ISEN 614 Project

Team – 6

Pooja Phadke (827006065)
Rohit Sonje (428000331)
Himanshu Gupta (426007282)
Naisarg Shah (927008715)

Table of Contents

Executive Summary

Quality control is essential to build a successful business that delivers products which meet customers' expectations. It also forms the basis of an efficient business that minimizes waste and operates at high levels of productivity and for this, it is essential to examine if the product output is in control or out of control.

The given data set has sample size equal to 1, observations equal to 552 and 209 attributes. The in-control data will be used to carry out Phase 1 analysis to find out the estimates of population mean and covariance matrix. Phase 1 analysis involves implementing control charts to the data, remove out-of-control points and then implement the chart to the remaining points. This procedure is reiterated till only in control point remain.

Phase 2 analysis involves testing any future samples on the charts developed in Phase 1 to test if they meet the quality requirements.

For this project, we used R - an open source software to analyze statistical data on the given data set to calculate the Upper Control Limit and Lower Control Limit to examine when a product is not meeting the expected quality specifications. Phase 1 analysis was carried out to estimate the in-control mean and variance, after which Phase 2 analysis can be carried out on any future samples.

The major give aways from performing the project were the following:

1. It provided hands on experience with many of the concepts which we have studied theoretically in class. We understood the application of concepts such as Principal Component Analysis, Multivariate quality control, T^2 chart to real-world problems.

2. The absence of physical significance to the data was a challenge because it made the decision of method selection difficult. It helped us understand other aspects in which we can approach a problem such as by visualisation, checking whether there is correlation or covariance etc.

3. It is hard to make covariance matrix of 209 predictors and analyze it through Hotelling T-square statistics. It was better to reduce number of predictors using Principal Component Analysis approach. We used covariance matrix to reduced dataset to 4 principal components.

4. Since the analysis was performed in R, it helped us gain in-depth knowledge about the software, its syntax, commands and packages.

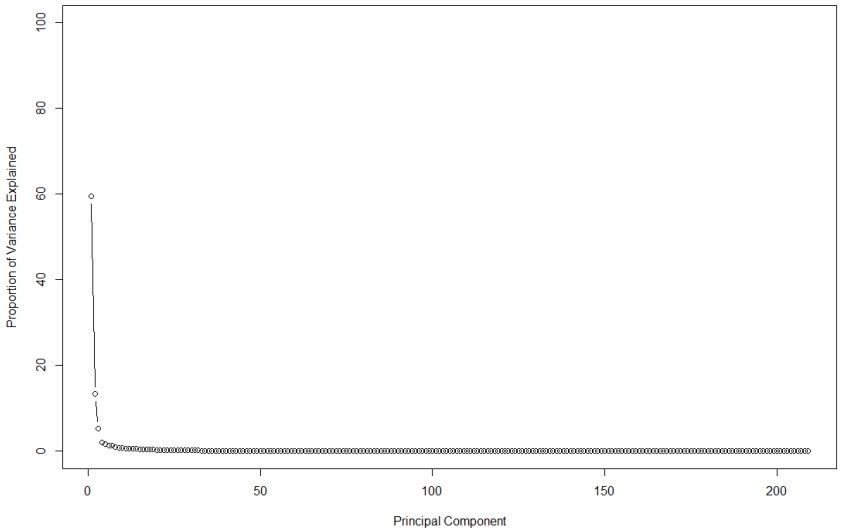
Principal Component Analysis

As physical meaning of each object is not specified, we can go ahead with building principal components from covariance matrix instead of correlation matrix.

Principal Component Analysis is carried out to find out the components that explain maximum variance. When there are too many attributes, the detection signal is overwhelmed by the noise in the data. This is called “Curse of Dimensionality.” Also, the principal of effect sparsity says that the “vital few” are more important than the “trivial many.”

The mean and variance of all the columns was calculated and a plot of “Principal Component” versus “Proportion of Variance Explained” was plotted.

Scree plot -



Here is the % variation explained by first 10 principal components.

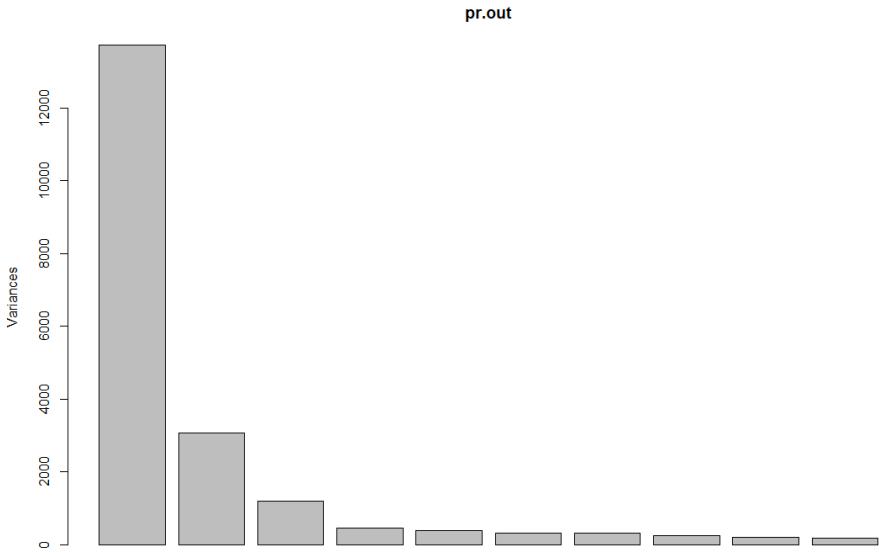
PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
59.4	13.5	5.2	2	1.7	1.4	1.3	1	0.8	0.8

We can see that top 4 principal components explain around 80.1% variation from the data.

The Scree plot and Pareto plot which are simpler, more intuitive and more interpretable were plotted to see the most important variables.

Below pareto plot gives value of variance that is explained by initial principal components.

Pareto plot -



Methods used

Multivariate detection using multiple univariate charts

As it has been suggested in lecture unit 32 *(Multivariate Detection Versus Multiple Univariate Detections)*, multiple univariate x-bar charts can be used to find out of control points in Multivariate distributions. In this case, if any of the univariate chart has out of control points, the whole process is called to be out of control.

In our sample data, we have selected 4 principal components that explain at-least 80% of the variation in the main dataset. We are going to plot univariate x-bar charts of all these 4 components. Below are the formulas to calculate upper and lower control limits of these charts.

$$UCL/LCL = \mu_0 \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma_0}{\sqrt{n}}$$

Here we have considered alpha value to be 0.0027 to detect a 3-sigma mean shift.

The problem with applying multiple univariate charts to multivariate distribution is that the correlation between the variables can not be detected using this methodology. Since PCA gives components which are uncorrelated, multiple univariate charts can be applied.

We used ‘qicharts’ package in R to calculate UCL/LCL and plotted univariate x-bar chart for each principal component.

Methodology -

1. First we will plot univariate charts for all 4 principal components and detect number of out of control points in each component.
2. We will then select out of control points from first principal component and then remove them from the dataset (selected 4 principal components).
3. We will plot control charts for all principal components again and check out of control points.
4. We will perform steps 1-3 and remove points for second, third and fourth principal component. This would be first iteration.
5. We will repeat another iteration by running steps 1-4 if there are any other out of control points.

The results of the iterations performed, alongwith the UCL, LCL and number of out-of-control points is summarised in the table below.

Results

Iteration 1

Number of out of control points detected from PC1 and removed from dataset = 43
UCL=171, LCL=-171.2

Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	8	60	2	6

Number of out of control points detected from PC2 and removed from dataset= 60
UCL = 80, LCL = -74

Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	12	10	2	6

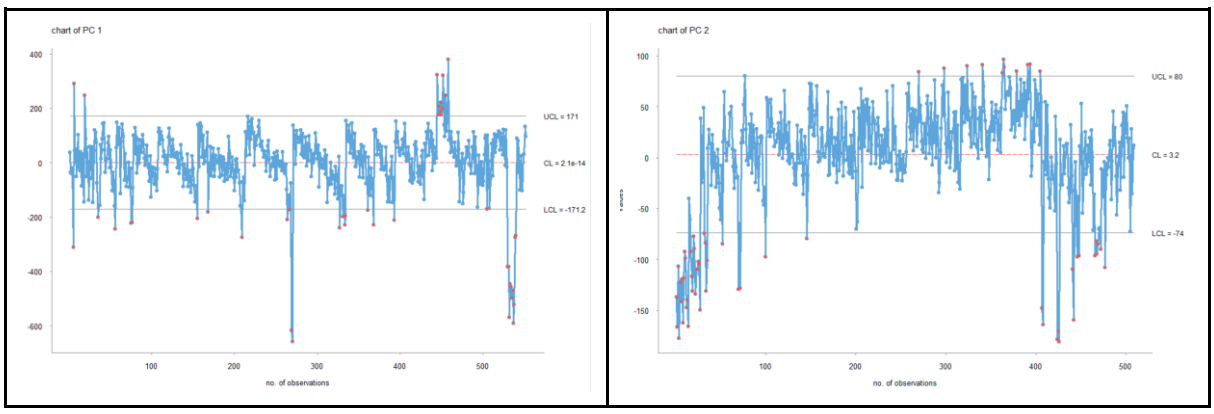
Number of out of control points detected from PC3 and removed from dataset = 2
UCL = 78, LCL = -87.2

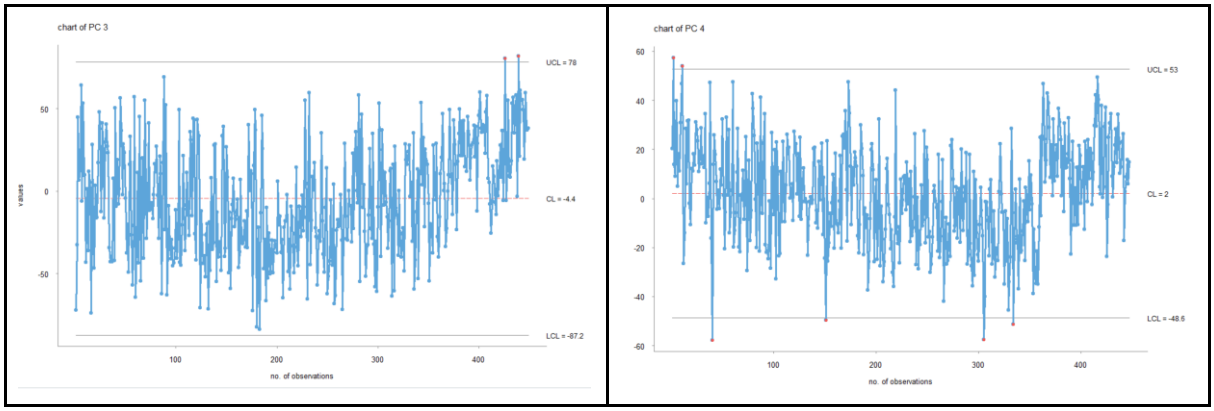
Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	12	10	0	6

Number of out of control points detected from PC4 and removed from dataset = 6
UCL = 53, LCL = -48.6

Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	11	9	0	0

Iteration 1 charts -





Iteration 2

Number of out of control points detected from PC1 and removed from dataset = 11
UCL = 172, LCL = -136.8

Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	4	6	0	0

Number of out of control points detected from PC2 and removed from dataset = 6
UCL = 86, LCL = -56.3

Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	1	2	0	0

Number of out of control points detected from PC3 and removed from dataset = 0
UCL = 77, LCL = -86

Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	1	2	0	0

Number of out of control points detected from PC4 and removed from dataset = 0
UCL = 53, LCL = -48.9

Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	1	2	0	0

Iteration 3

Number of out of control points detected from PC1 and removed from dataset = 1
UCL = 172, LCL = -126.4

Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	1	2	0	0

Number of out of control points detected from PC2 and removed from dataset = 2
UCL = 86, LCL = -54.3

Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	1	0	0	0

Number of out of control points detected from PC3 and removed from dataset = 0
UCL = 76, LCL = -85.1

Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	1	0	0	0

Number of out of control points detected from PC4 and removed from dataset = 0
UCL = 53, LCL = -49.3

Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	1	0	0	0

Iteration 4

Number of out of control points detected from PC1 and removed from dataset = 1
UCL = 172, LCL = -126.1

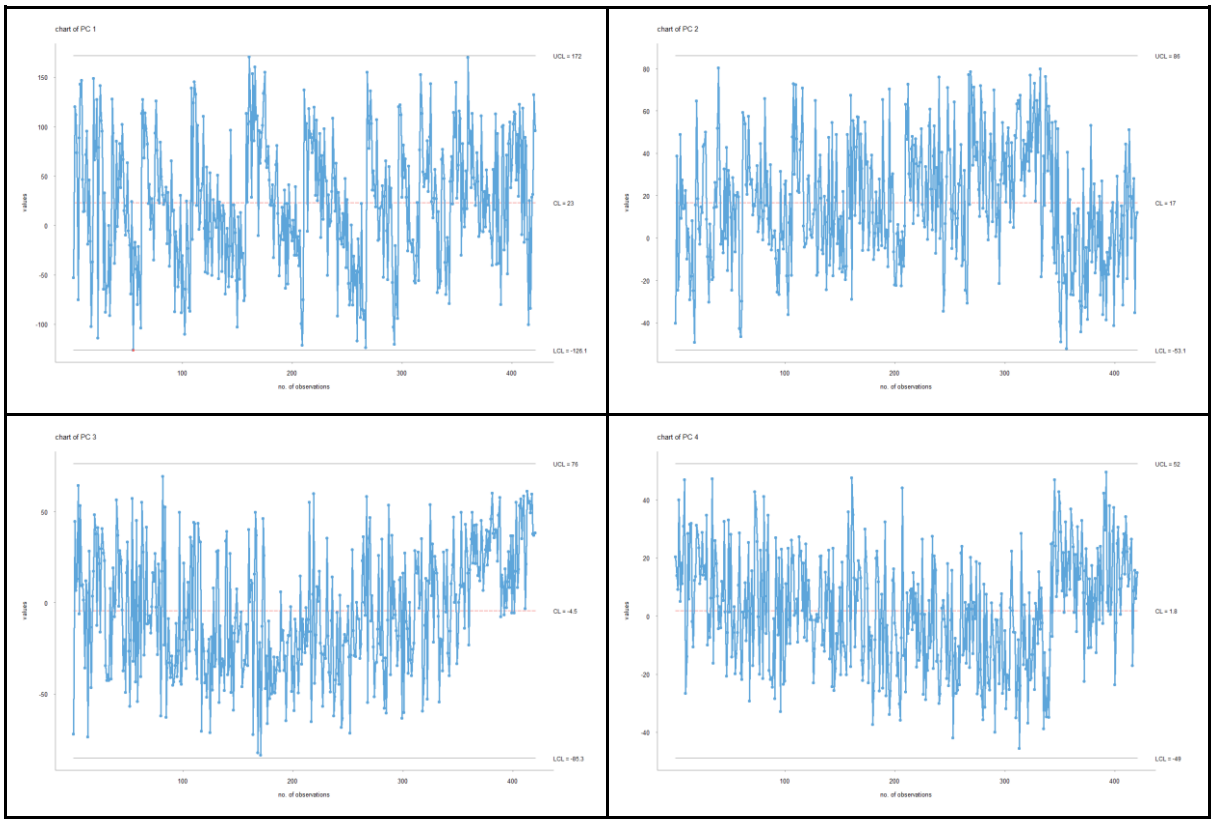
Remaining out of control points	PC 1	PC 2	PC 3	PC 4
	0	0	0	0

As it can be seen that no principal components have out of control points now, the process is in in-control. The modified control limits of all principal components are now -

Principal component 2 - UCL = 86, LCL = -53.1
Principal component 3 - UCL = 76, LCL = -85.3

Principal component 4 - UCL = 52, LCL = -49
After the four iterations, number of in-control data points are 420, whose plots are as shown below.

Iteration 4 charts -



Multivariate detection using Hotelling T-square chart

As it is discussed in lecture unit 34-37, Hotelling T-square charts can be used to detect out of control samples in Multivariate processes.

T2 chart has mainly 3 type of cases to detect out of control data points. In our project dataset, we do not have the population mean and covariance matrix, so we would have to estimate it through sample data points. This is called Phase 1 analysis and it fits in case 3(a) as taught in Lecture unit 37.
T2 statistic formula is given by -

$$T^2 = (\mathbf{x}_j - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}).$$

In Case 3(a), T2 statistics does not follow a F-distribution. Infact, when when n = 1 for Phase I analysis, T2 statistics follows a beta distribution. But in practice, it is often recommend to use a χ2(p) to approximate.

$$T^2 \overset{approx}{\sim} \chi^2(p) \text{ and } UCL = \chi^2_{1-\alpha}(p) \text{ for a given } \alpha.$$

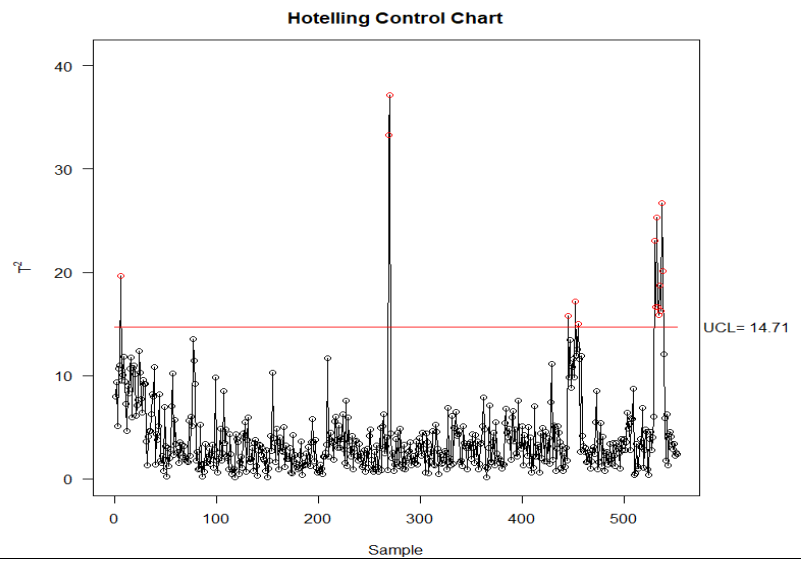
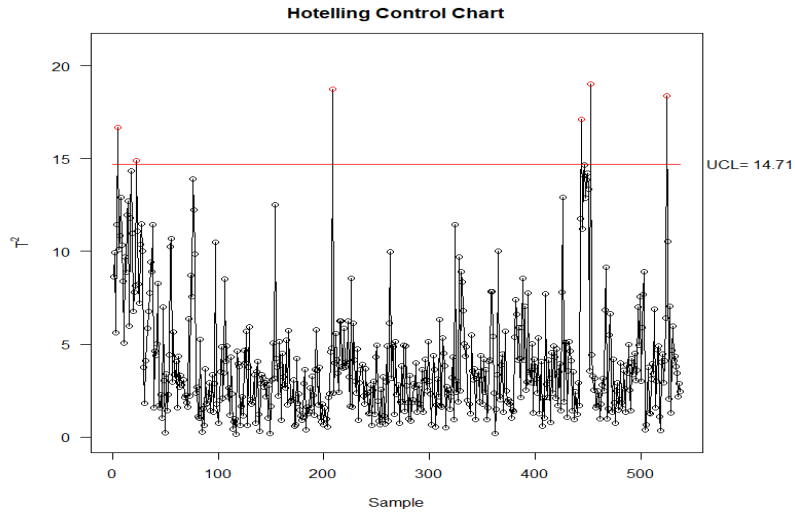
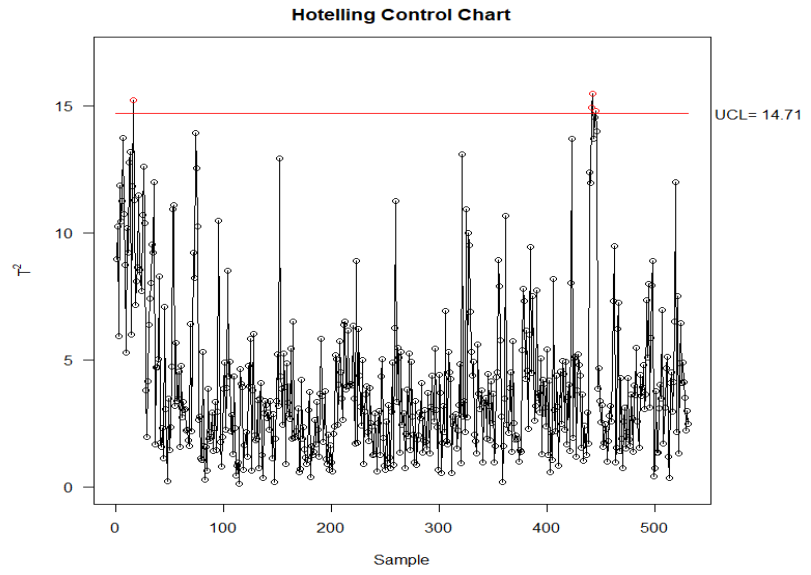
To calculate T-square statistics, “MSQC” package in R was used for this. The covariance matrix and x-bar statistics is calculated for all four Principal Components. The points above the Upper Control Limit and those below the Lower Control Limit were omitted from the data set and the T^2 chart was then plotted again on the remaining data points.

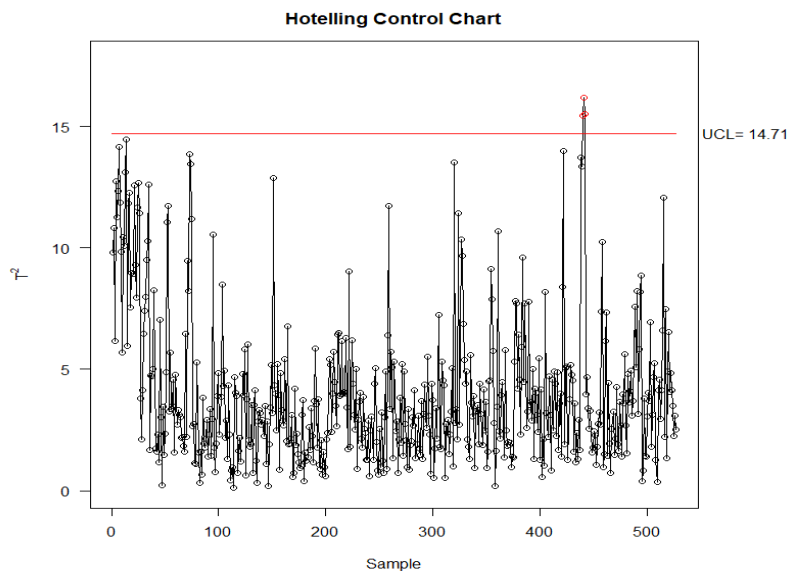
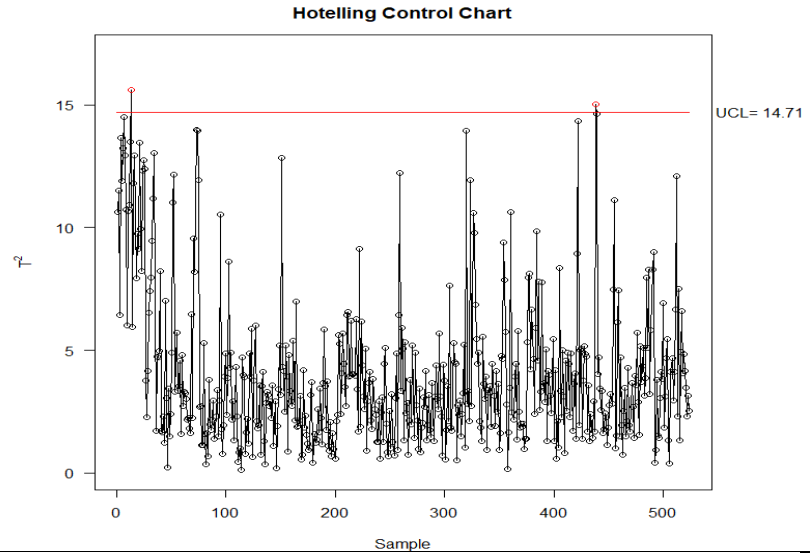
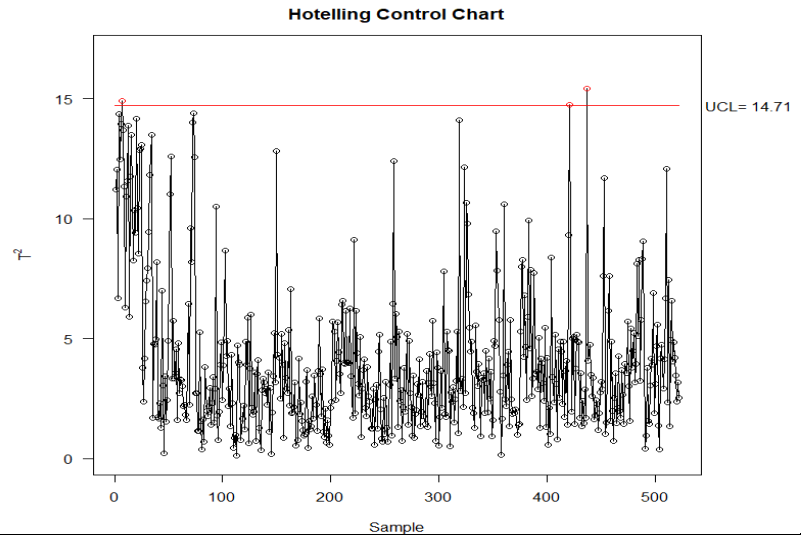
The level of confidence alpha = 0.005. The T^2 follows a chi-square distribution with p=4 degrees of freedom. UCL is calculated using the formula above.

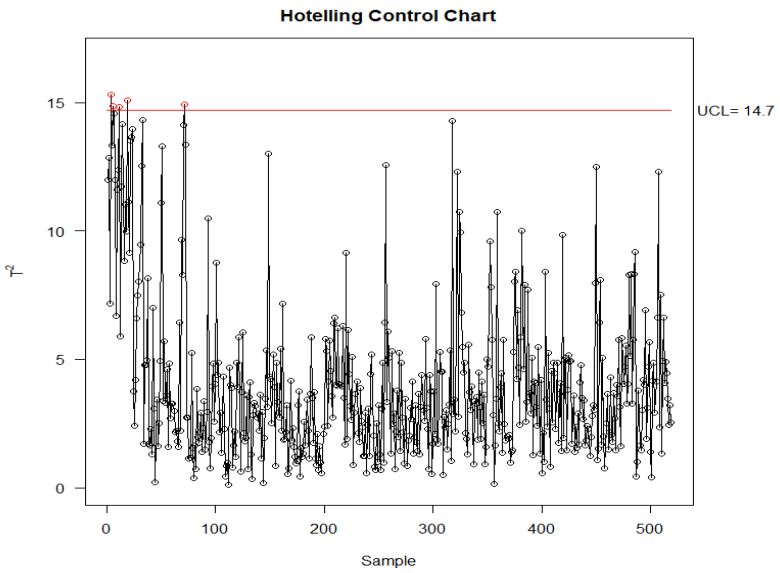
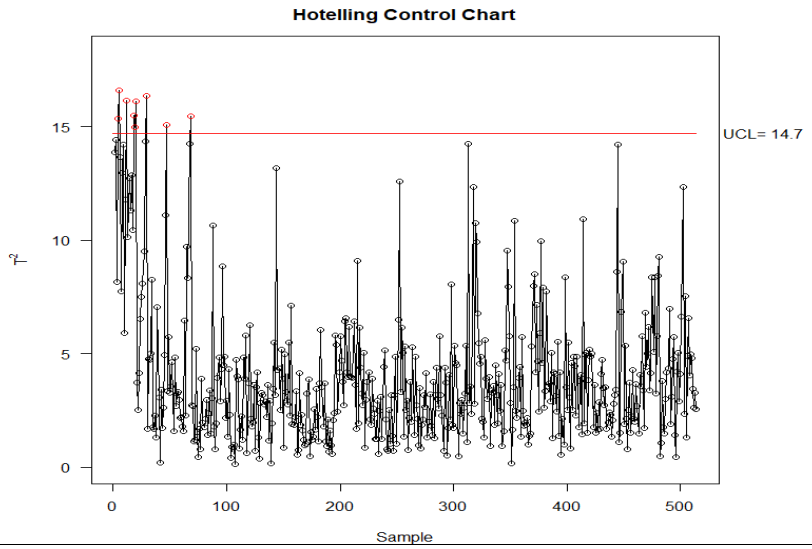
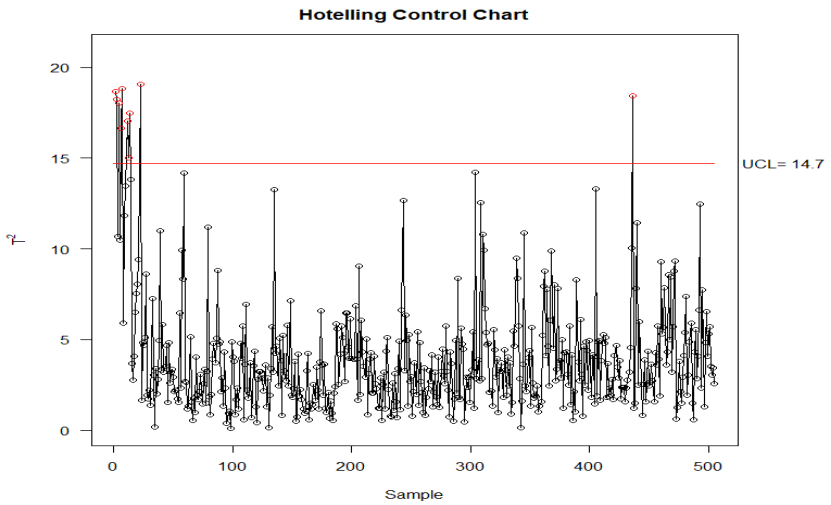
Methodology

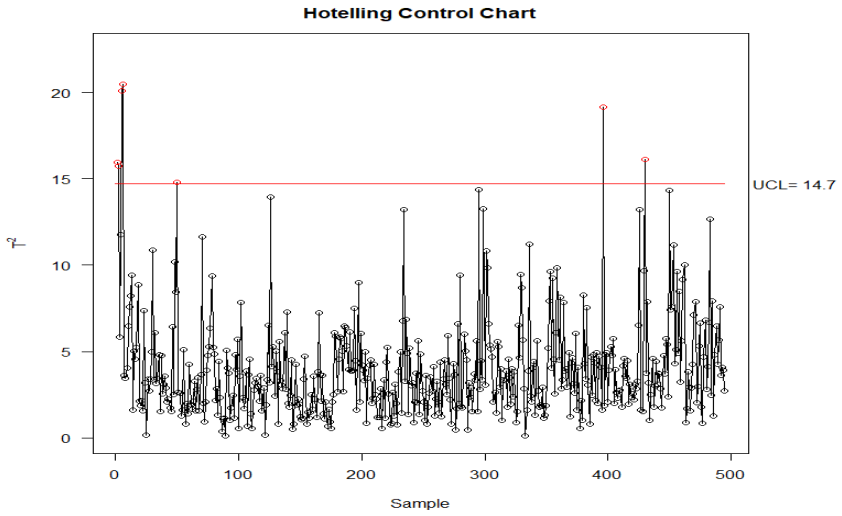
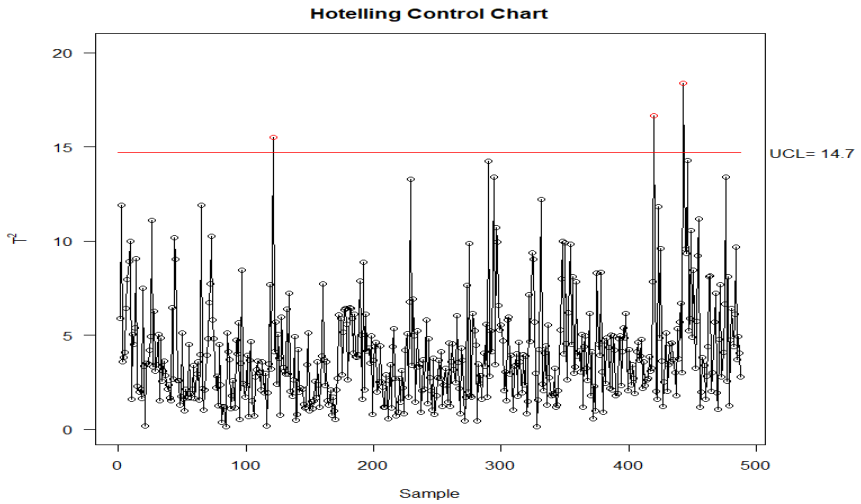
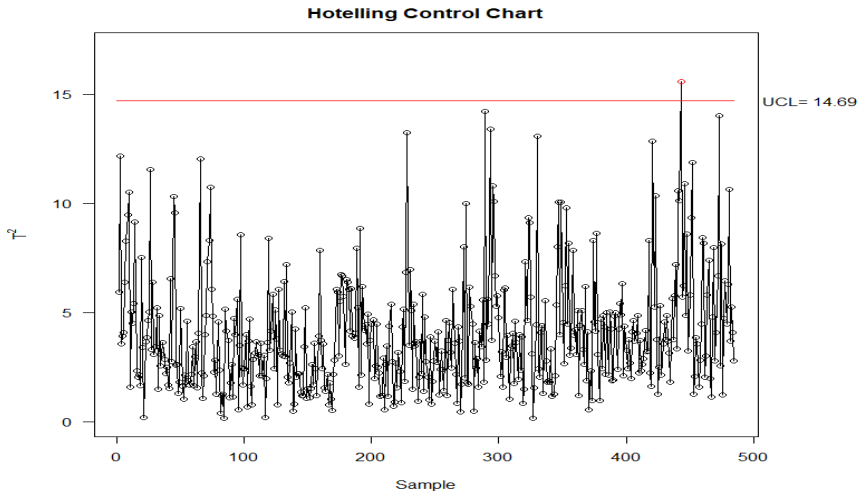
1. First we will plot T-square chart for dataset having 4 principal components and detect number of out of control points in chart.
2. We will then select out of control points from chart and remove them from the dataset.
3. We will repeat iterations by running steps 1-2 if there are any other out of control points.

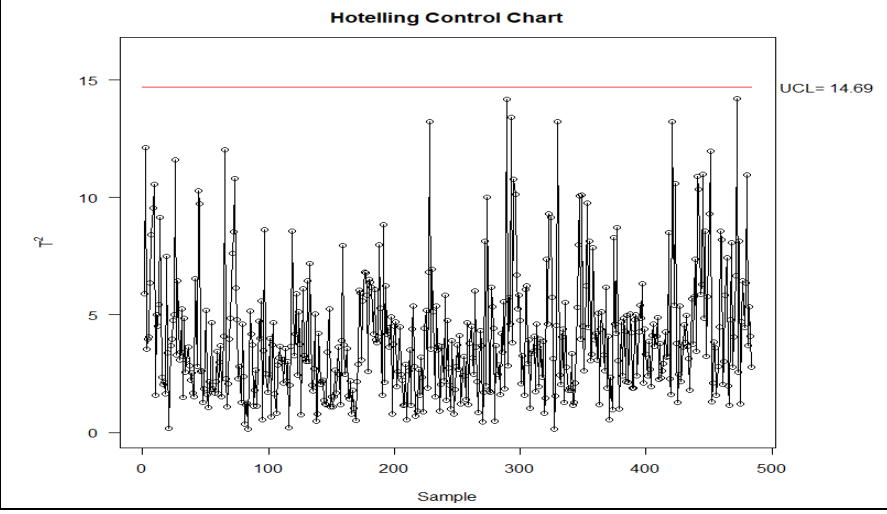
Results

Iteration 1, UCL = 14.71, LCL = 0, out-of-control points = 15	
Iteration 2, UCL = 14.71, LCL = 0, out-of-control points = 06	
Iteration 3, UCL = 14.71, LCL = 0, out-of-control points = 04	

Iteration 4, UCL = 14.71, LCL = 0, out-of-control points = 03	
Iteration 5, UCL = 14.71, LCL = 0, out-of-control points = 02	
Iteration 6, UCL = 14.71, LCL = 0, out-of-control points = 03	

Iteration 7, UCL = 14.71, LCL = 0, out-of-control points = 05	
Iteration 8, UCL = 14.71, LCL = 0, out-of-control points = 09	
Iteration 9, UCL = 14.71, LCL = 0, out-of-control points = 10	

Iteration 10, UCL = 14.71, LCL = 0, out-of-control points = 07	
Iteration 11, UCL = 14.71, LCL = 0, out-of-control points = 03	
Iteration 12, UCL =14.71, LCL = 0, out-of-control points = 01	

Iteration 13, UCL =14.71, LCL = 0, out-of-control points = 0	

484 data points were left in the data set at the end of all iterations and all these points were within the control limits.

Multivariate Exponential Weighted Moving Average(M-EWMA) chart

As suggested in unit 38, **Multivariate Exponential Weighted Moving Average(M-EWMA) chart** is used to detect small consistent mean shifts in the data. We used “Spc” library in R for detecting out of control data points in the dataset . Critical values selected were, the significance level = 0.005, lambda = 0.5, L0 = 200, p = 4 and smoothing parameter l = 0.5.The following formula is used for finding out the UCL for M-EWMA chart.

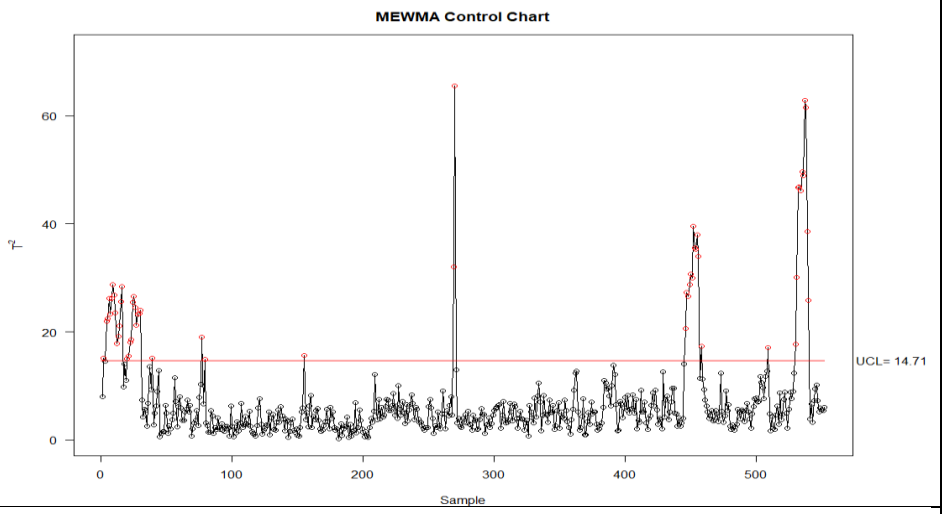
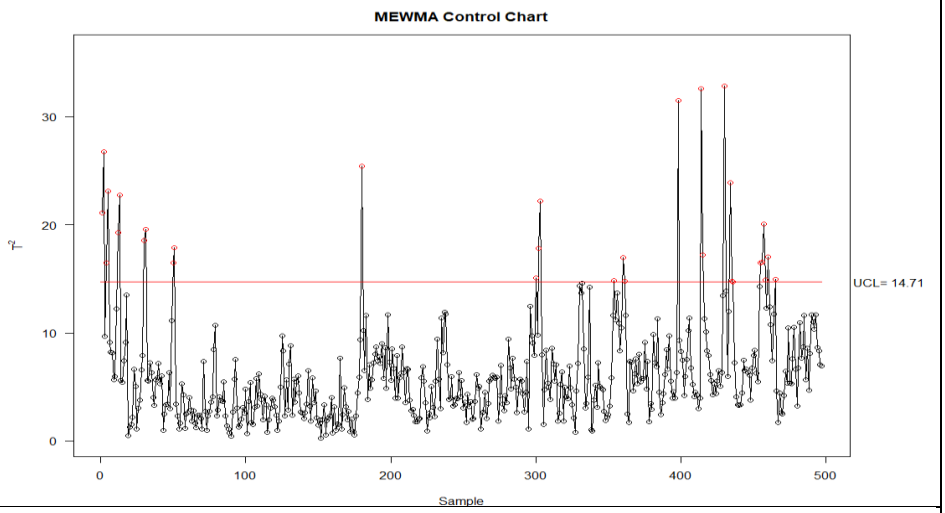
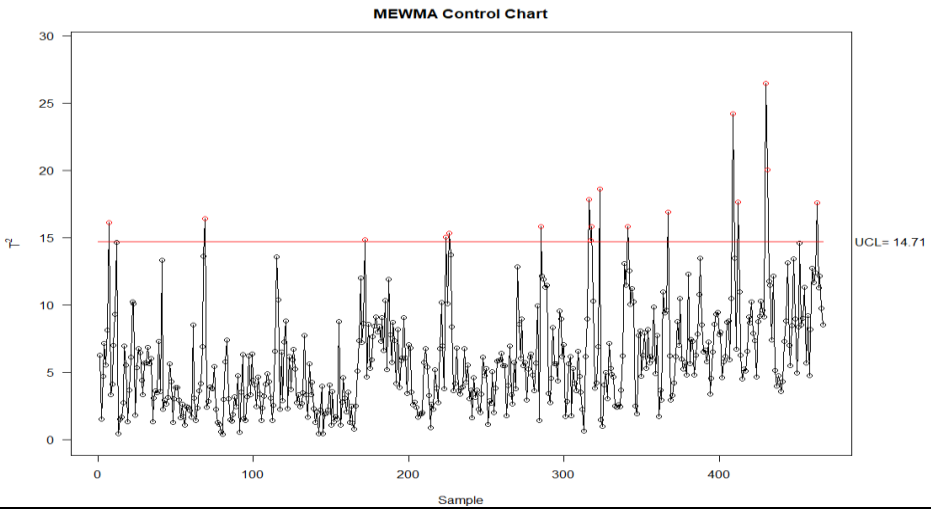
$$UCL = \chi^2_{1-\alpha}(p) \text{ for a given } \alpha.$$

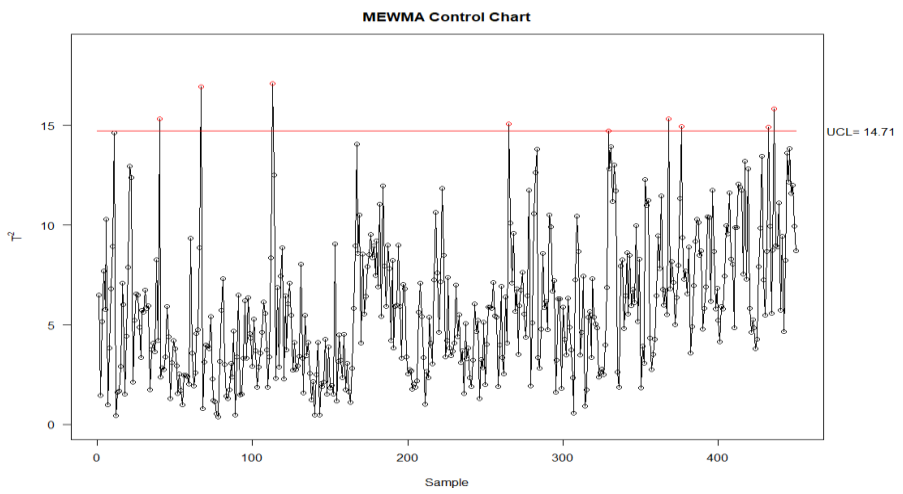
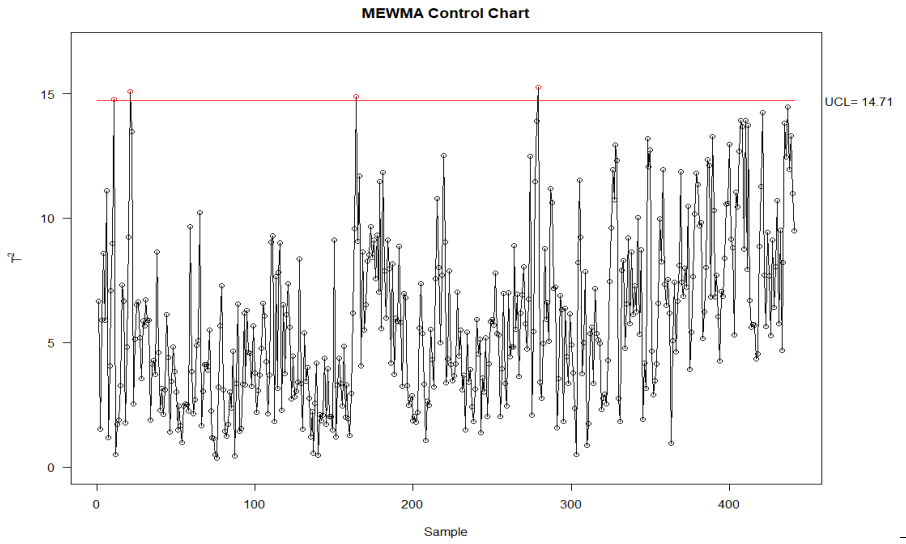
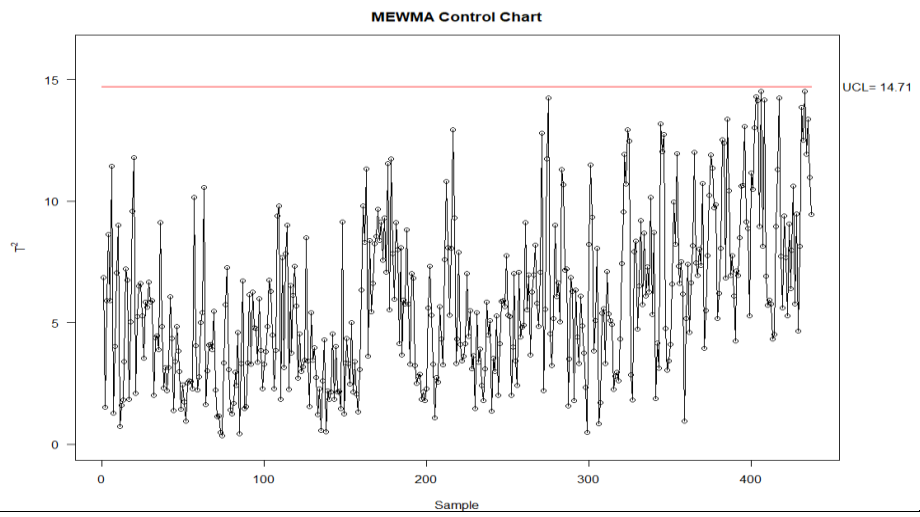
Methodology

1. First we will plot M-EWMA chart for dataset having 4 principal components and detect number of out of control points in chart.
2. We will then select out of control points from chart and remove them from the dataset.
3. We will repeat iterations by running steps 1-2 if there are any other out of control points.

Several iterations were required to remove all the out of control data points, the plots of which are as shown below.

Results

Iteration 1, UCL = 14.71, LCL = 0, number of out of control points = 55	
Iteration 2, UCL = 14.71, LCL = 0, number of out of control points = 30	
Iteration 3, UCL = 14.71, LCL = 0, number of out of control points = 17	

Iteration 4, UCL = 14.71, LCL = 0, number of out of control points = 09	
Iteration 5, UCL = 14.71, LCL = 0, number of out of control points =04	
Iteration 6, UCL = 14.71, LCL = 0, number of out of control points = 0	

At the end of this process, we had 437 in-control data points.

Multivariate Cumulative Sum (M-CUSUM) Chart

T2 chart does not perform well enough for detecting small mean shifts. The Multivariate CUSUM is a memory chart which can be used for this purpose.
The M-CUSUM procedure is as described below:
For the ith observation, define the CUSUM as

$$\mathbf{C}_i = \sum_{j=i-n_i+1}^{i} (\mathbf{x}_j - \mu_0)$$

where Ci = CUM SUM of all previous ni x's = ni(xbar - μ0), xbar = average over previous ni x's

$$MC_i = \max\{0, (\mathbf{C}_i^T \mathbf{\Sigma}_0^{-1} \mathbf{C}_i)^{\frac{1}{2}} - k \cdot n_i\}$$

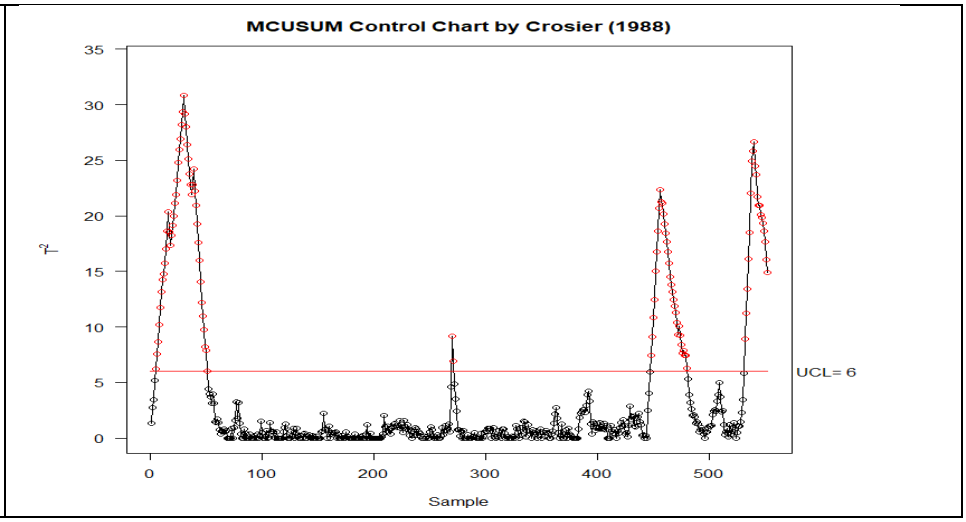
Where k is a user defined offset constant which is half of the statistical distance between the mean and the shift to be detected. The CUSUM chart signals when MCi>UCL

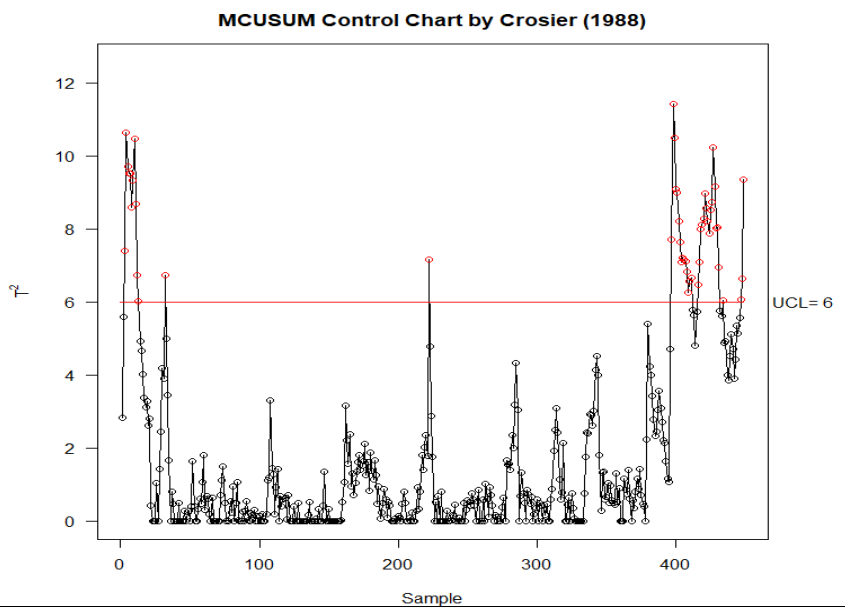
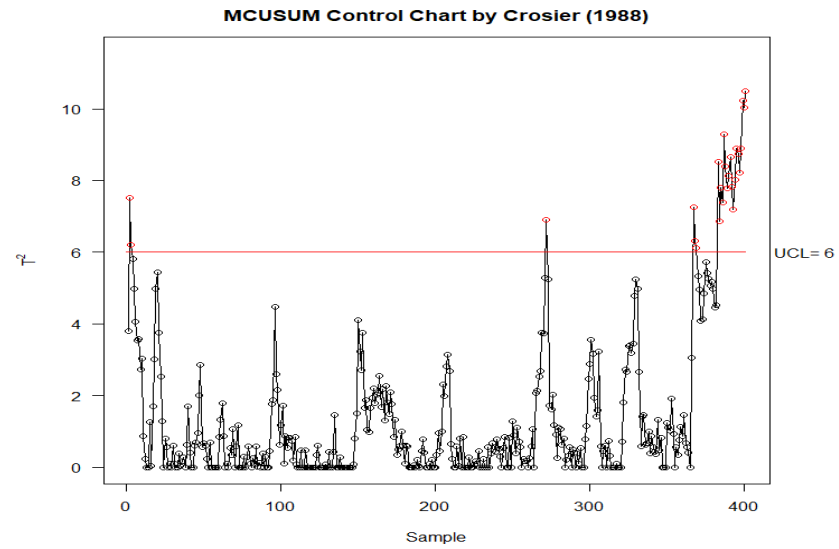
Methodology :

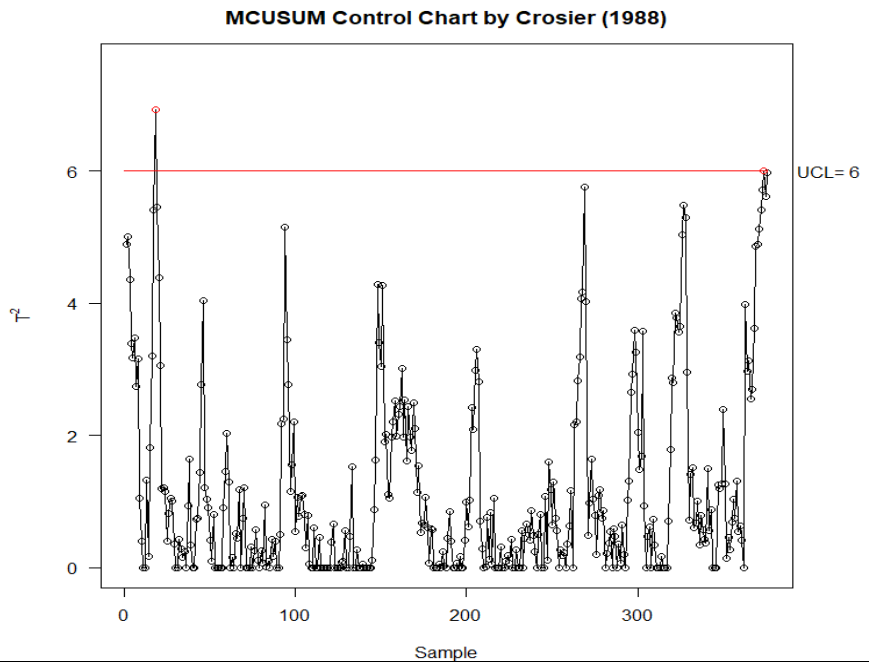
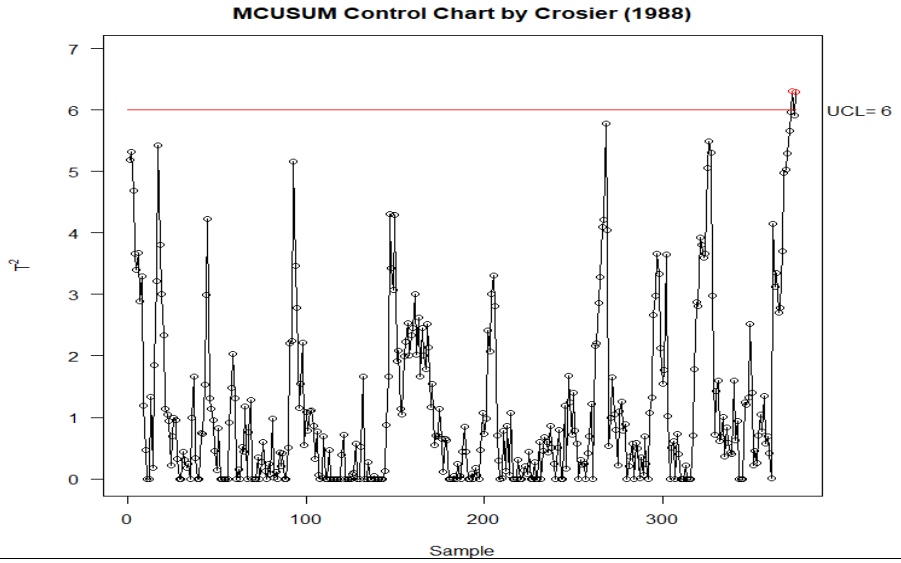
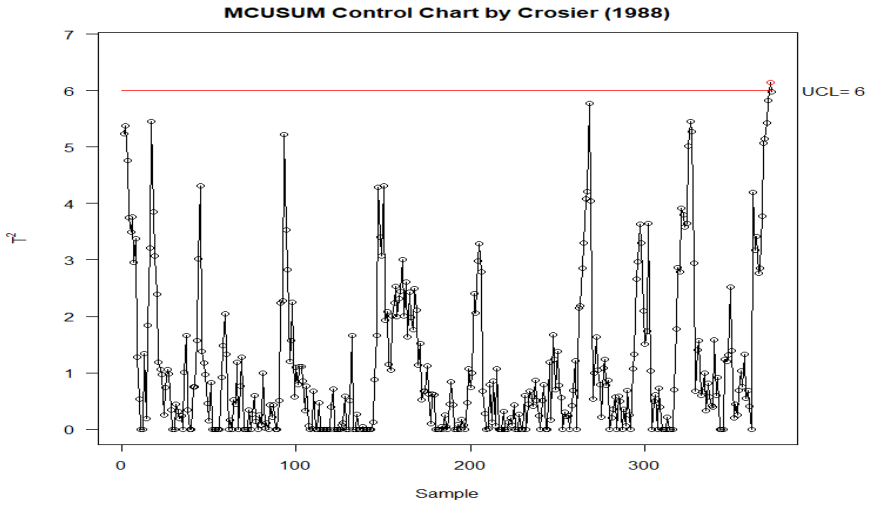
1. First we will plot M-CUSUM chart for dataset having 4 principal components and detect number of out of control points in chart.
2. We will then select out of control points from chart and remove them from the dataset.
3. We will repeat iterations by running steps 1-2 if there are any other out of control points.

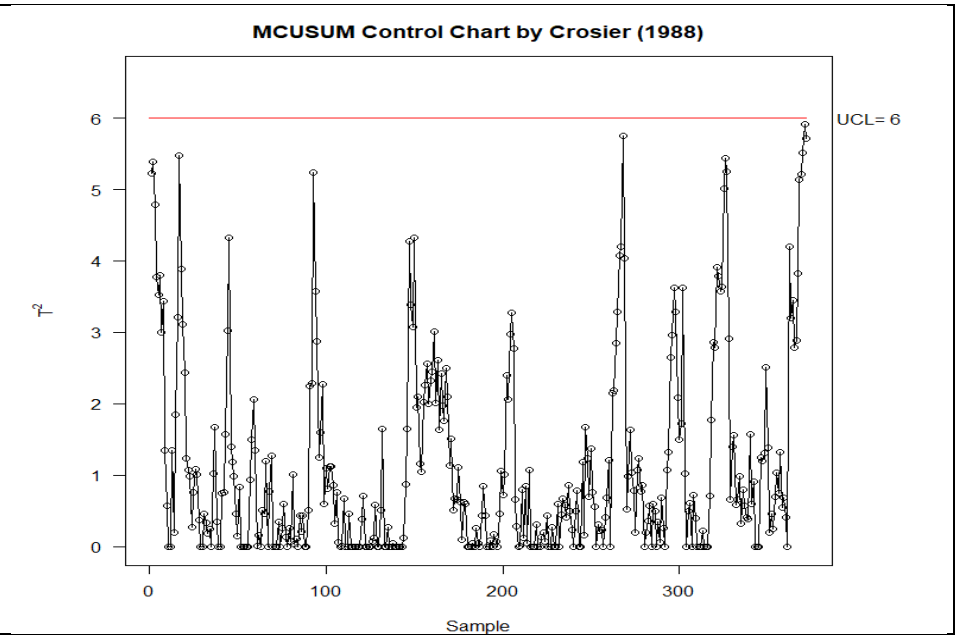
While applying Multivariate Cumulative Sum(M-CUSUM**)** chart to the given dataset, the critical parameters selected were k = 1.5 and h = 6.

Results

Iteration 1 UCL = 6, LCL = 0, Number of out of control data points = 103	

Iteration 2 UCL = 6, LCL = 0, Number of out of control data points = 48	
Iteration 3 UCL = 6, LCL = 0, Number of out of control data points = 25	

Iteration 4 UCL = 6, LCL = 0, Number of out of control data points = 1	
Iteration 5 UCL = 6, LCL = 0, Number of out of control data points = 2	
Iteration 6 UCL = 6, LCL = 0, Number of out of control data points = 1	

Iteration 7 UCL = 6, LCL = 0, Number of out of control data points = 0	

The number of in-control data points remaining at the end of five iterations were 372.

Combined T2 and M-EWMA chart

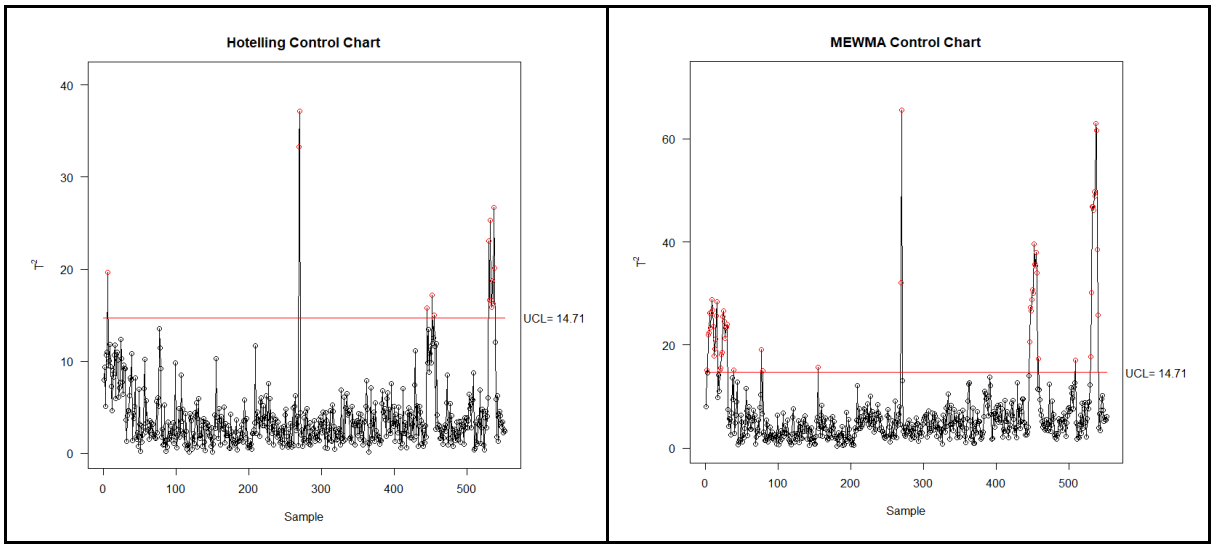
While T2 chart is good at detecting large spikes and M-EWMA and M-CUSUM are good at detecting sustained mean shifts. To detect small consistent mean shifts and large spikes together in the given dataset, we used combined T2 and MEWMA charts.
In a similar fashion as the above methods, out-of-control data points were removed in successive iterations.

Methodology :

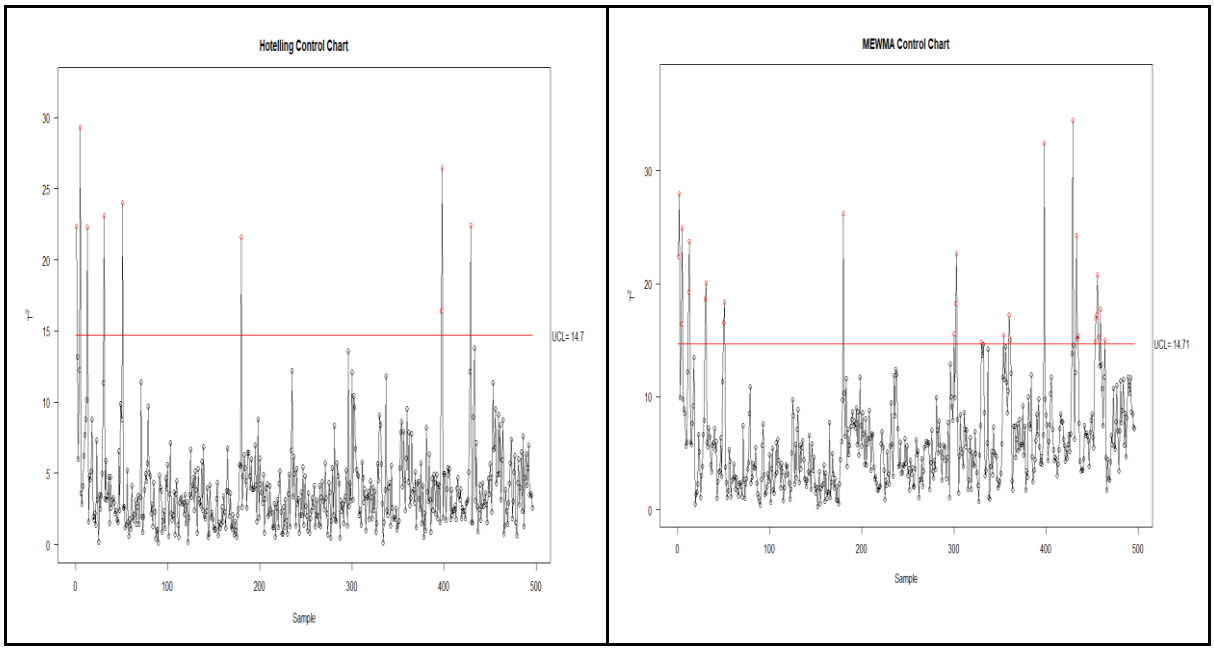
1. We will plot T2 chart and M-EWMA chart simultaneously for dataset having 4 principal components and detect number of out of control points in chart.
2. We will then select out of control points from charts and remove them from the dataset.
3. We will repeat iterations by running steps 1-2 if there are any other out of control points.

Results

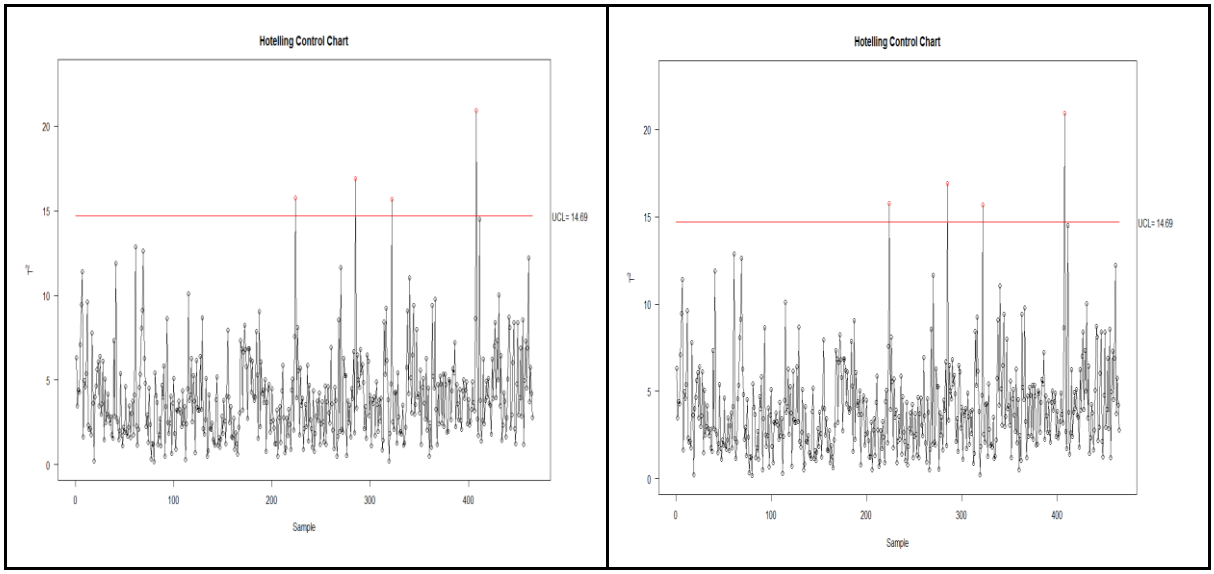
Iteration 1 UCL for T2=14.71 and UCL for EWMA=14.71 ,Combined No.Of OOC points =56



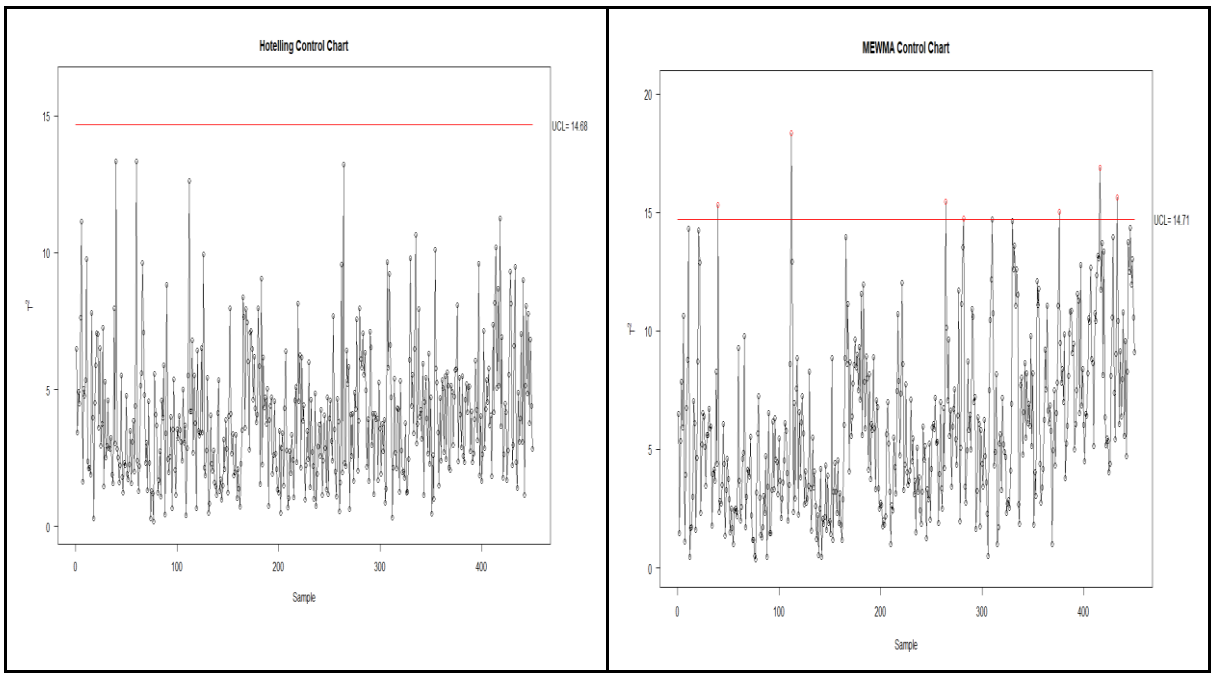
Iteration 2 UCL for T2=14.71 and UCL for EWMA=14.71 ,Combined No.Of OOC points =37



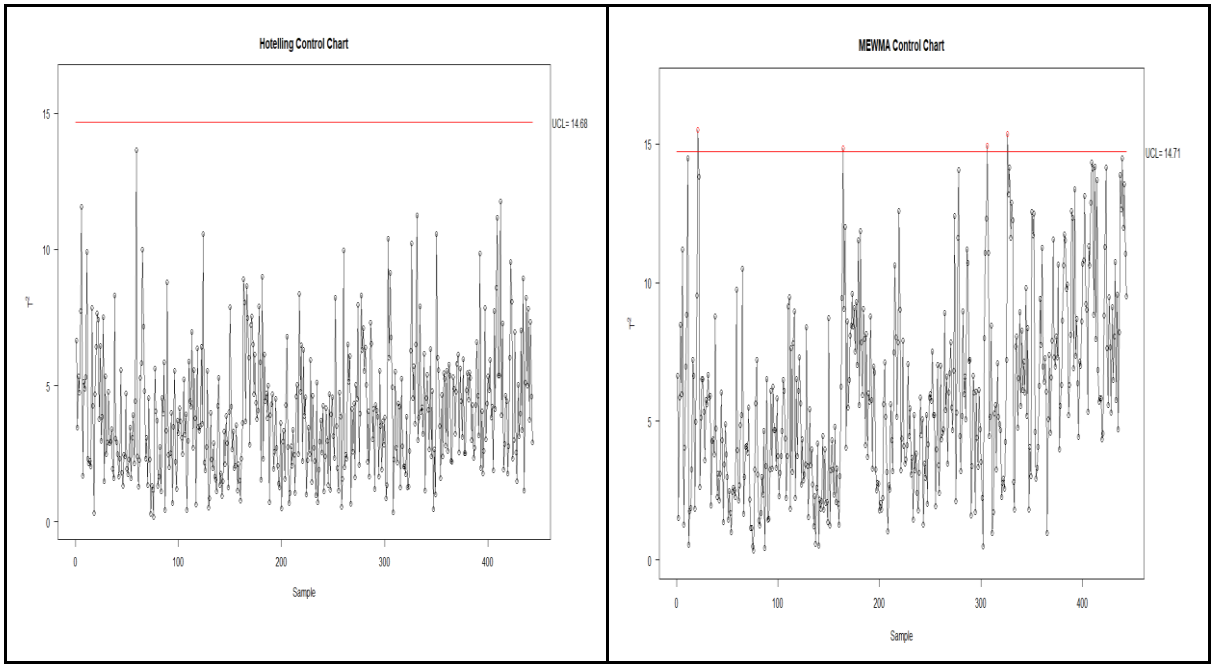
Iteration 3 UCL for T2=14.71 and UCL for EWMA=14.71 ,Combined No.Of OOC points =15



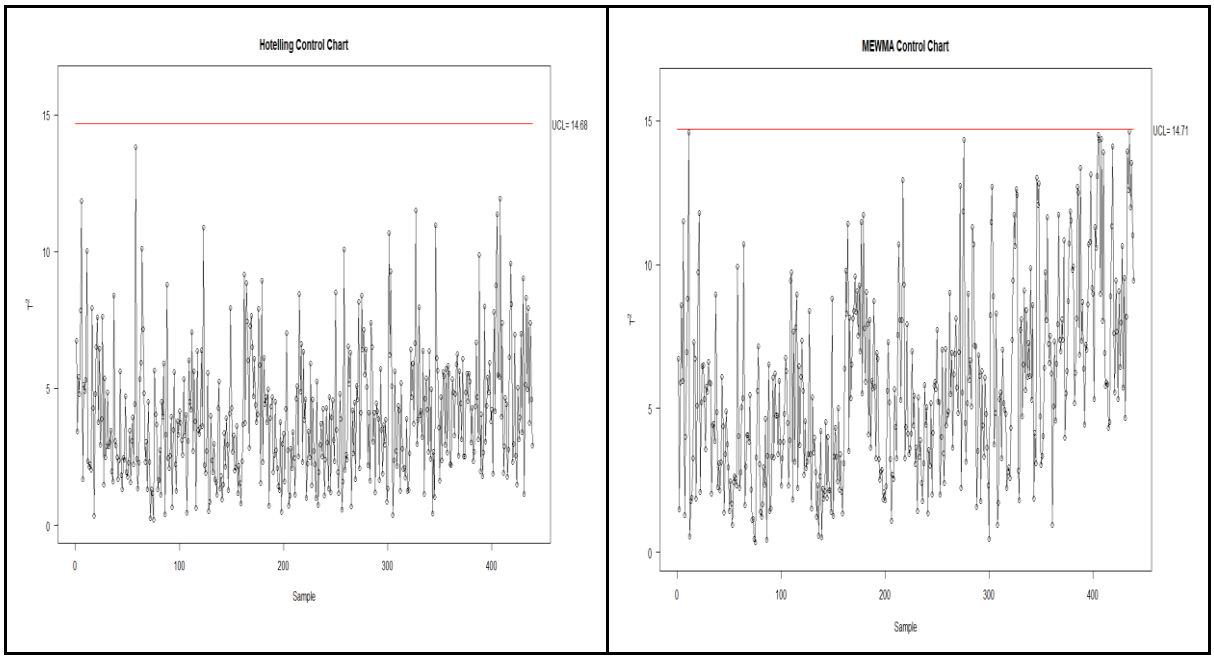
Iteration 4 UCL for T2=14.71 and UCL for EWMA=14.71 ,Combined No.Of OOC points =07



Iteration 5 UCL for T2=14.71 and UCL for EWMA=14.71 ,Combined No.Of OOC points =04



Iteration 6 UCL for T2=14.71 and UCL for EWMA=14.71 ,Combined No.Of OOC points =00



439 in control data points were left in the dataset at the end of these iterations which can be used to calculate estimates of mean and covariance.

Conclusion

Based on the table below, which compares different types of charts used in this project, we can say that T2 chart recommends the highest number of in-control data points, While M-CUSUM method recommend the lowest number of in-control points (372). However, T2 chart takes maximum number of iterations to reach at in-control sample. Also, T2 chart detects large spikes in the dataset.

So, if we want to select maximum number of in-control points out of the dataset and get an in-control sample without large spikes, we can select indexes of in-control points from T2 chart. Otherwise, if our objective is to eliminate small sustained mean shift then we can go ahead with M-EWMA approach in-control points which gives 437 in-control sample indexes.

And, if we want to eliminate both large spikes and small mean shifts, we can go ahead with a combined analysis of T2 and M-EWMA to select in-control points.

Type of Charts	ARL0	No.of Iterations for In control data	UCL for last iteration	In control data Points
Individual X bar chart	370	4	52	420
T2 chart	200	13	14.69	484
MEWMA	200	6	14.71	437
MCUSUM	200	7	6	372
Combined T2 & MEWMA	200	6	14.68 & 14.71	439

Through Individual X bar chart, we are getting ARL1 value as 2.00 while T2 chart ARL1 is around 1.74. So, depending on the application of ARL, we can decide to choose which method to use for developing control charts.

Appendix

R-codes

- Individual x-bar chart –

Individual_xbar_chart
.Rmd

- Hotelling T-square chart –

t2chartarl200.R

- M-EWMA chart –

mewma.R

- M-CUSUM chart –

mcusum.R

- M-EWMA and Hotelling T-square chart –

t2 and mewma
combined.R