INDUSTRIAL & SYSTEMS ENGINEERING
TEXAS A&M UNIVERSITY

ISEN 614 Project Report:

Multivariate Phase I analysis for monitoring manufacturing process

Prepared by:

- Arch Jignesh Desai **(UIN: 627006997)**
- Siddharth Devendra Rajopadhye **(UIN: 827005699)**
- Vatsal Nimeshkumar Thakkar **(UIN: 128003914)**

Acknowledgement

We as a team would like to firstly thank Professor Yu Ding, for assigning us such a challenging and holistic project. It is through this project we were able to apply our theoretical concepts and knowledge learned in the course and see its wide scale application in real-world problem solving.

Secondly, we would also like to thank Mr. Imtiaz Ahmed, teaching assistant for our subject, for his constant guidance throughout the project.

Thirdly, we would like to thank Texas A&M University for providing us with the resources and support, without which we would not be able to complete this project.

Through this medium we would also like to acknowledge the fact that this complied project report and the work we are submitting is our original work. We hold the ideals of our school Texas A&M in high regards and we abide by the code of honor.

Table of Contents

Executive Summary

The objective of the project is to identify the in-control distribution parameters of the given dataset. The dataset is obtained from a manufacturing process which has 209 variables and 552 data points. It contains both in-control and out-of-control data points. While the physical interpretation of the variables is not specified. our focus is mainly based on eliminating the out-of-control data points using multivariate control charts so that a control scheme can be set up for monitoring the future data points. This approach is essentially known as Phase I Analysis of Multivariate data.

As the number of variables in the dataset is large, firstly we used dimension reduction methods to reduce the dimension of the whole dataset. Principal Component Analysis (PCA) was used to reduce the dimension by which we obtained 4 PCs which explain the 80.097 % variability from the original dataset. The covariance matrix was used for PCA as the physical interpretation of variables has been omitted, the relative magnitude between them might be of importance.

Once we have the Principal Components, then Multivariate charts like Hotelling T^2 Chart and m-CUSUM charts were used to detect and eliminate the out-of-control points within the dataset. Firstly, to eliminate large spikes within dataset, we applied T^2 chart on the transformed variables till we got zero out-of-control data points. Then, to account for the small sustained mean shift, we used m-CUSUM charts. Multiple iterations were required of both the control charts to clean the data of any outliers that were present.

Finally, the distribution parameters can be set up using the cleaned data set and which can be used for monitoring future data. We used packages within R software as a tool for computational work.

Introduction

Problem Statement:

The dataset is obtained from a manufacturing process with in-control and out-of-control data points in it. The objective is to conduct a Phase I Analysis to eliminate all the out-of-control data points so as to set up control charts for monitoring future data from the same process. The in-control data obtained after the Phase I Analysis will yield the distribution parameters viz. Mean and Covariance Matrix. The properties of the dataset are:-

1. The dataset has 552 data records which are correlated with each other.
2. Each row from the spreadsheet is a data record with 209 values.
3. The dimension of dataset p = 209 and sample size n = 1
4. The physical meaning of each variable is omitted.

Dimension Reduction:

The large dimension of the dataset makes it difficult for us to conduct univariate analysis with the help of multiple univariate control charts. In doing so, we may end of inflating either of the errors namely α or β. Furthermore, the large aggregation of small noise within the data would make it even more difficult for us to reject the null hypothesis. Hence, it is utmost required to transform the current variables into a smaller number of variables which would not lose the variability within the original data. That is, we need to find out the 'vital few' from the 'trivial many' as per Pareto's Principle. We make use of Principal Component Analysis (PCA) to transform the variables into a system of linear combination having the maximum variability from the dataset. PCA would yield equal number of transformed variables as of those in the original dataset, but we would be able to eliminate the PCs which do not contribute to the variability of the dataset i.e. find out the 'vital few' with the help of Scree plots and Pareto Charts.

Multivariate Control Charts:

As the dataset contains 209 variables, it is not in our interest to use multiple univariate control charts as we would end up inflating either α or β errors leading to higher false alarms or high misdetections respectively. Also, we would not be able to detect any changes in the inherent correlation between the variables of the dataset. Hence to tackle the curse of dimensionality, we choose to use multivariate control charts namely T^2 chart and m-CUSUM charts for our analysis.

Methodology

Pre- Analysis of given dataset:

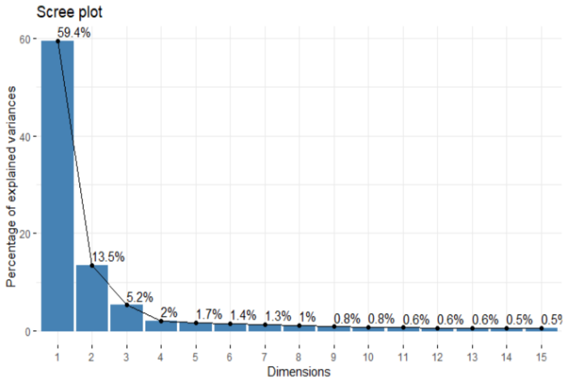
Analysis carried out on the dataset reveals that we have a unimodal function. Also, as this dataset is obtained from the manufacturing process, we should maintain the original relationship between the variables. This means that original magnitude of variables plays a significant role. Thus, PCA was carried out on the covariance matrix instead of correlation matrix.

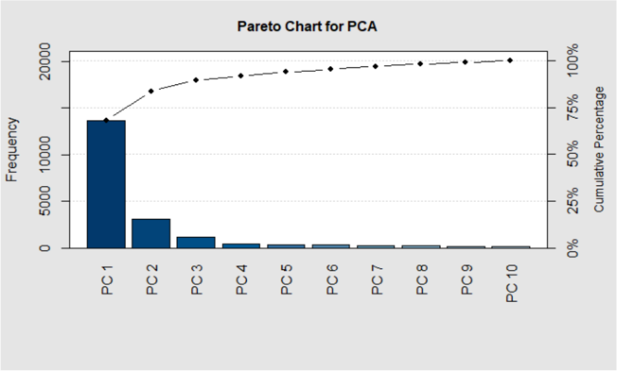
Data Reduction - Principal Component Analysis (PCA):

In order to reduce the dimension of our dataset, PCA was performed on the covariance matrix. The eigenvalue-eigenvector pairs of the covariance matrix were used to transform the original variables into a set of variables having maximum variance along their direction. Although, just transformation was not enough as it yielded 209 PCs which is equal to the number of Variables in the original dataset. Scree plots and Pareto charts need to be used to identify the 'vital few' PCs explaining the majority of the variability.

A Scree plot is a graph of eigenvalues plotted against the index of principle component. We look out for an elbow formation in the graph and retain the PCs equal to index of PCs at the bend. The remaining PCs below the elbow bend are small in magnitude and can be ignored for computational convenience of the objective.

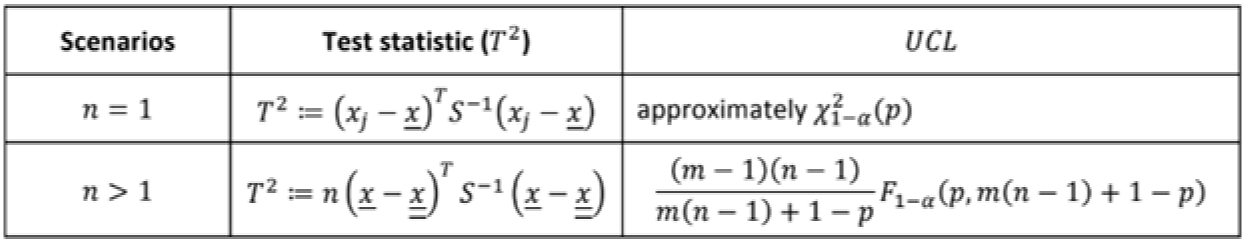
A Pareto Chart is a cumulative plot of variability against index of PCs. The number of PCs to retain is chosen by the variability we need in our transformed variable dataset (say 80%). Similar to Scree plot, the initial few PCs are retained while the rest are ignored. First 4 PCs were chosen which explained about 80.09% of the dataset's original variability. These 4 PCs would be used for training the data on T^2 chart and m-CUSUM chart.





Multivariate Detection using T^2 chart

Hotelling T^2 Chart is used to start our multivariate detection. We use Case III (a) from the T^2 chart where sample size n = 1 and mean and covariance matrix are not known. In this case, the T^2 statistic approximately follows a χ^2 distribution with the DOF equal to number of PCs used. The T^2 chart is used to identify and eliminate data points of a spiky nature. We used α=0.0027 for 3-sigma control limits. T^2 chart calculated utilizing the R-Code is used to identify some of the out of control data, and then move on to m-CUSUM.



Scenarios	Test statistic (T^2)	UCL
$n = 1$	$T^2 := \left(x_j - \underline{x}\right)^T S^{-1}\left(x_j - \underline{x}\right)$	approximately $\chi^2_{1-\alpha}(p)$
$n > 1$	$T^2 := n\left(\underline{x} - \underline{\underline{x}}\right)^T S^{-1}\left(\underline{x} - \underline{\underline{x}}\right)$	$\frac{(m-1)(n-1)}{m(n-1)+1-p}F_{1-\alpha}(p, m(n-1)+1-p)$

Multivariate Detection using m-CUSUM Charts

After the T^2 charts, we use the m-CUSUM charts to eliminate the small sustained mean shifts which cannot be detected by T^2 charts. In its essence, we can either use m-CUSUM or m-EWMA without having any benefits over the other. But as m-CUSUM is more popular within the industry, it was employed at our use. The value of offset constant *k* was set at 1.5 as we expect to detect a mean shift of unit magnitude in terms of statistical distance. Similarly, we have kept value of h which is used in calculating the m-CUSUM to 6. We have used the Information from Hamed (2016) and Santos-Fernandez (2017) in selecting the above mentioned values. α is set to 0.0027.

$$MC_i = \max\{0, \left(\mathbf{C}_i^T \mathbf{\Sigma}^{-1} \mathbf{C}_i\right) - \mathrm{k} \cdot n_i\}$$

$$\mathbf{C}_i = \sum_{j=i-n+1}^{i} X_j - \mu_0$$

$$n_i = \begin{cases} n_{i-1} + 1 & \textit{when } MC_{i-1} > 0, \textit{keep accumulating.} \\ 1 & \textit{otherwise, start CUSUM over.} \end{cases}$$

Iterations between T^2 and m-CUSUM

Now, we iterate between the two types of charts until no out-of-control points are observed on either of the charts. Once this condition occurs, we could be definite that the outlying data has been eliminated and the remaining data would most accurately describe the underlying distribution of its parent process.

Results and Conclusions

As described earlier, by utilizing 4 Principal Components we were able to explain nearly 80.1% of the variability in the system. Using these principal components, we ran iterations in R to distinguish between the in and out-of-control data. For accomplishing the given task, we iterated for both T^2 and m-CUSUM till there were no more out-of-control points removed detected by any of the charts. The results are as summarized below:

Iteration Number	Cycle Number	Control Chart Used	Out of Control Points
1	1	T^2	12
	2	T^2	6
	3	T^2	2
	4	T^2	0
	Total	**T^2**	**20**
1	1	m-CUSUM	61
	2	m-CUSUM	19
	3	m-CUSUM	2
	4	m-CUSUM	7
	5	m-CUSUM	1
	6	m-CUSUM	0
	Total	**m-CUSUM**	**90**
2	1	T^2	6
	2	T^2	2

	3	T^2	0
	Total	**T^2**	**8**
2	1	m-CUSUM	0
	Total	**m-CUSUM**	**0**

From the above table, it can be observed that a total of 118 were removed from the dataset after carrying out Phase 1 analysis utilizing T^2 and m-CUSUM charts in 2 iterations each. This left us with 434 data points which amounts to 78.6% of the original data.

We also present the initial and final charts in the T2 and m-CUSUM iterations, as a graphical representation of the before and after out-of- control data removal. The full set of graphs that detail all iterations of data removal are shown in the Appendix. The x-axis represents observations, and the y-axis is the T2 statistical index. As shown in Figure, outliers are found to locate in the beginning and tail section, and we will remove all these out-of-control data without further inspection since we don’t have the information of what kind of manufacturing data we are analyzing. The process will keep iterating until no outlier is found from control charts.

References

Yu, Ding, "ISEN614_CoursePack" ISEN 614-600 Handout. Handout Retrieved from https://tamu.blackboard.com

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr

Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

Scrucca, L. (2004). qcc: an R package for quality control charting and statistical process control. R News 4/1, 11-17.

Edgar Santos-Fernandez(2013). Multivariate Statistical Quality Control Using R. Springer, 14. URL http://www.springer.com/statistics/computational+statistics/book/978-1-4614-5452-6.

Lê S, Josse J, Husson F (2008). "FactoMineR: A Package for Multivariate Analysis." Journal of Statistical Software, 25(1), 1–18. doi: 10.18637/jss.v025.i01.

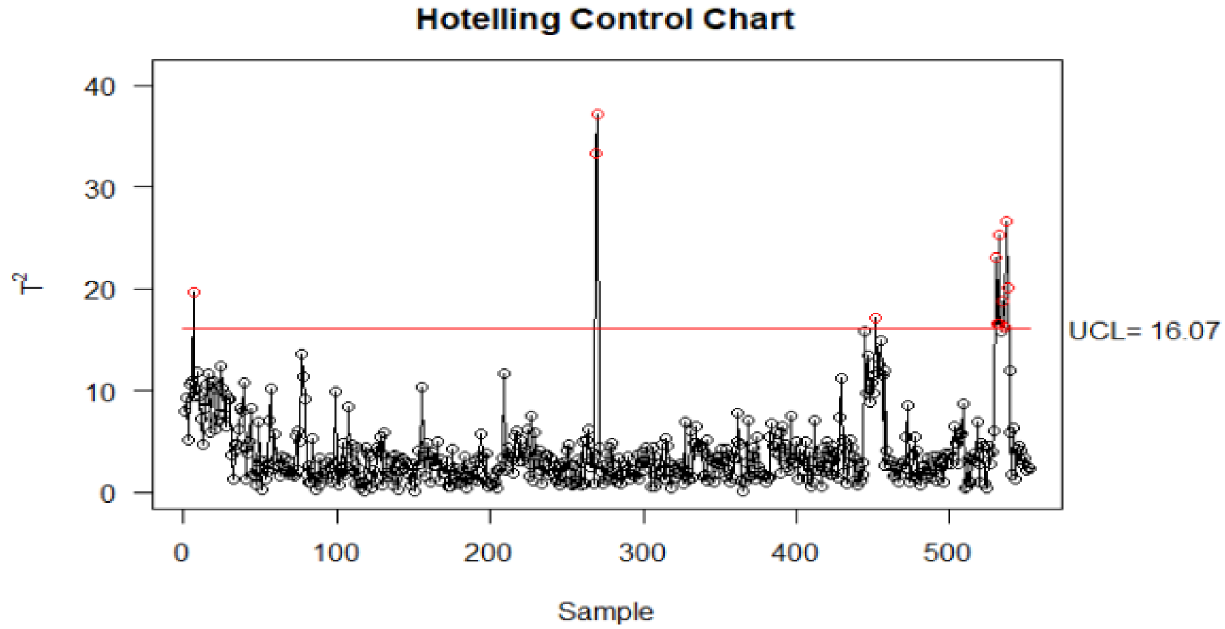
H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Joseph J. Pignatiello Jr. & George C. Runger (1990) Comparisons of Multivariate CUSUM Charts, Journal of Quality Technology, 22:3, 173-186, DOI: 10.1080/00224065.1990.11979237

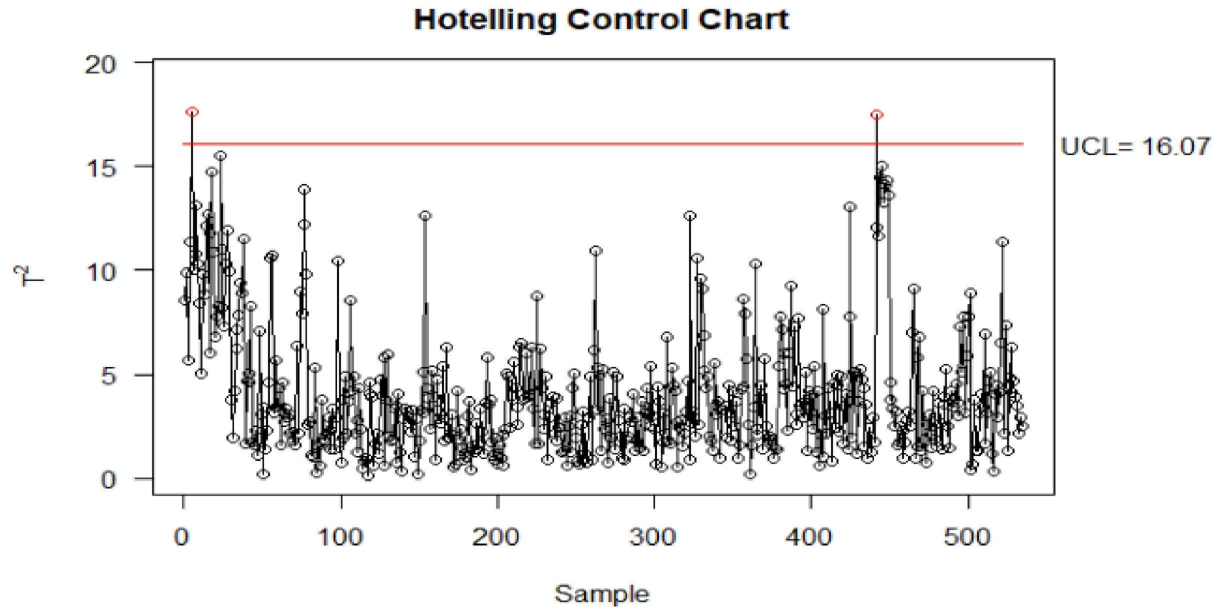
Appendix

Hotelling T^2 charts for Phase 1 Analysis: 1st Iteration

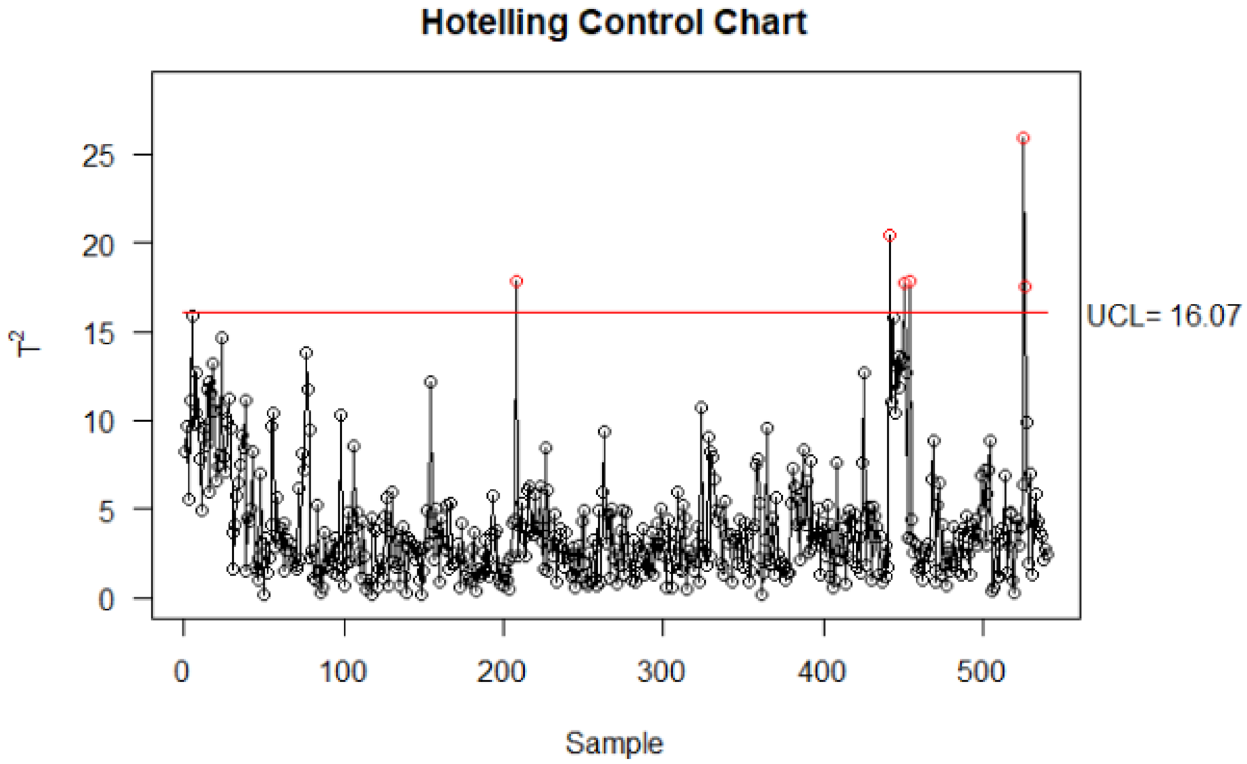
Cycle 1: 12 points removed



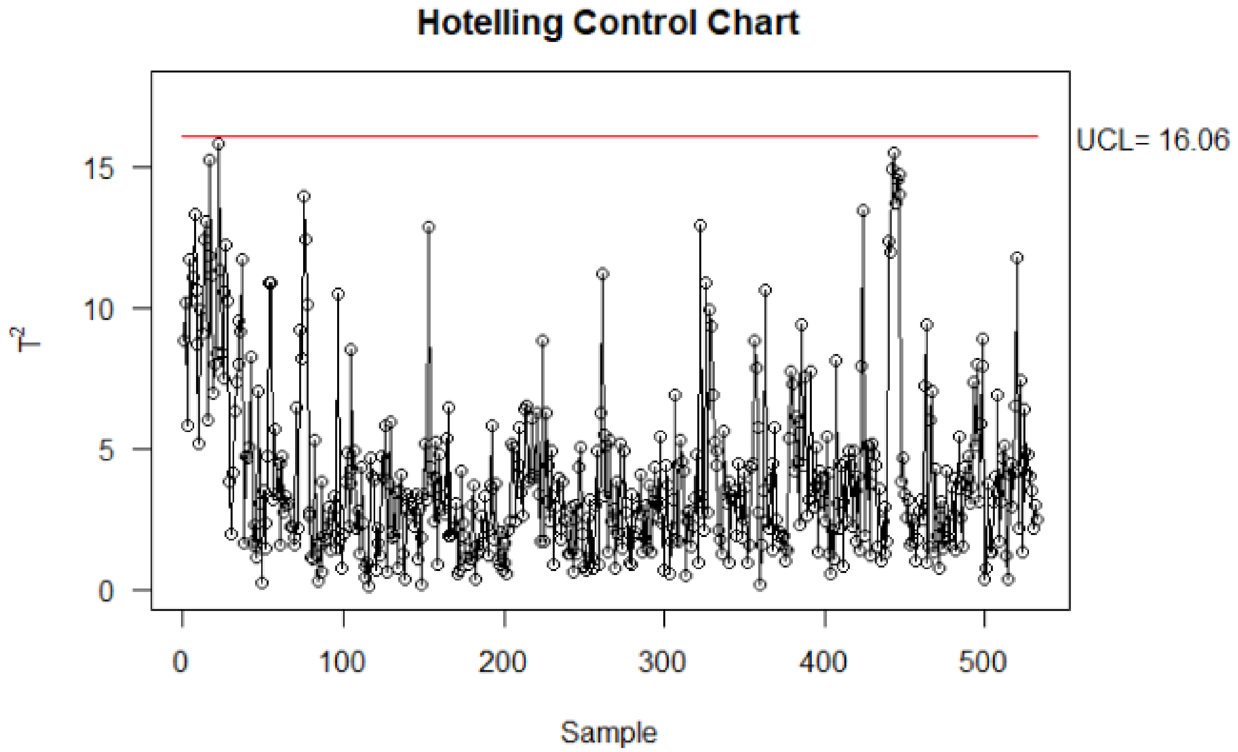
Cycle 2: 6 points removed



Cycle 3: 2 points removed

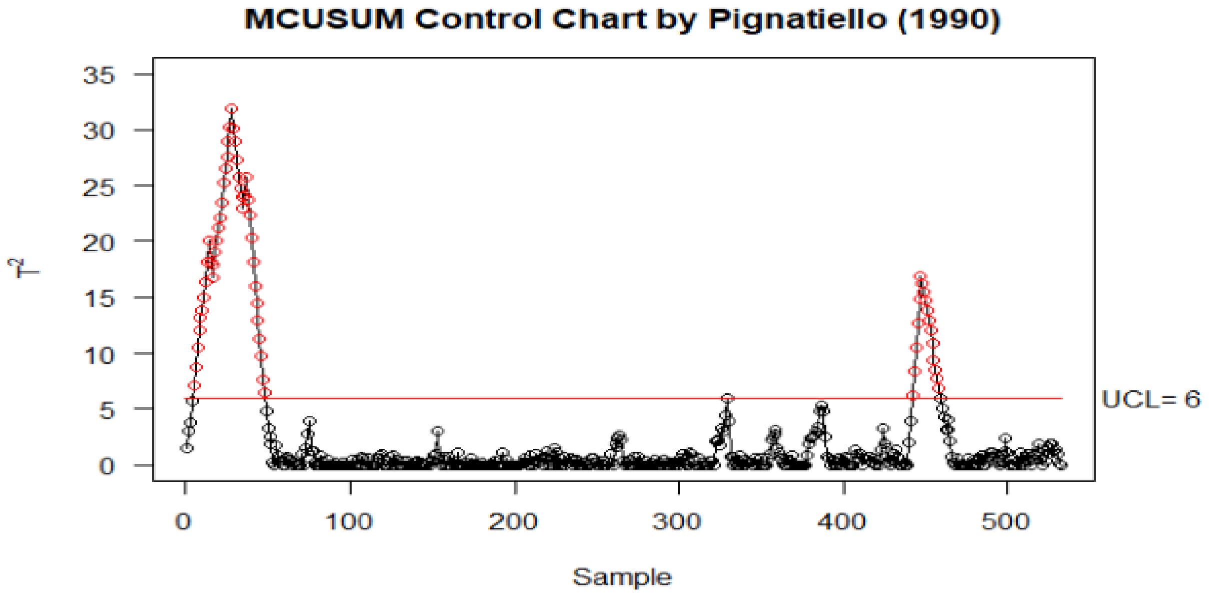


Cycle 4: 0 points removed

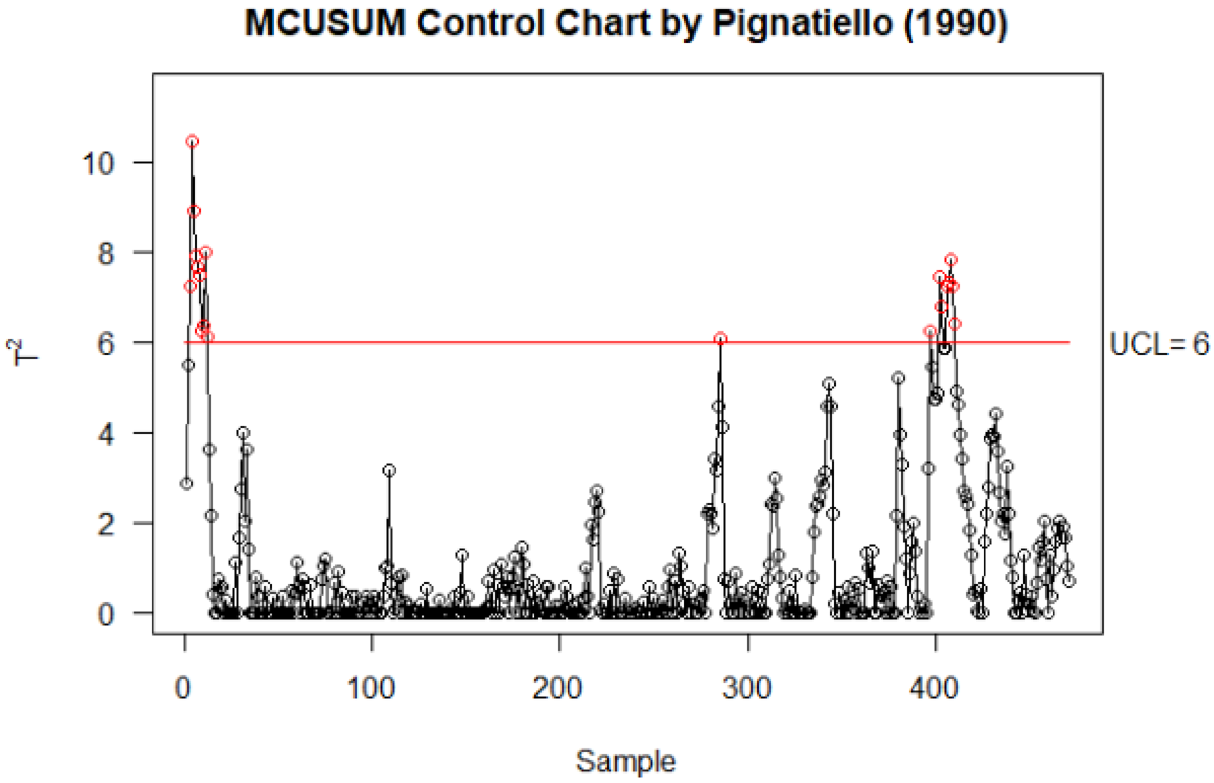


m-CUSUM charts for Phase 1 Analysis: 1st Iteration

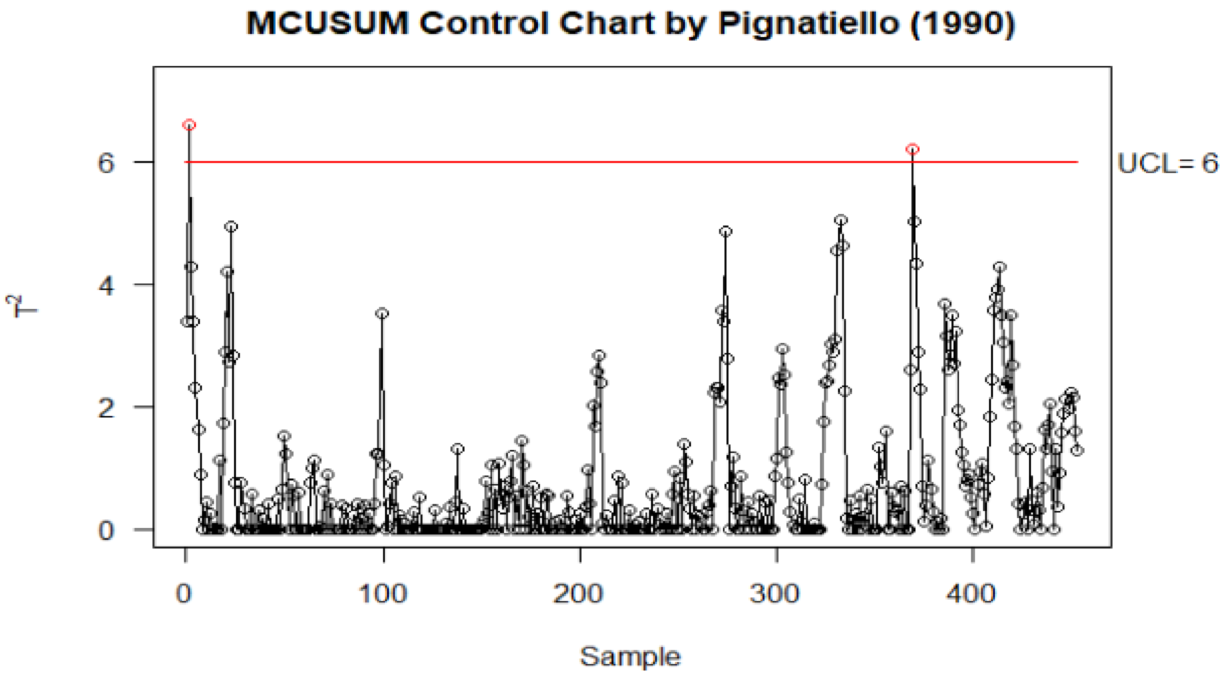
Cycle 1: 61 points removed



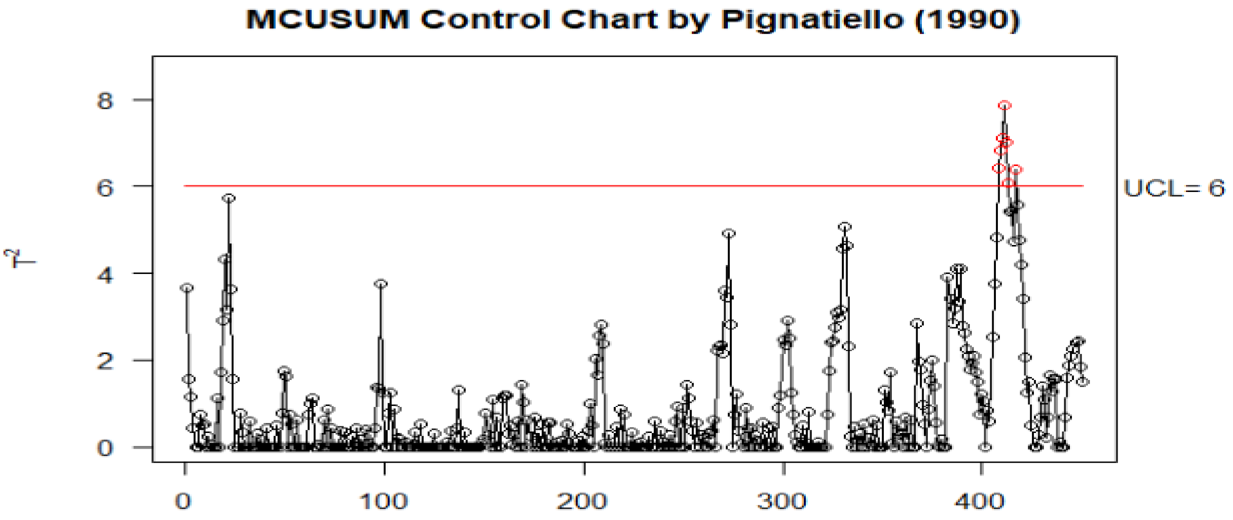
Cycle 2: 19 points removed



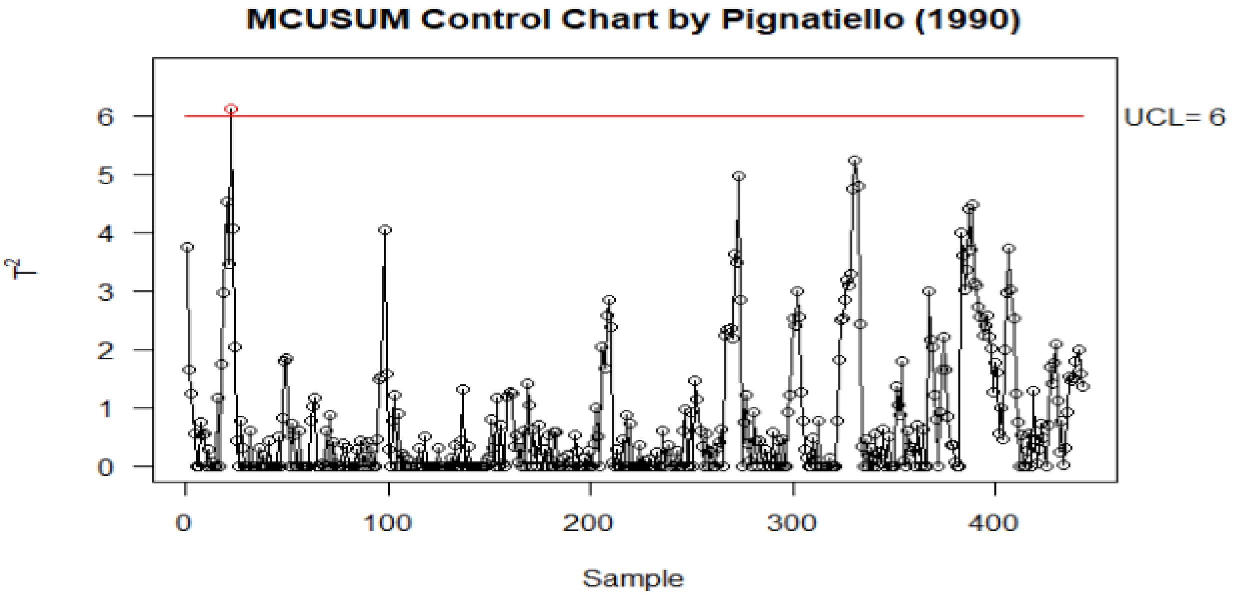
Cycle 3: 2 points removed



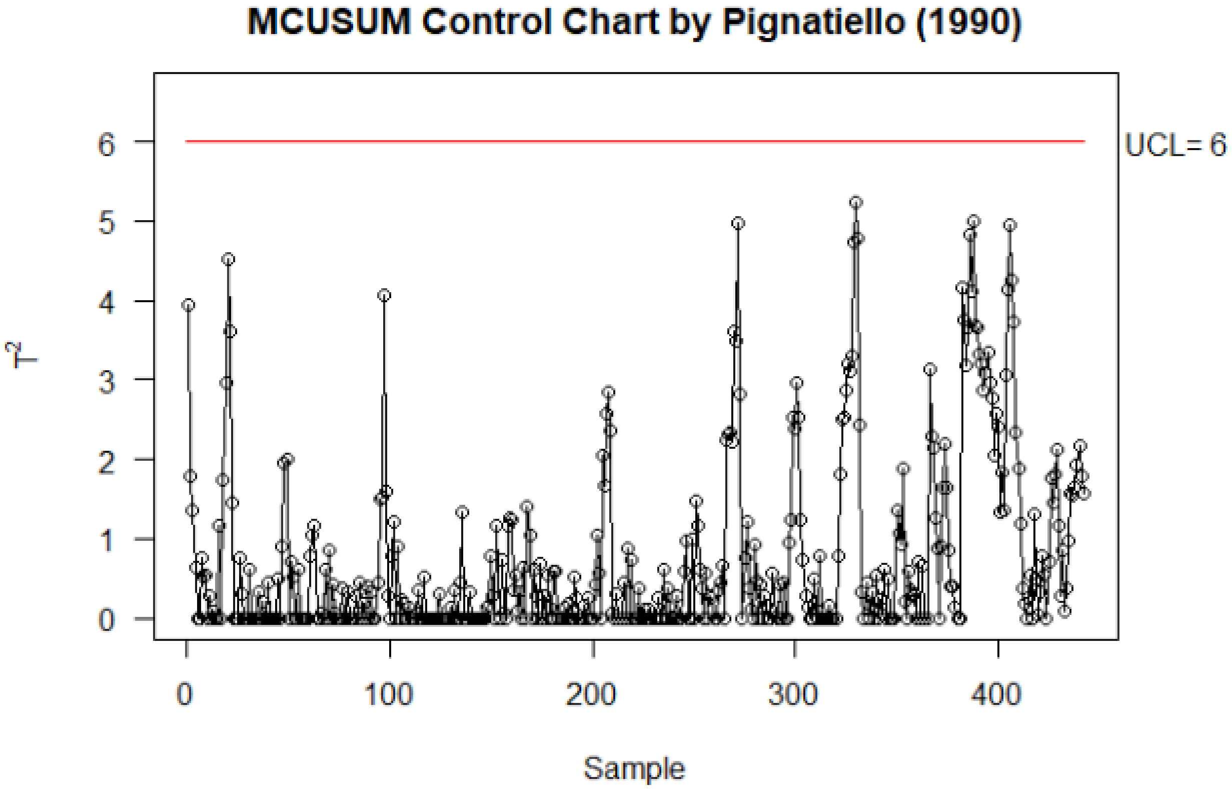
Cycle 4: 7 points removed



Cycle 5: 1 point removed

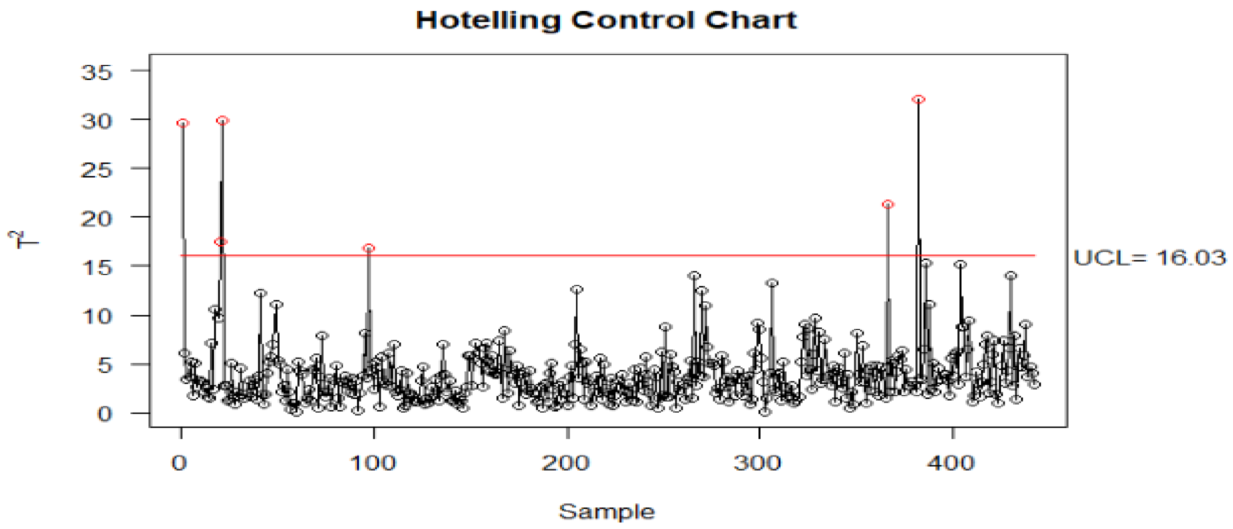


Cycle 6: 0 points removed

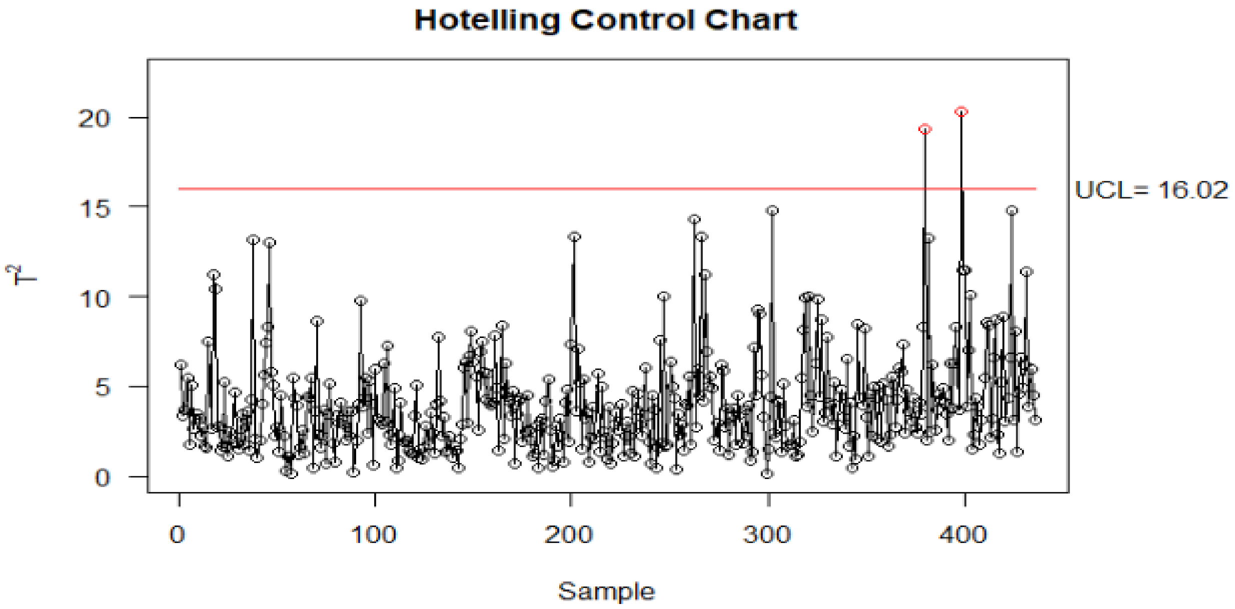


Hotelling T2 charts for Phase 1 Analysis: 2nd Iteration

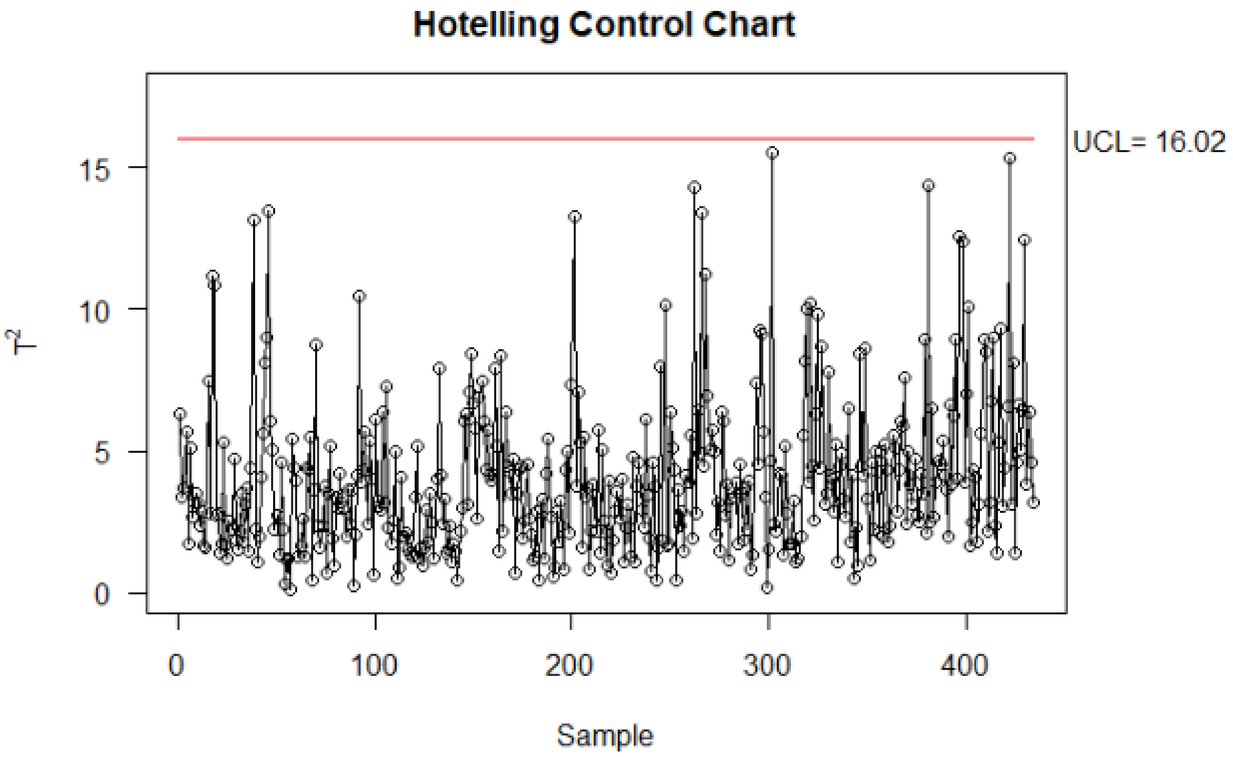
Cycle 1: 6 points removed



Cycle 2: 2 points removed

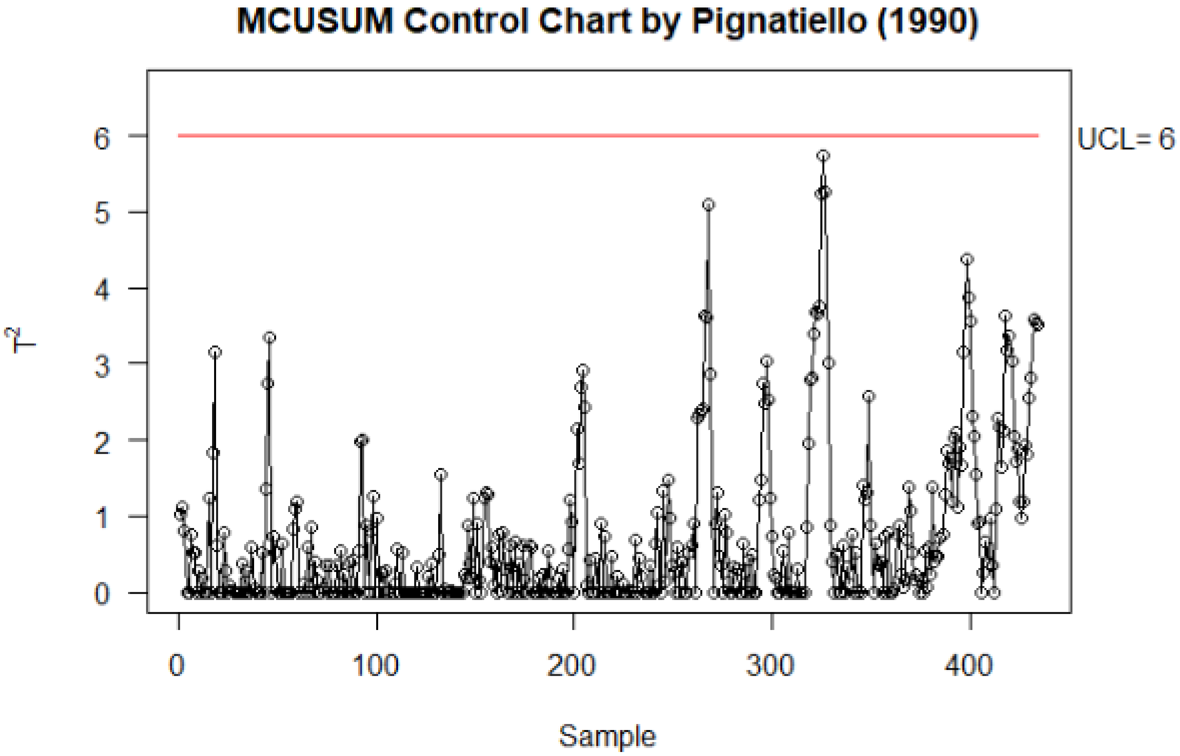


Cycle 3: 0 points removed

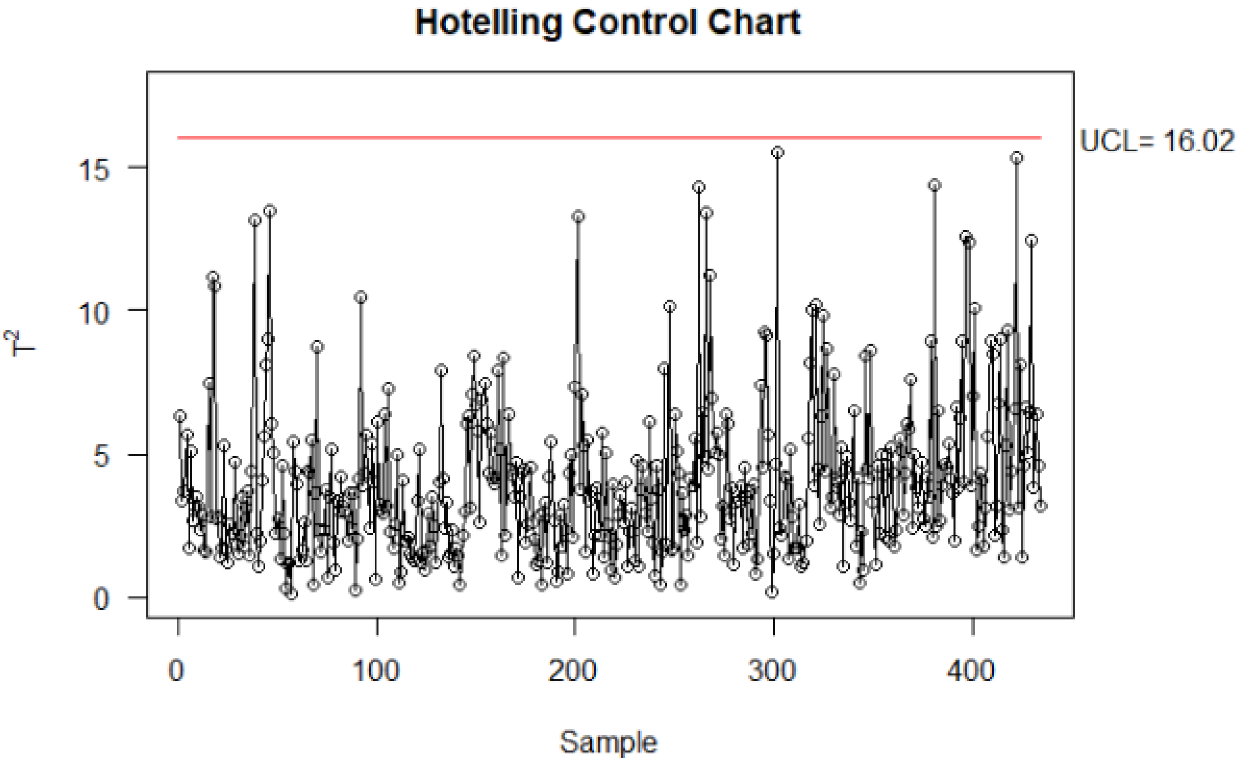


m-CUSUM charts for Phase 1 Analysis: 2nd Iteration

Cycle 1: 0 points removed



Final Hotelling T2 chart after Phase 1 Analysis


Hotelling Control Chart
UCL= 16.02
T^2
Sample