Phase-I Analysis for Manufacturing Process Control

Submitted in partial fulfillment of
the requirements of the course

ISEN 614 Advanced Quality Control

by

Ishank Gupta	832000398
Ronak Radadiya	731008599
Arpan Modi	331007975

Project Guide: Dr. Yu Ding

Wm Michael Barnes '64 Department of Industrial & Systems Engineering

Texas A&M University, College Station

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

Manufacturing units use a quality measurement, quality improvement, and the adoption of manufacturing field standards and best practices to continuously improve product quality. As per the quality standards established by the Quality Control Services United States, quality reporting and customer relations are critical to shaping and accelerating the industry's attempt to continuously monitor and improve product quality. The manager of a manufacturing business unit has provided the analytics team with a data set that includes both, in-control, and out-of-control samples. As core members of the analytics team, our primary objective is to eliminate the out-of-control data for estimating the in-control distribution parameters, so that a monitoring scheme can be set up for future missions. Because of the presence of high dimensionality in the data, the noise of each individual component can add up to a significant amount in the signal-to-noise ratio. This aggregated noise can overwhelm the signal effect, making it difficult to reject the null hypothesis ($\mu_1 \neq \mu_o$). This is referred to as the "curse of dimensionality". The given data's dependency on the multiple variables of the manufacturing processes led to the need to closely monitor each parameter.

In this project, exploratory analysis is conducted on the original data to learn more about the trends of the manufacturing process. The mean value of each feature is calculated and plotted. Based on the resulting graph, it is discovered that the data followed the trend of a unimodal function as the values increased to a certain point and then continued to curve downward for all samples. Because no identifying label for each feature is given and the scale of data is almost similar, a covariance matrix is chosen rather than a correlation matrix to preserve the original relationship between the features of the original data. There are a lot of features in the original data, so analyzing all of them would be inefficient. As a result, Principal Component Analysis (PCA) is used as a dimension reduction tool to identify the most important features using Minimum Description Length (MDL), scree plot, and Pareto plot. Based on this analysis, we identified 4 Principal components to be used in our detection process.

Two approaches were used to identify the in-control points and then final comparison on the best approach was made in such a way that analysis on future data will be efficient. In the first approach, multivariate analysis was executed by performing multiple iterations using a combination of T^2 and m-CUSUM charts to detect and eliminate all out-of-control points. A combination of T^2 and m-CUSUM chart is a good method to identify all out-of-control points because T^2 is effective in removing large spikes whereas m-CUSUM is efficient in detecting sustained mean shifts. In second approach, we performed multiple univariate analysis. Initially, we performed x-bar charts for individual 4 Principal components. Further, multiple iterations of a combination of T^2 and m-CUSUM charts was used to ensure that all out-of-control points are detected and removed. Proper attention was given to correct the problem of inflation in Type-I and Type-II errors, for this purpose multiple decisions rule were created to signal how to identify an out-of-control data point. Based on these decision rules combined Type-I and Type-II error, UCL, ARL_0 and ARL_1 were calculated. The UCL calculated was used to setup the T^2 and m-CUSUM chart whereas other parameters were helpful in identifying which decision rule is relevant to the business needs. Here, key assumption was made that the priority for the business is to ensure that production and shipping of defective parts is reduced i.e., focus is to ensure that ARL_1 is as small as possible, as once defective parts are shipped to customers it can result in damaging the reputation that made lead to huge loss in revenue. Based on the final approach and rule 1 selected, we calculated the parameters (mean and covariance matrix) that will be used in the analysis of future data.

2. PRINCIPAL COMPONENT ANALYSIS

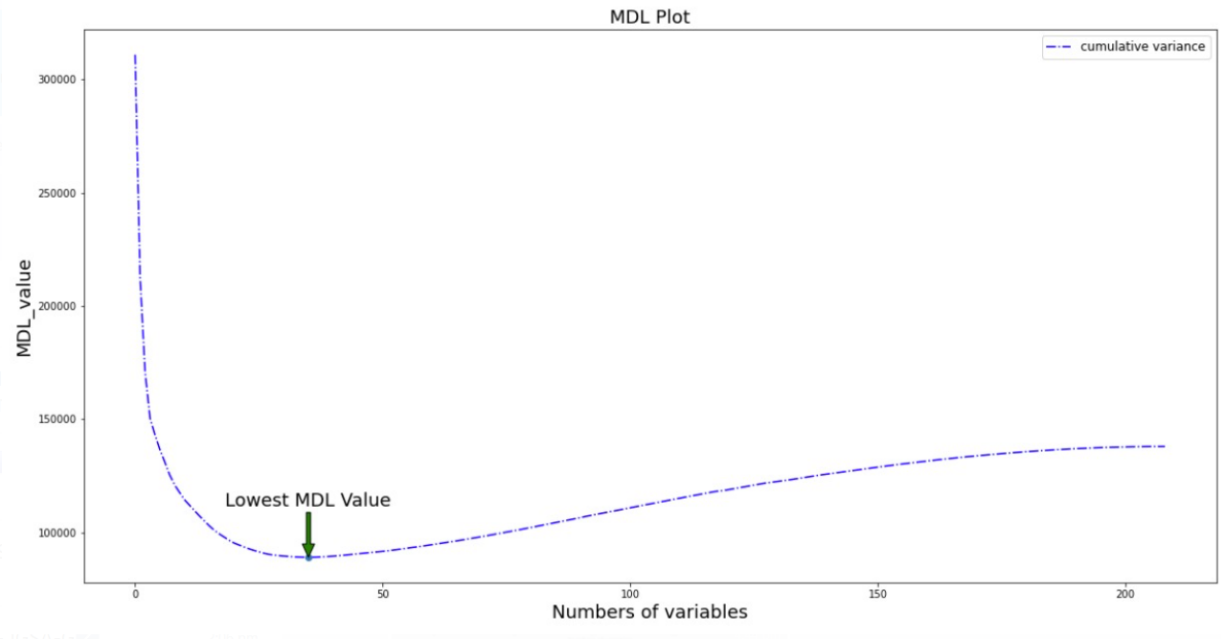
2.1 Implementation of PCA:

Initially, a single control chart was utilized to detect the control limit condition for many variables using multivariate statistical detection. However, when there are a large number of variables, as there are in our instance, it is quite difficult to keep track of the out-of-control process. Due to these factors, a need arose for a solution that minimizes the dimension to escape the dimensionality effect's curse. According to the impact sparsity principle, it is always the "vital few" who matter, not the "trivial many." Rather than tracking all 209 variables, we used Principal Component Analysis to discover the linear combination of factors with the largest variance. PCA is a statistical process that uses an orthogonal transformation to turn a set of observations of possibly correlated variables into a set of values that are linearly uncorrelated variables. The number of PCs created after applying PCA to the original data set is the same as the number of elements in the original vector. We can just keep the first few PCs, which correspond to the biggest eigenvalues, to reduce the dimensionality. The number of major components is set so that it accounts for over 80% of the total variations in all the parameters.

2.2 Determination of Principal Components:

The lowest description length, which reflects the full data with few variables, was computed as the first step in accomplishing dimension reduction. The MDL is 35, as shown in the graph below. However, 35 is still too many principal components, and the number of them can be lowered further. To further minimize the dimensions, it is preferable to utilize MDL in conjunction with a scree plot.

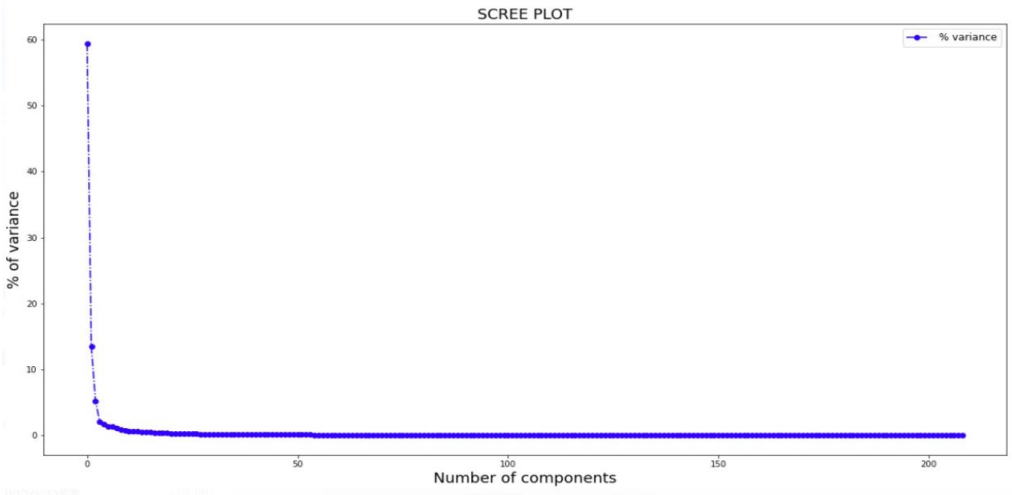
MDL Plot:



From the calculated MDL values, it can be seen that the minimum was achieved at $\iota = 35$

As the next step, a scree plot was plotted between the variable(i) number and the corresponding (λ_i) eigenvalues of these 35 variables, given that the eigenvalues are plotted in descending order. After observing the scree plot, the elbow(bend) was observed at 4. This indicates that the number of principal components can further be reduced to 4. After calculating the cumulative variances of these 4 principle components, it was found that it accounted for a total of 80.106 % of the variance exhibited by all the variables in the dataset provided.

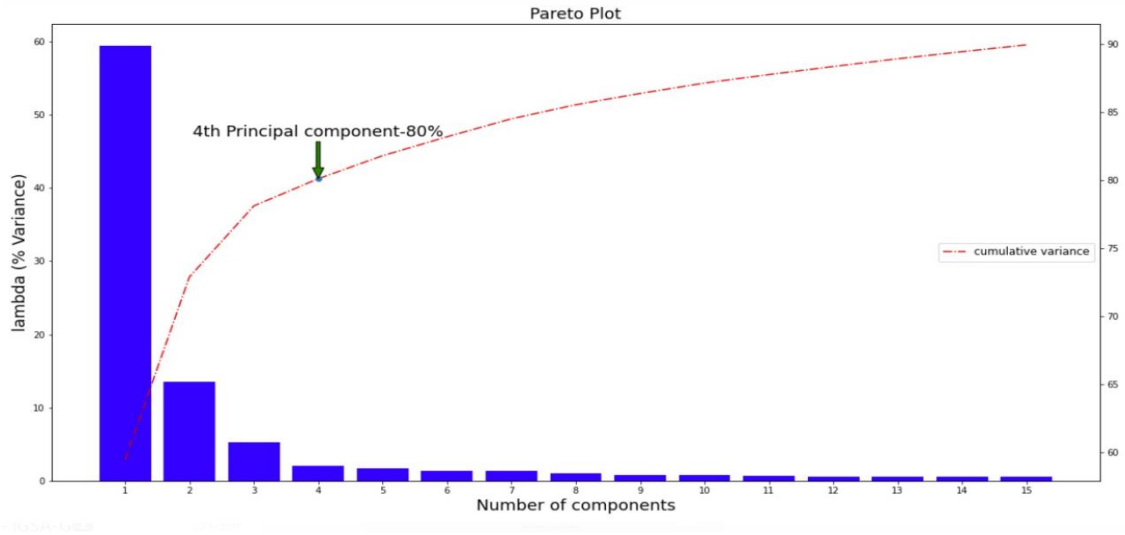
Scree Plot:



From the Pareto plot can see the accumulation of the effects in percentage. The table below shows the variation explained by all the principal components.

Principle Component	Proportion of Variance
PC1	59.413
PC2	13.467
PC3	5.225
PC4	2.000

Pareto Plot:



3. APPROACH 1

3.1 Hotelling T^2 chart:

The next step is to isolate the in-control data. After the selection of the principal components, the T2 chart was obtained by plotting the sample number on the x-axis and the corresponding T2 on the y-axis.
Phase I Analysis: It is used to identify the in-control data which are used to estimate the distribution parameters. In this case, Phase I analysis was conducted considering the sample size (n) equal to 1 as each variable has a unique observational value in each sample data set. At first, the chart is applied to the given data to verify the data points are in control and to identify all the out-of-control points. In the next step, we remove all the out-of-control data points and iterate until all the data points are in control.

The upper control limit (UCL) is determined as the value of $\chi_{1-\alpha}\ (p)$
Where, p = principal number of components = 4
α = 0.0027 (3-sigma control limits of Shewhart chart)

The formula for T2 statistic for the sample size of 1, is given as

$$T^2 = \left(x_J - \bar{x}\right)^T S^{-1}\left(x_j - \bar{x}\right)$$

Where Xj = matrix of the individual j data sample values
$\bar{x}$ = Average values of individual x_j matrix
S = Sample covariance matrix

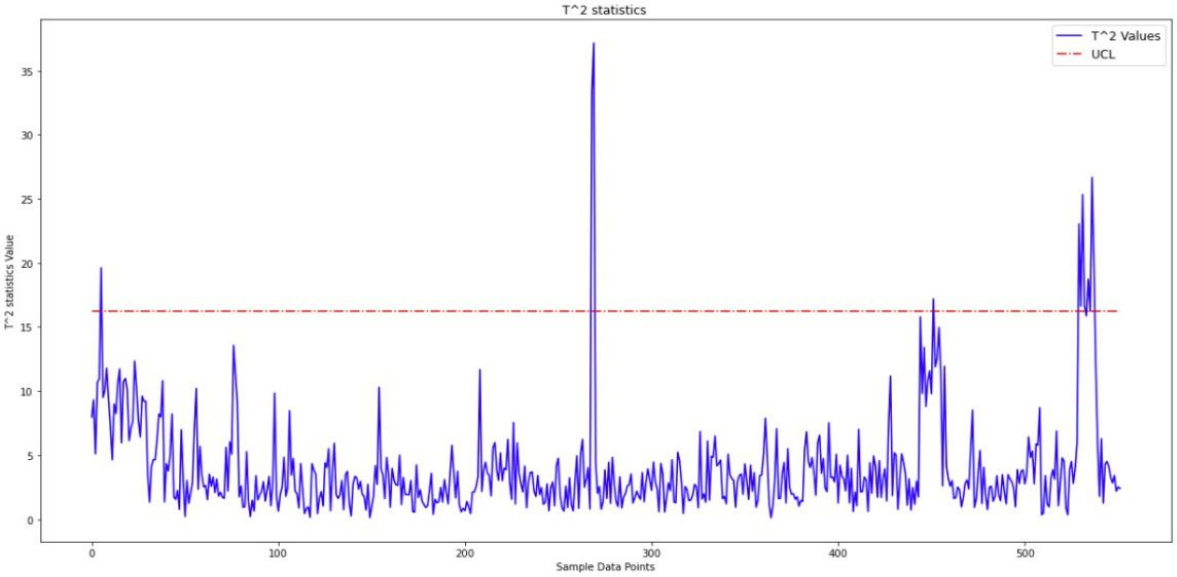
We chose α = 0.0027, we get,
UCL = $\chi^2$1-0.0027 (4) = 16.25
Based on the given data, the number of samples is 552 and the dimension p is 4. Hence, there will be 552 such statistics.

The T2 statistic was calculated for the reduced data vector y and compared it with the above UCL value. This statistic was then plotted in python. After plotting the available data values, it was found that 11 of these values were out-of-control (beyond UCL). The plot below shows all the 11 out-of-control data points. Now, the next step is to remove all the out-of-control data points as they are responsible for the process being detected as out of control. T2 statistic is calculated again after removing these 11 points. However, from the plot it is observed that there were 7 data points which were still beyond the UCL. To eliminate all the out-of-control points, the Phase I analysis was performed 4 times in total until all the data points obtained are within the upper control limit. The results of each iteration are as follows:

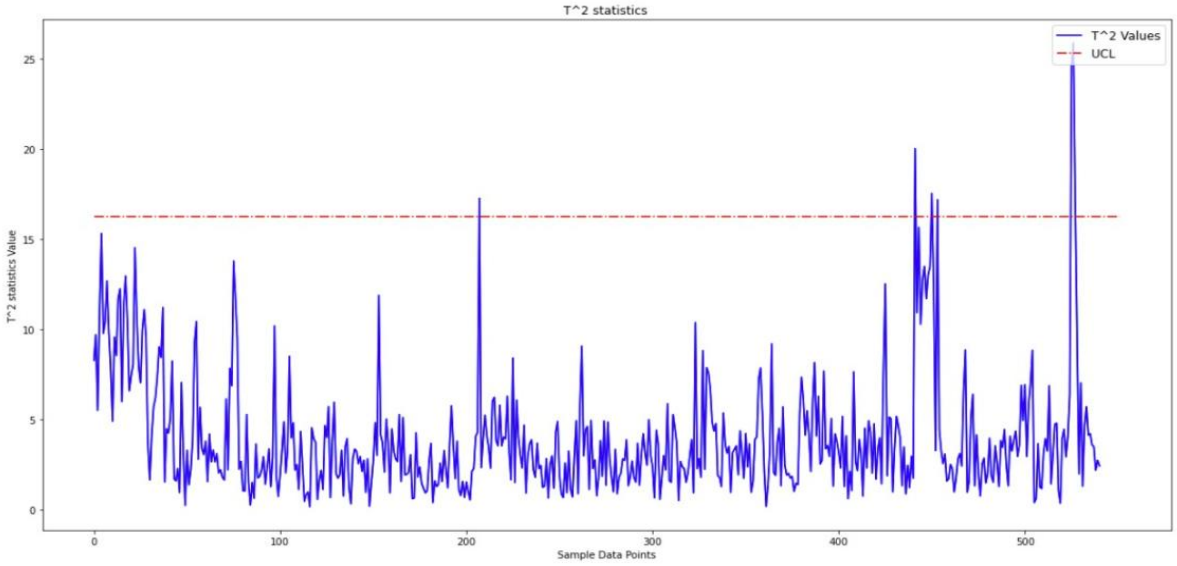
Iteration of Phase I Analysis	Number of Out-of-Control samples observed
1st Iteration	11
2nd Iteration	7
3rd Iteration	2
4th Iteration	0

Total Number of in-control samples after 4 iterations of Phase I Analysis using T2 control chart = 552-(11+7+2) = 532

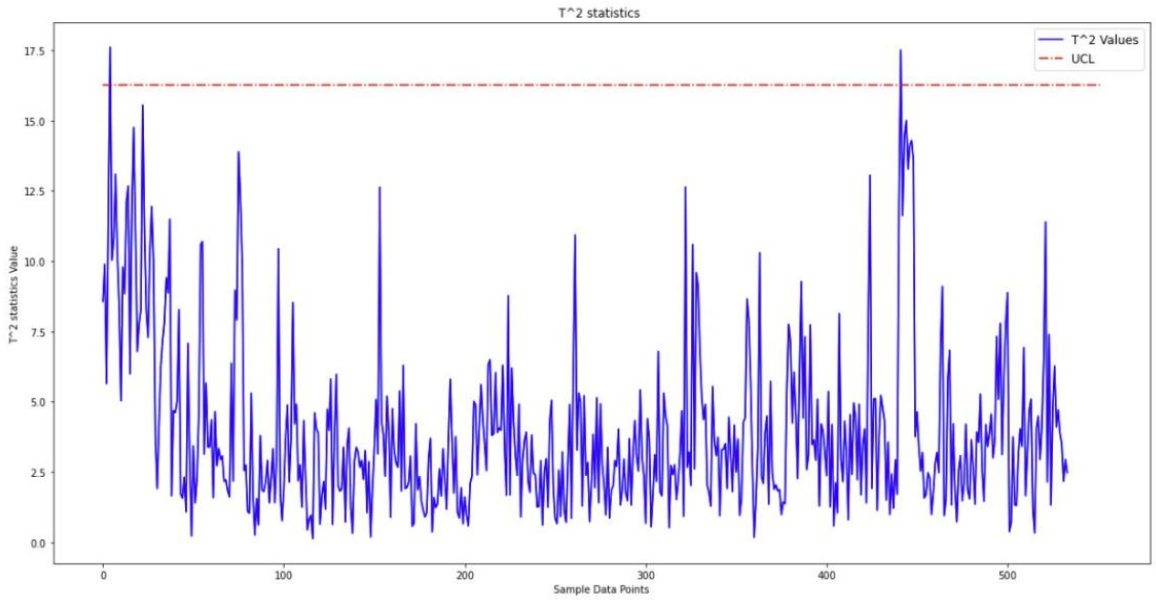
The below graphs show T^2 statistics:



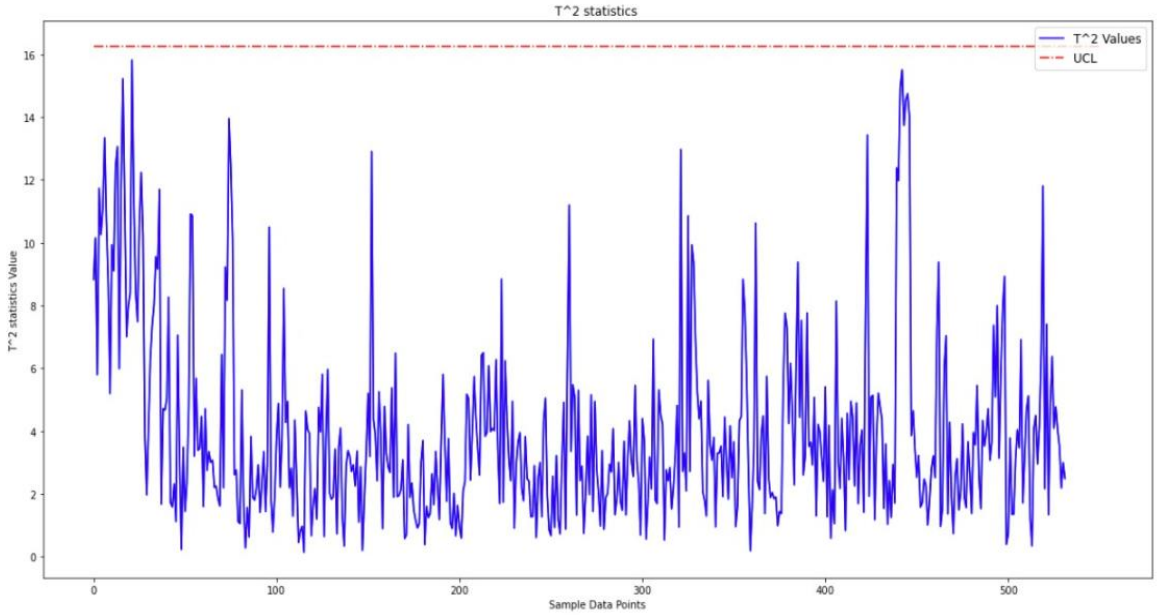
After the first iteration we removed the 11 out-of-control points



After the second iteration we removed the 7 out-of-control points



After the third iteration we removed the 3 out-of-control points



We can see there are no out of control points in the third iteration from the above plot.

All out-of-control samples are eliminated using the above iterative process. T2 has the advantage of being sensitive to spike-type change fluctuations in existing correlated data, but it cannot identify the disparity created by the process's sustained mean shift. This necessitated the use of additional multi-variate control charts, such as them-CUSUM or m-EWMA, which are sensitive to mean shift in a process. Using these charts, limitations faced in T2 can be eliminated. For further analysis, m-CUSUM chart is used to detect the mean shift.

3.2 m-CUSUM chart:

A cumulative sum (CUSUM) chart is a control chart that is used to enhance the sensitivity of detection in the presence of a small sustained mean shift. It calculates the total number of deviations from a threshold value. The cumulative sum of divergences from the target for individual measures or subgroup means is plotted on the CUSUM chart. All the points having a mean shift of $\lambda = 3$ is considered out-of-control. The value of k $= 1.5$ based on the chosen value of λ. m-CUSUM was done iteratively on the new data obtained after applying a hoteling T^2 statistics i.e., on 532 points. For this analysis with p = 4, the interpolation method is used to calculate the upper control limit and ARL_0.

P	UCL	ARL_0
2	5.5	406.42
3	6.25	388.87
10	11	488.93

Based on this table, interpolation technique is used to calculate UCL. The equation for the same is as follows:

UCL = $0.684211 * 4 + 4.162281$,
substituting p = 4, we get UCL as 6.899.

Similarly, ARL0 is calculated using equation : ARL0 = $11.78026 * 4 + 369.172$
The formula for m-CUSUM statistic (MCi)is:

$$MC_i = max\left\{0, \left(C_i^T \epsilon^{-1} C^T\right)^{1/2} - kn_i\right\}$$

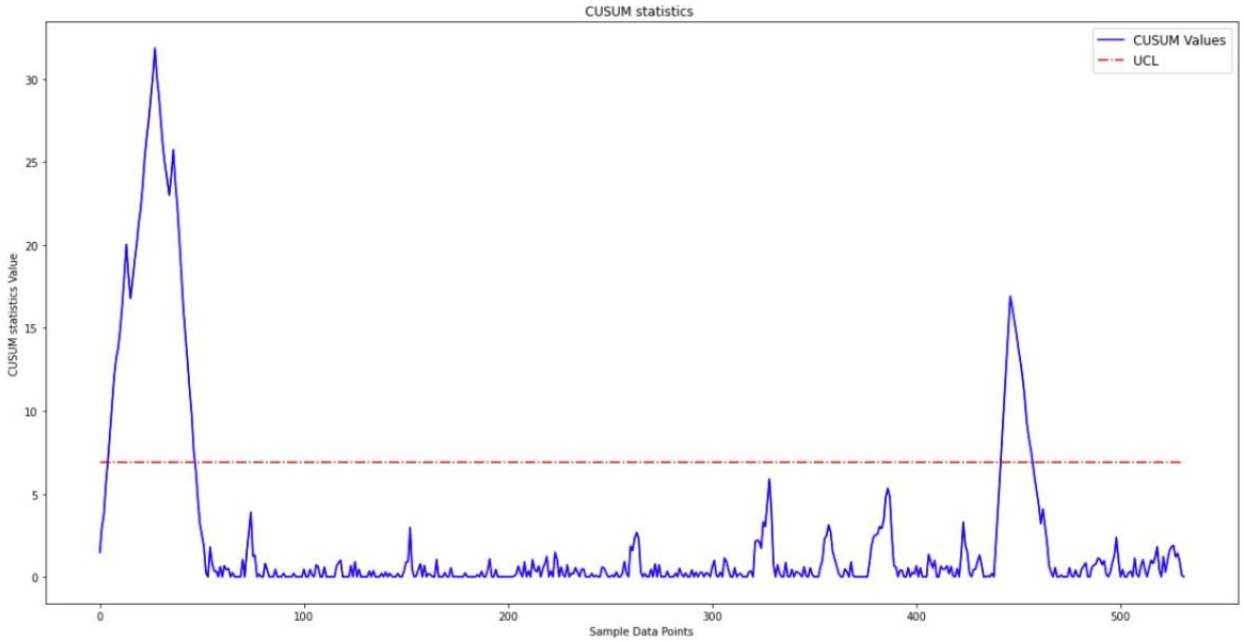
Where,

Ci = Cumulative sum of previous 'ni' number Xi's
$\in$ = Population covariance matrix

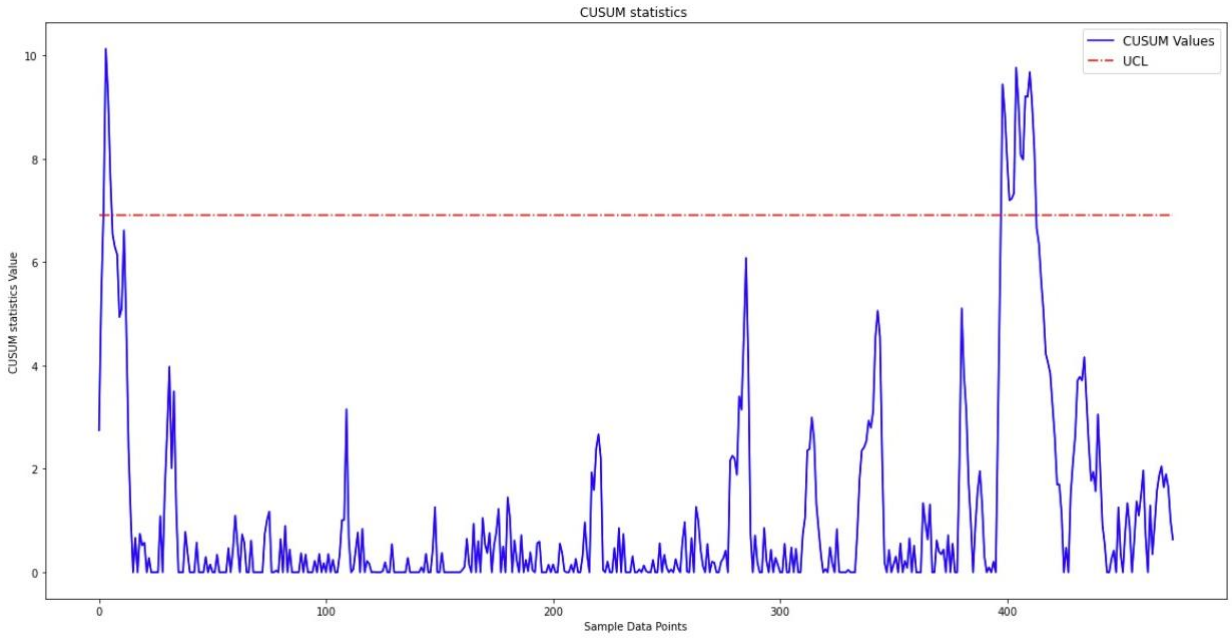
After performing 1st iteration on 532 data points using the m-CUSUM method, it was discovered that there are 58 out-of-control points. After removing these 58 points, 19 more samples are observed to be out of control. Now, 2nd iteration was performed to remove these 19 out-of-control data points. After 2 iterations all points were in control. The result of the analysis is summarized in the table below.

Iteration of Phase I Analysis of m-CUSUM chart	Number of Out-of-Control samples observed
1st Iteration	58
2nd Iteration	19
3rd Iteration	0

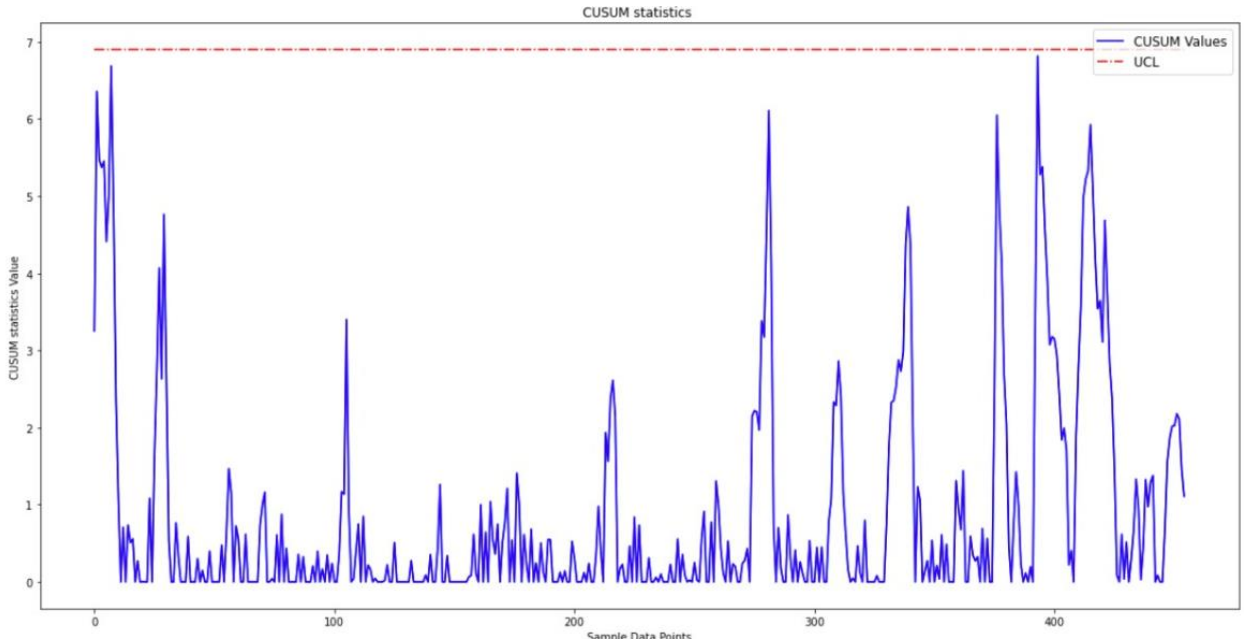
m-CUSUM plots for each iteration are as follows:



From the above plot, we can see that 58 samples are out of control. Further iterations is required to remove all the out-of-control data samples.



Further iteration is required to remove remaining out-of-control data samples.



From the final iteration, it is observed that all the out-of-control points are eliminated.

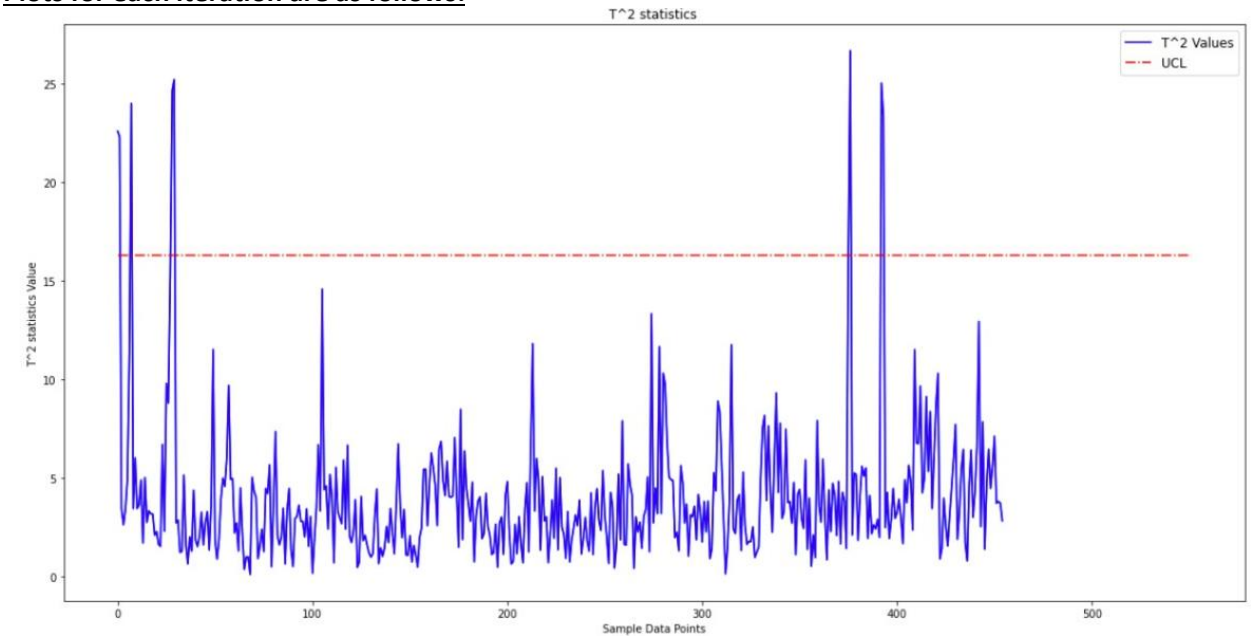
3.3 Final T^2 Statistics:

The T2 statistics of all 455 samples were plotted again after performing the Phase l Analysis of the m-CUSUM control chart. According to the graph below, 8 points are still out of control. As a result, Phase I Analysis must be repeated to guarantee that any out-of-control points have now been eliminated. In the second iteration, 3 points are still out-of-control. Finally, we have to do analysis 3 more times to eliminate all the out-of-control points. The results obtained in all the iterations are as follows:

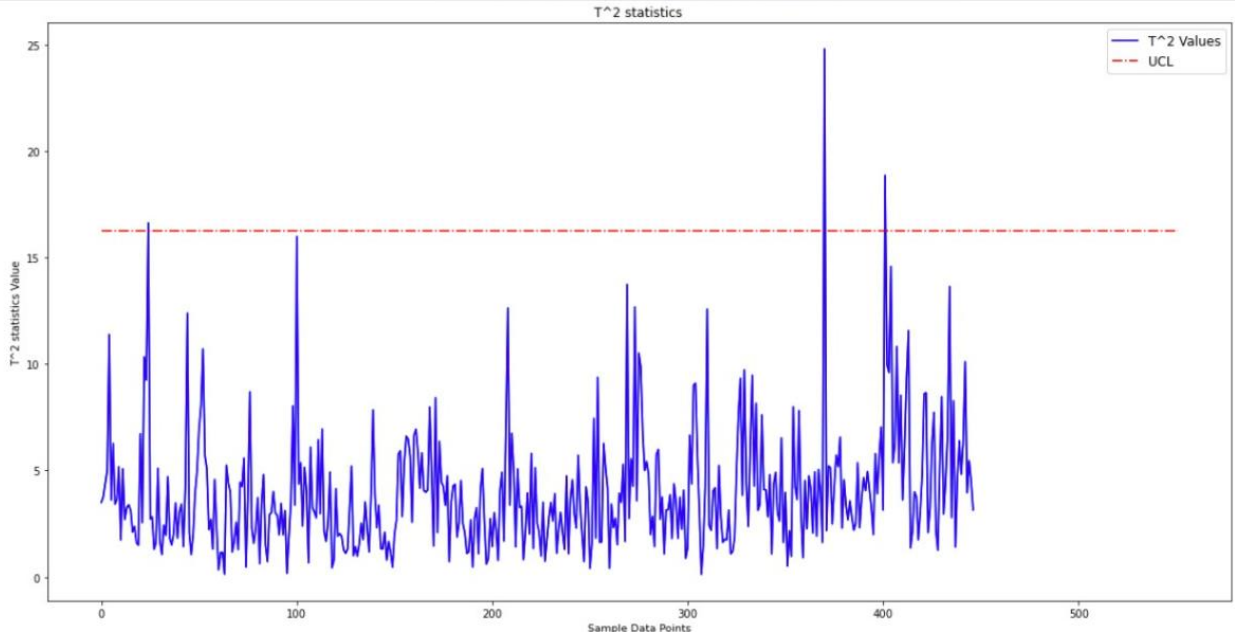
Iteration of Phase I Analysis	Number of Out-of-Control samples observed
1st Iteration	8
2nd Iteration	3
3rd Iteration	1
4th Iteration	1
5th Iteration	0

After finishing phase I analysis with the m-CUSUM control chart, total in-control samples are 455-(8+3+1+1) = 442.

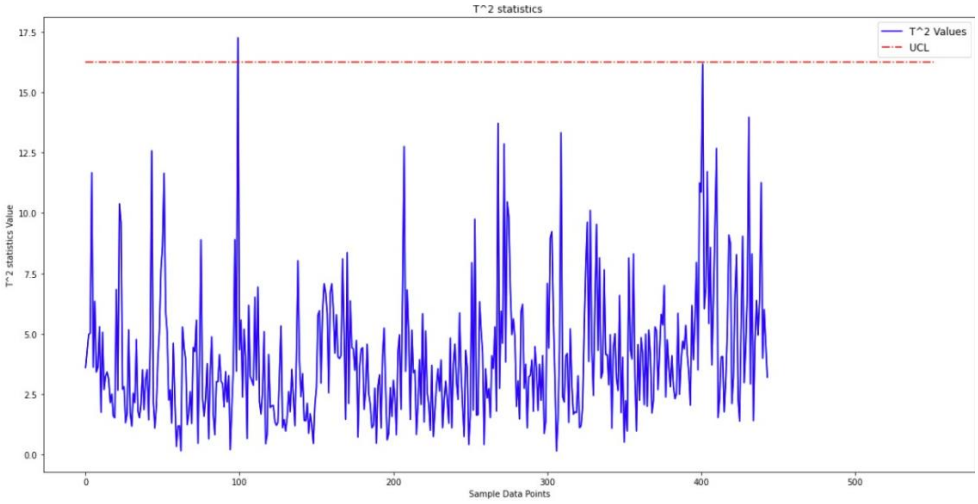
Plots for each iteration are as follows:

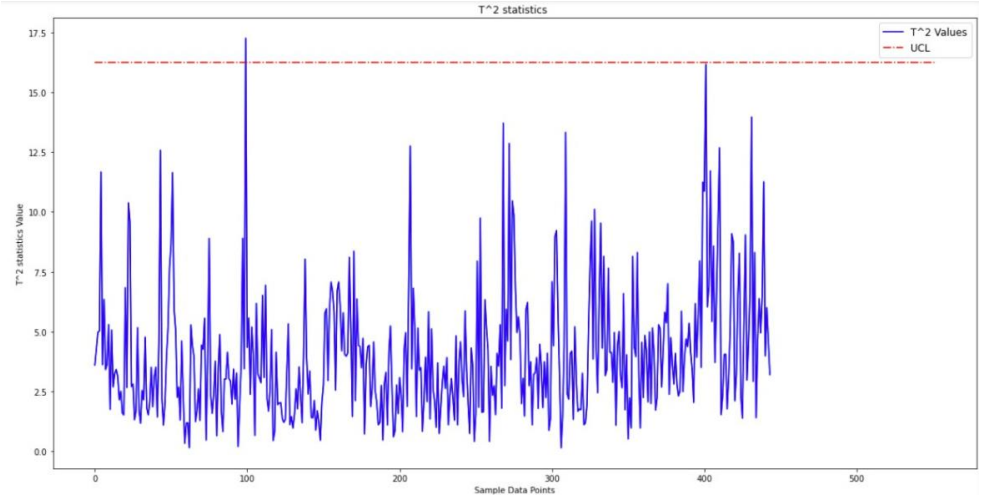


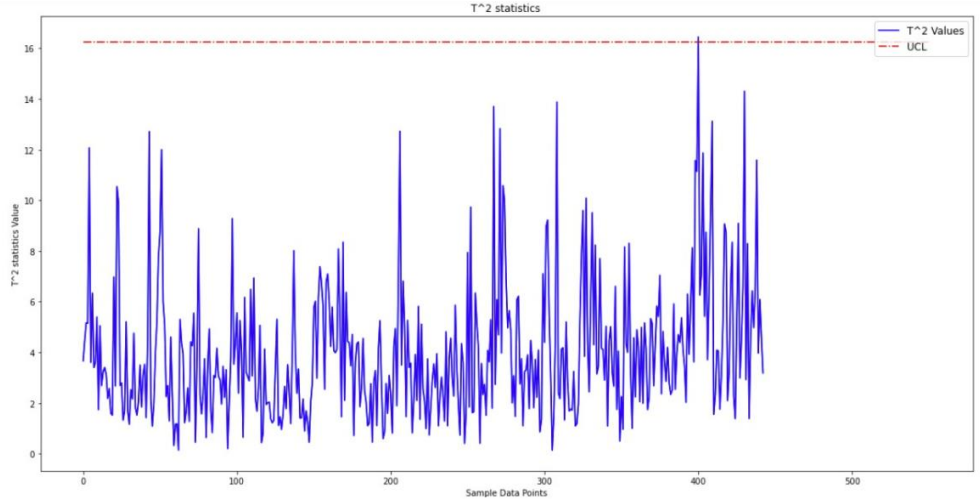
From the above iteration, it is observed that 8 points were out-of-control. Further iteration is required to remove the out-of-control points.

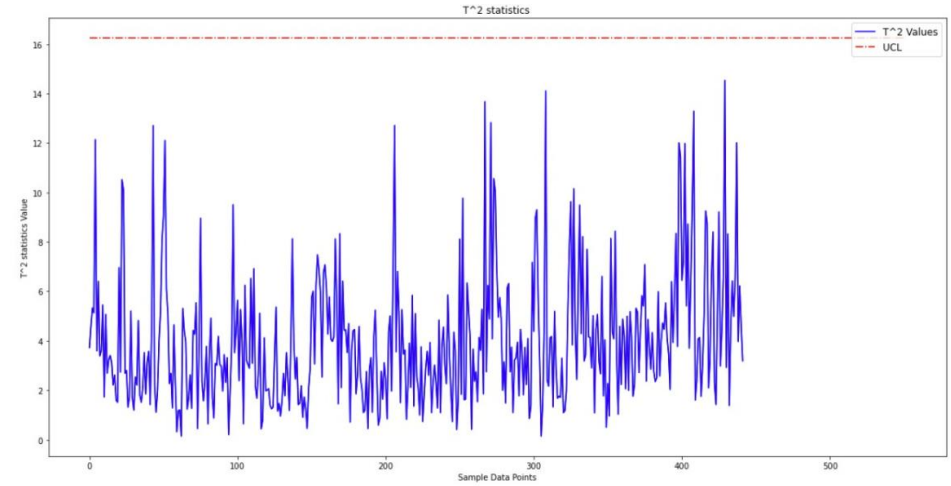


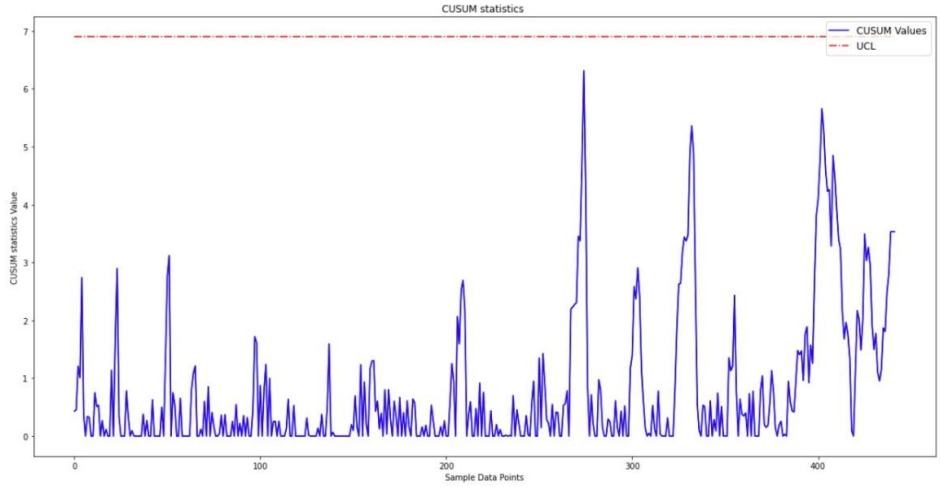
From the above plot, it can be observed that, 3 points are still out-of-control and more three iterations are performed to eliminate all the points. Plot for remaining three iterations is shown next slide:


T^2 statistics
T^2 Values
UCL
T^2 statistics Value
Sample Data Points


T^2 statistics
T^2 Values
UCL
T^2 statistics Value
Sample Data Points

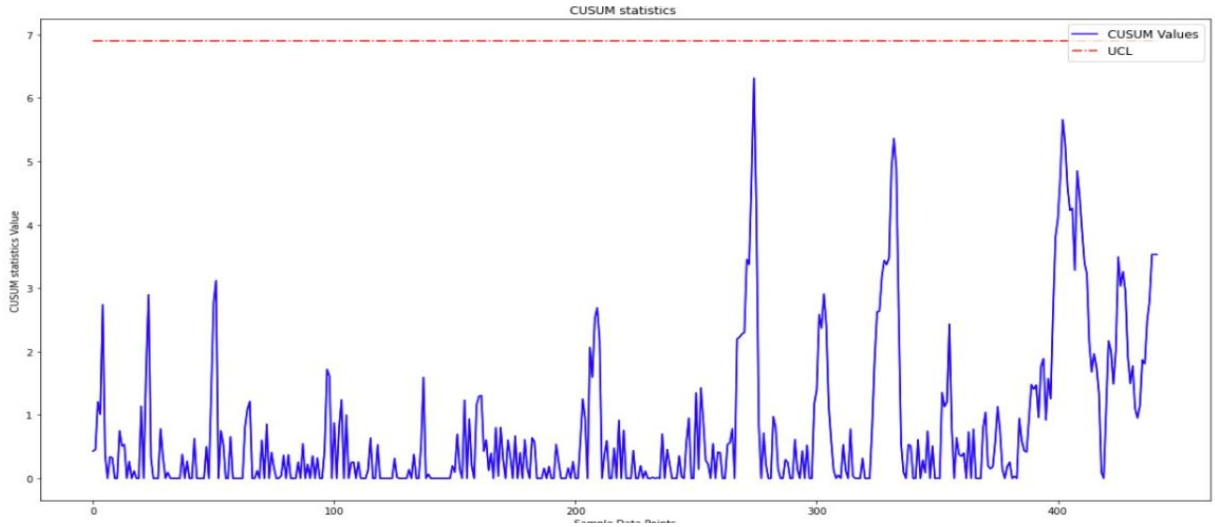

T^2 statistics
T^2 Values
UCL
T^2 statistics Value
Sample Data Points





Based on the aforementioned findings, it can be concluded that all 442 samples are in-control and do not show any spike-type alterations. However, in order to detect mean shift, m-CUSUM must be plotted again.

Final CUSUM Plot:



The T2 statistics of all 442 samples were plotted again after performing the Phase l Analysis of the m-CUSUM control chart. According to the graph below, 6 points are still out of control. As a result, Phase I Analysis must be repeated to guarantee that any out-of-control points have now been eliminated.

After finishing Phase I analysis with the m-CUSUM control chart, total in-control samples are 442

4. APPROACH 2

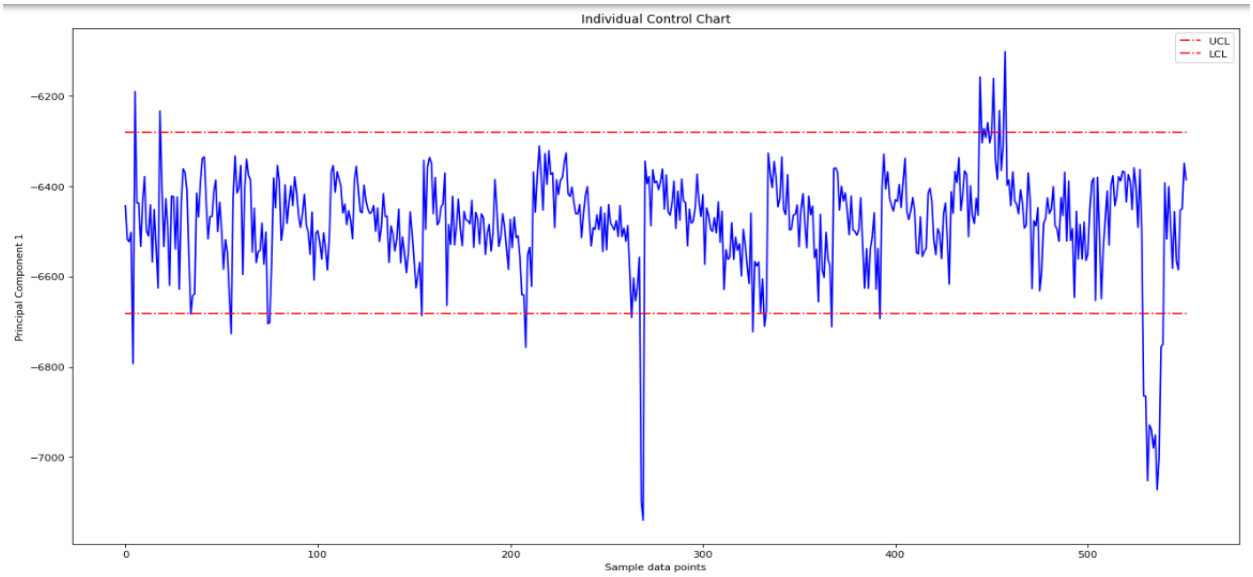
4.1 Multiple Univariate Chart

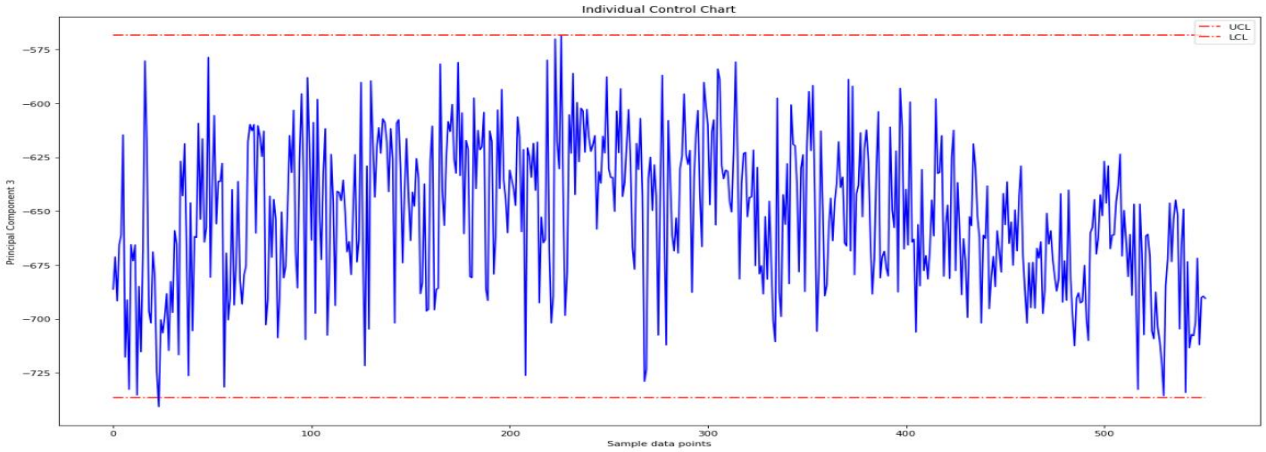
In this method, multiple univariate analysis was performed on the first four principal components derived from the prior results. The primary advantage of using multiple univariate charts on principal components is that it eliminates all correlations. As a result, monitoring individual PC's is advantageous since it ensures that any association between original data is not neglected. First, we need to calculate(estimate) the variance of these individual components by using the Moving Range(MR) method.

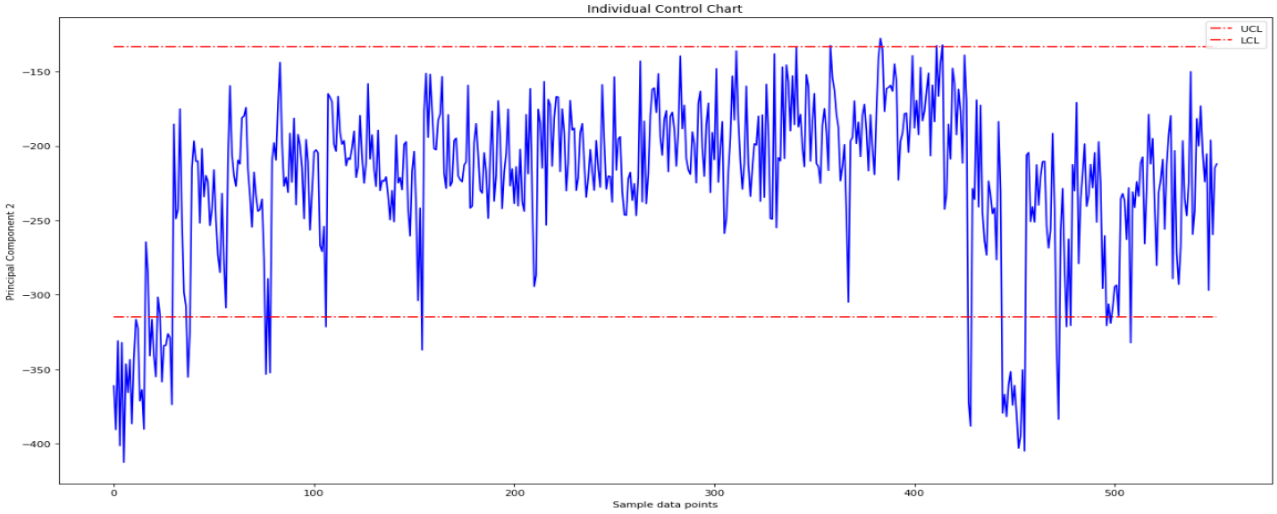
$$\frac{\overline{MR}}{d_2} = \sigma$$

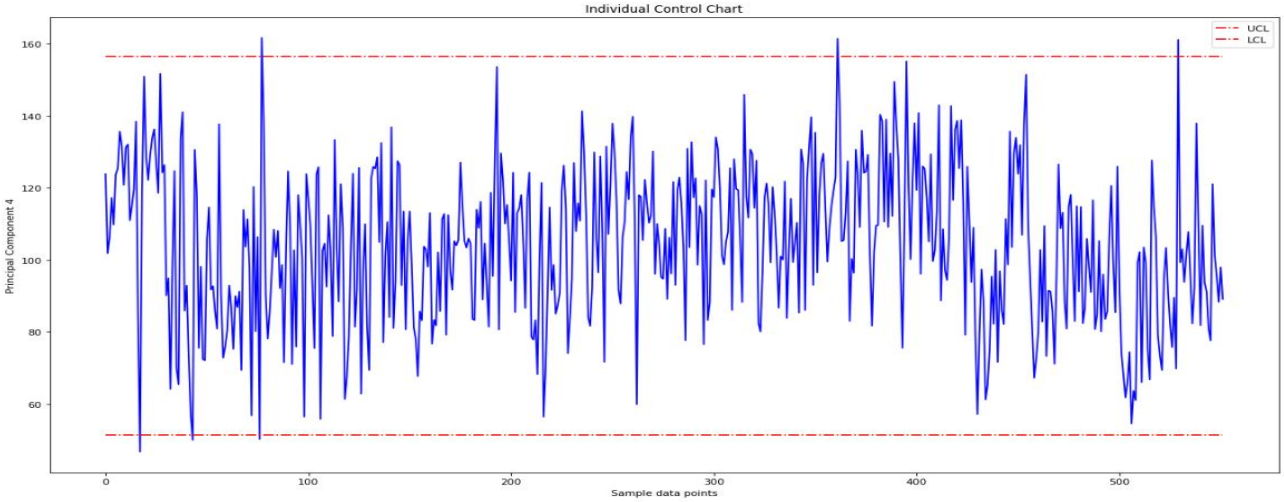
Further, we set UCL = $\bar{x} + L\sigma$
LCL = $\bar{x} - L\sigma$
α = 0.0027 and L = 3

<u>The four multiple univariate charts are plotted below:</u>




Individual Control Chart
UCL
LCL
Principal Component 3
−575
−600
−625
−650
−675
−700
−725
0
100
200
300
400
500
Sample data points


Individual Control Chart
UCL
LCL
Principal Component 2
−150
−200
−250
−300
−350
−400
0
100
200
300
400
500
Sample data points


Individual Control Chart
UCL
LCL
Principal Component 4
160
140
120
100
80
60
0
100
200
300
400
500
Sample data points

4.2 Decision Rules

To proceed further, several decision rules were created to identify the out-of-control points while ensuring that the inflation of Type-I and Type-II errors are minimized. The decision created are as follows:

Decision Rule	**Calculation of $\alpha_{combined}$**
Rule 1: either/or 1 plot out of control	$1 - Prob(None\ out\ of\ control \mid In\ control)$
Rule 2: Points from at least 2 plots out of control	$1 - Prob(None\ out\ of\ control \mid In\ control)$ – Prob (only 1 out of control | In control)
Rule 3: Points from at least 3 plots out of control	$Prob(3\ out\ of\ control \mid In\ control)$ + Prob (4 out of control | In control)
Rule 4: points from all 4 plots out of control	$Prob(all\ 4\ out\ of\ control \mid Process\ in\ Control)$

Above mentioned rules were used to identify out-of-control points, the results for the same are summarized below:

Decision Rule	**Number of Out-of-control Points identified**
Rule 1	86
Rule 2	11
Rule 3	0
Rule 4	0

From the above results, it can be observed that all the points are in-control as per Rule 3 and Rule 4. Therefore, we use Rule 1 and Rule 2 for further analysis. On the results obtained from multiple univariate plots, numerous iterations of Hotelling T2 statistics and m-CUSUM methods were applied to identify the out-of-control points.

Statistics Calculation:

UCL = $\chi^2_{1-\alpha_{\text{combined}}}(p)$ ……. (For T^2 statistics)

$p = 2$		$p = 3$		$p = 10$	
UCL_1	*ARL*	UCL_1	*ARL*	UCL_8	*ARL*
4.00	93.95	5.00	131.47	8.00	79.14
4.50	158.48	5.48	199.74	9.00	145.74
4.75	202.81	5.50	207.56	9.50	191.28
5.00	248.35	5.75	257.28	9.50	193.59
5.25	318.88	6.00	323.03	9.55	201.57
5.50	406.42	6.25	388.87	9.60	207.28
5.75	525.31	6.50	493.15	10.00	260.42
6.00	668.64	6.75	609.46	11.00	488.93
6.25	858.24	7.00	766.07	12.00	936.94
6.50	1054.82	7.25	976.63	13.00	1818.67
7.00	1748.86	7.50	1204.42	14.00	3514.75

…….. (For m-CUSUM)

The table above is used to calculate the value of UCL corresponding to the given ARL and p value using the interpolation method. The UCL corresponding to ARL = {93.95, 131.47, 79.14} and p = {2,3,10} is used to interpolate the value of UCL for Rule 1. Similarly, we calculate the UCL for Rule 2.

4.3 T^2 Statistics:

	UCL Using Rule 1	UCL Using Rule 2
m-CUSUM Statistics	5.19	8.605
T^2 statistics	13.099	26.6008

α = 0.0027

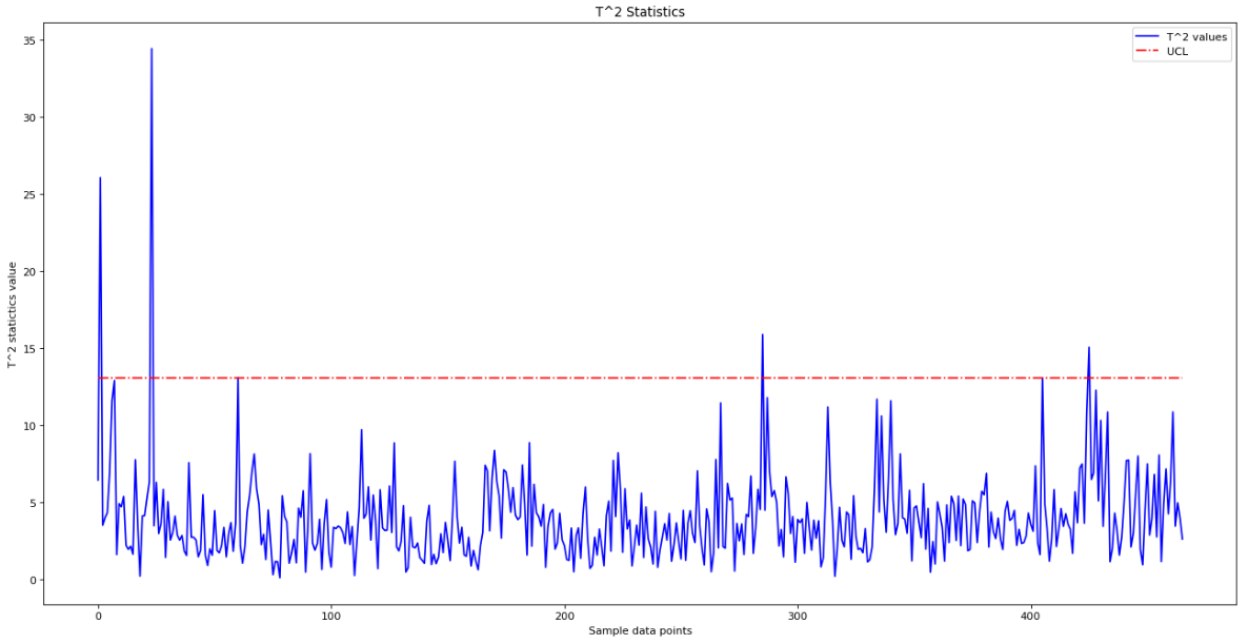
β is calculated using $\beta = \phi\left(z_{\alpha/2} - \delta^*\sqrt{n}\right) - \phi\left(-z_{\alpha/2} - \delta^*\sqrt{n}\right)$

Assuming, δ^* = 2 and n = 1

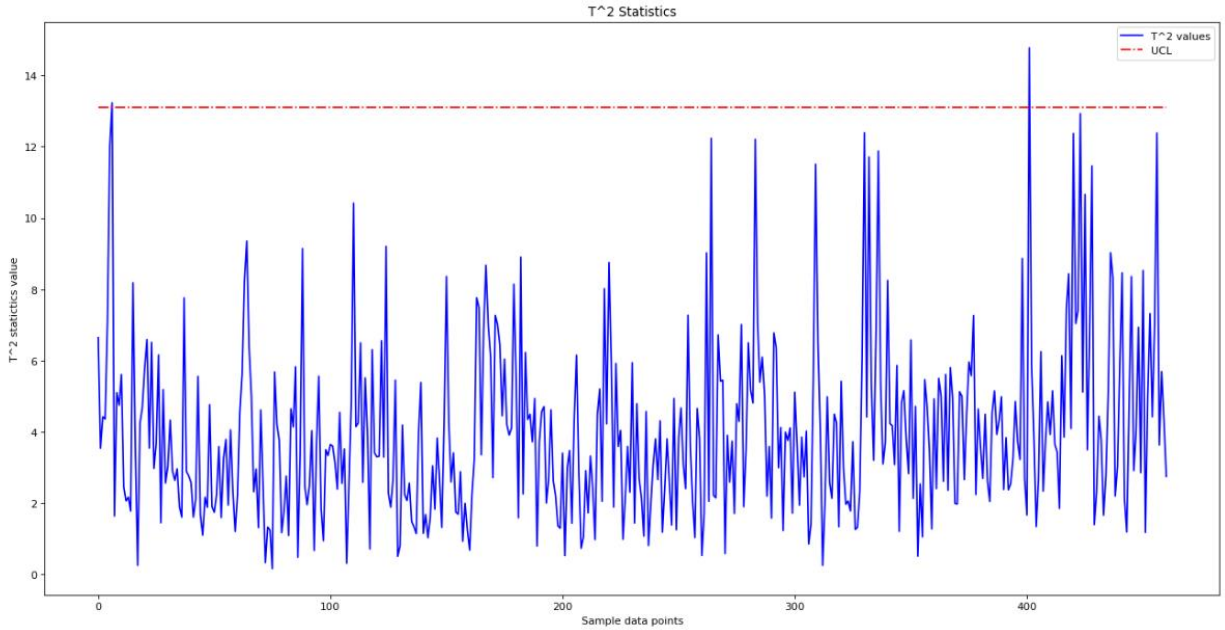
$ARL_0 = \frac{1}{\alpha}$ $ARL_1 = \frac{1}{1-\beta}$

Parameters	Rule 1	Rule 2	Original/Individual
Type - 1 error	4 * 0.0027 = 0.0108	0.70884	0.0027
Type – 2 error	4.3583e-05	0.98124	0.84134
ARL_0	92.592	22944.886	370.37
ARL_1	3.434	53.305	6.375

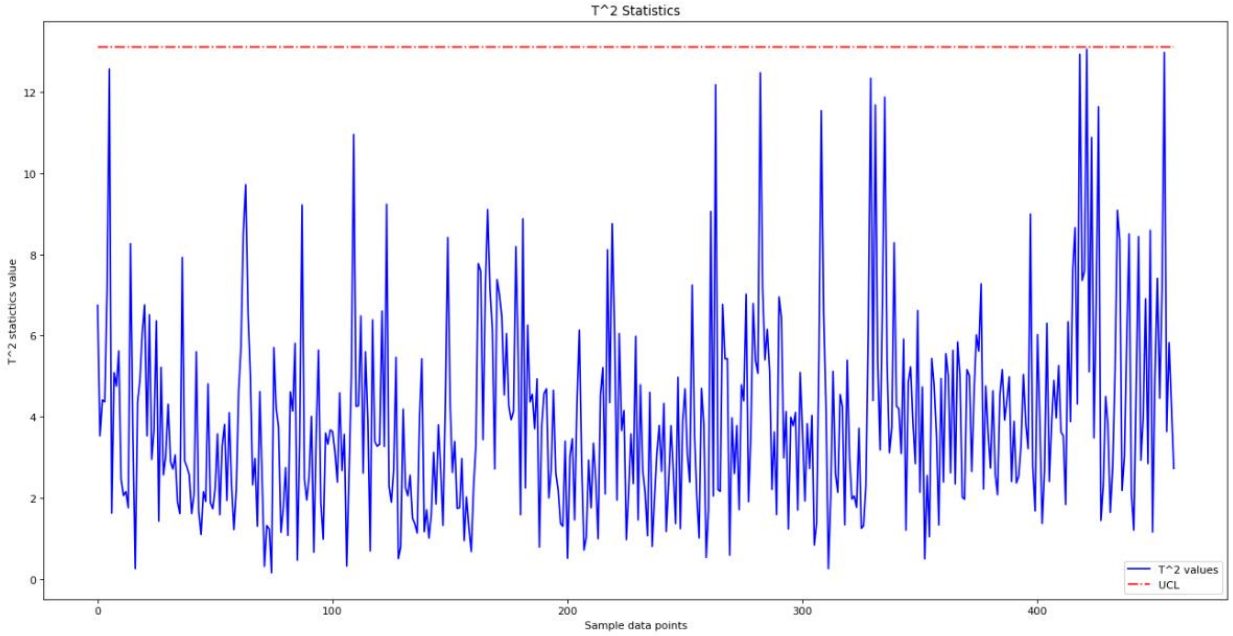
T^2 statistics plots for the result obtained based on Phase I Analysis of Rule 1 is given below :



From the above points it is observed that 5 points are still out-of-control after iteration. Hence, further iteration is required.



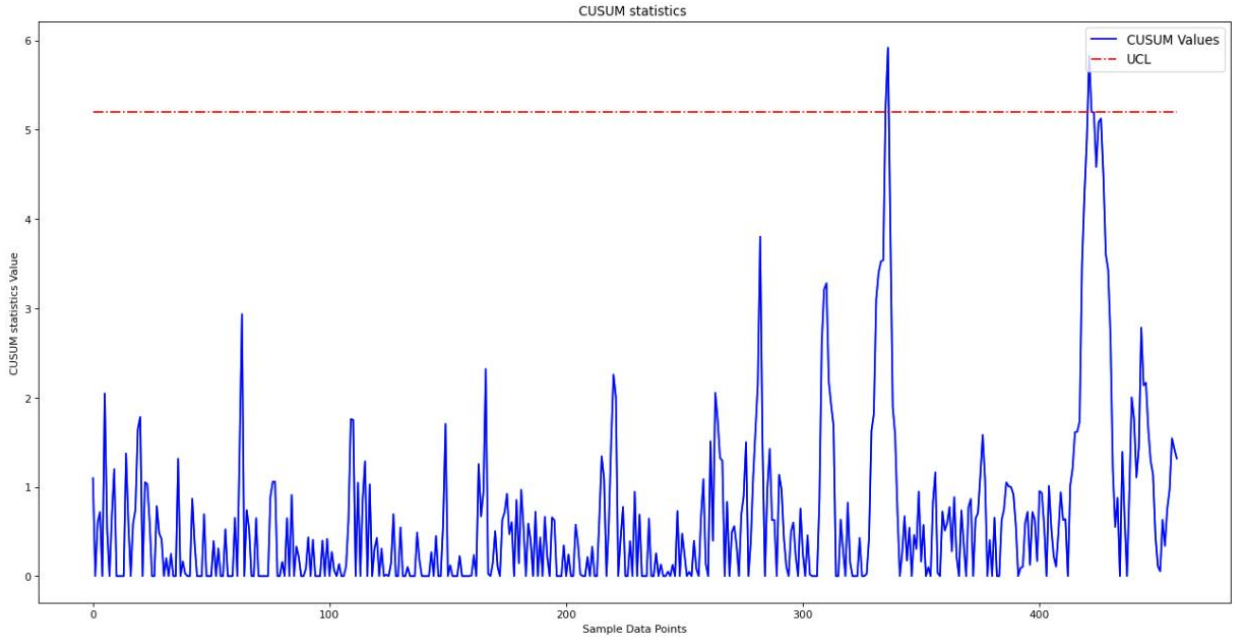
From the second iteration, still 2 points are out-of-control. Hence, one more iteration is required. All the data points are removed after completing the final iteration.



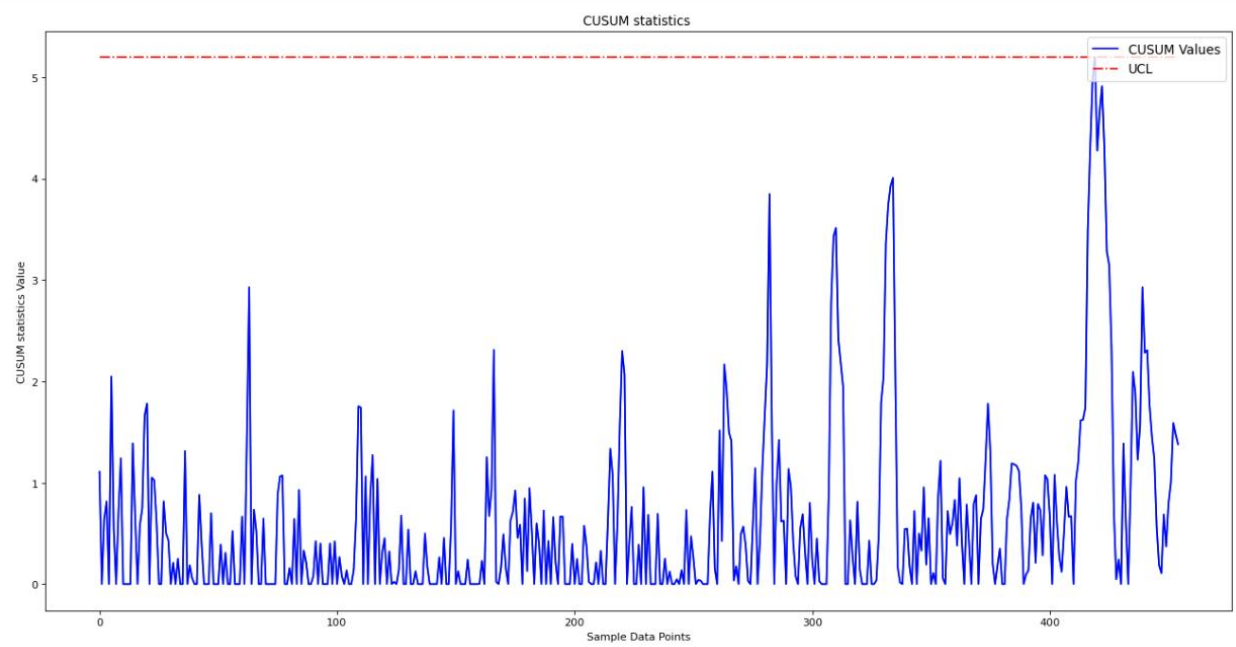
4.4 m-CUSUM method:

UCL was decided by first estimating ARL0 (ARL0 = 1/αcombined). Then UCL was estimated by interpolating the results for p=4 by using results of p =2,3,10.

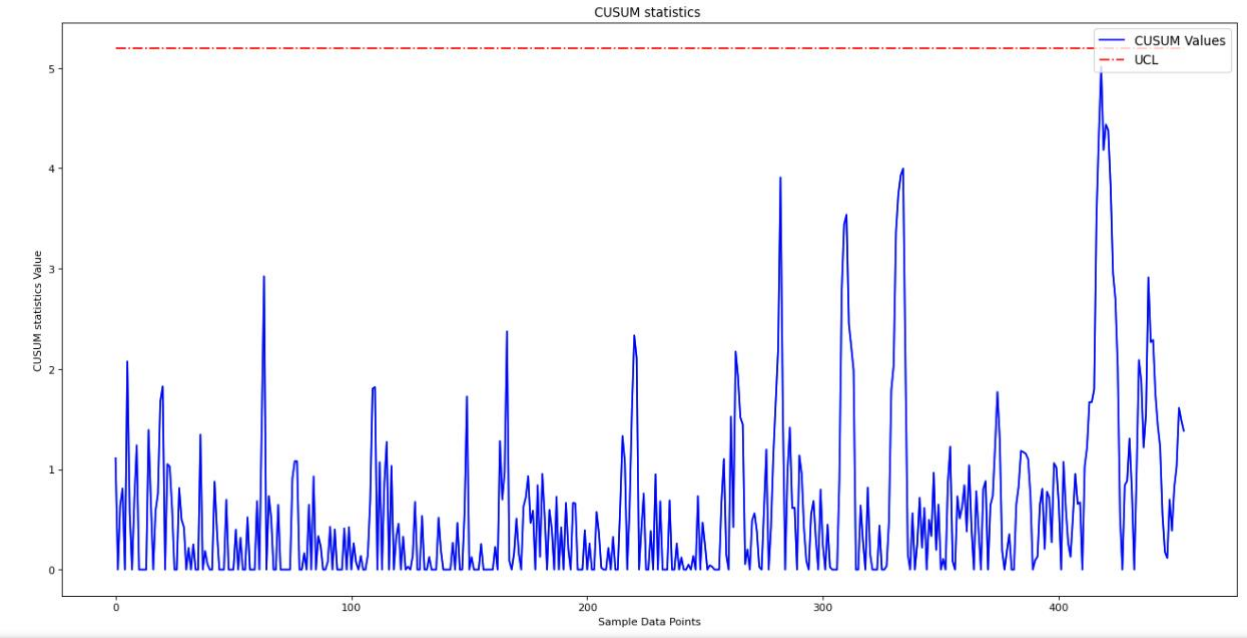
m-CUSUM plots for the result obtained based on Phase I Analysis of Rule 1 is given below :



From the above plot, it can be observed that after the first iteration 4 sample data points are still out-of-controls. As a result, further iterations is required to remove these 4 out-of-control data points.



After performing the 2nd iteration, it is observed that one data points is still out-of-control and hence final iteration is done to remove this remaining data point.



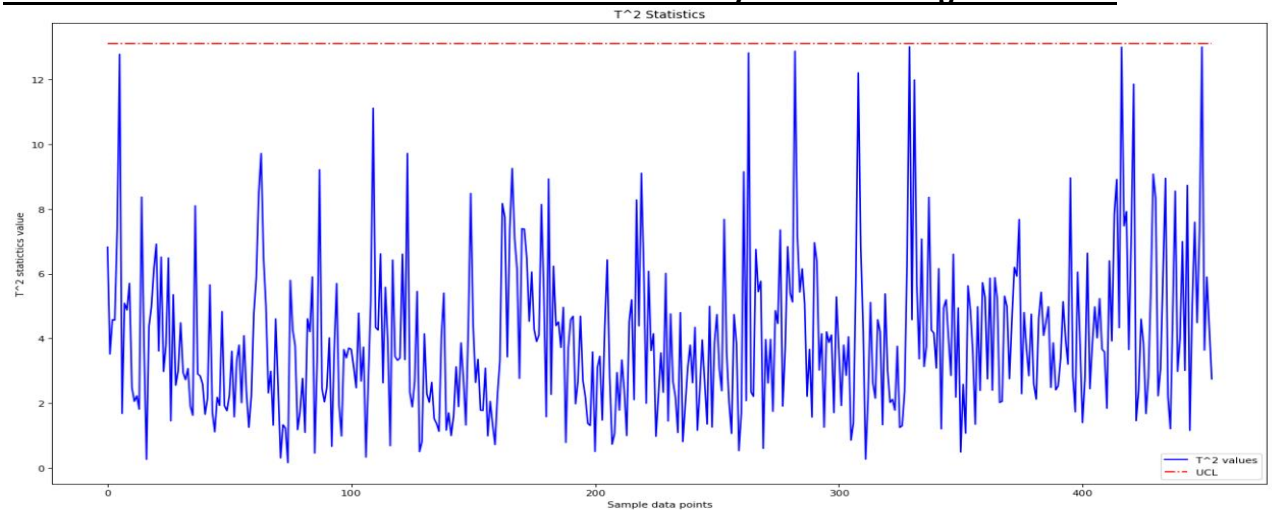
The results are summarized below:

Methods Applied	**Iterations**	**Rule 1**	**Rule 2**
T^2 (First time)	Iteration 1	5	15
	Iteration 2	2	14
	Iteration 3	0	16
	Iteration 4	(Not Required)	12
	Iteration 5	(Not Required)	4
	Iteration 6	(Not Required)	2
	Iteration 7	(Not Required)	0
m-CUSUM Method	Iteration 1	4	0
		1	(Not Required)
		0	(Not Required)
T^2 (Final)	Iteration 1	0	(Not Required)
Total points removed		12	63

So total number of points removed based on rule 1 is 86+12 = 98
So total number of points removed based on rule 2 is 11+63 = 74

Based on this analysis we decide to proceed with decision Rule 1 even though the Type 1 error is inflated because of the limitation of data table included in the slides to estimate UCL value based on ARL0 and p. To rectify this mistake in future Monte Carlo simulation can be done to estimate UCL for such high ARLo value.

Final T^2 Plots for result obtained based on Phase 1 analysis of Rule 1 is given below :



5. RESULTS AND CONCLUSION

5.1 Reason for choosing Approach 2 over Approach 1:

1. The application of multiple univariate analysis on principal components is beneficial because correlation among the process parameters is eliminated. Thus, analysis can be done more easily.

2. Another benefit is, in approach 2 smaller number of points are removed to achieve the in-control state as compared to approach 1, so parameters calculated for future state will be more accurate as we have relatively more data points for estimation.

5.2 Final Result:

Based on the method selected 98 no. of out of control were identified and removed, and the total In-Control samples points were = 552-. Based on the exploration of data we earlier had made a reasonable assumption that it will follow a normal distribution, so using the in-control points identified mean and covariance matrix were calculated which will serve as parameters for future data.

$$\mu_0 = [2.01262469,\ \ 4.17743157,\ -4.93914635,\ \ 0.48972496]$$

$$\Sigma_0 = \begin{bmatrix} 0.00887411 & 0.00714826 & -0.00414948 & 0.00158061 \\ 0.00714826 & 0.00775869 & 0.00707732 & 0.0145635 \\ -0.00414948 & 0.00707732 & 0.13669532 & 0.14708979 \\ 0.00158061 & 0.0145635 & 0.14708979 & 0.16534893 \end{bmatrix}$$

6. REFERENCES

- On multivariate monitoring: R. A. Johnson, R. A., and D. W. Wichern (2001). Applied Multivariate Statistical Analysis (5th Edition), Upper Saddle River, NJ: Prentice Hall.

- On statistical process control: D. C. Montgomery (2003). Introduction to Statistical Quality Control (3rd Edition). New York City: John Wily and Sons.

- https://www.rdocumentation.org/

- Dr. Yu Ding's course material