ISEN 614: Advanced Quality Control Project

Phase I Analysis of High Dimensional Data Using Multivariate Control Statistics

By

Team 2

Teja Gambhir (UIN: 525001619)

Pulkit Jain (UIN: 625005181)

Executive Summary

The purpose of this project is to apply statistical methods of quality control to conduct a Phase I analysis on given multivariate data by isolating the out-of-control data from the in-control data. The end goal is to identify the in-control mean and variance. MATLAB was used to analyze the data and solve for the in-control parameters as well as to generate the graphs.

Principal Component Analysis (PCA) was performed for two cases: 1) the data characteristics have similar physical meaning or their original variances are critical and 2) the data characteristics have different physical meaning so that their variances can be compared on the same level. Evaluating the variance plot for the original data in-conjunction with the Pareto plots for the reduced data helped over-rule the second case.

After ruling out the second case, the parameters for Hotelling T^2 and m-CUSUM control charts were determined based on their coherence with individual PC charts. Since information about the original dataset is sparse, comparing individual PCs with performances of T^2 and m-CUSUM helped establish viable control parameters after a prima facie analysis of the data. Following this, a Hotelling T^2 control chart was used in conjunction with the m-CUSUM control chart to identify the in-control data values through multiple iterations until both charts were found to be in-control.

The resulting values were then used to calculate the in-control mean and covariance matrices. The final results lead us to conclude that despite performing PCA on the covariance matrix, non-zero values in the non-diagonal elements of the covariance matrix were observed.

The project has increased our understanding of how the statistical methods we learned in this course can be applied to understand a given dataset. Even when the data is reduced using PCA, multiple iterations of the T^2 and m-CUSUM are needed simultaneously to identify all the out-of-control data because the T^2 control chart is better at identifying large changes while the m-CUSUM is better at detecting small, sustained changes in the mean shift.

In industry, relevant information about the data may not be available, which will require a subjective or perhaps even an intuitive understanding of the data. Since we used different methods of detection in-conjunction, we saw the relative advantages in utilizing one method over another. Overall, the project forced us to rely on our own judgment and to rationalize all methods used, similar to a possible industry scenario.

Approach and Justification

The given data set consists of 209 characteristics and 552 observations with a sample size of one. Because there are a large number of characteristics, the obvious preference is to use PCA and multivariate control charts instead of several univariate charts in order to decrease noise and to identify characteristics that contribute the most to the variance. The physical units of the data are unknown which makes it necessary to analyze two scenarios:

1. Either the data characteristics have similar physical meaning or
2. The data characteristics have different physical meaning

Before considering the cases, the mean and variance of the original data set were calculated. Then considering the first case, the covariance matrix was calculated and considering the second case, the correlation matrix was calculated. For the first case, PCA was performed on the covariance matrix while for the second case PCA was performed on the correlation matrix.

After calculating the Principal Components (PCs), Scree plots, Pareto plots and MDL plots were used simultaneously in order to identify the highest contributing PCs. Finally, the Pareto plot was used to infer the proportion of the cumulative variance to total variance the PCs yield for the two scenarios.

Selection of Principal Components (PCs)

When working with high dimensional data, there is greater uncertainty, which leads to poor sensitivity for detecting changes in the mean or correlation among individual components. The purpose of selecting PCs is to reduce the data set by identifying the PCs that account for at least 80% of the proportion of cumulative variance to total variance.

Reducing the dimension of data set using PCA decreases covariance, increases ease of change detection, and selects only the PCs with the highest proportions of information, this ensures that the data set is well represented. This also helps alleviate the curse of dimensionality.

Implementation of PCA on the Covariance Matrix:

In order to calculate the PCs, the mean and variance of the original dataset were first calculated. The eigenvalues and eigenvectors of the original dataset were then established from the variance matrix. Following this, the PCs for the covariance matrix were found using:

$$y_i = {e_i}^T * (x - \mu_x)$$

where y_i represents the PCs, $e_i$are the eigenvectors, x is the j_{th} column from the original data set, and μ_x is the estimated mean vector which was calculated as:

$$\widehat{\mu_x} = \bar{x} = \frac{1}{m}\sum_{j=1}^{m} x_j$$

where m is the total number of observations (552).

The Minimum Description Length (MDL) was used to identify if there are any regularities in the data and to compress those regularities so that fewer number of the data values represent the whole data set. The higher the regularities in the data, the more the data can be compressed, so fewer values are needed to represent the whole data set.

The results from the MDL plot show that the minimum occurs nearly at 36 PCs. However, the MDL retains too many eigenvalues, which is why further analysis with the Scree plot and Pareto plot was performed to deduce if a lower number of PCs can be used.

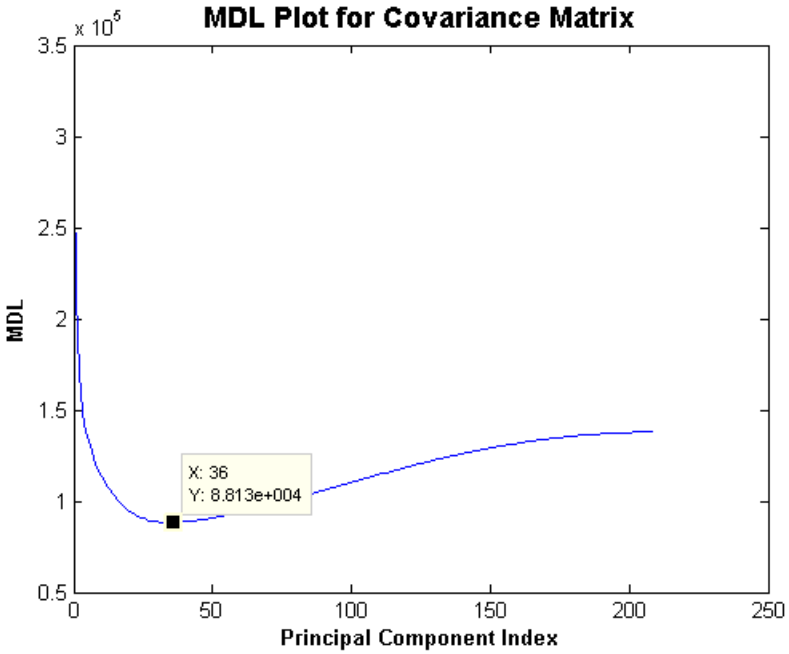

Figure 1: MDL Plot for the Covariance Matrix

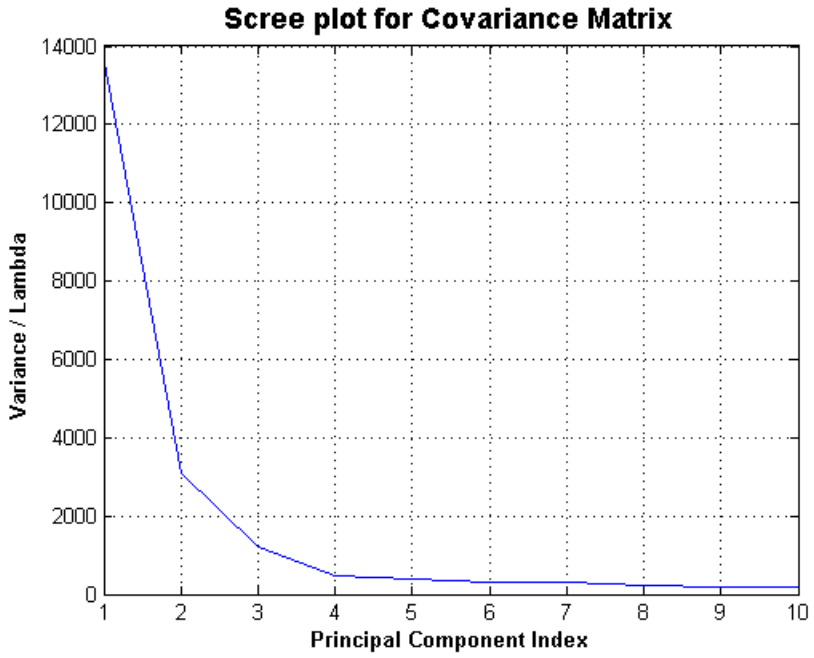

Figure 2: Scree Plot for PCA on the Covariance Matrix

The Scree plot helps to intuitively identify the principal components with the highest contribution to the variance. The Scree plot for the PCA of the covariance matrix yielded values with high contributions:

Index Value (i)	Eigenvalue / Variance
1	13,700
2	3105
3	1205
4	461.2
5	390.1

Table 1: Highest five Eigenvalues for PCs after PCA on the Covariance Matrix

The eigenvalues retained for PC1, PC2, and PC3 are significantly higher when compared to PC4 and PC5. Although there is a bend in the plot at index value 2, the plot does not start to level until PC4. The difference between the eigenvalues for PC4 and PC5 is much less and it can be observed from the graph that the values seem to decrease only slightly for PCs greater than 5, indicating that the plot has begun to level. Hence, it is sufficient to conclude that PC1, PC2, and PC3 are the 'vital few' that represent the data set.

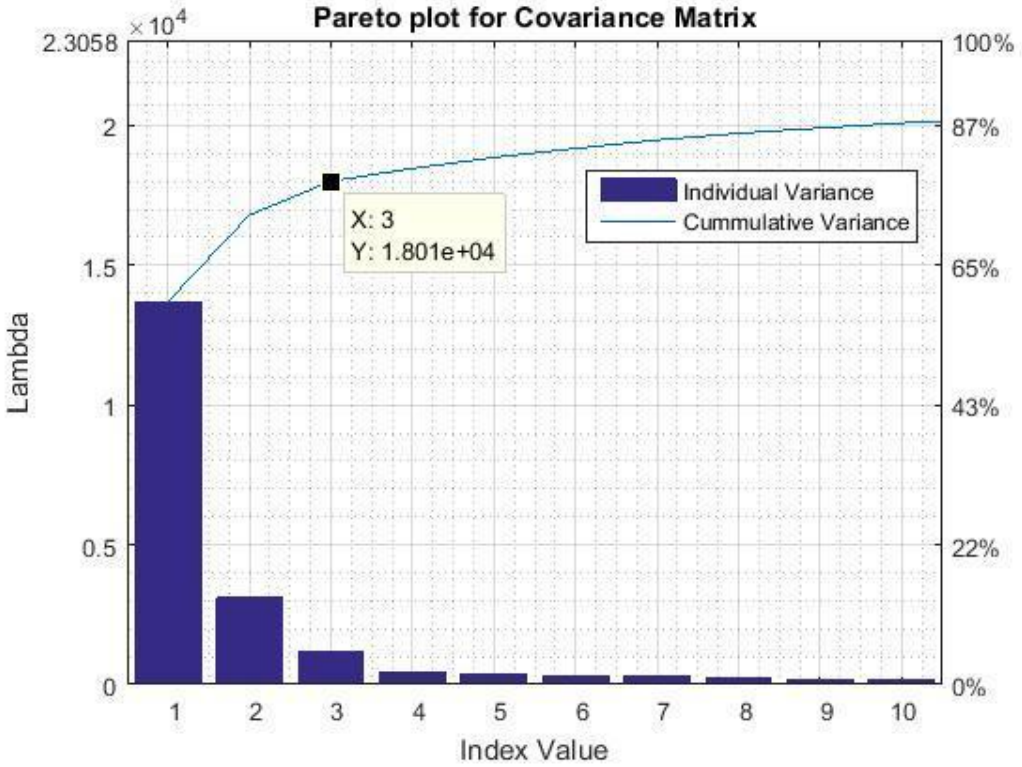

Figure 3: Pareto Plot for PCA on the Covariance Matrix

As mentioned before, we concede that the proportion of the cumulative variance of the first few PCs and their eigenvalues should account for at least 80% of the total variation in the data set. The Pareto

plot shows that PC1, PC2, and PC3 contributed to about 78%, which is very close to 80% and therefore sufficient. Subsequent PCs fail to contribute a significant percent. Hence, the results of the Pareto plot also agree with the conclusion from the Scree plot to choose PC1, PC2, and PC3 for further analysis.

PCA on the Correlation Matrix:

The number of PCs to retain for the correlation matrix were evaluated with a similar method. The PCs for the correlation matrix were found using:

$$y_i = (e_i^{\rho})^T * (V^{-0.5}) * (x - \mu_x)$$

where y_i represents the PCs, $e_i$are the eigenvectors of the correlation matrix ρ, V is the matrix containing the diagonal elements from the covariance matrix, x is the j_{th} column from the original data set, and μ_x is the estimated mean vector calculated as:

$$\widehat{\mu_x} = \bar{x} = \frac{1}{m}\sum_{j=1}^{m} x_j$$

where m is the total number of observations (552).

Evaluating the MDL for the correlation matrix, it was discovered that 24 PCs must be retained and because this is still a large amount, further analysis using Scree and Pareto plots was done.

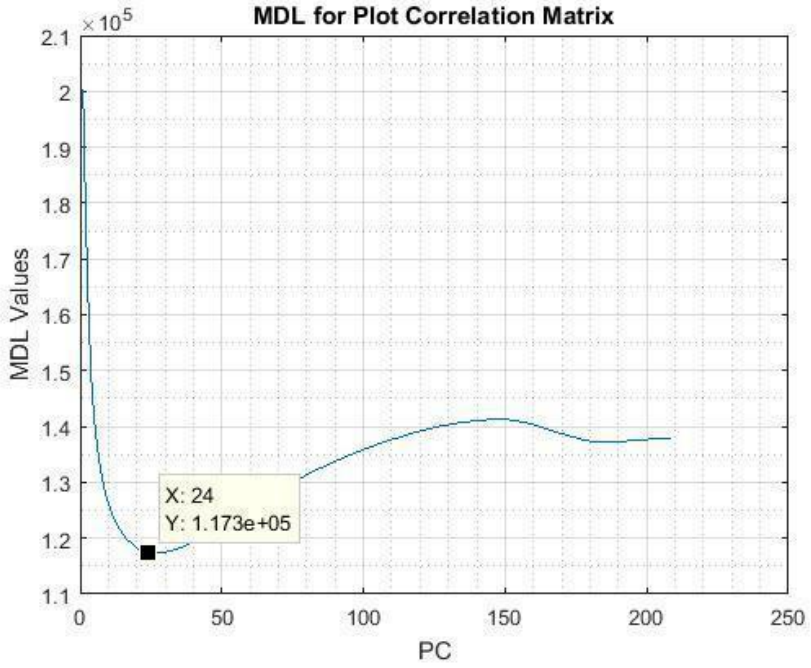

Figure 4: MDL plot for the Correlation Matrix

The Scree plot for the correlation matrix shows increased number of components when compared to the components retained by the Scree plot for the covariance matrix. Unlike the covariance matrix, the output for the correlation matrix does not start to level off until PC16.

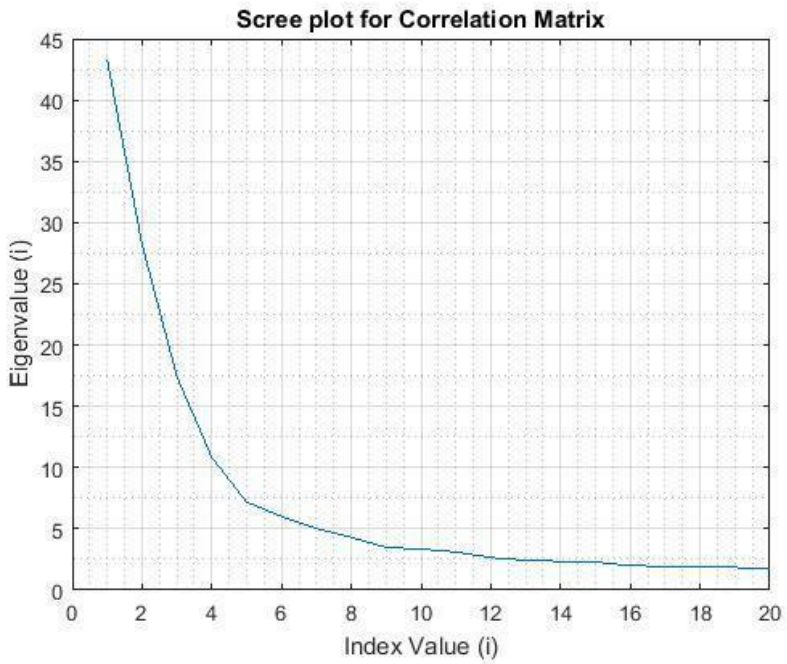

Figure 5: Scree Plot for PCA on the Correlation Matrix

Consecutive eigenvalues corresponding with index values of 1 to 16 have a fairly large difference. The plot seems to level off only after the first 16 index values. Therefore, at least the first 15 PCs will need to be retained, which is 12 more PCs than for the covariance matrix. Further evaluation with the Pareto plot for eigenvalues of correlation matrix is needed to reach a conclusion.

Index Value (i)	Eigenvalue / Variance
1	43.31
2	28.2
3	17.42
4	10.82
5	7.81
10	3.338
14	2.338
16	1.989
17	1.94

Table 2: Corresponding PCs and Eigenvalues after PCA on the Covariance Matrix

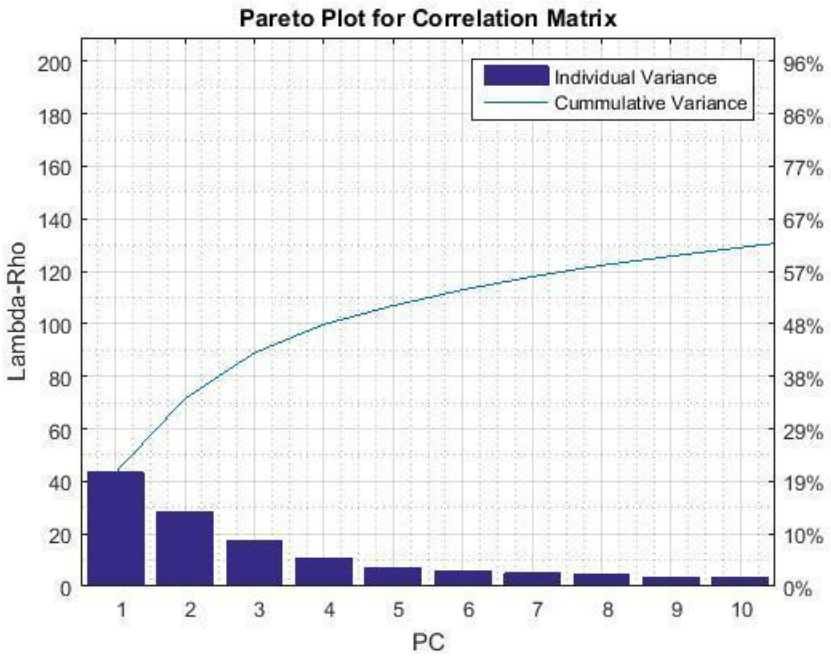

Figure 6: Pareto Plot for the Correlation Matrix

The Pareto plot for the correlation matrix shows that even the first ten PCs account for less than 67% of the proportion of the cumulative variance. The number of PCs needed to reach a cumulative variance of 80% is much higher. This is expected because the correlation matrix is normalized, which reduces the relative difference in covariance. Therefore, the significance placed on PC1 for the correlation matrix is much less than the significance placed on PC1 for the covariance matrix.

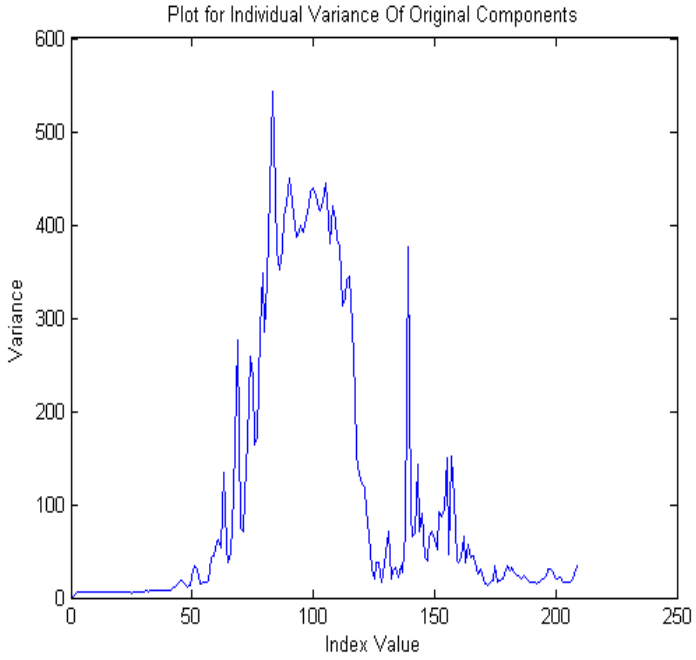

Figure 7: Plot of the Variance of the Original Dataset

In order to understand why the PCA drew such results, the variance of the original dataset was calculated (Figure 7). This deeper analysis of the data revealed that a wide range of differences in variance of the original components are present. If the components with smaller variance are relatively stable, then normalizing the variances will reduce the impact of the high varying components. Because the high varying components are what cause the system to go out-of-control, it is crucial to retain their influence on the data.

Keeping this in mind, it is necessary to forgo further analysis using the PCs from the correlation matrix. Recall that the aggregate variance of nearly 78% was achieved with just three PCs in the covariance matrix, which means a smaller number of eigenvalues in the covariance matrix account for a greater proportion of the total variance. Therefore, we choose to proceed with further analysis using only the first three PCs from the covariance matrix.

Isolation of Out-of-Control Data from In-Control Data

In the next step, analysis was done using the first three PCs of the covariance matrix by plotting the PC charts using 3-sigma control limits and identifying major out-of-control segments. T^2 and m-CUSUM control charts were then used simultaneously on this reduced number of PCs, to isolate the out-of-control data from the in-control data.

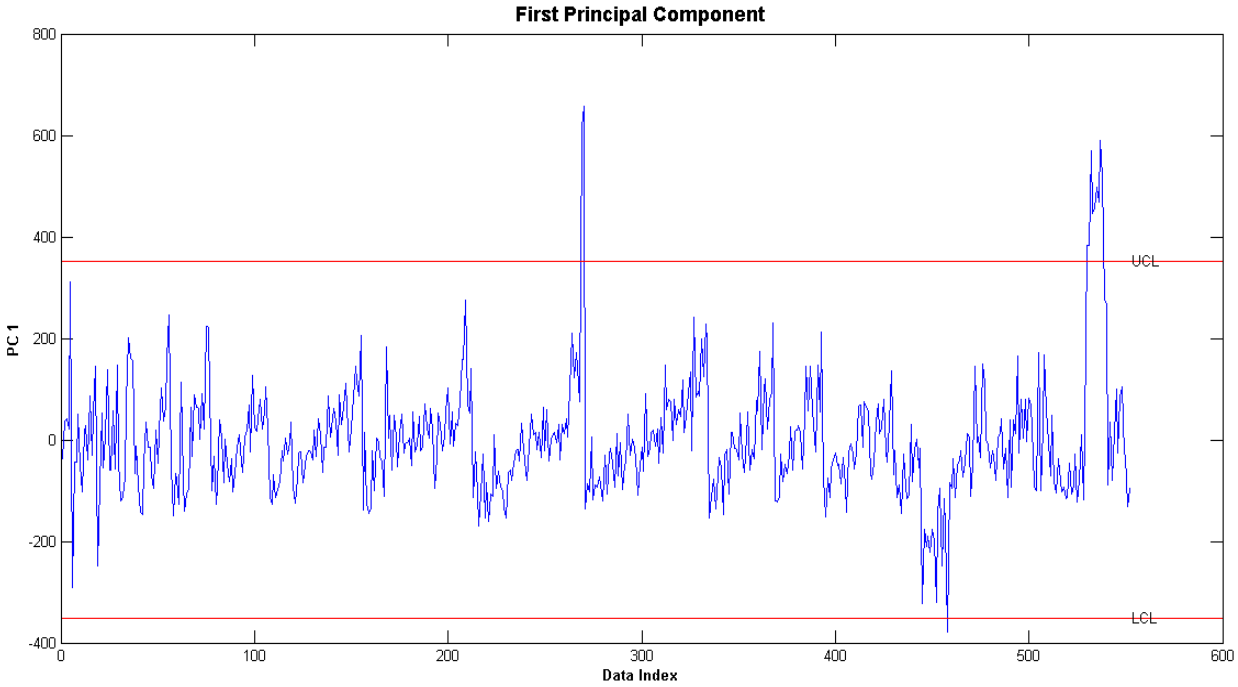

Figure 8: PC1 Chart for PCA on the Covariance Matrix

When compared to PC2 and PC3, PC1 has the most range for the y-axis (-800 to 400) which is expected because PC1 contains values that contribute the most to the variance.

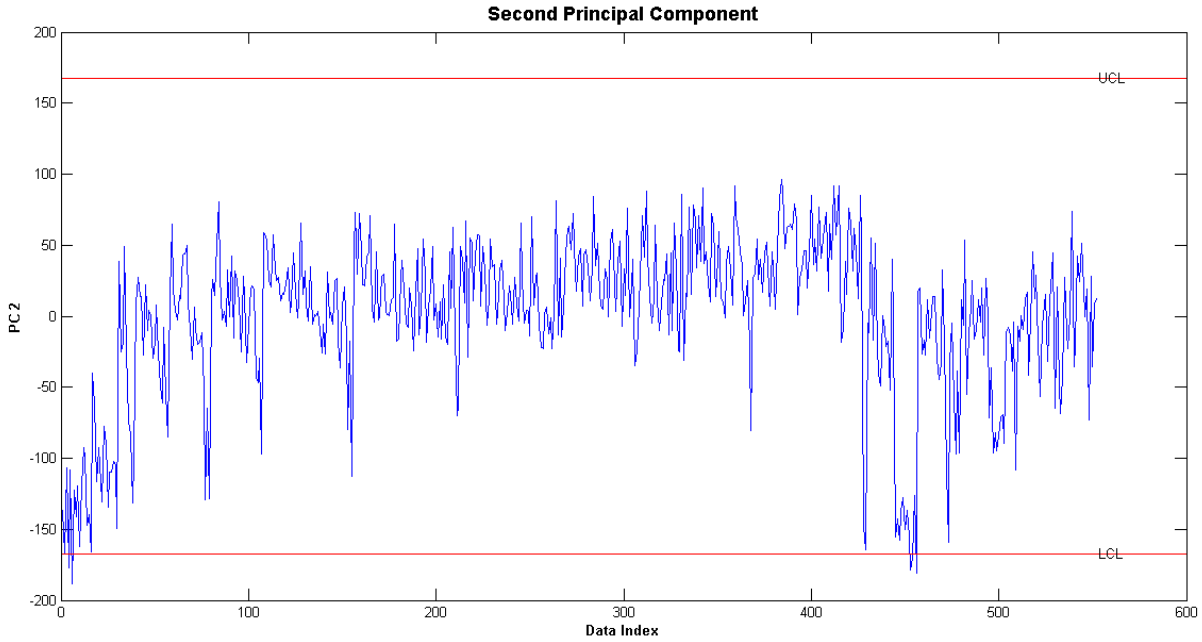

Figure 9: PC2 Chart for PCA on the Covariance Matrix

Compared to PC1 and PC3, PC2 has the most fluctuations. The fluctuations at the start of the graph can be attributed to a 'warm-up' period. The fluctuations seen near the end of the chart, which also have points going out-of-control, are followed by a slight change in mean shift. In contrast, all points are in-control for PC3.

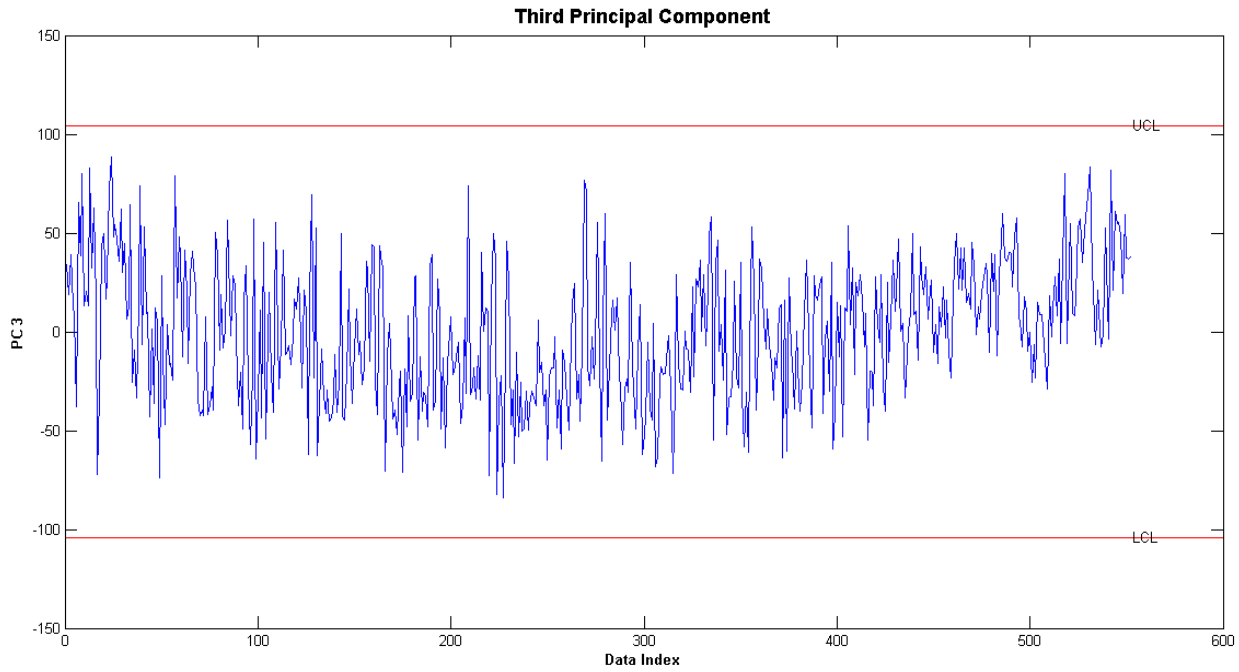

Figure 10: PC3 Chart for PCA on the Covariance Matrix

Selection of Control Chart Parameters

In order to identify whether detecting a mean shift of 2-sigma or 3-sigma will be more prudent, two different m-CUSUM charts were compared. The m-CUSUM control charts were generated using the data from the first three PCs.

Using Monte Carlo simulation for an Average Run Length for Type I error (ARL_0) of 200, the UCL was set as 5.48 for the 2-sigma and the 3-sigma mean shifts. Increasing ARL_0 decreases the probability of getting two or more points out-of-control, reducing the control chart's efficiency. Therefore, since there are a total of 552 observations, an ARL_0 of 200 will increase the likelihood of observing at least two out-of-control points.

The m-CUSUM control charts were compared with the PC charts (Figures 8-10) to see if the data behaved in a similar fashion. When compared with the plots for the first three PCs, the number of out-of-control points were more in accordance with the m-CUSUM chart for the 3 sigma mean shift. It appears that the smaller the mean shift, the more out-of-control points there are.

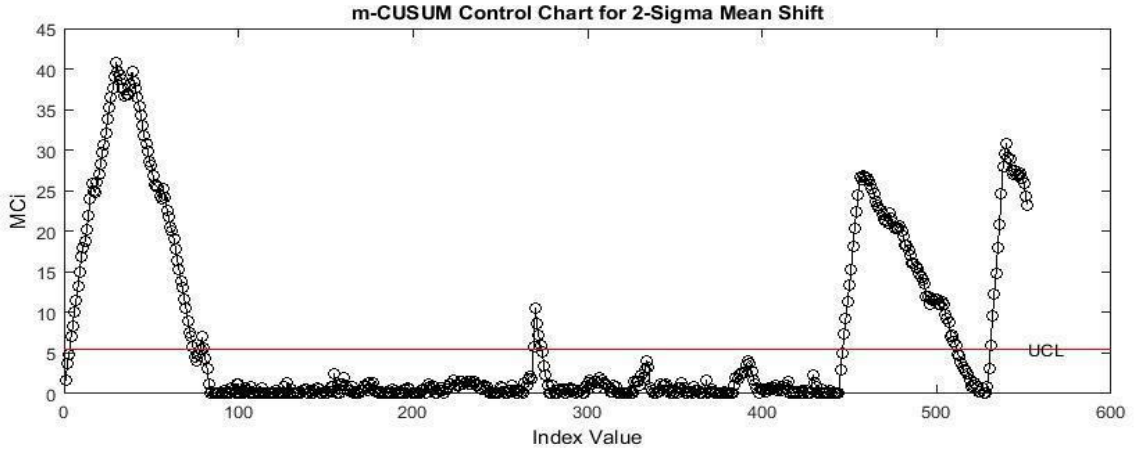

Figure 11: m-CUSUM Control Chart for Detection of a 2-Sigma Mean Shift

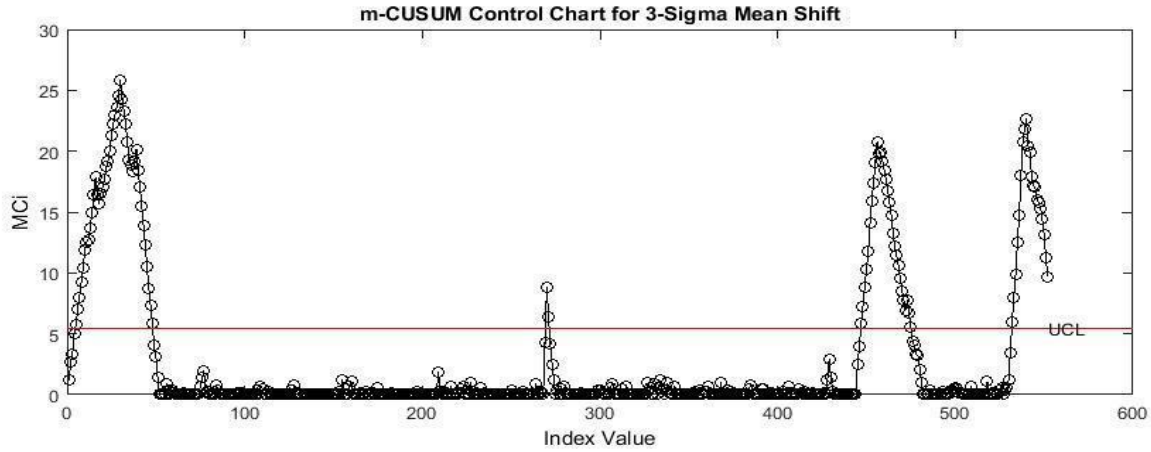

Figure 12: m-CUSUM Control Chart for Detection of a 3-Sigma Mean Shift

If the first 90 values are observed on the plot for PC2 (Figure 9), approximately 38 points out of 90 appear to be out-of-control or in transient phase. Figure 11 for the 2-sigma mean shift shows that between 4 to almost 90 points are out-of-control, but Figure 12 for the 3-sigma mean shift identifies that around 5 to 48 points are out-of-control, which most closely resembles the out-of-control trend for PC2.

The T^2 control chart control chart was used for additional comparison. To set an alpha error for the T^2 control chart which yields an ARL_0 of 200, the inverse of the ARL_0 (or 200^{-1}) can be taken as an approximation. Doing so gives an alpha error of 0.005. The plot for the T^2 control chart shows a similar range of out-of-control data values as observed on the m-CUSUM for the 3-sigma mean shift. This is expected because the T^2 control chart is able to detect sustained large changes. Hence, we elect to proceed with the detection of a mean shift of 3-sigma.

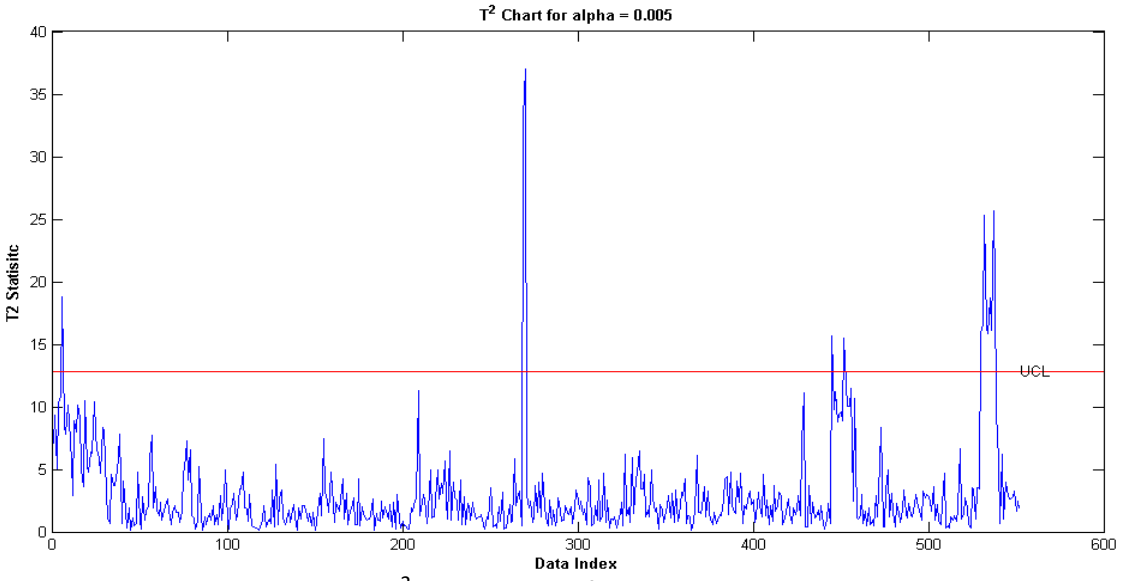

Figure 13: T^2 Control Chart for the First Three PCs

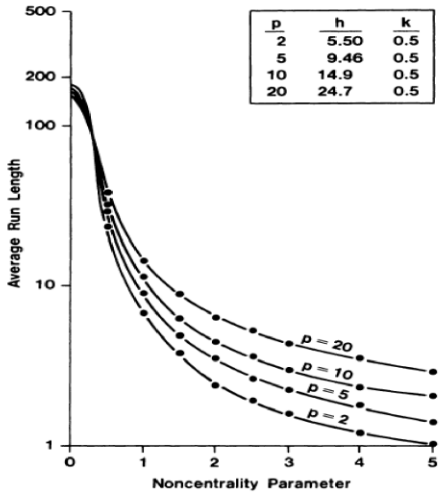

Figure 14: ARL Chart for m-CUSUM

The choice of an Average Run Length for Type II error (ARL_1) was not so easy to deduce because we cannot simply approximate by taking the inverse of the desired beta error. Research revealed that the ARL_1 for a mean shift of 3-sigma and three data characteristics $(p = 3)$ is approximately 3, as interpreted from Figure 13 (Crosier, 1998) where the 'Noncentrality Parameter' refers to the mean shift. For a Noncentrality Parameter of 3-sigma, the data characteristic of 3 falls in-between the data characteristics of 2 and 5. This corresponds to an ARL_1 of approximately 3.

Restoration of the Original Signal

Plotting the individual PC charts (Figures 8-10) revealed trends in the observations that helped divide the data into four major segments. Segment 1 refers to the first 38 points on PC2 (Figure 9). It can be observed that these first 38 points deviate from the trend and therefore they can be treated as a "warm up".

Segment 2 refers to index values between 444 and 457 on PC2. Comparing it with the m-CUSUM (Figure 12) shows that the data is out-of-control from 447 to 457, but m-CUSUM has a slight delay before signaling. PC2 captured the three out-of-control points that the m-CUSUM missed.

Segment 3 refers to index values on PC1 (Figure 8) signaling an out-of-control condition combined with the index values on m-CUSUM. There is a sudden change in trend in PC1 and have alarming fluctuations that also signal out-of-control values, which starts at 529. The m-CUSUM records this at 532 and continues to do till the last data point. It can also be seen that a slight mean shift occurs between index values after 529 on PC2.

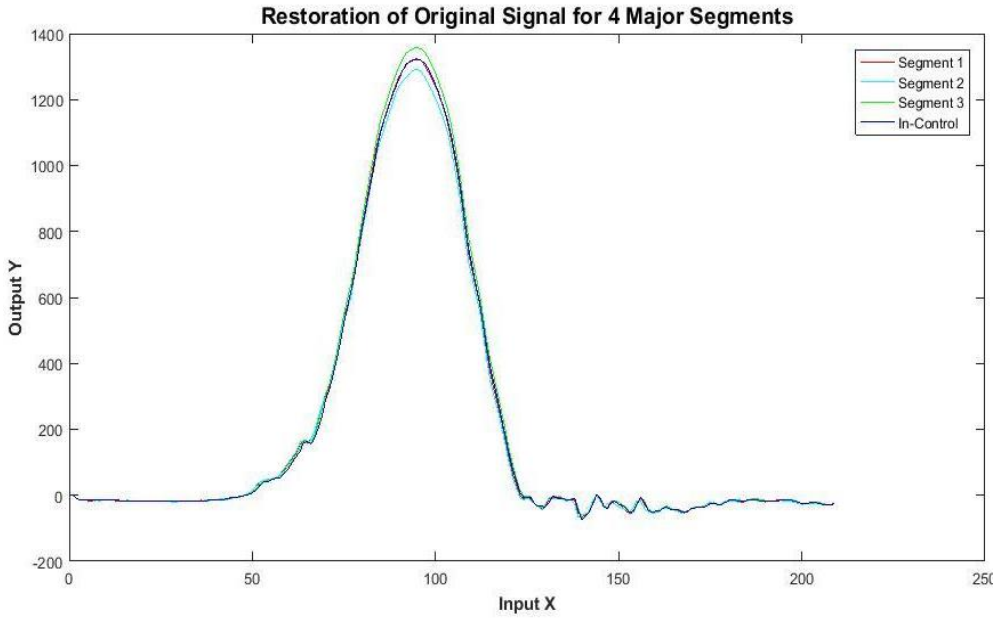

Figure 15: Restoration of the Original Signal

Averaging the four segments, the profile signal of the original dataset was restored as shown in Figure 8. The difference between Segment 3 and the In-Control curve is much greater than the other two

segments. This is expected because Segment 3 refers to fluctuations in PC1, which has the highest eigenvalues and therefore any change in PC1 will be large on the detection scale.

Generation of m-CUSUM and Hotelling T^2 Control Charts

Prior to generating the Hotelling T^2 and m-CUSUM control charts, the mean and variance of the selected first three PCs were calculated. The T^2 control chart is better at identifying large changes in the mean shift whether they are sustained or sudden (spike). On the other hand, the m-CUSUM control chart is better at detecting small and large sustained changes in the mean shift. Therefore, we opted to use both in-conjunction to isolate all possible out-of-control data.

After the first round of iterations for the T^2 control chart in which all the data values were established as in-control, the first round of m-CUSUM was generated. If data values were found to be out-of-control on the m-CUSUM control chart, then a second round of the T^2 control chart was needed.

Hence, multiple iterations were required in order to isolate the out-of-control data values from the in-control data values. The Hotelling T^2 and m-CUSUM control charts were used back and forth until values on both control charts were found to be in-control.

The T^2 statistic was calculated using:

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

where $\bar{x}$ is the mean of the x_j's, x_j is the individual observation, and S^{-1} is the sample covariance matrix.

The m-CUSUM statistic was calculated using:

$$MC_i = \max\{0, (C_i^T * \sum_0^{-1} * C_i)^{0.5} - k * n_i \}$$

where C_i is the cumulative sum of the previous $n_i's$ for the number of $x_i's$, $\sum_0$ was estimated using the sample covariance where m is the total number of observations:

$$\widehat{\sum_0} = S = \frac{1}{m-1} * \sum_{j=1}^{m} \left(x_j - \bar{x}\right) * \left(x_j - \bar{x}\right)^T$$

and the offset constant was defined as:

$$k = 0.5 * 3 = 1.5$$

After removing the out-of-control values identified after plotting the first three PCs, 467 observations remained of the 552 observations. Of these, further out-of-control points were identified and removed from the T^2 and m-CUSUM iterations:

T^2 and m-CUSUM Control Chart Out-of-Control Observations				
Round	**Control Chart**	**Iteration**	**Out-of-Control Observations**	**Remaining Observations**
One	T^2	1	9	458
		2	7	451
		3	1	450
	m-CUSUM	1	5	445
		2	1	444
Two	T^2	1	3	441
		2	1	440
	m-CUSUM	1	1	439
		2	1	438
Three	T^2	1	**0**	**438**

Table 3: T^2 and m-CUSUM Control Chart Iteration Results

Upon finalizing the in-control data set, the in-control mean and variance were calculated. This marked the conclusion of Phase I Analysis.

Results

Having isolated the out-of-control from the in-control data, the in-control parameters were calculated to be:

In-control mean: $\begin{bmatrix} 11.27567 \\ 16.49829 \\ 5.62782 \end{bmatrix}$

In-control covariance: $\begin{bmatrix} 5724.7 & 519.95 & -456.43 \\ 519.95 & 1180.8 & -354.18 \\ -456.43 & -354.18 & 1084.2 \end{bmatrix}$

Conclusion

Because in the original data set the physical meaning of the data was not given, statistical methods were used to figure out whether to use a covariance matrix or a correlation matrix. Analysis lead us to proceed with the PCA on the covariance matrix.

Isolating the out-of-control data from the in-control data by doing multiple iterations of the T^2 and m-CUSUM control charts helped identify six more data points during the second round of iterations that were out-of-control. Had we stopped after the round one of iterations, those points would not have been identified.

In addition, when PCA is performed on the covariance matrix, the diagonal elements should be zero because this implies the data is uncorrelated and thus independent. However, despite performing the PCA, the non-diagonal elements of the covariance matrix are non-zero, which implies the values are correlated and thus dependent. This may be attributed to aggregate noise which made it difficult to achieve a covariance of zero.

With the conclusion of Phase I analysis, the control charts established from the analysis with the in-control mean and covariance can be used to monitor future observations in Phase II analysis.

Reference

Crosier, Ronald B. “Multivariate Generalizations of Cumulative Sum Quality-Control Schemes.” *Technometrics*, vol. 30, no. 3, 1988, pp. 291–303.

Appendix

Round One Iterations: Hotelling T^2 Control Chart

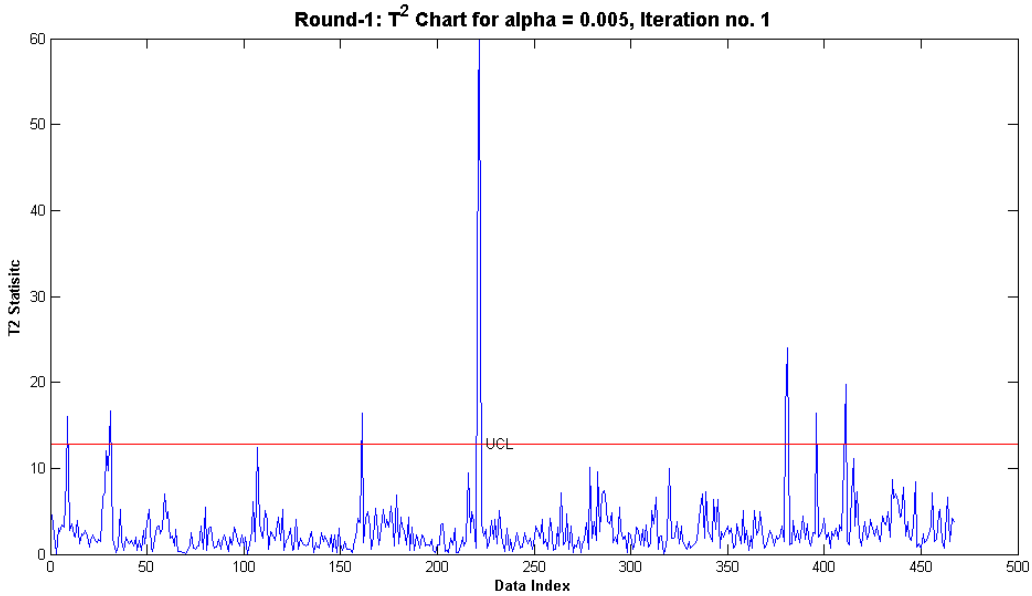

Figure 16: T^2 Round-1 Iteration-1

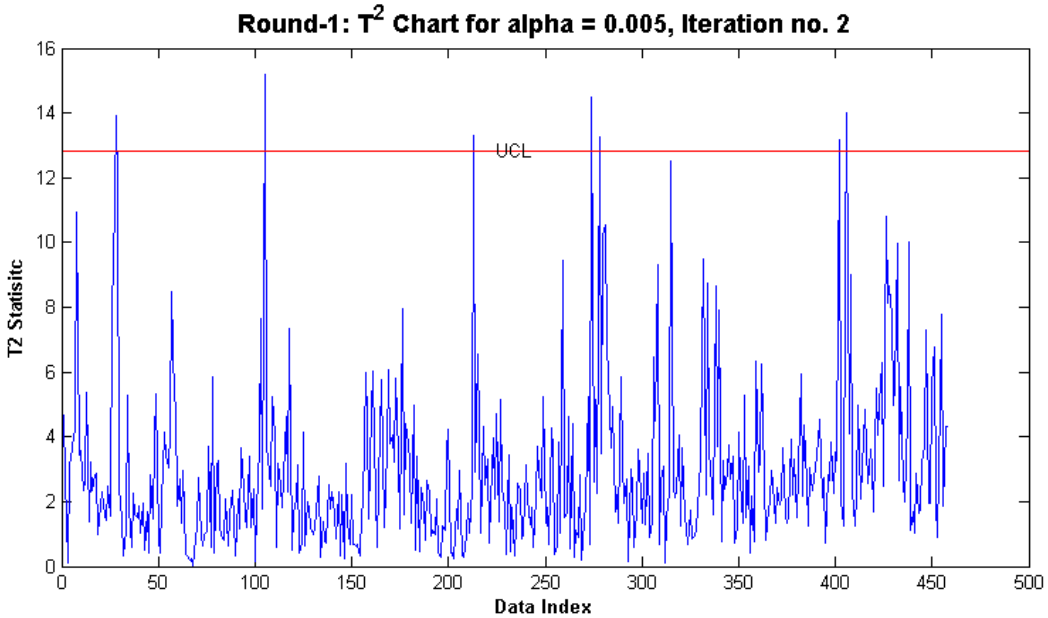

Figure 17: T^2 Round-1 Iteration-2

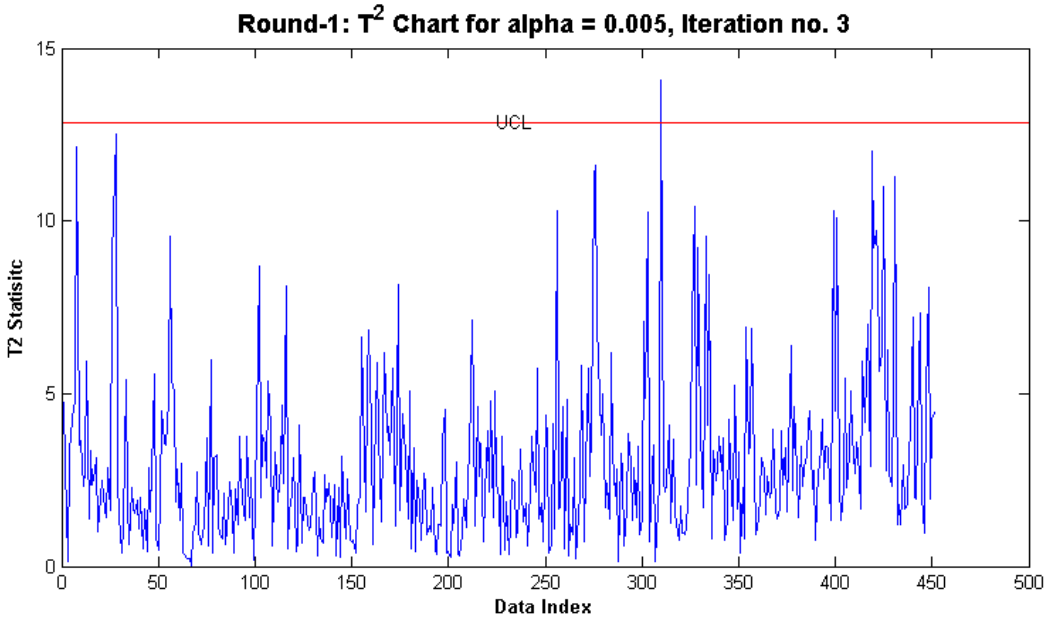

Figure 18: T^2 Round-1 Iteration-3

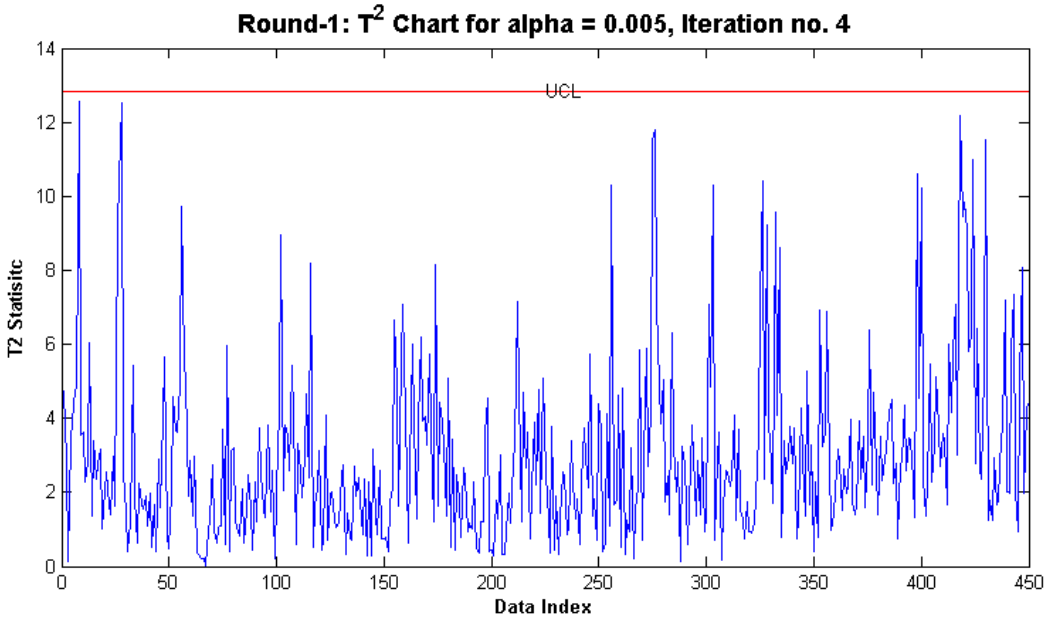

Figure 19: T^2 Round-1 Iteration-4

Round One Iterations: m-CUSUM Control Chart

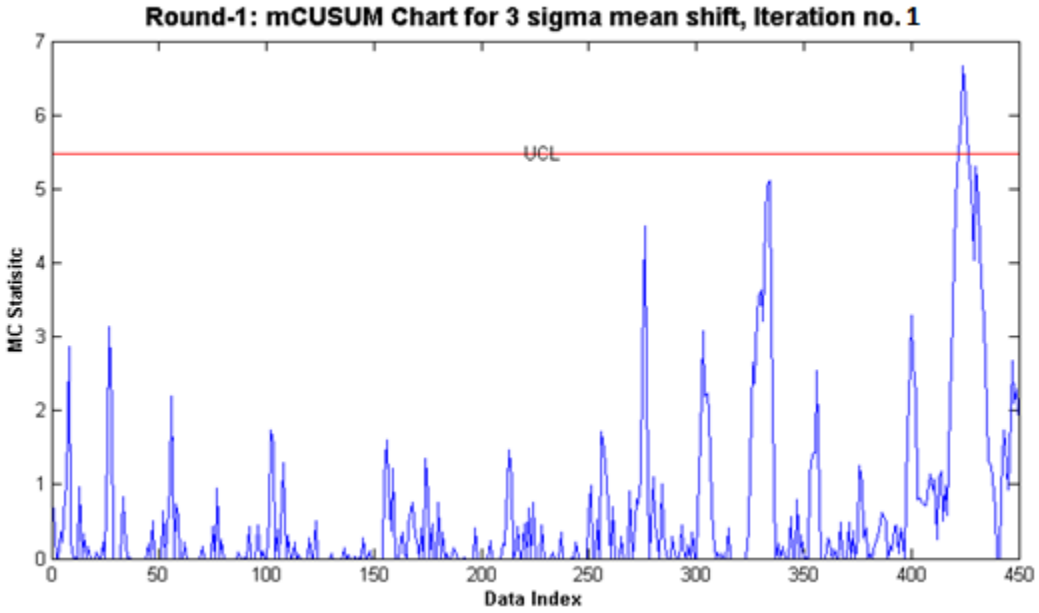

Figure 20: m-CUSUM Round-1 Iteration-1

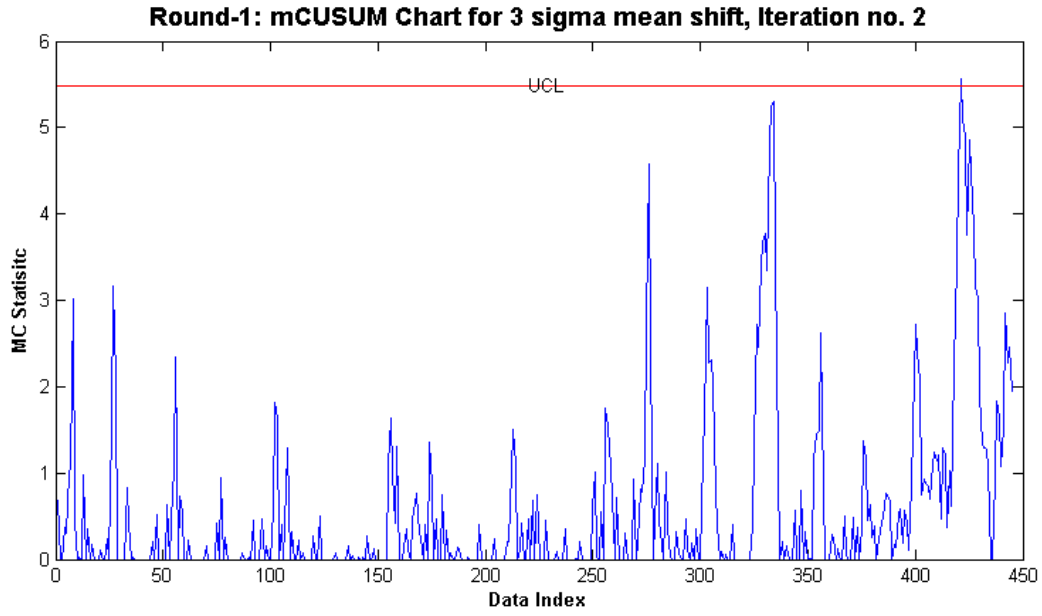

Figure 21: m-CUSUM Round-1 Iteration-2

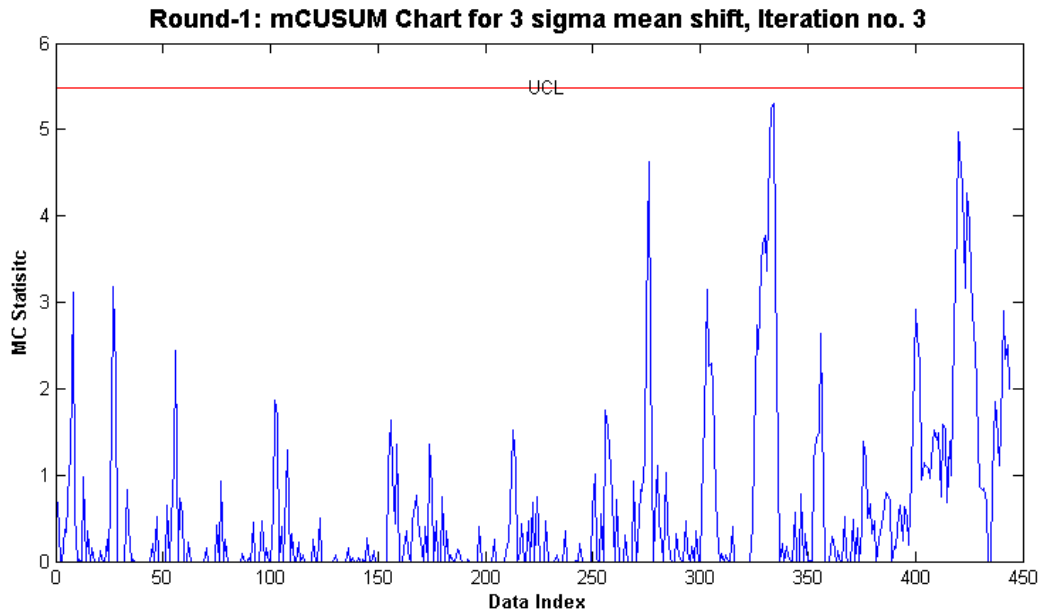

Figure 22: m-CUSUM Round-1 Iteration-3

Round Two Iterations: Hotelling T^2 Control Chart

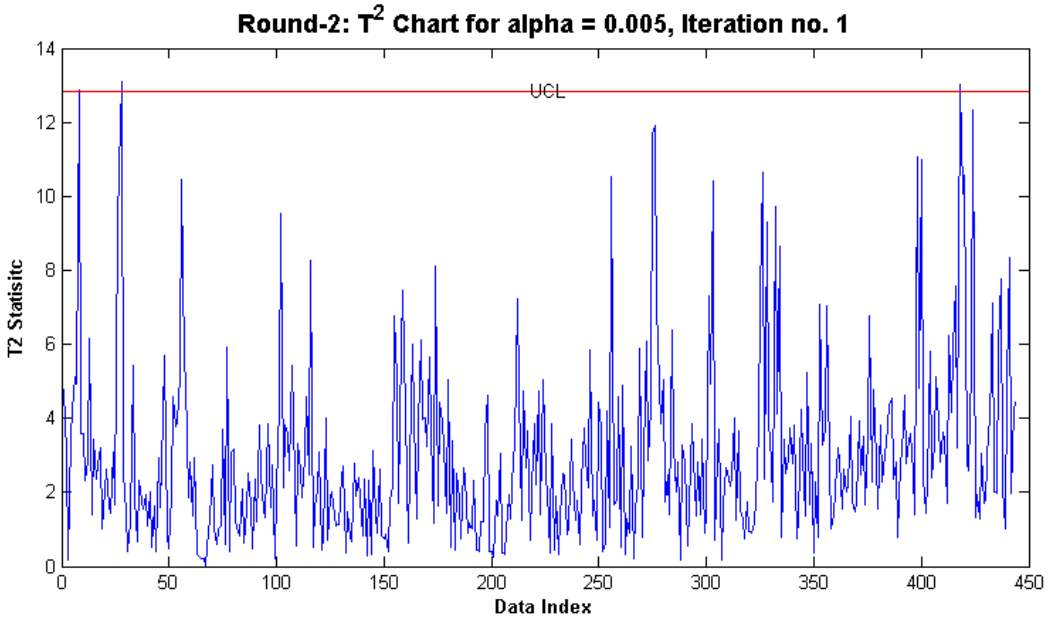

Figure 23: T^2 Round-2 Iteration-1

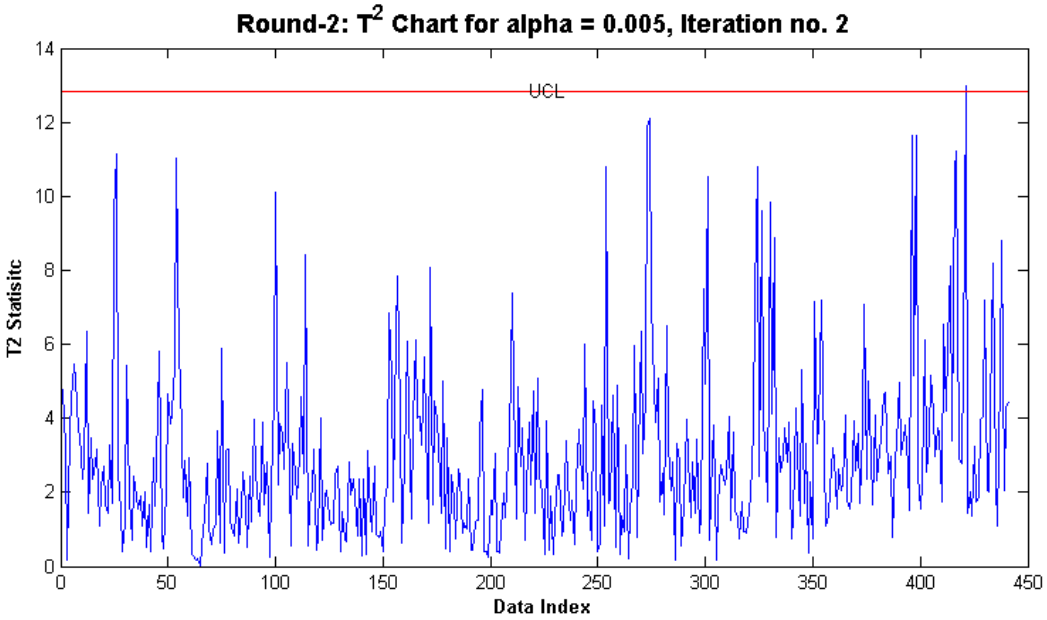

Figure 24: T^2 Round-2 Iteration-2

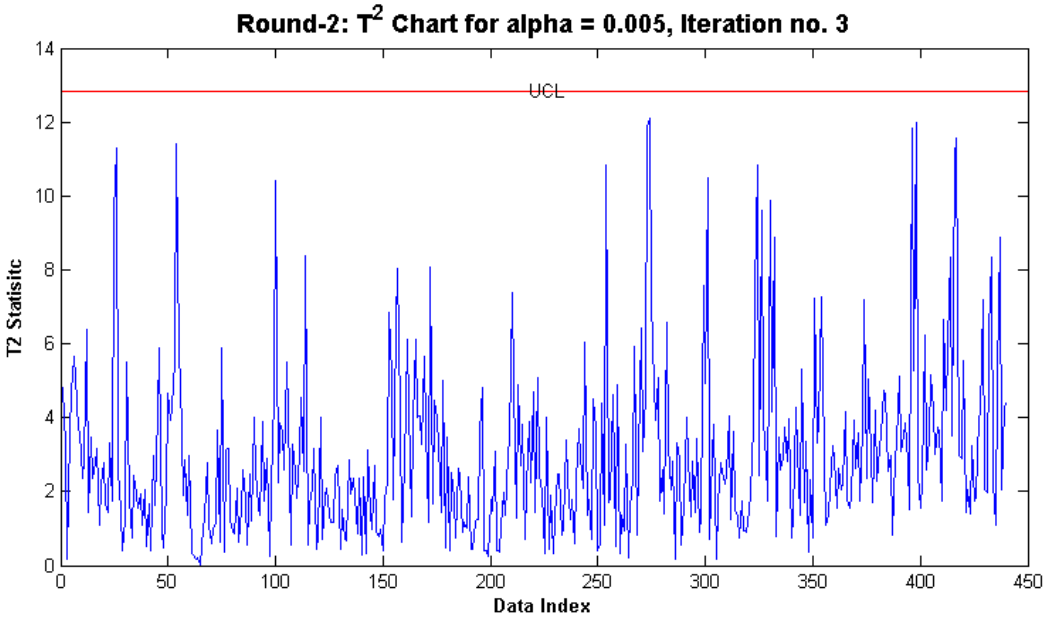

Figure 25: T^2 Round-2 Iteration-3

Round Two Iterations: m-CUSUM Control Chart

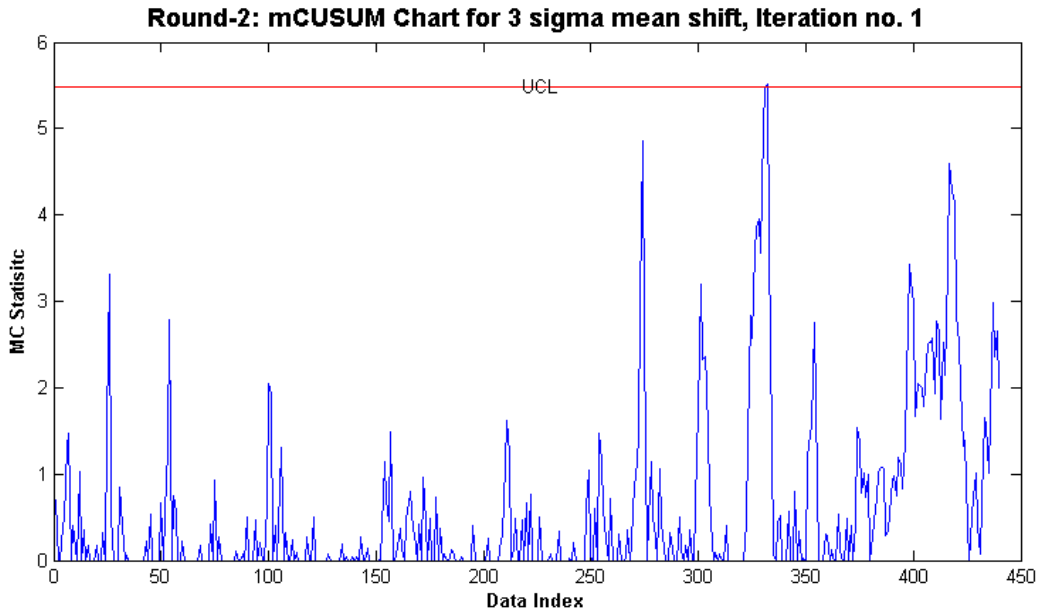

Figure 26: m-CUSUM Round-2 Iteration-1

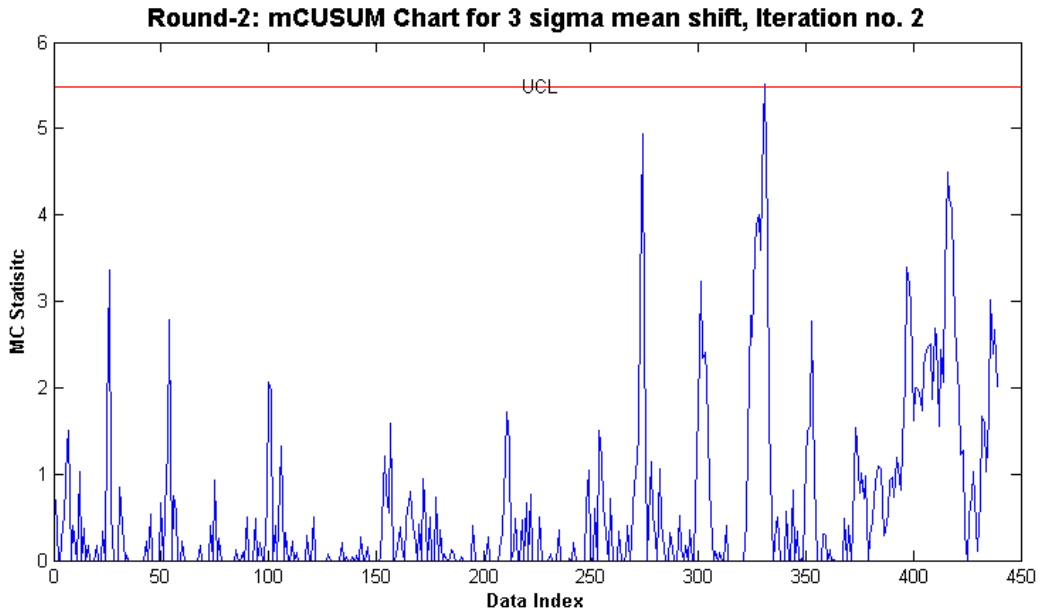

Figure 27: m-CUSUM Round-2 Iteration-2

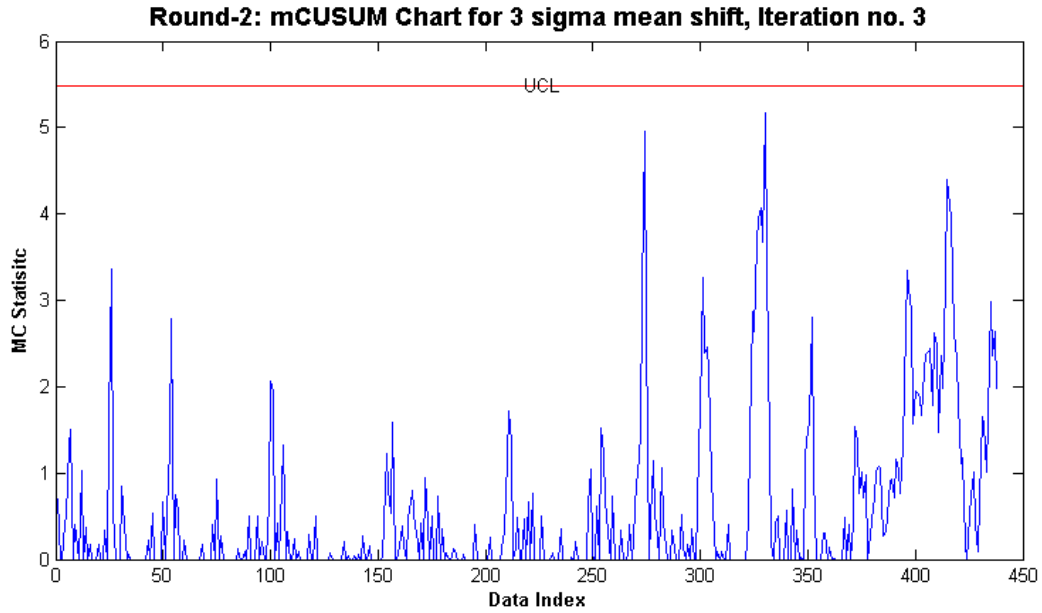

Figure 28: m-CUSUM Round-2 Iteration-3

Round Three Iterations - Hotelling T^2 Control Chart:

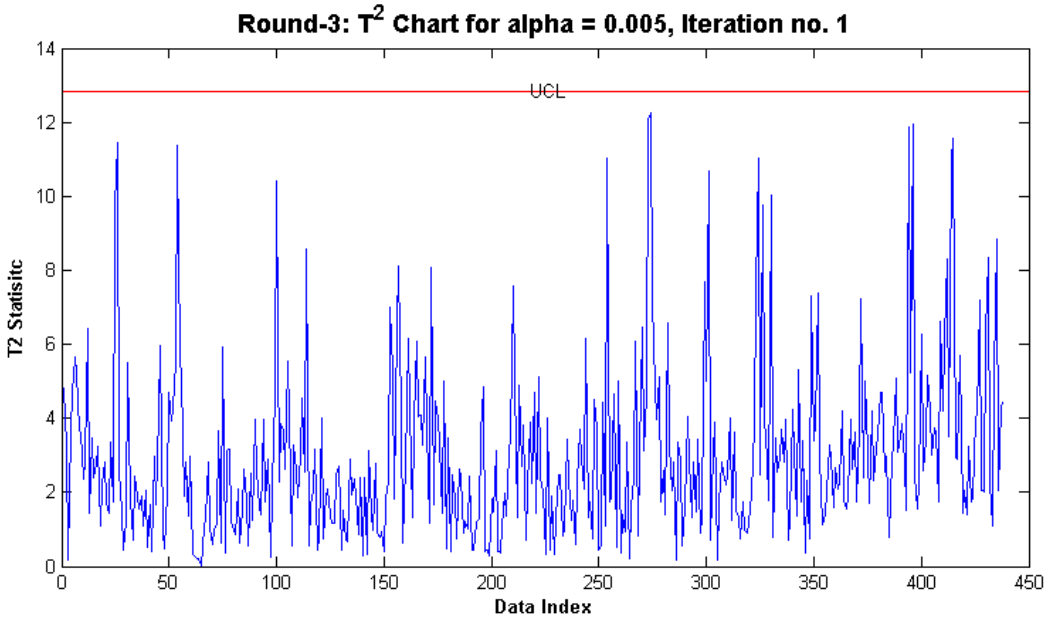

Figure 29: T^2 Round-3 Iteration-1