Original Research

Performance of Cluster-Based Logistic Profile Monitoring Under Existence of Different Linkage Functions

Davood Saremian[1]. Rassoul Noorossana[2*]. Sadigh Raissi[1]. Paria Soleimani[1]

Received: 03 Oct 2023 / Accepted: 04 Sep 2024 / Published online: 19 Sep 2024

* Corresponding Author, rnoorossana@uco.edu

1- Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran
2- Information Systems and Operations Management Department, College of Business, University of Central Oklahoma, Edmond, OK, 73034, United States

Abstract

During industrial process monitoring, a common practice involves analyzing the relationship between a measured outcome (response variable) and other relevant factors (descriptive variables), which is called a profile. However, the perceptible challenge in this issue is the reliable estimation of profile parameters that can deviate significantly under the influence of outliers. Saremian et al. (2021) addressed the challenge of parameter estimation within generalized linear profiles during Phase I of a research investigation. They proposed a robust methodology for this purpose. Their results showed that incorporating a clustering approach, particularly with a complete linkage function, yields superior control charts parameter for monitoring binary logistic profiles compared to the traditional, non-clustering method. The performance of cluster-based control charts in monitoring logistic profiles is evaluated under varying linkage function conditions in this paper. The aim is to improve the performance of cluster-based method by evaluating the effect of using different linkage functions, including complete, average, single, weighted, centroid, median, and ward linkage. The simulation runs demonstrated a significant improvement in the Hotelling $\boldsymbol{T^2}$ control chart's ability to detect process deviations when combined with a clustering approach. Furthermore, employing various linkage methods, such as average, centroid, and ward's linkage, demonstrably yields more accurate control chart parameter estimates compared to complete linkage. Therefore, the application of the linkage functions presented in this study has led to an enhancement in the performance of the cluster-based method.

Keywords - Binary logistic profiles**;** Linkage functions; Phase I analysis; Hotelling T^2;Cluster-based control chart

INTRODUCTION

Profile monitoring has garnered considerable attention within the industrial engineering research landscape over the past two decades, positioning it as a noteworthy statistical quality control method. This innovative statistical quality control methodology explores the functional relationship between a critical process output (response variable) and a set of control factors influencing its behavior (descriptive variables). Previous scholarly works have shown, different profile monitoring methods according to the type of relationship and phase of monitoring have been developed. Linear profiles were introduced by Kang and Albin (2000) and then other researchers including Soleimani et al. (2009), Chen and Nembhard (2010), Noorossana et al. (2016), Mahmood et al. (2019), Moheghi et al. (2020), Ghasemi et al. (2023), Pakzad et al. (2023), Adibfar et al. (2023) developed

this type of profiles. Researchers, including Williams et al. (2006), Steiner et al. (2016), Pan et al. (2019) (See the reviews of Noorossana et al. (2011) and Maleki et al. (2018).) and Cheng et al. (2023), have also explored the development and application of nonlinear profiles. Sometimes, it is necessary to monitor industrial or service processes whose response variables are discrete or follow an abnormal distribution. For example, the response variable of processes such as the number of alloy fastener failures in airplane structures when evaluated at a given load (Montgomery 2006) or the mortality rate of insects exposed to the different doses of a chemical gas (Dobson and Barnett 2008), is binary and distribution function is abnormal. Yeh et al. (2009) were the first to investigate the application of statistical process control methods to monitor binary logistic profiles during Phase I analysis. Five T^2Hotelling control charts were utilized to monitor profiles defined by binary outcomes. Among the five T^2 Hotelling control charts investigated, the one based on sample average and intra-profile pooling exhibited superior performance in detecting process deviations. Several researchers, including Shang et al. (2011), Koosha and Amiri (2012), Peynabar et al. (2012), Amiri et al. (2014), Shadman et al. (2014 & 2015), Noorossana et al. (2015), Shang et al. (2017), Izadbakhsh et al. (2018), and Bandara et al. (2020), have proposed various methods for monitoring profiles that follow a generalized linear family distribution. In all previous research, it has been assumed that the historical data set is free of outliers or contaminated data. However, if these assumptions are not met, this data will likely affect the accuracy of the estimated parameters of the generalized linear profiles and control charts. For this purpose, robust estimation methods that are more sensitive to this type of data have been proposed. Due to the focus of this paper on the field of monitoring generalized linear profiles, the most important papers that are developed to improve the estimation of such profiles in the presence of outlier data or contamination within the sample are presented. A limited body of research exists on the development of robust estimators for generalized linear profiles.

Hakimi et al. (2017, 2018) addressed the challenge of outlier contamination in logistic profile parameter estimation by proposing a trio of robust estimation methods: weighted maximum likelihood estimation (WMLE), redescending M-estimator (RM), and weighted redescending M-estimator (WRM). Their simulation results showed that the robust estimation methods have higher accuracy and precision for estimating profile parameters than the maximum likelihood estimation (MLE) method. The results also indicate that the weighted redescending M-estimator (WRM) outperforms the alternative robust methods in terms of estimation accuracy. Moheghi et al. (2020) proposed a robust estimation method based on the C-R estimator (Cantoni and Ronchetti (2001)) for generalized linear profiles when the outlier data is present in the data set. They evaluated the performance of their novel robust method against the widely used maximum likelihood estimation (MLE) method. The results showed that the proposed robust method performs better than the MLE method. Saremian et al. (2021& 2022) proposed a robust cluster-based method based on agglomerative hierarchical clustering for the monitoring of generalized linear profiles. In these papers, an agglomerative hierarchical cluster-based method with a complete linkage function has been used to estimate the parameters of control charts. Simulations revealed that employing a clustering approach based on the complete linkage function yielded superior performance compared to alternative non-robust methods and the robust estimators WRM, RM, and WMLE, particularly in scenarios with medium to large shifts. Phase I, as outlined previously, is primarily concerned with achieving accurate estimations of the true parameter values and then use the estimated parameters to determine the control limits. Consequently, biased estimations of the profile parameters can result in the establishment of inaccurate control limits, thereby compromising the effectiveness of control charts during Phase II. In order to enhance the performance of the clustering-based approach, this study employed linkage functions beyond the complete linkage function. The existing literature offers various linkage methods for cluster aggregation, including complete, average, single, weighted, centroid, median, and ward. In this paper, the impact of varying linkage functions on the performance of a cluster-based method is investigated. Furthermore, a comparative analysis of the cluster-based approach with a non-clustering alternative is conducted. The paper is structured as follows. In Section 2, the estimation process for the parameters of the binary logistic profiles is presented. Section 3 introduces a Hotelling's T^2 control chart constructed using the sample mean and and intra-profile pooling. A description of clustering and non-clustering methods for monitoring Phase I of logistic profiles is presented in Section 4. In Section 5, a case study to validate the proposed method's real-world applicably is presented. The performance of the clustering method based on different linkage functions and the non-clustering method is compared in Section 6. Finally, in Section 7, the conclusion and future research are presented.

BINARY LOGISTIC PROFILE PARAMETER ESTIMATION

Logistic regression profiles are core models within the generalized linear model family, characterized by response variables following Bernoulli or binomial distributions. Consider the dataset consisting of observations $\{\boldsymbol{X}_i, z_{ij}\}_{i=1}^{n}$, where $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ and z_{ij} is the j^{th} binary response variables in the i^{th} descriptive variable. The probability of success for $z_{ij}$is represented by π_i, where i ranges from 1 to n and j from 1 to k. Here, k denotes the number of Bernoulli variables at each level and $E(z_{ij}) = \pi_i$ and $Var(z_{ij}) = \pi_i(1-\pi_i)$. The probability of success, denoted as $\pi_i = \pi(\boldsymbol{X}_i)$, is a dependent variable whose value is determined by a function of the independent variable $\boldsymbol{X}_i$. Logistic regression models employ various link functions to

characterize the relationship between the response variable and its associated predictors(s). The logit link function, the most commonly employed link function, transforms the probability of an event occurring into the logarithm of the odds of that event:

$$g(\pi_i) = log\,\frac{\pi_i}{1-\pi_i} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \qquad (1)$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the model parameter vector. Usually the value of $x_{i1} \equiv 1$ so that β_1 is the intercept parameter.

$$\pi_i = \frac{\exp(X_i^T \beta)}{1+\exp(X_i^T \beta)} = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \qquad (2)$$

In Equation 2, η_i is equal to $\eta_i = X_i^T \beta = \sum_{j=1}^{p} \beta_i\, x_{ij}$.

It is presupposed that the data are clustered in a manner such that the ith group of descriptive variables contains mi observations, where i is an integer between 1 and n. The sum of the number of observations is denoted by $M = \sum_{i=1}^{n} m_i$. The response variable is $y_i = \sum_{j=1}^{m_i} z_{ij}$, where z_{ij} is equal to the j^{th} observation in the i^{th} set of the descriptive variables. Consequently, the data adheres to a binomial distribution characterized by parameters (m_i, π_i), $E(y_i) = m_i\pi_i$ and $var(y_i) = m_i\pi_i = m_i \times \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \times \frac{1}{1+\exp(\eta_i)}$.

Assuming independence among the binomial observations, the joint likelihood function for the variables $Y_1, Y_2, \ldots, Y_N$ is expressed as:

$$L(\pi; y) = \prod_{i=1}^{n} \binom{m_i}{y_i} {\pi_i}^{y_i} (1-\pi_i)^{m_i - y_i} \qquad (3)$$

where $\pi = (\pi_1, \pi_2, \ldots, \pi_n)^T$ and $y = (y_1, y_2, \ldots, y_n)^T$.

Taking the natural logarithm of Equation 3 results in Equation 4:

$$\iota(\beta : y) = \sum_{i=1}^{n} \log \binom{m_i}{y_i} + \sum_{i=1}^{n} \sum_{j=1}^{p} y_i \beta_i x_{ij} - \sum_{i=1}^{n} m_i \log[1 + \exp(\sum_{j=1}^{p} \beta_j x_{ij})] \qquad (4)$$

Differentiation of Equation 4 with respect to the parameter **β** yields Equation 5:

$$\frac{\partial \iota(\beta : y)}{\partial \beta} = X^T y - \sum_{i=1}^{n} m_i \pi_i X_i = X^T (y - \mu) \qquad (5)$$

Iterative weighted least squares is the predominant computational method for approximating maximum likelihood estimates, as introduced by McCullagh and Nelder in 1989. Parameter estimation for binary logistic models is conducted using this method.

T^2 HOTELING CONTROL CHART WITH INTRA PROFILE VARIANCE-COVARIANCE ESTIMATOR

A T^2 Hotelling control chart, employing sample means and intra-profile covariance, has demonstrated efficacy in monitoring binary logistic profiles, as validated by the research of Yeh et al. (2009) and Saremian et al. (2021); Furthermore, we utilize a control chart based on clustering and non-clustering modes The mean and covariance are calculated as follows:

$$T_{I,t}^2 = (\hat{\beta}_t - \bar{\beta})^T S_I^{-1} (\hat{\beta}_t - \bar{\beta}) \qquad (6)$$

$$\bar{\beta} = 1/k \sum_{t=1}^{k} \hat{\beta}_t, S_I = \frac{1}{k} \sum_{t=1}^{k} var(\hat{\beta}_t) = \frac{1}{k} \sum_{t=1}^{k} (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1}$$

MONITORING OF LOGISTICS PROFILES

I. Monitoring of logistics profiles in Phase I based on non-clustering method

In Phase I, a customary procedure is to gather a group of profiles to enable their ongoing tracking. The parameters of these profiles are estimated using the maximum likelihood estimator method and then the upper control limit is calculated. If all profiles are within the control limit, this limit is considered the final limit. Otherwise, we excluded the profiles that were outside of the control limit and revised the control limit. This method is repeated so that there is no out-of-control profile remain in the dataset and the process is stabilized.

II. Monitoring of logistics profiles in Phase I based on clustering method

- **Clustering method**

Clustering represents a data analysis technique that categorizes data points into distinct groups, termed clusters. This grouping process hinges on the principle of similarity – data points within a cluster exhibit a greater degree of resemblance compared to those in other clusters. Clustering algorithms can be broadly categorized into several distinct methodologies. These include connectivity-based methods (e.g., hierarchical clustering), centroid-based approaches, distribution-based techniques, density-based algorithms, and fuzzy clustering. As outlined in Section 1, the proposed method integrates hierarchical clustering with maximum likelihood estimation. Therefore, the subsequent sections will delve deeper into the specifics of hierarchical clustering.

- *Connectivity-based clustering (Hierarchical clustering)*

One of the most popular and easy-to-understand methods of segmentation is hierarchical clustering. This method follows two approaches, including bottom-up or top-down, based on the direction of progress. They are agglomerative and divisive hierarchical approaches, respectively. Agglomerative clustering starts with one cluster per observation, and during an iterative process, the two clusters that are most similar/ least different from each other are aggregated to form a larger cluster and continue to do so until only a single cluster is left. Hierarchical divisive algorithms are the opposite of hierarchical agglomerative algorithms, run from a single cluster that contains all observations, and then in a sequential process, the observations are split into the other clusters. A measure of dissimilarity is necessary to decide where clusters should be joined or where a cluster should be divided. This is accomplished by utilizing an appropriate distance metric in conjunction with a linkage function. The following are the steps of the agglomerative hierarchical clustering method:

Step 1: Assign a cluster to each of the profiles.
Step 2: Calculate the similarity matrix. (Calculate the distance between clusters)
Step 3: Employ a well-chosen linkage function to combine the two aforementioned clusters.
Step 4: Recalculate the similarity matrix based on the aggregated clusters
Step 5: Iterate through Steps 3 and 4 until a single cluster is obtained.
There are a variety of inter-group proximity measurements to aggregate clusters, such as complete, average, single, weighted, centroid, median, and ward, which are described below.

- Complete Linkage

This methodology involves determining the maximum inter-cluster distance between elements of clusters A and B.

$$D(A,B) = \max\{d(\boldsymbol{y}_i, \boldsymbol{y}_j), \text{for } \boldsymbol{y}_i \; in \; A \; and \; \boldsymbol{y}_j \; in \; B\} \quad (7)$$

- Average Linkage

This approach leverages the average inter-cluster distance between points in cluster A and cluster B.

$$D(A,B) = \frac{1}{n_A n_B}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d(\boldsymbol{y}_i, \boldsymbol{y}_j) \quad (8)$$

- Single Linkage

This approach utilizes the minimum inter-cluster distance between data points from clusters A and B, iteratively merging the closest data points at each step.

$$D(A,B) = \min\{d(\boldsymbol{y}_i, \boldsymbol{y}_j), \text{for } \boldsymbol{y}_i \; in \; A \; and \; \boldsymbol{y}_j \; in \; B\} \quad (9)$$

- Weighted Linkage

The weighted average method employs a recursive strategy to determine the distance between clusters. If cluster A is created by uniting clusters p and q, the distance between A and another cluster B is determined by calculating the mean of the distances between p and s, and q and s.

$$D(A,B) = \frac{(d(p,s)+d(q,s))}{2} \quad (10)$$

- Centroid Linkage

This method employs the minimum Euclidean distance between cluster centroids to assess their proximity.

$$D(A,B) = d(\bar{\boldsymbol{y}}_A, \bar{\boldsymbol{y}}_{\boldsymbol{B}})$$
$$\bar{\boldsymbol{y}}_{\boldsymbol{A}} = \sum_{\boldsymbol{i=1}}^{n_A} \boldsymbol{y}_i / n_A$$
$$\bar{\boldsymbol{y}}_{AB} = \frac{n_A \bar{\boldsymbol{y}}_{\boldsymbol{A}} + n_B \bar{\boldsymbol{y}}_{\boldsymbol{B}}}{n_A + n_B} \quad (11)$$

- Median Linkage

The distance between clusters is measured by determining the smallest Euclidean distance between their respective weighted means (centroids).

$$\mathbf{m}_{AB} = \frac{1}{2}(\bar{\boldsymbol{y}}_{\boldsymbol{A}} + \bar{\boldsymbol{y}}_{\boldsymbol{B}}) \quad (12)$$

- Ward Linkage

$$SSE_A = \sum_{i=1}^{n_A} (\boldsymbol{y}_i - \bar{\boldsymbol{y}}_{\boldsymbol{A}})\acute{}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_{\boldsymbol{A}})$$
$$SSE_B = \sum_{i=1}^{n_B} (\boldsymbol{y}_i - \bar{\boldsymbol{y}}_{\boldsymbol{B}})\acute{}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_{\boldsymbol{B}})$$
$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (\boldsymbol{y}_i - \bar{\boldsymbol{y}}_{\boldsymbol{AB}})\acute{}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_{\boldsymbol{AB}})$$
$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B) \quad (13)$$

- **Cluster-based Robust Methodology for Binary Logistic Profile Monitoring**

This section introduces the robust clustering method presented by Saremian et al. 2021. This method is described in several steps as follows:

Step 1: Employ maximum likelihood estimation to determine the parameters for all profiles within the dataset.
Step 2: Compute the mean vector and variance-covariance matrix utilizing the parameter estimates obtained in Step 1.
Step 3: Based on the variance-covariance matrix calculated in Step 2, calculate the similarity matrix (Equation 14).

$$S_{ij} = (\hat{\beta}_i \_ \hat{\beta}_j)^T \hat{V}_D^{-1} (\hat{\beta}_i \_ \hat{\beta}_j) \quad (14)$$

Step 4: Employ hierarchical clustering to group profiles based on the similarity matrix, utilizing an appropriate linkage function. Continue the hierarchical clustering algorithm until at least half of the profiles are in one cluster. Consider the profiles in this cluster as the main cluster.
Step 5: Compute the mean profile parameters of the main cluster.
Step 6: Calculate the value of the T_I^2 statistic (Equation 6) for all profiles outside the main cluster. Compare them with the upper control limit. If the values of the T_I^2 statistic is less than the upper control limit, add them to the main cluster.
Step 6: Iteratively recalculate the mean profile parameters of the main cluster until no further profiles can be incorporated.
Step 7: Calculate the mean vector and variance-covariance matrix of the primary cluster, thereby establishing the control chart parameters for Phase I.

CASE STUDY

In this section, insects' mortality rate data (Table 1) given by Annette J. Dobson and Adrian G. Barnett (2008) is used to demonstrate the proposed methodology's application in detecting contaminated data for logistic profiles. These insects were exposed to different doses of gaseous carbon disulfide for five hours, and their mortality rate was recorded. The descriptive variable is the carbon disulfide dose, measured in milligrams per liter (mg/L), and the response variable variable represents insect mortality in response to these chemical.

TABLE 1
INSECT MORTALITY DATA

Dose of carbon disulfide	Number of insects	Number of deaths
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

The first-order logistic regression with parameters β = (-60.72, 33.27) and variance-covariance matrix $\widehat{\boldsymbol{V}}_D = \begin{pmatrix} 26.8397 & -15.0821 \\ -15.0821 & 8.4805 \end{pmatrix}$, is fitted to this data using Eq.6. Then different step shifts in the parameters of profiles 3, 7, 11, and 15 of size $\beta_0 + \delta_1\sigma_1$ and $\beta_1 + \delta_2\sigma_2$ with $\delta_1 = 0.8655$ and $\delta_2 = 0$ are assumed to occur. Model parameters for the logistic regression are obtained using the maximum likelihood estimation technique (Eq. 5). The estimated parameters are tabulated in Table 2.

TABLE 2
ESTIMATED PARAMETERS OF BINARY LOGISTIC PROFILES

Profile	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	Profile	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$
1	-69.521	38.159	9	-63.685	34.736
2	-54.331	29.711	10	-60.531	33.044
3	-48.267	26.933	11	-55.025	30.564
4	-53.208	29.078	12	-61.236	33.553
5	-56.263	30.903	13	-78.820	43.182
6	-61.069	33.391	14	-64.953	35.485
7	-56.247	31.226	15	-59.217	32.856
8	-62.304	34.175	16	-54.039	29.502

The covariance matrix, denoted as $\widehat{\boldsymbol{V}}_D = \begin{pmatrix} 58.6149 & -32.2177 \\ -32.2177 & 17.7184 \end{pmatrix}$, is computed utilizing Equation 6 and the previously determined parameter estimates. Eq. 14 defines the method used to calculate pairwise similarity scores between data points, which are presented in Table 3. The profiles are clustered based on the hierarchical clustering method with a complete linkage function. The preliminary main cluster contains profiles 2, 4, 5, 6, 8, 9, 10, 12, 14, as depicted in Figure 1. Based on the data points within the main clusters, the mean vector is calculated to be (-38.8966, 4.7901). The T^2 values for the profiles belonging to the main cluster are determined by applying Equation 6.

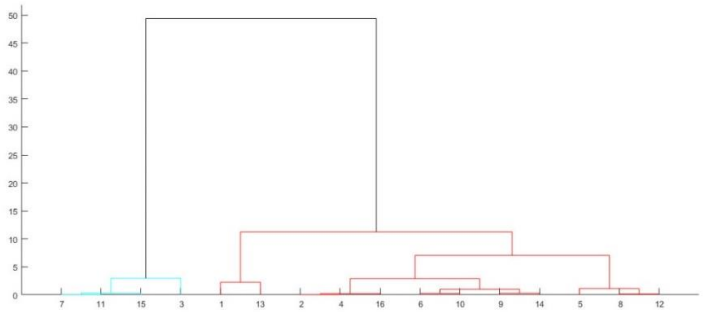

FIGURE 1
THE DENDROGRAM OF THE COMPLETE-LINKAGE HIERARCHICAL CLUSTERING FOR LOGISTIC PROFILES

The T^2 statistics are calculated for profiles 1, 13, 3, 7, 11, 15. These values are then compared to an upper control limit (UCL) set at 10.5534. The T^2 values for profiles 1 and 13 are smaller than the upper control limit; hence these profiles were deemed suitable for incorporation into the main data cluster, as shown in Table 4. The mean of the primary cluster is recalculated to accommodate newly incorporated profiles.

TABLE 4
THE T^2 VALUES FOR THE PROFILES EXCEED THE CONTROL LIMITS OF THE PRELIMINARY MAIN CLUSTER.

Profile	1	13	3	7	11	15
T^2	2.998	6.631	33.925	22.340	23.452	21.995

The T^2 values associated with profiles 3, 7, 11, 15 which are not in the main cluster are calculated again.

TABLE 5
THE T^2 VALUES FOR THE PROFILES SURPASS THE CONTROL LIMITS OF THE SECONDARY MAIN CLUSTER.

Profile	3	7	11	15
T^2	33.7726	21.7052	22.9000	21.1113

As shown in Table 5, all T^2 values are greater than UCL. Consequently, no additional profiles qualify for inclusion in the primary cluster, resulting in algorithm termination. The results show that the clustering method correctly identified profiles 3, 7, 11, and 15 as the out-of-control profiles and other profiles as the in-control profiles.

PERFORMANCE EVALUATION OF CLUSTERING METHOD BASED ON DIFFERENT LINKAGE FUNCTIONS AND NON-CLUSTERING METHOD IN PHASE I, UNDER STEP SHIFT

As mentioned, Saremian et al. (2021) compared the performance of a clustering approach versus the non-clustering approach for generalized linear profiles. The authors restricted their evaluation of the proposed method's performance to the complete linkage function. A comparative analysis is conducted to evaluate the performance of a clustering-based approach against a non-clustering-based method within Phase I. The clustering approach incorporates seven distinct linkage functions. The performance of these methods is compared based on Monte Carlo simulations for different step shifts in profile parameters. For this purpose, six indicators (see Saremian et al. (2021) and Chen et al. (2015)) are used as follows:

- **Probability of signal (POS)**

The probability of signal is the most common indicator for determining the power of control charts in Phase I. The probability of signal (POS) serves as a measure of the likelihood that at least one out-of-control signal will be detected within the entire sample set.

- **Fraction correctly classified (FCC)**

This index reflects the control chart's power in correctly classifying profiles. Fraction correctly classified is calculated as the ratio of correctly identified in-control and out-of-control profiles to the total number of analyzed profiles.

- **Sensitivity**

This indicator, known as Sensitivity, reflects the control chart's ability to correctly identify out-of-control profiles. It is calculated as the proportion of correctly identified out-of-control profiles relative to the total number of out-of-control profiles.

- **Specificity**

This index focuses on the control chart's ability to correctly identify in-control profiles. It is calculated as the number of these profiles to the total number of in control profiles.

- **False Positive Rate (FPR)**

This index, known as the False Positive Rate (FPR), quantifies the weakness of the control chart in accurately detecting out-of-control profiles. The FPR is expressed as the ratio between the number of misidentified out of control profiles and the total number of actual out-of-control profiles.

- **False Negative Rate (FNR)**

It indicates the fault of the control chart in correctly identifying the control profiles. FNR is computed by dividing the count of misidentified in-control profiles by the total count of in-control profiles.

TABLE 3
SIMILARITY MATRIX

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.0000	4.9354	28.5688	5.8726	3.0964	2.7224	16.1872	0.9194	5.2652	4.4186	17.5124	1.4479	2.2600	3.0587	14.8386	6.2909
2	4.9354	0.0000	31.6184	0.0455	1.7846	0.8260	21.5002	1.7631	2.8496	1.2093	22.3944	1.0374	10.2442	2.3404	21.6483	0.2360
3	28.5688	31.6184	0.0000	33.1564	19.1968	36.3634	1.9643	25.8600	49.3724	42.3942	1.4773	28.8150	45.6649	43.3329	2.9637	37.1854
4	5.8726	0.0455	33.1564	0.0000	2.3106	1.0594	23.0538	2.3693	2.8926	1.2618	23.9324	1.4941	11.2633	2.5932	23.3090	0.1202
5	3.0964	1.7846	19.1968	2.3106	0.0000	2.7614	11.0106	0.8609	7.0737	4.5373	11.7195	1.1251	10.1165	5.1159	11.0187	3.3104
6	2.7224	0.8260	36.3634	1.0594	2.7614	0.0000	23.9857	1.1296	0.9967	0.2725	25.2068	0.4906	5.4916	0.4319	23.4413	0.9095
7	16.1872	21.5002	1.9643	23.0538	11.0106	23.9857	0.0000	15.1142	34.5246	29.1938	0.0358	17.7035	29.0914	29.1147	0.1509	26.2395
8	0.9194	1.7631	25.8600	2.3693	0.8609	1.1296	15.1142	0.0000	3.9868	2.5114	16.1740	0.1423	5.1579	2.2755	14.4894	2.8763
9	5.2652	2.8496	49.3724	2.8926	7.0737	0.9967	34.5246	3.9868	0.0000	0.3468	36.0450	2.7857	5.5423	0.2981	33.6752	2.0507
10	4.4186	1.2093	42.3942	1.2618	4.5373	0.2725	29.1938	2.5114	0.3468	0.0000	30.4973	1.4903	6.4422	0.3436	28.6819	0.7865
11	17.5124	22.3944	1.4773	23.9324	11.7195	25.2068	0.0358	16.1740	36.0450	30.4973	0.0000	18.7897	30.9840	30.5825	0.3149	27.2196
12	1.4479	1.0374	28.8150	1.4941	1.1251	0.4906	17.7035	0.1423	2.7857	1.4903	18.7897	0.0000	5.4045	1.4853	17.1720	1.8006
13	2.2600	10.2442	45.6649	11.2633	10.1165	5.4916	29.0914	5.1579	5.5423	6.4422	30.9840	5.4045	0.0000	3.8553	26.7627	10.8345
14	3.0587	2.3404	43.3329	2.5932	5.1159	0.4319	29.1147	2.2755	0.2981	0.3436	30.5825	1.4853	3.8553	0.0000	28.1589	2.0581
15	14.8386	21.6483	2.9637	23.3090	11.0187	23.4413	0.1509	14.4894	33.6752	28.6819	0.3149	17.1720	26.7627	28.1589	0.0000	26.3965
16	6.2909	0.2360	37.1854	0.1202	3.3104	0.9095	26.2395	2.8763	2.0507	0.7865	27.2196	1.8006	10.8345	2.0581	26.3965	0.0000

The control limit is calculated based on 100,000 simulation runs and the probability of false alarm is equal to 0.05 (Jensen et al., Yeh et al., Saremian et al.). Building upon the findings of Saremian et al. (2021), the T^2 statistic calculated with sample average and intra-profile pooling emerges as the most effective method for monitoring processes following logistic distributions. Therefore, this control chart is used. We used the same simulation settings applied by Saremian et al. (2021). Suppose $p = 2$ and the in-control chart parameters are $\boldsymbol{\beta}_0 = (3, 2)^T$. The design matrix is as follows:

$$X = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ \log(0.1) & \log(0.2) & \cdots & \log(0.8) & \log(0.9) \end{pmatrix}^T$$

The upper control limits (UCLs) are determined based on three different repetition levels: 30, 50, and the sample size is equal to 30. It is presented in Table 6. An analysis of the simulated results presented in Table 6 reveals that the upper control limits (UCLs) derived using the clustering method are consistently lower than those obtained with the non-clustering method. Furthermore, for both clustering and non-clustering T_I^2 control charts, the UCLs tend to increase as the sample size grows, while the number of iterations remains constant.

I. Impact of Step Shifts in Binary Logistic Profiles

To evaluate the performance of control charts, we create different step shifts in the profile parameters. Step shifts are defined as $\tilde{\boldsymbol{\beta}}_0 = \boldsymbol{\beta}_0 + \boldsymbol{\Delta}$, and the shift level based on the non-central parameter is defined as NCP=$\boldsymbol{\Delta}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Delta}$ and $\boldsymbol{\Delta} = (\delta_1\sigma_1, \delta_2\sigma_2)^T$. This paper assumes a step shift in the data, affecting the last third of the profiles. The investigation employed different m and k values to evaluate the control chart's performance against the predetermined criteria. This section only presents results for m = 30 and k = 30, 50, and 100 for brevity.

TABLE 6
THE SIMULATED UPPER CONTROL LIMITS FOR T_I^2 CONTROL CHART

m	*k*	T_I^2							
		Non- cluster	Complete	Average	Single	Weighted	Median	ward	Centroid
30	30	15.080	12.652	12.557	12.487	12.640	12.630	12.682	12.537
	50	13.935	11.882	11.787	11.734	11.865	11.858	11.917	11.796
	100	13.113	11.302	11.247	11.195	11.311	11.281	11.361	11.265

II. Assessing Control Charts Performance via Probability of Signal Index

In Phase I of control chart analysis, the probability of a signal index is a commonly employed metric to assess the chart's effectiveness in detecting out-of-control conditions. This section investigates the performance of the T_I^2 control chart under various step shift scenarios. The results demonstrate that a control chart grounded in clustering, regardless of the linkage function used, performs better than a non-clustering control chart at identifying out-of-control profiles. Consistent with the data presented in Figures 2, 3, and 4, the cluster-based (CB) T_I^2 control chart employing average and centroid linkage functions exhibited superior performance compared to both other CB control charts and the non-cluster-based (NCB) control chart. However, the difference is small between the values of the probability of signal index based on different linkage functions.

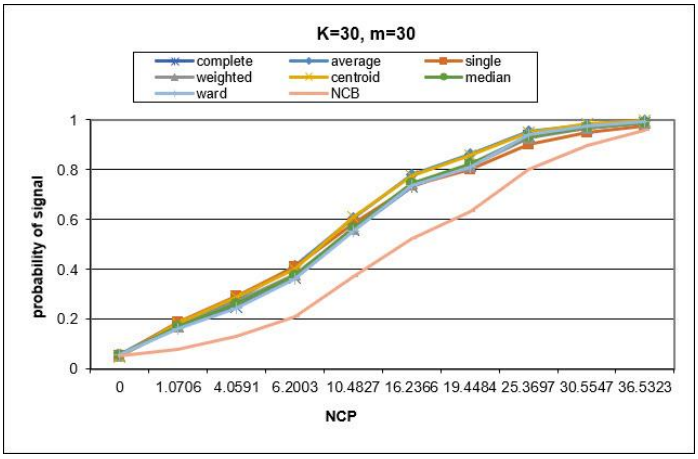

FIGURE 2
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING PROBABILITY OF SIGNAL INDEX (K = 30, M = 30)

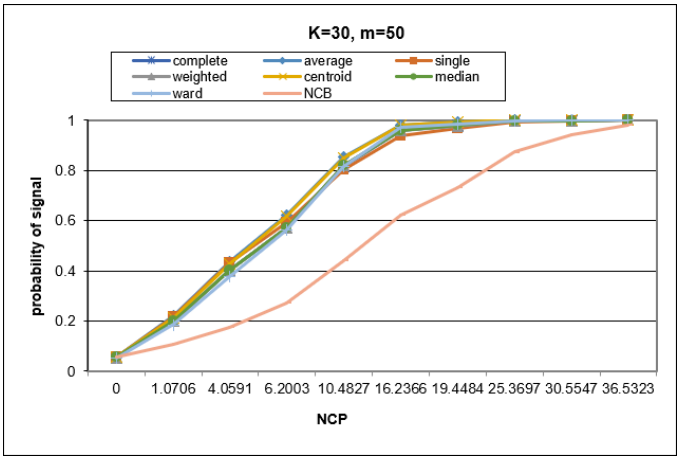

FIGURE 3
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING PROBABILITY OF SIGNAL INDEX (K = 30, M = 50)

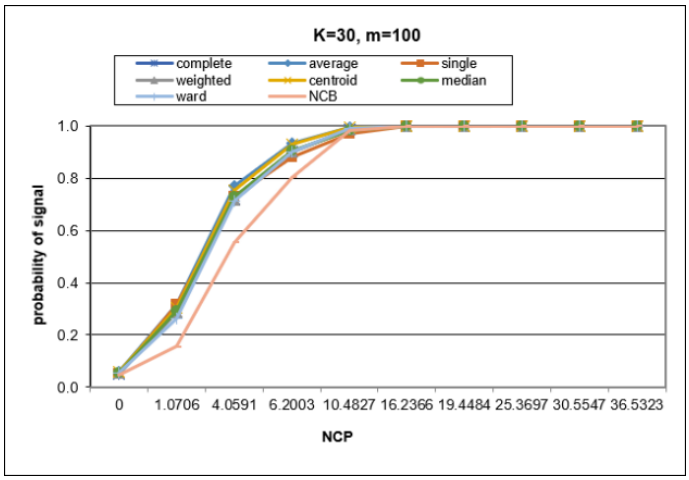

FIGURE 4
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING PROBABILITY OF SIGNAL INDEX (K =30, M = 100)

III. Assessing Control Charts Performance via Fraction Correctly Classified Index

As depicted in Figures 5, 6, and 7, the results indicate that the index values for small shifts exhibit comparable outcomes for both CB and NCB control charts. However, when faced with medium to large shifts, CB control charts, regardless of the linkage function employed, consistently outperform NCB control charts. These results also show that the performance of the CB control charts based average, centroid and ward linkage functions is better than NCB control chart.

IV. Assessing Control Charts Performance via Sensitivity Index

The simulation results show that as the amount of step shift increases, the values of the index also increase. As shown in Figures 8 to 10, the T_I^2 control chart performs better in clustering mode compared to the non-clustering mode. Furthermore, the cluster-based control chart utilizing average, centroid, and ward linkage functions exhibited superior performance compared to both methods employing other linkage functions and the non-clustering approach altogether.

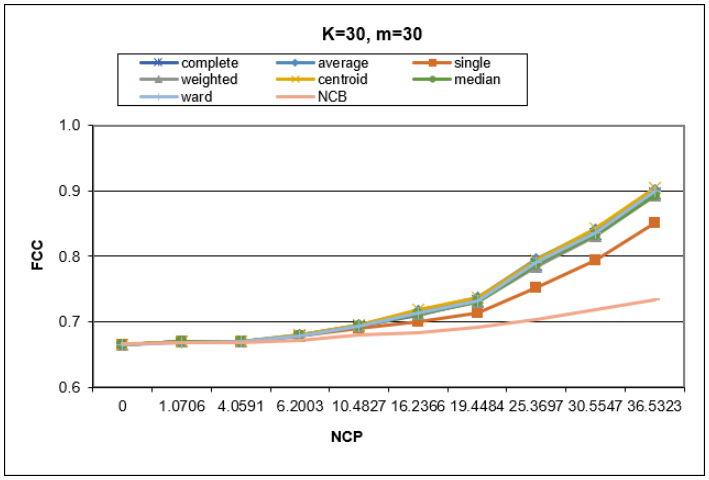

FIGURE 5
EVALUATION OF CONTROL CHART PERFORMANCE UNDER A STEP SHIFT USING THE FRACTION CORRECTLY CLASSIFIED INDEX (K = 30, M = 30)

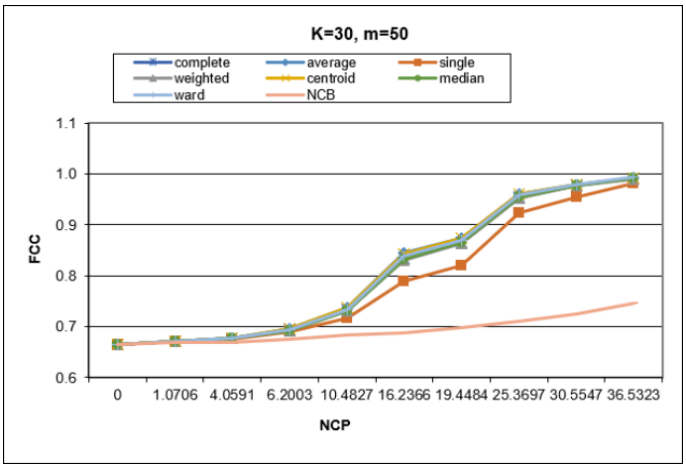

FIGURE 6
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING FRACTION CORRECTLY CLASSIFIED INDEX (K = 30, M = 50)

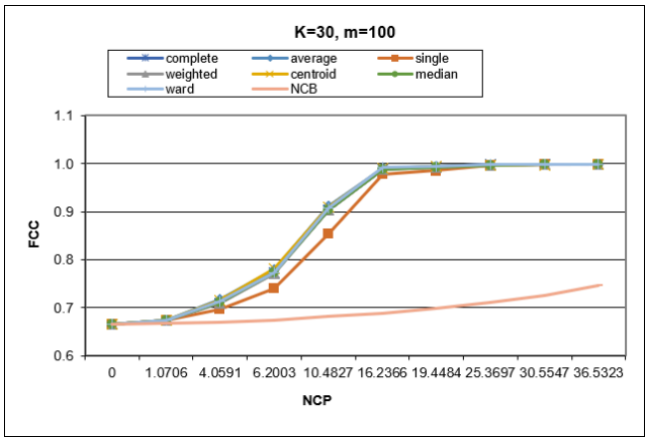

FIGURE 7
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING FRACTION CORRECTLY CLASSIFIED INDEX (K =30, M = 100)

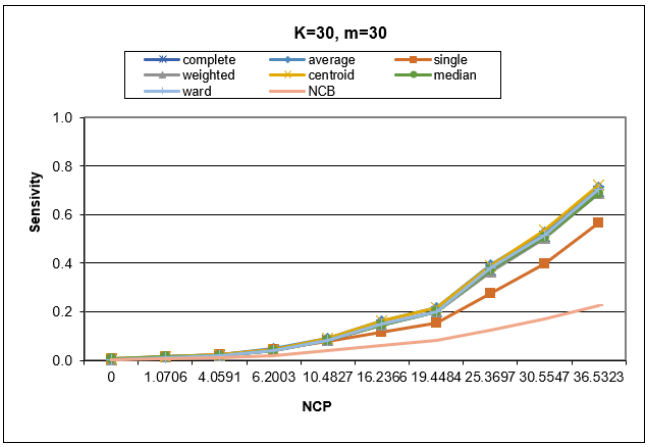

FIGURE 8
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING SENSITIVITY INDEX (K = 30, M = 30)

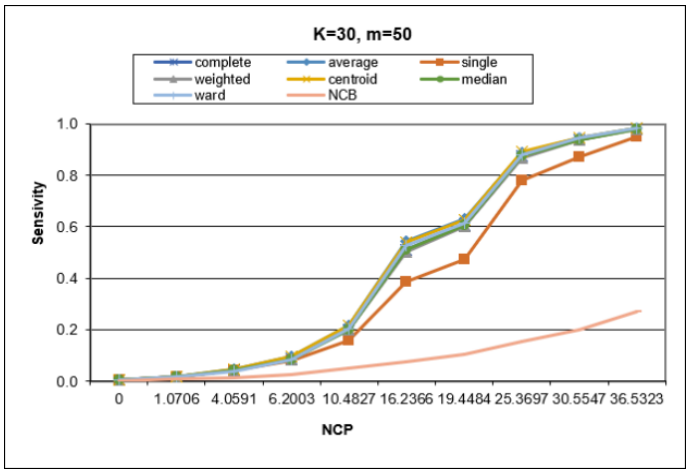

FIGURE 9
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING SENSITIVITY INDEX (K =30, M = 50)

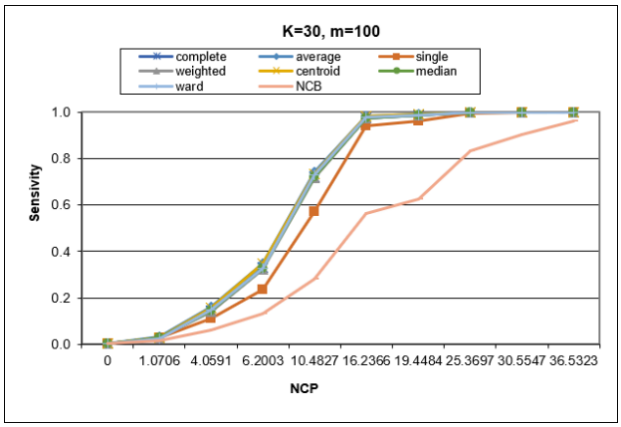

FIGURE 10
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING SENSITIVITY INDEX (K = 30, M = 100)

IV. Assessing Control Charts Performance via Specificity Index

Figures 11 to 13 show that for small and medium shifts, there's little difference between the clustering and non-clustering methods. The cluster-based control chart exhibits progressively superior performance compared to the non-clustering control chart as the magnitude of the step shifts increases. The simulation results revealed that the clustering method employing average, centroid, and ward linkage functions achieved superior performance compared to methods utilizing other linkage functions.

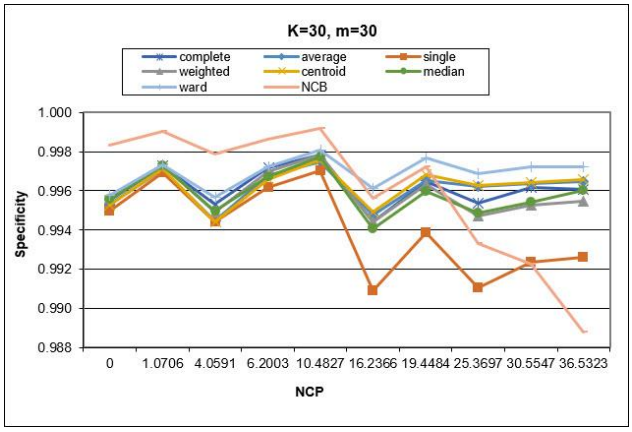

FIGURE11
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING SPECIFICITY INDEX (K = 30, M = 30)

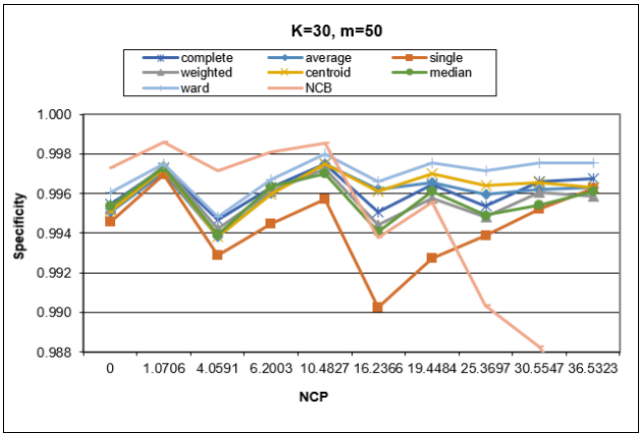

FIGURE12
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING SPECIFICITY INDEX (K = 30, M = 50)

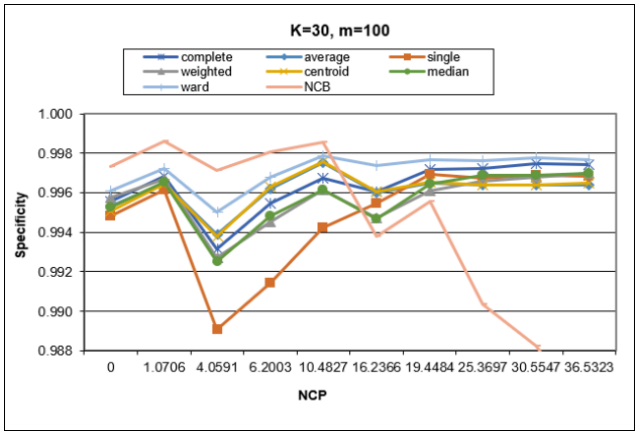

FIGURE13
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING SPECIFICITY INDEX (K = 30, M = 100)

V. Evaluation of the performance of control charts based on false positive rate index

To assess the impact of data clustering, the false positive rate index values were calculated for both the T_I^2 control chart in its clustered and non-clustered forms. Analysis of the simulation data indicated that the clustering method achieved statistically significantly better results than the non-clustering method. The results revealed a consistent downward trend in the index across both clustering and non-clustering methods as sample size and the number of iterations increased. Our findings indicate that the clustering method (CB) utilizing ward, centroid, and average linkage exhibited superior performance compared to CB methods employing complete, single, median, and weighted linkage approaches. The performance of the cluster-based method, as evidenced by Figures 14-16, surpasses that of the non-cluster based method for all linkage functions.

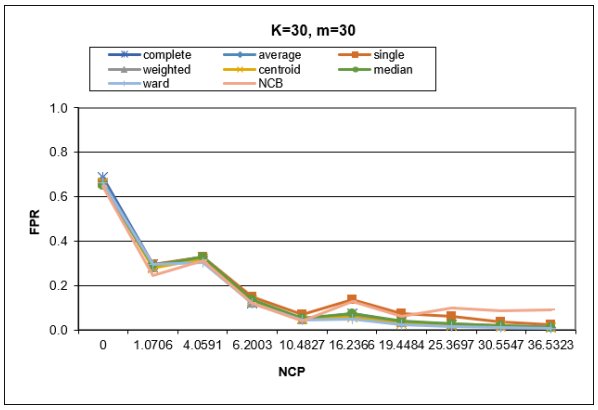

FIGURE14
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING FALSE POSITIVE RATE INDEX (K = 30, M = 30)

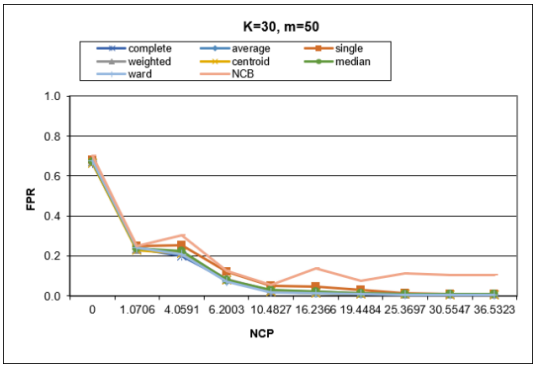

FIGURE15
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING FALSE POSITIVE RATE INDEX (K = 30, M = 50)

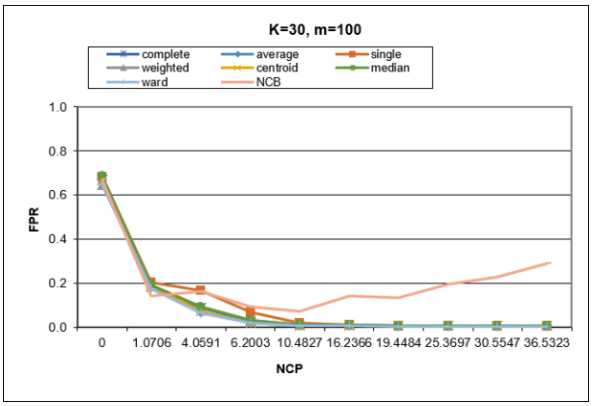

FIGURE16
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING FALSE POSITIVE RATE INDEX (K = 30, M = 100)

VI. Evaluation of the performance of control charts based on false negative rate index

Figures 17, 18, and 19 depict the simulation results, revealing a downward trend in the T_I^2 statistic for both clustering and non-clustering control charts. Overall, cluster-based control charts demonstrate superior performance compared to non-cluster-based approaches, as consistently revealed by the analysis. Also, the clustering method using the centroid, ward, and average linkage functions performs better than the complete linkage function. The weighted linkage function performance is almost identical to the complete linkage. Moreover, the complete linkage function performs better than the single and median linkage functions.

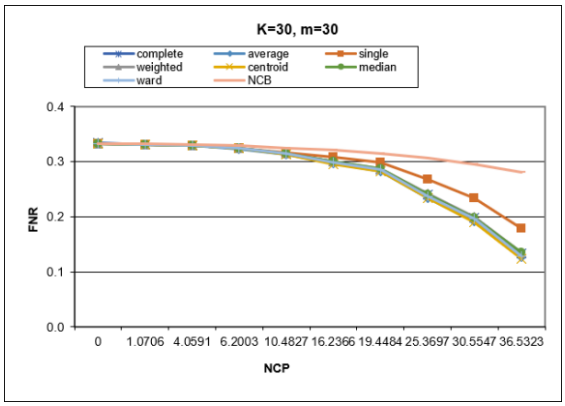

FIGURE17
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING THE FALSE NEGATIVE RATE INDEX. (K = 30, M = 30)

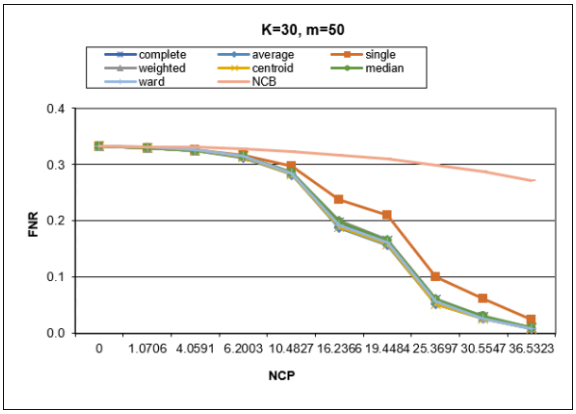

FIGURE18
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING THE FALSE NEGATIVE RATE INDEX. (K = 30, M = 50)

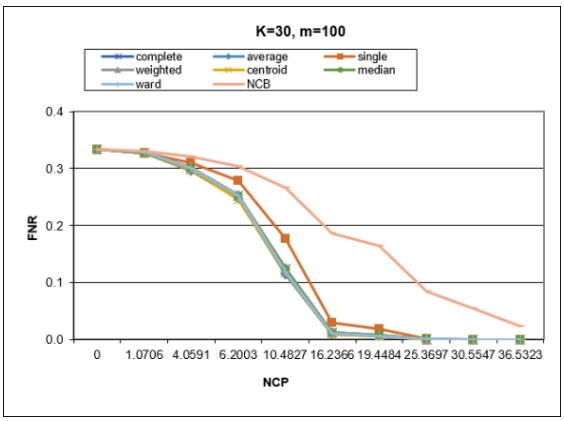

FIGURE19
PERFORMANCE OF CONTROL CHART UNDER A STEP SHIFT BY EMPLOYING THE FALSE NEGATIVE RATE INDEX. (K = 30, M = 100)

CONCLUSION

The presence of outliers within the historical dataset may compromise the precision of the control chart parameter estimations during Phase I analysis. To mitigate the influence of outliers on parameter estimation during Phase I control chart analysis, Saremian et al. (2021) introduced a groundbreaking clustering methodology. This approach strategically combines the maximum likelihood method with hierarchical clustering, aiming to enhance the robustness of parameter estimations for logistic profiles. The results of Saremian et al. (2021), as well as Chen et al. (2015), showed the satisfactory performance for the clustering method using a complete linkage function compared to the non-clustering method. This study examines the influence of seven distinct linkage functions on the performance of the proposed cluster-based methodology. The linkage functions under investigation include complete, average, single, weighted, centroid, median, and Ward's method. Within the context of Phase I analysis, this investigation assessed the efficacy of Cluster-Based (CB) control charts compared to their Non-Cluster-Based (NCB) T_I^2 counterparts for monitoring logistic profiles subjected to a step shift. This study extends the evaluation of control chart performance beyond the traditional probability of signal. To evaluate the charts' power in distinguishing between in-control and out-of-control profiles, five additional performance metrics are employed: fraction correctly classified, sensitivity, specificity, false negative rate, and false positive rate. The simulation results indicate that the T_I^2control chart, regardless of the specific linkage function employed, exhibits superior performance compared to the Non-Cluster-Based (NCB) control chart. These results further reveals that the cluster-based method achieves superior performance when employing the average, ward, or centroid linkage functions compared to the complete linkage function. Therefore, the use of these linkage functions will lead to a more accurate estimation of the parameters of control charts. The study also includes a real-life example using data on insect mortality rates. This demonstrates the practical applicability of the novel method. Also, a case study on insects' mortality rate data to shows how the new method can be applied in practice is presented. For future research, it is recommended that other clustering methods than hierarchical clustering be used and that their effect on the estimation of control chart parameters be investigated.

REFERENCES

[1] Kang, L., & Albin, S.L. (2000). On-line monitoring when the process yields a linear profile. *Journal of Quality Technology*, 32(4): 418-426. DOI: 10.1080/00224065.2000.11980027
[2] Soleimani, P., Noorossana, R. & Amiri, A. (2009). Simple linear profiles monitoring in the presence of within profile autocorrelation. *Computers and Industrial Engineering*, 57(3):1015-1021. DOI: 10.1016/j.cie.2009.04.005
[3] Chen, S., & Nembhard, H. B. (2010). A high-dimensional control chart for profile monitoring. *Quality and Reliability Engineering International*, 27(4), 451–464. DOI:10.1002/qre.1140
[4] Noorossana, R., Aminmadani, M., & Saghaei, A. (2016). Effect of phase I estimation error on the monitoring of simple linear profiles in Phase II. *IntJ Adv Manuf Technol*, 84:873-884. DOI:10.1007/s00170-015-7078-2
[5] Mahmood, T., Abbasi, S.A., Riaz, M., & Abbas, N.(2019). An efficient Phase I analysis of linear profiles with application in photo-voltaic system. *Arab J Sci Eng*. 2019;44(3):2699-2716. DOI:10.1007/s13369-018-3426-5
[6] Moheghi, H.R., Noorossana, R., Ahmadi, O. (2020) Phase I and Phase II analysis of linear profile monitoring using robust estimators. *Commun Stat Theory Methods*. 49:1-18. DOI:10.1080/03610926.2020.1758724
[7] Ghasemi, Z., Zeinal Hamadani, A., & Ahmadi Yazdi, A. (2023). New methods for phase II monitoring of multivariate simple linear profiles. *Communications in Statistics-Simulation and Computation*, 1-25. DOI:10.1080/03610918.2023.2249268
[8] Pakzad, A., Adibfar, S., Razavi, H., & Noorossana, R. (2023). Process capability analysis for simple linear profiles. Qual Quan. DOI: 10.1007/s11135-023-01726-4

[9] Adibfar, S., Noorossana, R., & Ahmadi, O. (2023). Process capability analysis for multivariate simple linear profiles in a multistage process. *Journal of Industrial and Systems Engineering, 14(4), 158-173.* DOI:10.18187/pjsor.v20i1.4410

[10] Williams, J.D., Woodall, W.H., Birch, J.B., & Sullivan, J.H. (2006). Distribution of Hotelling's T2 Statistic Based on the Successive Differences Estimator. *Journal of Quality Technology*. 38(3), 217–229. DOI:10.1080/00224065.2006.11918611

[11] Steiner, S., Jensen, W.A., Grimshaw, S.D., & Espen, B.(2016). Nonlinear profile monitoring for oven-temperature data. *J Qual Technol*. 48(1):84-97. DOI:10.1080/00224065.2016.11918153

[12] Pan, J.N., Li ,C.I., Lu, M.Z. (2019). Detecting the process changes for multivariate nonlinear profile data. *Qual Reliab Eng Int.*35:1890-1910. DOI:10.1002/qre.2482.

[13] Noorossana, R., Saghaei, A., Amiri, A.(2012). Statistical Analysis of Profile Monitoring. *Hoboken, NJ: John Wiley & Sons*.

[14] Maleki, M.R., Amiri, A., Castagliola, P. (2018) An overview on recent profile monitoring papers based on conceptual classification scheme. *Comput Ind Eng*. 126:705-728. DOI:10.1016/j.cie.2018.10.008

[15] Cheng, C. S., Chen, P. W., & Wu, Y. T. (2023). Phase I Analysis of Nonlinear Profiles Using Anomaly Detection Techniques. *Applied Sciences*, *13*(4), 2147. DOI:10.3390/app13042147

[16] Montgomery, D.C., Peck, E.A, and Vining, G,G.(2008). Introduction to Linear Regression Analysis, fourth edition, *Wiley, Hoboken, NJ* .

[17] Yeh, A.B., Huwang, L., & Li, Y.M. (2009). Profile monitoring for a binary response. *IIE Transactions*. 41(11):931-941. DOI: 10.1080/07408170902735400

[18] Shang, Y., Tsung, F., & Zou, C. (2011). Profile Monitoring with Binary Data and Random Predictors. *Journal of Quality Technology.* 43(3), 196–208. DOI:10.1080/00224065.2011.11917857

[19] Koosha, M., & Amiri, A. (2012). Generalized linear mixed model for monitoring autocorrelated logistic regression profiles. *International Journal of Advanced Manufacturing Technology*. 64(1-4):487-495.DOI: 10.1007/s00170-012-4018-2

[20] Paynabar, K., Jin, J., & Yeh, A.B. (2012). Phase I Risk-Adjusted Control Charts for Monitoring Surgical Performance by Considering Categorical Covariates. *Journal of Quality Technology*. 44(1), 39 53. DOI:10.1080/00224065.2012.11917880

[21] Amiri, A., Koosha, M., Azhdari, A. & Wang, G. (2014). Phase I monitoring of generalized linear model-based regression profiles. *Journal of Statistical Computation and Simulation*. 85(14), 2839-2859. DOI: 10.1080/00949655.2014.942864

[22] Shadman, A., Mahlooji, H., Yeh, A.B., & Zou, C. (2015). A change-point method for monitoring generalized linear profiles in Phase I. *Quality and Reliability Engineering International*. 31(8):1367-1381. DOI: 10.1002/qre.1671

[23] Shadman, A., Mahlooji, H., & Yeh, A.B. (2014). A Change Point Method for Phase II Monitoring of Generalized Linear Profiles. *Communications in Statistics - Simulation and Computation*. 46(1): 559–578. DOI:10.1080/03610918.2014.970698

[24] Noorossana, R., Niaki, S., Izadbakhsh, H. (2015). Statistical monitoring of nominal logistic profiles in phase II. *Commun Stat Theory Methods*. 44(13):2689-2704. DOI:10.1080/03610926.2013.788712

[25] Shang, Y., Wand, Z., He, Z., & He, S. (2017). Nonparametric change-point detection for profiles with binary data. *J Qual Technol*. 49(2):123-135. DOI:10.1080/00224065.2017.11917984

[26] Izadbakhsh, H., Noorossana, R., & Niaki, S. T. A. (2018). Monitoring multinomial logistic profiles in Phase I using log-linear models. *International Journal of Quality & Reliability Management.* 35(3), 678–689. DOI:10.1108/ijqrm-04-2017-0068

[27] Bandara, K., Abdel-Salam, A.S. G., & Birch, J. B. (2020). Model robust profile monitoring for the generalized linear mixed model for Phase I analysis. *Applied Stochastic Models in Business and Industry.* 36(6), 1037-1059. DOI:10.1002/asmb.2587

[28] Hakimi, A., Amiri, A., & Kamranrad, R. (2017). Robust approaches for monitoring logistic regression profiles under outliers. *International Journal of Quality & Reliability Management.* 34(4), 494–507. DOI:10.1108/ijqrm-04-2015-0053

[29] Hakimi, A., Amiri, A., & Kamranrad, R. (2018). Robust Method for Logistic Profiles Monitoring in Phase I. *Production and Operations Management*. 9(16). DOI: 10.22108/JPOM.2018.92335.0

[30] Moheghi, H.R., Noorossana, R., & Ahmadi, O. (2020). GLM profile monitoring using robust estimators. *Quality and Reliability Engineering International*. 1– 17. DOI:10.1002/qre.2755

[31] Cantoni, E., Ronchetti, E. (2001). Robust inference for generalized linear models. *J Am Statist Assoc*. 96(455):1022-1030. DOI: 10.1198/016214501753209004

[32] Saremian, D., Noorossana, R., Raissi, S., & Soleimani, P. (2021). Robust Cluster-Based method for monitoring generalized linear profiles in phase I. Journal of Industrial Engineering, International. 17(1), 88-97. DOI: 2021.1920761.1085

[33] Chen, Y., Birch, J.B., & Woodall, WH. (2015). Cluster-Based Profile Analysis in Phase I. *Journal of Quality Technology* 2015. 47(1):14–29. DOI:10.1080/00224065.2015.11918103

[34] Saremian, D., Noorossana, R., Raissi, S., Soleimani, P. (2022). Monitoring logistic profiles in phase I using robust cluster-based method. *Qual Reliab Eng Int*. 38: 1977– 1993. DOI:10.1002/qre.3054

[35] McCullagh, P., Nelder, J.A. (1989) Generalized Linear Models (2nd edn), *Chapman & Hall*, London, UK.

[36] Dobson, A.J., Barnett,A.G. (2008). An Introduction to Generalized Linear Models, Third Edition, *CRC Press,Taylor & Francis Group.*