INDUSTRIAL & SYSTEMS ENGINEERING
TEXAS A&M UNIVERSITY

PROJECT REPORT

FALL 2018

Phase-I analysis for Manufacturing Process Control

TEAM 06

NITESH KUMAR (UIN: 327003765)

VASU KUMAR (UIN: 427007395)

Table of Contents

1.0 EXECUTIVE SUMMARY

The project is based on a set of data collected from a manufacturing process, in which both in control and out-of-control data are present. The objective is to develop a method or procedure in order to identify data falling into respective categories, understand the distribution pattern and set up control limits so that change and anomaly detection process can be applied effectively on future process. The manufacturing process related dataset available consists of 552 records of 209 features/dimensions. Due to the presence of high dimensionality, the aggregated noise effect can overwhelm the signal effects thus making it harder to reject the null hypothesis. This phenomenon is known as curse of dimensionality.

The principal component analysis (PCA) as the data reduction tool to reduce the data points is used. By using the PCA 'vital few' dimensions were selected over the 'trivial many'. The selection of principal components is then done by analyzing the pareto plot, minimum description length (MDL) and scree plot in conjunction. It was found that 9 PCAs using correlation matrix were enough in order to explain 60% of the variability in the given data set.

Multivariate univariate charts for the selected principal components were used to find out the out of control data points and eliminate them. These charts can be applied on the PCAs since they are un-correlated. Multiple iterations in the process of establishing control limits for the process too place. The established control limits can be used to monitor the future data points of the manufacturing process.

Another way to approach the identification of out of control points in order to set control limits, T2 control chart was applied on principal components. This was done to explain and eliminate the spike changes in the dataset. Now in order to eliminate the sustained mean shift points, m-EWMA control charts were applied. After this, the iterated points are then again verified with T2 control chart in order to identify any spikes in the data points after removal of sustained mean shift data points. After this, the data points are passed through the T2 control charts for removal of any large spikes observed after mean shift detection done by m-EWMA.

The exploration of the real high dimensional manufacturing data set and development of the quality control parameters was quite insightful. We were able to understand the process of establishing and implementing PCAs in the industry. In order to develop phase 1 statistics, we used multivariate univariate control charts, T2 control chart and m-EWMA chart in order to identify the outliers and develop statistical control limits for the future data points. Collectively, this enhanced our understanding of multivariate process control methods and phase-1 analysis. The project was a good exercise in developing the phase-1 analysis for a high dimension data set. This project we were able to learn and practice MATLAB, Minitab and R software.

2.0 INTRODUCTION

In the advanced manufacturing world driven by industry 4.0 and digitization initiatives, the statistical quality control data collection and analysis is not a challenge anymore. There exists a dependency of multiple features in contribution towards its final quality. This has led the need for monitoring all features/ parameters to achieve the desired quality. The effective detection and monitoring system when applied on the process can supplement in early detection of change and a better controlled process. Quality control of multivariate data sets or features starts with the development of Phase-I analysis for the setting up of control limits. Most widely used quality control-monitoring charts for multivariate data are T^2, m-CUSUM and m-EWMA control charts. If the analysis done by observing the control charts shows no outliers, the data points can be selected with no further corrections/ iterations to the process control parameters. In addition, the control limits set can be used to predict the future performance of the manufacturing process.

This report discusses the Principal Component Analysis (PCA) as the dimension reduction tool. The use of this technique ensures the conversion of any correlated parameters to uncorrelated variables/ principal components by spectral decomposition. After selection of principal components, they are applied to determine the control limits using the Phase-1 analysis. The subsequent sections in this report will discuss the interpretations of data, approaches used, justification, results and conclusion. The primary advantage of a multivariate data analysis over the univariate one is that it eliminates the need to prepare the univariate charts for each monitoring characteristic.

2.1 Dataset Interpretation

The manufacturing process dataset given consists of a multivariate continuous with 209 features (columns) and 552 observations (rows). Each observation has a sample size of 1 i.e m=552, n=1, p=209. The data is perceived to follow a normal distribution by plotting the mean profile. Figure 1 shows that the distribution resembles normal distribution, hence it can be considered as continuous multivariate data for the analysis.

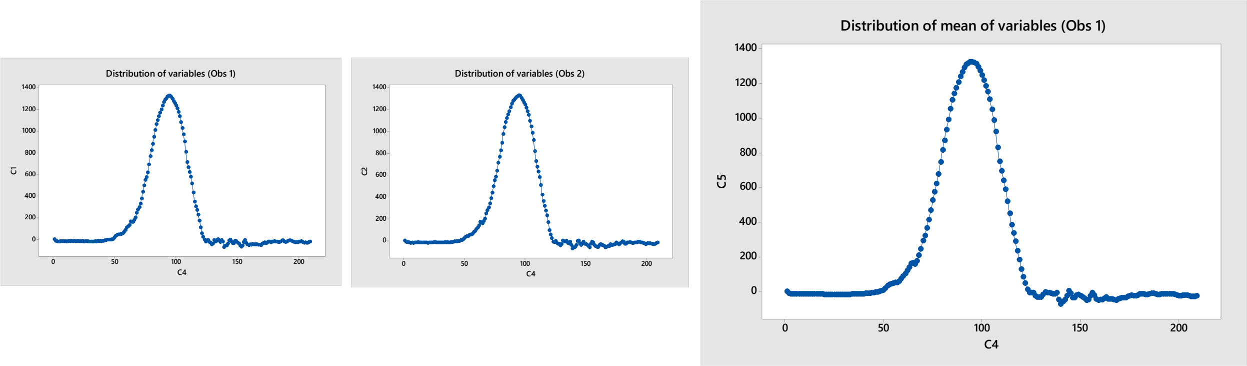


Figure 1. Data distribution and mean distribution

3.0 APPROACH

This section describes the methodology used in analysis of the manufacturing process data set in order to do the Phase 1 analysis and determine its control limits. Since there are 209 features/ variables, it is impossible and taxing to prepare 209 control charts in order to determine which variable contributes to the process change. Multivariate statistical analysis is used over univariate detection. Also, the α and β errors will now be needed to be adjusted based on the inflation caused during application of multiple univariate charts.

3.1 Principal Component Analysis

The Principal Component Analysis (PCA) helps to determine the variables/ features which contribute the maximum towards explaining the distribution variance and determining the observations lying outside the control limits. The information on the units of variables is unknown and limited, therefore PCAs were calculated based on cumulative variance matrix and select principal components. Application of PCA to the original data set, the total number of principal components are the same number in the original vector. In order to reduce it further, PCs with largest eigen values can be retained.

For our problem we needed to estimate the covariance matrix (S). This was done using the equations:

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Where $\bar{x}$ is the mean of individual variable, xi is the observation value and n is the data sample size.

3.2 Minimum Description Length (MDL) Criterion

The MDL plot shown in Figure 2 shows 30 principal component features to be selected in order to explain the data set as the vital few.

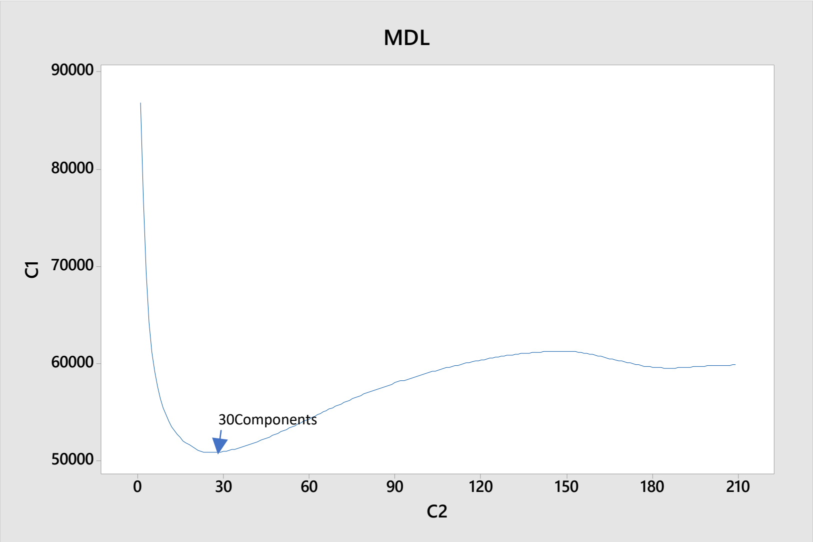


Figure 2. MDL Plot

The formula used for identifying the MDL is:

$$MDL\ (l) = n(p-l)\log\left(\frac{a_l}{g_{\ l}}\right) + l(2p-1)\log(n)\ /2$$

n=data sample size
p=dimension of the covariance/correlation matrix.
a_l=arithmetic mean of p-1 smallest eigen values
$g_{l=}$ geometric mean of the p-l smallest eigenvalues

MDL(l) is evaluated for l = 0, 1, ..., p − 1 and the number of PC to retain is chosen as the l that minimizes MDL(l). Sometimes MDL(l) can retain too many eigenvalues. So it is a better practice to use MDL together with the scree plot and pareto plot.

3.3 Pareto Plot

Figure 3 shows the pareto plot for the principal components. It shows that 33 PCAs represent 80% of the variance of data. Using 33 PCAs for multivariate univariate control charts is a tedious job, so an alternate way to find minimal PCAs is via scree plot explained in the next section.

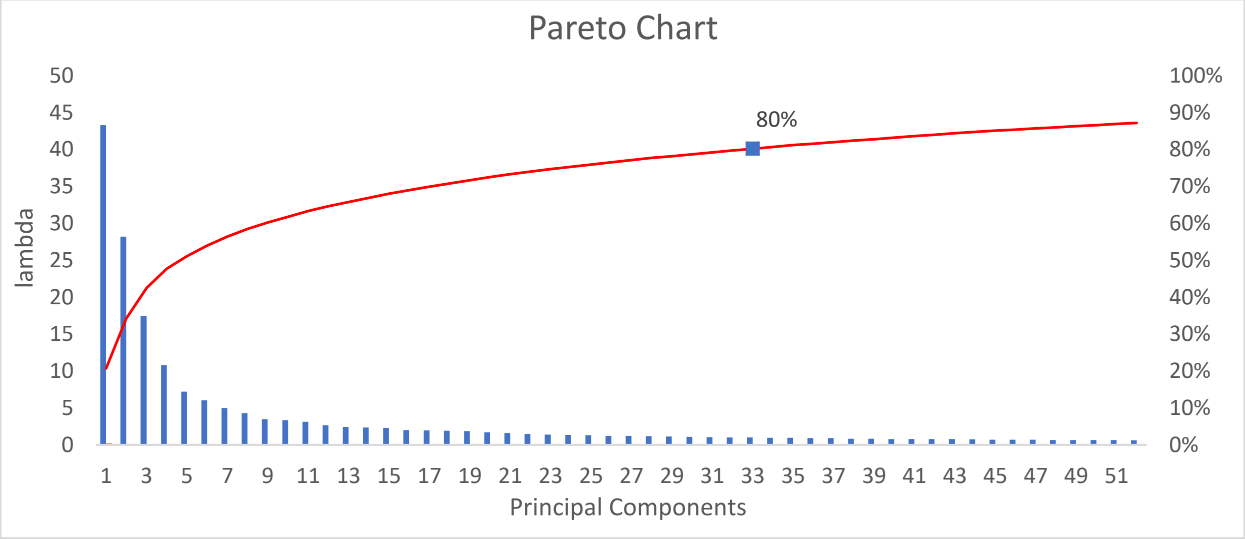


Figure 3. Pareto Plot

3.4 Scree Plot

Figure 4 shows the scree plot for the PCAs. It shows that the elbow is formed at 09 Principal components.

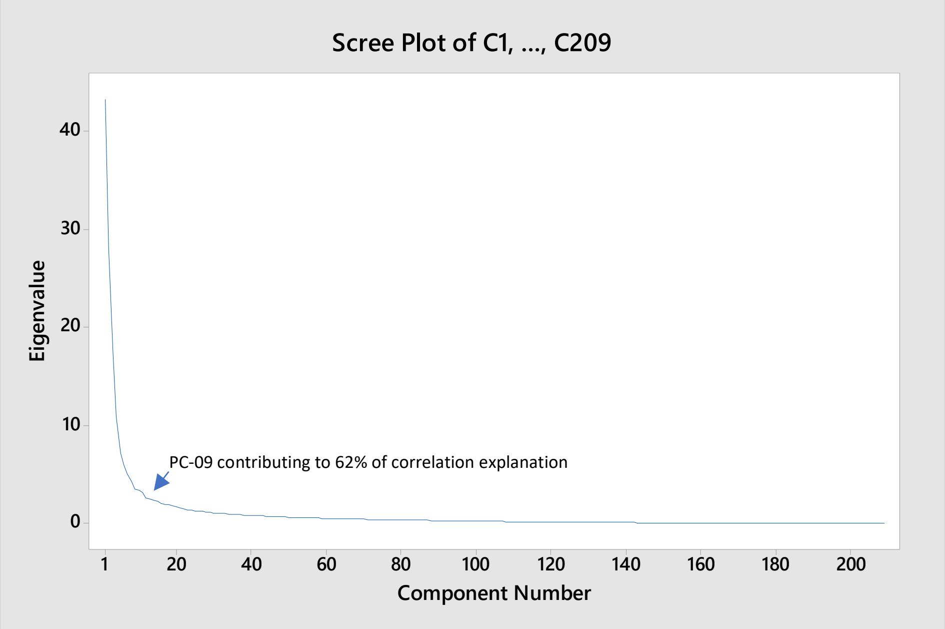


Figure 4. Scree Plot

3.5 Determination of Principal Components

The ith principal component of data can be calculated using:

$$y_i = (\mathbf{e}_i^{\rho})^T(\mathbf{V}^{-\frac{1}{2}})(\mathbf{x} - \boldsymbol{\mu}_x)$$

where $\mathbf{V} = \begin{pmatrix} \sigma_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{pp} \end{pmatrix}$ is the matrix having the diagonal elements from $\boldsymbol{\Sigma}$ and

$$\mathbf{V}^{-\frac{1}{2}} = \begin{pmatrix} \frac{1}{\sqrt{\sigma_{11}}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sqrt{\sigma_{pp}}} \end{pmatrix}.$$

Based on the analysis carried out for PCAs we are selecting 09 PCAs. So the i=1…9. Since we have 209 variables there will be 209 eigen vectors. After the identification of PCAs, they can be selected to go ahead with phase 1 analysis to identify any out of control data points and eliminate them. Since PCA was performed to bring out the variables that explains maximum variability, multiple univariate charts were used to do the phase-I analysis. PCAs obtained have no correlation among each other, therefore it makes the use of multiple univariate charts for PCAs in the manufacturing setup more acceptable.

3.6 Univariate charts for monitoring PCAs

Since the principal components are uncorrelated, it is easier to use multivariate univariate charts for Phase-1 analysis. 3 sigma limits are considered for each characteristic. These charts play a key role in identification of out of control points in the respective control charts for each PCA and determining the values to be removed for the subsequent iterations.

3.7 T2 Control Chart and m-EWMA Control Chart

The given manufacturing process data set may contain spike type or data or sustained mean shift in conjunction with it. In order to identify and remove the large spike type of changes T2 control chart is applied on it. Since, EWMA control charts are good in identifying sustained mean shifts, the data points are passed through it and the sustained mean shift points are removed. The data set is again passed through the T2 control chart in identify the out of control points and then finalize out control limits.

4.0 JUSTIFICATION

The given data is very high dimensional contributing to high noise components and less signal ratio. This high dimensional data induces high noise components which add up to a great magnitude and reduce our signal to noise ratio. This is called as curse of dimensionality. We used correlation for choosing our principal components since the data dimensions were not given.

The principle effect of sparsity, the ‘vital few’ matters instead of ‘trivial few’. Detection and monitoring are effective on lower data dimension. PCA effectively reduced the data dimensions. Since, the data units were not given, consideration of correlation matrix was done. The PCAs selected were then applied to multivariate univariate control charts for the phase 1 analysis. For identification of large spikes in the data, T2 control chart is used and to control the sustained mean shit m-EWMA chart is implemented.

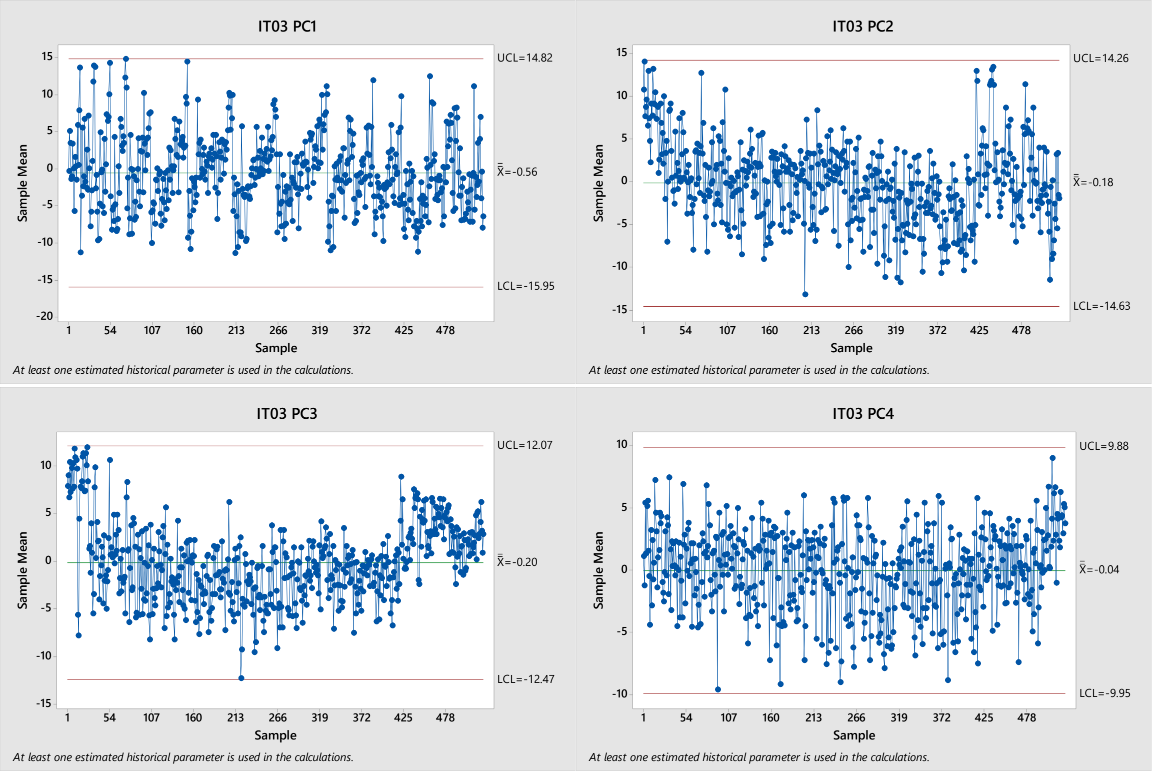
5.0 RESULT

After selecting the principal components, we observe the multivariate univariate charts using individual PCs. Following are the univariate control charts for the individual PCAs using correlation matrix.

Table 1. Principal components removed after iterations (for Multivariate Univariate Chart)

Phase-I Analysis	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
After 0 Iteration	11	3	1	0	0	0	0	1	3
After 1st Iteration	4	4	0	0	0	0	0	0	0
After 2nd Iteration	0	2	0	0	0	0	0	0	0
After 3rd Iteration	0	0	0	0	0	0	0	0	0

Finally, after 4 steps for identification of out of control points, the dataset is obtained within control limits with the recalculated UCL/LCL. The points which lie in the out of control samples are then removed. Table 1 shows the number of data points removed after every iteration. This is the end of Phase-1 analysis. This plots for the multivariate univariate charts after the final iterations are shown below.



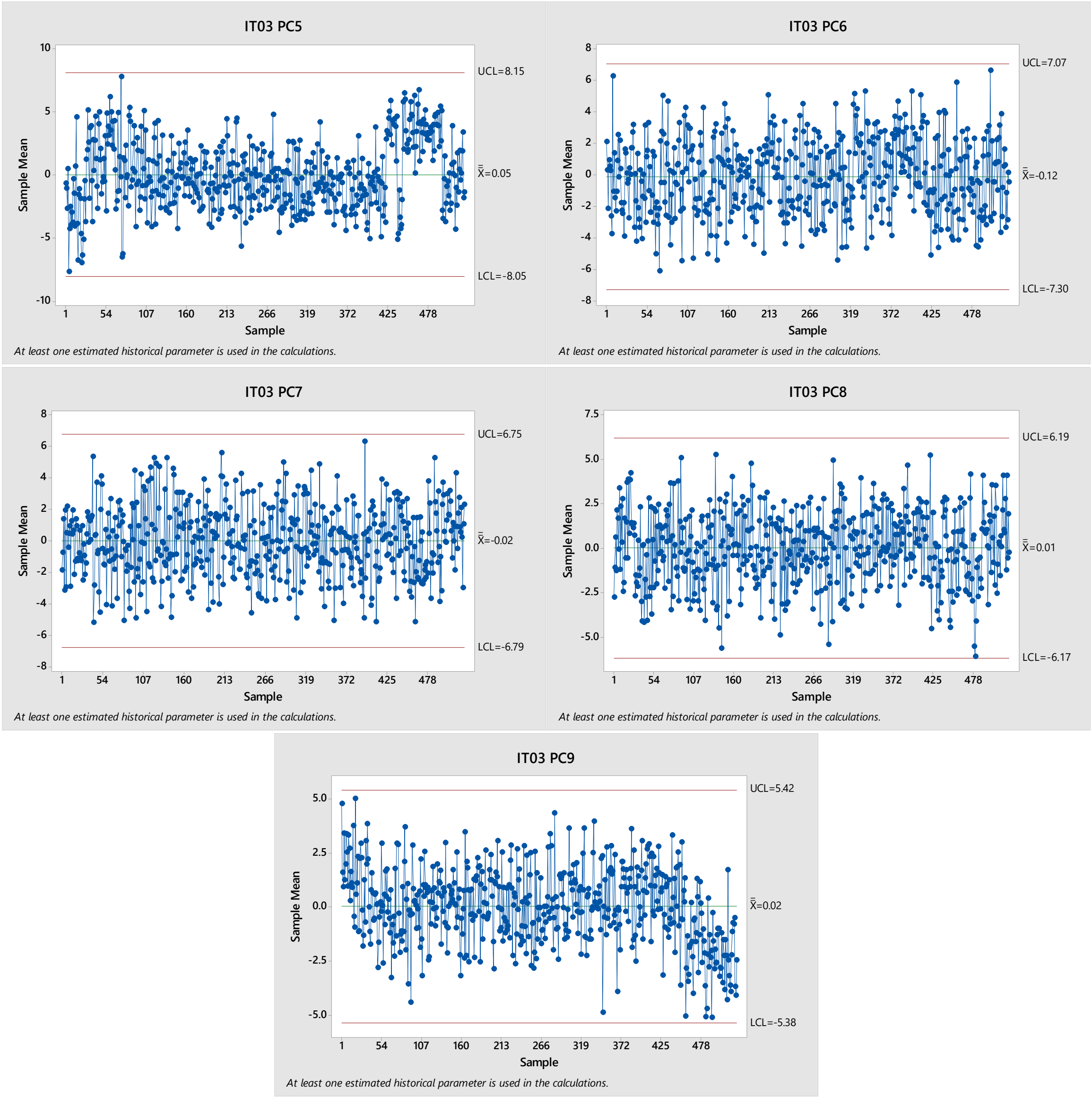


Figure 5. Final Multivariate Univariate Charts

Now, doing the second analysis for identification of large spikes in the data set, the Hoteling T2 charts were applied. Table 2 shoes the data points found as outliers after each iteration/step. Table 2 shows the data points identified and removed after the iterations.

Table 2. Data points removed after iterations (for T^2 Charts)

Phase-I Analysis	Total Data points identified as outliers
After 0 Iteration	59
After 1st Iteration	9
After 2nd Iteration	5
After 3rd Iteration	4
After 4th Iteration	3
After 5th Iteration	0

Below, the charts of each iterations are shown for T2 statistics. It is clearly observed that at step 5, there are no out of control points. This statistical quality control for phase 1 analysis gives us the outliers which are large spikes in nature. Now in order to refine the data for setting the control limits, m-EWMA chart is used. Multivariate EWMA chart are efficient in determining sustained mean shifts. The charts below shows the data points when applied to T2 in first step and the data set obtained after the removal of out of control points after the final iteration.

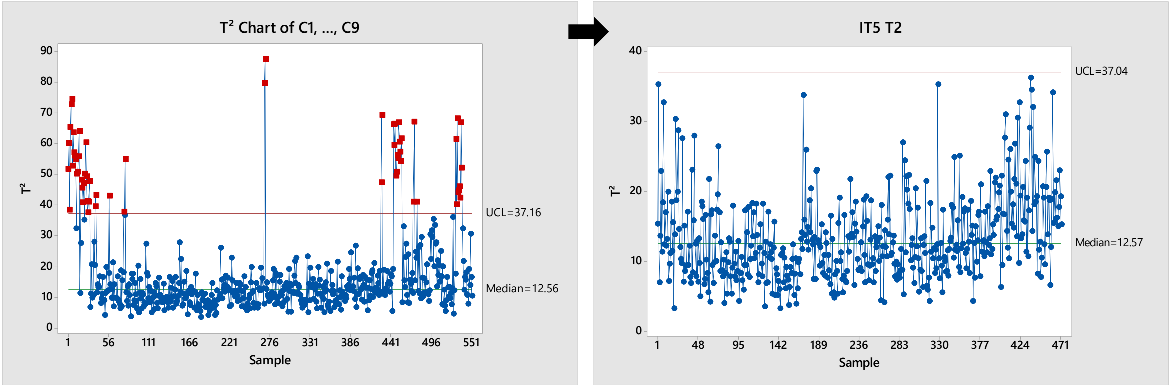


Figure 6. (a) Data obtained in iteration 0 (b) Control Chart obtained after removal of all ooc points

The data points are then passed through m-EWMA control chart. The figure 7 shows that there are no points identified in the m-EWMA control chart thereby concluding the final result.

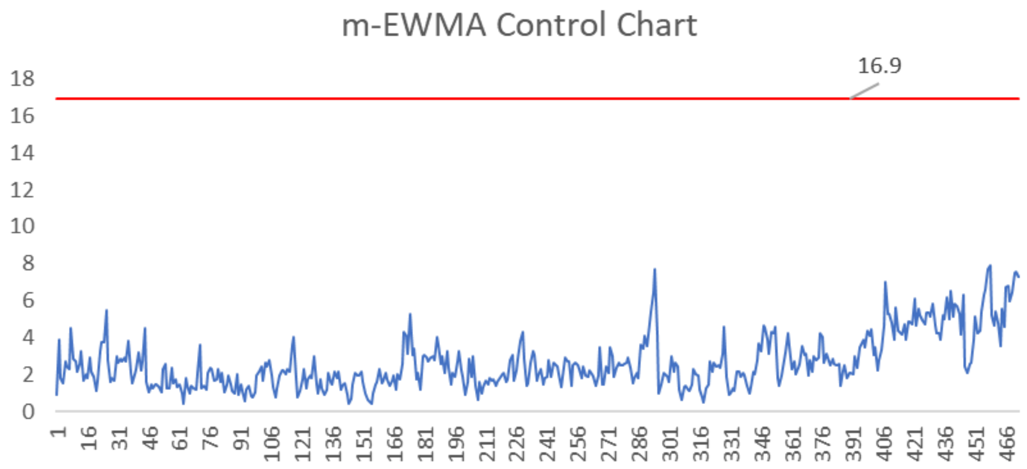


Figure 7. m-EWMA applied on

6.0 CONLUSION

The phase-1 analysis performed on the principal components involved multiple iterations for setting up of the control limits. Since we don't know whether the relative magnitude of deviations of each of the magnitude of deviations of each of the dimension are of relative importance or not, it is better to use correlation matrix for monitoring purpose. On application of PCA we identified 9 PCA contributing and explaining the data set by the analysis of MDL, Scree and Pareto analysis. The elbow on scree plot analysis was found at 09 sample hence, 9 PCAs were taken into consideration.

The out of control (OOC) data points were removed after 4 steps and the data cleaning is done which can be used for determining the control limits for the future observations for the values. We can conclude that with the result obtained with the phase-1 analysis, we can use the new data for carrying out the phase-2 analysis. The alpha value for the detection is corresponding to 3 sigma limits. The first univariate chart analysis on PCAs lead to the removal of 29 variables and the s T2 control chart eliminated 80 out of control data points leaving us with 472 data points. The final control limits are:

Table 3. Control limits from Multivariate Univariate in PCA, T2 and m-EWMA control charts

Multivariate Univariate Control Chart (PCAs)	Control Limits	
PC1	UCL : 14.82	LCL: -15.95
PC2	UCL : 14.26	LCL: -14.63
PC3	UCL : 12.07	LCL: -12.47
PC4	UCL : 9.88	LCL: -9.95
PC5	UCL : 8.15	LCL: -8.05
PC6	UCL : 7.07	LCL: -7.3
PC7	UCL : 6.75	LCL: -6.79
PC8	UCL : 6.19	LCL: -6.17
PC9	UCL : 5.42	LCL: -5.38
T2 Control Chart	UCL : 37.04	LCL: 0
m-EWMA	UCL : 16.9	LCL: 0

7.0 REFRERENCES

[1] Firat, Seniye & Aricigil Cilan, Cigdem. (2000). Multivariate Quality Control: A Historical Perspective.

[2] https://onlinecourses.science.psu.edu/stat505/

[3]https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/control-charts/how-to/multivariate-charts/tsquared-chart/before-you-start/example/

[4]https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/control-charts/how-to/multivariate-charts/tsquared-chart/methods-and-formulas/methods-and-formulas-for-tsquared-chart/

8.0 SOFTWARES USED

1. R
2. Minitab
3. MATLAB