`

Phase I Analysis of Manufacturing Process Dataset

Team ID: 20

Gaurav Rai, Gaurav Burman and Atul Srivastava
Data Analyst, Industrial & Systems Engineering
Texas A & M University, TX- 77840

Executive Summary

Quality Control is the process of change or anomaly detection in a process. We use statistical procedures to compare standard population parameters to the sample parameters and declare that the process is out-of-control when there is deviation from the standard 'in-control' parameters of more than a specified level. However, many times the 'in-control' parameters, i.e mean and variance, are not available to us and we require to estimate them from some historical data. But what is authenticity that the historical data is really in-control and does not consists of out-of-control data points. Here comes the significance of Phase-I analysis which constitutes identifying and removing out-of-control data points from the historical data and then using the in-control data points to calculate control chart parameters for the future data. we performed Phase-I analysis over the dataset provided to us from a manufacturing firm.

This is a multivariate data with 209 attributes. We have used two approaches for isolating the in-control observations. The first approach uses Hotelling's T^2 statistics over all 209 variables of the dataset. The out-of-control data points in the T^2 chart are removed from the dataset and the process is re-iterated for the remaining data points until we end up with all in-control data points. However, we believe that 209 attributes can make the noise components add up to a great magnitude resulting in weak signal to noise ratio making it hard to reject the null hypothesis. Hence, T^2 may not be that effective in detecting out of control data points.

The second approach is the Principal Components Analysis (PCA) approach, wherein we have transformed the dimension of our dataset by focusing on only the 'vital few' principal components, that offer the maximum variability and reduce the effect of cumulative noise in high dimensions. Using Pareto chart, Scree plot and MDL plot in conjunction, we decided to choose the top 10 principal components for our analysis. Additionally, we have plotted multi-univariate charts for these principal components since the data obtained through PCA is uncorrelated. We have used the following rule - If either of the 10 charts detected an out-of-control point, the entire observation is out-of-control. Like the T^2 chart approach, we have performed an iterative study here by carrying out PCA until we end up with all in-control data points. Since we are not given the details of units of each attribute, the original relative magnitude in deviation is not important to the subsequent inference, and we standardized the data before PCA which brought all the variables to same scale with a mean value of 0.

Through the T^2 chart approach on the raw data, we performed 13 iterations until we got a set of 446 in-control data points from the initial 552 data points. After applying the PCA to the data, in multi-univariate analysis we performed 4 iterations to get 515 in- control data points and on applying T^2 chart to the data obtained after PCA we got 438 in-control data points in 8 iterations. These results illustrate that T^2 chart after PCA detected more out of control observations as compared to the multi-univariate analysis because of the curse of dimensionality individual significance levels α had to be kept high to get a composite decision α of 0.05 resulting in wide confidence intervals for the multi-univariate charts hence less out of control points detected. So we will use Hotelling's T^2 charts on first 10 principal points to remove out of control data points to find parameters of our control charts for future observations.

1. Hotelling T^2- chart

We have been provided a dataset with 552 observations, p= 209 variables, and sample size n = 1. We must perform phase-I analysis on this multivariate dataset. Our desired significance level is $\alpha = 0.05$. This alpha level implies that we want only 5% chance of false alarm or a detection even if the process is in control. T^2 formula is given by

$$T_i^2 = (X_i - \mu)^T S^{-1}(X_i - \mu)$$

Where Xi is the observation matrix for ith observation, μ is the calculated mean for each variable and S is the calculated 209X209 covariance matrix for the data.

1.1 Approach

- We assume that the multivariate data follows a normal distribution. We find this assumption to be reasonable as any process with many continuous variables eventually approximate to a normal distribution.
- T^2 statistic was calculated for all observations and plotted along the index. UCL is $\chi^2_{1-}\ \alpha(209)$, where 209 is our degree of freedom. Plot is shown in Fig.1.
- The first T^2 Chart gave us 44 out of control points. To establish the in-control mean and variance, we need to remove these points and perform Hotelling T^2 test repeatedly until all observations are below UCL.
- It took 13 iterations and a total of 106 observations were filtered out as out-of-control after the process. The observation indexes with the T^2 statistics for all values have been attached in Appendix.
- The T^2 Charts for all iterations are as shown below:

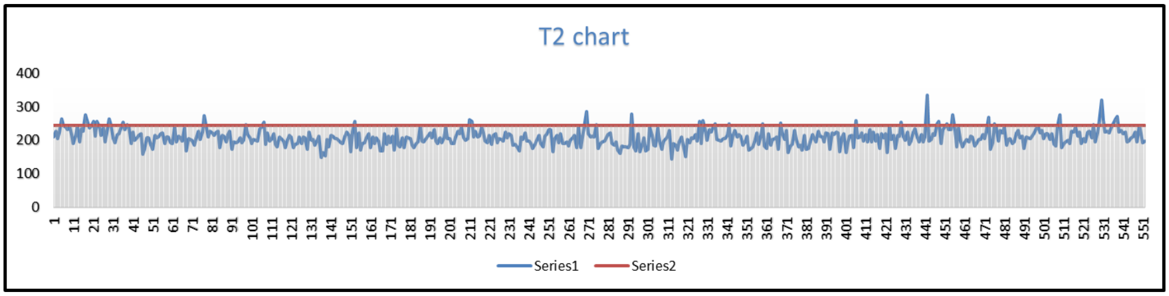


Fig.1A: Iteration1 - Hotelling T^2 Chart for all 552 observations. 44 observations are found to be above UCL.

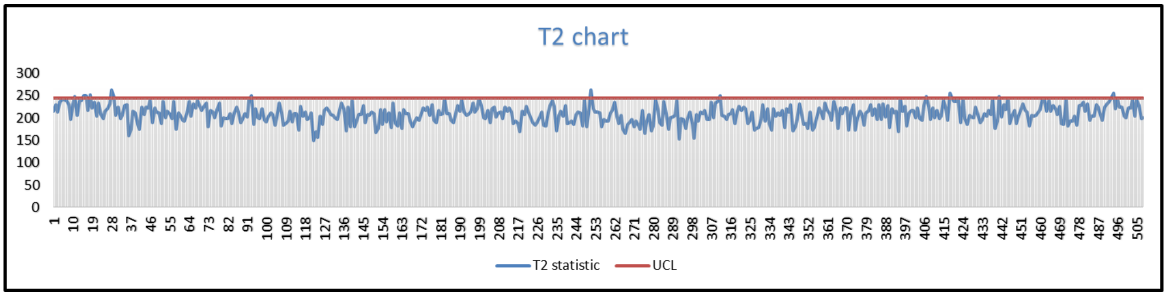


Fig.1B: Iteration2 - Hotelling T^2 Chart for all 508 observations. 19 observations are found to be above UCL.

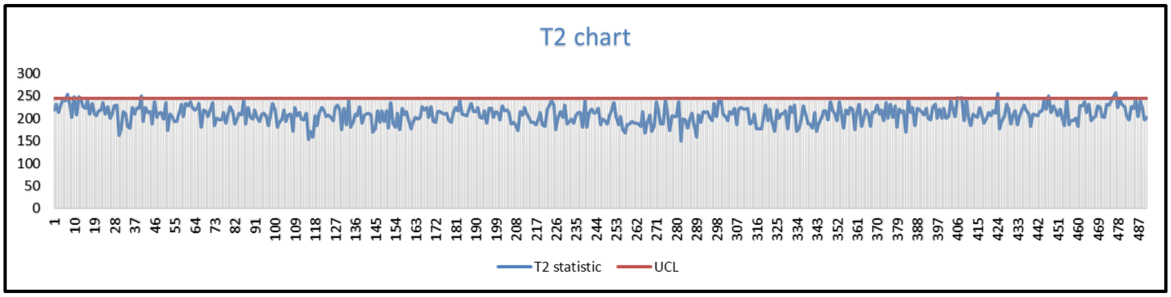


Fig.1C: Iteration3 - Hotelling T^2 Chart for all 489 observations. 15 observations are found to be above UCL.

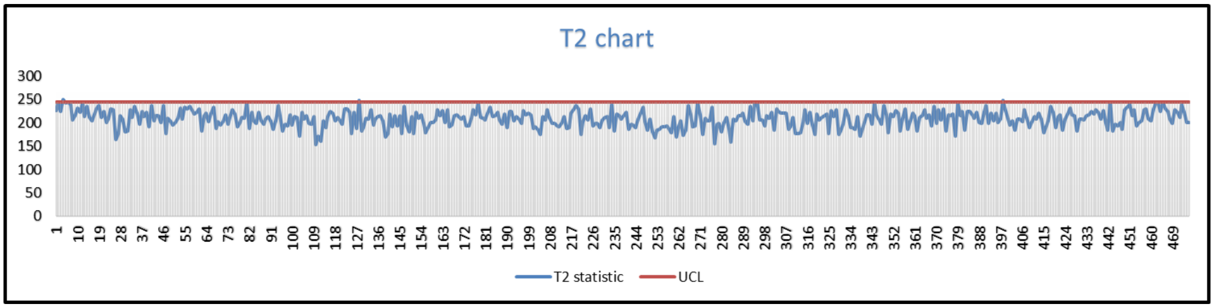


Fig.1D: Iteration4 - Hotelling T^2 Chart for all 474 observations. 11 observations are found to be above UCL.

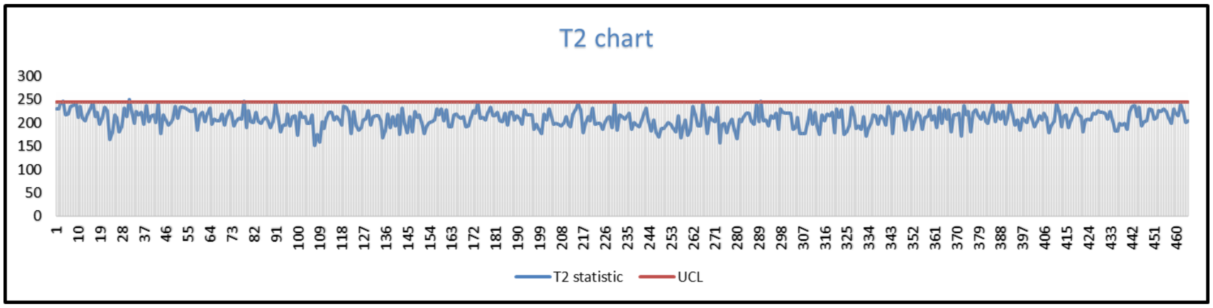


Fig.1E: Iteration5 - Hotelling T^2 Chart for all 463 observations. 5 observations are found to be above UCL.

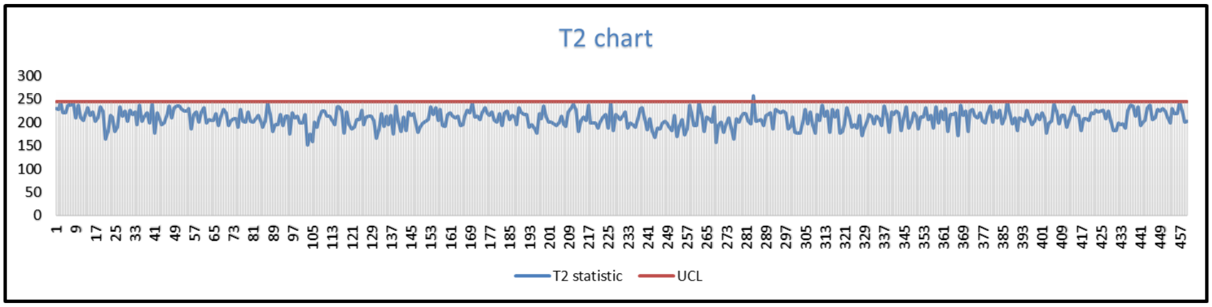


Fig.1F: Iteration6 - Hotelling T^2 Chart for all 458 observations. 3 observations are found to be above UCL.

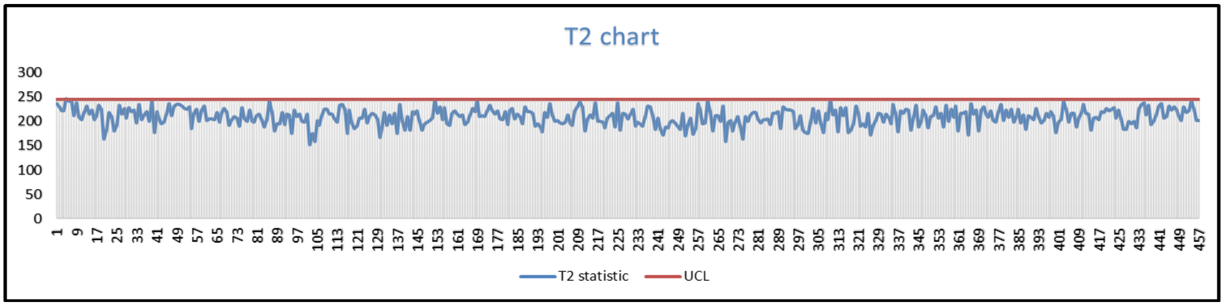


Fig.1G: Iteration7 - Hotelling T^2 Chart for all 455 observations. 3 observations are found to be above UCL.

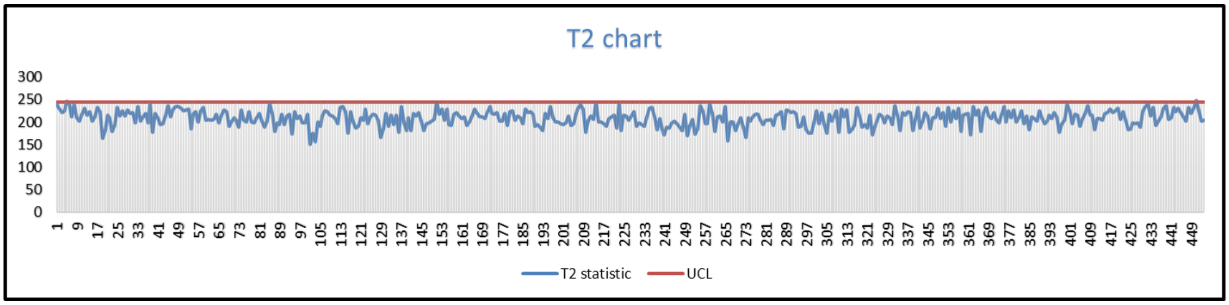


Fig.1H: Iteration8 - Hotelling T^2 Chart for all 452 observations. 2 observations are found to be above UCL.

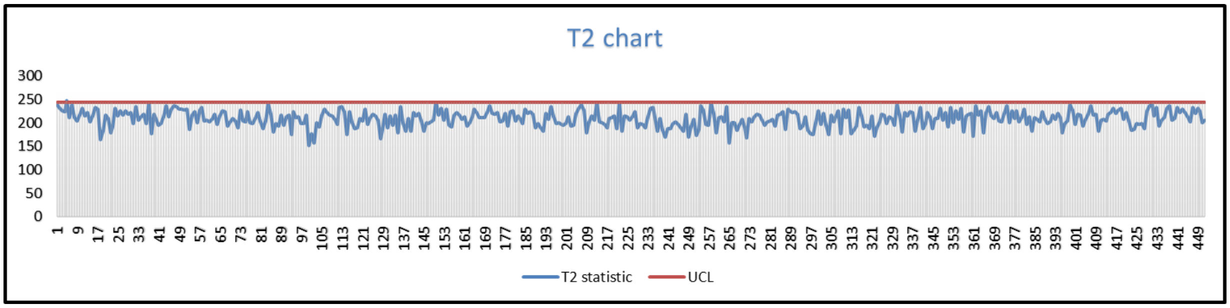


Fig.1I: Iteration9 - Hotelling T^2 Chart for all 450 observations. 1 observation is found to be above UCL.

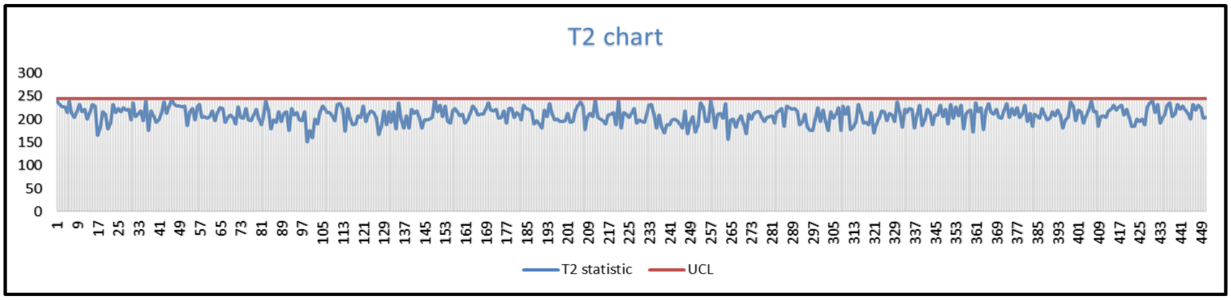


Fig.1J: Iteration10 - Hotelling T^2 Chart for all 449 observations. 1 observation is found to be above UCL.

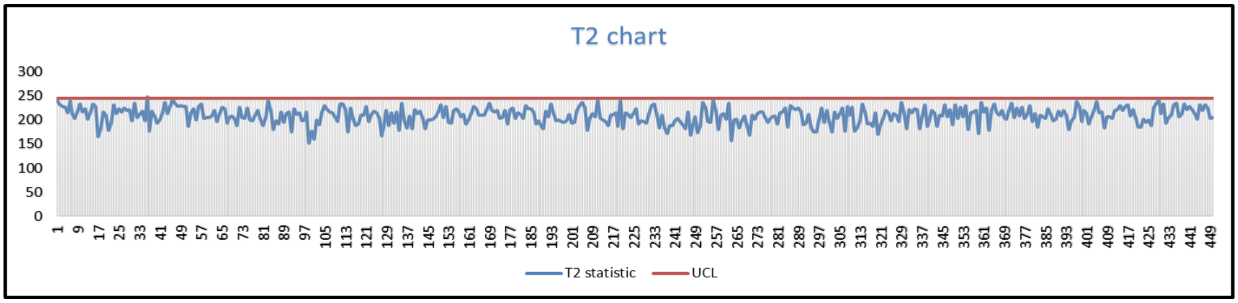


Fig.1K: Iteration11 - Hotelling T^2 Chart for all 448 observations. 1 observation is found to be above UCL.

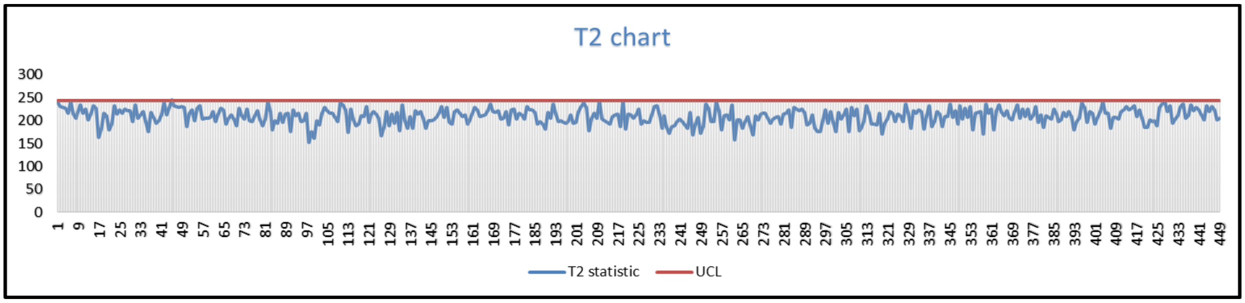


Fig.1L: Iteration12 - Hotelling T^2 Chart for all 447 observations. 1 observation is found to be above UCL.

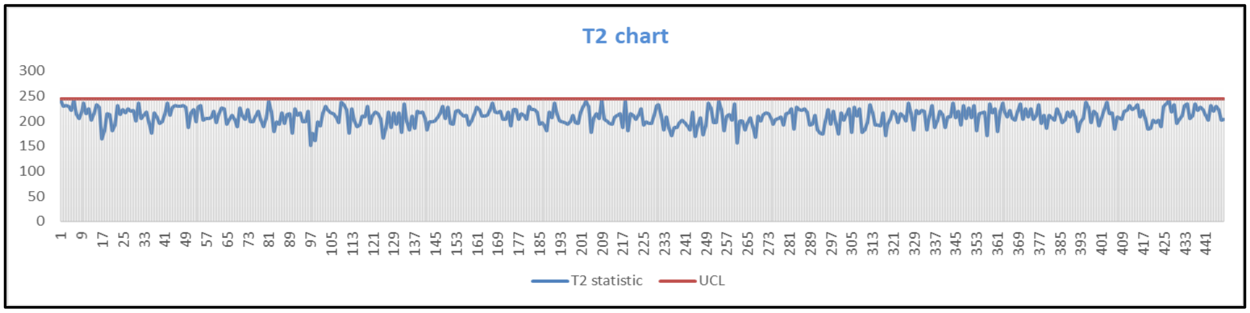


Fig.1M: Iteration13 - Hotelling T^2 Chart for all 446 observations. All observations are under control

1.2 Justification

The multivariate T^2 is preferred here over the univariate analysis as it gives us 209 different control charts, studying which is a tedious task. Besides, the following reasons add more weight to the choice of multivariate analysis using Hotelling T^2 Statistic.

- T^2 chart takes care of the covariance between the variables which cannot be observed in individual univariate charts.
- The combined significance level($\alpha_{composite}$) of 209 variables will be calculated from individual α levels and our criteria of selecting these observations out-of-control as per the control chart. Getting the desired $\alpha_{composite}$ and $\beta_{composite}$ values for the control charts is very difficult because of the very high dimensionality of the data. This adjustment of α and β values could be avoided using T^2 Chart.

1.3 Results

Performing 13 iterations on T^2 Chart, we eliminated 106 observations as follows:

Iteration	Total No. of observations	No. of OOC observations
1	552	44
2	508	19
3	489	15
4	474	11
5	463	5
6	458	3
7	455	3
8	452	2
9	450	1
10	449	1
11	448	1
12	447	1
13	446	0

1.4 Conclusion

Application of T^2 Control charts eliminated almost 20% of the observations as out of control. The results of T^2 are not reliable as we have 209 variables and as mentioned earlier, the noise levels may add up and overwhelm the detection process. Which means that the number of points can be even more than those detected by T^2 Chart. To overcome this characteristic, we applied Principal Component Analysis in next section to reduce the dimensions of this data set. So that we ca use multi-univariate control charts to analyze the data.

2. Principal Component Analysis

Principal Component Analysis is the process of determination of uncorrelated variables having largest variances from correlated variables of a given data. These uncorrelated variables are called Principal Components(PCs) and are derived as a linear combination of 209 variables of the data. First PC explains the largest amount of variance in the data, second PC explains the second largest amount of variance in the data, and so on ith PC explains the ith largest amount of data. After PCA, we use these PCs as our individual variables. To achieve data reduction, we select the first few PCs which explain approx. 75% of variance in data. In other words, we can say that we were able to derive the same amount of information from far lesser number of variables than that in original data. This would save us from the 'curse of dimensionality' and the effect of aggregation of noise that subdues the detection efficiency in T^2 chart.

2.1 Approach

- Since no information/units is provided for the variables in the data, we decided to standardize the dataset prior to PCA so that all the variables get equal importance in determination of PCs.
- Performing PCA, we obtained 209 PCs for 552 observations. To choose the optimum number of PCs we made use of MDL plot in conjunction with scree plot and pareto plot. Fig2 ,Fig3 & Fig4. illustrate MDL Plot, Scree plot and Pareto plot for different PCs.
- We chose the number of PCs explaining at least 70% variance in data and analyzed the PCs using univariate Shewhart chart and multivariate T^2 chart for the combined PCs. This much variance was captured by first 10 PCs. It should be noted that the MDL plot gives the number of PCs as 12. The pareto chart returns the number 15, if we choose to consider the PCs that contribute to 80% of the variability. The scree plot returns the number 5. After doing an intuitive analysis along with the results got from the above three plots, we have chosen the number of PCs as 10.
- Since we wanted the overall significance level to be $\alpha = 0.05$, the α` for individual control charts was calculated by composite decision rule of at least one of all the sample means falls outside the control limits which gives alpha approx. α` ~ $\alpha/10 = 0.005$.
- This brought the UCL for the univariate chart from the standard 3-sigma to a lower value of 2.81-sigma.
- T^2 chart was also applied for multivariate analysis treating PCs as variables.
- Both the aforementioned methods (T^2 & Univariate charts) have been discussed later in this report.
- The plots for choosing the optimum number of principal components are as shown below:

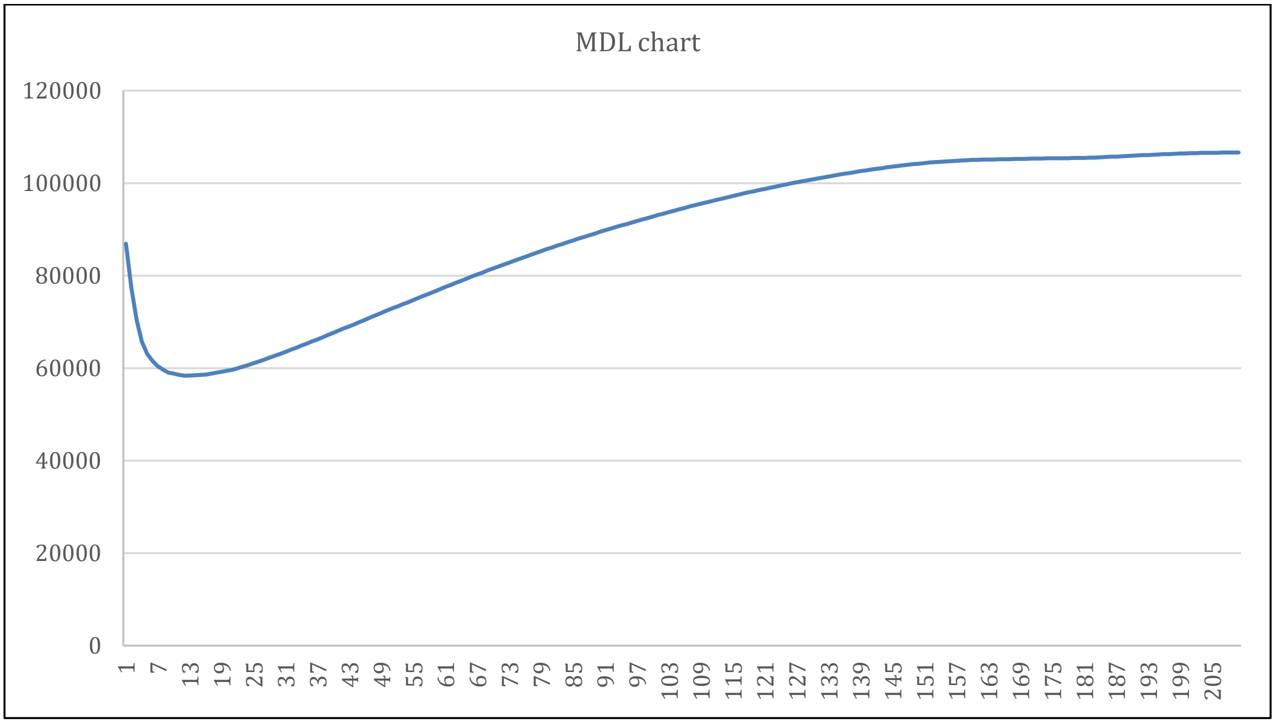


Fig.2: MDL plot for Eigen values

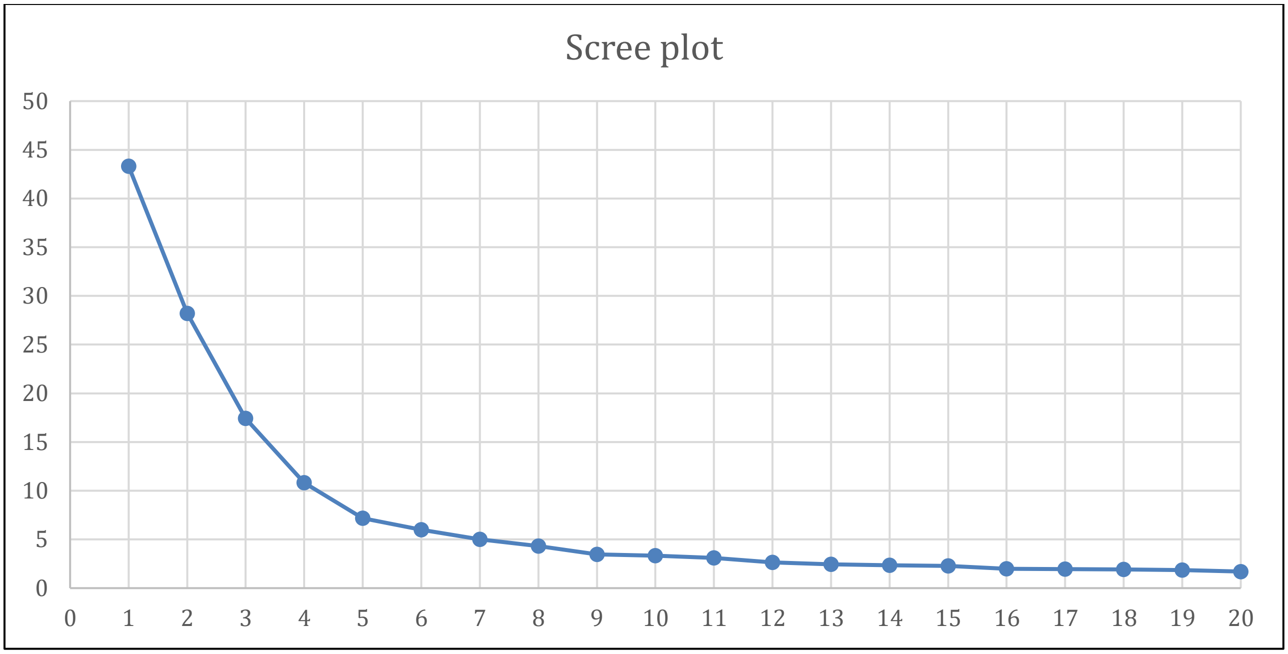


Fig.3: Scree plot for Eigen values

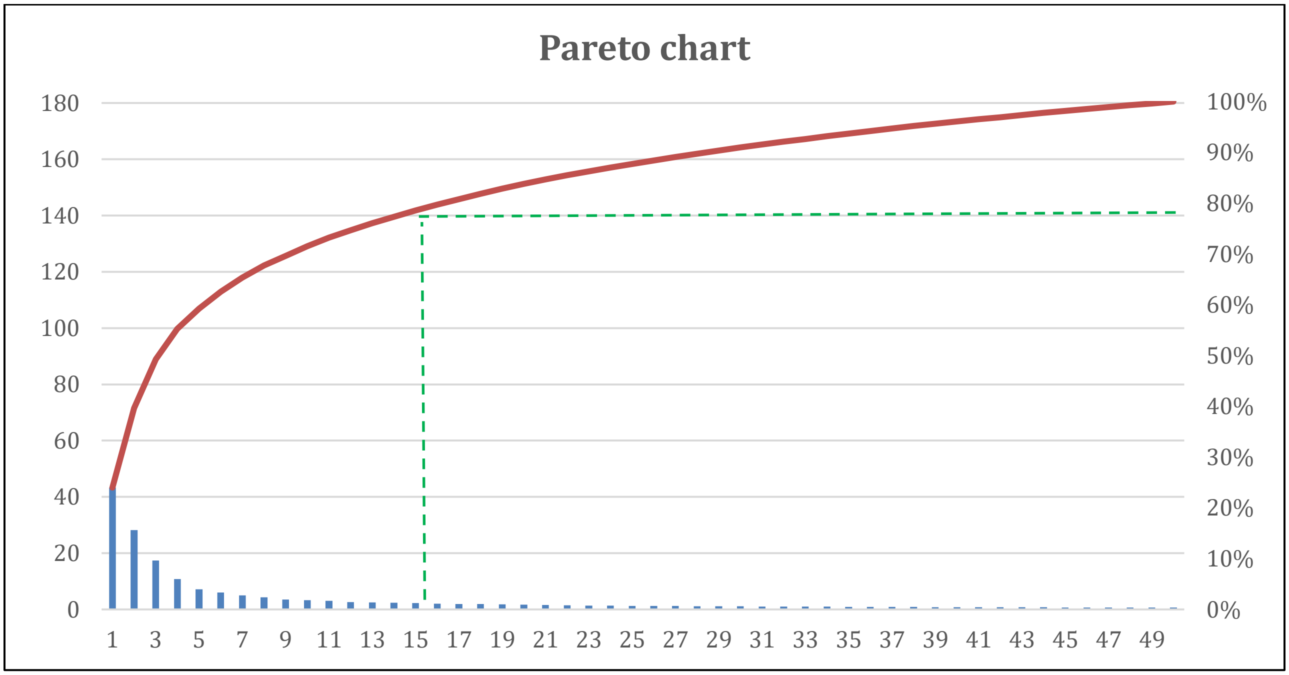


Fig.4: Pareto chart for Eigen values

2.1.2 Multi-Univariate chart analysis on first 10 PCs.

- We have performed multiple iterations using the univariate charts of all the first 10 PCs until all the OOC points were removed.
- While performing this analysis, we have treated the PCs as an individual set of data and performed Shewhart control chart on it.
- UCL was set at 2.81 sigma which we got using composite decision rule for an overall 95% confidence interval.
- We used minitab software for the multi univariate alalysis.

The first iteration of X chart on the first 10 PCs are as shown below:

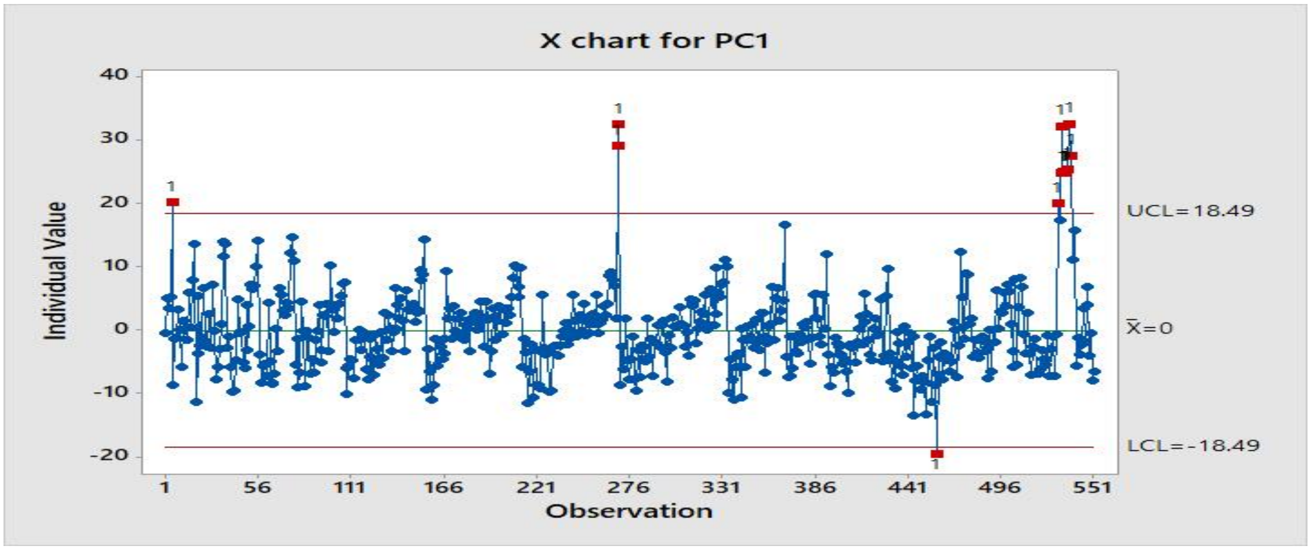
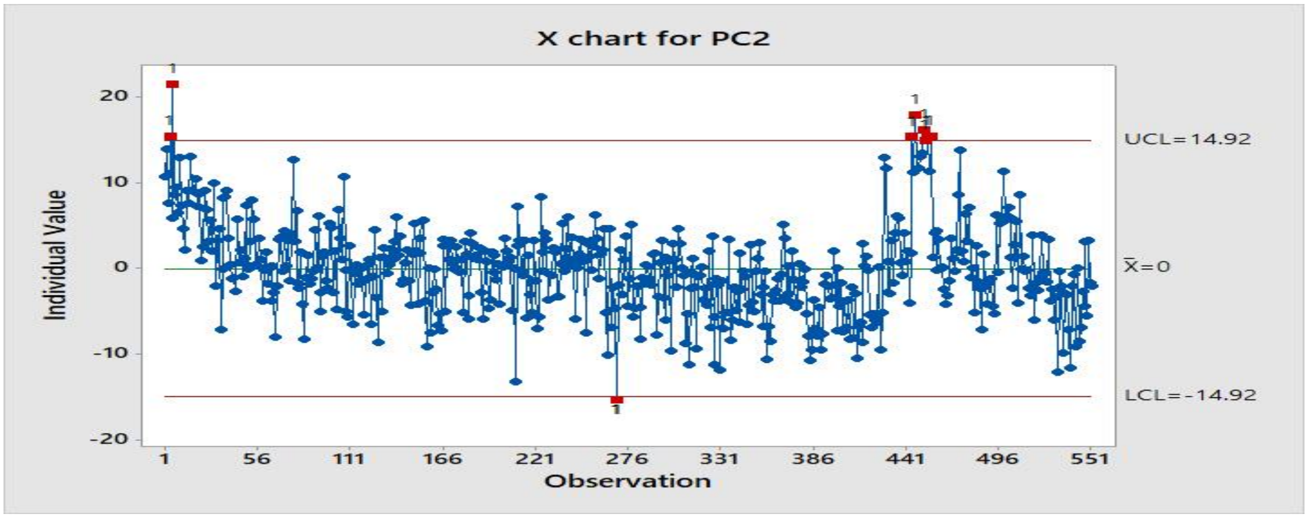
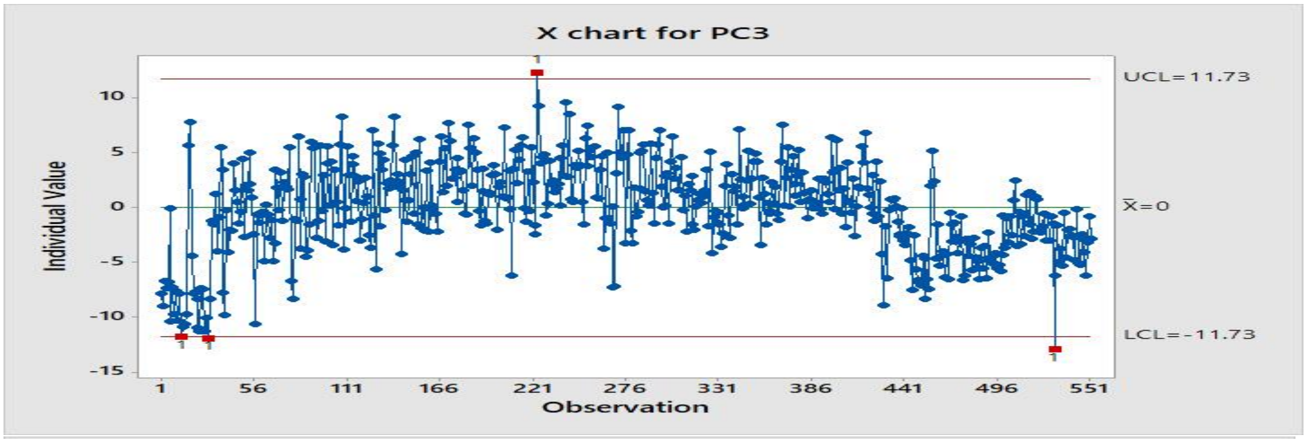
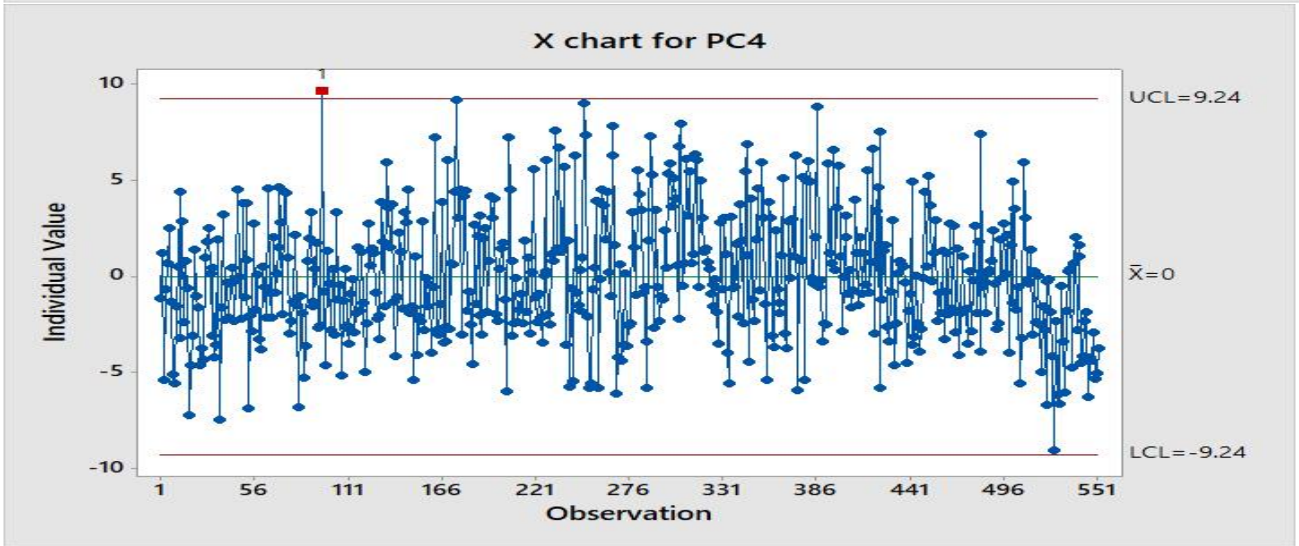
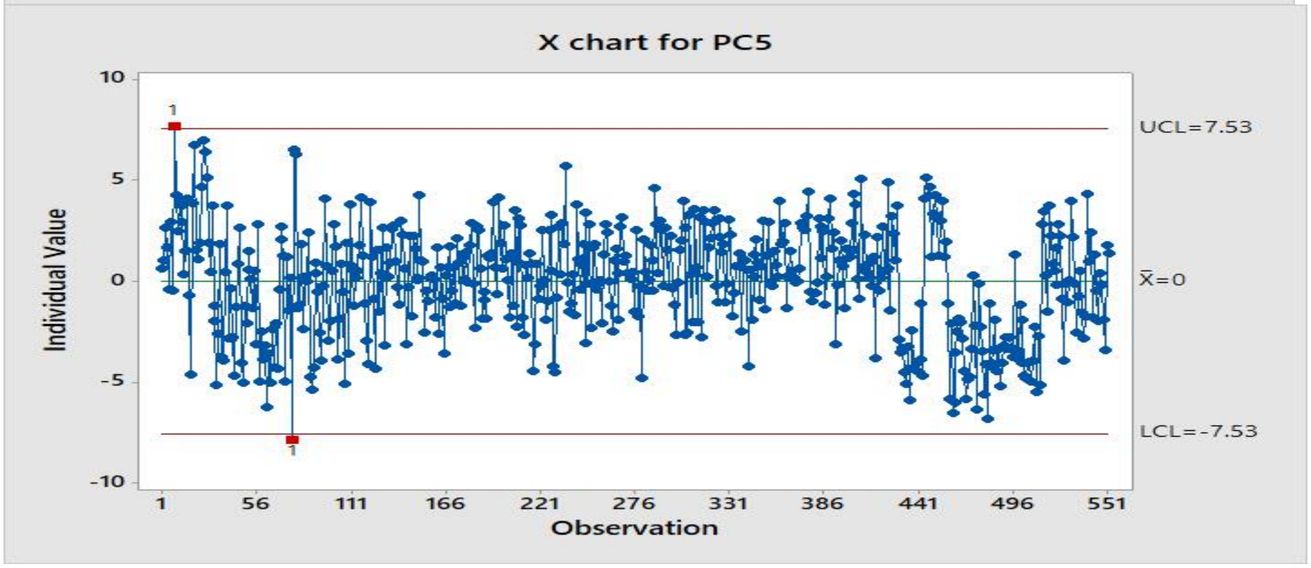


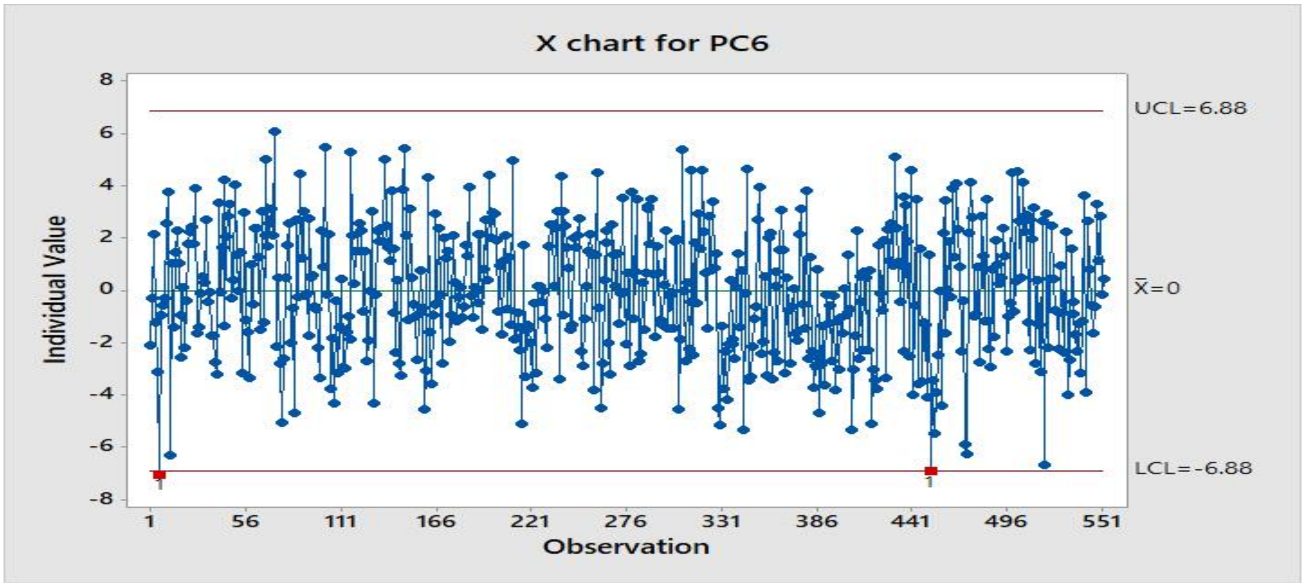
Fig.5A.

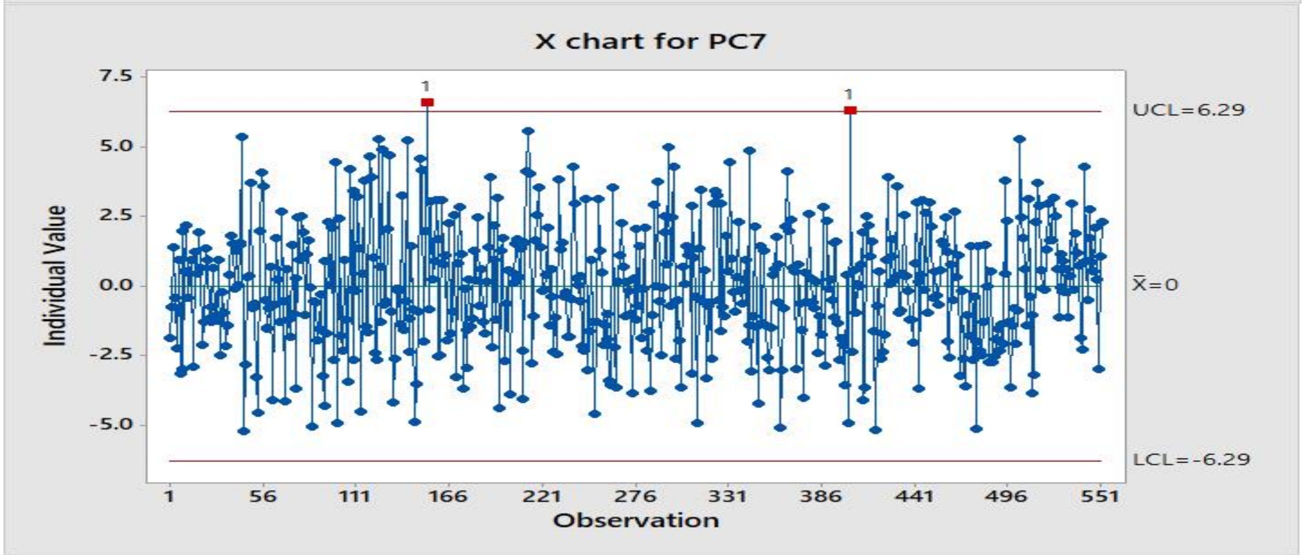


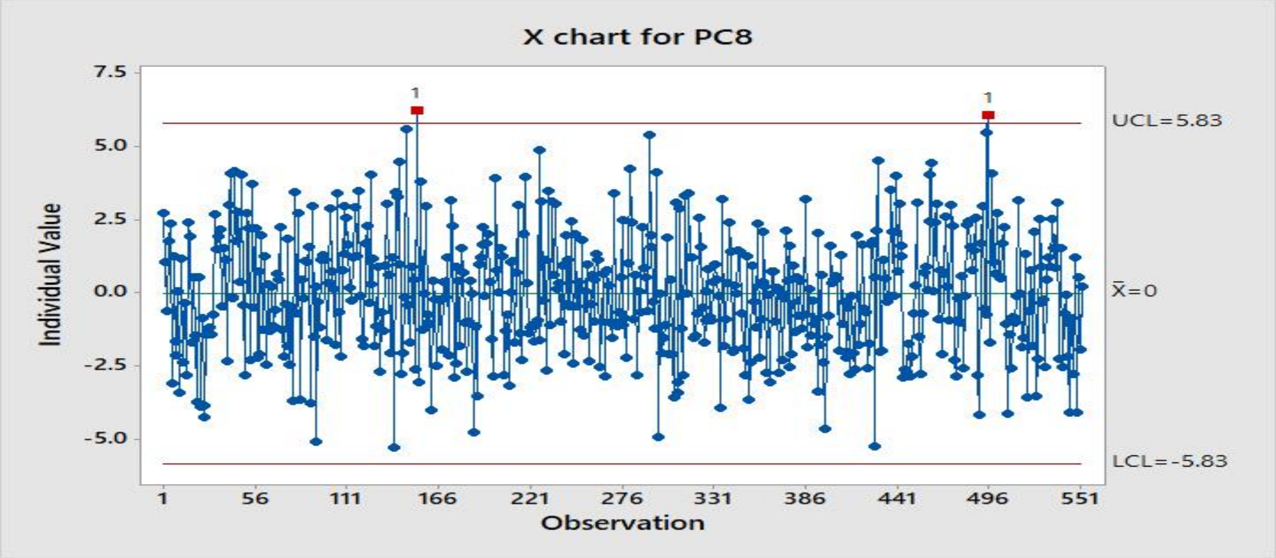

X chart for PC3
Individual Value
10
5
0
-5
-10
-15
UCL=11.73
X̄=0
LCL=-11.73
1
1
1
1
1
56
111
166
221
276
331
386
441
496
551
Observation


X chart for PC4
Individual Value
10
5
0
-5
-10
UCL=9.24
X̄=0
LCL=-9.24
1
1
56
111
166
221
276
331
386
441
496
551
Observation


X chart for PC5
Individual Value
10
5
0
-5
-10
UCL=7.53
X̄=0
LCL=-7.53
1
1
1
56
111
166
221
276
331
386
441
496
551
Observation


X chart for PC6
8
6
4
2
0
-2
-4
-6
-8
Individual Value
UCL=6.88
X̄=0
LCL=-6.88
1
1
1
56
111
166
221
276
331
386
441
496
551
Observation


X chart for PC7
7.5
5.0
2.5
0.0
-2.5
-5.0
Individual Value
1
1
UCL=6.29
X̄=0
LCL=-6.29
1
56
111
166
221
276
331
386
441
496
551
Observation


X chart for PC8
7.5
5.0
2.5
0.0
-2.5
-5.0
Individual Value
1
1
UCL=5.83
X̄=0
LCL=-5.83
1
56
111
166
221
276
331
386
441
496
551
Observation

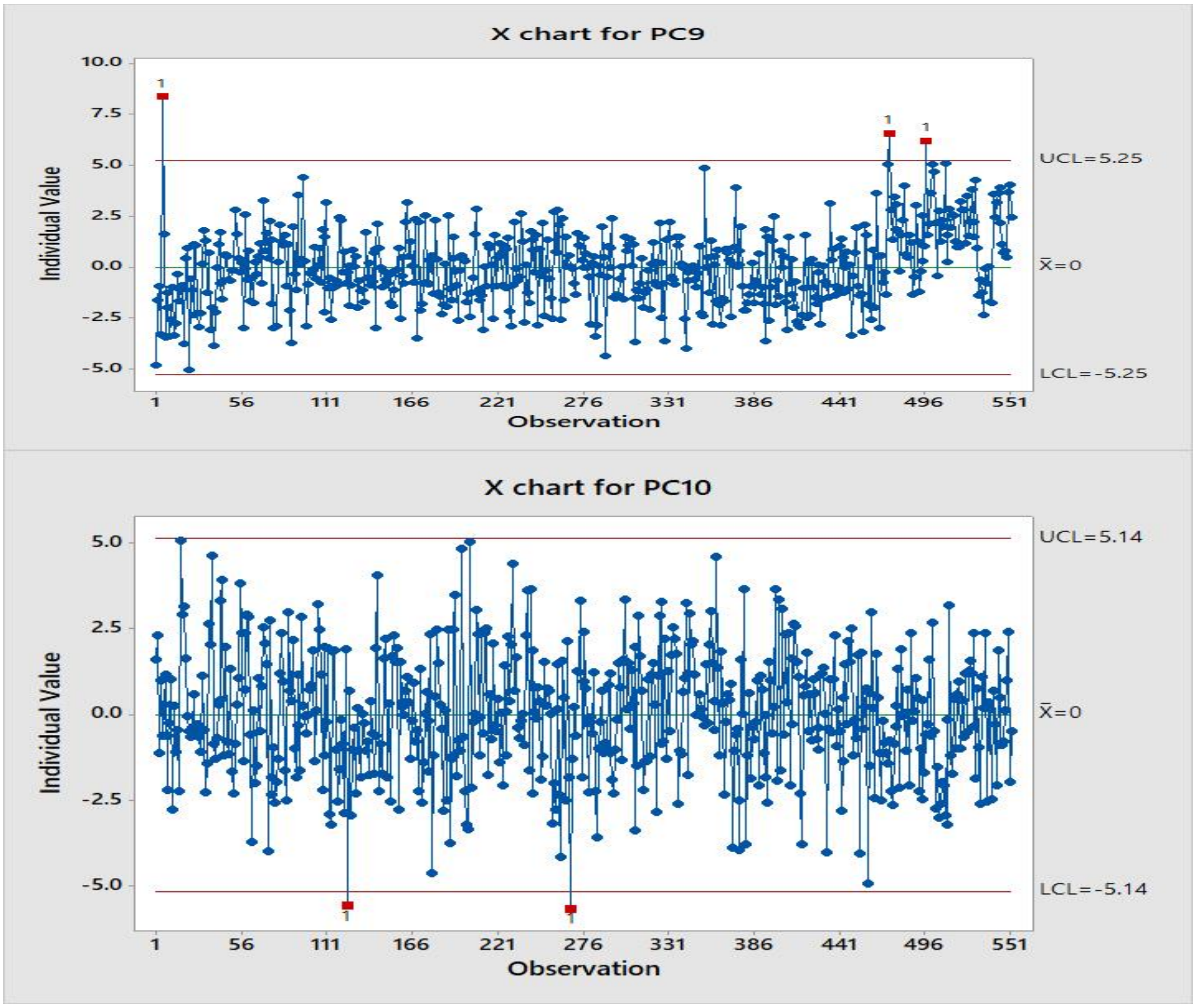
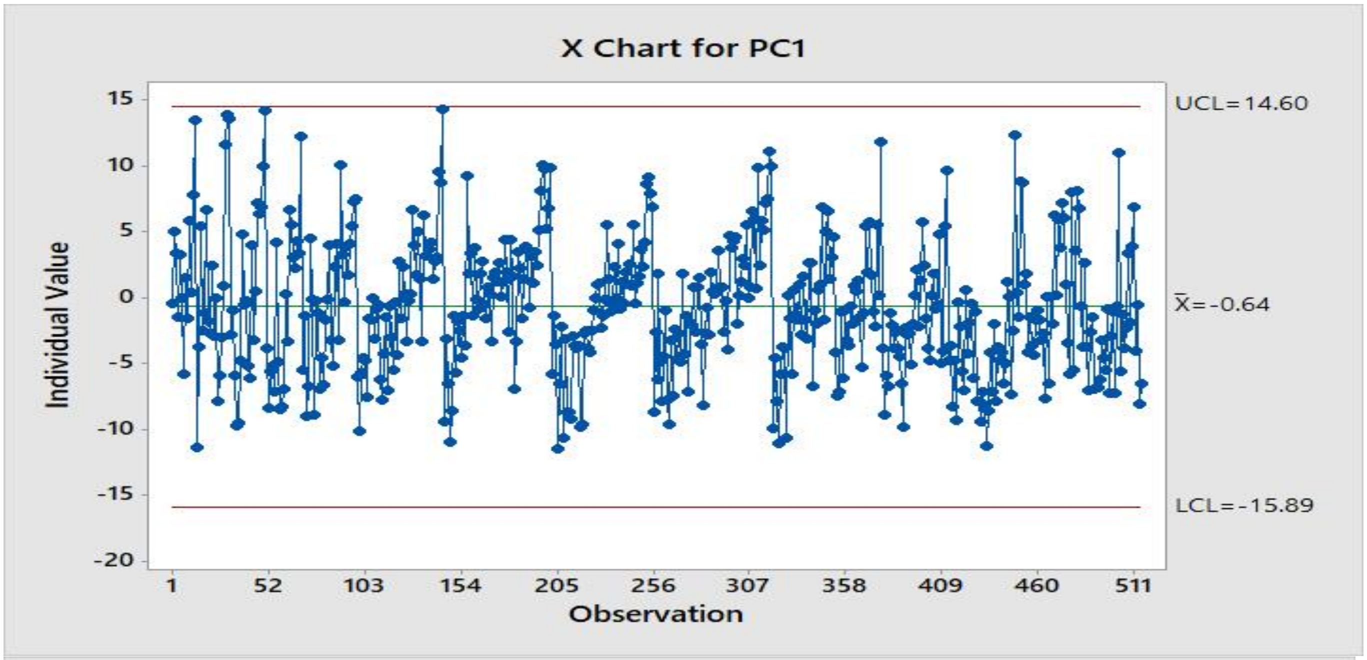
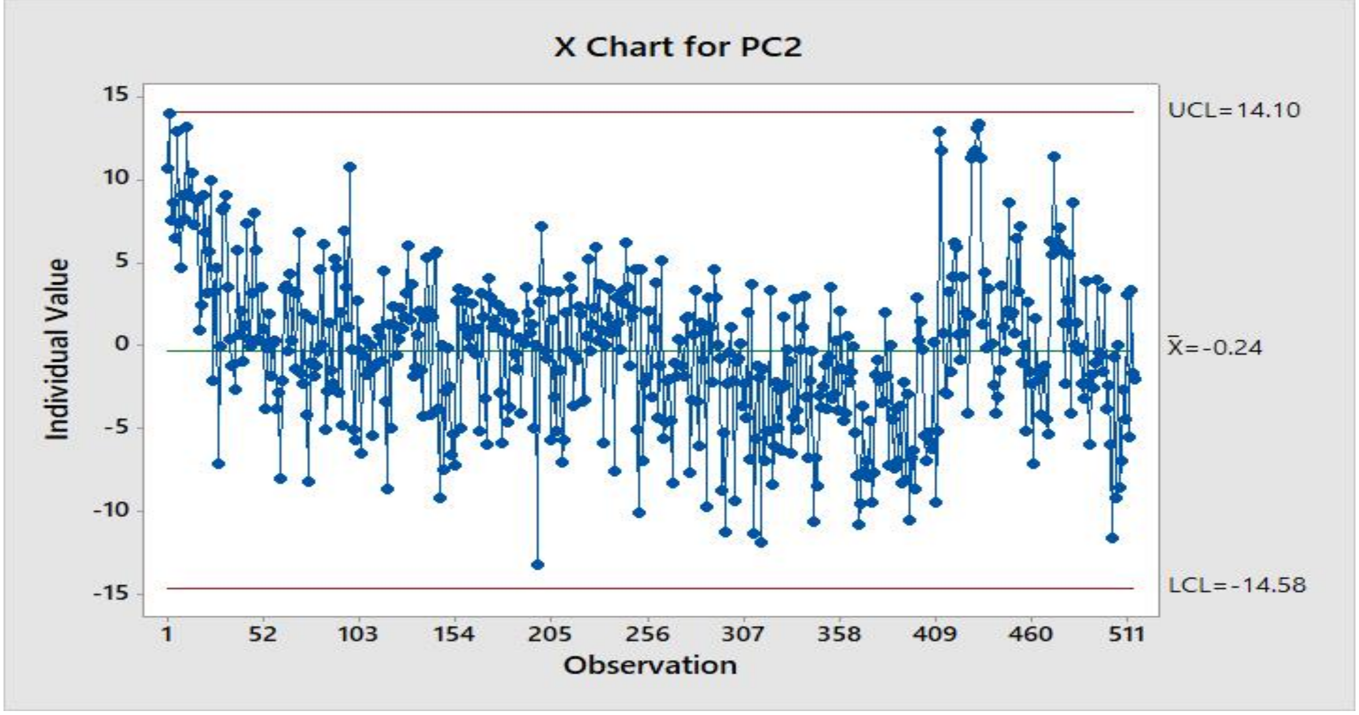
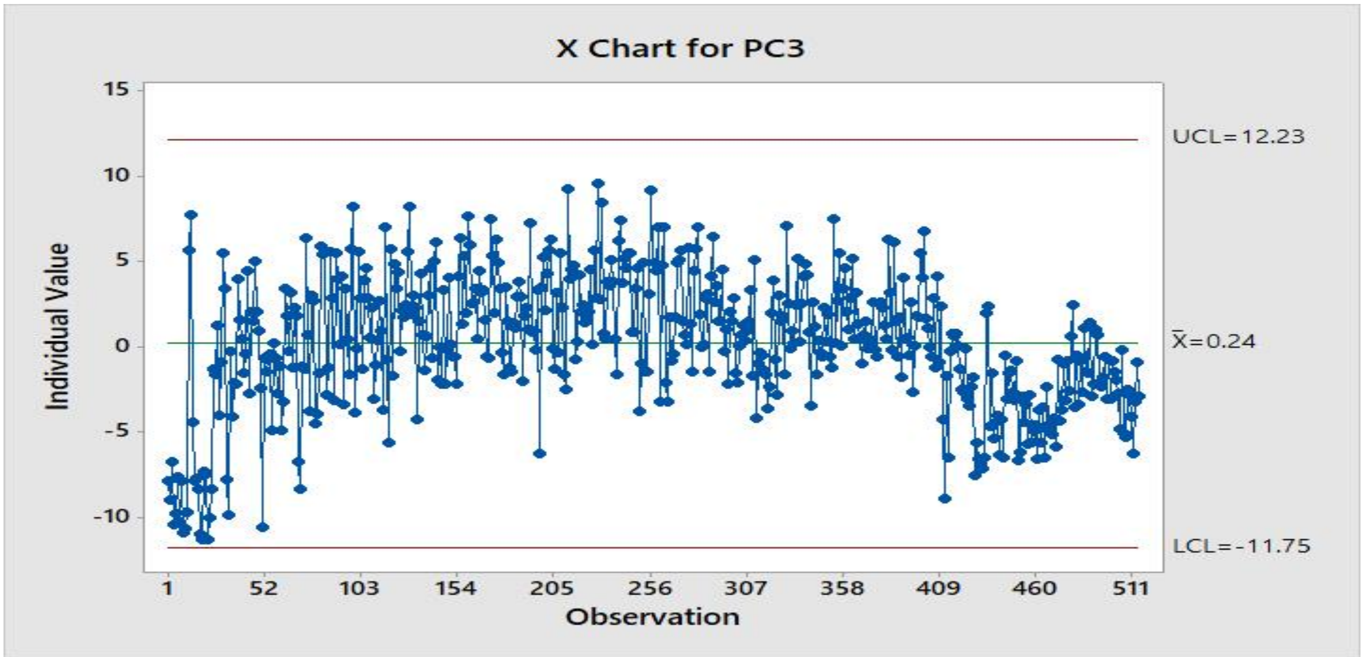

X chart for PC9
Individual Value
10.0
7.5
5.0
2.5
0.0
-2.5
-5.0
1
1
1
UCL=5.25
X̄=0
LCL=-5.25
1
56
111
166
221
276
331
386
441
496
551
Observation
X chart for PC10
Individual Value
5.0
2.5
0.0
-2.5
-5.0
UCL=5.14
X̄=0
LCL=-5.14
1
1
1
56
111
166
221
276
331
386
441
496
551
Observation

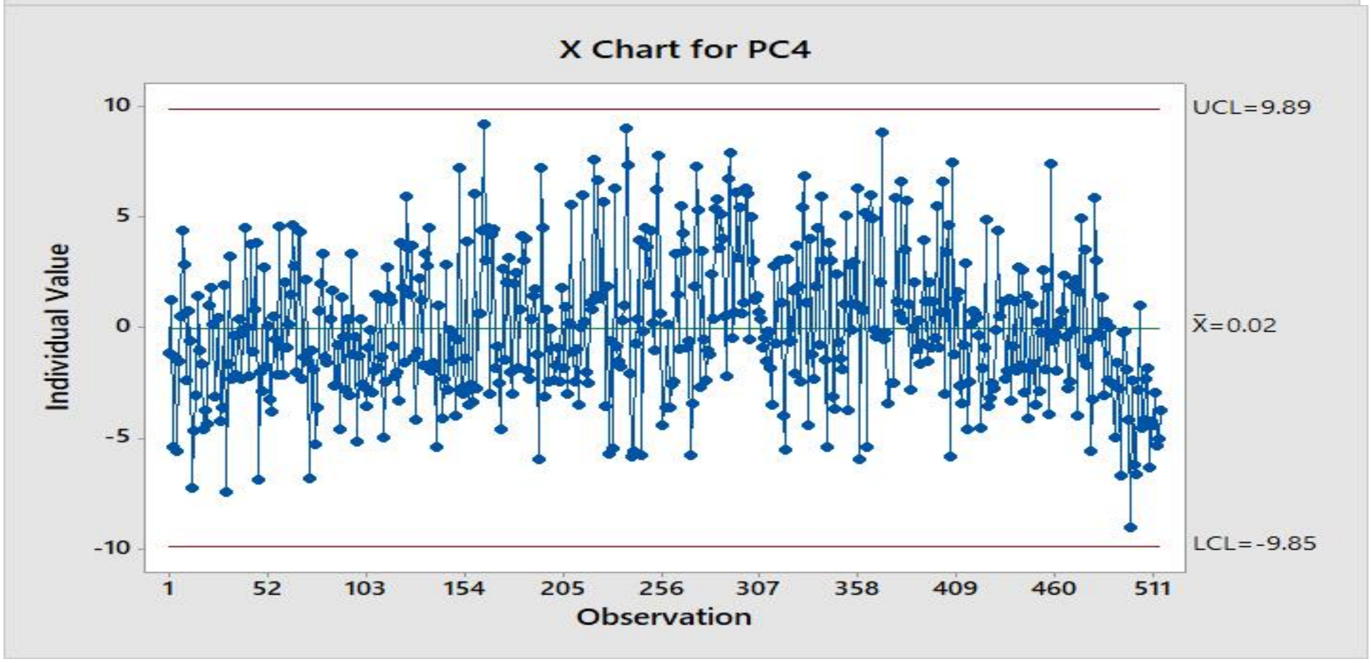
Fig 5A

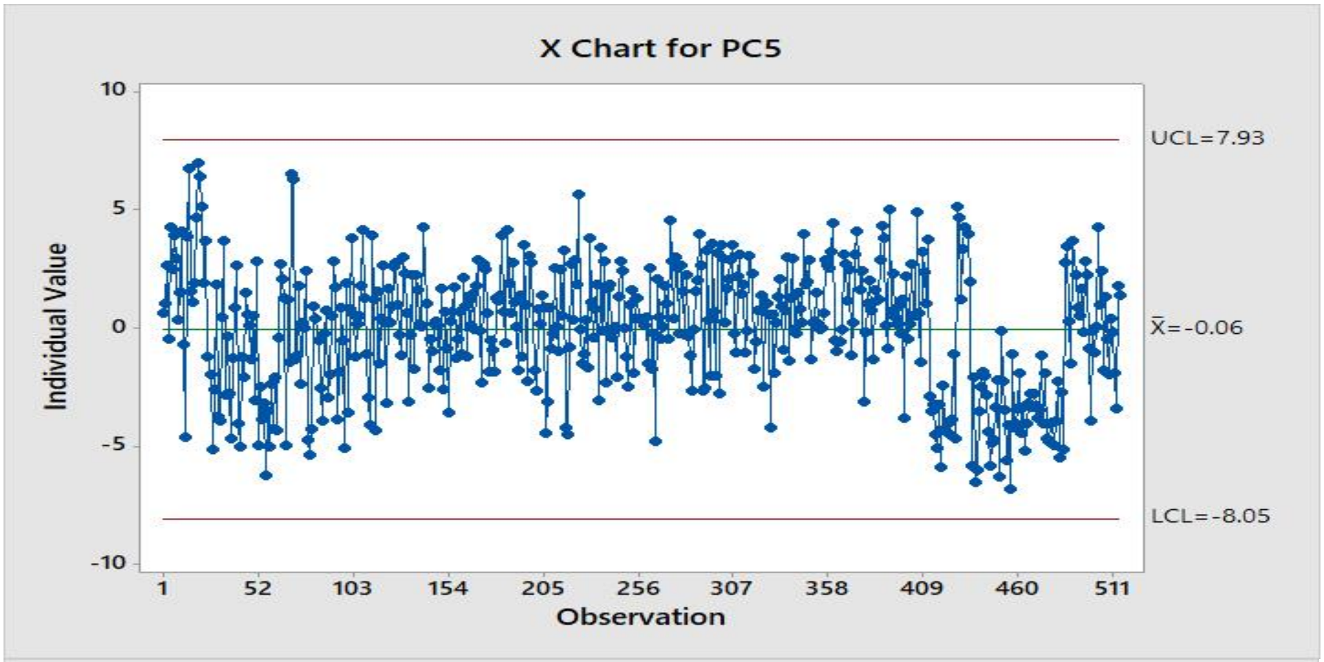
The final iteration(4th)of X chart on the first 10 PCs are as shown below:

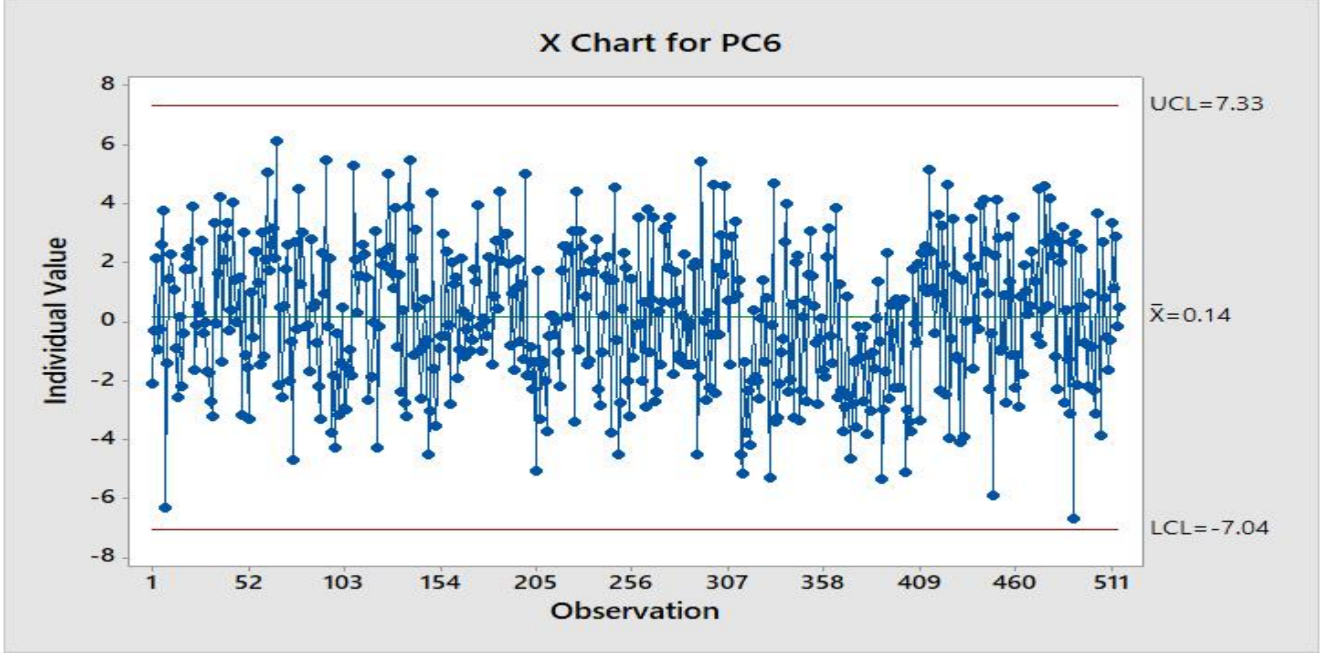


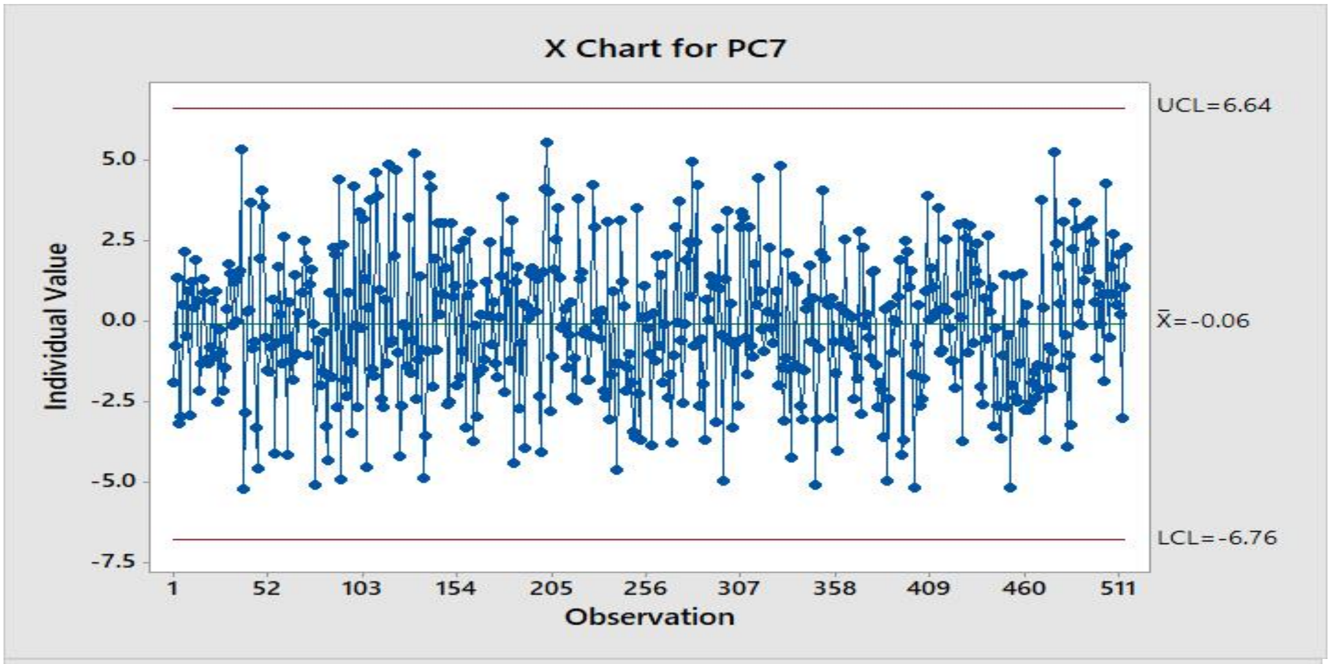


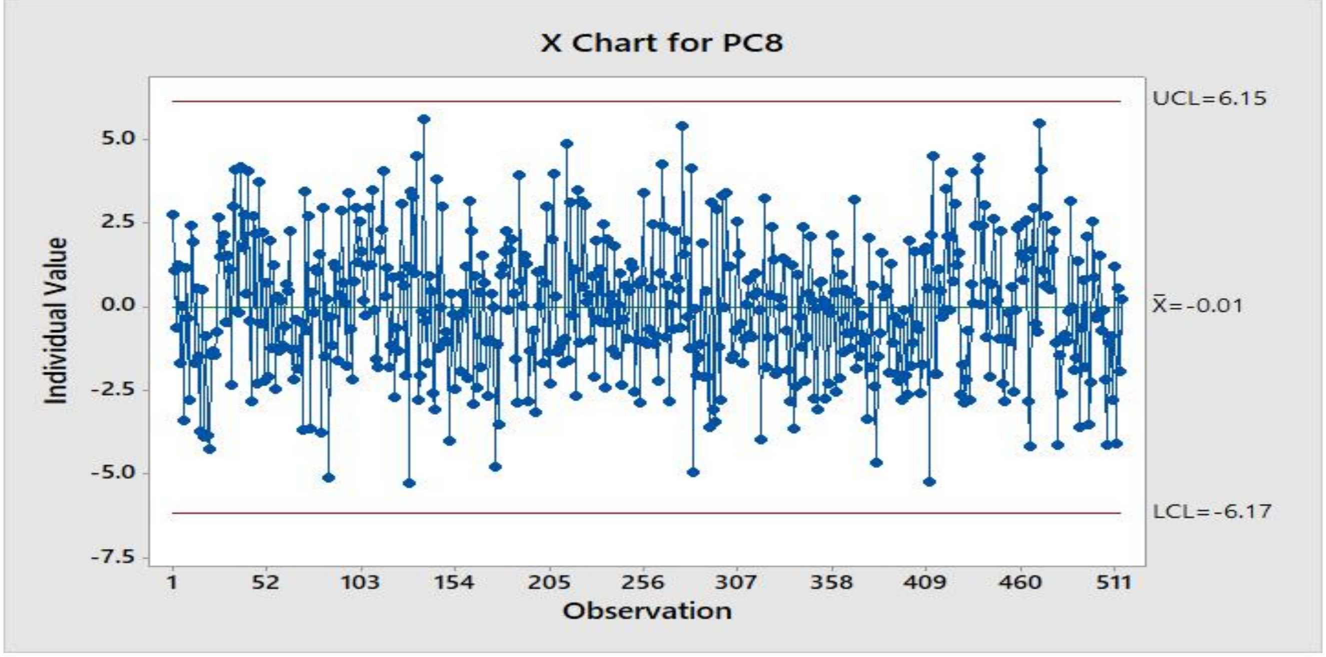

X Chart for PC3
15
10
5
0
-5
-10
Individual Value
UCL=12.23
X̄=0.24
LCL=-11.75
1
52
103
154
205
256
307
358
409
460
511
Observation

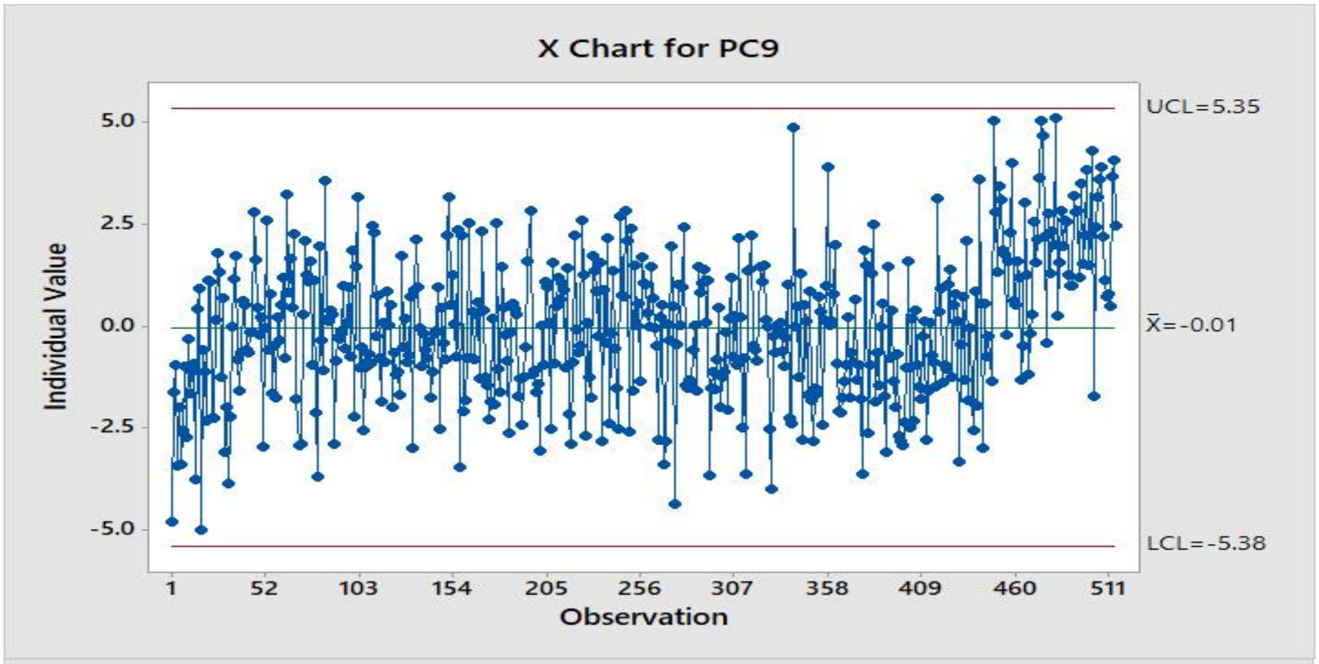

X Chart for PC4
10
5
0
-5
-10
Individual Value
UCL=9.89
X̄=0.02
LCL=-9.85
1
52
103
154
205
256
307
358
409
460
511
Observation

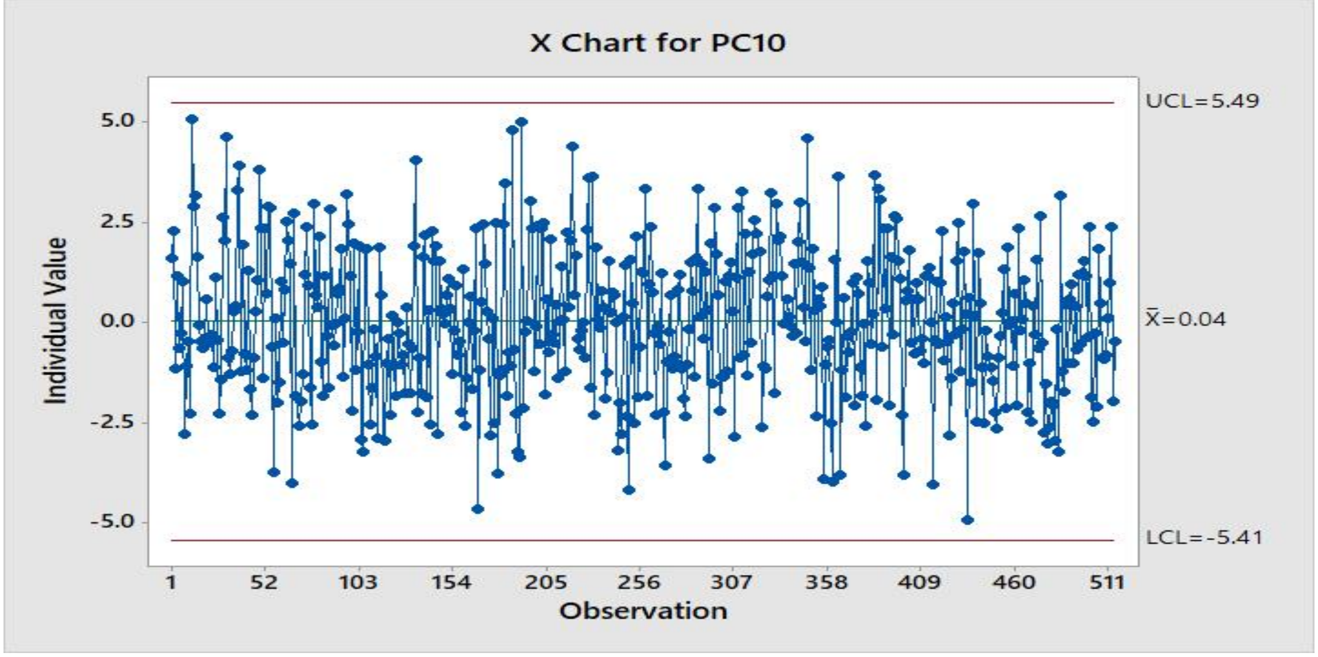

X Chart for PC5
Individual Value
10
5
0
-5
-10
UCL=7.93
X̄=-0.06
LCL=-8.05
1
52
103
154
205
256
307
358
409
460
511
Observation


X Chart for PC6
Individual Value
8
6
4
2
0
-2
-4
-6
-8
UCL=7.33
X̄=0.14
LCL=-7.04
1
52
103
154
205
256
307
358
409
460
511
Observation


X Chart for PC7
Individual Value
5.0
2.5
0.0
-2.5
-5.0
-7.5
UCL=6.64
X̄=-0.06
LCL=-6.76
1
52
103
154
205
256
307
358
409
460
511
Observation


X Chart for PC8
Individual Value
5.0
2.5
0.0
-2.5
-5.0
-7.5
UCL=6.15
X̄=-0.01
LCL=-6.17
1
52
103
154
205
256
307
358
409
460
511
Observation


X Chart for PC9
Individual Value
Observation
5.0
2.5
0.0
-2.5
-5.0
1
52
103
154
205
256
307
358
409
460
511
UCL=5.35
X̄=-0.01
LCL=-5.38


X Chart for PC10
Individual Value
Observation
5.0
2.5
0.0
-2.5
-5.0
1
52
103
154
205
256
307
358
409
460
511
UCL=5.49
X̄=0.04
LCL=-5.41

2.1.2 T^2 chart analysis on top 10 PCs

- We have performed multiple iterations using the T^2 chart analysis on the top 10 PCs until all the OOC points are discarded.
- While performing this analysis, we have treated the PCs as another set of multivariate data and performed Hotelling T^2 on it.
- UCL is $\chi^2_{1-}\alpha(10)$, where α is 0.05. UCL = 18.307
- The T^2 charts of the top 10 PCs are as shown below:

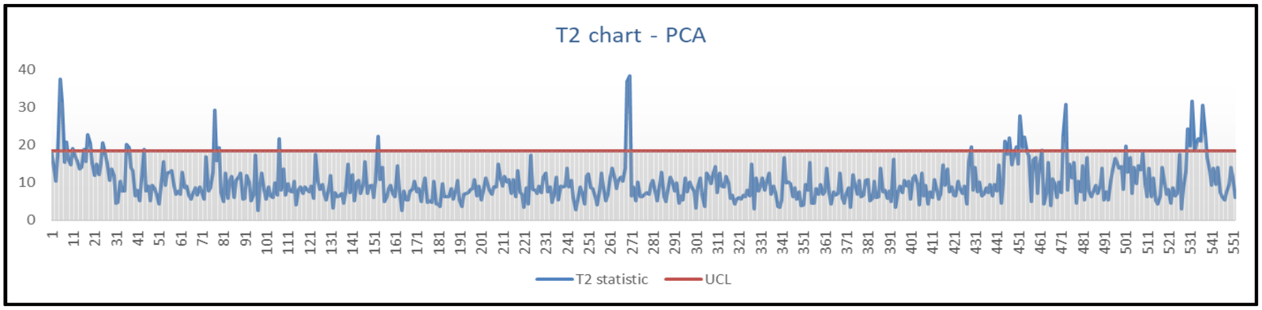


Fig.6A: Iteration1 - Hotelling T^2 Chart for all 552 observations. 38 observations are found to be above UCL.

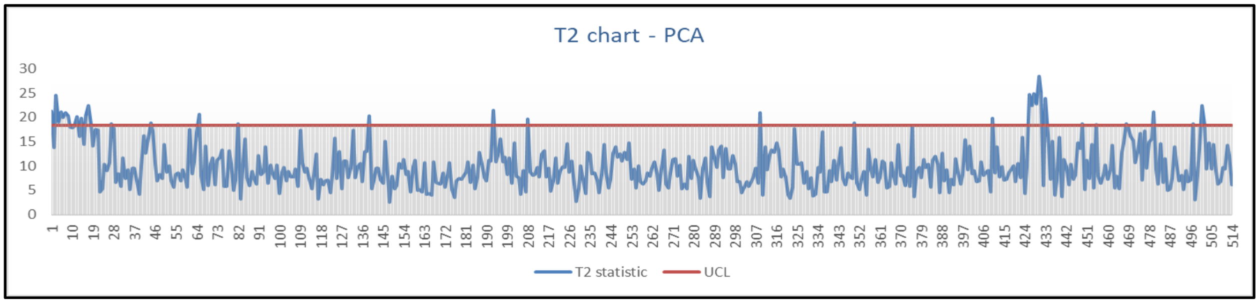


Fig.6B: Iteration2 - Hotelling T^2 Chart for all 514 observations. 36 observations are found to be above UCL

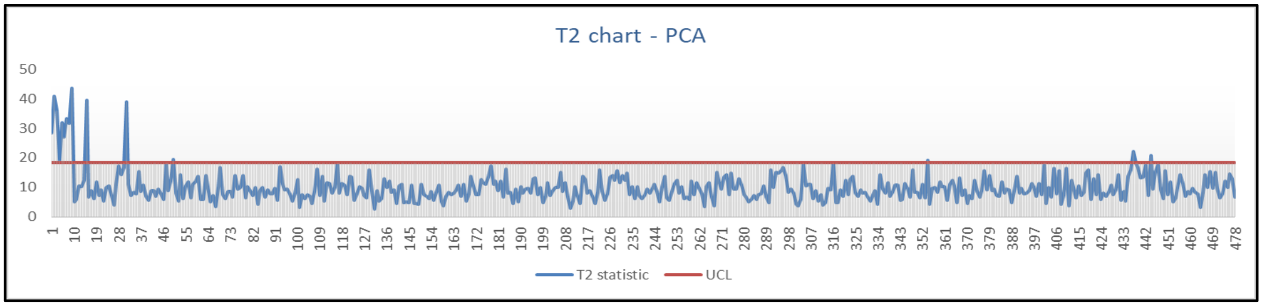


Fig.6C: Iteration3 - Hotelling T^2 Chart for all 478 observations. 16 observations are found to be above UCL

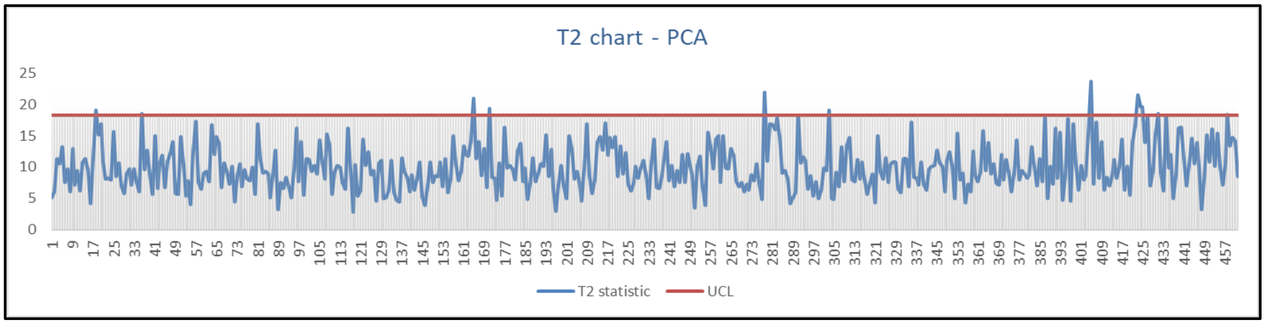


Fig.6D: Iteration4 - Hotelling T^2 Chart for all 462 observations. 12 observations are found to be above UCL

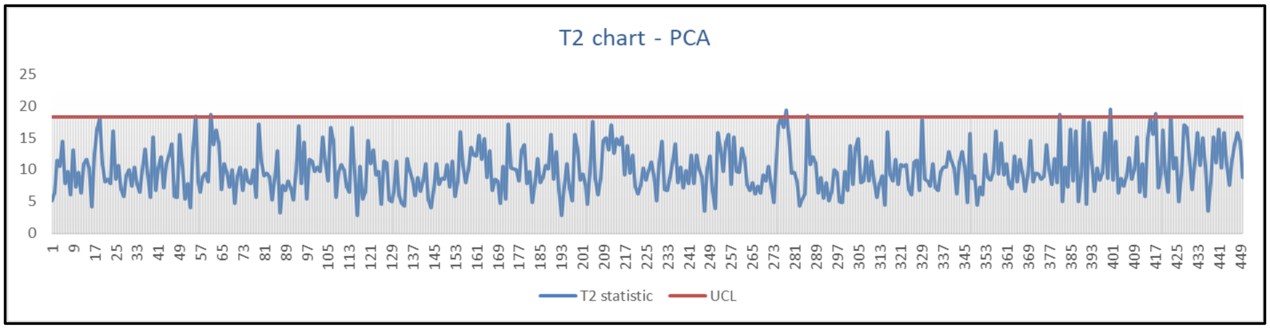


Fig.6E: Iteration5 - Hotelling T^2 Chart for all 450 observations. 7 observations are found to be above UCL

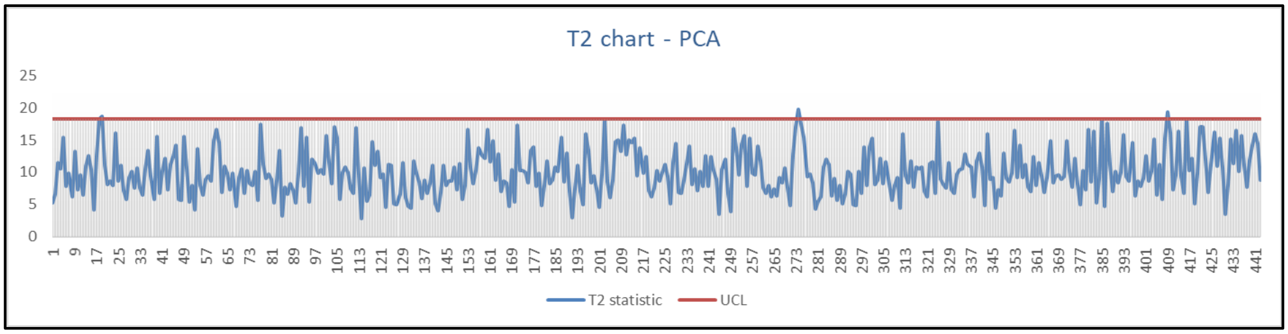


Fig.6F: Iteration6 - Hotelling T^2 Chart for all 443 observations. 3 observations are found to be above UCL

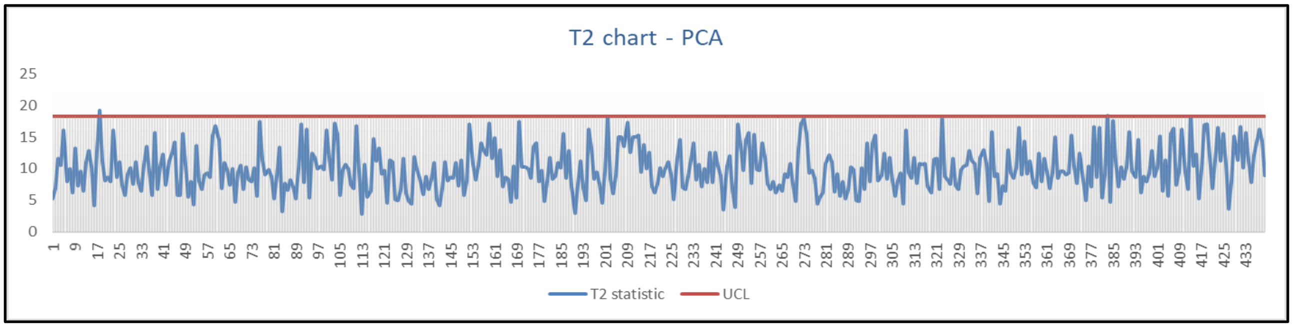


Fig.6G: Iteration7 - Hotelling T^2 Chart for all 440 observations. 2 observations are found to be above UCL

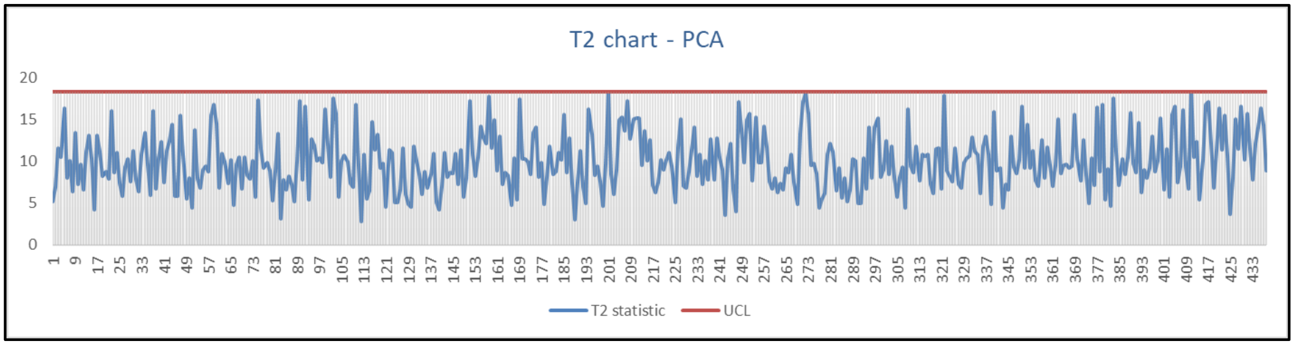


Fig.6H: Iteration8 - Hotelling T^2 Chart for all 438 observations. All observations are under control

2.2 Justification

Principal Component Analysis approach helps our analysis in the following ways:

- Dimension Reduction – Our no. of attributes has been reduced from 209 to 10 which makes our analysis our analysis less tedious and more meaningful..
- Uncorrelated variables- PCA transforms the variables to perpendicular directions and hence removing any correlation among variables, if present.
- More reliable detection – Since the dimension has been reduced, the noise will not add up to large values leading to more reliable detections.

2.3 Results

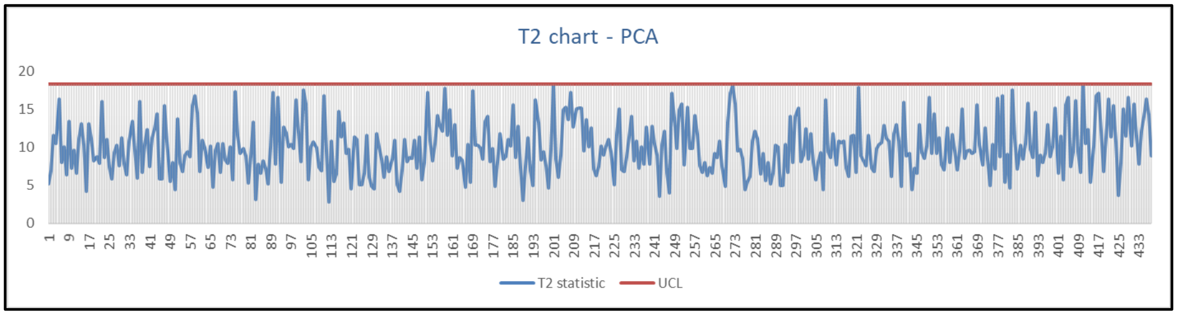
1. Using T2 Hotelling charts we removed 114 observations points from the data.

Iteration	Total No. of observations	No. of OOC observations
1	552	38
2	514	36
3	478	16
4	462	12
5	450	7
6	443	3
7	440	2
8	438	0

2.4 Conclusions

- We find that multivariate analysis using T^2 chart after PCA gives us more number of out-of-control points as compared to T^2 chart applied to original data. This is due to reduced effect of noise caused by reduction in number of variables.
- One more insight that we gained from the study is that increasing the number of PCs we had to reduce the individual α level to approx. 0.05/n, where n = number of PCs.
  This reduction in α led to increase in UCL and hence lesser number of points were detected from each chart when we applied the criterion of reject the observation if either of the chart signals. This established a better understanding of the 'curse of dimensionality' with increasing number of variables in case of univariate analysis of multiple variables.
- We accept PCA approach as a more reliable method as it takes care of correlation and the aggregate effect of noise. The final results of the PCA approach can be used as standard parameters determined through phase-I analysis and used for future missions.
- Finally, the in-control parameters for T^2 chart after PCA are -

$$\mu_0 = [\ 0.48 \quad -9.9 \quad -13.6 \quad - - - - - - - \quad -29.9 \quad -28.6 \quad -25.6\ ]^T$$



References

1. Md. Shamim Reza, Sabba Ruh, Science Journal of Applied Mathematics and Statistics Volume 3, Issue 4, August 2015, Pages: 171-176.
2. Analysis Of Accuracy Multivariate Control Chart T2 Hotelling Free Distribution With Outlier Removal, Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management Detroit, USA, September 23-25, 2016