ISEN-614
Final Project Report

Industrial and Systems Engineering

TEXAS A&M UNIVERSITY

Team Members

George Paul 627003548

Muhammad Hussain Anwaar 827008798

Jeremiah Prakash 627005933

Yash Mehta 228002156

Executive Summary

The project dataset excel file contained 209 columns corresponding to the number of predictor variables (p) as well as 552 rows corresponding to the number of samples (m). Here sample size is 1 (n). Due to large dimensions, the noise components can add up and overwhelm the signal effects in the dataset, thereby make it harder to reject the null hypothesis. Thus, we resorted to the data reduction technique to derive the "vital few" that matters out of the data. We employed the Principal component Analysis as the tool for Dimensionality Reduction task. Which generated the low-dimensional data by projecting original data into vital few dimensions that explained most of the variance. These dimensions have zero correlation, thus we implemented both Hotelling T^2 and multiple univariate charts for Phase one Analysis. The final step was to iteratively remove the out of control points from the data and decide the in-control parameters for future detection purposes.

Data Analysis

The dataset was analyzed to explore the distribution of different variables and their correlations with each other. Where, we discovered that all the predictor variables follow a normal distribution, thus it would be safe to model the data using multivariate normal distribution. Secondly, we generated the correlation matrix that depicted high correlations within the feature. Thereby, posing a dire need to reduce the data dimensions for further analysis.

Methodology

In order to cater the effect of aggregated noise due to Curse of Dimensionality and high number of correlated features, we used the Principal Component Analysis for dimensionality reduction. This is because with high dimensions, it is harder to systematically search and signal to noise ratio is very small thereby reducing the detection power. After projecting the data into vital few dimensions, we determined the number of Principal components by using a Scree plot (looking for the elbow bend) and a Pareto Plot (observed at 80%). After data reduction we employed T^2 Chart to determine the in control and out of control points by setting the Upper Control Limit using Chi-Square statistics.

Analysis and Conclusion

Based on the Scree and Pareto plot, we decided to use four Principal Components which would account for about 80% of the total variance in the data. We used the Hotelling T^2 chart statistics to determine the OOC points by setting an upper control limit determined by Chi-square statistic. We considered several candidate values (0.0027, 0.05 and 0.01) for alpha level and proceeded with alpha=0.05. The UCL for the corresponding alpha level of 0.05 is 9.44.

Using the above-mentioned parameters, we set the T^2 chart and iteratively removed out of control points. It took 8 iterations to remove all of the 141 out of control points from a total sample of 552. Using the remaining values, we were then able to determine the in control mean $\mu 0$ and the covariance matrix $\Sigma 0$. We also used the 4 individual x charts for preliminary analysis of the data. To determine the source of the OOC signal we use the Hotelling T^2 chart.

Concepts used

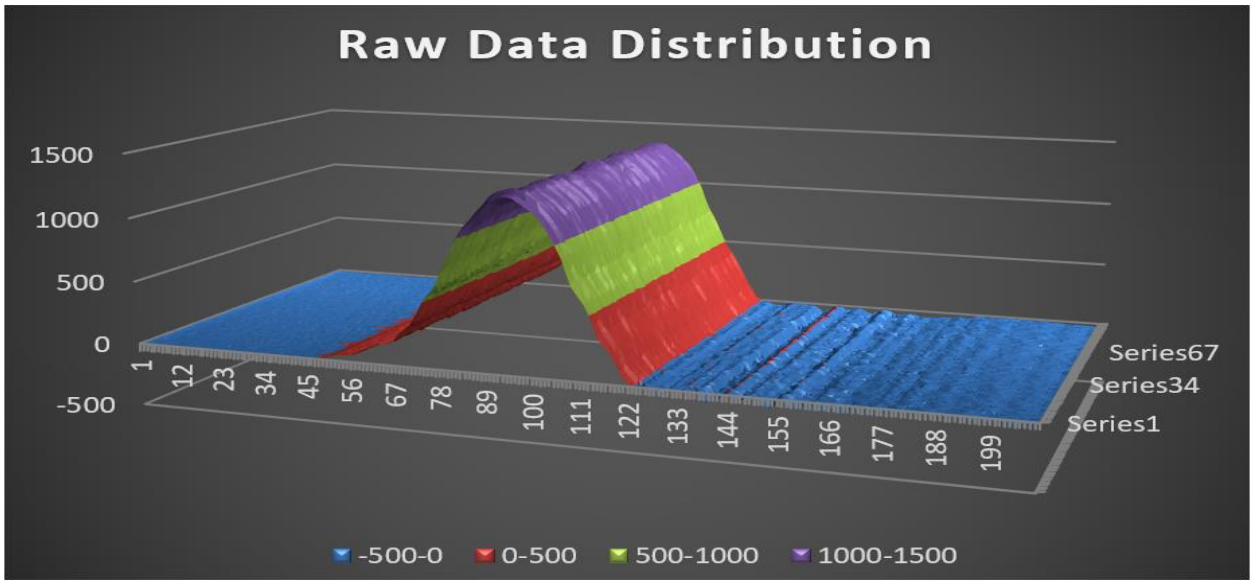
1) Principal Component Analysis (PCA) for dimensionality reduction.
2) Individual X chart and hotelling T^2 chart to determine OOC points.
3) Removing OOC points through iteration to determine IC parameters.
4) Application of CUSUM chart through Monte Carlo simulation to test the In-Control Parameters for future monitoring (Phase-II analysis).

Introduction

In a manufacturing setting it is important to deliver the products with high quality. In this project, we worked on a real time data generated from a system/machine. There will be variations in the product dimensions due to inconsistencies from human and machine error. Where, our aim is to set a control system to detect these variations. We used PCA, univariate and Hotelling T^2 charts and performed Phase-1 analysis in order to generate the in-control conditions for our detection process.

Approach

We did the initial data exploration in excel. We found that data is normally distributed for all 209 predictors. The correlation matrix suggests that the variables are highly correlated. We analyzed that due to these high correlations among high dimensions, the data could be projected in fewer uncorrelated dimensions which would explain most of the variance in the data. Various methods have been tried and tested such as multiple univariate charts and Hotelling T^2 chart for different alpha levels in order to adjust the ARL0 values which will in turn give us the best results for phase 1 analysis.



Principal Component Analysis

It is difficult to analyze or perform any statistical procedure on a large data set with 209 variables and 552 observation for each series. Hence, we adopted principal component analysis as the dimensionality reduction technique, in order to reduce the number of variables to be considered for our analysis. As stated earlier, PCA is performed to derive few dimensions that explains most of the variance. PCA can be performed on both the covariance matrix and the correlation matrix. But since there is enough evidence from the initial analysis of raw data that there are significant correlation between successive variables we can move forward in our principal component analysis considering the covariance matrix as the difference between the correlation matrix and covariance analysis is normalization of its elements.

Once the Principal component analysis was performed we used a combination of scree plot and pareto chart to identify the principal components that would help us gain us most of the information by explaining most of the variance within the data.

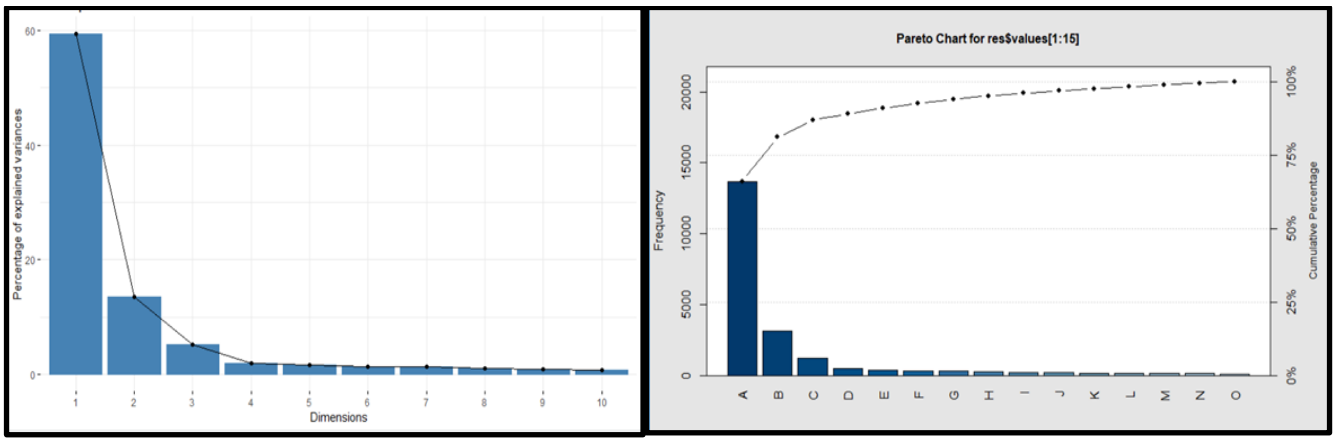


Figure 1: Scree Plot and pareto charts of the principal components. We consider 4 PC's to explain 80% of variation.

Analysis

We considered the threshold of 80% of the total variance explained. Thus, for pareto chart we drew the line on 80% of the total variance to get the correct number of corresponding PC's. Whereas, in the Scree Plot, we employed the elbow bend technique to figure out the number of PC's. Then, we compared the output from both the charts to select the number of Principal Components to be used.

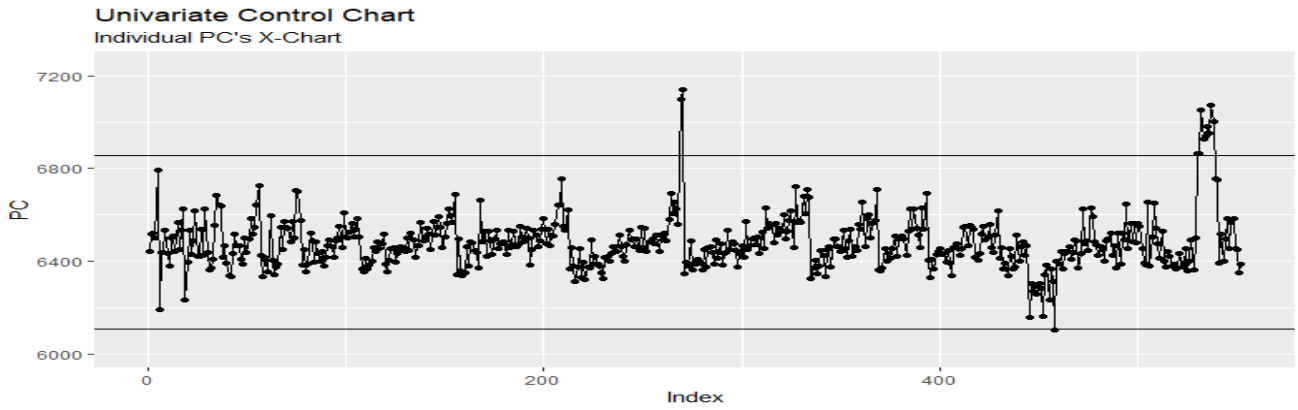
From the Figure 1 we can see that the elbow bends at around the third or fourth principal component. Whereas, in Pareto Chart the 80% mark gives us 4 PC's. Thus, we went ahead with four principal components and regenerated the dataset of 552 observations with 4 Principal Components as predictor variables.

Univariate Control Chart Approach:

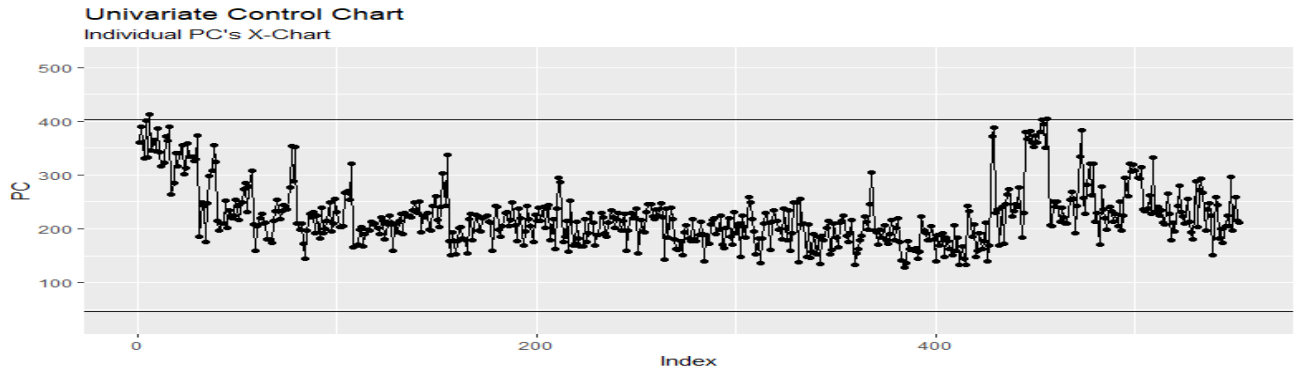
Since the regenerated data is uncorrelated among its features, we can use multiple univariate control charts to detect the out of control points. We used the composite rule that "If any of the chart signals, the whole control system would signal". We decided earlier to use the alpha=0.05 for our analysis w.r.t T^2 chart and univariate ones. We adjusted the individual chart alpha level such that the combined alpha would be 0.05. Thus, at individual alpha=0.0127, we get the total alpha=0.05. So, for individual chart L=2.5 corresponding to the individual alpha level.

Firstly, we used X-Chart to plot all the principal components to visualize the initial condition of the process. The graphs below give an overview that there are many out of control points in all four principal component charts.

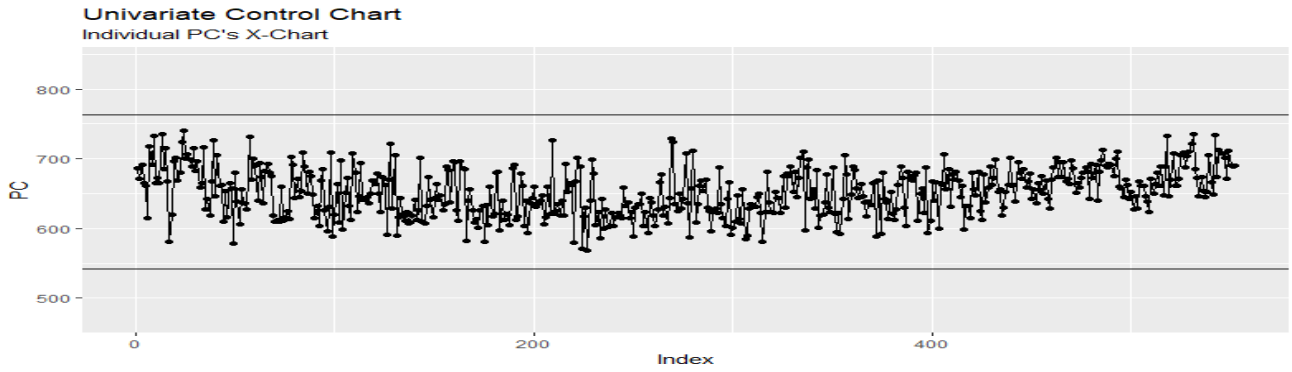
PC1:



PC2:



PC3:



PC4:

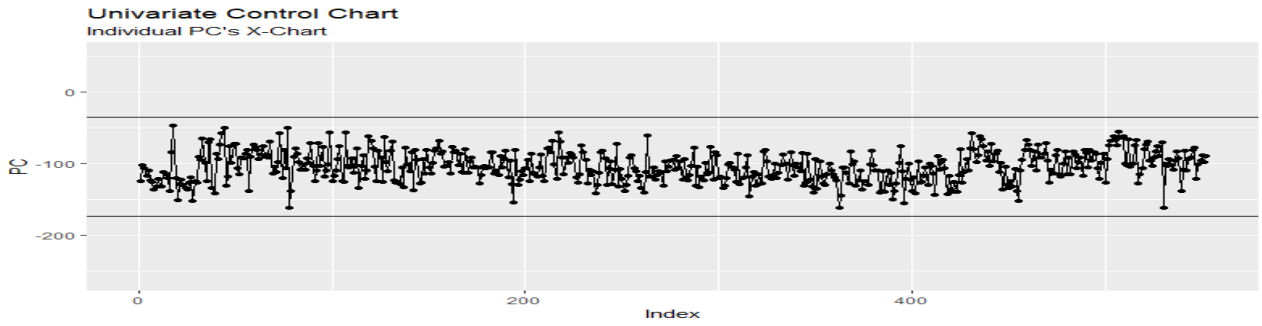
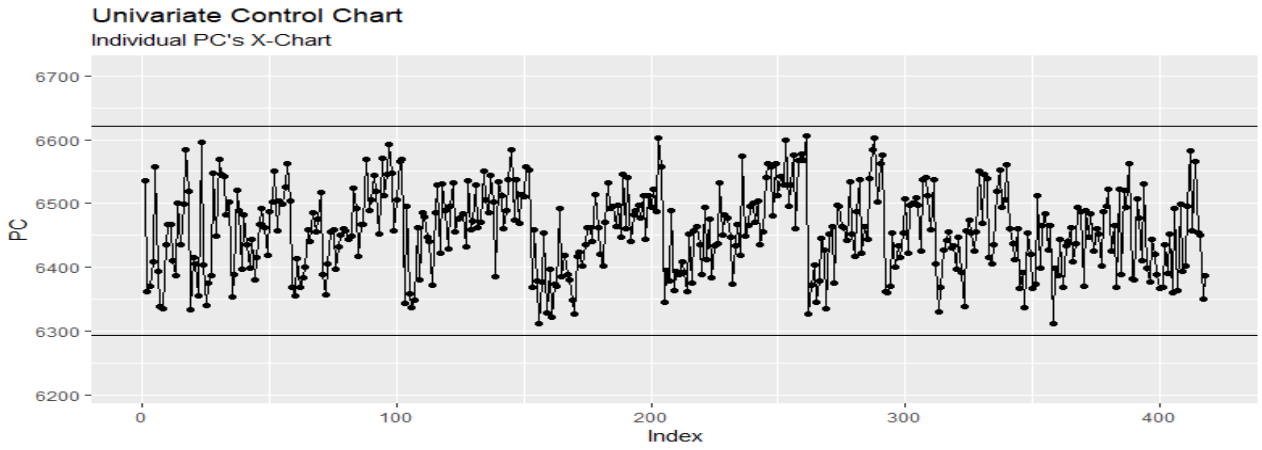


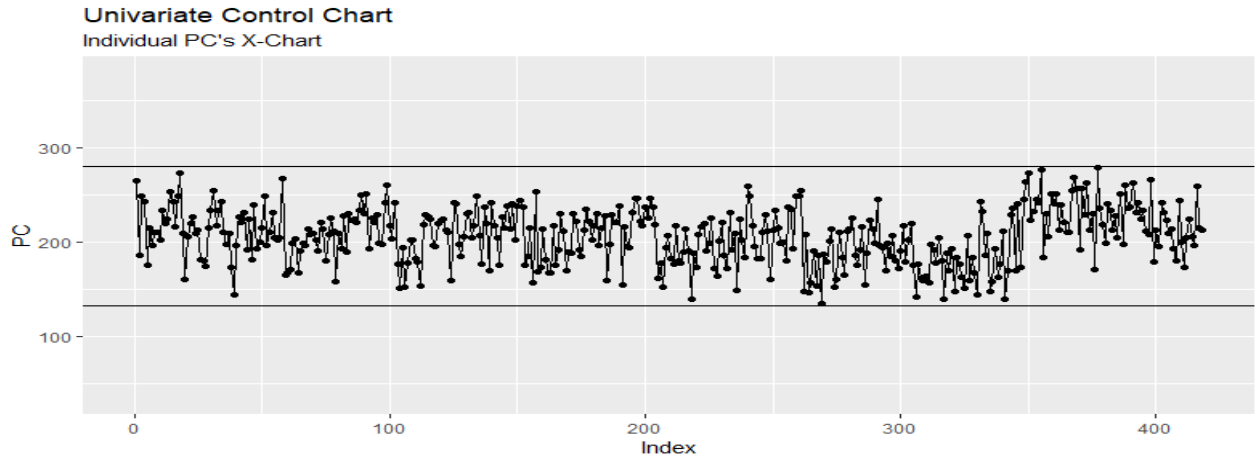
Figure 2: Univariate approach results for each principal component analysis. Here the results of each PC is set to a separate control limits.

After that we performed the iterative procedure of removing the out of control points, it took 16 iterations to remove all the out of control points. In these 16 iterations we removed 128 points. The plots below depict the final in control data w.r.t all the principal components.

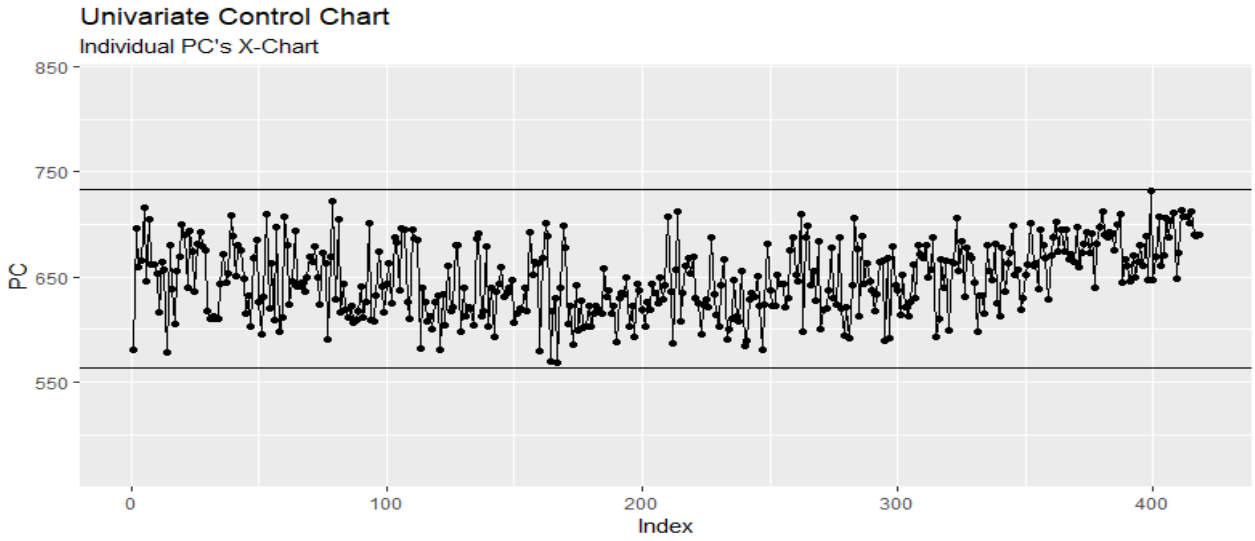
PC1:



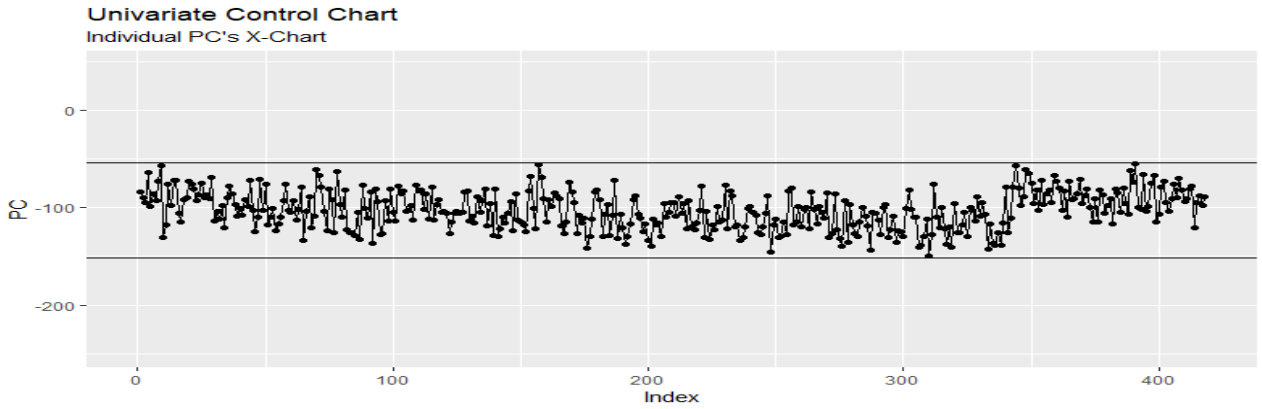
PC2:



PC3:



PC4:



Hotelling T^2 Chart

Since this is a Phase-I analysis with n=1, we used chi-squared distribution with alpha set to 0.05, with the corresponding UCL at 9.44. We used this UCL to determine the OOC points and eliminate them from the dataset. Once the OOC points are removed, the T^2 statistic is re-calculated for all the remaining data points as the mean and covariance changes after each iteration. This process is repeated until no points fall outside the control limit. Figure 3 is the plot obtained for the raw data, we can clearly see that a lot of points fall outside the UCL. Thus, we removed these points by performing multiple iterations until we obtained a chart where no points lie above the UCL.

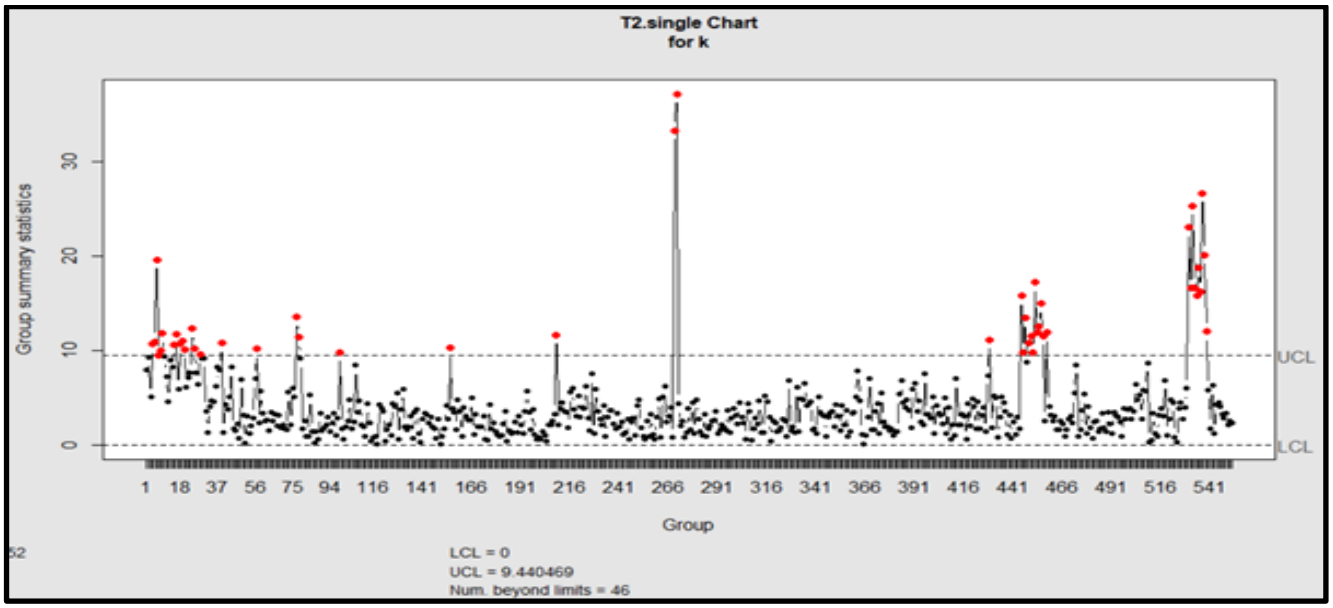


FIGURE 3: T2 plot obtained for the dataset using four PC's at UCL set at 9.44.

Figure 4 is the plot obtained after iterating the process 8 times. Here we can see that all the data points fall below the UCL and hence we can state that all OOC points have been removed. Therefore, this data can now be used to estimate in-control mean and covariance matrix for analysis on future observation. We observed 141 points that were out of control in all of the iterations combined. Thus, these points were removed from the training dataset. From this in-control data we estimated the in-control parameters that can be used for Phase-II Analysis.

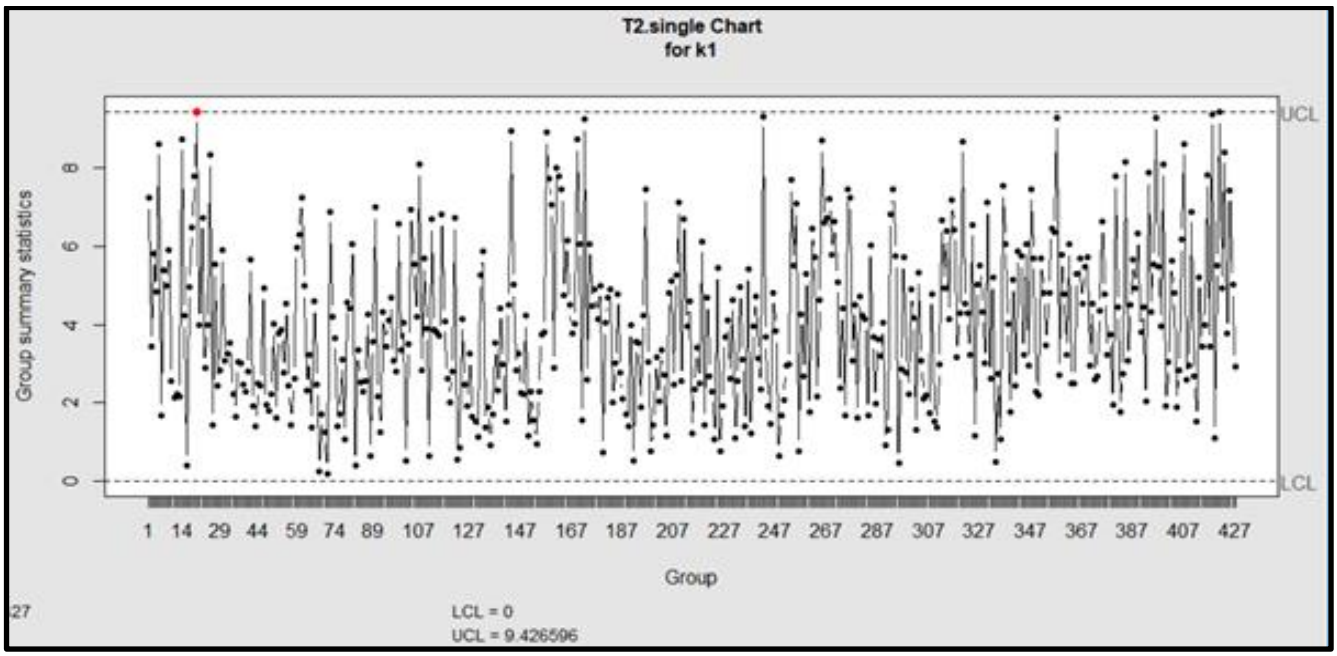


FIGURE 4: T2 plot obtained for the dataset using four PC's at UCL set at 9.42. Here no OOC points are seen.

Comparison of the two approaches:

The In-Control parameters can be estimated from the training data set from Phase-I Analysis.

Where, after performing two separate analysis we had two different in-control parameters. The in-control mean from T^2 chart is "6463.2617 206.6251 647.5817 -102.7600". Whereas, the mean from univariate method is "6458.8653 207.0119 648.0444 -102.7712".

Here, we analyzed that the difference between the in-control mean from two different approaches is pretty small. The T^2 chart eliminated all the out of control points in 8 iterations, whereas the univariate method took 16 iterations to do the same. Thus, we concluded that the former is a less computationally expensive method for performing Phase-I Analysis on the dataset provided. Which is the reason we went ahead with the in-control parameters estimated from the T^2 chart method.

Future Observation Monitoring

After finding the in control mean and covariance matrix, we simulated data using these parameters to find out what the ARL0 and the ARL1 would be for our data. For finding the ARL0, we used the in-control parameters to simulate the data points and find the RL. This was repeated to get enough data, to get the ARL0. Similarly, to detect the ARL1 we increased the mean by 5% to assess the detection capability in the event of a small mean shift. The results obtained are as follows:

For alpha = 0.05 and UCL = 6.06

ARL0 = 215, ARL1 = 4

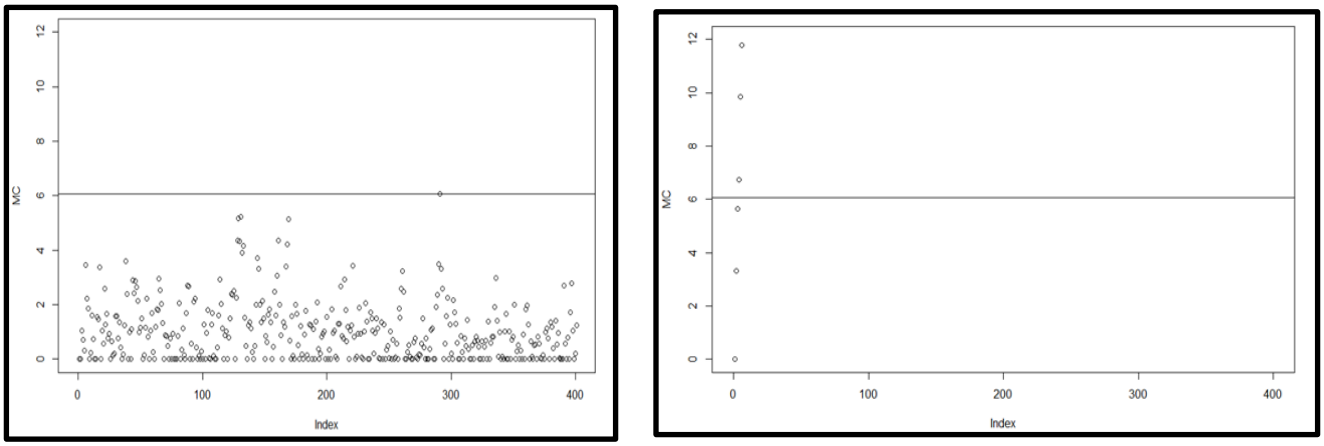


Figure 5: CUSUM charts. Simulated data with IC parameters (left). Simulated data with 5% mean shift (right)

Conclusion

In this project we have learned how to apply and set control chart limits to real life manufacturing data. Dimensionality reduction is a handy method to initiate the multivariable problem, as it helps to retain all the vital information without actually changing the actual meaning of the process. Tried different iterations and methods for the data to achieve the best ARL0 value for phase one analysis. We realized that Hotelling T^2 chart will help us set the training data by removing the out of control points for phase 1 analysis. While, univariate control charts will help to set controls it's difficult to compare and monitor multiple graphs at the same time when the features dimension is large. There is inflation in ARL values which makes it difficult to detect. We have also done a M-CUSUM analysis to ensure the results obtained are comparable and acceptable.