidimensional visualization technique to process fault detection and diagnosis. Yan *et al* (2016) proposed a robust multivariate process monitoring via stable principal component pursuit. Their goal was to increase the PCAs robustness against gross outliers. In a different study, Jiang *et al* (2016) investigated a Gaussian mixture model and optimal principal component-based Bayesian method for multimode fault diagnostics.

Most of these studies showed that their methods work well with the Tennessee-Eastman process that has 52 process variables. The Tennessee-Eastman process is a benchmark process that consists of five main process units: a two-phase reactor where an exothermic reaction occurs, a separator, a stripper, a compressor, and a mixer. This is a nonlinear open-loop unstable process that has been used in many studies as a case study for plant-wide control, statistical process monitoring, sensor fault detection, and identification of data-driven network models. However, it is unclear whether these methods are capable of scaling up for a production facility with more than a thousand process parameters.

3.2.2 Group Control Charts

The control charting methods reviewed so far often focus on a single product with one or multiple quality characteristics. Boyd (1950), recognizing the need for applying process monitoring for multiple-stream processes, proposed the use of group control charts. Multiple streams are defined as multiple input sources of the same product. Any chart in a group is based on a pair of average (i.e. X-bar) and range (i.e. R) charts. All streams are sampled and each is monitored by a pair of X-bar and R charts. The group control chart framework only records the largest, smallest mean, and the maximum range of the streams with the understanding that if these are within the control limits, the other streams must be too. Specifically, the TUFF method adopts the idea of monitoring the worst-case scenarios (Montogomery, 2012).

3.2.3 Pre-control

Pre-control is a technique used to detect shifts or upsets in the process that may result in the production of nonconforming units (Satterthwaite, 1954). The technique differs from a statistical process control in that conventional control charts are designed for real-time monitoring while pre-control is mainly used to assure process capability. Pre-control uses the normal distribution in determining changes in the process mean or standard deviation that could result in increased production of nonconforming units. Only three statistics related to green, yellow, and red zones are required to provide control information as shown in Figure 6.

The original pre-control method assumes that the process is normally distributed and the natural tolerance limits ($\mu \pm 3\sigma$) exactly coincide with the specification limits. Therefore, this process produces 0.27% of fallout in the red zone. The control works by setting an upper and lower process control limit at ¼ and ¾ of the specification limits. Two consequent samples are drawn

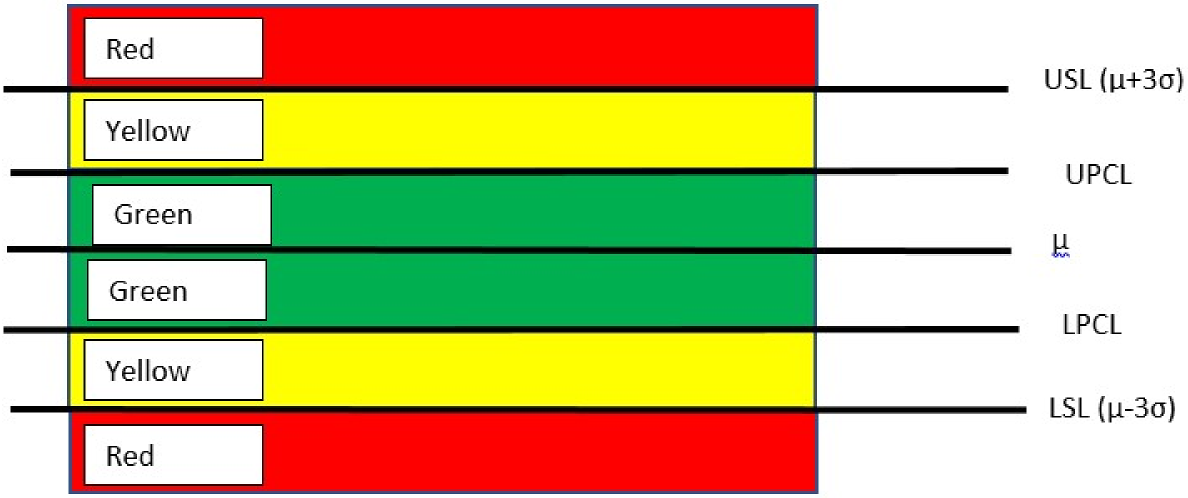


Figure 6. The basic set-up of the pre-control chart

and compared to the UPCL and LPCL. If they fall within the green zone, the process has not changed. If two samples fall to the same side of between control limit and specification limit (i.e. the yellow zone), the mean might have changed. If the samples are on the opposite sides between control and specification limits, the standard variation might have changed. If they are outside the specification limits (i.e. the red zone), the process has produced non-conforming parts. Comparing the traditional control charts, pre-control needs to be executed at the beginning of a shift and about six times during a shift for quality assurance purposes. In addition, traditional control charting usually requires a pair of charts – one for mean shift and the other for variance changes while pre-control only requires one chart. It is not an emphasis in the pre-control chart to connect the dots of consecutive samples as the traditional control chart does. This proposed work uses the color scheme of pre-control to classify samples while defining different limits for determining the classification of each sample. The upper and lower control limits on the TUFF method are set at 1 sigma from the mean rather than 1.5 as in the traditional method. Users can set their desired limits when dealing with cases where more precision is required such as 6 sigma processes.

3.3 A Proposed Framework for System/Enterprise-Level Monitoring

The following framework is the TUFF method for enterprise-level system monitoring. Note that existing SPC methods such as PCA are usually limited to approximately 50 variables at once or need to solve large variance-covariance matrix (Hotelling T^2). There are also possibilities of using machine learning algorithms, artificial neural networks or other methods, which are usually computer-time consuming. The TUFF method can be used as a tool to reduce the size of the dataset before it is analyzed with machine learning algorithms.

An enterprise-level system monitoring method should possess the following properties:

1. Able to detect there is a change in the system of interest
2. Able to detect changes in both QCs and their corresponding process parameters
3. Able to detect the location of the change
4. Able to detect the timing of the change
5. Able to work with different types of data: continuous, profile and binary data
6. Able to be easily modeled (or model-free) and implemented
7. Able to be scaled up for big data applications

Enterprise-level system monitoring is assumed to aid different levels of managers. The department head is interested if all machines in the department are performing as expected or on the same level as previously, the factory manager is interested if all the departments are performing on needed levels, the area manager is interested if all the factories are performing on the needed level, etc. Therefore, the monitoring does not focus on finding the cause of the change, but simply on detecting it. The machine operator or process engineer is also expected to correct any deviations promptly. The monitoring system allows detecting changes on the level that a particular manager or investigator is interested in and gives a starting point for further analysis.

With the growth of sensors in a system, all the process parameters and quality characteristics can be monitored and treated as variables. The TUFF method takes into account all variables and monitors all of them. It is data-driven. No knowledge of distribution is needed as the basis of the TUFF method is to compare results of the period currently under investigation to results from previous periods. Two main strategies are followed to maintain the scale of the problem formulated. First, only the variables that exhibit changes according to the pre-control rules are flagged for further investigation. Second, the values of all variables in different data types are transformed into a standard scale so that changes are easy to identify with familiar units.

The TUFF method is a two-layer method: the first layer is the bottom-level entities such as individual machine performance and the second layer is the aggregation of results over all bottom-level entities such as machines at the department level or factory level. The same application can be applied to the department and factory level and so on. The machine-level layer looks at all of the variables connected to the particular machine (process parameters, quality characteristics) and categorizes them based on “distance from target”. This can be used with both continuous data and profile data. Then the output of the variables is categorized. Summarization is done by using the group control chart idea. The base here is to look at all the variables for each timeframe, that is, samples are taken (10 per second, every second, or every minute, etc.), and then the worst outcome is chosen to represent the status of the machine at that timepoint.

The higher level generates summary statistic values over time periods and machine groups that the user has identified. These statistic values are then compared with similar values from previous time periods and conclusions are made based on the results of the comparison.

When change is detected on a higher level, indexing is used to change the resolution of the report to identify the location of the change (machine, department) and the timing of the change. More precise analysis with a much smaller dataset can then be started to identify the cause of the change.

3.3.1 Formulation of the TUFF method

Consider a small production system of *d* departments each containing *m* machines. The machines generate raw data on quality characteristics and process parameters. Each machine generates *v* variables. The number of machines in each department and the number of variables in each machine do not have to be the same across the factory. Assume that the system produces two different products that have different process paths and different target values for each quality characteristic and process parameter.

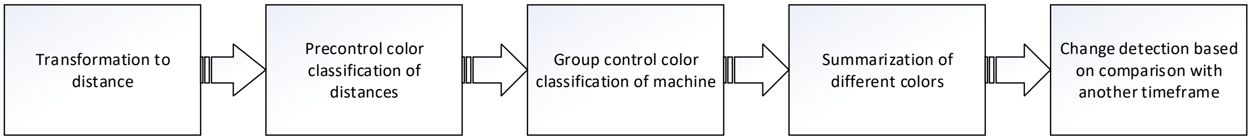

Transformation to distance
Precontrol color classification of distances
Group control color classification of machine
Summarization of different colors
Change detection based on comparison with another timeframe

Figure 7. The flowchart of the TUFF method

The TUFF method consists of five steps (Figure 7). The main purpose is to identify if there has been a change, when the change occurred and where the change occurred. This is achieved by using the indexing method and different resolutions of time and space. The indexing method uses the timestamps associated with each measurement to assign time and location identification (machine ID, department ID and factory ID) to all measurements.

The first step is to transform all the raw data into a distance measure from the target. The assumptions are that all the process parameters are continuous variables and quality characteristics have target values. The calculation of the target is as follows

$$d_{jklt} = \left. |x_{jkltp} - Ta_{jklp}| \middle/ s_{jklp} \right. \tag{2}$$

where *d* – distance

x – raw continuous data measurement

Ta –target value

s – target standard deviation

j – index for departments in each factory (*j*=1,2,…, *d*)

k- index for machines in each department (*k*=1,2,….., *m*)

l – index for variables in each machine (*l*=1,2,…., *v*)

t – index for the count of measurement (*t*=1,2,….., *n*)

p – product identifier

The second step is to classify each distance from the measurement using the pre-control idea. The classification assigns one of four colors to each distance-based as follows:

$$c_{jklt} = \begin{cases} green, if the\ d_{jklt}\ is\ within\ one\ target\ standard\ deviation\ from\ 0\ for\ product \\ yellow, if\ the\ d_{jklt}\ is\ between\ 1\ and\ 3\ target\ standard\ deviations\ from\ 0\ for\ product \\ red, if\ the\ d_{jklt} is\ more\ than\ 3\ times\ the\ target\ standard\ deviation\ or\ machine\ is\ down \\ white, if\ the\ machine\ is\ scheduled\ to\ be\ down \end{cases}$$

The setup of the limits can be done using different principles. In the presented case higher value was assigned to more precision, so anything under 1 sigma shift was rewarded with green in the TUFF method. The classification limits could also be set up based on the original precontol charts, where the “acceptable” area is divided equally between green and yellow; and the division is at 1.5 times the standard deviation.

The third step uses the group control chart idea to offset within machine variable dependency. The worst classification will be reported. The machine will be assigned into a category for each sample row *ti* based on the following rules

Green count $$ti_{jkt}^{g} = \begin{cases} 1, when\ all\ the\ c_{jklt} = green \\ 0, otherwise \end{cases}$$

Yellow count $$ti_{jkt}^{y} = \begin{cases} 1, when\ at\ least\ one\ of\ the\ c_{jklt} = \ yellow\ and\ none\ in\ red \\ 0, otherwise \end{cases}$$

Red count $$ti_{jkt}^{r} = \begin{cases} 1, when\ at\ least\ one\ of\ the\ c_{jklt} = \ red\ or\ unexpected\ stop \\ 0, otherwise \end{cases}$$

White count $$ti_{jkt}^{w} = \begin{cases} 1, when\ the\ machine\ has\ a\ scheduled\ stop \\ 0, otherwise \end{cases}$$

The overall count $$ti_{jkt} = \ ti_{jkt}^{g} + ti_{jkt}^{y} + ti_{jkt}^{r} + ti_{jkt}^{w}$$

The fourth step is to summarize all color counts and generate a statistic that is used for comparison and detection of changes. The summarization is completed as follows:

All the available counts:

$$T = \sum_{t=1}^{n} \sum_{j=1}^{n_j} \sum_{k=1}^{n_k} ti_{jkt} \tag{3}$$

Count of each color category:

$$T^{c} = \sum_{t=1}^{n} \sum_{j=1}^{n_j} \sum_{k=1}^{n_k} ti_{jkt}^{c} \tag{4}$$

where *c* is the index of the color, either *g* for green, *y* for yellow, *r* for red, or *w* for white

The statistic is the ratio between the counts of each color category divided by the overall available count of samples. Based on the calculations following table is created for each period under investigation

Green	${T^g}/{T}$
Yellow	${T^y}/{T}$
Red	${T^r}/{T}$
White	${T^w}/{T}$

Step 5 – the final step is to detect changes. The previous four steps are used on the period under investigation and also on two other periods, for example, average values of the ratios over year-to-date and average values of the ratios over historical data from the same periods over the previous years. When the change is detected (i.e. the percentages of colors are different between the periods), the location and the period of change can be pinpointed by segmenting the data further and running the same method. For example, if the time frame under investigation was a week, the segmentation would be to generate the same table for each day of that week and for each department of the system. The comparison would show when and where the change occurred.

Change Detection Criteria. The selection of the criteria depends on many aspects, such as the availability of historic data, the distributions of variables, etc. In the case of historic data, the user can choose the timeframe that is known to be acceptable and generate acceptable thresholds based on those timeframes and compare the results with the results under investigation. This process is very similar to the Phase I operation of control charting. The examples in this chapter were generated by assuming a normal distribution for each variable. The results showed that the percentage of red was the best indication of change. Depending on if there were 10 or 20 variables grouped together in the third step, the threshold for change was 3.417% and 6.305% of red respectively. The process behind these suggestions can be found in the simulation study section of this chapter. In practice, these thresholds should come from the computation based on a historical data set with the considerations of both Type I and Type II errors.

3.3.2 Examples

The following samples are presented to show how the TUFF method works. The first example shows the step-by-step process of how to locate the change and the second example shows how to pin-point the timeframe of the change.

3.3.2.1 Example 1

Assume there is a small system of three departments that are producing two products. Each department has three machines that have six critical process variables each. The goal is to detect if the system is operating on the same level one day as it did on the previous day. The proposed system-wide monitoring framework is implemented for the assessment.

An example of one machine from one department of raw data and the target value is in Table 2. Process variables x1, ..., x6 are raw data for each of the six variables. TA1, ..., TA6 are the target values for each variable. Ten sample periods are shown in this table. The measurements are assumed to be recorded at the same time for each variable in this case for the sake of simplicity, but in real life that does not have to be the case. This will be addressed in the discussion section.

The distance for each variable is calculated based on the equation in the first step of the TUFF method. The results are listed in Table 3.

Table 2. Raw data example for one machine

Sample Period no	x1	TA1	x2	TA2	x3	TA3	x4	TA4	x5	TA5	x6	TA6
1	20.645	20.000	81.491	100.000	48.659	50.000	1079.103	1000.000	69.268	70.000	71.991	66.000
2	19.062	20.000	93.879	100.000	51.037	50.000	1039.937	1000.000	69.828	70.000	66.385	66.000
3	19.690	20.000	101.773	100.000	51.288	50.000	1021.800	1000.000	69.210	70.000	68.343	66.000
4	20.441	20.000	94.206	100.000	53.557	50.000	1060.937	1000.000	69.359	70.000	67.117	66.000
5	20.654	20.000	113.126	100.000	49.391	50.000	972.231	1000.000	68.941	70.000	65.596	66.000
6	20.403	20.000	87.734	100.000	49.528	50.000	1037.427	1000.000	71.952	70.000	62.247	66.000
7	20.475	20.000	93.849	100.000	49.716	50.000	933.001	1000.000	69.856	70.000	63.205	66.000
8	21.967	20.000	91.667	100.000	51.327	50.000	1010.222	1000.000	69.861	70.000	62.996	66.000
9	14.953	20.000	83.017	100.000	49.024	50.000	1039.758	1000.000	71.692	70.000	65.149	66.000
10	20.532	20.000	117.300	100.000	47.953	50.000	1070.887	1000.000	70.254	70.000	66.333	66.000

Table 3. The results of the distance calculation based on the first step of the TUFF method. d1, ..., d6 are the distance values for each variable

Sample Period no	d1	d2	d3	d4	d5	d6
1	0.645	18.509	1.341	79.103	0.732	5.991
2	0.938	6.121	1.037	39.937	0.172	0.385
3	0.310	1.773	1.288	21.800	0.790	2.343
4	0.441	5.794	3.557	60.937	0.641	1.117
5	0.654	13.126	0.609	27.769	1.059	0.404
6	0.403	12.266	0.472	37.427	1.952	3.753
7	0.475	6.151	0.284	66.999	0.144	2.795
8	1.967	8.333	1.327	10.222	0.139	3.004
9	5.047	16.983	0.976	39.758	1.692	0.851
10	0.532	17.300	2.047	70.887	0.254	0.333

The distances are classified based on the logic presented in the second step of the TUFF method using the pre-control methodology. In this example case, all the values for all process variables for machine 1 are within one standard deviation from the target except for variable 1 at row 9 which is 3 or more standard deviations from the target. Table 4(a) shows the color classifications c1, c2, …, c6 for corresponding process variables of machine1.

In the next step -- the third step of the method, each machine is given an overall color classification based on the group control chart idea. The worst classification across all

Table 4. (a) Pre-control classification for Machine 1; (b) overall color code for Machine 1

Sample period no	c1	c2	c3	c4	c5	c6
1	🟩	🟩	🟩	🟩	🟩	🟩
2	🟩	🟩	🟩	🟩	🟩	🟩
3	🟩	🟩	🟩	🟩	🟩	🟩
4	🟩	🟩	🟩	🟩	🟩	🟩
5	🟩	🟩	🟩	🟩	🟩	🟩
6	🟩	🟩	🟩	🟩	🟩	🟩
7	🟩	🟩	🟩	🟩	🟩	🟩
8	🟩	🟩	🟩	🟩	🟩	🟩
9	🟥	🟩	🟩	🟩	🟩	🟩
10	🟩	🟩	🟩	🟩	🟩	🟩

(a)

ti
🟩
🟩
🟩
🟩
🟩
🟩
🟩
🟩
🟥
🟩

(b)

process variables of a machine is chosen as the performance for a machine. Table 4(b) shows the overall color classification for machine one for each sample.

Next, the overall available number of samples is calculated based on equation 3 in the fourth step of the TUFF method as well as the count of each color classification found in the sample list. In this case for Machine 1: $\sum ti_{jkt}^{g} = 9$ and $\sum ti_{jkt}^{r}$=1. The results for machine 1 are then summarized according to equations in step 4 shown in Figure 8. Note that Table 4 is the result of one machine while the overall system would generate Table 5.

Level	Statistic
🟥	10.0%
🟨	0.0%
🟩	90.0%
IDLE	0.0%

Figure 8. The final results in term of percentage of sample in each color class

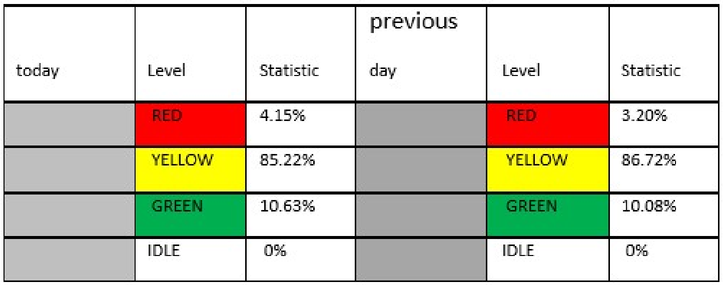
3.3.2.2 Example of detecting the time and location of the change

The TUFF method is also designed to identify the location and the time of changes by changing the resolution of the output. Let's assume there is a department with 10 machines that

Table 5. The system-wide look at the perfomance

Sample period no	Machine 1	Machine 2	Machine 3	Machine 4	Machine 5	Machine 6	Machine 7	Machine 8	Machine 9
1	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
2	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
3	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
4	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
5	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
6	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
7	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
8	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
9	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
10	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

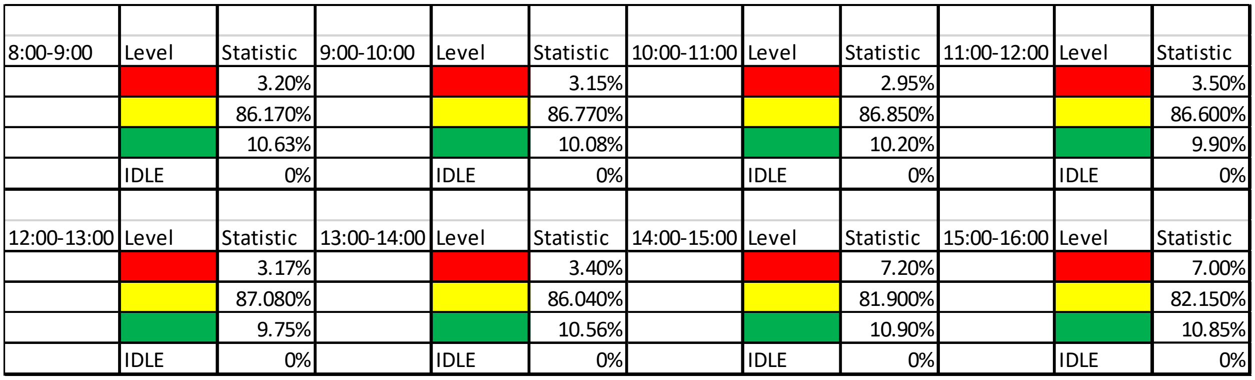
are producing different products. The goal is to determine if the department is producing on a similar level as the previous day. Over an 8 hour shift the data is collected, targets are deducted from the raw data, all the variables and time points are classified into color classes with the help of pre-control part of the TUFF method. Then the machines are assigned to color classes based on the group control part. All the data is summarized and the output for the department is reported in figures 9-11.



today	Level	Statistic	previous day	Level	Statistic
	RED	4.15%		RED	3.20%
	YELLOW	85.22%		YELLOW	86.72%
	GREEN	10.63%		GREEN	10.08%
	IDLE	0%		IDLE	0%

Figure 9. Comparison of two days of production of the department

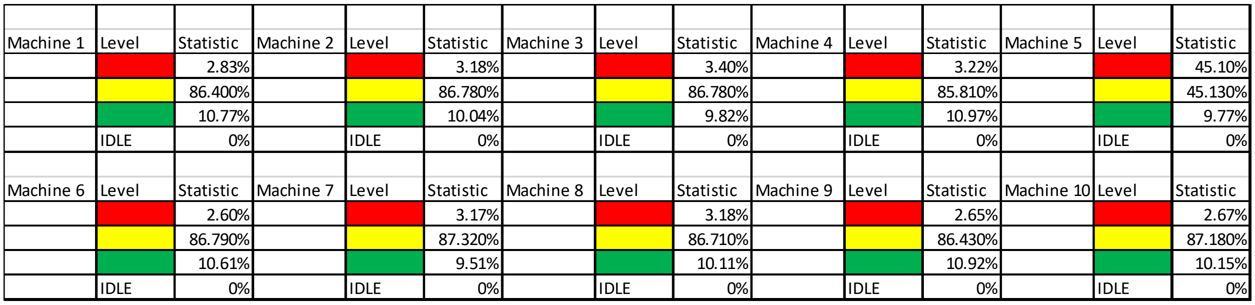
In figure 9, the red percentage has changed from 3.2% to 4.15 % which means more measurements were beyond 3 times the deviation from the target value. The recommendations for the decision criteria will be discussed under the simulation portion of this chapter. Obviously, there has been a change somewhere and some time. To determine when the change happened, the results are viewed in higher resolution. Specifically, the categorized machine output is divided into hourly blocks with the help of timestamps and presented in the series of outputs.



8:00-9:00	Level	Statistic	9:00-10:00	Level	Statistic	10:00-11:00	Level	Statistic	11:00-12:00	Level	Statistic
	🟥	3.20%		🟥	3.15%		🟥	2.95%		🟥	3.50%
	🟨	86.170%		🟨	86.770%		🟨	86.850%		🟨	86.600%
	🟩	10.63%		🟩	10.08%		🟩	10.20%		🟩	9.90%
	IDLE	0%		IDLE	0%		IDLE	0%		IDLE	0%
12:00-13:00	Level	Statistic	13:00-14:00	Level	Statistic	14:00-15:00	Level	Statistic	15:00-16:00	Level	Statistic
	🟥	3.17%		🟥	3.40%		🟥	7.20%		🟥	7.00%
	🟨	87.080%		🟨	86.040%		🟨	81.900%		🟨	82.150%
	🟩	9.75%		🟩	10.56%		🟩	10.90%		🟩	10.85%
	IDLE	0%		IDLE	0%		IDLE	0%		IDLE	0%

Figure 10. Hour-to hour production of the whole department on higher resolution

In figure 10, it is clear that the change happened somewhere between 14:00 and 15:00 in the department. The next step would be to identify the machine/machines that are responsible for the change. Since the time of the change is known, the time is limited only to that slot and the categorized machine output is summarized over that time slot. The machines are not summarized into department level to identify the culprit. The results are presented in a series of outputs.



Machine 1	Level	Statistic	Machine 2	Level	Statistic	Machine 3	Level	Statistic	Machine 4	Level	Statistic	Machine 5	Level	Statistic
	🟥	2.83%		🟥	3.18%		🟥	3.40%		🟥	3.22%		🟥	45.10%
	🟨	86.400%		🟨	86.780%		🟨	86.780%		🟨	85.810%		🟨	45.130%
	🟩	10.77%		🟩	10.04%		🟩	9.82%		🟩	10.97%		🟩	9.77%
	IDLE	0%		IDLE	0%		IDLE	0%		IDLE	0%		IDLE	0%
Machine 6	Level	Statistic	Machine 7	Level	Statistic	Machine 8	Level	Statistic	Machine 9	Level	Statistic	Machine 10	Level	Statistic
	🟥	2.60%		🟥	3.17%		🟥	3.18%		🟥	2.65%		🟥	2.67%
	🟨	86.790%		🟨	87.320%		🟨	86.710%		🟨	86.430%		🟨	87.180%
	🟩	10.61%		🟩	9.51%		🟩	10.11%		🟩	10.92%		🟩	10.15%
	IDLE	0%		IDLE	0%		IDLE	0%		IDLE	0%		IDLE	0%

Figure 11. Time 14:00-15:00 based on individual machines on higher resolution

In figure 11, the results show that machine 5 has started to produce higher percentages of signals that are classified as red and now a more thorough analysis of the reasons behind that can start with a much smaller time window.

3.4 Simulation Studies

Three simulation studies were carried out to identify the different characteristics of the TUFF method. The goal of the first study is to determine what threshold should be used for making a decision if there has been a change. The second study aims to determine the maximum number of variables that can be grouped together in the third step of the method. Finally, the third study determines how sensitive this method is. All the simulation studies were carried out with an assumption that the data was normally distributed for simplicity and demonstration purposes. In real-life applications, the distribution does not have to be predetermined and the decisions would be made based on the comparison with historic data.

3.4.1 Determining threshold for Decision Making

Let's assume there is a machine with 10 variables. Each variable measures a different parameter of the machine. There are 3600 data points in each variable. For simplicity, all the variables have been normalized so that the unchanged variable would be normally distributed with N(0,1). In order to determine the threshold for change detection and Type II errors for different scenarios, all but one variable were left at N(0,1) and one was changed according to a series of parameter changes. The changes introduced were mean shifts of 0, 0.5, 1, 2 and 3 while standard deviations' changes of 1, 1.5, 2 and 3, and the combination of both. After each iteration of the simulation, the TUFF method was applied to the new dataset and the results of the color percentages were recorded. Each combination was repeated 10,000 times. One of the additional findings of this simulation was that in this case, the red percentage was the best indicator of change. The green and yellow percentages were more random and together mirroring the red color percentage. The results of the red color percentages on the machine level are presented in the following box-plots.

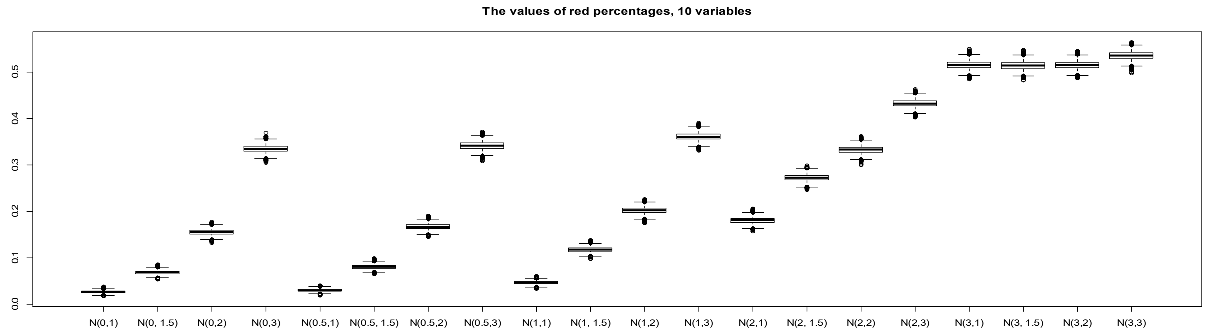


Figure 12. Boxplots of red zone percentages from various out-of-control situations at the machine level

Figure 12 shows all the scenarios on the same graph. As can be seen, most of the scenarios have no overlapped results with the unchanged scenario, which is the process is at N(0,1). The only overlapping scenarios to N(0,1) are N(0.5,1) and N(1,1). Figure 13 shows a more detailed look at the results of those three scenarios.

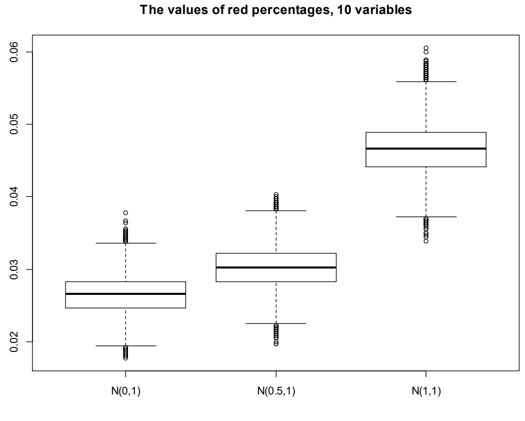


Figure 13. Box-plots of red zone percentages of processes N(0,1), N(0.5,1), and N(1,1)

A more detailed look reveals that scenario N(1,1) has much fewer overlaps than N(0.5,1). The smaller change results are much closer to the unchanged variable results. To determine what threshold should be used in the decision making, the overall Type I error will be set at 0.0027. In the case of 10 variables, the threshold of red color percentage would be 3.417%. Anything over that would be considered as a changed variable.

Table 6 shows the Type II errors of different scenarios when considering the set threshold. The results are drawn from the 10 000 repetition results.

Table 6. Type II errors for different scenarios

St.Dev \ Mean	0	0.5	1	2	3
1	■	0.914	0.0001	0	0
1.5	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0

The table confirms the previous finding that only N(0.5, 1) and N(1,1) have overlaps with the unchanged variable and therefore are also the only ones with Type II error. Given the overall type I error at 0.0027, the TUFF monitoring method is not capable of detecting a very small process mean shift.

The second part of this simulation study was to determine if the threshold will change when the number of variables in each machine is different. The simulation steps and scenario parameters stayed the same, only this time the machine was assumed to have 20 variables with 3600 data points each and one of those is responsible for change based on the scenario. Each scenario was repeated 10,000 times. Similarly to the first part, the red color percentage was found to be the best indicator of change. The overall results of the study are shown in the box-plot of Figure 14.

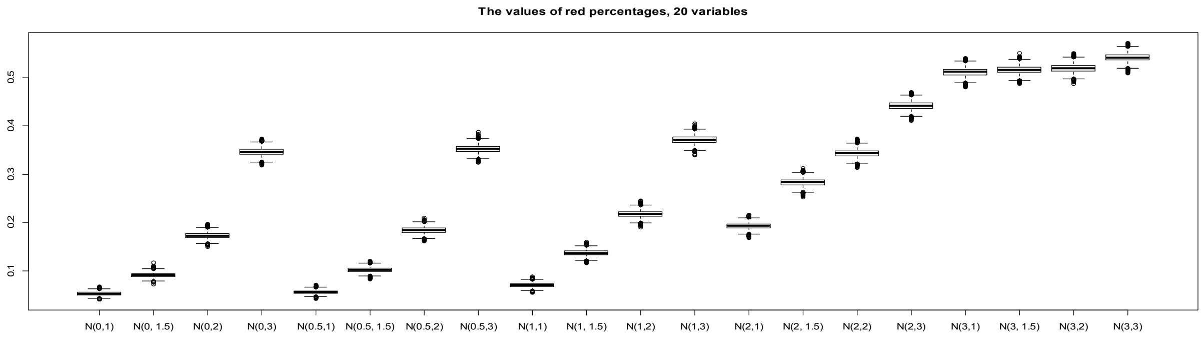


Figure 14. Box-plots of red zone percentages from various out-of-control situations with 20 variables

Once again only the small mean shifts of N(0.5, 1) and N(1, 1) seem to have overlapping parts with unchanged variables. More detailed boxplot of the three is provided in Figure 15

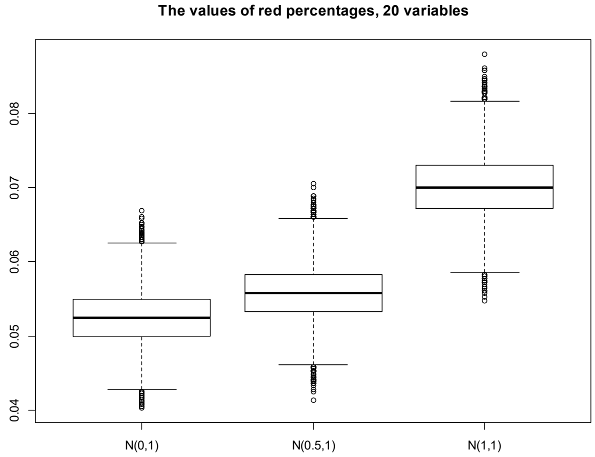


Figure 15. Box-plot of red zone percentages of processes N(0,1), N(0.5,1) and N(1,1)

The threshold, in this case, is again based on the Type I error of 0.0027. When the machine uses 20 variables and is assumed to be normally distributed, the red percentage threshold for change is 6.305%. Additionally, there were two revelations. First, the unchanged variable produces much more “red” colored signals which can be explained with Bonferroni curse of dimensions. The threshold is almost two times larger than on the 10 variable cases. The second is that there is less difference or more overlapping between scenarios.

Table 7. Type II errors of different scenarios with 20 variables

St.Dev \ Mean	0	0.5	1	2	3
1	■	0.9664	0.0411	0	0
1.5	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0

Table 7 shows the Type II errors of the scenarios based on the 10,000 repetitions and the threshold set previously at 6.305%. As can be seen, the Type II error for both N(0.5,1) and N(1,1) has increased. The larger changes are detected with 100% accuracy, while the small mean shift is virtually undetectable.

3.4.2 Determining the max number of variables in one machine

This simulation study was inspired by the first study. The purpose of this study was to determine how many variables can be grouped together in the group control part of the TUFF method without losing too much detection power.

For this simulation, the statistic monitored was the percentage of "red" classifications as the previous study indicated to be the best indicator. The scenarios were based on 30-100 variables in 10 variable steps. Each number of variables was simulated with variable changing N(0,1), N(0.5,1), N(1,1), N(0, 1.5), N(0.5,1.5) and N(1,1.5) because the previous study showed that smaller changes are more prone to be overlapping with the unchanged variable. For 30, 40, 50 and 60 variables, each combination of mean and standard deviation was recorded 10,000 times. For 70, 80, 90 and 100 variables, each combination was recorded 1,000 times due to the high demand in computational time.

Tables 8 and 9 show the results for this simulation. Both Type I and Type II errors are considered to reach a balance performance in setting the threshold values. Table 8 shows the threshold values in the second column Type II errors of a different number of variables when the Type I error is set at 0.0027 while Table 9 shows the results of Type I errors when the Type II error is set at 0.05.

Table 8. Type II errors of different combinations of scenarios and number of variables with Type I error of 0.0027

No of variable	Threshold	N(0,1.5)	N(0.5,1)	N(0.5,1.5)	N(1,1)	N(1,1.5)
30	0.0905	0	0.9735	0	0.0496	0
40	0.1175	0.0002	0.9878	0	0.3711	0
50	0.1417	0	0.9821	0	0.2756	0
60	0.164	0.0085	0.992	0	0.6904	0
70	0.193	0.008	0.993	0	0.574	0
80	0.222	0.03	0.989	0	0.776	0
90	0.24	0.017	0.99	0	0.571	0
100	0.255	0.036	0.986	0.001	0.742	0

Table 9. Type II errors of different combinations of scenarios and numbers of variables with Type I error of 0.05

No of var	Threshold	N(0,1.5)	N(0.5,1)	N(0.5,1.5)	N(1,1)	N(1,1.5)
30	0.0852	0	0.7796	0	0.0021	0
40	0.1113	0	0.8445	0	0.0659	0
50	0.136	0	0.8556	0	0.0568	0
60	0.1565	0.0002	0.888	0	0.2274	0
70	0.1844	0	0.866	0	0.137	0
80	0.2119	0.001	0.897	0	0.217	0
90	0.2315	0	0.863	0	0.261	0
100	0.248	0.003	0.907	0	0.353	0

As shown in Tables 8 and 9, the more variables are used during the group control part of the TUFF method, the higher the threshold is for keeping the type I error at 0.0027 and 0.05 respectively. Again, this is explained by the Bonferroni curse of dimensions. What also can be seen is that in each combination there is a trend for Type II error to grow when the number of variables increases. The more variables are used, the less obvious the difference between unchanged and changed variable is. For the recommendation of how many variables could be used in the group control part of the proposal, decision criteria must be established. The proposal is to base the decision on the N(1,1) scenario because in the previous simulation study that scenario showed a small Type II error and it is an important change that needs to be captured. The second side of the decision criteria is where to draw the line of what is acceptable. The recommendation here would be to use 5% of Type II error because, at that level, a lot of changes are still detected. Based on the decision criteria, the recommended maximum number of variables to be used in the group control part of the algorithm would be 30 variables for the TUFF method to be effective with the consideration of both Type I and Type II errors.

3.4.3 Determining the sensitivity of the TUFF method

This simulation study on the sensitivity of the TUFF method aims to determine how sensitive it is by leveraging the knowledge gained from the previous two studies. Assume that there is a factory with 10 departments. Each department consists of 10 machines and each machine monitors 10 variables. Overall, there are 1000 variables to monitor. Each variable has 3600 data points or rows. The simulation is done with a small mean shift of N(1,1) because that was used in the number of variables simulation study decision criteria. In the first scenario, 1% of random variables changed by N(1,1). This means 10 random variables out of 1000. The selection was made with a uniformly distributed random number generator. Each variable had an equal chance of being selected. Also, the changed variables could have appeared on the same machine. The selected variables changed from iteration to iteration. After each dataset was generated, the TUFF method was applied. First, the pre-control part of the algorithm classified each data point into either red, yellow or green color. Then the group control part grabbed the worst classification for each row and assigned that classification to the machine for that row. The next steps were to sum up the results on the machine, department and factory level and report the percentages for each color. This was repeated 1000 times. The red color percentage was used for the identification statistic. The same process was carried out with 5% of variables (50 random variables out of 1000) and 10% of variables (100 random variables out of 1000) experiencing changes.

Two thresholds were used for alpha: 3.417% for 0.0027 and 3.111% for 0.05. The results can be found in Table 10. The Type I error reported in this table represents the amount of machines/departments/factory that were labeled as changed but in fact, they did not have the changed variable in them. On the other hand, the Type II error represent machines/departments/factory that were labeled unchanged but had a changed component in them.

Table 10 shows that the TUFF method is capable of detecting changes at the machine level well. The method failed to detect changes in department and factory level when Type I is 0.0027 in the 1% case. In the case of 5% variables change case, the departmental detection is getting much better. In the 10% case, the factory level and the department level are almost always labeled as changed. The revelation is that on any managerial level the change is brought to attention if there is either a large change or a lot of small changes.

3.5 Discussions

The TUFF method is capable of practical implementation in many different environments. The example given in the previous chapter was when all the samples were taken at the same time point. In most real-life cases that might not be possible. For example, the measurement of each part in height might happen 10 times a second, but the temperature is measured only every 2-3

Table 10. The sensitivity of the TUFF method

N(1,1)		alpha = 0.05		alpha=0.0027	
		Type I	Type II	Type I	Type II
1% variables changed	Factory (1)	0.0000	1.0000	0.0000	1.0000
	Department (10)	0.0000	0.6613	0.0000	0.9585
	Machine (100)	0.0612	0.0000	0.0047	0.0010
5% variables changed	Factory (1)	0.00000	0.00000	0.00000	0.00000
	Department (10)	0.00000	0.02536	0.00000	0.14874
	Machine (100)	0.00897	0.00002	0.00074	0.00052
10% variables changed	Factory (1)	0.0000	0.0000	0.0000	0.0000
	Department (10)	0.0000	0.0001	0.0000	0.0009
	Machine (100)	0.0084	0.0000	0.0006	0.0004

seconds. The recommendation would be to assign the less frequent measurement to each row of the sample that is between the less frequent sampling times. In cases where there is so much data that the traditional computational tool cannot handle the data, the use of big data applications might be useful. Another facet of the TUFF method worth mentioning is that it is not an in-time monitoring method where the sample in hand is the item under investigation. The method can be used as near-time since the focus is on the results from a period of time.

3.5.1 Implementation in an Enterprise environment

The TUFF method is scalable and can be adapted to a system ranging from a factory to a supply chain. The recommendation of the implementation of the method in a real-world enterprise would consist of the following steps:

1. Determine the target value for each variable in the system for each product produced in the facility to calculate the distances from the target:
    a. If the targets can be set by requirements, those could be used

   b. If not, the historical data must be used to calculate the targets

   c. Not all process variables are continuous. Some may be discrete and the others may be in the form of a profile. The method introduced in section 3.1 is based on the fraction of time as shown in equations (2), (3), (4) and (5) computed by the number of samples falling into each color zone. Discrete variables can be easily accommodated using the same set of equations. However, profile variables require additional procedures to convert. It will be introduced in another study.

2. Determine the standard deviations to be used in the pre-control part of the algorithm

   a. If the standard deviations are set by requirements, those should be used

   b. If not, the historical data of "good" products should be used. This is often referred to as the Phase I of control charting.

3. Group variables together for the group control part of the TUFF method:

   a. If there are more than 30 variables in the machine, then the variables should be separated and grouped together to form subgroups. For example, if the machine has 10 variables that are connected to temperature measurements, 20 connected to speeds and 10 connected to the shape of the product, then 40 variables together will not produce good output. However, grouping them according to function within the machine would give the desired result.

   b. While it would be advisable to keep the variable groups similar in size, it is not required

4. Run the TUFF method on the full set of historic data. This step might help with solving problems in the future. The data should be analyzed for both "good" parts as well as "bad" parts. If in the future a certain set of changes has been detected, then comparing it to

historical data and finding a similar set of changes with the current issue might help to solve the problem and also predict the quality of the products or even predict an imminent failure.

5. Start the real-time monitoring using the target values from step 1 and deviations from step 2.
6. For reporting purposes, the recommendation would be to have three output charts side-by-side that show the current period, averages of year-to-date and averages of historical data from a similar timeframe (for example all the Julys in history). This might help understand the trends and also helps in the case of autocorrelated data.
7. Different levels of management are assumed to be interested in different levels of data. The department head is interested in the machine level report in their department. The factory manager is interested in the department-level report and CEO is interested in the factory level report. Each level can dig deeper if more precise analysis is needed.
8. Prior knowledge of the distributions of the variables is not required, because the decisions are based on the comparison with the historical data that can be chosen on the basis of "good" products.

The addition of new machines is not an issue, because all the statistics are generated based on all the available time and they are a percentage of that. New machines add to the total time.

3.5.2 Implementation in Big Data environment

MapReduce is a framework for executing highly parallelizable and distributable algorithms across large datasets using hundreds or thousands of commodity computers (Lublinsky *et al*, 2013). A MapReduce algorithm does parts of the calculations in the server that the data segment is stored in parallel (hence the name “parallel computing”).

MapReduce consists of two procedures that the user must develop: mapping and reducing. The system manages the parallel execution, coordination of tasks that execute mapping or reducing and also deals with possible failures of some of the task execution. In the mapping procedure, the data segment in each server is split, sorted and filtered. If needed, other calculations are also carried out. Users must define two critical parameters that are used as the input and output of each server: key and value. The key is the identification parameter that depends on the goal of the algorithm and the value is the output of the segment in that server. All the key-value pairs are collected by the master controller and divided among all Reduce tasks in a way that all the pairs with same key end up in the same Reduce task (Rajaraman & Ullman, 2012)

In the reducing procedure, the outputs of the mapping procedures are shuffled and sorted based on the key defined in the mapper and then reduced by combining the values defined previously in some manner defined by the user.

The TUFF method could be turned into a MapReduce function using the following logic. The assumption is that all the targets and standard deviations are stored in the top-level computer so they can be accessed by the program at any time. The data is assumed to have a timestamp and variable identification for each measurement. In the mapper function, the distances are calculated, colors are assigned, the machine level is assigned. The output is a timestamp, machine ID and color assignment.

In the reducer function, the summarization of the colors is made based on the level of interest of the user. This is where the resolution of the report is set. The output would be the traffic light for the timeframe and location under investigation. Details of the MapReduce algorithm applied to the TUFF method will be reported in another study.

3.6 Conclusions

An enterprise-level monitoring system called Technique of Uniformally Formatted Frequencies (TUFF) is proposed. All the raw data is transformed into distances from the target. The distances are then classified into colored groups with the help of the pre-control chart idea. All the variables associated with a machine are then used to classify the machine into a color class according to the group control chart idea. The resulting counts are turned into percentages of time and that statistic is used to determine if the process is changed or not. The simulation studies show that the TUFF method is capable of handling 1000 variables per department and produce usable results with different scenarios. A recommendation for implementation of it in real-life situations is proposed as well as a recommendation for big data application with the help of the MapReduce method.

The TUFF method has the following characteristics:

1. It is capable of detecting changes, also identify the spatial and temporal space of the change
2. It does not use a covariance matrix, which makes the calculation much easier
3. It is easily scalable from a few to 1000 process variables
4. It can be extended in big data format
5. It does not require previous knowledge of distributions

Future studies include the implementation of the method on different types of data, such as profile data and binary data. In addition, the TUFF method may be able to be integrated into maintenance planning - usually, a topic traditionally studied from the field of reliability. Additionally, the method may be applied to supply chain applications.

Chapter 4 - Monitoring Profile Data in a System-Wide Monitoring Framework in the Big Data Era

4.1 Introduction

The continuous development of sensors and the adoption of data storage solutions such as the Internet of Things (Bruner, 2013) has allowed the collection of a large amount of data in production facilities. The collected data could be categorized as big data, which is defined as data collection so large that it can not be analyzed with traditional methods (Megahed and Jones-Farmer, 2015). It is usually described to have at least one of the three Vs: Volume, Variety, and Velocity (Megahed and Jones- Farmer, 2015). The data collected in production facilities can be considered as big data because it is usually in large volume and have many varieties. Production data may come from different sources such as process parameters and quality characteristics with different formats in discrete (or categorical), continuous, profile, or images.

As the amount of raw data collected rises, the monitoring of the production systems becomes more difficult. Traditional statistical process monitoring methods such as univariate or multivariate control charts (Montgomery, 2012) aims to monitor local areas in an entire system. Additionally, traditional methods usually use sample sets of quality characteristics with the assumption that if the quality characteristics are good, the process parameters must be good as well and that the current stage is independent of previous stages.

Traditional process monitoring methods are not designed to take the full advantage of new opportunities in the big data era. For example, it may be possible for defect prevention through the monitoring of both process parameters and product quality characteristics. Analyzing full data sets may also reveal hidden correlations between different stages and timeframes. An analogy is heavy snowfall in the upstream of the Mississippi River in the winter months may cause flooding in downstream cities in the spring. While this example is obvious, it may be hidden in a hierarchical production system where some of the rogue parameters in the upstream stages may cause production or quality issues downstream. In summary, alternative methods are needed to monitor the data abundant production systems to elevate understanding of the system and allow for faster reaction to problems. We have developed a system-wide monitoring framework called Technique of Uniformally Formatted Frequencies (TUFF) to tackle this challenge in our previous work (Chapter 3). However, that work focuses on a continuous variable in the big data environment. Other types of process parameters such as profiles over time or space have not been studied. In this study process parameters characterized by profiles are integrated into the TUFF framework.

The purpose of this study is to provide a method that can monitor profiles as a part of a smart system monitoring framework for a larger production system. The TUFF method is not in direct competition with current profile monitoring methods which focus on monitoring one or more profiles in real-time (Chang *et al.,* 2014; Zou *et al.*, 2008; Chou *et al.*, 2014; Chou *et al.*, 2020, etc.). These real-time methods are still needed to monitor the profiles on the local level to establish statistics that the proposed method uses to generate system-wide results. The method is complementary to the real-time methods as a managing tool for summarizing the system's performance, identifying the timeframe and space of change and monitoring the process parameters to prevent quality deterioration. For example, a decline in the yield level in a semiconductor production line may trigger an investigation into causes. Thousands of variables need to be examined to find the causes for dropping yield. The analysis would be much faster if the parameters and characteristics that have changed can be narrowed down. The TUFF method focuses on the system-wide performance of the profile data in near time which means that the method analyses data over some time period, so that all the data from that time period is readily available. The same approach can also be applied to process parameters of all types in addition to profile data.

Assume that minor process deviations were not detected on real-time control charts. However, combinations of these deviations in various stages in the process may cause problems in overall product quality. The TUFF method would need to be able to detect changes, indicate the timing and location of the change, and easy to model and implement. Finally, the proposed method should be scalable for big data applications.

4.2 Literature review

A lot of manufacturing processes contain profiles to characterize some of their processes' performance as well as quality characteristics. Therefore, profile monitoring has drawn attention over the last 10-15 years (Woodall, 2007; Maleki *et al*., 2018). A profile is defined as a relationship between a response variable and the explanatory variable(s) (Woodall, 2007). An explanatory variable is usually either time or space.

Woodall (2007) reviewed various profile monitoring methods in quality control prior to 2007. He reviewed over 50 articles and discussed the use of simple linear regression, multiple and polynomial regression, nonlinear regression, mixed models and wavelets. More recently, Maleki *et al.* (2018) did an overview of papers published between 2008 and 2018 on profile monitoring and classified them based on the characteristics of the profiles and methods, such as monitoring single and multiple linear profiles, non-linear profiles, etc. A total of 195 articles were identified that proposed solutions to different aspects of profile monitoring. The main findings were that considerable attention has been given to different areas of profile monitoring over the last decade, most of the work has been focused on statistical design of the profile monitoring control charts, the most common profile regression model is a simple linear model and that most studies have been devoted to univariate response variable case. The most popular methods for single profile monitoring over the last few years are different variations of Exponentially Weighted Moving Average (EWMA). Zou *et al*. (2008) proposed a control scheme that could monitor both linear and nonlinear regression models by integrating the EWMA with the Generalized Likelihood Ratio (GLR) test based on local linear regressions. In a different study, Abdella *et al.* (2016) considered the effectiveness of double EWMA and double multivariate EWMA statistics in the multivariate statistical process control (SPC) applications and extended these statistics to Phase II polynomial profile monitoring. Fasso *et al.* (2016) introduced two types of functional EWMA control charts which differed for the stopping rule rationale. The functional data analysis uses observed functional data as a single object rather than a sequence of single observations. The focus was on using Multivariate EWMA control charts as functional EWMA on the random effects of a mixed linear model. In another study, Huwang *et al.* (2016) proposed two Phase II Multivariate EWMA-type of control charts based on the observed and in-control Fisher information for monitoring profiles that can be characterized by proportional odds models where the response variable was both categorical and ordinal. Chiang *et al.* (2017) used a multivariate EWMA chart for a simple linear process in the presence of within-profile autocorrelation. Another study by Abbas *et al.* (2017) introduced the Bayesian control charting structure for linear profile monitoring under phase II using double EWMA charts. Yang *et al.* (2017) proposed a kernel-based control scheme that integrated the multivariate EWMA procedure with the dynamic probability control limits. Closely related to the

EWMA methods is the Cumulative Sum (CUSUM) methods. Zhang *et al.* (2017) proposed two-sided Cumulative Sum (CUSUM) schemes with two separate or one single statistic to detect small shifts in pre-specified changes.

The second most popular method is wavelets. Chang and Yamada (2010) studied the monitoring of non-linear profiles using wavelet filtering and B-Spline approximation. Shahiari *et al.* (2016) proposed the use of estimators insensitive to outlying samples in monitoring complicated profiles using wavelet transformation. The estimators were based on the clustering of the estimated wavelet coefficients and a type S-estimators. In a different study, Koosha *et al.* (2017) applied image data in the SPC context using a nonparametric profile monitoring approach based on wavelet transformation for feature extraction.

Other methods have been proposed. Zeng *et al.* (2014) studied Phase I monitoring of profile data in non-normality assumption cases. These authors used independent component analysis to transform multivariate coefficient estimates to independent univariate data and then used univariate nonparametric control charts to detect changes. Chang *et al.* (2014) proposed an SPC framework to detect potential changes in wave profiles before the entire information on the profile of interest was fully available by converting each wave profile using the exponential-decayed function as cutting line into a statistic that could then be monitored with a univariate control chart. In another study, Shang *et al.* (2016) looked at change point detection with binary data profiles and random predictors using a logistic model.

Charkhi *et al.* (2016) proposed two methods to calculate Process Capability Indices (PCI) for logistic regression profile. In one method the PCI was calculated for certain levels, then overall PCI was determined for each level by calculating the percentage of nonconforming items and then the percentage was used to estimate the overall PCI. The second method calculated PCI for all levels under a continuous state. Shi *et al.* (2016) presented a manifold learning-based approach to identify and visualize the nature of profile-to-profile variation in a sample of profile data. In a different study, Wu *et al.* (2016) proposed a Bayesian Hierarchical Linear Modelling (HLM) with level -2 variance heterogeneity to build a relationship between profiles data, the explanatory variables and the microstructural parameters for quality inspection and control.

Zang *et al.* (2016) provided a framework for monitoring unaligned profiles based on robust Dynamic Time Wrapping (DTW) and penalized likelihood estimation by calculating a baseline profile from aligned in-control profiles, aligning the new profile to be monitored with the baseline profile and estimating the true mean of the aligned profile using several penalization-based methods for example fused LASSO. In another study, Koppel and Chang (2017) proposed the use of modified Sample Entropy value as an indicator of a possible change in the profile. Fan *et al.* (2017) chose the hyperbolic tangent function to model and monitor the aluminum alloy vacuum heat treatment process data. The monitoring was proposed to be executed by using two T^2 and three metric control charts. Ding *et al.* (2017) used proportional odds ratio models to monitor profiles with an ordinal response and random predictors. In another study, Awad *et al.* (2018) presented a data-driven methodology for fault detection of complex arrays such as structural systems via multivariate SPC and the use of artificial neural networks. Liu *et al.* (2018) proposed a mixed-effect profile monitoring scheme to achieve effective out-of-control profiles detection for spatial data with patent inter-cluster variations using the spatial Dirichlet process. Darbani and Shadman (2018) proposed to add a variable sampling interval to the generalized likelihood ratio control chart for monitoring linear profiles. The above-mentioned profile monitoring studies provide a good variety of approach to the single profile problem.

In a multiple profile case, several profiles are under investigation simultaneously. As is the case in the continuous variable cases, Principal Component Analysis (PCA) is one of the widely used methods. Lei *et al.* (2010) developed a feature selection and hierarchical classification method for missing part detection in the multi-operational forging process using data segmentation and Principal Component Analysis (PCA). Noorossana *et al.* (2010) extended the likelihood ratio, Wilk's lambda, T^2 and PCA to monitor multivariate multiple linear regression profiles in Phase I. They found that likelihood ratio and Wilk's ratio were the best in detecting sustained and outlier shifts. In a different study, Zou *et al.* (2012) proposed the use of variable-selection based multivariate control scheme that is capable of monitoring the regression coefficients and profile variations. Paynabar *et al.* (2013) proposed the use of uncorrelated multilinear principal component analysis for multichannel profiles that utilizes information from each profile channel as well as takes into consideration the interrelationship among different channels. In another study, Paynabar *et al.* (2016) adopted multidimensional functional principal component analysis for multivariate profile monitoring. Wang *et al.* (2018) proposed a thresholded multivariate PCA for multichannel profile monitoring. The method first reduces high-dimensional multichannel profile to a reasonable number of features using PCA and then uses soft-thresholding techniques to further select informative features under the out-of-control state. Chou *et al.* (2014) researched simultaneous process monitoring for multiple linear or non-linear profiles. Their approach was based on a multivariate EWMA control chart. Jahani *et al.* (2018) studied the modeling and monitoring of multivariate profiles using multivariate Gaussian process modeling.

As can be seen from this literature review, a myriad of methods have been proposed over the years. The problem with most of these methods is that the calculations are usually computationally demanding and not scalable for big data level problems. The other issue is that the vast majority of current methods deal with one profile at a time or one machine with a few different profiles that characterize the performance and then detect changes. A real-world production system might have thousands of variables some are characterized by profiles scattered all over a facility. To the best of our knowledge there is no study that links multiple profile variables scattered on different machines or locations in a hierarchical system. The proposed method aims to fill this void.

4.3 The monitoring of a profile with the TUFF method

Consider an hierarchical production system of h departments each containing m machines. The number of machines can vary from department to department. The system can be any production system where the product is moved from station to station and modified at each station. Assume that the machines contain process parameters only in the profile format. This assumption does not fully reflect a real-world situation but is used for simplicity in explaining the proposed method. The TUFF approach would generate a pre-control color distribution over certain predetermined timeframe similar to that presented previously in the continuous variable case. The timeframe can be hour, day, week etc. and depends on the specifics of the production system. During each timeframe o profiles is generated by any machine in any department. The number of profiles for each machine can vary from machine to machine similarly to number of machines in each department. At the department level, there are m machines, each is associated with a color-distribution. On the factory level, each department would have its own color-distribution table. The purpose of this method is to detect changes in the entire production system over the time period so that more advanced analyses of fault identification can take place.

The TUFF method consists of two stages which are similar to Phase I and Phase II in traditional SPC control charts. Traditionally in Phase I, process data is gathered and analyzed to construct control limits. When users confirm that the control limits from Phase I represent in-control process performance, Phase II begins. In Phase II the incoming sample drawn from the process is compared to the established control limits in Phase I to monitor the behavior of the process. Similarly, the first stage of the TUFF method aims to establish standards leading to color-distributions for each department or machine. In the second stage, each new data point is compared to the color-distributions established in the first stage and ultimately decide whether the system has changed or not. The steps in both stages are very similar and the differences will be pointed out in the explanation of the proposed method.

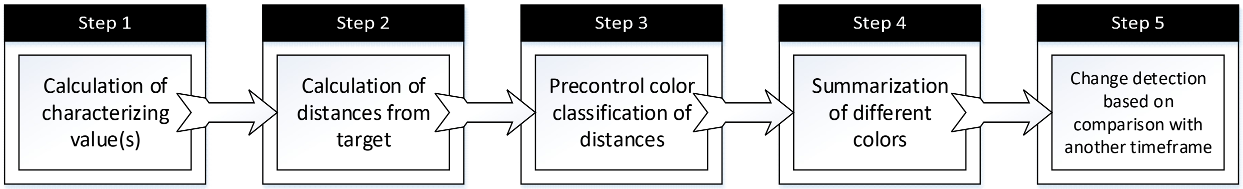

Step 1
Calculation of characterizing value(s)
Step 2
Calculation of distances from target
Step 3
Precontrol color classification of distances
Step 4
Summarization of different colors
Step 5
Change detection based on comparison with another timeframe

Figure 16. The Flowchart of the TUFF method

Each stage consist of five steps (Figure 16). The purpose of the method is to identify: a) if there have been changes, b) when the changes occur, and c) where the changes occur. This is achieved by using an indexing method of different resolutions of time and space. The indexing method uses the timestamps associated with each measurement to assign time and location identification using machine ID, department ID, and factory ID to identify all measurements.

The first step is to acquire a statistical measurement from the profile. As was stated before, the TUFF method is not competing with the methods presented in the literature review. Any of the traditional methods that create a statistic that helps to identify changes can be used. In cases of simple profiles, a piecewise regression could be beneficial to determine change points in the profile. In more elaborate profiles, b-spline regression or wavelet analysis could be used to identify the change points and get the coordinates for those points (for example, see Chang and Yamada, 2010). Other methods such as PCA, modified Sample Entropy or methods from the literature review would also work. The simplest method is for users to assign some control points on the x-axis that are critical and record the y-axis value for those points. For more precise monitoring, the profile can be divided into segments and a different number of control points can be assigned to each segment. The result is a group of tuples of values or coordinates. Each machine and each process need to have its own profile identification statistic.

The second step is to calculate the distance measured from the target value for all the statistics generated in the first step. In essence, the monitoring shifts from tracking the raw data into monitoring the characterizing values or statistics of the profiles. In the first stage, the target value is calculated by either using the pre-set specification values or averaging the characterizing values that were acquired in the first step. In the case where the user can set certain critical points on the x-axis and uses the reading of those points (y-axis value), the difference between the reading and the target value is calculated. The same is true in the case where the profile has been characterized by one value (for example modified Sample Entropy as proposed Chapter 2) where the distance is defined as the distance from the "good" characterizing value. If certain control points or change points are established, the distance from the calculated or set target needs to take into account both x-axis and y-axis distances. A simple triangle calculation would be sufficient in those cases.

The calculated distances are then further analyzed. In parameter setting stage (stage 1), the standard deviation of the distances is also calculated. Then in both stages, the data can be normalized using the standard deviation and mean value.

To level the simpler profile monitoring methods (using the control points) with one-value methods, only one characterizing value should be assigned to the whole profile. Two methods could be recommended: a worst-case scenario that is based on group control charts idea or average distance over all control points. This decision must be made during stage 1 to help with setting up the color- distributions because the calculation order differs.

In the average over control point method, the average value of *u* control point distances is calculated in stage 1 and used as the standard in stage 2.

$$d_{jks} = \sum_{r=1}^{n_r} b_{jksr} \Big/ p \tag{5}$$

where d is the profile characterizing value (average distance from target)

j – index for departments (j = 1,2,…,h)

k – index for machines in department j (k = 1,2,…,m)

s– index for identifying the profile sequence number (s = 1,…,o)

b is the calculated distance from the target for each control point

r – index for control point (*r*=1,2,…,*u*)

p is the number of control points in each profile

The worst-case scenario skips the formula (5) and is continued in the third step.

The third step is to classify each distance calculated in the previous step using the pre-control idea. If the average distance from check point method is used, the classification adds one to the counter representing one of the four colors based on the following criteria:

$$C_{jks}^{i} = C_{jks}^{i} + 1 \tag{6}$$

where

$$\begin{cases} i = g\ for\ green, if the\ d_{jks}\ is\ within\ 1.5\ times\ the\ target\ standard\ deviation\ from\ the\ target \\ i = y\ for\ yellow, if\ the\ d_{jks}\ is\ between\ 1.5\ and\ 3\ times\ the\ \ target\ standard\ deviations\ from\ the\ target \\ i = r\ for\ red, if\ the\ d_{jks} is\ more\ than\ 3\ times\ the\ target\ standard\ deviation\ or\ machine\ is\ down \\ i = w\ for\ white, if\ the\ machine\ is\ scheduled\ to\ be\ down \end{cases}$$

Note that these counters preset to 0 at the beginning of a period are used to generate the frequencies (i.e. counts) for all profile variables over a period of time. For the worst-case control point method, eq (6) conditions can be easily modified by replacing the average distance d_{jks} by the worst distance. Similar methods can be applied to the other process parameters with types other than profile. In fact, continuous parameters are very similar to the worst-case profile method.

The fourth step summarizes all the categorized color machine counts and generates a statistic for comparison and detection of the change. Total numbers of occurrences for green, yellow, red, and white zone are computed as follows:

$$T^i = \sum_{s=1}^{o} \sum_{j=1}^{h} \sum_{k=1}^{m} C_{jks}^i, for\ i = \{g, y, r, w\} \tag{7}$$

All available counts over the period under investigation are tallied as:

$$T = T^r + T^y + T^g \tag{8}$$

While the monitoring of white color (scheduled maintenance or no orders) is important for the higher management, the red, yellow and green are vital for any level as they provide information on the actual performance of the system. Therefore, the three colors are used for performance indicators. However, a large proportion of $T^w/(T + T^w)$ as shown in eq (9) means a long idle time during the investigation period.

The statistics of the ratio between the counts of each color category divided by the overall available count of samples represent the frequencies for each color zone. The following formulas (9) are used to create a report for each period under investigation.

Green	T^g/T
Yellow	T^y/T
Red	T^r/T
White	$T^w/(T+T^w)$

9)

The formulas 7-9 are shown here to be used in the system-wide monitoring case. However, if more detailed look is needed, the resolution could be heighten and the formulas could be used to summarize the results for each department or machine as needed. The difference is that classified samples and available counts are only used if they are part of that particular department or machine.

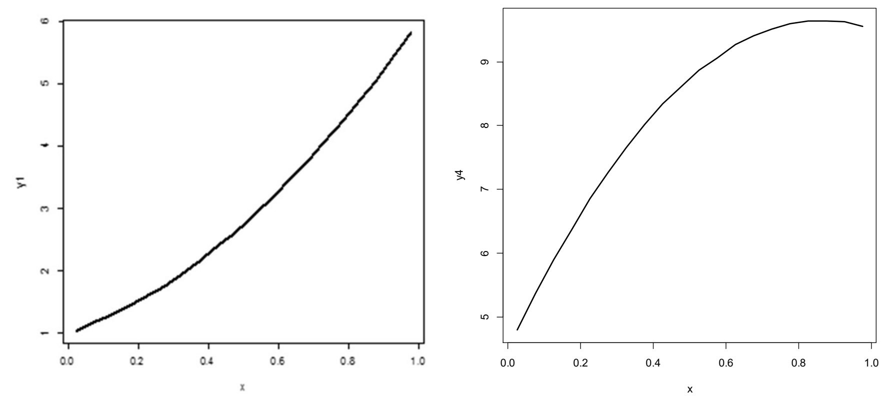
The final step is to detect changes. The previous four steps are used in the period under investigation and also in two other periods, for example, average values of the ratios over year-to-date and average values of the ratios over historical data from the same periods over the previous years. The results from the first stage calculations can also be used for comparison. The decision criteria for the change detection can be calculated by either using historical data to determine the best indicator. When a change is detected, the location and the period of change can be pinpointed by segmenting the data further and re-analyzing using the same method. For example, if the time frame under investigation is weekly, the segmentation would be to generate the same table for each day of that week and for each department of the system. The comparisons of the daily patterns in term of the ratios in eq. (9) would reveal when and where any change might occur.

4.4 An Illustrative Example

In this section, the TUFF method is applied to a small simulated production system. The assumption is that the system consists of three stations each containing one machine that modifies the product in different ways and the process parameters are characterized by profiles. Specifically, the first station applies exponential pressure over time, the second applies similar pressure on different angles, and the third station heats the product over time as shown in Table 11. The representation of underlying models without error term is shown in Figure 17.

Table 11. Underlying models of illustrative example

Station number	Underlying model
1	$y_{1,t} = 1 + 2x_1 + 3x_2^2 + \varepsilon$
2	$y_{2,t} = 1 + 2x_1 + 3x_2^2 + \varepsilon$
3	$y_{3t} = y_{1,t-2} + (-10(x - 0.5)^2 + 6) + \varepsilon$



(a) (b)

Figure 17. The underlying profiles for (a) stations 1 and 2; (b) station 3

The simulated system assumes that the first and third stations are correlated as can be seen in Table 1. In real-life situations, this information is usually not known. The error terms ε are assumed to be normally distributed with a mean of 0 and a standard deviation of 0.1. Twenty equally spaced points between 0 and 1 were used as x values. The example is laid out as a weekly production run. Each week 100 products are processed in all three stations.

Setup Stage. One hundred simulated profiles for all three stations over one week were analyzed with the TUFF method. The profiles were simulated to represent one week of work where the production results were deemed to be good. Since the profiles were simple, seven control points were chosen along the x-axis that is capable of defining the shape. The average values for 100 profiles at those seven points were calculated, which were used as the target value at each control point. Next, the target values were subtracted from an observed profile value at each profile's control point to calculate the distance from the target. Using eq. (5), mean distances, as well as their standard deviations, were calculated. Based on the classification and summarization of the distances explained in steps 3 and 4 of the TUFF method using eqs. (6) to (9), the process monitoring results were obtained in Table 12.

Table 12. The results of process monitoring using the TUFF method (setup stage)

Class	Overall	Station 1	Station 2	Station 3
Green	87.05%	86.57%	87.43%	87.14%
Yellow	12.67%	13.14%	12.29%	12.57%
Red	0.28%	0.29%	0.28%	0.29%

The main interest of the system-wide monitoring would be in the overall section, in which, 87.05%, 12.67%, and 0.28% of profiles are deemed in the green, yellow, and red zone, respectively. In this example, this result reflects what the monitoring system should report when the underlying process from all three stations is under control. Note that all stations shown in table 12 also carry similar color distributions.

Monitoring Stage. Two weeks of production data is simulated next. The first week represents the no-change scenario and the second week assumes that a small change in the error term was introduced in the first station. Instead of the error term being N(0,0.1), it becomes N(0, 0.11). The new profiles for each station are analyzed by running the TUFF method using the target values, mean distances and standard deviations of distances calculated in the setup stage as shown in Table 13.

Table 13. The results of the process monitoring using the TUFF method (monitoring stage): week 1, week 2 and a detailed look at week 2

	Setup Wk	Week 1	Week 2	Detailed Week2		
Class	Overall	Overall	Overall	Station 1	Station 2	Station3
Green	87.05%	83.81%	84.24%	78.86%	88.71%	85.14%
Yellow	12.67%	15.67%	14.71%	19.14%	11.14%	13.86%
Red	0.28%	0.52%	1.05%	2.00%	0.15%	1.00%

As shown in Table 13, the first week's overall results showed small changes, but these are not considered problematic when the threshold for issues was set at 1% of the red percentage.

However, the results from week 2 shows that the red percentage is more than 1%. The diagnostics using the details in the station level shows that the 1% threshold has been exceeded in both Station 1 and Station 3. More precise investigation may be initiated in more details in the machine or parameter level in a more complex system. While this three-machine system is not difficult to analyze, real-world applications may have thousands of variables. Since the TUFF method is easy to scale up, it can manage much bigger challenges.

4.5 Simulation Studies

This section presents a couple of simulation studies regarding the numerical performance of the TUFF method. To the best of our knowledge, the vast majority of existing literature focuses on detecting changes in a single profile and not on the system of different profiles. Therefore, these simulation studies were used to explore the properties of the TUFF method in a hierarchical system.

6.1 Study 1

The purpose of the first simulation was to study how the color distributions of one simple profile behaves facing various shifts while all the profiles in the system exhibit no changes during the production. The profile in the first study is a quadratic model $y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_{ij} \quad i = 1, .., n$ was considered. This simulation was also intended to identify the best practice for single machine level of the TUFF method. In each run, 1000 profiles were generated to mimic a certain daily or weekly production run on one machine. Six out-of-control cases were considered as shown in Table 14. The underlying shape of the model changes by small amounts for each of the parameters of β in cases 1-3. This is comparable to the mean change in continuous variable cases. Case 4 has a variance increase, while in case 5, the variance decreased. Case 6 was a combination of the variance change and parameter change.

Several approaches may be possible for monitoring a single machine in the TUFF method. The first possible approach for summarization uses the average deviation distance from

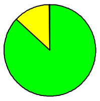
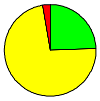
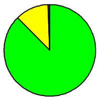
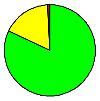
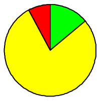
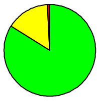
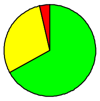
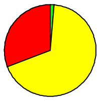
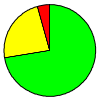
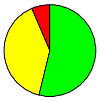
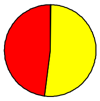
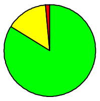
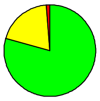
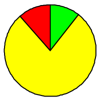
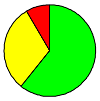
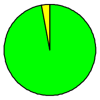
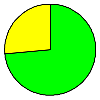
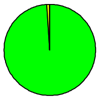
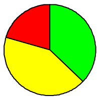
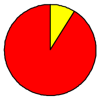
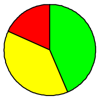
Table 14. The parameters for changed profiles in simulation study 1

	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{St.Dev}$
Base case	1.0	2.0	3.0	0.1
Case 1	1.0	2.0	3.1	0.1
Case 2	1.0	2.1	3.1	0.1
Case 3	1.1	2.1	3.1	0.1
Case 4	1.0	2.0	3.0	0.11
Case 5	1.0	2.0	3.0	0.05
Case 6	1.1	2.1	3.1	0.11

10 control points on the graph to generate the color distribution using eq. (6)-(9). The second possible approach uses the group control chart idea to choose the worst deviation distance among 10 control points. Then eq. (6)-(9) are used to generate the color distribution. The third possible approach uses the modified Sample Entropy (Chapter 2) in the first step to generate one value that represents the profile. Sample Entropy in short is often applied to a time series for assessing its variability (Grassberger and Procaccia, 1983). The sensitivity of the original Sample Entropy to detect small changes was enhanced by the introduction of modified Sample Entropy (Xie *et al*, 2010). More detail of the modified Sample Entropy method can be found in Chapter 2. Once a modified sample entropy value is generated for each profiles, eqs. (6)-(9) can be carried out to generate the color distribution.

Table 15 summarizes the colors distributions generated from the approaches mentioned above. The numerical percentages were replaced with pie-charts for ease of visualization of the

Table 15. The results from simulation study 1 with 6 cases of profile change

ALL	M1. Average over 10 variables	M2. Worst case scenario	M3. Modified Sample Entropy
Base case			
Case 1			
Case 2			
Case3			
Case4			
Case 5			
Case 6			

changes. The base case represents the in-control state where no change occurs. Comparing three methods M1 to M3, M2 is the worst case method based on the group control chart idea so M2 is the most sensitive to any shifts. Even in the base case situation, the norm is a large portion of yellow comparing to the green and red zone. On the other hand, both M1 and M3 are based on the "average" concept considering the entire profile, the color distributions of M1 and M3 are much less sensitive for detecting shifts as shown in the base case. The green zone dominates the other color zones.

Comparing the detection performance within each methods, small changes in the underlying model (cases 1-3) are basically not detected by M1 and M3 as shown in Table 15 although case 3 shows that M1 is a little better than M3. Both M1 and M2 cannot detect a change in the variability change (case 4), however the sample entropy model M3 is able to catch that change. In all the other cases, the distribution of the classes is visibly different from the base case. Even in case 5 where the variation decreased, the color distributions of all methods are different from their corresponding base cases. Since the change, in that case, is for better, it is probably not too important from the monitoring point of view, but nevertheless, the change has been detected.

The worst-case scenario method M2 is different from the two other methods in the base case in that the yellow and red zones have much larger proportions. This can be explained by the theory that the more variables (or in this case, control points) there is in the system, the higher is the probability to find at least one point that is out of control limits even though the process might be in control. Since the focus of the method is to monitor the system and not to focus on every single profile, the exact ratios identified in the in-control situation do not need additional analysis. As can be seen in Table 15, the worst-case method M2 is capable of detecting all cases of out-of-control sets because all color distributions from case to case differ. However, the drawback of M2 is that as was shown in the continuous variable case (Chapter 3), if the number of control points increases over 30, the method loses its detection power because of the Bonferroni curse of dimensionality.

6.2 Study 2

The goal of this simulation study is to establish the sensitivity of the TUFF method through Type I and Type II errors when a portion of the underlying profiles exhibit changes. In this study, the system consisting of 100 machines is explored. Assume that each machine generates one profile over a period. The goal is to monitor the system and detect changes. Since the underlying model is not important since the proposed method focuses on deviation from the profile control points, simple models were chosen to represent the output of the machines. The worst-case control point method M2 was chosen to represent simple profiles and the average over control points method M1 was used to represent the complicated profiles. M2 was chosen because the previous simulation study showed that it had the best detection capability while M1 was chosen because it showed similar detection power as modified Sample Entropy method M3. The profiles simulated were based on two models in Zou *et al*, 2008.

6.2.1 First stage: Setting of the parameters

The quadratic model was used as the base model

$$y_{ij} = 1 + 2x_i + 3x_i^2 + \varepsilon_{ij} \quad i = 1, .., n$$

The error term ε_{ij} was assumed to be independent and identically distributed N(0,0.01) in both models. The error term was chosen to be 10 times smaller than the one used in Zou *et al* (2008) because the original error terms did not seem to be realistic. Twenty equally spaced points $x_i = \frac{i-0.5}{n}, i = 1, ..$, were chosen for the implementation of the TUFF process.

First, 10,000 profiles representing 100 profiles for 100 machines were generated to calculate mean values for control point targets. Given the shape of the profiles and for the sake of simplicity, seven control points were chosen. More points were assigned to steeper curve areas for more precision. The control points represent the x values and then the deviation distances represent the corresponding y values. The target values were calculated by averaging all y values at corresponding x points.

Next, the distances from the target were calculated based on the preliminary 10,000 profiles where the mean values were subtracted at each point in each profile. Since the x values were set, the distance was calculated only on the y-axis. The mean and standard deviation of the distances was calculated for each control point to help the classification in the following step.

The distances were then classified based on the guidelines from the third step and ratios of the color classes were calculated. This process was repeated 10,000 times to help establish thresholds in the percentage table for change detection. Four thresholds were recorded to determine the performance: average of control point ratios for the green color, average of control point ratios for red color, group control-based "worst-case from each profile" ratios of green color and of red color. The threshold was set based on Type I error (for false alarm rate) of 0.02. The thresholds for the scenario 1 were 0.835, 0.0071, 0.3205 and 0.032 and those for scenario 2 were 0.8385, 0.007, 0.33 and 0.02901, respectively. The differences between the scenarios were small, but they still need to be accounted for.

6.2.2 Second stage: a Sensitivity study

The purpose of the simulation study is to determine the sensitivity of the TUFF methods. For each repetition, 100 profiles for 100 machines were generated. Several different sensitivity levels were examined: 0.5%, 0.6%, 0.7%, 0.8%, 0.9%, 1%, 5%, 10% and 20% of the profiles changed. The same control points were used in the parameter setting segment calculations.

Three different out of control models were chosen similarly to Zou *et al* (2008). The models were as follows:

(I): $y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_{ij}$

(II): $y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_{ij}$

(III): $y_{ij} = 1 + 2x_i + 3x_i^2 + \beta_1 \sin(2\pi\beta_2 x_i) + \varepsilon_{ij}$

These three out of control models represent three cases: (I) the parameters shift, but the structure of the regression relationship remains; (II) the regression relationship changes to another linear model and (III) the regression relationship changes to a nonlinear model.

Table 16. The parameters for changes used in the simulation

Scenario 1

Out-of-control model 1					Out-of-control model 2						Out-of-control model 3			
	β_0	β_1	β_2	$St.Dev$		β_0	β_1	β_2	β_3	$St.Dev$		β_1	β_2	$St.Dev$
Case 1	1	2	3.1	0.1	Case 1	0.8	4.4	-3	4	0.1	Case 1	0.1	1	0.1
Case 2	1	2.1	3.1	0.1	Case 2	0.8	4.4	-3	4.1	0.1	Case 2	0.2	1	0.1
Case 3	1.1	2.1	3.1	0.1	Case 3	0.8	4.4	-3.2	4.1	0.1	Case3	0.2	0.8	0.1
Case 4	1	2	3	0.11	Case 4	1	4.4	-3.2	4.1	0.1	Case 4	0.2	1.3	0.1
Case 5	1	2	3	0.05	Case 5	0.8	4.4	-3	4	0.11	Case 5	0.3	1.5	0.1
Case 6	1.1	2.1	3.1	0.11	Case 6	0.8	4.5	-3	4	0.11	Case 6	0.3	1.5	0.11

In scenario 1 model 1 the first three cases show changes in the shape of the model, fourth case had a variance increase, fifth case showed variance decrease and the sixth case was the mixture of variance and parameter change. In the second model, the first four cases were shape change, the fifth was variance change and the sixth was a mixture. In model 3 the first five has shape change and the sixth was the mixture case.

The output of this simulation was captured in the form of red and green percentages in each case. After recording the percentage data, Type II error for each combination was calculated based on the thresholds established in the first stage of the study.

In the case of profiles represented with single value (for example, SamplEn, worst case of control points, etc) (table 17), the results show that the method is capable of detecting changes already at 0.6% of the profiles being changed. Smaller changes than that are not picked up well. The reduction of variability is not noticed (case 5 in model 1), but as was explained with the first simulation study, the change is usually considered to be better. On the color comparison, red color monitoring is much better than green color monitoring. The red can pick up 0.6% of the profiles changing while the green color starts to read at 5% changes.

In the case of the simpler profiles when using the worst classification of the control points, the results (table 18) show that if a small number of profiles were affected by the change (one machine or less), the method had difficulties to capture the change. As could be expected, when the changes were in larger number of machines, the method became better in detecting changes. When comparing the detection power of the color class, then monitoring of red color was superior to green color monitoring. The red color was capable of picking up the changes when they affected more than one machine, while green color monitoring started to be effective after 10 machines had been changed. However, since not all the possible combinations could be presented here and there is another color in the mix, the recommendation would be to monitor both colors

Table 17. The results of simulation study 2 for profile monitoring with single value in terms of Type II error

		Percent of change								
RED		0.5%	0.6%	0.7%	0.8%	0.9%	1%	5%	10%	20%
Model 1	Case 1	0.289	0.002	0	0	0	0	0	0	0
	Case 2	0.313	0.003	0	0	0	0	0	0	0
	Case 3	0.284	0.001	0	0	0	0	0	0	0
	Case 4	1	1	1	1	1	0	0	0	0
	Case 5	1	1	1	1	1	1	1	1	1
	Case 6	0.295	0.002	0	0	0	0	0	0	0
Model 2	Case 1	0.672	0.108	0.001	0	0	0	0	0	0
	Case 2	0.29	0	0	0	0	0	0	0	0
	Case 3	0.269	0.002	0	0	0	0	0	0	0
	Case 4	0.292	0.001	0	0	0	0	0	0	0
	Case 5	0.819	0.274	0.013	0	0	0	0	0	0
	Case 6	0.284	0.005	0	0	0	0	0	0	0
Model 3	Case 1	0.881	0.431	0.057	0.005	0	0	0	0	0
	Case 2	0.285	0.001	0	0	0	0	0	0	0
	Case 3	0.284	0.003	0	0	0	0	0	0	0
	Case 4	0.274	0	0	0	0	0	0	0	0
	Case 5	0.321	0.003	0	0	0	0	0	0	0
	Case 6	0.282	0.003	0	0	0	0	0	0	0
		Percent of change								
GREEN		0.5%	0.6%	0.7%	0.8%	0.9%	1%	5%	10%	20%
Model 1	Case 1	1	1	1	1	1	1	0.01	0	0
	Case 2	1	1	1	1	1	1	0.022	0	0
	Case 3	1	1	1	1	1	1	0.027	0	0
	Case 4	1	1	1	1	1	1	0.018	0	0
	Case 5	1	1	1	1	1	1	1	1	1
	Case 6	1	1	1	1	1	1	0.026	0	0
Model 2	Case 1	1	1	1	1	1	1	0.018	0	0
	Case 2	1	1	1	1	1	1	0.011	0	0
	Case 3	1	1	1	1	1	1	0.018	0	0
	Case 4	1	1	1	1	1	1	0.019	0	0
	Case 5	1	1	1	1	1	1	0.027	0	0
	Case 6	1	1	1	1	1	1	0.018	0	0
Model 3	Case 1	1	1	1	1	1	1	0.018	0	0
	Case 2	1	1	1	1	1	1	0.015	0	0
	Case 3	1	1	1	1	1	1	0.016	0	0
	Case 4	1	1	1	1	1	1	0.019	0	0
	Case 5	1	1	1	1	1	1	0.024	0	0
	Case 6	1	1	1	1	1	1	0.012	0	0

Table 18. The results of simulation study 2 for profile monitoring using worst case of control points in terms of Type II error

		Percent of change							
RED		0.6%	0.7%	0.8%	0.9%	1%	5%	10%	20%
Model 1	Case 1	1	1	0.997	0.994	0.957	0	0	0
	Case 2	1	1	0.999	0.992	0.97	0	0	0
	Case 3	1	1	0.998	0.995	0.968	0	0	0
	Case 4	1	1	1	1	0.981	0	0	0
	Case 5	1	1	1	1	1	1	1	1
	Case 6	1	1	0.999	0.994	0.969	0	0	0
Model 2	Case 1	1	1	0.999	0.998	0.955	0	0	0
	Case 2	1	1	1	0.998	0.966	0	0	0
	Case 3	1	1	1	0.995	0.971	0	0	0
	Case 4	1	0.999	1	0.999	0.969	0	0	0
	Case 5	1	1	1	0.994	0.974	0	0	0
	Case 6	1	1	0.999	0.992	0.967	0	0	0
Model 3	Case 1	1	1	0.999	0.997	0.975	0	0	0
	Case 2	1	1	0.999	0.998	0.963	0	0	0
	Case 3	1	1	0.999	0.994	0.962	0	0	0
	Case 4	1	1	1	0.996	0.972	0	0	0
	Case 5	1	1	1	0.995	0.976	0	0	0
	Case 6	1	1	1	0.999	0.967	0	0	0
		Percent of change							
GREEN		0.6%	0.7%	0.8%	0.9%	1%	5%	10%	20%
Model 1	Case 1	1	1	1	1	1	1	0.905	0
	Case 2	1	1	1	1	1	1	0.875	0
	Case 3	1	1	1	1	1	1	0.903	0
	Case 4	1	1	1	1	1	1	0.887	0
	Case 5	1	1	1	1	1	1	1	1
	Case 6	1	1	1	1	1	1	0.906	0
Model 2	Case 1	1	1	1	1	1	1	0.867	0
	Case 2	1	1	1	1	1	1	0.897	0
	Case 3	1	1	1	1	1	1	0.891	0
	Case 4	1	1	1	1	1	1	0.847	0
	Case 5	1	1	1	1	1	1	0.908	0
	Case 6	1	1	1	1	1	1	0.896	0
Model 3	Case 1	1	1	1	1	1	1	0.89	0
	Case 2	1	1	1	1	1	1	0.882	0
	Case 3	1	1	1	1	1	1	0.894	0
	Case 4	1	1	1	1	1	1	0.897	0
	Case 5	1	1	1	1	1	1	0.863	0
	Case 6	1	1	1	1	1	1	0.886	0

4.7 A Real-World Example

In this section, the TUFF method is applied to a real dataset obtained from an industrial rubber hose manufacturing facility. High-pressure hose products are made out of alternate layers of rubber and metal wires. Toward the end of the production process, various high-pressure hose reels are loaded and cured. A curing process typically consists of a sealed, heated chamber, called an *autoclave* or *vulcanizer.* Most vulcanizers are equipped with multiple thermocouples inside their chambers and/or parts. The most important information gathered from the thermocouples is the air temperature, represented as a nonlinear profile during the curing cycle. A programmable logical controller is used to control temperature based on sensor reading or a fixed time interval according to a curing recipe. Figure 5 shows a typical air temperature profile divided into three sections – heat-up stage, curing stage and cool-down stage. It is important for the flat section (the curing stage) of this profile to hold for a fixed period of time. The thermocouples are calibrated from time to time. The data collection unit of the PLC is capable of collecting multiple temperature readings per second. (Chang *et al*, 2012).

The data analyzed for the illustrative example contained profiles from five days. The first day had 26 profiles and the following days had 26, 44, 24 and 33 profiles respectively. The first day was chosen to be the benchmark day to which all the other days are compared to. Piecewise regression was chosen to be the single profile analysis method for the simplicity of the algorithm and the shape of the profile. R function *segmented()* from package *segmented* was used for the piecewise regression analysis because the function is capable of producing estimated change points, which were then used for the distance calculations and all the other steps of the proposed method.

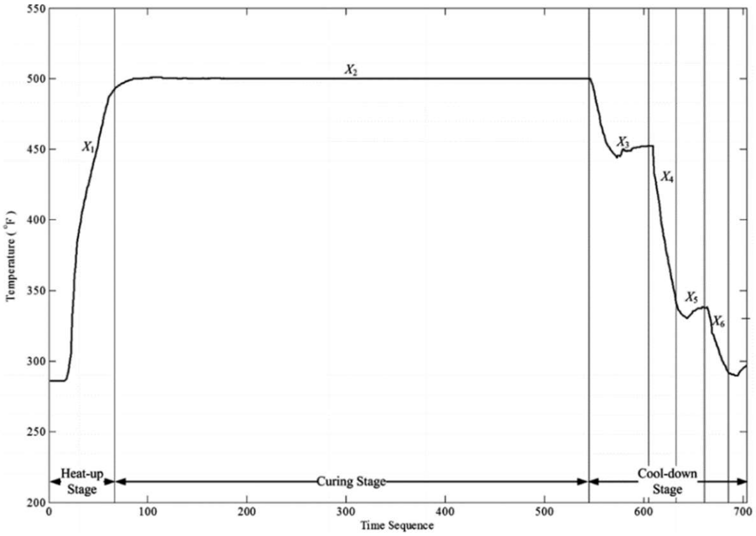


Figure 18. Example of a profile generated in the vulcanization process

Stage I. The profiles from the first day were analyzed with the *segmented()* function. The results indicated that there were 4 change points along the time axis of the profiles. For each profile, the change points and the temperature reading at that point were recorded, resulting in 8 records for each profile. Then the average and standard deviation for each record were calculated over the 26 profiles of the first day. The resulting values were then used to color-code each record and summarized according to step 4 of the proposed method (table 19).

Stage II. Each of the remaining 4 days was analyzed separately. The same principals were used. Each profile was analyzed with the *segmented()* function to establish change points. The same amount of change points and respective temperature readings were detected and recorded. The average values for each point established in Phase I were subtracted and the results were classified based on the suggestions from the third step of the method. Each day was then summarized and the result table 20 was generated.

Table 19. The results of the first day of vulcanizer profiles

	Percentage
Green	90.38%
Yellow	6.25%
Red	3.37%

Much larger sample size than 5 days would be useful for more precise change detection, but when comparing each day to the first day, the results show that the second day has changed since the green portion of the results of the day is higher and the yellow and red portions are lower. This means more change points were close to the targets that were established on the first day. The third and the fourth day indicate also changes, but they are different, indicating fewer change points were close to the targets. The fifth day does not show many changes when compared to the first day. The more specific follow up analysis should be done on the third and fourth day, because of the change detected. Since the assumption for simplifying the example here was that day 1 was “acceptable”, then no analysis is necessary for day 1.

Table 20. The results of 5 days of vulcanizer profiles

	day1	day2	day3	day4	day5
	Percentage	Percentage	Percentage	Percentage	Percentage
Green	90.385%	92.308%	86.080%	89.063%	90.152%
Yellow	6.250%	5.769%	11.364%	8.854%	6.061%
Red	3.365%	1.923%	2.557%	2.083%	3.788%

4.8. Conclusions and Future Studies

In this chpater, a system-wide monitoring method called the Technique of Uniformally Formatted Frequencies (TUFF) is proposed for the monitoring of profile data in enterprise-level quality control. It is adapted from the pre-control chart idea for classifying all measurements over a period of time into three classes according to how close to their target. This study examines three methods to convert a profile to a measurement ready for classification. The TUFF method is easily scalable to big data problems and can be used for both linear and non-linear profiles. The method is not designed for real-time monitoring, but rather for near-time quality assurance over a certain period. A simulation study was carried out to investigate the sensitivity of the proposed method. An illustrative example was presented based on the curing process of rubber products.

The TUFF method is designed as a system process monitoring tool first and foremost. However, it can be implemented for other purposes as well. The following are two applications where the proposed method could be effective.

Process capability study

A process capability study measures the uniformity of a process. A process capability analysis helps quantify the process variability, analyze this variability relative to product requirements in terms of specification limits, and assist in eliminating or greatly reducing this variability (Montgomery, 2012). The TUFF method could be applied to process capability studies. The proposed method can first convert the profiles used as quality characteristics quality into colored pie charts. Observing the changes in the red portion in the pie charts in a fixed sample frequency(e.g. weekly), decision makers would be able to glean more information in process stability as oppose to just one process capability index number from the traditional process capability study.

Prognoses

The TUFF method can be used for prognoses when all the historical data has been analyzed. For example, the proposed framework can pinpoint the time and place of potential changes for a new set of data. Association rule mining methods (Rajaman and Ullman, 2012) could be applied to this new data set to search for similar occurrences in the past. If similarities are found, past corrective actions linked to this data pattern may provide immediate feedback for adjustments to restore process stability impacting product quality.

One of the future research tasks is to extend the proposed profile monitoring method for an all-inclusive, enterprise-level, system-wide process monitoring framework. The main task involves the integration of all process data types such as continuous, attribute, and profile variables. This study and the previously published works in the profile monitoring are based mainly on the assumption that the profiles are not affected by external factors. However, this assumption may not hold for real-world situations. For example, process temperatures might be affected by the ambient temperature of the factory. It could be beneficial to study how that affects the TUFF method for profiles and what additional steps might be needed to address that type of change. The examples given in this article have been generated using normal distribution on the error term of the models. It would be interesting to see if the method works similarly for non-normal data.

Chapter 5 - From Data to Knowledge: MADIC- a Six Sigma Implementation Strategy in Big Data Environments

5.1. Introduction

The competition for market shares and consumer satisfaction has been one of the major driving forces for many innovations in manufacturing companies. Businesses try to find any advantages that can help them succeed. Many process improvement methods have been proposed over the years for companies to improve their operational performance. One of the most utilized methods is the Six Sigma approach (Alexander *et al,* 2019; Gupta *et al,* 2018; Noori *et al,* 2018; Anthony *et al,* 2017; Gijo *et al,* 2014, etc). It is a data-driven method that focuses on the reduction of variability of the controllable variables in a system with the ultimate goal of reducing process variability (Alexander *et al*, 2019).

Six Sigma methodology was first introduced in the Motorola company around 1987 as the competition from overseas manufacturers started to mount extreme pressure (Anthony *et al*, 2017). The engineers in Motorola started improvement projects that followed a "roadmap" of Measure, Analyze, Improve and Control (MAIC). This roadmap provides a generic approach for a variety of problems. The results of this process attracted the attention of other companies such as General Electric (GE) during the mid-1980s. When GE decided to adapt the methodology, they also added an additional level to the beginning of the process (Define) and DMAIC was born (Anthony *et al*, 2017). In recent years, DMAIC has been used in many applications in manufacturing companies and also in various scientific fields, some of which will be discussed.

Attention towards Six Sigma has been increasing still to this day. Raval *et al* (2016) reported in a comprehensive literature review on the Lean Six Sigma trends and themes and pointed out that there is an increasing trend of numbers of papers on the topic since 2002. Alexander *et al* (2019) reviewed the topic of Lean Six Sigma for small and medium-sized manufacturing enterprises. Pathirante *et al* (2018) reviewed critical success factors for Six Sigma in service and manufacturing companies and identified 48 different factors, one of these was information technology and innovation. Shamsi *et al* (2018) focused on the implementation barriers of Lean Six Sigma in the Information Technology industry. They identified that the biggest barriers were time consumption, staff turnover, difficulty in data collection, and difficulty in deciding on the scope of the project. Anthony *et al* (2017) pointed to big data as one of the future trends for Six Sigma and projected that big data would breathe new life into Six Sigma standards by providing entry to a data-rich environment.

The rapid evolution of sensors, data storage, and computational resources has enabled the processing of large datasets that can contain different data types. Since Six Sigma adopts a data-driven philosophy, the opportunities provided by big data environments would be a natural fit. Big data tools would allow the search process for the areas needing most improvement to be automated, therefore speeding up the project-identification process. The other benefit would be to minimize the guesswork on which area to focus on because the identification of the improvement project could be initiated by the data, rather than a person. Most of the published case studies that focus on projects in the manufacturing industry follow the traditional path of DMAIC, where the goal is defined by the process owner, the team is gathered and then data is collected on the issue to identify the source of variation that needs improvement (Martinez Leon *et al*,2012; Gijo *et al*, 2014; Gupta *et al*, 2018; Noori *et al*, 2018; Sharma *et al*, 2018). Big data tools could help both the Define and Measure steps.

A few research fields that examine the use of big data in Six Sigma include education (Laux *et al*, 2017), digital curation (Arcidiacono *et al*, 2016), and financial institutions (Zwetsloot *et al*, 2018). In the manufacturing field, Gaudard *et al*. (2009) proposed using historical datasets for a rotogravure printing process to shrink the number of variables that might cause issues by using a decision tree method called recursive partitioning. Thirty-nine variables used in the dataset were either continuous or categorical. The decision tree process identified the variables that were possibly the most influential in producing non-conformity. These variables were used to perform more experiments in a controlled environment to confirm the need for improvement. Stojanovic *et al* (2015) proposed a method of big data analytics for continuous process improvement in manufacturing. They elaborated on using big data-driven clustering for the efficient discovering of unusual occurrences in real-time. They claim that extending traditional clustering algorithms with methods for better understanding the nature of clusters through big data processing will pave the way for empowering Lean Six Sigma. Their example showed finished-product testing using three profile variables. In another study, Stojanovic *et al* (2017) argued that that Six Sigma comes from the era of small data. They proposed a big data platform for enabling self-healing manufacturing, which could be used in continuous improvement projects.

One of the bigger obstacles of implementing Six Sigma processes in a big data environment lies in the fact that real-life systems may have more than 1000 variables and these variables can be a mixture of continuous, categorical, binomial, and profile variables. The use of big data allows analyzing a much bigger set of historical data. When big data is available, the identification of the problematic areas may be automated, and the order of activities in the continuous improvement method could be altered. Therefore, the first goal of this research is to renovate one of the most used Six Sigma tools DMAIC (Define, Measure, Analyze, Improve and

Control) in a manufacturing environment and show how big data analytics can help identify and prioritize continuous improvement projects. The second goal of this study is to identify continuous improvement projects according to changes identified in the system-wide monitoring phase. Specifically, a pair of attribute control charts namely p-charts are applied to red lights and green lights at any hierarchical level of a manufacturing system. When either chart indicates the process at that point is out of control, the identification of the possible improvement candidates is just a few steps away. Details of the proposed framework are described in the following sections.

One of the tools used in the continuous improvement projects in the Measure section is statistical process monitoring (or SPM, formerly also statistical process control, SPC) (Pande *et al,* 2000). The purpose of SPM is to monitor the process to determine if the process is under control. The evolution of the methods started with the univariate control charts such as Shewhart charts for detecting median to large process shifts and exponential weighted moving average or EWMA charts and cumulative sum or CUSUM charts for small shifts. To simplify the inspection, pre-control charts (Satterthwaite, 1954) were established, where data points were plotted based on how far they were from the mean with a set of rules on how to make a decision based on the location of the data points. In addition to univariate control charts, multivariate solutions such as Hotelling T^2 charts, principal component analysis, and Partial Least Squares, etc were also introduced. In the case of multiple stream processes, group control charts were introduced (Boyd, 1950). When the product quality characteristics follow a discrete distribution, p- and np- charts for attribute monitoring. (Montogomery, 2012). These are all traditional methods that are still used in manufacturing facilities. While they have been proven useful on the local machine level, they are not suitable for use at the system level. The traditional methods do not take into account the whole set of data, focus mainly on the quality characteristics, not on the process parameters and disregard previous stages of the process. Therefore, a system-wide monitoring method that takes into account all the data and uses both quality characteristics and process parameters would be very useful in the Six Sigma improvement portfolio.

5.2. Redefining DMAIC for the manufacturing sector

One of the most popular techniques of Six Sigma continuous improvement projects is DMAIC, which stands for Define, Measure, Analyze, Improve, and Control. In the first step (Define), the objective is to identify the project opportunity and to verify or validate that it represents legitimate breakthrough potential. The purpose of the Measure step is to evaluate and understand the current state of the process. In the second step (Measure), the defined metrics are measured. In the Analyze step, the objective is to use the data from the Measure step to begin to determine the cause-and-effect relationships in the process and to understand the different sources of variability. Next, improvements are suggested and carried out, and finally, they are monitored for performance in the Control step (Pande *et al*, 2000). While the method is data-driven, there is a large human presence in the Define step. Traditionally, in the define step of DMAIC, the experts weigh in on what variables or areas need to be monitored and what variation is allowed for measurements

As stated before, a modern manufacturing facility can generate a large amount of data. Almost all the facets of production are covered with sensors and measures. Unfortunately, not all of the data is used, but it is gathered. More often than not, the unused data is called 'dark data' because opportunities for the discovery of improving a process or situation are lost. One of the main reasons for dark data is the sheer volume of data generated. The first step of DMAIC – "Define" is often initiated by a Six Sigma team. Once a subject is deemed worthwhile, the data related to this subject is then collected. Since everything or almost all data is measured, the DMAIC technique could be redefined.

We propose to redefine a Six Sigma approach in a big data environment as MDAIC (Measure, Define, Analyze, Improve, and Control) as shown in figure 20. The first step assumes that all process and quality-related data is collected and measured. The data is constantly analyzed to identify any abnormalities in the streams of data using SPM tools for regular day-to-day operations of a factory. The proposed method also allows for system performance comparisons at any time scale such as by quarters. Based on the information gathered, issue areas and expert opinions are used to classify the severity and the priorities of improvement candidates. In essence, the novelty of this approach is that the signal for improvement need is originated by data and in turn triggers process owners to initiate Six Sigma projects. Human input is still needed to determine if the issue or the source of variation is actually critical. The data-driven approach also supports the forming of the team as relevant operators or engineers from the floor can be assigned to the team. After the Define step is completed, analyses are carried out to identify the root cause of the issue and all the usual steps (Improve and Control) are performed.

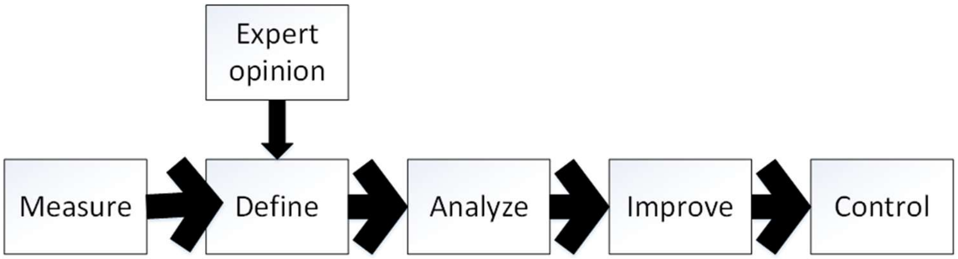


Figure 19. Redefined DMAIC proposal

5.3. The proposed application of the redefined MDAIC process

As stated previously, the traditional DMAIC steps dictate that a problem is first identified and then the data is collected based on the scope. The newly proposed redefinition of MDAIC requires a different method that could help trigger an investigation. The proposed method uses a few different SPM tools, such as continuous and profile monitoring algorithms and the attribute control chart idea to monitor the process and identify issue areas. The following figure outlines the steps of the proposed method (figure 21).

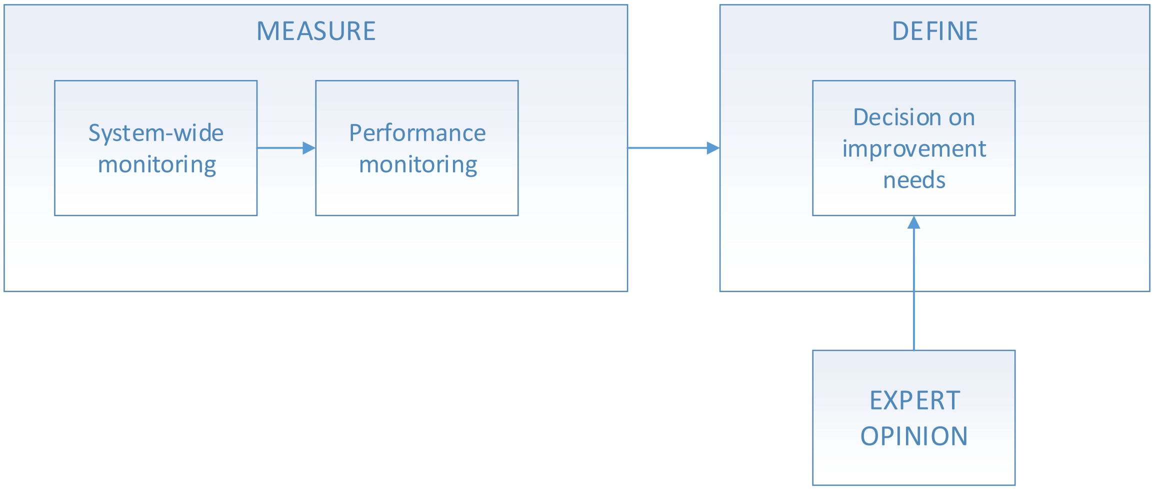


Figure 20. The flow diagram of the proposed application of the redefined MDAIC process

5.3.1 System-wide monitoring

Traditionally only raw data on quality characteristics (QC) from a product such as dimensions of a finished part is used for SPM. Some SPM methods are designed for univariate monitoring while others are for multivariate QCs. The SPM methods used in this study, namely monitoring of continuous variable data (Chapter 3) and profile variable data (Chapter 4), were proposed in the context of system monitoring. This study provides an integration of all types of process variables and product variables in a unified platform for monitoring system performance using the Technique of Uniformally Formatted Frequencies (TUFF). The reason for using this method is that all types of data should be assessed in the automated Six Sigma project to identify the candidate for improvement. The TUFF method is capable of monitoring continuous, attribute and profile variables. The methodology presented in this manuscript uses a top-down monitoring strategy, which means that the data is summarized to the top level and monitored there rather than monitor all the variables at their own local level**.** The reasoning is that top-down monitoring is less sensitive to false alarms at the variable level. While some of the errors might not be detected as fast as those when monitoring at the variable level, the variable level is more prone to overreaction to false alarms and a lot of effort might be directed to a non- issue. The TUFF method collects data over a period of time. Therefore, it is not a real-time process monitoring approach. Traditional process monitoring tools should still be implemented for detecting shifts at local levels. The TUFF system monitoring provides additional monitoring of system performance at a global level in that it counts the numbers of non-conformities gleaned from its sub-levels. It is a supplement to the traditional process monitoring that is carried out on the variable or machine level.

While the proposed research aims to refine the DMAIC principle in data abundant environment, the proposed method of system-wide SPM can also extend additional diagnostic information to traditional SPM methods. The first point is that the classic SPM relies on the sampling principles because of the gathering and analyzing of the data used to be expensive. This means that the data in traditional settings is usually scarce, the collection is slow and the frequency is also low. Big data applications contain abundant data, fast collection, and real- or near-time observations. It also offers the chance to analyze all observations, not just a set of samples. While this could mean collecting less useful data and recording outliers, the chance of catching problems is much higher when analyzing the full dataset. The TUFF method is capable of detecting changes and pinpointing issue areas in addition to traditional control charting methods.

The second point is that the classical SPM usually monitors just the quality characteristics, assuming that: if those are good, then all the other facets of production are also good. This circles back to the price of sampling and analyzing. Big data applications offer the chance to analyze everything. The TUFF method takes advantage of that and analyses both quality characteristics as well as the process parameters. In this way, pending problems might be caught earlier before they have had a chance to affect product quality. This practice may lead to savings in rework, possibly in preventive maintenance, etc.

The third point is that classic SPM methods are only applied to local quality characteristics without the consideration of its prior or posterior steps. The TUFF method is designed to summarize the system performance of the full system over a user-defined period of time. Coupled with product functional flowcharts, quality engineers can examine the entire production steps to pinpoint an issue related to time or places quickly.

5.3.1.1 Data manipulation

The measure phase of the continuous improvement process starts with gathering the data in the form of the results of continuous, profile and/or attribute monitoring data. These results may be collected hourly, daily, weekly at a different frequency depending on the specifics of the manufacturing process. Users can choose which timeframe is sufficient for their application; however, the data collection is done automatically as a part of day-to-day operations.

The generalized method for TUFF system-wide monitoring consists of five stages. First, raw data is transformed into the characterizing statistic over a preset sampling period. During the data collection phase, a different number of values will be collected for continuous variables and profile variables as well as attribute variables. In continuous variable cases, all the values would be recorded. In profile cases, only a few characterizing values are collected, for example, distance to the target at the sampling point or modified Sample Entropy value (Chapter 2). Similarly, the attribute results would have few recordings during the sampling period, for example, the percentage of non-conformity over the period between sampling points. This creates an imbalanced set where continuous variable readings would overwhelm the profile and attribute variables. The latter two variables would be lost and the monitoring would be almost exclusively on the continuous variable. Since the purpose is to monitor all the variables, the number of representations for each variable should be the same or very similar. In order to achieve a balanced representation of profile variables and attribute variables as well as continuous variables in the system-wide process monitoring, some modification is needed. The TUFF method would calculate and monitor the mean and standard deviation of the segment of the continuous variable that is similar in length as the profile using equations 10 and 11. A segment is defined as the period of time to produce on profile. The reason for using both mean and standard deviations on the continuous variable is that both are needed to characterize the segment, otherwise some changes are not captured.

$$\overline{\boldsymbol{x}}_{\boldsymbol{it}} = \frac{\sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{k}} \boldsymbol{x}_{\boldsymbol{ijt}}}{\boldsymbol{k}} \tag{10}$$

$$\boldsymbol{s}_{\boldsymbol{it}} = \sqrt{\frac{\boldsymbol{k} * \sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{k}} \boldsymbol{x}_{\boldsymbol{ijt}}^{\boldsymbol{2}} - \left(\sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{k}} \boldsymbol{x}_{\boldsymbol{ijt}}\right)^{\boldsymbol{2}}}{\boldsymbol{k} * (\boldsymbol{k}-\boldsymbol{1})}} \tag{12}$$

where $\overline{\boldsymbol{x}}_{\boldsymbol{it}}$ is the mean value of the *i*-th continuous variable during the segment which the *t*-th profile characterizing value is generated; *k* is the number of continuous variable readings during the segment which the *t*-th profile characterizing value is generated; $\boldsymbol{x}_{\boldsymbol{ijt}}$ is the *j*-th single value of the *i*-th continuous variable during the segment which the *t*-th profile characterizing value is generated; $\boldsymbol{s}_{\boldsymbol{it}}$ is the standard deviation of the *i*-th continuous variable during the segment which the *t*-th profile characterizing value is generated

After this transformation, the monitoring table is no longer based on raw data, but instead, on statistics generated by all variables which have similar weights. The new objects for monitoring would be the mean values and standard deviation values in case of continuous variables, the sampling points across the profile, and the nonconforming fraction values for attributes. The variables have been reduced to the same time scale and can be monitored simultaneously in the proposed framework.

The main sampling strategy is to generate the same number of statistics for all data types. We choose to use the least number of statistics from a profile variable as a base. For example, if the profile is characterized by a single Sample entropy value, then the mean and standard deviation of the continuous variable and the percentage of nonconformity of attribute variable are calculated over the same timeframe that was used to create the profile. If the characterization of the profile is done by merely setting certain control or sampling points, then the other variable characterization values are established over the period between those control or sampling points.

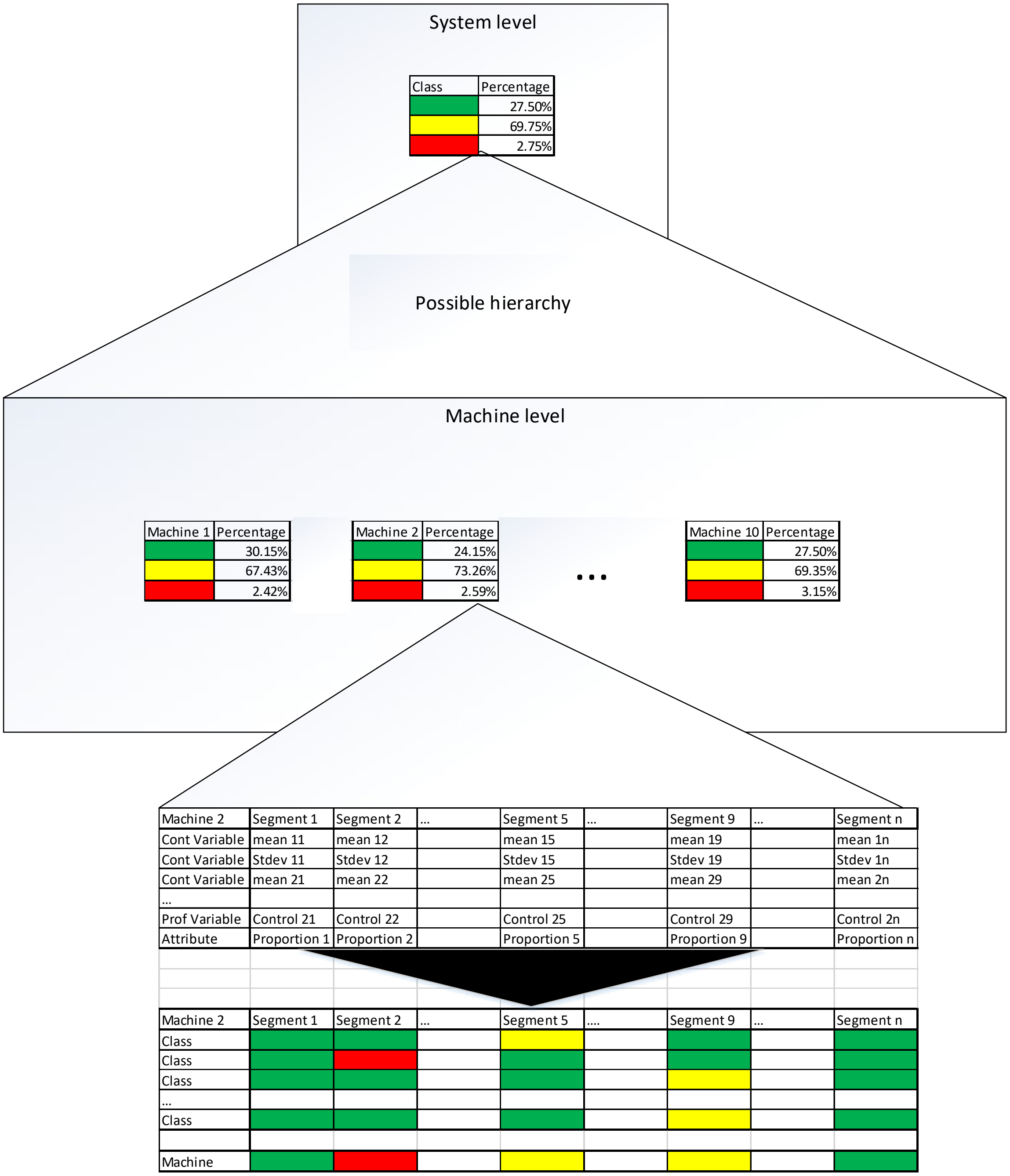
5.3.1.2. Classification and summarization

After the characterizing values have been established, the method classifies each of these using ideas from pre-control charts (Satterthwaite, 1954). Each value is classified into a color group based on how far the characterizing value is from its mean value. Let w be one of the characterizing values, then $\mu_w$would be the mean of that particular characterizing value and σ_w would accordingly be the standard deviation. The mean and standard deviation would be estimated based on historical data, which has been deemed satisfactory. The classification follows guidelines from Table 21.

Table 21. The generic classification of each of the characterizing values

Value	Classification
$\mu_w - 1.5 * \sigma_w < w < \mu_w + 1.5 * \sigma_w$	Green
$\mu_w - 3 * \sigma_w < w < \mu_w - 1.5 * \sigma_w$ or $\mu_w + 1.5 * \sigma_w < w < \mu_w + 3 * \sigma_w$	Yellow
$w < \mu_w - 3 * \sigma_w$ or $w > \mu_w + 3 * \sigma_w$	Red
The machine has scheduled stop or is not used in the production schedule	White

The third stage uses the group control chart idea of recording the worst-case observation (Boyd, 1950). For each row of classes over all the variables within one machine or another predetermined group, the “worst” class is selected for that particular segment and the machine or group will be assigned to that color class. For example, if all variables are assigned to the green class except for one in yellow, then the class of this machine or group is assigned the color yellow. If there is at least one red, then the machine or group is classified as red. The next stage summarizes all the classification instances over all segments in the period under investigation by selecting the worst cases of the machines on the investigators’ level of interest (department level, factory level, etc.). In this way, the algorithm generates a frequency table (Figure 22). The last stage detects changes by comparing the current frequency table to a similar table from historic data.



Class	Percentage
🟩	27.50%
🟨	69.75%
🟥	2.75%

Machine 1	Percentage
🟩	30.15%
🟨	67.43%
🟥	2.42%

Machine 2	Percentage
🟩	24.15%
🟨	73.26%
🟥	2.59%

Machine 10	Percentage
🟩	27.50%
🟨	69.35%
🟥	3.15%

Machine 2	Segment 1	Segment 2	...	Segment 5	...	Segment 9	...	Segment n
Cont Variable	mean 11	mean 12		mean 15		mean 19		mean 1n
Cont Variable	Stdev 11	Stdev 12		Stdev 15		Stdev 19		Stdev 1n
Cont Variable	mean 21	mean 22		mean 25		mean 29		mean 2n
...								
Prof Variable	Control 21	Control 22		Control 25		Control 29		Control 2n
Attribute	Proportion 1	Proportion 2		Proportion 5		Proportion 9		Proportion n

Machine 2	Segment 1	Segment 2	...	Segment 5	....	Segment 9	...	Segment n
Class	🟩	🟩		🟨		🟩		🟩
Class	🟩	🟥		🟩		🟩		🟩
Class	🟩	🟩		🟩		🟨		🟩
...								
Class	🟩	🟩		🟩		🟨		🟩
Machine	🟩	🟥		🟨		🟨		🟩

Figure 21. The hierarchical structure of the TUFF method

For example, we now consider a small workshop, shown in Figure 22, composed of 10 machines (m1,m2,…, m10) that each has five different variables (m1.1, m1.2, …, m10.5). These numbers can vary in practice but for the purpose of illustration, we fix them at ten and five. We also assume that these variables, as well as product quality, are monitored at all times during the production. In each machine, one variable is characterized by a profile, while all the other variables are continuous. Assume the continuous variables produce 1 data point per second for 8 hours each day. The continuous variable values are simulated using a normal distribution with a mean of 0 and a standard deviation of 1. Each profile variable generates one profile every 20 seconds. Each profile consists of 20 points, matching the sampling speed of continuous variables. The underlying model in all profiles is

$$\boldsymbol{y = 1 + 2x + 3x^2 + \varepsilon} \tag{32}$$

where x is twenty equally spaced points between 0 and 1 calculated by

$$\boldsymbol{x_i = \frac{i-0.5}{20}, i = 1, 2, \dots, 20} \tag{13}$$

The $\boldsymbol{\varepsilon}$ is the error term assumed to be normally distributed with a mean of 0 and a standard deviation of 0.1 (Chapter 5). The values recorded for each segment are the two critical values at point 5 and point 15 with the assumption that these are important points in the profile.

During the transformation, each of the continuous variables is divided into sets of 20 to mimic the profile variables, followed by a calculation of means and standard deviations over each of the sets of 20 observations for each continuous variable. At the same time, the profiles are monitored by recording y values on two critical points of each run.

The outcome is a new base for monitoring with raw data transformed into statistics (Table 2). For each continuous variable (m1.1, m.1.2, etc.) two statistics are generated: m1.1mean, m1.1st.dev, m1.2mean, etc., and for each profile variable two critical values are recorded: m1.5cr.val 1, m1.5cr.val2, etc. In essence, each machine would have 10 values for each segment.

Table 22. Transformed data for system-wide monitoring

t	m1.1 mean	m1.1 st.dev	m1.2 mean	m1.2 st.dev	…	m1.5 cr.val 1	m1.5 cr.val 2	…	m10.5 cr.val1	m10.5 cr.val 2
1	120	2	3600	25	…	1.0678	4.547	…	1.0854	4.552
2	121	1.9	3585	20	…	1.0858	4.752	…	1.0674	4.514
….	….	….	…	…	…	…	…	…	…	…
1440	120	2.1	3621	21	…	1.0644	4.523	…	1.0687	4.651

Next, each new variable (m1.1 mean, m1.1 st.dev … m10.5 cr.val 2) is classified into green, yellow, or red class based on the Table 21 rules. Assume that the variable m1.1st.dev had the upper threshold of 1.03 and a lower threshold of 0.97 for the green color. Then the value of the m1.1st.dev for the first segment would be classified as yellow, while the value for the second segment would be classified as green based on the values in Table 22. The mean and standard deviation values used for Table 21 classification rules are generated using historical data on the same characterizing values from the period when the system was deemed in-control. The worst color (green is better than yellow and yellow is better than red) from all the variables in a machine is chosen to represent the machine in each segment. Next, the colors for each machine are then summarized to generate the machine level frequency table of each color. The last step is to identify the worst machines in terms of red color to represent the system-level frequency table and identify the "worst-performing" machines for each day. The monitoring of the lights will happen in the performance monitoring section.

It is important to point out that while the red percentage is an indicator of non-compliant values, it may also contain false alarms. Every distribution allows for some level of non-conformity. In this method, the probability of false alarms and a small number of out-of-control samples is acknowledged and accepted. If the current results compared to a historic timeframe were similar in terms of green and red percentages, then we would conclude that the system performance stays the same without changes. However, the red proportion exceeds its upper control limit of the p chart, the process is deemed out of control. Details will be described in the next section.

5.3.2 Performance monitoring

The Measure phase continues with the automatic monitoring of the systems' performance over a certain period of time. This is achieved by recording and monitoring the worst-case machine over a period of time, more specifically, the green and red percentages generated in the first step during day-to-day manufacturing operations. The reason for not monitoring all three percentages is that they will always add up to 100% and therefore it would be redundant to use all three. For example, if the red percentage does not change, but the green declines, the yellow percentage is increasing. Traditionally traffic lights reflect the idea that green color means good and red color means bad. In this research the colors do not mean exactly that, rather they are used just as indicators based on the pre-control chart. Color green means that a process observation statistic is close to its mean, while the color red means that the process statistic is far from its mean. Previous research (Chapter 4) has also shown that the percentage of red and percentage of green are good indicators for changes.

In the proposed method, the red and green percentage data that was collected during the monitoring is monitored using an attribute control chart with upper and lower control limits defined by formulas (14) and (15) (Montgomery, 2012).

$$UCL = p + \sqrt{\frac{p(1-p)}{n}} \tag{44}$$

$$LCL = p - \sqrt{\frac{p(1-p)}{n}} \tag{55}$$

where p is the percentage of red or green color from historical dataset deemed in-control and n is the number of percentage measurements in the historic dataset deemed in-control

First, overall or system-wide data is monitored over a period of time, for example over one month. The worst performing machines are identified by the system and recorded for each day. The machine that has the highest count of records could be considered for the improvement candidate. If there are several machines with high counts over the month, all of them can be the candidate for improvement. The system gives an alarm to the process owner and the decision step of the MDAIC process can start.

Continuing with the example of 10 machines with 5 mixed variables, the performance monitoring would start with the generation of a "good" dataset that is used for set-up. The dataset is assumed to run over 30 days. The dataset is simulated using standardized values to simplify the coding. Each continuous variable is simulated using a normal distribution with a mean of 0 and a standard deviation of 1. The profiles are generated using the underlying model (12). Two critical points are identified as explained in the previous section. Similarly, the mean and standard deviations of the continuous variables during each profile run are calculated. For each day 28800 (8 h*60 min *60 sec) measurements are recorded for each continuous variable. Since the profile generates two measurements every 20 seconds, 1440 (8 h*60 min*60 sec /20) values are calculated for each characterizing variable. The values are classified and summarized at the machine and the system level. A single worst-case machine is identified for each day. When the process is in control, the assumption is that each machine has an equal probability of being labeled as "worst-case” for each day. Therefore, the expectation is that the list of daily under-performers would include almost all the machines. The results shown in Table 23a show that as expected, almost all machines are “worst-case” at some point in the month and no machine has significantly more appearances.

Table 23. The frequency table for daily "worst-case" machines. (a) in-control, (b) Machine 1 had a small mean shift in one continuous variable

Machine	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**10**
Frequency	**3**	**2**	**4**	**5**	**3**	**4**	**2**	**4**	**3**

(a)

Machine	**1**	**4**	**7**	**8**
Frequency	**24**	**4**	**1**	**1**

(b)

Let’s now assume that there is a small shift in one variable in machine 1 and all the other machines perform on the previous levels. The shift occurs on the 7th day. The mean changes from 0 to 0.5. All the other variables are generated with the old parameters. After running through the proposed method, the frequency table (Table 23b) is generated again. Obviously, one machine has a much higher count of being “worst-case” on a daily bases. In this case, the Machine 1 frequency of 24 meant that Machine 1 was the “worst-case” machine on 24 occasions over the observed month. This would then trigger the alarm to the process owner and the decision step of MDAIC can start. Let’s also assume that we would implement a pair of control charts, say X-bar and R charts on this machine 1 variable. Since the control charts would not respond to a shift instantaneously, the frequency count for machine 1 should still be higher than those for the other machines under normal operating conditions.

5.3.3 Decision on improvements needs

Decision-makers of a system of interest can leverage the traffic light statistics generated to identify opportunities for improved in the Define step of the Six Sigma process. Based on the example used previously, machine 1 was identified as the worst performer according to Table 23b while machine 4 is the second-worst performer. Two attribute control charts should be applied to the red and green light statistics respectively on both machines 1 and 4 using formulas (14) and (15). These control charts are established to define the upper and lower control limits as well as the centerlines using the SPC Phase I process which assumes that the red and green color percentages are stationary in a certain range when the process is in control. Because the system-wide monitoring relies on retrieving the worst case from the machine level values, most of the calculations are already done in the background and can be easily retrieved. The p charts indicate that both Machine 1 and Machine 4 are within the control limits when no changes are introduced, so no immediate action is required.

As shown in Figure 24, Machine 1 shows points beyond control limits after day 6 while the signals for machine 4 are well within the limits. Based on this information, the process owner can assemble the project personnel whose expertise is connected to the machine 1. At this point, the proposed framework already leads the Six Sigma team to a continuous improvement candidate. However, a more in-depth analysis of a cause can be performed for machine 1.

5.3.4 Sensitivity of the TUFF method in MDAIC

To show the sensitivity of the TUFF method, two more simulations were carried out. In the first simulation, the same small change from mean 0 to mean 0.5 was introduced in one variable in Machine 1, while other machines performed at the original levels. The difference was that instead of the fault occurring every day after the initial incident, the fault happens every 3 days. In essence, this simulation is more in line with the actual manufacturing environment in that when the issue arises, the operators will try to correct it. However, if the root cause is not addressed and

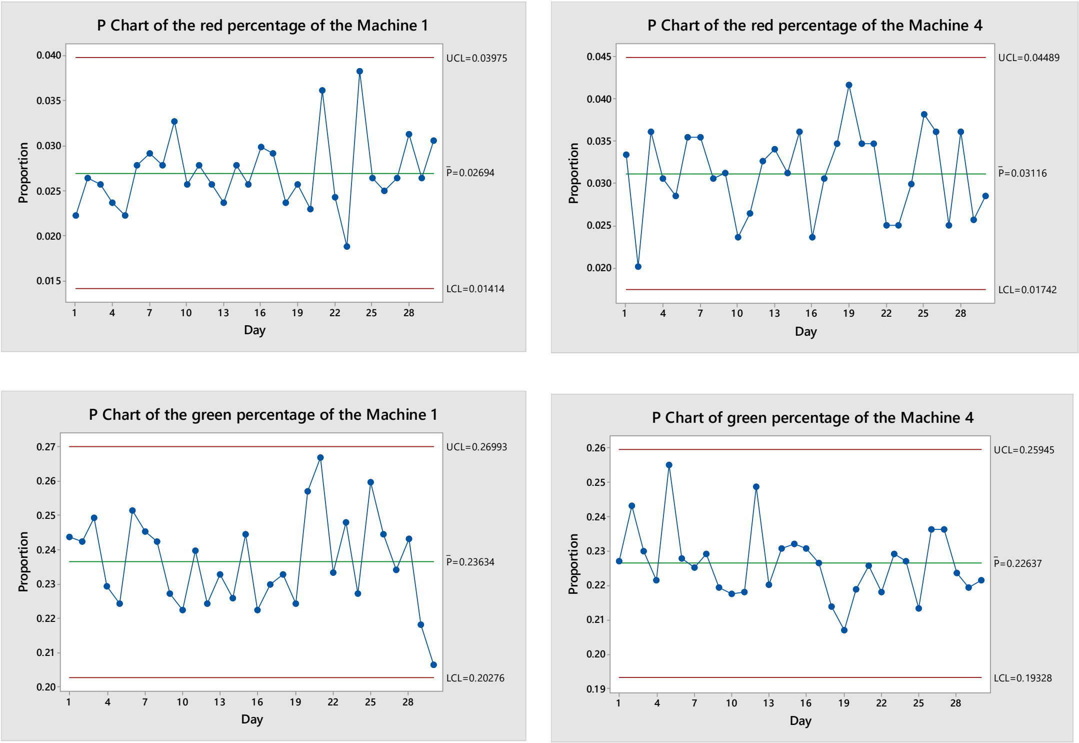


Figure 22. Attribute control charts on machine level for the red and green class for machines 1 and 4 in the in-control dataset

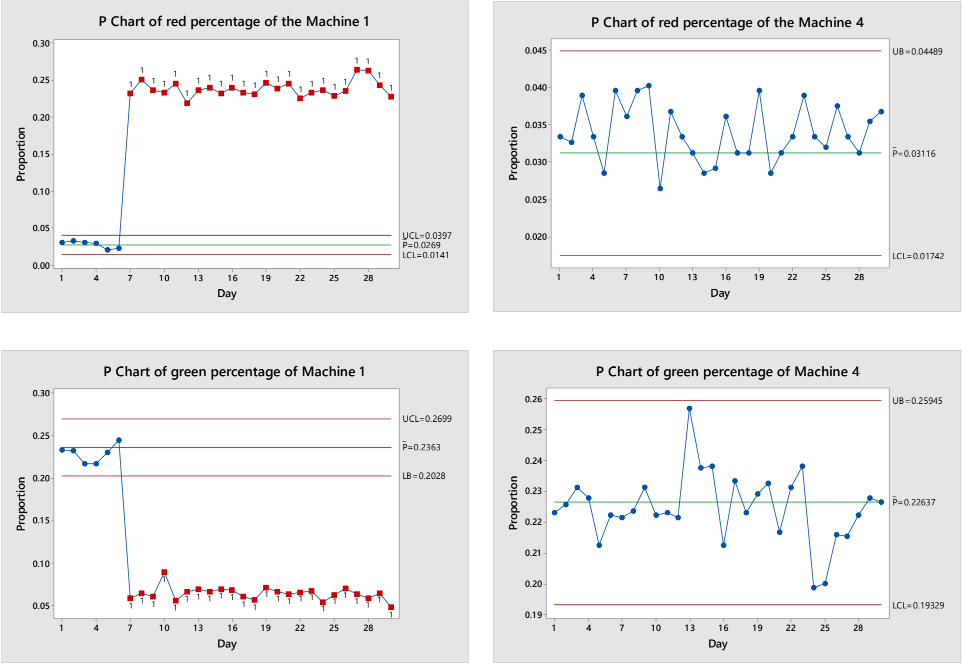


Figure 23. Attribute control charts for machines 1 and 4 for the red and green class when one variable change on day 7 in machine 1

the same issue is reoccurring, the process owner might look into more permanent solutions that are beyond the capabilities of the operators.

Table 24. The frequency table for daily "worst-case" machines after machine 1 had a small mean shift in one continuous variable every 3 days

Machine	1	2	4	5	6	7	8	9	10
Frequency	9	3	1	1	2	5	5	3	1

The frequency table for the “worst-case” machine (Table 24) shows that while there are quite a few machines such as Machines 7 and 8 exhibiting high counts, Machine 1 still has the highest count. This would trigger the Decision step for deeper analyses. The results of the attribute control chart show that even in the case when the problem is reoccurring over 3 days, the method is capable of picking up the change (Figure 6). Every three days, both red and green percentages are not within the control limits. Machine 1 would be a candidate for an Improvement project.

The second simulation was carried out to identify how well the TUFF method in MDAIC reacts to the change in a profile variable, rather than in a continuous variable. Suppose the underlying model for the profile variable was changed to equation (17) from equation (16).

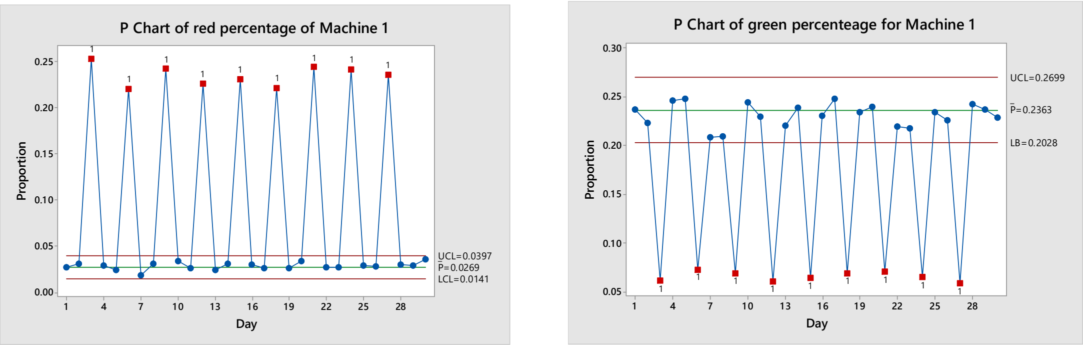


Figure 24. Attribute charts for machine-level monitoring of red and green percentages when one variable has a change once every 3 days in Machine 1

$$\boldsymbol{y = 1 + 2x + 3x^2 + \varepsilon} \tag{66}$$

$$\boldsymbol{y = 1.1 + 2.1x + 3.1x^2 + \varepsilon} \tag{77}$$

The change happened similarly to the original change sample on the 7th day. The frequency table (Table 25) shows again the large count for Machine 1 as the daily worst-case machine. The attribute control charts show that the change was caught by both red and green percentages at the machine level (Figure 26). In this case, the problem was identified after collecting one month's worth of data. This example demonstrates that the proposed framework is designed to collect system-wise problem counts in addition to locally implemented SPC methods. If a process is reset multiple times without addressing the underlying issues, p charts will identify these persistent problems.

Table 25. The frequency table for daily "worst-case" machines after machine 1 had a shift in profile variable

Machine	**1**	**4**	**5**	**6**	**7**	**9**
Frequency	**24**	**1**	**1**	**2**	**1**	**1**

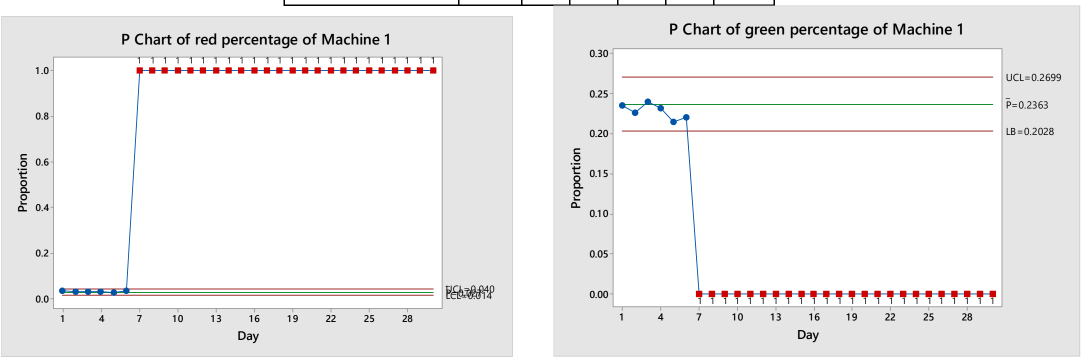


Figure 25. Attribute control charts for system-wide monitoring of red and green percentages when one profile changed on the 7th day in machine 1

5.3.5 Other scenarios with solutions

The above-mentioned examples have dealt with scenarios where all the machines are the same and perform on the same level. The purpose of the monitoring is to identify the machine that is showing most frequently to be out-of-control for the purpose of MDAIC project identification. In real life, this assumption might not be true. The following sections provide two additional scenarios that might occur as well.

5.3.5.1 Mixture of old and new machines

In this case, the assumption is that there are a few old machines and a few new machines in a manufacturing facility. Obviously, these machines would perform at different levels. Old machines are more prone to issues and breakdowns than the new ones. If the process owner runs the proposed method as-is, the assumption would be that usually, the old machines would be the ones that will be deemed as Improvement candidates. On the one hand, it is completely expected. But on the other hand, if the machines cannot be replaced or cannot be improved to perform at the level of the new machines (assuming they perform at the level that is acceptable for their purpose), the process owner might want to monitor the newer machines as well to detect if there might be candidates for Improvement. The solution for this type of scenario would be to alter the proposed method slightly. In the last steps, instead of picking the "worst-case" machine as the representative of a department, for example, the method would average each color over the machines to represent the department. Furthermore, the average each light over all departments represents the factory. In this way, if the old machines are operating at the same level as usual and one or a few of the new machines are operating worse than that previously, the proposed method would be able to detect the changes. The system performance monitoring would start from the top-level percentages. If a change is detected on the factory level using attribute control charts over a period of time, for example, 30 days, the method should look for changes on the department level and after identifying the department, on the machine level of that department and so on. Because the average of light count instead of a worst-case statistic is used for a system performance measure, we can detect changes in the new machines as well on a daily basis.

5.3.5.2 Heterogeneous of machines

So far the examples examined have three levels: machine, department, and factory. It is viable to consider the overall factory performance assuming the departments are equally important. However, the departments may be very different so it makes no sense to generate performance statistics at the factory level. For example, there are three or more groups of machines. Each group has similar performance levels and similar tasks within the group that are different between the groups, says, a group of metal sheers, a group of stamping presses and a group of bending breaks. While the monitoring of the overall worst machine would be useful, more information might be available for each group. So the method would identify the "worst-case" machine for each group. If some of the machines are identified more often, while maybe not necessarily as the "worst-case" of the overall facility, these machines might be candidates for the Improvement project to enhance the overall quality performance of the whole factory.

5.4. Big Data MapReduce Strategies and Algorithms

In this section, the focus is on proposing a solution on how the TUFF method in MDAIC would be applied in the big data environment. Assume that the file for the manufacturing data over 100 days is larger than the traditional software such as Excel is capable of managing, therefore big data methodology is used. The solution uses a MapReduce algorithm to analyze large sets with the proposed method.

MapReduce is a framework for executing highly parallelizable and distributable algorithms across large datasets using hundreds or thousands of commodity computers (Lublinsky *et al*, 2013). A MapReduce algorithm does parts of the calculations in the server that the data segment is stored in parallel (hence the name "parallel computing").

MapReduce consists of two procedures that users must write: mapping and reducing. The system manages the parallel execution, coordination of tasks that execute mapping or reducing and also deals with possible failure handling. In the mapping procedure, the data segment in each server is split, sorted and filtered. If needed, other calculations are also carried out. Users must define two critical parameters that are used as the input and output of each server: key and value. The key is the identification parameter that depends on the goal of the algorithm and the value is the output of the segment in that server. All the key-value pairs are collected by the master controller and divided among all Reduce tasks in a way that all the pairs with the same key end up in the same Reduce task (Rajaraman and Ullman, 2012).

In the reducing procedure, the outputs of the mapping procedures are shuffled and sorted based on the key defined in the mapper and then reduced by combining the values defined previously in some manner defined by the user.

A MapReduce method is applied to the workshop example presented earlier (Figure 27). The proposed method is similar to the word count problem in MapReduce. The goal of traditional word count is to generate a list of all words in a certain text and count how many times each word occurs. In this case, the goal is to calculate mean and standard deviations from variable data with date and time information on machine and date level. In the mapping procedure, the variable data is split into smaller segments for servers to process. Then each machine and date in the segment is defined as key and has a value of the reading. In the reducing process, all the key-value pairs generated in the mapping procedure are shuffled together based on the keys and then all the averages and standard deviations for each key are calculated. Finally, all the reduced key-value pairs are merged and reported for the final output.

The MapReduce paradigm is powerful, but it does not provide a general solution to all big data problems. While it works particularly well on some problems, others are more challenging. It is shown to work well on summarization problems including mean and standard deviation calculations and counting; filtering problems including distinction problems; data organization problems like changing structured data to hierarchical, partitioning, binning, shuffling and total order sorting; and joining problems, etc. (Miner & Shook, 2013).

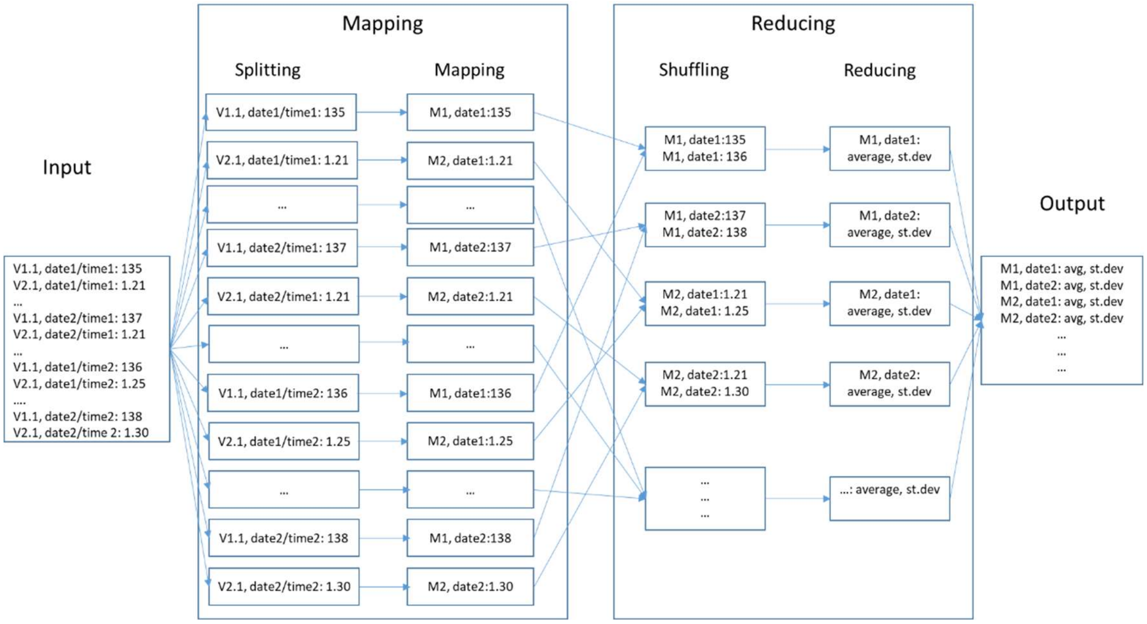


Figure 26. Flowchart of a MapReduce method applied to a manufacturing facility

5.4.1 Indexing in the Step of Measure

The identification of the location and the timeframe of change are of crucial importance in the first step (**M**easure) of the proposed MDAIC framework, especially in big data applications. This information allows an investigator to identify the areas that need more in-depth analysis and that need to be monitored more closely during the SPM operations often used in the Control step. It also helps to group data for further summarization. For example, a process engineer may be interested in the performance of one machine during week 25 as compared to week 23 and 24. On a different level, a factory manager may be interested in comparing the performance of line 1 to line 3 in March and April. The data should be prepared so that this information would be easily accessible and summarized for different levels of interest. This opportunity-seeking stage can be automated by computing various time intervals called scales. Comparisons of the statistics generated in each scale should be sent to managers when a certain number of changes over a threshold is reached.

The method uses timestamps and machine identification as the key to identifying the place and the time of changes. The timestamps must be translated into regular mm/dd/yyyy hh:mm:ss format, which allows fast identification of any timeframe of interest. It also allows multiple-scale analysis since the timescale can be set as hourly, daily, weekly, monthly, or yearly bases. The same thing must be done with the machine-id, which should contain various identifying features, such as machine identification, department identification, factory identification, etc. Again, depending on what an investigator is interested in, a summation of the data according to various scales can be easily accomplished even for a large amount of data. This scalable data identification method used in the big data environments can help identify the variables/machines/departments in the

Measure step of the redefined Six Sigma method that can identify potentially good candidates for Improvement based on data collected.

The indexing is used in the “key” of the MapReduce algorithm described previously. Depending on the length of time the investigator has set, the timestamp containing date/time information is separated and the set level of time is assigned to serve as a part of the “key” in identifying each raw or calculated value that is later summarized in the reduce section. The second part of the “key” would be the identification of the machine.

5.4.2 Indexing and MapReduce Example

Assume that production data is in the JavaScript Object Notation (JSON) file format (json.org). To analyze the data, the data file first needs to be converted into separated value files such as in the CSV format. A JSON file is often organized in key and value pairs in which each value could be paired. It is one of the most often used data formats in browser/server communications. Each row in the dataset contains different values of states such as machine_id, department_id, sequence number, event begin time, event end time, event description, a reason for an event, and so on. For example, a row might contain the following syntax:

{ "_id" : { "$oid" : "4f6abdbcf437c071d940a3a2" }, "tm" : 1332395360, "category" : "Utilization", "component" : "v2", "dataitemid" : "", "machine_id" : 3, "mt_name" : null, "mt_value" : "ACTIVE", "sequence" : 2037056, "subtype" : null, "type" : "Execution", "virtual_flag" : "N", “Reading”: 102.547 }

In this structure, the “_id” is the key and the rest of the terms are values. The interest of an investigator is in the value parts of the file. In the value part, the first important component for the method purposes is “tm” which is the time stamp identifying the time of the reading. Next, the machine id and the component id need to be extracted and finally the reading field. It is very important to note, that these key and value pairs are not defined the same way as the MapReduce key and value pairs, which will be extracted from JSON or CSV format and redefined.

Assume the same example of 10 machines with 5 variables in each machine that was used previously. In order to achieve the data input for the Define stage of the newly proposed method, different MapReduce algorithms would be used. Two MapReduce codes are used when all the data is the same, either continuous, attribute or profile type; three are used when the data is a mixture of continuous, attribute and profile. The first code would calculate the averages and standard deviations in case of mixed data types. The second code would classify the values while the third code would summarize the classes to generate the traffic lights for monitoring.

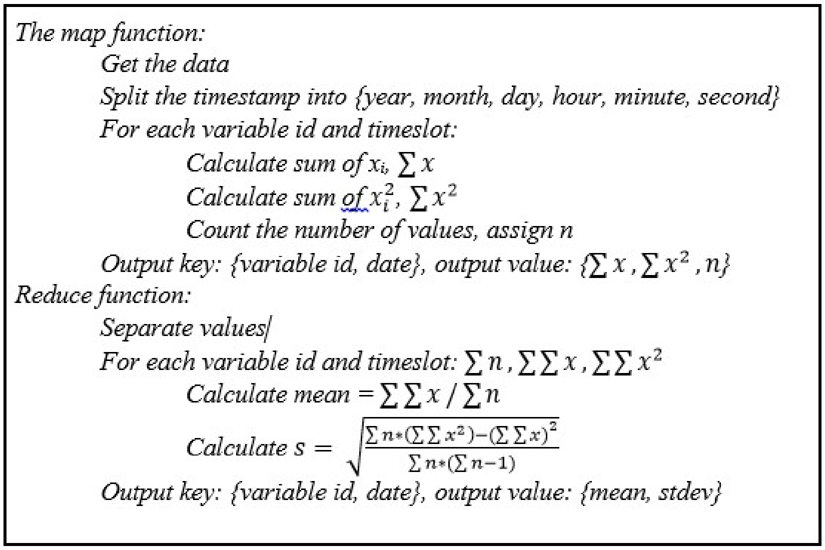
The first MapReduce code would use the indexing proposed previously to extract the machine id, variable id, date and time info as well as the actual reading from the distributed JSON file and map the extracted data into key and value format. The key would be the machine id, the variable id and the date/time combo that is the same length as the profile in the same machine. Then the reducer would shuffle all the keys together and reduce it by calculating the mean and standard deviation for each variable in each machine. The results are saved as a new distributed JSON file. The mean and the standard deviation can be calculated using formulas (18) for mean and (19) for standard deviation. The MapReduce pseudocode for that calculation is presented in Figure 28.

$$\overline{\boldsymbol{x}} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i}{\boldsymbol{n}} \tag{88}$$

$$\boldsymbol{s} = \sqrt{\frac{\boldsymbol{n} * \sum \boldsymbol{x}^2 - \left(\sum \boldsymbol{x}\right)^2}{\boldsymbol{n} * (\boldsymbol{n} - \mathbf{1})}} \tag{99}$$

where $\boldsymbol{x_i}$ is individual the i-th reading

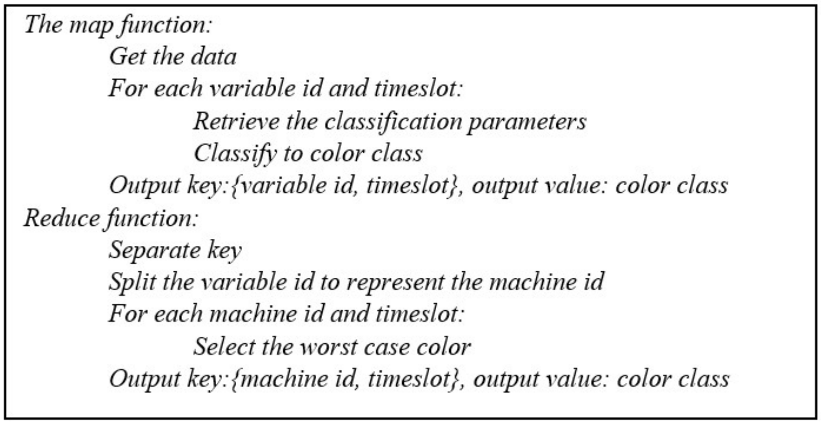
n – number of individual readings



```
The map function:
        Get the data
        Split the timestamp into {year, month, day, hour, minute, second}
        For each variable id and timeslot:
                Calculate sum of $x_i$, $\sum x$
                Calculate sum of $x_i^2$, $\sum x^2$
                Count the number of values, assign n
        Output key: {variable id, date}, output value: {$\sum x$ , $\sum x^2$ , $n$}
Reduce function:
        Separate values|
        For each variable id and timeslot: $\sum n$ , $\sum\sum x$ , $\sum\sum x^2$
                Calculate mean $= \sum\sum x \,/\, \sum n$
                Calculate $s = \sqrt{\frac{\sum n*(\sum\sum x^2)-(\sum\sum x)^2}{\sum n*(\sum n-1)}}$
        Output key: {variable id, date}, output value: {mean, stdev}
```

Figure 27. Mapreduce pseudocode for calculating the mean and standard deviation of a large dataset with multiple variables and timeframes

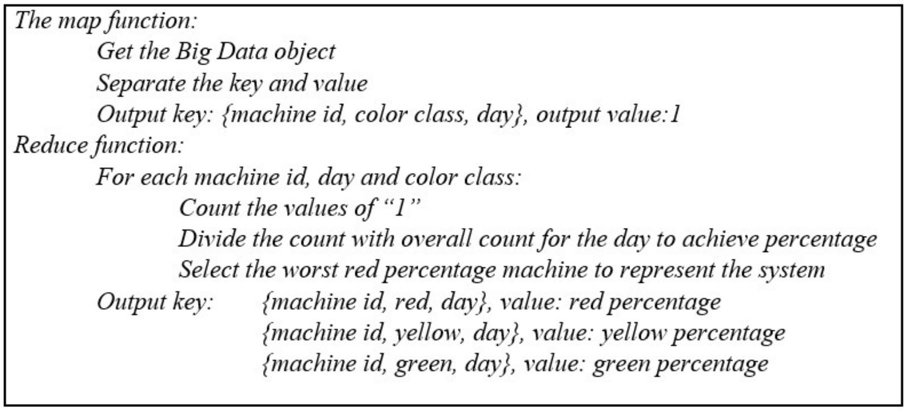
In the second MapReduce code, the mapper would extract the data from the JSON files similarly to the first code. The machine id, the variable id, and the time and date data are used as the key and reading (or statistics such as averages, standard deviation and the characterizing values for the mixture case) are the values. For each variable, the classification parameters as explained in Table 21 are retrieved with the assumption that these are previously established. Each variable is classified to either red, yellow or green for each timeslot and the color class is used as the output value of the mapping stage, while the key has identifications for variable and timeslot. For the reducer operation, the similar keys are shuffled together, the worst case is established, and each machine is assigned that color class for that time period. The key is now machine id, date and time, while the output value is the color class. This is saved as a big data object in the Hadoop File System (HDFS, 2019).



```
The map function:
        Get the data
        For each variable id and timeslot:
                Retrieve the classification parameters
                Classify to color class
        Output key:{variable id, timeslot}, output value: color class
Reduce function:
        Separate key
        Split the variable id to represent the machine id
        For each machine id and timeslot:
                Select the worst case color
        Output key:{machine id, timeslot}, output value: color class
```

Figure 28. Mapreduce pseudocode for classifying the characterizing values and machines

In the third MapReduce code, the mapper would use the big data object generated in the previous step. The first step would be to change the key into machine id, color class, and day identifier. The value is assigned 1. In the reducer side, the keys are shuffled together again and counts of green, yellow and red are generated and transformed into percentages. The worst daily performer is chosen as the representation of the system for the day. The output would give the traffic lights for each day, which would then be fed into attribute control charts and then analyzed.



```
The map function:
        Get the Big Data object
        Separate the key and value
        Output key: {machine id, color class, day}, output value:1
Reduce function:
        For each machine id, day and color class:
                Count the values of "1"
                Divide the count with overall count for the day to achieve percentage
                Select the worst red percentage machine to represent the system
        Output key:     {machine id, red, day}, value: red percentage
                        {machine id, yellow, day}, value: yellow percentage
                        {machine id, green, day}, value: green percentage
```

Figure 29. MapReduce pseudocode for summarizing the color to produce daily frequenzy table

5.5. Conclusions

In this research, a new Six Sigma process called MDAIC (Measure, Define, Analyze, Improve and Control) is proposed for the manufacturing discipline. In the era of big data where all the variables in the factory can and are monitored and stored in large amounts, the proposed framework is more data-driven so that the opinion of the expert on the area needing improvement is considered after the issue has been identified based on the data. A Statistical Process Monitoring framework called Technique of Uniformally Formatted Frequencies (TUFF) is presented that can take into account continuous, profile and attribute data to help identify the machine that could be a candidate for the Improvement section of the MDAIC from the data point of view. An example based on simulated datasets is provided to demonstrate how the proposed method works and the sensitivity of the proposed method reacting to process variable shifts.

The proposed method is expected to be applied in a big data environment. A solution is provided that uses MapReduce methodology to implement the monitoring method in big data. Pseudocodes for different steps are provided.

While the proposed redefinition of the DMAIC principles is applied to the manufacturing industry because of the abundance of measurements and data available, it is not the only industry where the statistical process monitoring could be used. For example, a large fast-food chain could also benefit from the proposed method. In this case, several performance variables may include time from order to payment, time from ordering to receiving, the portion of mess-up orders to all orders, etc. If these variables are monitored, then they can be classified into three colors and summarized. Restaurant locations are similar to machines in the manufacturing case. Additional levels of hierarchy may include city, state, region level, and country level. On any level, one month's (or quarterly) worth of data may be analyzed using the proposed method and the worst performing location could be determined. These locations at various levels would then be candidates for continuous improvement.

Future research may include the extension of the proposed approach to other application areas of Six Sigma such as healthcare. Other interesting paths would be to investigate if adding weights to variables based on the criticality of the measurement would help in any way with more accurate monitoring. The current method assumes all process variables are of equal importance. Finally, it would be interesting to investigate if separating quality characteristics and process variables into two monitoring groups would have provided more diagnostic information.

Chapter 6 - A Visualization tool for Multivariate Process Monitoring in Data-abundant environment using Adjusted Modified Sample Entropy

6.1 Introduction

The advance in sensor and data storage technologies has increased possibilities for monitoring and diagnosis in different production processes. Processes are becoming more complex, and monitoring devices are generating additional data. Although data analytics possibilities have grown with the introduction of big data, inherent statistical control challenges including process capability analysis have also grown. Process capability analysis is often used to compensate for variability, but the increasing amount of data has complicated process capability analysis (Montgomery, 2012). Traditional methods such as process capability ratios cannot manage large volumes of process data over time because these methods assume that the process has a normal distribution and does not change. There have been some proposals for using entropy types of methods for process capability studies. The reasoning is that entropy is not dependent on distribution and so it can still provide accurate process assessment in changing distribution environments.

6.2 Background

6.2.1 Modified Sample Entropy

In this report, we use entropy to detect changes in time series. Entropy is defined as the average amount of information contained in each message received. *Message* refers to an event, sample, or character drawn from a distribution or data stream. Entropy is also used to measure chaos in the data signal. The fundamental algorithm for entropy was derived from Shannon's theory published in 1948 (Shannon, 1948) by Andrei Kolmogorov in 1963 (Kolmogorov, 1998). The base of the algorithm was the calculation of probabilities of chances that a particular message was actually transmitted, and the entropy of the message system was a measure of the average amount of information in a message.

Several proposed methods that use heuristics to compute approximate entropy for finite input data have been derived from Kolmogorov's algorithm. The two most common modifications are Approximate Entropy and Sample Entropy (SampEn) (Grassberger, 1988; Grassberger &Procaccia, 1983). Because studies have shown SampEn to be more accurate when data length varies, (Grassberger, 1988; Grassberger &Procaccia, 1983) it was used in this application.

Different modifications of SampEn were reviewed by Humeau-Heurtier (2015). One modification, Modified Sample Entropy (mSampEn) by Xie *et al* (2010), showed more promise for stable performance and precision in our simulation studies. However, Sample Entropy typically does not have an output when the sample size is small (less than 100 samples; a sample size of 20 has been suggested to obtain readings. In addition, tolerance *r* cannot be very small: the entropy gives outputs until $r = 0.2$ if the sample size is larger than 200 observations, but it does not provide outputs below that value of r. If the sample size is 100–200, tolerance must be approximately 0.6* standard deviation. Both of these shortcomings occur because the process in Sample Entropy contains unit step function. When the algorithm compares the data points, it provides an output of 1 (if the data point is less than the threshold) or 0 (if the data point is larger than the threshold). This function has a strict distinction for membership and also is not continuous. Therefore, Xie *et al* (2010) proposed an alternative function, the Modified Sample Entropy algorithm, to be used instead of the Heaviside function. They proposed the use of the fuzzy membership function in order to represent the similarity degree between two data points.

Calculation of Modified Sample Entropy requires the following steps. For *N* points normalized time series $\{u(i){:}\, 1 \le i \le N\}$, the vector sequence takes a form similar to the definition of SampEn:

$$X_i^m = \ (u(i), u(i+1), \dots, u(i+m-1)) - u0(i) \qquad 1 \le i \le N-m+1 \tag{20}$$

X_i^m is generalized by removing a baseline

$$u0(i) = \frac{1}{m}\sum_{j=0}^{m-1} u(i+j) \tag{21}$$

Then distance d_{ij}^m between vectors X_i^m and X_j^m is defined as

$$d_{ij}^m = d\left[X_i^m, X_j^m\right] = \max|u(i+k) - u0(i) - (u(j+k) - u0(j))|,\ k \in (0, m-1)\,, i \ne j$$

(22)

The similarity degree D_{ij}^m between X_i^m and X_j^m is determined by a fuzzy membership function

$$D_{ij}^m = u(d_{ij}^m, r) \tag{23}$$

A fuzzy membership function such as Gaussian or Sigmoid bell shape can be used if the function is continuous in order to prevent the similarity from changing abruptly. The function must also be convex, ensuring that self-similarity is maximized. In this report we use the following Sigmoid function:

$$u\left(d_{i,j}^m, r\right) = \frac{1}{1 + \exp\left(\dfrac{d_{ij}^m + 0.5}{r}\right)} \tag{24}$$

where $u\left(d_{i,j}^{m}, r\right)$ is the similarity output, d_{ij}^{m} is the distance between any x_i and x_j, and r is the threshold.

Similar to the definition of SampEn, for each vector X_i^m, averaging all similarity degrees of its neighboring vectors X_j^m results in

$$C_r^m(i) = \frac{1}{N-m-1} \sum_{j=1, i \neq j}^{N-m} D_{ij}^{m} \tag{25}$$

The cumulative probability is

$$C_r^m = \frac{1}{N-m} \sum_{i=1}^{N-m} C_r^m(i) \tag{26}$$

Then $(m+1)$-dimensional embedding vectors $X_i^{m+1} = \{u(i), u(i+1), \dots, u(i+m)\}$ are formed, and C_r^{m+1} is defined using X_i^{m+1} and steps described previously. For finite datasets, *mSampEn* can be estimated from

$$mSampEn = -ln\, \frac{C_r^{m+1}}{C_r^m} \tag{28}$$

6.2.2 Adjusted Modified Sample Entropy

SampEn can detect variance change by counting the number of data points that fall within the threshold of the value r. This counting mechanism, which is based on the overall variance of a time series, enables SampEn to handle variance-change detection. Unfortunately, however, the algorithm does not contain an element to react to mean-level shifts. Therefore, Adjusted Sample Entropy (AdSEn) was derived (Kong *et al*, 2015) in which the core component is the transformation of original time series based on input in order to synthesize the mean shift and variance changes. The mean shift in a sample dataset must be converted into variance change via input transformation. The algorithm transforms the original time series of data into a new time series by multiplying the series as follows:

$$\boldsymbol{y_{ij}} = \boldsymbol{x_{ij}} \left(\left| \frac{\boldsymbol{\overline{x}_i - \mu}}{\boldsymbol{\sigma}} \right| + \mathbf{1} \right) \tag{29}$$

where $\boldsymbol{\overline{x}_i}$ is the estimated mean of i^{th} segment set, $\boldsymbol{\mu}$ is the desired mean of the variable of interest, and $\boldsymbol{\sigma}$ is the desired standard deviation of the variable of interest. Simulations for this study showed that similar transformation as described above can be applied to Modified Sample Entropy to achieve mean shift and detect variance change detection.

6.2.3 Visualization

The visualization is done by using two concepts: trellis display and star glyphs

6.2.3.1 Trellis Display

A trellis display is a lattice-like arrangement that organizes plots into rows and columns on multiple pages. Plots on the different fields can be histogram, kernel density plot, theoretical quantile plot, two-sample quantile plot, strip chart, bar plot, scatterplot, or parallel coordinate plot (Sarkar, 2008). Each panel contains a subset of the data graphed by the plots.

6.2.3.2 Star Glyphs

Star glyphs (also called star plots) display a multivariate dataset in a geometrical shape, such as a hexagon for six-dimensional data. In a star glyph, each graphic represents a vector of observations at a particular time. A star contains *n* spikes (e.g., *n* = 6) that evenly radiate from the center of the graph. The angle between each spike is equal to 360/*n* degree, and each spike has a value with the same proportion of the variable for that observation, meaning that observations must be standardized before stars are constructed. Line segments usually connect the end of the point of each spike to neighboring spikes. In this study, each spike on a star glyph represented the ratio of each segment's measure of AdmSEn or mSampEn to the template measure.

The lattice-like arrangement of the trellis display was applied to the proposed multivariate visualization tool in this study. The block assignment was replaced by the time sequence, and the star glyphs were adopted to represent each multivariate observation on panels.

6.3. Proposed Multivariate Adjusted Modified Sample Entropy (AdmSEn) Visualization Graphs

This section presents the proposed visualization tool. The proposed method, which integrates multiple AdmSEn outcomes into one glyph, is a visualization tool for a decision support system that encounters multivariate quality characteristics. Glyphs plotted over time are organized in panels of a trellis, and every column contains one time series glyph. Each vertex in the glyph represents a response that shares equal space around 360 degrees. The corresponding vertex would either grow or shrink in length depending on the output of the algorithm. The following steps describe the main procedure of the proposed AdmSEn multivariable glyphs.

Step 1: Import data from collection devices.

Step 2: If a known target and control standard deviation are present, use the existing μ and σ for each variable. Otherwise, select a reference sub dataset that is representative of the process under study and calculate the sample mean and sigma in the transformation function.

Step 3: Input transformation: For each variable transform the original vector observations $\boldsymbol{x}$ into $\boldsymbol{y}$, that is,

$$\boldsymbol{y_{ij}} = \boldsymbol{x_{ij}}\left(\left|\frac{\overline{\boldsymbol{x}}_{\boldsymbol{i}} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}\right| + \mathbf{1}\right) \tag{30}$$

Step 4: Set manipulation parameters for mSampEn. Users can define parameters such as threshold r, length m of comparing vector, delay τ, and select multiscale resolution k, which determines how many subsets will be broken into in the original dataset. Threshold r is the selection distance between vectors, m is the dimension of the vector, and τ is determined according to the selected vectors (τ is recommended to be 1 for best use of full-size dataset), and scale k is the number of slices resulting from segmentation. If the specific parameters are unknown, a user must define scale number k; the rest of the parameters will be set as $r = 0.6$, $m = 2$, and $\tau = 1$ as default while $k = 2$ as the minimal number for two segments on the dataset.

Step 5: Graph the outcomes as glyphs for results.

The size of the spike is the value of the output. With a stable and in-control process, a glyph on the trellis panel should be presented as a circle (when the number of spikes is more than 6), because the entropy or the amount of new data is less than thresholds determined in a separate simulation study that is waiting for publishing. Variable names are plotted in red or green depending on if the threshold has been crossed or not, respectively, in order to identify the possible out-of-control variable. If a spike shrinks toward the center of the circle, the process is in control because the entropy is smaller in value. However, if the spike expands toward the outside rim of the glyph, the variable is demonstrating a higher output of entropy. High entropy on the transformed vector $\boldsymbol{y}$ means that the segment under consideration has changed due to mean shifts or variation changes. High entropy on the original data vector $\boldsymbol{x}$ indicates that only the variance has changed.

6.4. Method Demonstration

Ten multivariate processes are considered in this section to demonstrate the proposed method. Each variable contained 1200 observations. This section also explores various mean shift and variance change combinations. The dataset was generated from multivariate normal distributions. The proposed AdmSEn algorithm was used on the transformed dataset ***y***, while the mSampEn algorithm was used on the raw dataset ***x***. The code for SampEn was obtained from R computer language package *pracma*, and the AdSEn code was coded by Zhang (2015). Both codes were modified to fit the mSampEn algorithm. Visualization codes in R computer language were adopted from Vaughn (2013) and modified to fit the needs of this work.

Table 26 shows a simulated time series when the process is in control. Four segments were generated for each dataset of 1200 observations. Entropy values were then computed for each segment using AdmSEn and mSampEn algorithms. The first row in Figure 31 contains entropy values for the AdmSEn performed on the transformed dataset ***y***; the second row shows entropy values for the mSampEn on the original dataset ***x***.

Table 26. Process parameters for in-control circumstances

Segment (i)	1	2	3	4
Samples	1-300	301-600	601-900	901-1200
Mean	0	0	0	0
Standard deviation	1	1	1	1

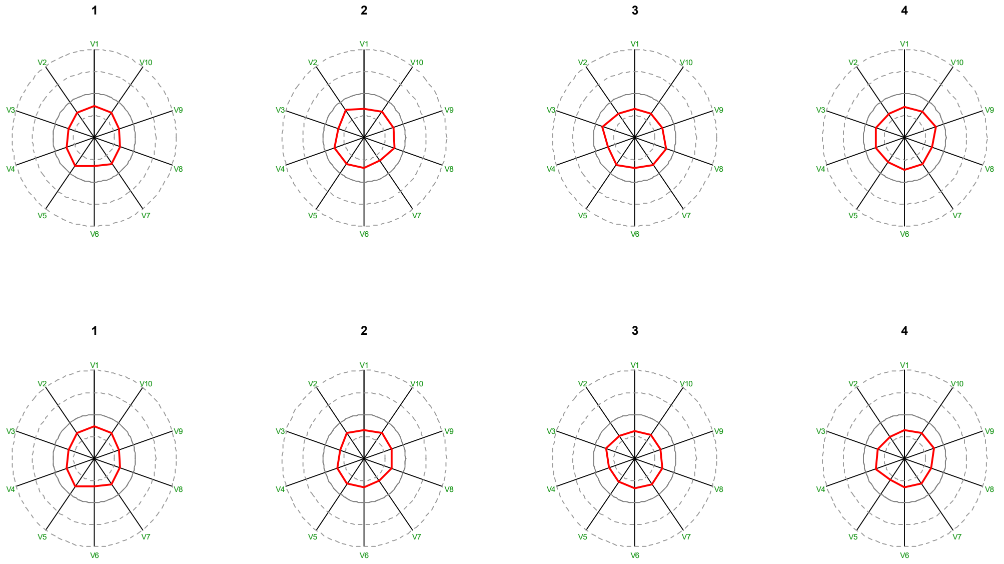


Figure 30. Multivariate visualization glyphs when the process is in conrol

As shown in Figure 31, all outcomes are under the threshold and all the variables are green, leading to the assumption that all processes were under control. In addition, the shapes representing the outputs are similar in the figure, so no deviation from the mean or standard deviation was detected since the simulated data was set up to have characteristics of no change in mean and variation.

Next, the conditions when either process means or variation levels or both processes mean and variation levels changed are analyzed. The proposed framework identified these changes and variables that contribute to changes.

Table 27 shows the cases in the second visualization demonstration. The first segment consisted of 300 observations that were simulated as an in-control dataset. The rest of the segments represented cases when process parameters of either mean or variance changed. The first segment was used as a template, the second segment represented mean change, the third segment represented variance change, and the fourth segment contained mean shift and variance change.

Table 27. Process parameters for out-of-control circumstances

Segment (i)	1	2	3	4
Samples	1-300	301-600	601-900	901-1200
Mean	0	1	0	1
Standard deviation	1	1	1.5	1.5

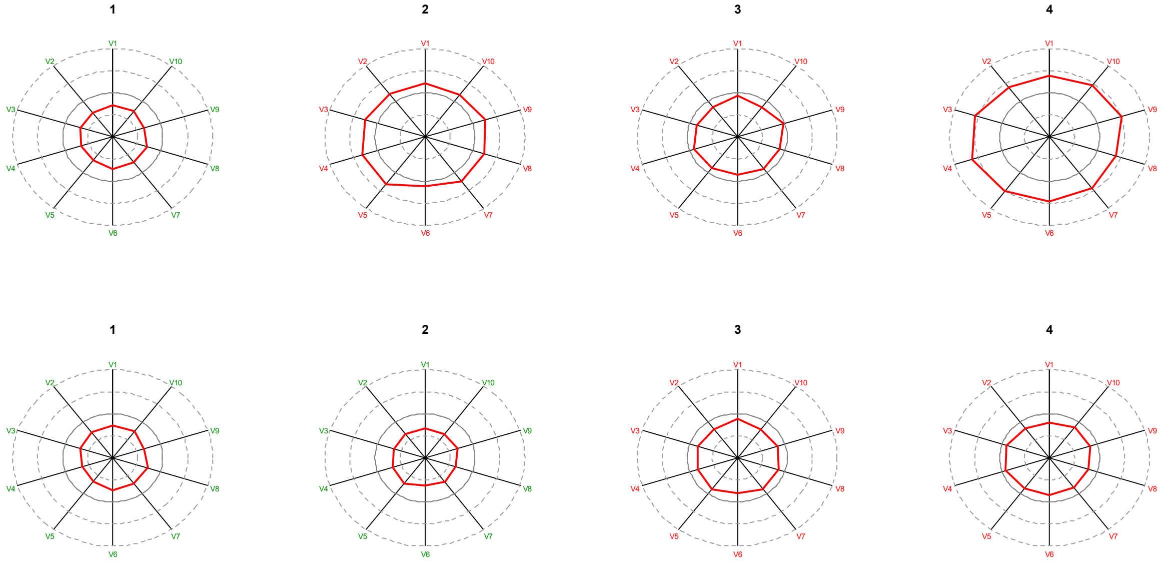


Figure 31. Multivariable visualization glyphs when the process has out-of-control signals

As shown in Figure 32, the AdmSEn algorithm on the top row detected changes in mean and variance, while mSampEn on the bottom row detected only variance changes. The figure also clearly demonstrates which variable was out of control (colored red on the graphs). The change from segment 1 to segment 2 (labeled as 2) was mean shift because the shape of the entropy outputs became much larger on the top row, but the shape in the bottom row remained constant. The change from segment 1 to segment 3 (labels as 3) was variance change because shape changes were identified by AdmSEn and mSampEn: the sizes of the shape changes were approximately the same for the top row and the bottom row. The change from segment 1 to 4 (labeled as 4) was a combination of mean shifts and variance changes. The magnitude of shape change on the top row was much bigger than the shape changes on the bottom, indicating that mean shifts and variance changes occurred.

6.5. Conclusions and Future Studies

This study provided a visualization tool to help analyze time-series data in order to increase effective, quality-related decision making when the number of dimensions is large (i.e., up to 50 variables). The tool helps users perceive mean shifts and variance changes and identify which out-of-control variable(s) are responsible for the shifts or changes.

Future research should consider that because this tool assumes that variables have no correlation, the proposed star-glyph does not clearly indicate if the variable measured at one station is directly correlated with a variable measured at another station. Also, in this study, the highest number of dimensions explored was 50. The visualization tool should be expanded to accommodate more dimensions without compromising the tool's effectiveness. For example, the angles between spikes for 100 variables in this study were 3.6 degrees, but the more spikes within the circle, the more clutter the spikes cause. Factories may want to simultaneously monitor more than 100 sensors or variables, thereby requiring additional studies.

Chapter 7 - Conclusions

7.1 Summary

The big data applications have widened the possibilities for many lines of business. The proliferation of sensors and the expansion of data storage capability has enabled increasing amounts of manufacturing data to be collected. This modern phenomenon, coupled with the complexity in the manufacturing processes, has revealed shortcomings of traditional quality monitoring methods. First, the traditional methods rely on taking a sample from the overall pool of readings and base the recommendations on that sample. The reason for sampling has historically been the high cost of gathering and analyzing the data. The result is that the data is usually scarce, the collection rate is low and the frequency of data points is also low. The second issue is that the traditional methods more often than not monitor only the quality characteristics of products with the assumption that if these are in control, the process parameters are also in control as well. The third problem is that the traditional methods only look at characteristics at the level on which they occur, assuming independence from previous steps. The opportunities offered by big data analytics can change that.

Since current or historic data is stored in the cloud, the proposed big data quality monitoring framework can analyze each reading on various scales. The proposed system-wide monitoring approach called Technique of Uniformally Formatted Frequencies (TUFF) reduces the chance that hidden or small issues are overlooked. It also allows monitoring of all the parameters on both product and process characteristics found in the factory, so the issues related to these characteristics could be detected before they emerge as major problems. By monitoring the process parameters, the TUFF methods could help improve the preventive maintenance schedule, by predicting the need for maintenance, which lowers costs. Thirdly, the TUFF methods look at the system in a comprehensive way in that they take into account the entire production steps and identify hidden correlations. Additionally, the dissertation investigates the application of the proposed system-wide monitoring method into Six Sigma continuous improvement methodology of DMAIC (Define, Measure, Analyze, Improve and Control).

In the second chapter of the dissertation, a novel method for profile monitoring was proposed. The method uses modified sample entropy to generate a single value or, in case of segmenting the profile, set of values that can be used to characterize the profile. When these values are compared to each other, the changes can be detected. The simulation study showed that the proposed method is capable of detecting the changes in the underlying model and in the variation along the profile.

The third chapter of the dissertation first identified the requirements for system-wide monitoring method. Secondly, a novel statistical system monitoring method called TUFF for continuous variables was proposed. The method relies on the ideas from the pre-control chart to classify measurements into red, yellow, green and white classes which are further used to characterize all the machines using a group control chart idea of the worst case. Over the user-specified period (e.g. hour, day, or week), a frequency table for each color class in the system is generated in the form of traffic lights. A comparison of traffic lights from the current period and historical “good” periods would show if the system has changed or not. A step-by-step example was provided to show how the method could be implemented in a big data environment. Additional simulation studies were conducted to examine the sensitivity of the method and the thresholds that decision-makers can use.

In Chapter 4 a novel method is proposed to monitor profile data in the TUFF framework. The method transforms the profile data into the same color frequency table used in the continuous variable method by characterizing the profile with values (for example the modified sample entropy from the first section) and classifying these values into color classes using pre-control chart idea. Simulation studies were carried out to identify the characteristics and study the sensitivity of the method.

In Chapter 5 the TUFF method is integrated into the Six Sigma continuous improvement methods specifically one of the more popular Six Sigma methods of DMAIC (Define, Measure, Analyze, Improve, Control). Since all or almost all variables in a manufacturing setting are monitored, the specific signal for improvement could come from data rather than initiated by a Six Sigma team. Therefore, the redefined method reorders the steps into Measure, Define, Analyze, Improve, Control (MDAIC). In addition, the statistical system monitoring methods for continuous, profile and attribute data were also integrated to form a comprehensive TUFF method using the traffic lights from the previous chapters. A big data implementation solution using MapReduce algorithms was proposed to detect changes and to identify time and space where the changes occur.

In Chapter 6 a novel visualization tool based on star glyph is proposed to examine a large number of variables simultaneously. The statistics displayed are the modified sample entropy. Visualization helps to identify possible changes. The proposed method works in different settings. The main contributions of this dissertation are:

- The redefinition of Six Sigma method of DMAIC to MDAIC where the unit in need of improvement is identified automatically by data
    - A solution for the method to be used in big data environment based on the MapReduce algorithm is proposed

    - Simulation studies that explore the sensitivity and other characteristics of the proposed method are presented
- A comprehensive monitoring method (TUFF) that merges continuous variable, profile variable and attribute variable data to detect changes
    - The method monitors both quality characteristics and process parameters
    - The method takes into account all the readings

This proposed framework re-defines the way quality control is implemented in the modern-day manufacturing industry. The results from this research have the potential to rejuvenate data processing, reduce the production of non-conforming products, and increase the efficiency of the facility inundated with abundant data. Further, this research can be seen as the next step in the fourth industrial revolution (Schwab, 2016) as a method to monitor smart factories.

7.2 Future Studies

For future studies, the focus could be shifted to supply chain monitoring, where there would be additional variables that are not necessarily measured already and that might be more difficult to obtain. The second focus could be to investigate how the method performs in different distribution cases. In this dissertation, the variables had either normal distribution or binomial distribution. It would be beneficial to learn if other distributions have similar detection rates. The third focus could be on improving the sensitivity of the method by different classification or summarization algorithms.

The TUFF method has been presented in this dissertation to work on static databases of production data. However, since every day new sets of data points are generated, the method should be capable of working with data streams. The problem would be how to update the results when new readings become available

It would also be interesting to see if weighing variables differently based on their importance toward the process would be beneficial in monitoring the system. The more important the variable is, the higher the weight it should be assigned. Similarly, if the quality characteristics and process parameters are monitored separately, the proposed framework may provide more insights. The future research could also investigate how to use the visualization tool from Chapter 6 within the proposed method.

References

Abbas, T., Qian, Z., Ahmad, S. & Riaz, M. (2017) Bayesian Monitoring of Linear Profile Monitoring Using DEWMA Charts. *Quality and Reliability Engineering International.* Vol 33, pp: 1783-1812

Abdella, G.M., Kim, J., Al-Khalifa, K.N. & Hamouda, A.M. (2016) Double EWMA-based Polynomial Quality Profiles Monitoring. *Quality and Reliability Engineering International.* Vol 32, pp: 2639-2652

Alexander, P., Anthony, J. & Rodgers, B. (2018) Lean Six Sigma for Small- and Medium-sized Manufacturing Enterprises. *International Journal of Quality and Reliability Management*. Vol 36(3), pp: 378-397

Anthony, J., Snee, R. & Hoerl, R. (2017) Lean Six Sigma: Yesterday, Today and Tomorrow. *International Journal of Quality and Reliability Management*. Vol 7, pp: 1073-1093

Arcidiacono, G., De Luca, E.W., Fallucci, F. & Pieroni, A. (2016) The Use of Lean Six Sigma Methodology in Digital Curation. In: *CEUR Workshop Proceedings. 1st Workshop on Digital Humanities and Digital Curation DHC 2016.* Goettingen Germany November 22, 2016. ISSN**:** 1613-0073

Awad, M.I., AlHamaydeh, M. & Faris, A. (2018) Fault Detection via Nonlinear Profile Monitoring Using Artificial Neural Networks. *Quality and Reliability Engineering International.* Vol 34, pp: 1195-1210

Boyd, D.F. (1950) Applying the Group Control Chart for –x and R. *Industrial Quality Control*. Vol 7(3), pp: 22–25

Bruner, J. (2013) *Industrial Internet.* 1st ed. Sebastopol: O'Reilly Media, 2013

Chang, S.I. & Yadama, S. (2010) Statistical Process Control for Monitoring Non-linear Profiles Using Wavelet Filtering and B-spline Approximation. *International Journal of Production Research.* Vol 48(4), pp: 1049-1068

Chang, S.I., Tsai, T.R., Lin, D.K.J, Chou, S.H. & Lin, Y.S. (2012) Statistical Process Control for Monitoring Nonlinear Profiles: A Six Sigma Project on Curing Process. *Quality Engineering*. Vol 24, pp: 251-263

Chang, S.I., Tavakkol, B., Chou, S.H. & Tsai, T.R. (2014) Real-time Detection of Wave Profile Changes. *Computers and Industrial Engineering.* Vol 75, pp: 187-199

Chang, S.I. (2017) Approaches to Implement Statistical Process Control for Manufacturing in Big Data Era. *ASME 2017 12th International Manufacturing Science and Engineering*

Conference, MSEC 2017 collocated with the JSME/ASME 2017 6th International Conference on Materials and Processing. 2017, Vol 3

Charkhi, M.R.A., Aminnayeri, M. & Amiri, A. (2016) Process Capability Indices for Logistic Regression Profile. *Quality and Reliability Engineering International.* Vol 32, pp: 1655-1661

Chiang, J.Y., Lio, Y.L. & Tsai, T.R. (2017) MEWMA Control Chart and Process Capability Indices for Simple Linear Profiles with Within-profile Autocorrelation. *Quality and Reliability Engineering International.* Vol 33, pp: 1083-1094

Chou, S.H., Chang, S.I. & Tsai, T.R. (2014) On Monitoring of Multiple Non-linear Profiles. *International Journal of Production Research.* Vol 52(11), pp: 3209-3224

Chou, S.H, Chang, S.I., Tsai, T.R., Lin, D.K.J., Xia, Y. &Lin, Y.S. (2020) Implementation of Statistical Process Control Framework with Machine Learning on Waveform Profiles with no Gold Standard Reference. *Computers & Industrial Engineering.* Vol 142, article no 106325

Crosier, R.B. (1988) Multivariate Generalizations of Cumulative Sum Quality Control Schemes. *Technometrics.* Vol 30, pp: 291-303

Cuentas, S., Penabaena-Niebles, R. & Garcia, E. (2016) Support Vector Machine in Statistical Process Monitoring: A Methodological and Analytical Review. *International Journal of Advanced Manufacturing Technology.* Vol 91(1-4), pp: 485-500

Daryabari, S.A., Malmir, B. & Amiri, A. (2018) Monitoring Bernoulli Processes Considering Measurement Errors and Learning Effect. *Quality and Reliability Engineering International*. Vol 35(4), pp: 1129-1143

Darbani, F.H. & Shadman, A. (2018) Monitoring of Linear Profiles Using Generalized Likelihood Ratio Control Chart with Variable Sampling Interval. *Quality and Reliability Engineering International.* Vol 34(8), pp: 1828-1835

DeKetelaere, B., Hubert, M. & Schmitt, E. (2015) Overview of PCA-based Statistical Process-monitoring Methods for Time-dependent, High-dimensional Data. *Journal of Quality Technology.* Vol 47(4), pp: 318-335

DeKetelaere, B., Rato, T., Schmitt, E. & Hubert, M. (2016) Statistical Process Monitoring of Time-dependent Data. *Quality Engineering.* Vol 28(1), pp: 127-142

Ding, D., Tsung, F. & Li, J. (2017) Ordinal Profile Monitoring with Random Explanatory Variables. *International Journal of Production Research.* Vol 55(3), pp: 736-749

Fan, S.S., Jen, C.H. & Lee, T.Z. (2017) Modeling and Monitoring the Nonlinear Profile of Heat Treatment Process Data by Using an Approach Based on a Hyperbolic Tangent Function. *Quality Engineering*. Vol 29(2), pp: 226-243

Fasso, A., Toccu, M. & Magno, M. (2016) Functional Control Charts and Health Monitoring of Steam Sterilizers. *Quality and Reliability Engineering International.* Vol 32, pp: 2081-2091

Gajjar, S., Kulachi, M. & Palazoglu, A. (2016) "Use of Sparse Principal Component Analysis (SPCA) for Fault Detection. *IFAC-PapersOnLine.* Vol 49, pp: 693-698

Gajjar, S. & Palazoglu, A. (2016) A Data-driven Multidimensional Visualization Technique for Process Fault Detection and Diagnosis. *Chemometrics and Intelligent Laboratory Systems*. Vol 154, pp: 122-136

Gajjar, S., Kulahci, M. & Palazoglu, A. (2018) Real-time Fault Detection and Diagnosis Using Sparse Principal Component Analysis. *Journal of Process Control.* Vol 67, pp: 112-128

Gaudard, M., Ramsey, P. & Stephens, M. (2009) Interactive Data Mining Informs Designed Experiments. *Quality and Reliability Engineering International*. Vol 25, pp: 299-315

Ge, Z., Song, Z. & Gao, F. (2013) Review of Recent Research on Data-based Process Monitoring. *Industrial and Engineering Chemistry Research.* Vol 52(10), pp: 3543-3562

Gijo, E.V., Bhat, S. & Jnanesh, N.A. (2014) Application of Six Sigma Methodology in a Small-scale Foundry Industry. *International Journal of Six Sigma*. Vol: 5(2), pp: 193-211

Grassberger, P. (1988) Finite Sample Corrections to Entropy and Dimension. *Physics Letters A.* Vol 128(6–7), pp: 369–373

Grassberger, P. & Procaccia, I. (1983) Estimation of the Kolmogorov Entropy from a Chaotic Signal. *Physical Review A.* Vol 28(4), pp: 2591–2593

Gupta, V., Jain, R., Meena, M.L. & Dangayach, G.S. (2018) Six-sigma Application in Tire-manufacturing Company: A Case Study. *Journal of Industrial Engineering International.* Vol: 14(3), pp: 511-520

The Hadoop Ecosystem Table (2017), https://hadoopecosystemtable.github.io/ Accessed March 22, 2019

Haq, A., Brown, J. & Moltchanova, E. (2015) A New Exponentially Weighted Moving Average Control Chart for Monitoring Process Dispersion. *Quality and Reliability Engineering International.* Vol 31(8), pp: 1337-1357

Hotelling, H. (1947) Multivariate Quality Control Illustrated by Air Testing of Sample Bombsights. In: *Techniques of Statistical Analysis*. New York, NY: Mcgraw - Hill, pp: 11–184

Hu, J., Runger, G. & Tuv, E. (2007) Tuned Artificial Contrasts to Detect Signals. *International Journal of Production Research.* Vol 45(23), pp: 5527-5534

Humeau-Heurtier, A. (2015) The Multiscale Entropy Algorithm and its Variants: A Review. *Entropy*. Vol 17(5), pp: 3110–3123

Huwang, L., Wang, Y.H.T., Yeh, A.B. & Huang, Y.H. (2016) Phase II Profile Monitoring Based on Proportional Odds Models. *Computers and Industrial Engineering.* Vol 98, pp: 543-553

Huwnag, L., Lin, J.C. & Lin, L.W. (2018) A Spatial Rank-based EWMA Chart for Monitoring Linear Profiles. *Journal of Statistical Computation and Simulation.* Vol 88(18), pp: 3620-3649

Jahani, S., Kontar, R., Veeramani, D. & Zhou, S. (2018) Statistical Monitoring of Multiple Profiles Simultaneously Using Gaussian Processes. *Quality and Reliability Engineering International*. Vol 34, pp: 1510-1529

Jiang, Q. & Yan, X. (2014) Monitoring Multi-mode Plant-wide Processes by Using Mutual Information-based Multi-block PCA, Joint Probability and Bayesian Interference. *Chemometrics and Intelligent Laboratory Systems*. Vol 136, pp: 121-137

Jiang, Q., Yan, X. & Huang, B. (2016) Performance-driven Distributed PCA Process Monitoring Based on Fault-relevant Variable Selection and Bayesian Interference. *IEEE Transactions on Industrial Electronics.* Vol 63(1), pp: 377-386

Jiang, Q., Huang, B. & Yan, X. (2016) GMM and Optimal Principal Components-based Bayesian Method for Multimode Fault Diagnosis. *Computers and Chemical Engineering* Vol 84, pp: 338-349

Jiang, Q., Yan, X., Huang, Biao. (2019) Review and Perspectives of Data-driven Distributed Monitoring for Industrial Plant-wide Processes. *Industrial & Engineering Chemistry Research.* Vol: 58, pp: 12899-12912

Kolmogorov, A. (1998). On Tables of Random Numbers. *Theoretical Computer Science.* Vol 207, pp: 387–395

Kong, X., Chang, S.I. & Zhang, Z. (2015) A Novel Method Based on Adjusted Sample Entropy for Process Capability Analysis in Complex Manufacturing Processes. In: ASME 2015 International Manufacturing Science and Engineering Conference, MSEC 2015

Koosha, M., Noorossana, R. & Megahed, F. (2017) Statistical Process Monitoring via Image Data Using Wavelets. *Quality and Reliability Engineering International.* Vol 33, pp: 2059-2073

Koppel, S. & Chang, S.I. (2017) A Process Capability Analysis Method Using Adjusted Modified Sample Entropy. *Procedia Manufacturing.* Vol 5, pp: 122-131

Laux, C., Li, N., Seliger, C. & Springer, J. (2017) Impacting Big Data Analytics in Higher Education Through Six Sigma Techniques. *International Journal of Productivity and Performance Management*. Vol: 66(5), pp: 662-679

Lee, J., Bagheri, B. & Kao, H.A. (2015) A Cyber-physical Systems Architecture for Industry 4.0-based Manufacturing Systems. *Manufacturing Letters*. Vol 3, pp: 18-23

Lei, Y., Zhang, Z. & Jin, J. (2010) Automatic Tonnage Monitoring for Missing Part Detection in Multi-operation Forging Process. *Journal of Manufacturing Science and Engineering, Transactions of the ASME.* Vol 132(5), Article no 051010

Li, G. & Quin, S.J. (2016) Comparative Study on Monitoring Schemes for Non-Gaussian Distributed Processes. *Journal of Process Control.* Vol 67, pp: 69-82

Liu, Y., Pan, Y., Wang, Q. & Huang, D. (2015) Statistical Process Monitoring with Integration of Data Projection and One-class Classification. *Chemometrics and Intelligent Laboratory Systems.* Vol 149, pp: 1-11

Liu, J.P., Jin, R. & Kong, Z.J. (2018) Wafer Quality Monitoring Using Spatial Dirichlet Process Based Mixed-effect Profile Monitoring Scheme. *Journal of Manufacturing Systems*. Vol 48, pp: 21-32

Lowry, C.A., Woodall, W.H., Champ, C.W. & Rigdon, S.E. (1992) A Multivariate Exponentially Weighted Moving Average Control Chart. *Technometrics*. Vol 34(1), pp: 46–53

Lublinsky, B., Smith, K. & Yakubovich, A. (2013). *Professional Hadoop Solutions*. Indianapolis, IN: John Wiley & Sons, pp: 63-97

Maleki, M.R., Amiri, A. & Castagliola, P. (2017) Measurement Errors in Statistical Process Monitoring: A Literature Review. *Computers & Industrial Engineering*. Vol 103, pp: 316-329

Maleki, M.R., Amiri, A. & Castagliola, P. (2018) An Overview on Recent Profile Monitoring Papers (2008-2018) Based on Conceptual Classification Scheme. *Computers and Industrial Engineering.* Vol 126, pp: 705-728

Martinez Leon, H. C., Temblador Perez, M.C., Farris, J.A. & Beruvides, M.G. (2012) Integrating Six Sigma Tools Using Team-learning Processes. *International Journal of Six Sigma*. Vol 3(2), pp: 133-156

Megahed, F.M., Jones-Farmer, L.A. (2015) Statistical perspective on "big data". *Frontiers in Statistical Quality Control.* Springer International Publishing: Switzerland

Menafoglio, A., Grasso, M., Secchi, P. & Colosimo, B.M. (2018) Profile Monitoring of Probability Density Functions via Simplicial Function PCA with Application to Image Data. *Technometrics.* Vol 60(4), pp: 497-510

Miner, D & Shook, A. (2013). *MapReduce Design Patterns*. Sebastopol, CA: O'Reilly Media Inc.

Montgomery, D. C. (2012). *Introduction to Statistical Quality Control* 7th edition, New York, NY: John Wiley & Sons.

Noori, B. & Latifi, M. (2018) Development of Six Sigma Methodology to Improve Grinding Processes. *International Journal of Lean Six Sigma*. Vol: 9(1), pp: 50-63

Noorossana, R., Eyvzian, M., Amiri, A., Mahmoud, M.A. (2010) Statistical Monitoring of Multivariate Multiple Linear Regression Profiles in Phase I with Calibration Application. *Quality and Reliability Engineering International.* Vol 26, pp 291-303

Pande, P.S., Neuman, R.P. & Cavanagh, R.R. (2000) *The Six Sigma Way*. New York, NY: McGraw - Hill

Pathirante, S.U., Khatibi, A. & Md Johar, M.G. (2018) CSF for Six Sigma in Service and Manufacturing Companies: An Insight in Literature. *International Journal of Six Sigma*. Vol: 9(4), pp: 543-561

Paynabar, K., Jin, J., Pacella, M (2013) Monitoring and Diagnosis of Multichannel Nonlinear Profile Variations using Uncorrelated Multilinear Principal Component Analysis. *IIE Transactions.* Vol 45, pp 1235-1247

Paynabar, K., Zou, C. & Qiu, P. (2016) A Change –point Approach for Phase I Analysis in Multivariate Profile Monitoring and Diagnosis. *Technometrics.* Vol 58(2), pp: 191-204

Pignatiello, J.J. & Runger, G.C. (1990) Comparisons of Multivariate CUSUM Charts. *Journal of Quality Technology.* Vol 22, pp: 173-186

Rajaraman, A. & Ullman, J.D. (2012) *Mining of Massive Datasets*. Cambridge, UK. Cambridge Printing House.

Rato, T., Reis, M., Schmitt, E., Hubert, M. & De Ketelaere, B. (2016) A Systematic Comparison of PCA-based Statistical Process Monitoring Methods for High-dimensional, Time-dependent Processes. *AlChE Journal.* Vol 62(5), pp: 1478-1493

Raval, S.J., Kant, R. & Shankar, R. (2018) Revealing Research Trends and Themes in Lean Six Sigma: From 2000 to 2016. *International Journal of Six Sigma.* Vol: 9(3), pp: 399-443

Runger, G.C. & Testik, M.C. (2004) Multivariate Extensions to Cumulative Sum Control Charts. *Quality and Reliability Engineering International.* Vol 20(6), pp: 587–606

Saleh, N., Zwetsloot, I.M., Mahmoud, M.A. & Woodall, W.H. (2016) CUSUM Charts with Controlled Conditional Performance Under Estimated Parameters. *Quality Engineering.* Vol 28(4), pp: 402-415

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R (user),* 1st ed. New York, NY: Springer

Satterthwaite, F.R. (1954) A Simple, Effective Process Control Method. *Report 54-1*. Rath&Strong Inc, Boston, MA

Schwab, K. (2016) The Fourth Industrial Revolution: What it Means, How to Respond. Accessed on April 3, 2017. Can be found: https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Shahriari, H., Ahmadi, O. & Samimi, Y. (2016) Estimation of Complicated Profiles in Phase I, Clustering and S-estimation Approaches. *Quality and Reliability Engineering International.* Vol 32, pp: 2455-2469

Shamsi, M.A., Alam, A. (2018) Exploring Lean Six Sigma Implementation Barriers in Information Technology Industry. *International Journal of Six Sigma.* Vol: 9(4), pp: 523-542

Shannon, C. E. (1948) A Mathematical Theory of Communication. *Mobile Computing and Communications Review*. Vol 5(1), pp: 3–55

Sharma, P., Malik, S.C., Gupta, A. & Jha, P.C. (2018) A DMAIC Six Sigma Approach to Quality Improvement in the Anodising Stage of the Amplifier Production Process. *International Journal of Quality and Reliability Management*. Vol: 35(9), pp: 1868-1880

Shang, Y., Man, J., He, Z. & Ren, H. (2016) Change-point Detection in Phase I for Profiles with Binary Data and Random Predictors. *Quality and Reliability Engineering International*. Vol 32(7), pp: 2549-2558

Shewhart, W.A. (1930) Economic Quality Control of Manufactured Product. *Bell Systems Technical Journal.* Vol 9(2), pp: 364-389

Shi, Z., Apley, D.W. & Runger, G.C. (2016) Discovering the Nature of Variation in Nonlinear Profile Data. *Technometrics*. Vol 58(3), pp: 371-382

Stojanovic, N., Dinic, M. & Stojanovic, L. (2015) Big Data Process Analytics for Continuous Process Improvement in Manufacturing. In: *Proceedings 2015 IEEE International Conference on Big Data IEEE Big Data 2015.* Santa Clara, USA, October 29- November 1, 2015, pp: 1398-1407

Stojanovic, L. & Stojanovic, N. (2017) PREMIuM: Big Data Platform for Enabling Self-healing Manufacturing. In: *2017 International Conference on Engineering, Technology and Innovation: Engineering, Technology and Innovation Management Beyond 2020: New Challenges, New Approaches, ICE/ITMC 2017 – Proceedings.* Madeira, Portugal, June 27-June 29 2017, pp: 1501-1508

Suman, S. & Das, A. (2019) Stratified Statistical Process Monitoring Strategy for Multi-product Manufacturing Facility with Early Detection Approach. *Computers & Industrial Engineering.* Vol 130, pp: 551-564

Tong, C. & Yan, X. (2015) A Novel Decentralized Process Monitoring Scheme Using Modified Multiblock PCA Algorithm. *IIE Transactions on Automation Science and Engineering.* Vol 14(2), pp: 1129-1138

Tong, C., Yan, T., Yu, H. & Peng, X. (2019) Distributed Partial Least Squares Based Residual Generation for Statistical Process Monitoring. *Journal of Process Control.* Vol 75, pp: 77-85

Vaughn, A. (2013). http://statisticstoproveanything.blogspot.com/2013/11/spider-web-plots-in-r.htm

Wang, G. & Yin, S. (2015) Quality-related Fault Detection Approach Based on Orthogonal Signal Correction and Modified PLS. *IEEE Transactions on Industrial Informatics.* Vol 11(2), pp: 398-405

Wang, M., Yan, G. & Fei, Z. (2015) Kernel PLS Based Prediction Model Construction and Simulation on Theoretical Cases. *Neurocomputing.* Vol 165, pp: 389-394

Wang, Y., Mei, Y, Paynabar, K. (2018) Thresholded Multivariate Principal Component Analysis for Phase I Multichannel Profile Monitoring. *Technometrics.* Vol 60(3), pp 360-372

Weese, M. Martinez, W., Megahed, F. & Jones-Farmer, L.A. (2016) Statistical Learning Methods Applied to Process Monitoring: An Overview and Perspective. *Journal of Quality Technology*. Vol 48(1), pp: 4-27

Woodall, W. H. (2007) Current Research on Profile Monitoring. *Production*. Vol 17(3), pp: 420-425

Wu, J., Liu, Y. & Zhou, S. (2016) Bayesian Hierarchical Linear Modeling of Profile Data with Applications to Quality Control of Nanomanufacturing. *IIIE Transactions on Automation Science and Engineering.* Vol 13(3), pp: 1355-1366

Xie, H.B, Guo, J.Y, & Zheng, Y.P. (2010). Using the Modified Sample Entropy to Detect Determinism. *Physics Letters A*. Vol 374(38), pp: 3926-3931

Yan, J., Chen, C.Y., Yao, Y. & Huang, C.C. (2016) Robust Multivariate Statistical Process Monitoring via Stable Principal Component Pursuit. *Industrial and Engineering Chemistry Research.* Vol 55, pp: 4011-4021

Yang, W., Zou, C. & Wang, Z. (2017) Nonparametric Profile Monitoring Using Dynamic Probability Control Limits. *Quality and Reliability Engineering International.* Vol 33, pp: 1131-1142

Yin, S., Ding, S.X., Xie, X. & Luo, H. (2014) A Review on Basic Data-driven Approaches for Industrial Process Monitoring. *IEEE Transactions on Industrial Electronics.* Vol 61(11), pp 6418-6128

Zang, Y., Wang, K. & Jin, R. (2016) Unaligned Profile Monitoring Using Penalized Methods. *Quality and Reliability Engineering International.* Vol 32, pp: 2961-2776

Zeng, L. Neogi, S. & Zhou, Q. (2014) Robust Phase I Monitoring of Profile Data with Application in Low-E Glass Manufacturing Processes. *Journal of Manufacturing Systems*. Vol 33(4), pp: 508-521

Zhang, Z. 2015, A Study of Sample Entropy Towards Process Capability. Master dissertation, Kansas State University

Zhang, X. & Woodall, W.H. (2016) Dynamic Probability Control Limits for Lower and Two-sided Risk-adjusted Bernoulli CUSUM Charts. *Quality and Reliability Engineering International*. Vol 33(3), pp 607-616

Zhang, Y., Shang, Y., He, Z. & Wang, Q. (2017) CUSUM Schemes for Monitoring Prespecified Changes in Linear Profiles. *Quality and Reliability Engineering International.* Vol 33, pp: 579-594

Zhang, C., Gao, X., Xu, T., Li, Y. & Pang, Y. (2018) Fault Detection and Diagnosis Strategy Based on a Weighted and Combined Index in the Residual Subspace Associated with PCA. *Journal of Chemometrics.* Vol 32(11), article no e2981

Zheng, Y., Wang, Y., Wong, D.S.H. & Wang, Y. (2015) A Time Series Model Coefficients Monitoring Approach for Controlled Processes. *Chemical Engineering Research and Design.* Vol 100, pp 228-236

Zhou, J.L., Zhang, S.L., Zhang, H. & Wang, J. (2018) A Quality-related Statistical Process Monitoring Method Based on Global Plus Local Projection to Latent Structures. *Industrial & Engineering Chemistry Research.* Vol 57, pp 5323-5337

Zhu, J., Ge, Z., Song, Z. & Gao, F. (2018) Review and Big Data Perspectives on Robust Data Mining Approaches for Industrial Process Modelling with Outliers and Missing Data. *Annual Reviews in Control*. Vol 46, pp 107-133

Zou, C., Tsung, F. & Wang, Z. (2008) Monitoring Profiles Based on Nonparametric Regression Methods. *Technometrics*. Vol 50(4), pp: 512-226

Zou, C., Ning, X., Tsung, F. (2012) LASSO-based Multivariate Linear Profile Monitoring. *Annals of Operations Research.* Vol 192(1), pp. 3-19

Zwetsloot, I.M., Kuiper, A., Akkerhuis, T.S. & de Koning, H. (2018) Lean Six Sigma Meets Data Science: Integrating Two Approaches Based on Three Case Studies. *Quality Engineering*. Vol: 30(3), pp: 419-431.