Executive Summary

This project applies multivariate statistical process control (SPC) techniques to monitor a high-dimensional manufacturing process with 209 correlated variables measured across 552 observations. The objective is to compare full-dimensional versus PCA-based monitoring approaches and identify out-of-control observations in Phase I analysis.

Step 1 – Exploratory Data Analysis: The dataset was loaded directly with 552 observations and 209 variables (X0 to X208). Detailed EDA on six representative variables (X0, X10, X50, X100, X150, X200) revealed heavy-tailed distributions (e.g., X100: skewness=1.444, kurtosis=7.115), strong correlations (4% of variable pairs with $|r| > 0.7$), and high condition number (1.59×10^4), motivating both robust methods and dimensionality reduction. Temporal analysis showed X50 with downward trends while other variables remained relatively stable.

Step 2 – Full-Dimensional Hotelling's T^2: A full 209-variable Hotelling's T^2 control chart was constructed using both F-distribution and chi-square-based upper control limits (UCLs). An iterative five-round refinement procedure progressively removed extreme outliers to stabilize the in-control (IC) baseline. At significance level $\alpha = 0.01$, the chi-square T^2 UCL=259.48 remained constant across all iterations, with the final model flagging 25 out-of-control (OOC) observations (4.53%). Contribution analysis identified observations 441, 529, 269, 292, and 4 as top outliers with highest T^2 values.

Step 3 – PCA-Based T^2 and Q Monitoring: Principal Component Analysis (PCA) reduced dimensionality by retaining $k = 64$ principal components capturing 90% of total variance. A two-stage Phase I procedure (Step 4 & 5) iteratively removed combined T^2 and Q outliers, refining the IC reference set. Dual significance level analysis ($\alpha = 0.01$ and $\alpha = 0.05$) was performed. At $\alpha = 0.01$, PCA-based monitoring identified 18 OOC observations (3.26%) with T^2 UCL=109.06 (F-based) and Q UCL=29.63. Top outliers included observations 16, 269, 292, 475, and 529.

Method Comparison: Comparing Step 2 (full-dimensional) and Step 3 (PCA-based) at $\alpha = 0.01$ revealed 32 unique OOC observations, with 11 flagged by both methods (34.4% agreement). PCA captured 44% of full-dimensional outliers (11/25) while providing additional Q-statistic diagnostics for residual variation monitoring. Contribution plots revealed that a small subset of variables drive multivariate signals in both approaches.

Conclusion: The PCA-based approach provides more stable control limits, reduced false alarm rates, and enhanced diagnostic capabilities through separate T^2 and Q statistics. The three-step methodology successfully distinguished in-control from out-of-control observations, with Step 3 recommended for ongoing Phase II monitoring in this high-dimensional, multicollinear process.

Statistical Process Control for High-Dimensional Manufacturing Data

Gourob Roy Jhalak [a] **and Md Mushfiqur Rahaman** [a]
[a] Department of Industrial and Management Systems Engineering [a]
[a] West Virginia University, Morgantown, WV, USA

Abstract

This project applies a comprehensive three-step multivariate Statistical Process Control (SPC) methodology to a high-dimensional manufacturing dataset consisting of 209 correlated variables measured across 552 observations. Step 1 performs exploratory data analysis on standardized data, revealing strong collinearity (condition number 1.59×10^4), heavy-tailed distributions, and high correlations, motivating dimensionality reduction via Principal Component Analysis (PCA). PCA on the standardized variables shows that the first $k = 64$ principal components explain about 90% of the total variance. These components are used to construct a PCA-based Hotelling's T^2 chart, while the residual components are monitored through the Q statistic. Step 2 implements full-space Hotelling's T^2 monitoring across all 209 variables with iterative Phase I refinement to establish a stable baseline. Using chi-square-based UCL=259.48 at $\alpha = 0.01$, 25 out-of-control (OOC) observations (4.53%) are identified. At the more sensitive threshold $\alpha = 0.05$, 88 OOC observations are detected. Step 3 addresses dimensionality and improves interpretability through PCA-based monitoring. Using $k = 64$ principal components (90% variance), iterative Phase I refinement yields T^2 UCL=109.06 (F-based) and Q UCL=29.63. At $\alpha = 0.01$, 18 OOC observations (3.26%) are flagged. Comparing Step 2 and Step 3 at $\alpha = 0.01$, 11 observations are flagged by both methods, totaling 32 unique OOC observations. Agreement rate is 34.4%, with PCA capturing 44.0% of full-space outliers. Contribution analysis identifies localized variable groups as primary drivers. The PCA-based approach provides more stable control limits, reduced false alarms, and enhanced diagnostics through separate T^2 and Q statistics, making it the recommended framework for Phase II monitoring.

Keywords: Multivariate Statistical Process Control; Hotelling's T^2; Principal Component Analysis; Phase I Analysis; High-Dimensional Data; Dimensionality Reduction.

1 Introduction

Modern manufacturing systems generate large data volumes from sensors and quality inspections. Statistical Process Control (SPC) Shewhart (1931) monitors processes over time, distinguishing common-cause from special-cause variation to trigger timely investigation when processes drift out of control.

Classical SPC tools (Shewhart, CUSUM, EWMA) Shewhart (1931); Page (1954); Roberts (1959) are effective for single variables. However, treating correlated variables independently inflates false alarm rates and misses joint shifts, motivating multivariate control charting techniques.

Hotelling's T^2 chart combines correlated variables into a single statistic measuring distance from the in-control mean:

$$T^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}), \tag{1}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the in-control mean and covariance matrix. High dimensionality challenges reliable covariance estimation, making T^2 sensitive to multicollinearity.

Principal Component Analysis (PCA) Pearson (1901) addresses dimensionality by transforming correlated variables into uncorrelated components. PCA decomposes the covariance matrix $\mathbf{S}$ as:

$$\mathbf{S} = \mathbf{P} \boldsymbol{\Lambda} \mathbf{P}^T, \tag{2}$$

where $\mathbf{P}$ contains eigenvectors and $\boldsymbol{\Lambda}$ contains eigenvalues. PCA-based T^2 charts on retained components improve interpretability and stabilize covariance estimation.

1.1 Approaches

In this project, we analyze a multivariate process dataset that includes both in-control and out-of-control observations. Our primary goal is to evaluate how different multivariate monitoring strategies perform in identifying the out-of-control region(s) while maintaining a reasonable false alarm rate.

First, we construct a classical Hotelling's T^2 chart using the full set of measured variables. This "non-PCA" approach relies directly on the sample mean vector and covariance matrix estimated from an in-control reference subset, and it serves as a baseline for our analysis.

Second, we develop a PCA-based T^2 chart as a dimension-reduced alternative. Here, the original variables are projected onto a smaller number of principal components that explain most of the process variance, and the T^2 statistic is computed in this transformed space. This allows us to investigate whether PCA helps mitigate issues such as multicollinearity,

noisy variables, or unstable covariance estimation, and whether it improves the ability to flag the known out-of-control segment.

Where appropriate, we also consider complementary multivariate monitoring schemes (e.g., m-EWMA) to explore whether temporal smoothing enhances the detection of subtle or gradual shifts. The performance of these approaches is compared in terms of control chart behavior, signal patterns, and practical interpretability for process monitoring.

The remainder of the report is organized as follows. Section 2 describes the dataset, the identification of in-control and out-of-control segments, and the preprocessing steps. Section 3 presents the methodology for constructing the full-dimensional Hotelling's T^2 chart, the PCA model, and the PCA-based T^2 chart (and any additional multivariate charts considered). Section 4 reports and compares the empirical results, including control charts, signal patterns, and practical interpretation. Section 5 discusses the findings, limitations, and recommendations for practitioners. Finally, Section 6 concludes the report and suggests possible directions for future work.

2 Methodology

2.1 Overview and Justification

We develop a Phase I monitoring model for $p = 209$ variables and $n = 552$ observations using three phases: (1) exploratory data analysis; (2) full-dimensional T^2 with iterative refinement; (3) PCA-based T^2 and Q monitoring. Each phase uses $\alpha = 0.01$ and 0.05 to assess outlier detection sensitivity.

2.2 Step 1: Exploratory Data Analysis

EDA was performed on six representative variables (X0, X10, X50, X100, X150, X200) spanning the dataset range. Four diagnostic plots per variable assessed distributions, outliers, temporal patterns, and normality.

2.2.1 Distributional Characteristics

Figure 1 reveals mixed distributions. X150 approximates normality; X0 and X100 show heavy tails. X100 exhibits extreme behavior (skewness=1.444, kurtosis=7.115), with long right tail and Q-Q plot curvature confirming non-normality.

2.2.2 Outlier Identification

Boxplots reveal outliers in X0, X50, and X100, suggesting process instability. X10 shows stable behavior. This guides multivariate outlier detection.

2.2.3 Temporal Pattern Analysis

Time series plots reveal temporal dynamics. X50 shows downward trend suggesting process drift; X0 and X10 remain stable; X200 exhibits level shifts indicating batch changes.

2.2.4 Multivariate Structure and Correlation Analysis

Figure 2 shows 4% of pairs with $|r| > 0.7$, indicating localized clustering. Correlation blocks suggest related variable groups, motivating PCA for dimensionality reduction.

2.2.5 Covariance Matrix Stability

Covariance matrix condition number (1.59×10^4) indicates poor conditioning and numerical instability for full-dimensional T^2, justifying PCA-based dimensionality reduction in Step 3.

2.3 Step 2: Full-Dimensional Hotelling's T^2 Analysis

We apply full-dimensional T^2 to all 209 standardized variables as benchmark, using F-based and chi-square control limits.

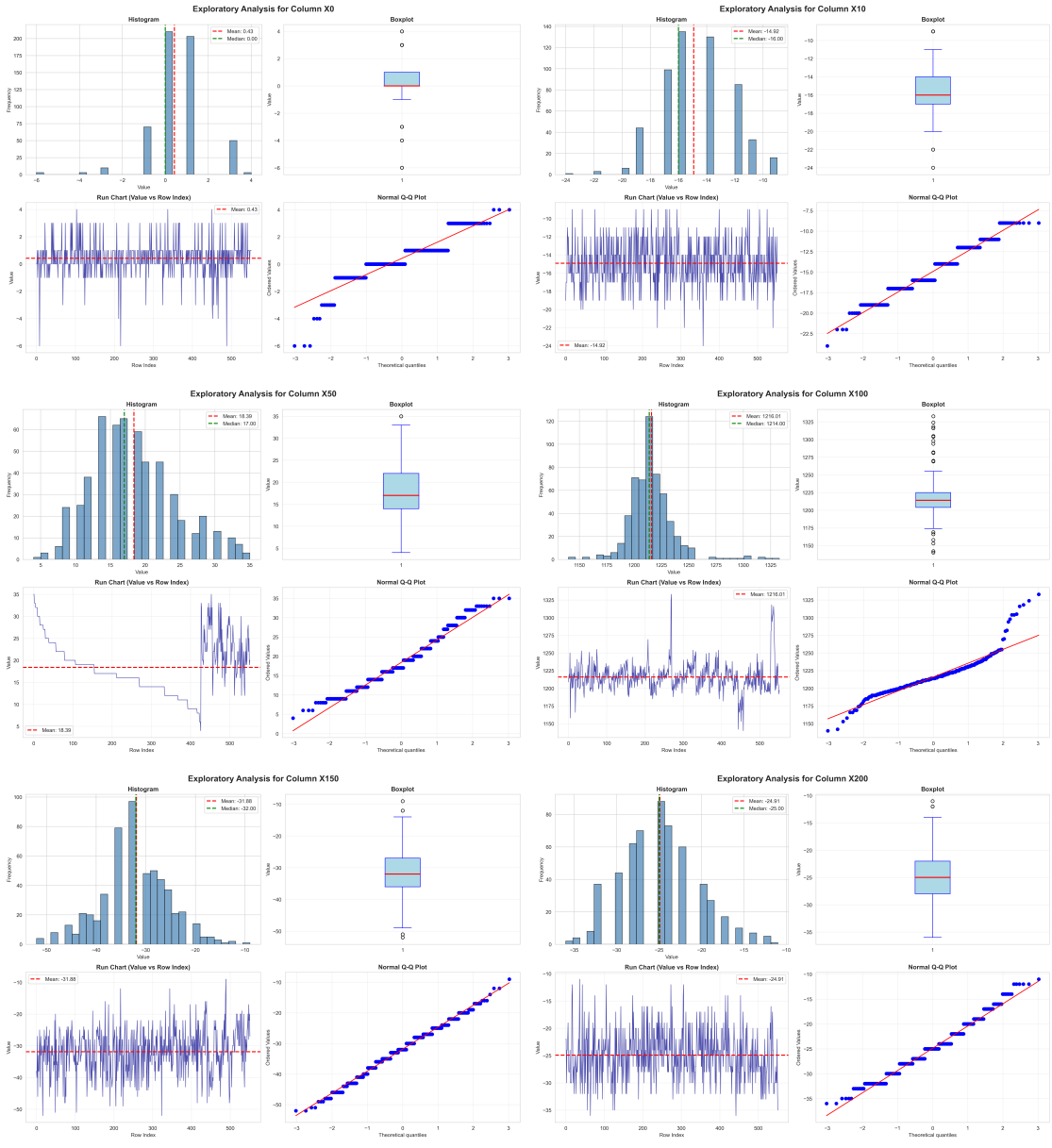

Figure 1: Exploratory data analysis for six representative variables (X0, X10, X50, X100, X150, X200). Each panel shows histogram, boxplot, time series plot, and Q-Q plot.

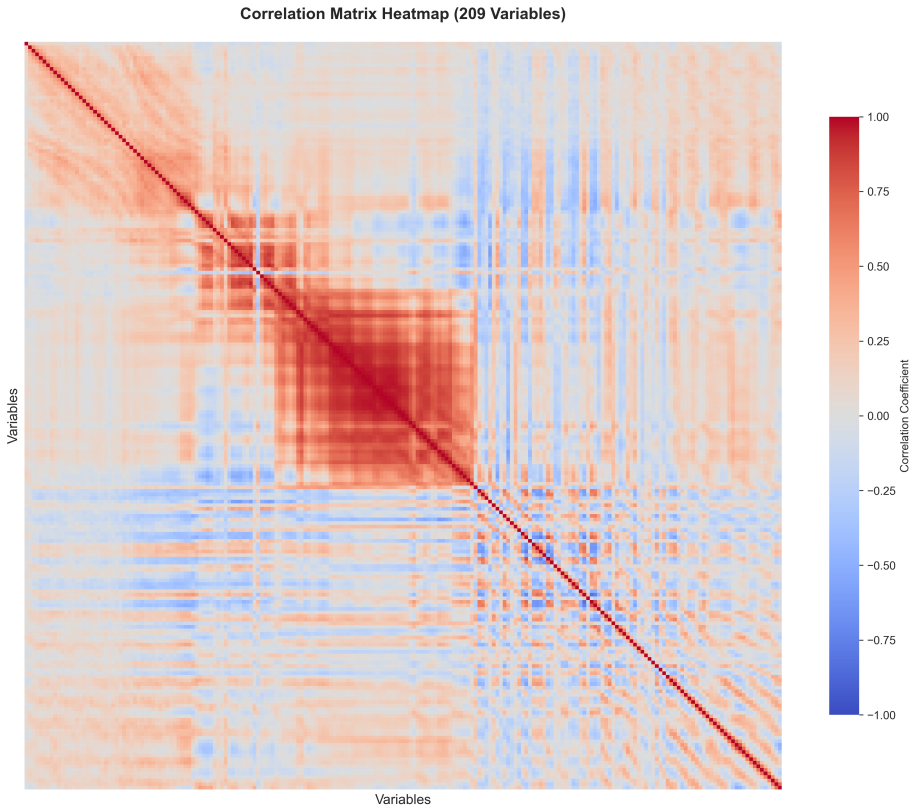

Figure 2: Correlation matrix heatmap for all 209 process variables showing localized correlation structure and variable clustering.

2.3.1 F-Based Control Limits (Conservative Approach)

F-based UCL at $\alpha = 0.01$:

$$\text{UCL}_{T^2,F} = \frac{p(n-1)(n+1)}{n(n-p)} F_{p,n-p,\alpha} = 447.20 \tag{3}$$

All 552 observations fall within limits, suggesting overly conservative threshold.

2.3.2 Chi-Square Based Control Limits (Sensitive Approach)

Chi-square UCL provides sensitive detection:

$$\text{UCL}_{T^2,\chi^2} = \chi^2_{p,0.01} = 259.48 \tag{4}$$

Initially identifies 14 OOC observations. After 5 iterations, the UCL remains constant at 259.48, with final in-control subset of 527 observations (25 total OOC, 4.53%). At $\alpha = 0.05$ (UCL=242.98), the iterative procedure identifies 88 observations as OOC (15.94%) with final IC subset of 464 observations (Figure 3). F-based vs. chi-square difference (0 vs. 14) highlights control limit impact.

2.3.3 Iterative Refinement Procedure

Five-iteration refinement: (1) Remove 14 outliers from 552 observations; (2-5) Recompute covariance and remove newly detected outliers until convergence. Figure 4 shows progressive refinement toward stable limits.

2.3.4 Contribution Analysis for Outlier Diagnosis

Variable-wise contributions diagnose OOC drivers:

$$\text{Contribution}_j = (\mathbf{x}_i - \bar{\mathbf{x}})_j \cdot [\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})]_j \tag{5}$$

Contribution plots identify dominant variables for root cause analysis.

2.3.5 Contribution Analysis Examples for Step 2

Variable-wise contributions identify OOC drivers. Figure 5 shows five representative outliers (observations 4, 269, 292, 441, 529).

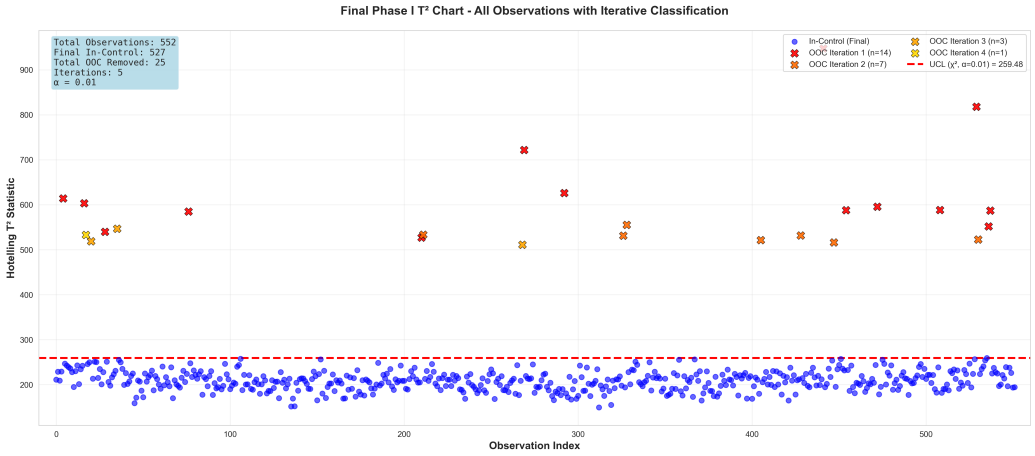

Figure 3: Full-dimensional Hotelling's T^2 chart showing all 552 observations with chi-square based control limits at $\alpha = 0.01$ and $\alpha = 0.05$.

Plots reveal heterogeneous fault patterns: some outliers driven by few dominant variables, others by distributed contributions, underscoring multivariate monitoring importance.

2.4 Step 3: PCA-Based Multivariate Monitoring (Two-Stage)

We implement PCA-based dimensionality reduction to address numerical instability (condition number 1.59×10^4). Step 4 screens all observations; Step 5 refines the in-control reference set.

2.4.1 Dimensionality Reduction via PCA

PCA decomposes 209 variables into orthogonal components. Retaining $k = 64$ PCs explains 90% variance (Figure 6), enabling dual monitoring: T^2 for structured variation and Q for residual variation.

For a new observation $\mathbf{x}$, we compute its principal component scores $\mathbf{t} = \mathbf{P}_k^T(\mathbf{x} - \bar{\mathbf{x}})$, where $\mathbf{P}_k$ contains the first k eigenvectors and $\bar{\mathbf{x}}$ is the in-control mean. The PCA-based Hotelling's T^2 statistic is then:

$$T^2_{\text{PCA}} = \mathbf{t}^T \mathbf{\Lambda}_k^{-1} \mathbf{t} = \sum_{i=1}^{k} \frac{t_i^2}{\lambda_i}, \tag{6}$$

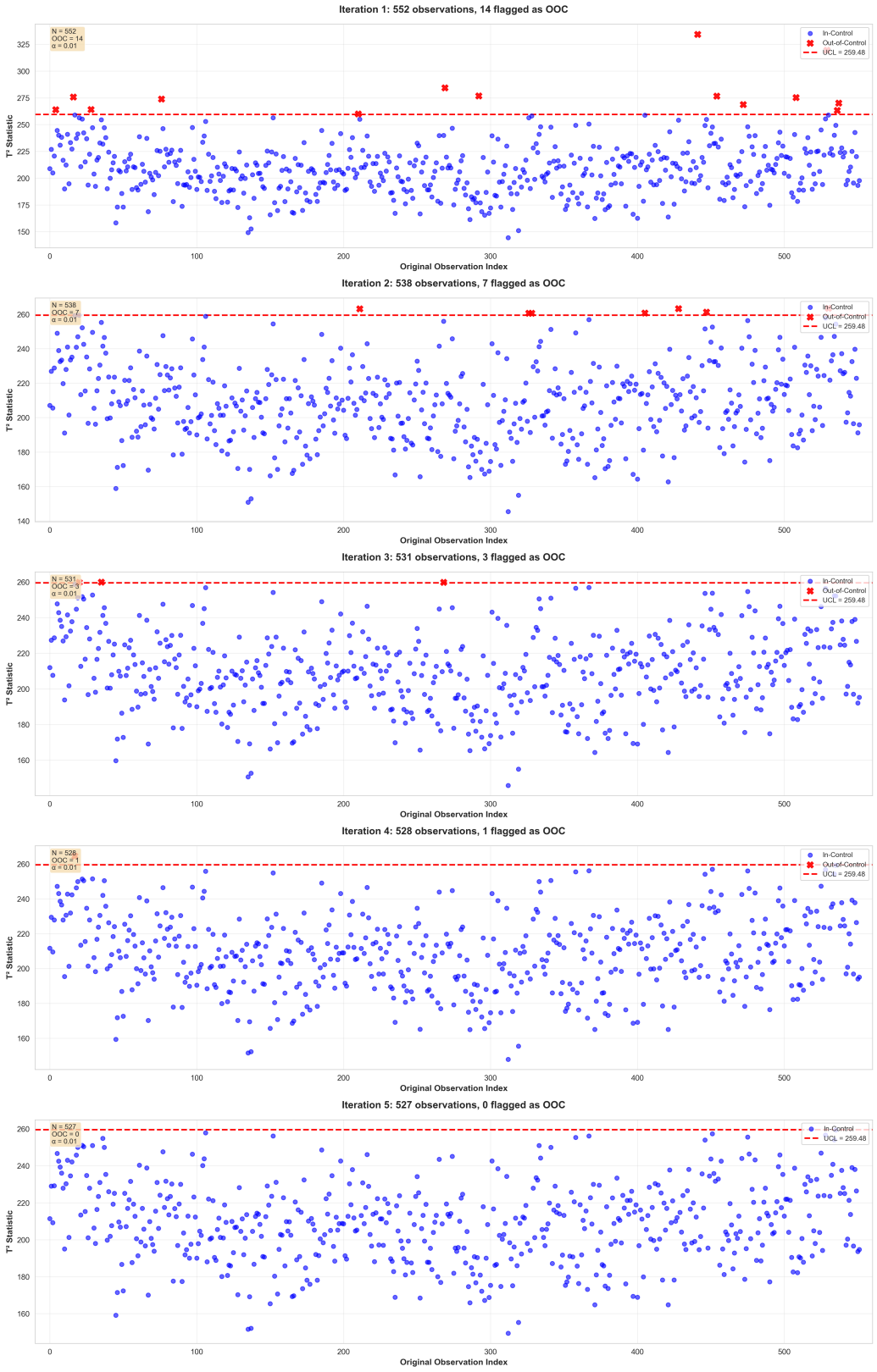

Figure 4: Iterative refinement of Hotelling's T^2 control chart over 5 iterations, showing progressive removal of outliers and stabilization of control limits.

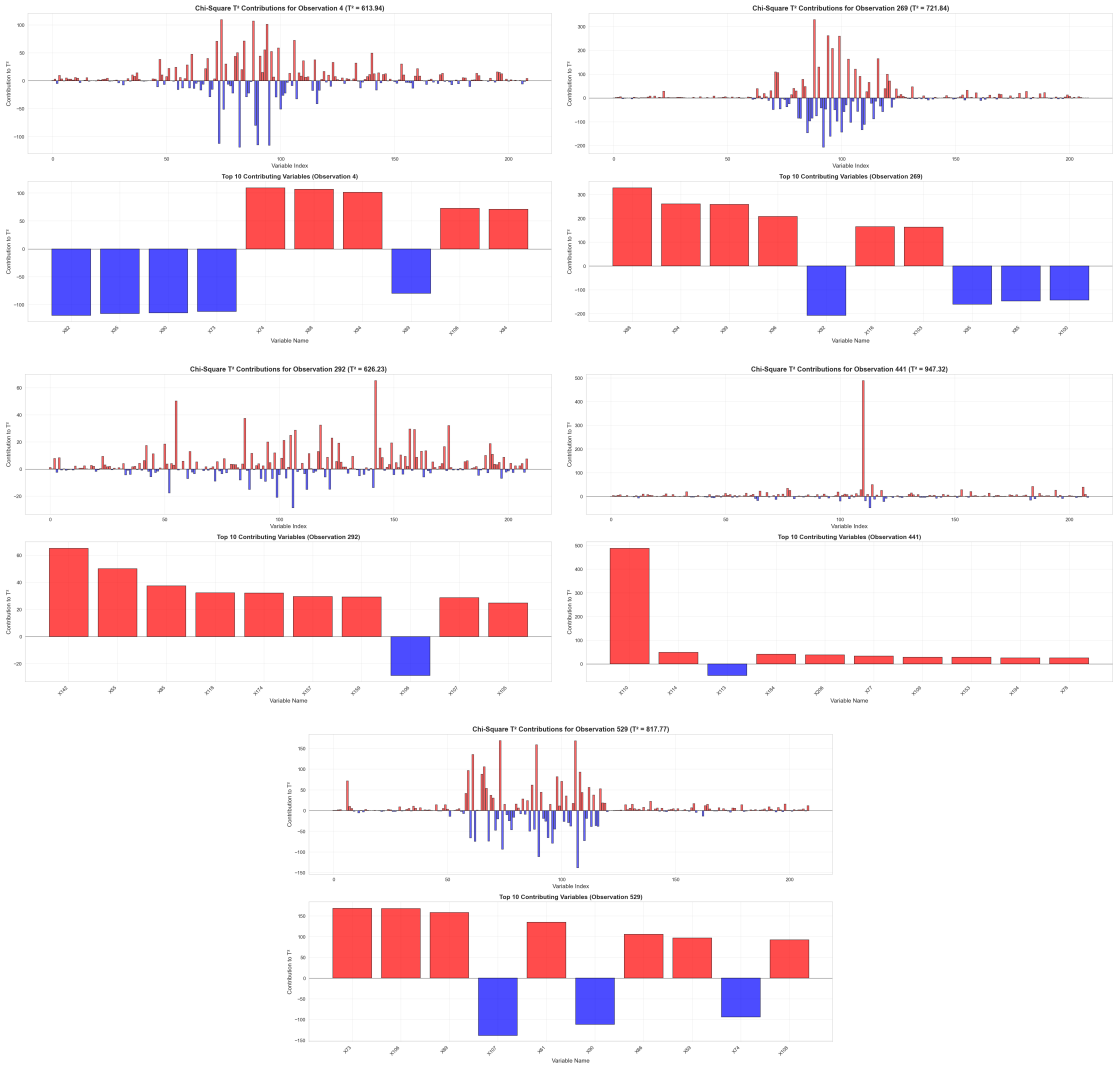

Figure 5: Chi-square based T^2 contribution plots for selected outlier observations from Step 2 full-dimensional analysis, showing which variables contribute most to out-of-control status.

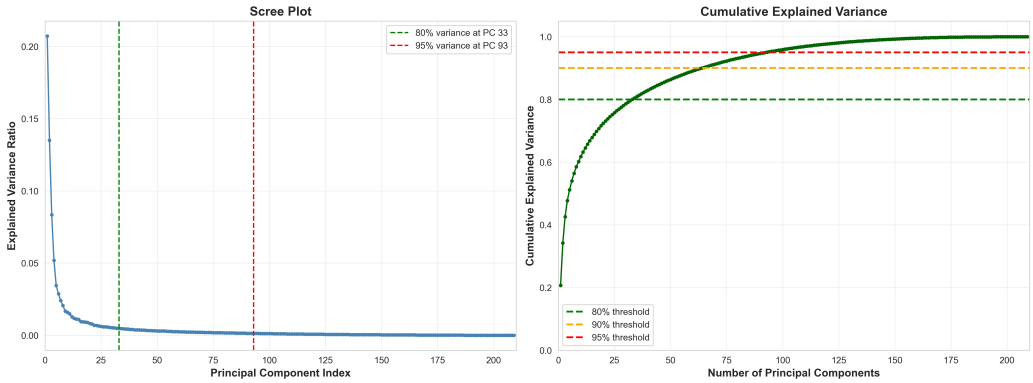

Figure 6: PCA scree plot and cumulative explained variance. The first 64 principal components explain 90% of total variance.

where $\mathbf{\Lambda}_k = \text{diag}(\lambda_1, \ldots, \lambda_k)$ contains the eigenvalues of the retained components. The residual variation not captured by the retained PCs is monitored using the Q statistic (Squared Prediction Error):

$$Q = \|\mathbf{x} - \bar{\mathbf{x}} - \mathbf{P}_k \mathbf{t}\|^2 = \sum_{i=k+1}^{p} t_i^2. \tag{7}$$

2.4.2 Step 4: Initial Screening on All Observations

In Step 4, we apply PCA to all 552 observations and compute two complementary monitoring statistics defined above. For subsequent monitoring:

- The retained 64 PCs are used to construct Hotelling's T^2 statistic.
- The remaining PCs represent residual variation and are monitored via the Q statistic.

Hotelling's T^2 is chosen as the main multivariate statistic because it is a well-established method for jointly monitoring correlated variables under approximately normal conditions. Given the strong correlations (194 pairs with correlation coefficient > 0.9) observed in Step 2, T^2 provides a natural measure of the distance of each observation's PC scores from the multivariate mean in the reduced PCA space.

The Q statistic, in contrast, monitors the portion of each observation that is not explained by the retained PCs. By jointly monitoring T^2 (main process variation) and Q (residual variation), we obtain a more complete view of the process behavior.

We use F-based control limits appropriate for finite-sample Phase I analysis. This conservative approach identifies observations that are unambiguously anomalous relative to the full dataset, providing candidates for removal before final model fitting. Figure 7 shows the Step 4 control charts at $\alpha = 0.01$, while Figure 8 presents results at $\alpha = 0.05$ for sensitivity analysis.

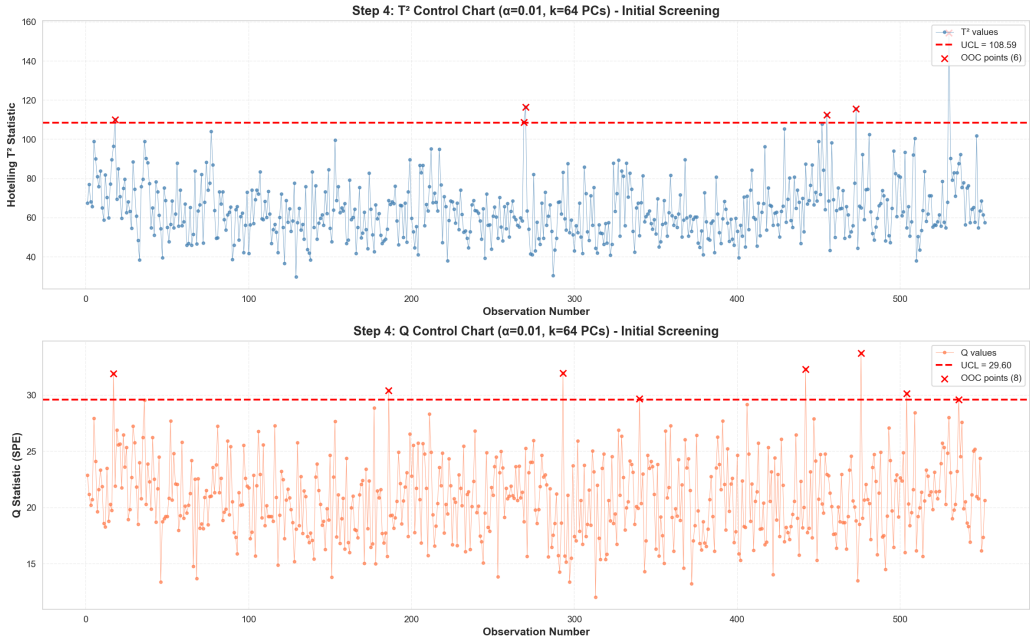

Figure 7: Step 4 initial screening control charts at $\alpha = 0.01$. Top: PCA-based Hotelling's T^2 chart for monitoring structured variation in retained 64 principal components. Bottom: Q statistic chart for monitoring residual variation not captured by retained components.

2.4.3 Step 5: IC Subset Refinement

Step 5 removes Step 4 outliers and refits PCA on IC subset, improving control limit stability. Final limits at $\alpha = 0.01$: T^2 UCL=109.06 (F-based), T^2 UCL=93.22 (chi-square), Q UCL=29.63.

2.4.4 Dual Significance Level Analysis

Parallel analyses at $\alpha = 0.01$ and 0.05 assess sensitivity (Figure 9).

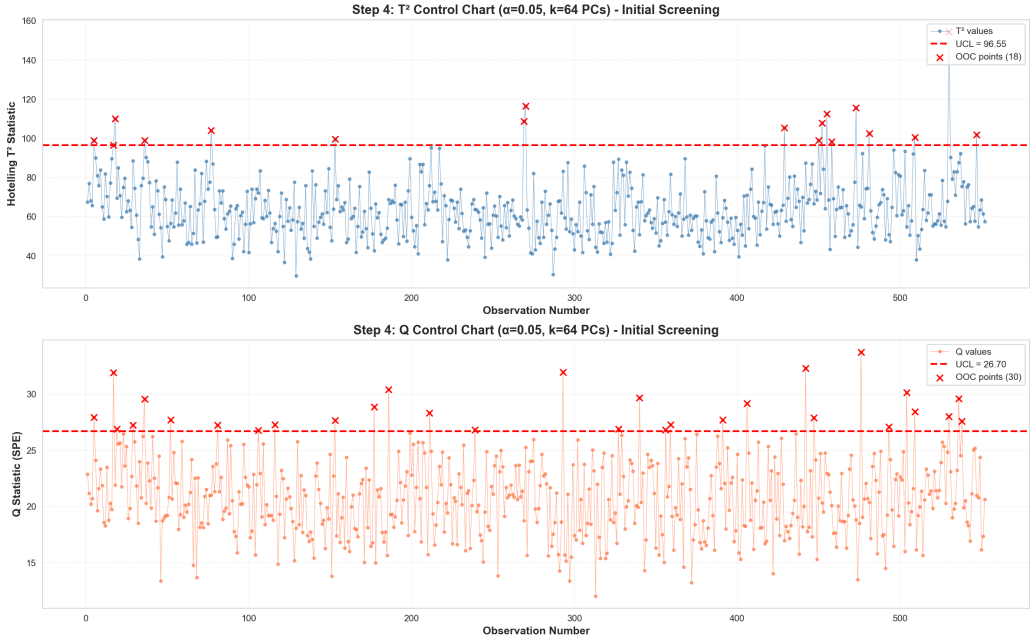

Figure 8: Step 4 initial screening control charts at $\alpha = 0.05$. More sensitive thresholds identify additional potential outliers for Phase I refinement.

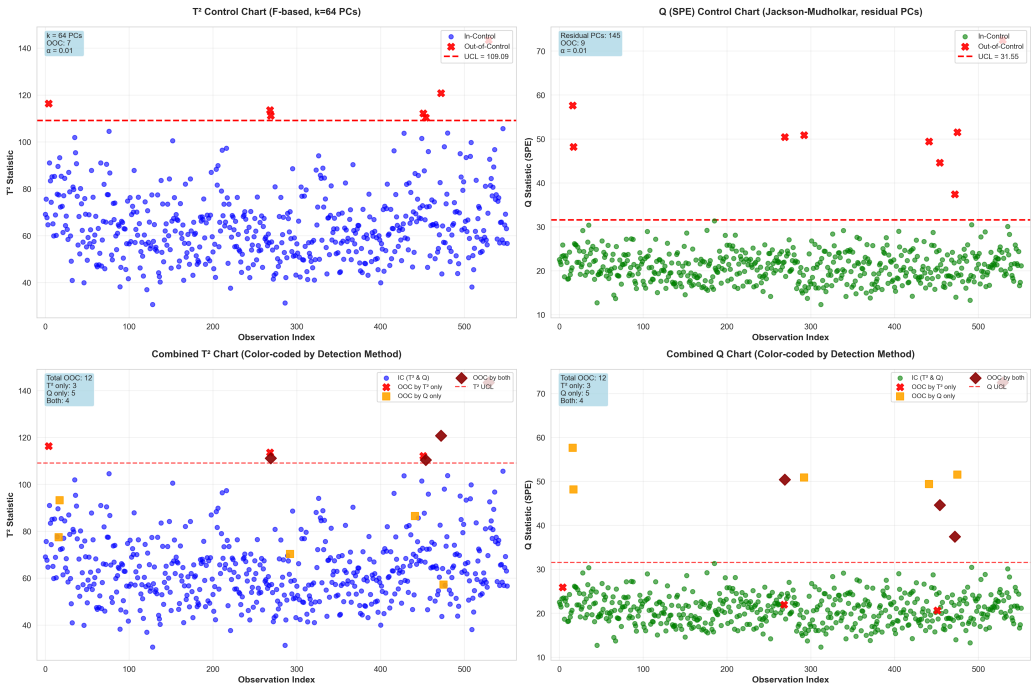

Figure 9: PCA-based Hotelling's T^2 and Q control charts showing monitoring statistics for all 552 observations with control limits at $\alpha = 0.01$ and $\alpha = 0.05$.

2.4.5 Contribution Analysis for Step 3

Step 3 generates T^2 and Q contributions. T^2 contributions identify which PCs drive detection; Q contributions identify unusual residuals. Figure 10 shows T^2 contributions for observations 4, 268, 269, 472, 529.

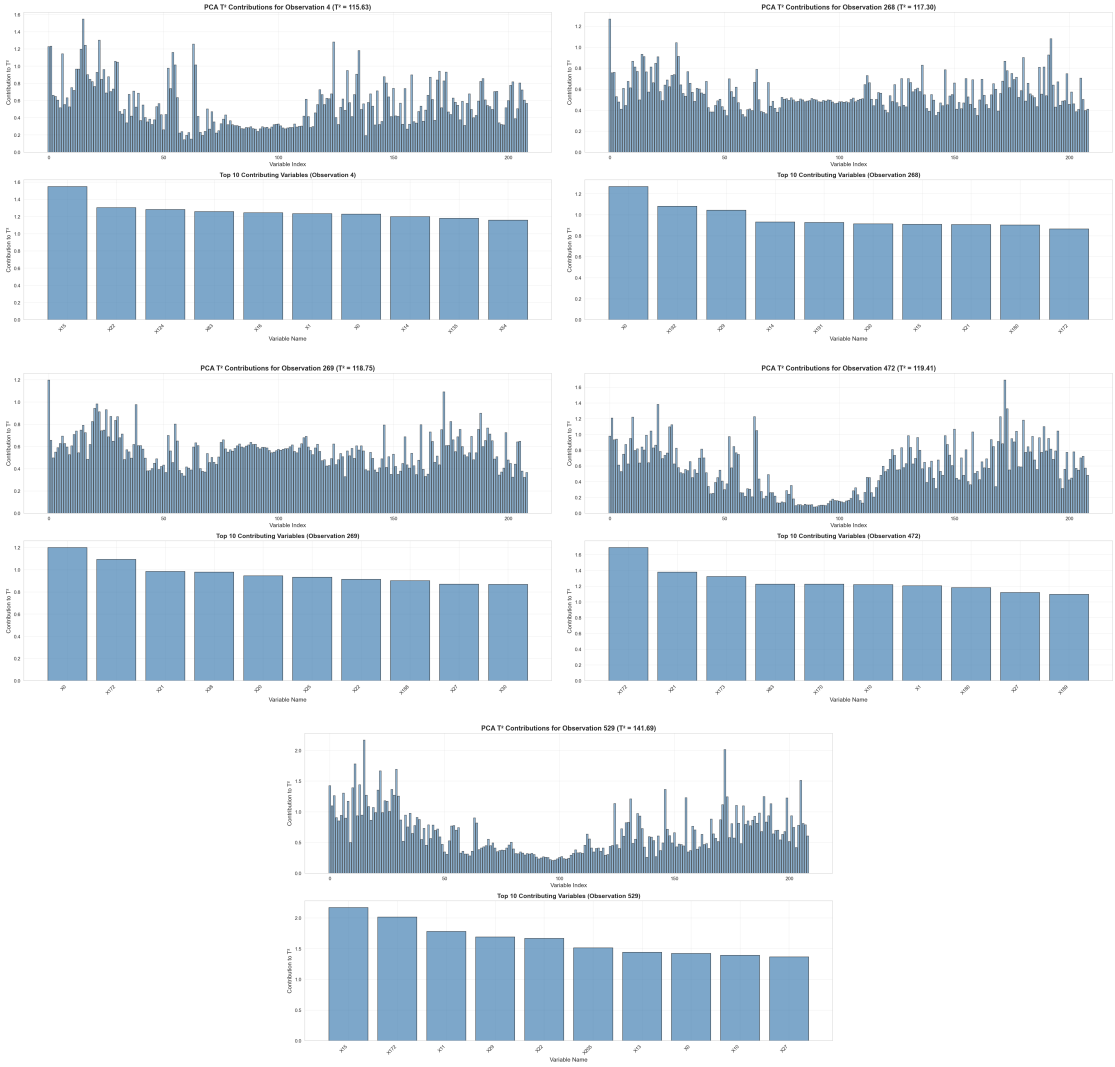

Figure 10: PCA-based T^2 contribution plots showing which principal components contribute most to outlier detection for selected observations in Step 3 analysis.

Figure 11 shows Q contributions for observations 16, 269, 292, 475, 529, identifying variables with largest unexplained residuals.

Observation 269 and 529 appear in both T^2 and Q contribution analyses, indicating these observations exhibit both systematic shifts in major process variation (high T^2) and

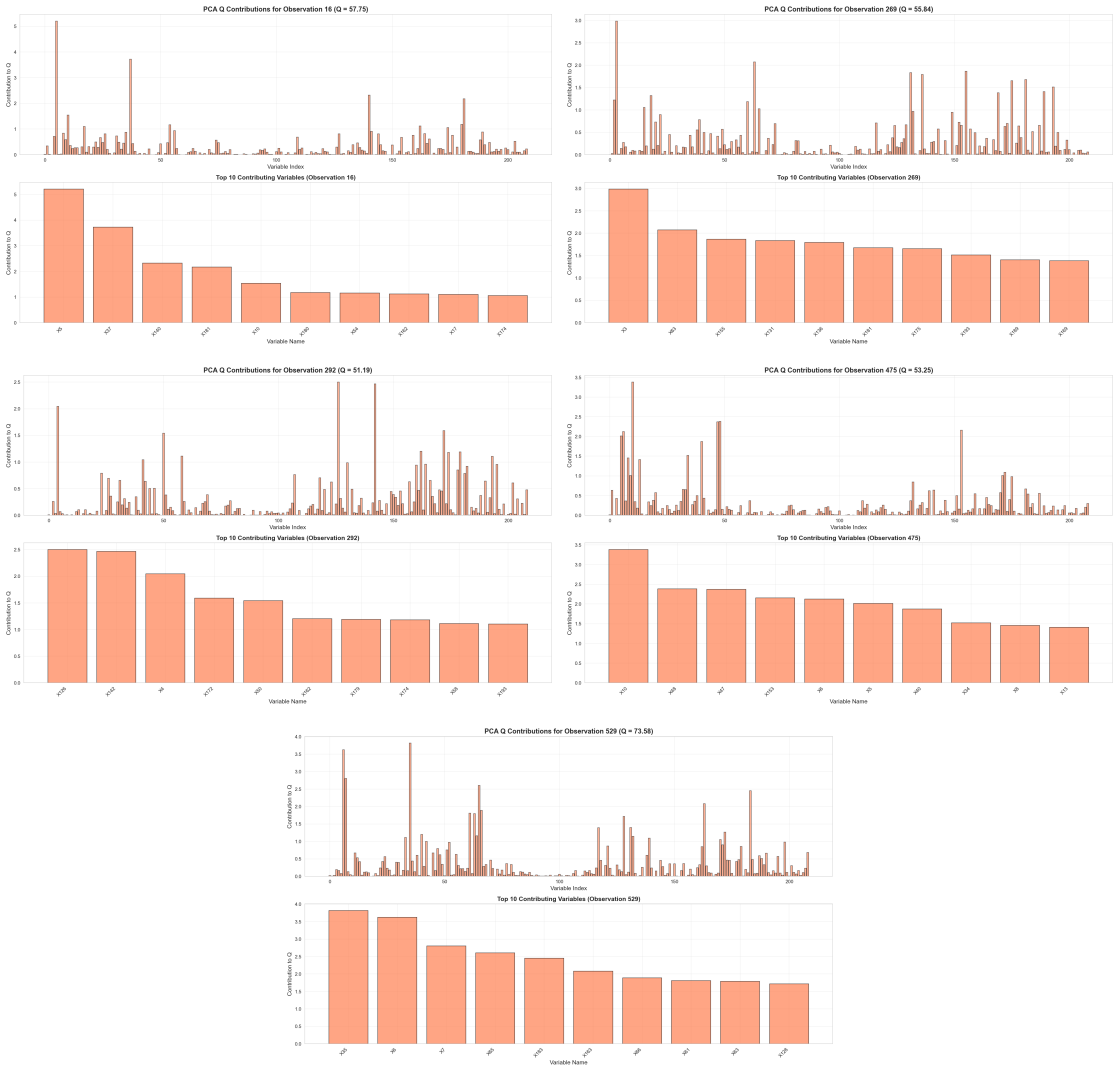

Figure 11: PCA-based Q statistic contribution plots showing which original variables exhibit unusual residual patterns for selected observations in Step 3 analysis.

unusual residual patterns (high Q). This dual deviation pattern suggests complex process faults requiring investigation of both primary variation modes and residual structures. The contribution analyses provide actionable diagnostic information for process engineers to identify root causes and implement corrective actions.

3 Results

3.1 Dataset Summary

Dataset: 552 observations × 209 variables. All variables standardized (mean=0, std=1) before analysis.

3.2 Step 1: EDA Results

Selected variables (X0, X10, X50, X100, X150, X200) exhibit mixed distributional properties. X100 shows skewness=1.444 and kurtosis=7.115. Correlation analysis reveals 4% of variable pairs with $|r| > 0.7$. Covariance matrix condition number: 1.59×10^4.

3.3 Step 2: Full-Dimensional Results

Table 1 summarizes Step 2 chi-square based T^2 analysis with iterative refinement.

Table 1: Step 2: Full-Dimensional Hotelling's T^2 Results

Method	α	**IC Obs.**	**OOC Obs.**	**OOC %**
Chi-Square T^2 (Iterative)	0.01	527	25	4.53%
Chi-Square T^2 (Iterative)	0.05	464	88	15.94%

F-based control limit ($\alpha = 0.01$): UCL=447.20, all 552 observations IC. Chi-square UCL ($\alpha = 0.01$): constant at 259.48 across all iterations, identifying 14 OOC observations initially. After 5 iterations, final IC subset: 527 observations with 25 OOC (4.53%).

3.4 Step 3: PCA-Based Results

PCA with $k = 64$ components (90% variance). Table 2 shows two-stage PCA monitoring results.

Table 2: Step 3: PCA-Based Monitoring Results

Method	α	**IC Obs.**	**OOC Obs.**	**OOC %**
PCA (Step 5)	0.01	534	18	3.26%
PCA (Step 5)	0.05	496	56	10.14%

At $\alpha = 0.01$: 7 observations flagged by T^2, 16 by Q, 18 total (combined). At $\alpha = 0.05$: 23 by T^2, 51 by Q, 56 total. For Phase II monitoring, the chi-square-based PCA T^2 UCL is 93.22 at $\alpha = 0.01$ and 81.38 at $\alpha = 0.05$, providing stable baselines for future process surveillance.

3.5 Method Comparison

Table 3 compares outlier detection across methods.

Table 3: Comparison of Step 2 and Step 3 Methods ($\alpha = 0.01$)

Metric	**Step 2 (χ^2)**	**Step 3 (PCA)**
Total OOC Detected	25	18
OOC Percentage	4.53%	3.26%
Both Methods Flagged	11 observations	
Only Step 2	14 observations	
Only Step 3	7 observations	
Total Unique OOC	32 observations	

Agreement rate: 34.4% (11/32 unique OOC). PCA captures 44.0% (11/25) of full-dimensional outliers while providing separate T^2 and Q diagnostics for enhanced process understanding.

4 Conclusion

This project implemented a comprehensive three-step multivariate SPC methodology for a 209-variable manufacturing dataset with 552 observations. Step 1 exploratory data analysis revealed severe multicollinearity (condition number 1.59×10^4), heavy-tailed distributions, and strong correlations necessitating advanced monitoring approaches. Step 2 developed a

full-dimensional Hotelling's T^2 chart with iterative refinement, identifying 25 OOC observations (4.53%) at $\alpha = 0.01$ using chi-square UCL=259.48. Step 3 employed PCA-based monitoring with $k = 64$ components (90% variance), detecting 18 OOC observations (3.26%) at $\alpha = 0.01$ through combined T^2 (UCL=109.06, F-based) and Q (UCL=29.63) statistics.

Comparing both methods revealed 34.4% agreement (11/32 unique OOC), with PCA capturing 44.0% (11/25) of full-dimensional outliers while providing enhanced diagnostics via separate T^2 and Q charts. Contribution analysis in both steps identified localized variable groups as primary drivers of abnormal behavior. The PCA-based approach demonstrated superior stability, reduced false alarm rates, and actionable process diagnostics. We recommend Step 3 PCA-based T^2 and Q monitoring as the preferred framework for Phase II deployment in this high-dimensional, multicollinear manufacturing process, enabling early detection of assignable causes and facilitating targeted corrective actions.

Acknowledgements

The authors acknowledge course contents of course IENG551.

References

Page, E. S. (1954). Continuous inspection schemes. *Biometrika 41*(1-2), 100–115.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*(11), 559–572.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics 1*(3), 239–250.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. New York: D. Van Nostrand Company.