mathematics

MDPI

Article

Robust Multivariate Shewhart Control Chart Based on the Stahel-Donoho Robust Estimator and Mahalanobis Distance for Multivariate Outlier Detection

Ishaq Adeyanju Raji [1], Nasir Abbas [2,*], Mu'azu Ramat Abujiya [3] and Muhammad Riaz [2]

[1] Dammam Community College, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia; Ishaq.raji@dcc.kfupm.edu.sa
[2] Department of Mathematics and Statistics, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia; riazm@kfupm.edu.sa
[3] Preparatory Year Mathematics Program, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia; abujiya@kfupm.edu.sa
* Correspondence: nasirabbas@kfupm.edu.sa

Abstract: While researchers and practitioners may seamlessly develop methods of detecting outliers in control charts under a univariate setup, detecting and screening outliers in multivariate control charts pose serious challenges. In this study, we propose a robust multivariate control chart based on the Stahel-Donoho robust estimator (SDRE), whilst the process parameters are estimated from phase-I. Through intensive Monte-Carlo simulation, the study presents how the estimation of parameters and presence of outliers affect the efficacy of the Hotelling T^2 chart, and then how the proposed outlier detector brings the chart back to normalcy by restoring its efficacy and sensitivity. Run-length properties are used as the performance measures. The run length properties establish the superiority of the proposed scheme over the default multivariate Shewhart control charting scheme. The applicability of the study includes but is not limited to manufacturing and health industries. The study concludes with a real-life application of the proposed chart on a dataset extracted from the manufacturing process of carbon fiber tubes.

Keywords: multivariate control charts; Mahalanobis distance; control chart; Hotelling T^2; Stahel-Donoho robust estimators; outlier detection

Citation: Raji, I.A.; Abbas, N.; Abujiya, M.R.; Riaz, M. Robust Multivariate Shewhart Control Chart Based on the Stahel-Donoho Robust Estimator and Mahalanobis Distance for Multivariate Outlier Detection. *Mathematics* **2021**, *9*, 2772. https://doi.org/10.3390/math9212772

Academic Editor: Ioannis S. Triantafyllou

Received: 8 October 2021
Accepted: 29 October 2021
Published: 1 November 2021

1. Introduction

Outliers are those observations at both extremes, which do not follow the majority of observations pattern in a dataset. Outlier detection is of concern in data analysis and scientific areas, of which statistical process control (SPC) is not an exemption [1]. This is because outliers have a major influence on any statistical analysis as they increase the error variance, reduce the power of statistical tests, and cause bias in estimation, hence leading to incorrect inferences and conclusions, and sometimes, ending with deadly decisions, take the health sector as an example. With little percentage and magnitude present in data (big or small), outliers will grossly distort the performance and analysis of the data. Therefore, the art of outlier detection is a prominent and important aspect of data analysis, even more so now that more and more data are being analyzed simultaneously, such as with multivariate control charting.

Control charts are the most widely used tool amongst the seven tools of SPC [2]. Their vast applicability in different fields and sectors give them an edge over other tools of SPC for process monitoring. Control charts, however, can have a univariate or multivariate setup, a memory or memory-less type, and/or monitoring location or dispersion in an ongoing process. Readers are referred to [2] for more information about control charts and

their types. Furthermore, control charts are of two stages: phase-I (the prospective stage) and phase-II (the retrospective stage). The process parameters are used to set the chart's control limits in phase-I. Moreover, if the process parameters are unknown, they are estimated with some preliminary samples, whereas the monitoring and correction of unnatural causes of variation occur in the retrospective stage. The choice and amount of preliminary sample employed in estimating the unknown parameters in phase-I vary among practitioners and as result affect the performance of the chart in the monitoring stage. These samples often contain some unusual observations and outliers, which exert a disproportionate pull on the parameter estimated, making the chart less efficient in detecting anomalies. The multivariate Shewhart chart that has been studied in this paper is a memory-less type for monitoring location parameters, while the process parameters are known and estimated from phase-I samples. Over the years, SPC researchers have investigated the effect of parameter estimation on control charts in both univariate and multivariate setups. To mention a few, reference [3] gave an up-to-date review on parameter estimation effects on control charts. Saleh et al. [4] evaluated the parameter estimation's effect on an exponentially weighted moving average (EWMA) control chart with its run length properties. A similar study was conducted by Jones [5].

Many research works in the literature have studied outlier detection in the univariate setup, some of which are applied to control charts in the univariate setup. References [6–8] have independently proposed outlier detection models in the univariate setup of control charts either for location or dispersion monitoring. They found that the control charts based on detection models require fewer phase-I samples to detect anomalies, as these charts are quicker and more sensitive to contamination. Guarnieri et al. [9] developed control charts for individual observation and exponentially weighted moving averages based on residues to detect outliers in autoregressive models. Bakar et al. [10] also conducted a comparative study for outlier detection techniques in control charts with application in data mining. As Vidmar and Blagus [11] applied different outlier detection approaches to healthcare quality monitoring. Zhang and Albin [12] employed a chi-square chart method for detecting outliers in complex profiles. Other research in this direction include, among others, [13,14]. While there are models for detecting multivariate outliers, few of them have been applied to SPC. Examples include the robust multivariate control chart for outlier detection by Fan et al. [15] and robust estimates, residuals, and outlier detection with multi-response data by Gnanadesikan and Kettenring [16]. The authors of [17] considered minimum volume ellipsoid (MVE) and/or weighted mean vector and mean square successive differences (WD) to decrease the impact of outliers on multivariate control charts. Hubert et al. [18] reviewed the minimum covariance determinant (MCD) methods and their extension as competent tools for outlier detection. Other researchers have approached the outlier detection problem with robust multivariate estimators. The pioneer of this idea was Stahel [19] where he studied the breakdown of covariance estimators; Maronna and Yohai [20] further extended the research of Stahel. Rousseeuw and Hubert [21] also studied the robust multivariate location and scatter estimators. Similar studies include but are not limited to [22–24].

In the aforementioned references, none of the studies that applied multivariate robust estimators to control charts have focused on detecting and screening outliers of the phase-I samples. Therefore, this paper focuses on detecting multivariate outliers in the multivariate Shewhart control chart. It employs a Stahel-Donoho robust estimator incorporated with the Mahalanobis distance for detecting and screening out the outlying observations in the preliminary samples, from which the process parameters are estimated. This paper reports the effect of parameter estimations on a multivariate Shewhart chart's control limits and performance. Reporting parameter estimations' effect is not the main goal of this study; however, it helps readers to better understand the positive impact of the outlier detection process.

The remainder of this article is organized as follows. Section 2 entails the methodology with an insight to the multivariate Shewhart control chart when the process parameters are known and estimated, the presence of outliers in the preliminary samples, and the proposed multivariate outlier detection process. Results and discussion appear in Section 3, while Section 4 gives an illustrative example with a real-life dataset extracted from the manufacturing process of carbon fiber tubes. Section 5 concludes the study with a summary of the findings and future recommendations.

2. Methodology

The aim of this study is to detect and screen outliers of the m preliminary samples employed for parameter estimation, especially when the samples are outlier prone. This section explains in detail the multivariate Shewhart control chart for location monitoring, both when the parameters are known and estimated from phase-I preliminary samples. Then it demonstrates the effect of practitioners' variability in the samples employed for estimation, and its effect on the chart's performance. In addition, this section presents how outliers in those samples distort the chart's efficacy and become less sensitive, then concludes the section with the proposed outlier detection-based multivariate Shewhart chart, and its application on a real-life data set extracted from the carbon fiber tubes manufacturing industry.

2.1. Multivariate Shewhart Control Chart

Let $\boldsymbol{X} = (X_1, X_2, X_3 \ldots, X_p)$, a vector of p-correlated quality characteristics, each of size n subgroups, drawn from a p-variate normal distribution be the characteristic of interest for monitoring in a multivariate process. The probability distribution function of $\boldsymbol{X}$ is given as follows:

$$f(\boldsymbol{X}) = \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{1/2}}\, e^{\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right)};\ -\infty < X_i < \infty, i = 1, 2, \ldots, p. \tag{1}$$

The resulting multivariate Shewhart chart statistic termed Hotelling T^2, for monitoring the location parameter of the random process $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ X, is given as follows:

$$T_i^2 = n(\overline{\boldsymbol{X}}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{X}}_i - \boldsymbol{\mu}). \tag{2}$$

where $\overline{\boldsymbol{X}}_i$ is the mean vector of the ith observation, n is the sample size, $\boldsymbol{\mu}' = (\mu_1, \mu_2, \ldots, \mu_p)$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

is the mean vector and variance-covariance matrix of the process. The chart signals an alarm when the T_i^2 statistic is plotted beyond the upper control limit (UCL) of the chart, i.e., $(T_i^2 > \text{UCL} = \chi^2_{\alpha,p})$. This is the case when the process parameters (μ, Σ) are known. However, when the parameters are unknown, they are estimated from m phase-I preliminary samples. The Hotelling T^2 statistics then become

$$T_i^2 = n(\overline{\boldsymbol{X}}_i - \widehat{\boldsymbol{\mu}})'\boldsymbol{S}^{-1}(\overline{\boldsymbol{X}}_i - \widehat{\boldsymbol{\mu}}), \tag{3}$$

where $\widehat{\boldsymbol{\mu}} = \sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{X}_{i,j}/mn$ and $\boldsymbol{S} = \sum_{i=1}^{m}\sum_{j=1}^{n}(\boldsymbol{X}_{i,j} - \overline{\boldsymbol{X}}_i)(\boldsymbol{X}_{i,j} - \overline{\boldsymbol{X}}_i)^T/m(n-1)$ are the estimates of the in-control mean vector and variance-covariance matrix emerging from the phase-I samples. It is important to note that the amount of m phase-I sample and the choice of estimators employed for estimating the parameters vary amongst practitioners, hence the variability in the efficacy and performance of their charts. Subsequently, the corresponding UCL of the T_i^2 statistic in (3), for the monitoring stage, phase-II, is given as follows:

$$\text{UCL} = \frac{p(m+1)(n-1)}{mn-m-p+1} F_{\alpha,p,mn-m-p+1} \quad (4)$$

Again, if $T^2 > \text{UCL}$, the chart sends a signal, so the practitioner tends to the unnatural cause of variation. The *i*th observation on which a signal was sent is the run length. The run length is simply the number of observations plotted within the limit before recording the first out-of-control (OoC). With many iterations, run length becomes a variable whose properties will be used for evaluating the chart.

All this explains the traditional method for constructing the multivariate Shewhart chart for location monitoring. The next section establishes parameter estimation effects on the multivariate Shewhart chart. Section 2.3, on the other hand, reveals how the outliers emanating from the phase-I sample negatively affect the chart's performance, while Section 2.4 highlights the need for incorporating multivariate robust estimators for outlier detection.

2.2. Effect of Practitioners' Variabilities on the Multivariate Shewhart Chart

In this section, the study reveals how the practitioners' variability in the choice of m samples affects the multivariate Shewhart chart's performance. Through intensive Monte-Carlo simulation, we demonstrate how different m phase-I samples for estimating the unknown parameters play a vital role in the performance of the multivariate Shewhart chart as compared to the known parameter case. This study considers m of 25, 100, and 500 to represent small, medium, and large samples, respectively. An algorithm was developed in R language to simulate the multivariate Shewhart chart defined in (2) for the known parameter case and in (3) for the unknown case. For the known case, it was assumed that the mean vector was zero, variances were unity, and the covariance was 50% (i.e., $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.5$). With $p = 2, 3, \alpha = 0.0027$, the in-control (IC) average run length (ARL_0) corresponded to 370. While for the unknown cases, the process parameters were estimated from $m = 25, 100, 500$ samples with sample mean vector $\hat{\boldsymbol{\mu}}$ and covariance matrix $\boldsymbol{S}$**.** The algorithm also considered the OoC situations, when the mean vector increased over a range of shift $\delta \in [0,5]$. The first effect of estimation began with the UCL; the UCL varied as the m sample varied, to yield the nominal ARL_0 of 370 as in the known case. The simulation results are presented in Tables 1 and 2. The detailed discussion of these results is in Section 3.

Table 1. ARL and SDRL values of the multivariate Shewhart control chart with $p = 2$.

Unknown Case: Parameters Estimated							**Known Case**	
	$\boldsymbol{m = 25}$		$\boldsymbol{m = 100}$		$\boldsymbol{m = 500}$			
$\boldsymbol{\delta}$	**ARL**	**SDRL**	**ARL**	**SDRL**	**ARL**	**SDRL**	**ARL**	**SDRL**
0.00	369.38	529.24	369.101	392.986	370.2419	392.7475	370.50	370.19
0.50	230.28	338.60	207.208	228.794	207.52	210.96	201.90	202.24
1.00	85.73	128.35	72.408	79.334	68.79	69.71	67.28	66.57
1.50	29.91	42.31	25.617	27.083	23.58	23.57	23.28	22.94
2.00	11.82	15.50	9.918	10.088	9.63	9.37	9.45	8.92
2.50	5.20	5.84	4.733	4.482	4.66	4.12	4.59	4.12
3.00	2.87	2.72	2.653	2.124	2.57	2.01	2.57	1.97
3.50	1.86	1.42	1.756	1.173	1.71	1.13	1.70	1.09
4.00	1.39	0.80	1.351	0.693	1.33	0.66	1.32	0.65
4.50	1.16	0.45	1.143	0.412	1.14	0.40	1.13	0.38
5.00	1.06	0.27	1.053	0.239	1.05	0.23	1.05	0.22
	UCL = 12.27		UCL = 11.96		UCL = 11.87		UCL = 11.83	

Note: p is the number of charactrsitics, δ is the shift, m is the phase-I sample, UCL is the upper control limit, ARL is the average run length, and SDRL is the standard deviation run length.

Table 2. ARL and SDRL values of the multivariate Shewhart control chart with $p = 3$.

Unknown Case: Parameters Estimated							Known Case	
	$m = 25$		$m = 100$		$m = 500$			
δ	ARL	SDRL	ARL	SDRL	ARL	SDRL	ARL	SDRL
0.00	369.71	511.73	370.28	404.67	370.85	375.25	370.35	368.60
0.50	252.41	354.01	237.52	264.45	233.12	236.52	229.50	229.21
1.00	108.87	158.98	89.98	96.39	87.09	89.19	86.11	85.38
1.50	40.93	57.93	33.15	36.22	31.57	32.24	30.98	30.38
2.00	16.28	21.76	13.38	13.94	12.52	12.29	12.37	11.90
2.50	7.20	8.63	6.12	5.94	5.80	5.29	5.71	5.21
3.00	3.71	3.81	3.23	2.76	3.10	2.58	3.10	2.55
3.50	2.25	1.94	2.06	1.53	2.02	1.42	1.98	1.39
4.00	1.57	1.05	1.49	0.86	1.46	0.83	1.45	0.81
4.50	1.26	0.60	1.21	0.50	1.21	0.49	1.19	0.48
5.00	1.11	0.36	1.08	0.30	1.08	0.29	1.08	0.29
	UCL = 15.16		UCL = 14.43		UCL = 14.22		UCL = 14.16	

Note: p is the number of charactrsitics, δ is the shift, m is the phase-I sample, UCL is the upper control limit, ARL is the average run length, and SDRL is the standard deviation run length.

2.3. Effect of Outliers on the Multivariate Shewhart Control Chart with Estimated Parameters

Having noticed the estimation effect on the multivariate Shewhart chart's performance in the previous section, we demonstrate how outliers in the m phase-I samples worsen the chart's performance in the monitoring stage. To achieve this aim, we generated m phase-I samples from a mixed distribution, a$(1-\theta)100\%$ from the normal distribution and the remaining $\theta 100\%$ from a chi-square distribution with v degrees of freedom as follows:

$$\boldsymbol{X} \sim (1-\theta)N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \theta[N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \omega\boldsymbol{\chi}^2_{(v)}] \quad (5)$$

where $\theta > 0$ represents the percentage of outliers present in the data, $\omega \geq 1$ is the magnitude of the outliers, and $\boldsymbol{\chi}^2_{(v)}$ represents the outlier added to the normal distribution. The study estimated the parameters $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{S}$ from the m sample, and then computed the Hotelling T^2 statistic as in (3). The same algorithm, process parameters, and control limits employed in Section 2.2 were used to compute the IC run length properties alone to observe the outliers' effect. The results are presented in Tables 3 and 4 for magnitudes $\omega = 1{,}2$, respectively. With just 10% of outliers ($\theta = 0.10$), the ARL_0 increased by more than 600% of its expected value when $\omega = 1$ and close to 3000% when $\omega = 2$.

The findings from the results in this section and the previous section suggest the following options:

- The m phase-I sample should be sufficiently increased until results similar to those of the known case are achieved.
- The process should prevent the occurrence of unnatural variations and outliers with smaller m phase-I samples

These options are practically impossible in real life scenarios, because increasing samples is typically uneconomical. More so, a process cannot be freed from variations with a natural or assignable cause. Hence, there is the need to incorporate robust multivariate estimators for better estimation and screening of the outliers.

Table 3. ARL_0 and $SDRL_0$ values of the multivariate Shewhart control chart with outliers ($\omega = 1$).

$\omega = 1$	$m = 25$		$m = 100$		$m = 500$	
θ	ARL	SDRL	ARL	SDRL	ARL	SDRL
0.00	370.22	506.49	369.05	406.10	370.39	376.06
0.01	479.59	824.67	486.10	549.93	494.51	503.12
0.02	630.37	1301.69	648.14	775.72	653.54	677.66
0.03	780.66	1614.26	816.27	1045.23	832.28	866.55
0.04	959.31	2445.42	1041.62	1316.05	1041.87	1111.20
0.05	1167.80	2772.21	1267.57	1754.63	1312.05	1400.20
0.06	1492.42	3676.36	1526.53	1987.07	1591.60	1692.19
0.07	1844.59	5011.12	1778.28	2371.01	1841.32	1978.55
0.08	2053.30	4823.31	2098.60	2889.33	2169.25	2256.39
0.09	2349.65	5314.44	2523.52	3514.10	2476.13	2649.50
0.10	2766.19	6210.28	2736.63	3800.61	2749.25	2914.35
	UCL = 15.16		UCL = 14.43		UCL = 14.22	

Note: ω and θ are the magnitude and percentage of outliers, respectively; m is the phase-I sample; ARL is the average run length; and SDRL is the standard deviation run length.

Table 4. ARL_0 and $SDRL_0$ values of the multivariate Shewhart control chart with outliers ($\omega = 2$).

$\omega = 2$	$m = 25$		$m = 100$		$m = 500$	
θ	ARL	SDRL	ARL	SDRL	ARL	SDRL
0.00	373.15	528.37	370.30	407.83	376.98	382.58
0.01	810.48	1994.63	945.97	1420.80	1084.37	1246.99
0.02	1600.06	4562.27	2142.58	4030.07	2449.51	2999.48
0.03	2772.41	7427.22	4021.04	7220.47	4876.12	6078.31
0.04	3957.63	9442.19	6533.33	10,703.47	7752.17	9110.05
0.05	5523.83	11,857.39	9510.68	14,211.41	11,506.65	13,324.94
0.06	6896.48	13,435.25	11,889.15	16,306.08	15,134.80	16,454.69
0.07	8376.87	15,180.61	14,364.71	18,529.89	18,707.26	19,675.92
0.08	9705.77	16,708.27	16,011.60	19,649.58	21,367.60	21,353.52
0.09	10,826.71	18,000.62	17,978.87	21,409.00	23,325.17	22,914.61
0.10	11,452.42	18,314.76	19,179.15	22,606.24	25,097.16	24,509.93
	UCL = 15.16		UCL = 14.43		UCL = 14.22	

Note: ω and θ are the magnitude and percentage of outliers, respectively; m is the phase-I sample; ARL is the average run length; and SDRL is the standard deviation run length.

2.4. Proposed Multivariate Shewhart Chart Based on Stahel-Donoho Robust Estimators (SDRE)

From the results in Tables 3 and 4, it is apparent that increasing the m samples cannot suppress the negative impact of the outliers on the chart. Hence, there is a need to employ robust location and dispersion estimators as substitutes to the default $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{S}$ that are not sensitive to outliers. Therefore, this study proposes a multivariate Shewhart chart based on the Stahel-Donoho robust estimator. Like any robust estimator, the SDRE estimators were able to retain their efficiency in the presence of outliers. This feature makes them able to detect the presence of outliers no matter how small or large the m samples are. Readers are referred to [25–27] for more information about the merits of robust estimators.

Stahel [19] and Donoho [22] were the first to develop a robust equivariant estimator of multivariate location and dispersion with a considerable high breakdown point of any p-variate multivariate data. However, it became well known with the analysis of Maronna and Yohai [20]. Maronna and Yohai [20] assumed $\boldsymbol{X} = \{x_{1,} x_{2,} \ldots, x_n\}$ to be a set of n data points in $\Re^p$, and defined the "outlyingness" r for any $y \epsilon \Re^p$ as $r(\boldsymbol{y}, \boldsymbol{X}) = \sup_{\boldsymbol{a}} r_1(\boldsymbol{y}, \boldsymbol{a}, \boldsymbol{X})$,

where $r_1(\boldsymbol{y}, \boldsymbol{a}, \boldsymbol{X}) = |a'y - \mu(\boldsymbol{a}'\boldsymbol{X})|/\sigma(\boldsymbol{a}'\boldsymbol{X})$ and $\mu()$ and $\sigma()$ are the robust univariate location and dispersion statistics. The Stahel-Donoho robust estimators (SDRE), denoted as $(\boldsymbol{t}, \boldsymbol{V})$, are defined as weighted mean and weighted covariance matrix, each with weights of the form $w(r)$, where w_i is the weight function of each observation and inverse proportional to the "outlyingness" of the observation, r, obtained by considering all univariate projections of the data. Mathematically, SDRE is written as follows:

$$\boldsymbol{t}(\boldsymbol{X}) = \frac{\sum_{i=1}^{n} w_i \boldsymbol{X_i}}{\sum_{i=1}^{n} w_i} \; and \; \boldsymbol{V}(\boldsymbol{X}) = \frac{\sum_{i=1}^{n} w_i (\boldsymbol{X_i} - \boldsymbol{t})(\boldsymbol{X_i} - \boldsymbol{t})^{\boldsymbol{T}}}{\sum_{i=1}^{n} w_i} \tag{6}$$

where $w_i = w(r(x_i, \boldsymbol{X}))$**.** The SDRE is then used to estimate the process parameters from the m phase-I samples instead of $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{S}$**.** Furthermore, $(\boldsymbol{t}, \boldsymbol{V})$ estimators are employed in the Mahalanobis distance to screen out the potential outliers present in the m samples as in (7).

$$D(\boldsymbol{X}, \boldsymbol{t}) = n * \sqrt{(\boldsymbol{X} - \boldsymbol{t})^T \boldsymbol{V}^{-1} (\boldsymbol{X} - \boldsymbol{t})} \tag{7}$$

2.5. The Algorithm

This section explains in detail the algorithm and performance evaluation adopted in this study. The major performance measure of a control chart is the run length properties: average run length (ARL) and the standard deviation of the run length (SDRL). Through the Monte-Carlo simulation approach, the run length properties of both the IC (ARL_0 and $SDRL_0$) and OoC (ARL_1 and $SDRL_1$) of the scheme were computed. The following is the algorithm developed in R language to achieve this aim:

1. Generate 10^6 random variables of p-variate quality characteristics, each of sample size $n = 5$ from a multivariate normal distribution to be monitored in the phase-II stage.
2. (a) **Known case**: Define the mean vectors and covariance matrix, then proceed to step 3.
   (b) **Unknown case**: Generate some m phase-I samples from the same distributions to compute the default mean vector and covariance matrix estimators ($\hat{\boldsymbol{\mu}}$ and $\boldsymbol{S}$), then proceed to step 3 (see Section 2.2).
   (c) **Unknown case with outliers**: Generate some m phase-I samples from a mixed distribution as defined in (5), then compute the default mean vector and covariance matrix estimators ($\hat{\boldsymbol{\mu}}$ and $\boldsymbol{S}$) and then proceed to step 3 (see Section 2.3).
   (d) **Unknown case with outliers screened:** Generate some m phase-I samples from a mixed distribution as defined in (5), compute the SDRE $(\boldsymbol{t}, \boldsymbol{V})$ in (6), employ the SDRE to screen the outliers as explained in (7), and then compute $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{S}$ of the remaining dataset after screening. Then, proceed to step 3 (see Section 2.4).
3. Calculate the T_i^2 statistic in (2) for the known parameter case and (3) for the unknown cases, as the case may be.
4. Plot the T_i^2 statistic against the control limit, UCL, until the first ith observation plots beyond UCL. For known cases, $UCL = \chi^2_{\alpha,p}$, while for the unknown cases, use the UCL defined in (4).
5. Record the ith observation where the signal occurred as the run length.
6. Repeat the steps from 1–5 for 10^5 iterations. Record the run length for each iteration. Then, calculate the average and standard deviation of the run length as ARL_0 and $SDRL_0$, respectively.

The algorithm is summarized with a flowchart presented in Figure 1 for easy readability.

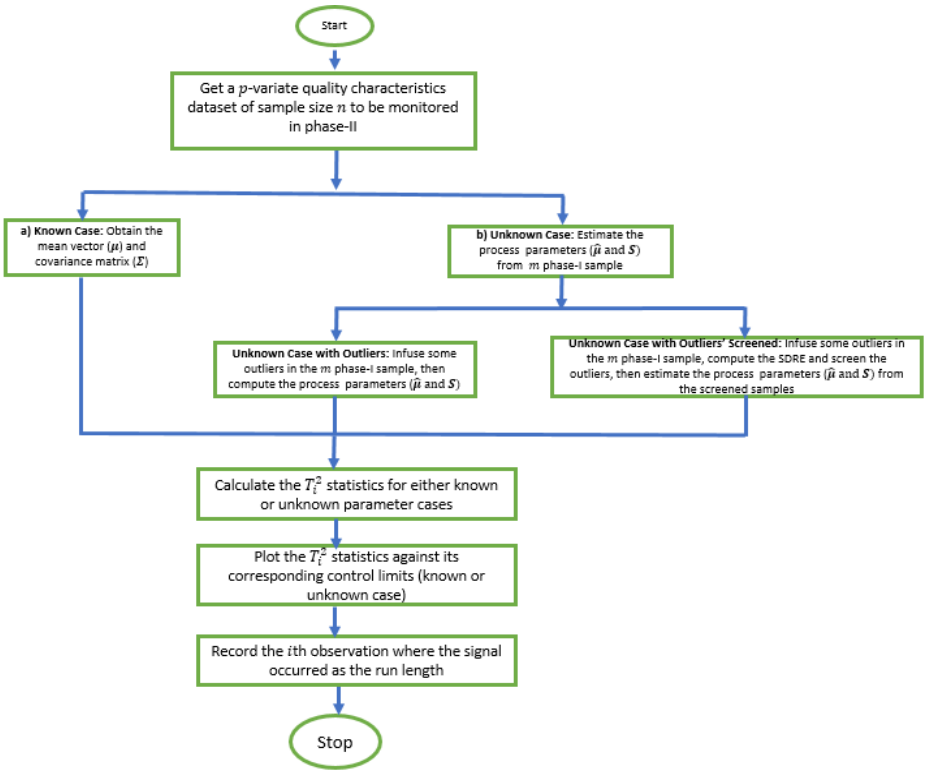

Figure 1. The flowchart of the methodology.

3. Results and Discussions

This section presents and discusses the results and findings of the methodologies explained in Section 2, in three categories: (a) the effect of practitioners' variabilities on the chart, (b) the effect of outliers on the chart's performance, and (c) the improvement of the proposed SDRE-based multivariate chart. The performance measure of this study was the run-length properties. The IC ARL_0 and $SDRL_0$ were expected to be sufficiently large as the nominal ARL_0 = 370, while the OoC ARL_1 and $SDRL_1$ were expected to be significantly small, implying the chart's ability to quickly detect anomalies in the process.

3.1. Practitioners' Variability Effect on Multivariate Shewhart Chart

Following the algorithm in Section 2.5, Tables 1 and 2 contain the ARLs of the multivariate Shewhart chart for the known parameter case and the estimated parameters with $p = 2$ and 3, respectively. The parameter estimation effect on the chart's performance is evident from the ARL and SDRL values. The different m phase-I samples represent the variabilities in practitioners' choice, ranging from small to medium to large. The larger the m samples, the better the chart's performance as compared to the known case. When $\delta = 0$, both the ARL_0 and $SDRL_0$ were expected to cluster around the nominal value 370. The ARL values did so, but the SDRLs of the estimated parameter scenarios did not. They dispersed from 370, and the disparity became even wider as the m samples got smaller. Another major effect was how the charts with the estimated parameters were less sensitive to shift as their ARL_1 and $SDRL_1$ imply. This effect also worsened as the m samples reduced (see Tables 1 and 2).

3.2. Effect of Outliers on the Multivariate Shewhart Chart's Performance

Tables 3 and 4 depict the in-control ARL_0 and $SDRL_0$ of the multivariate Shewhart chart from the mixed distribution in (5), with $p = 3$, some percentages of outliers $\theta = [0\%, 10\%]$, and the magnitude $\omega = 1{,}2$. Here, all values should be approximate to the nominal ARL of 370, since the environment was IC. When $\theta = 0\%$, it implies the absence

of outliers in the process, so the ARL values for all different m samples clustered around 370, while their SDRL values did not. It can be easily observed from the two tables that the outliers' effect on the chart worsened as the percentage and magnitude of outliers increased. Also, the effect on the ARL values was more obvious as the m sample increased, and vice-versa for the SDRL values. In general, there was more than a 600% increment in the ARL and SDRL values when $\omega = 1$ and a more than 3000% increment when $\omega = 2$. All of these were due to less than 10% outliers in the data.

3.3. Improvement of the Proposed SDRE-Based Multivariate Shewhart Chart

Here, we present and discuss the results of the proposed multivariate Shewhart chart based on SDRE and Mahalanobis distance for detecting and screening out the multivariate outliers, as described in Section 2.4. Tables 5 and 6 contain the IC ARL and SDRL as a remedy to the results in Tables 3 and 4, respectively. These results were obtained by applying the algorithm given in Section 2.5 (with part (d) of step 2). The improvement in the multivariate Shewhart chart's performance is easily noticeable. When magnitude $\omega = 1$, there was a more than a 25% decrement in comparison with when the outliers were not screened, while a decrement of more than 70% was achieved when $\omega = 2$ for the ARL values; the recoveries in the SDRL were even better. The SDRE-based multivariate Shewhart could not restore the chart's performance clustering around the nominal ARL=370; however, the recorded improvements are remarkable.

Table 5. ARL_0 and $SDRL_0$ values of the proposed SDRE multivariate Shewhart control chart ($\omega = 1$).

$\omega = 1$	$m = 25$		$m = 100$		$m = 500$	
θ	ARL	SDRL	ARL	SDRL	ARL	SDRL
0.00	361.92	495.73	375.75	406.13	369.15	372.24
0.01	449.23	701.48	457.44	506.28	458.50	464.16
0.02	517.30	786.42	566.53	646.91	559.66	569.57
0.03	652.16	1304.65	648.51	751.88	689.62	714.77
0.04	769.83	1581.16	803.33	948.51	802.49	821.11
0.05	899.75	1853.83	943.29	1110.27	950.73	991.70
0.06	1079.96	2177.20	1126.66	1421.11	1124.37	1179.38
0.07	1260.20	2722.94	1285.69	1573.09	1325.64	1377.60
0.08	1375.70	2832.76	1463.11	1810.17	1532.29	1626.18
0.09	1638.25	3739.77	1686.52	2192.34	1690.58	1752.70
0.10	1858.54	4088.50	1856.54	2425.25	1889.24	1968.77
	UCL = 15.16		UCL = 14.43		UCL = 14.22	

Note: ω and θ are the magnitude and percentage of outliers, respectively; m is the phase-I sample; ARL is the average run length; and SDRL is the standard deviation run length.

Table 6. ARL_0 and $SDRL_0$ values of the proposed SDRE multivariate Shewhart control chart ($\omega = 2$).

$\omega = 2$	$m = 25$		$m = 100$		$m = 500$	
θ	ARL	SDRL	ARL	SDRL	ARL	SDRL
0.00	361.78	499.51	370.12	400.86	375.70	384.66
0.01	549.31	1196.01	568.50	655.42	565.91	573.83
0.02	751.84	1868.04	815.21	1003.31	840.93	887.36
0.03	1066.50	2371.92	1147.92	1560.90	1204.67	1248.53
0.04	1461.44	3778.42	1584.77	2122.19	1652.91	1775.20
0.05	1933.98	4755.13	2115.68	3075.28	2239.99	2431.66
0.06	2457.79	6024.81	2795.44	3957.07	2837.89	2999.14
0.07	2997.34	7025.49	3539.72	5263.74	3533.50	3838.51

0.08	3645.17	8120.49	4207.78	6211.58	4222.54	4591.99
0.09	4499.18	9722.21	4962.75	7126.19	4958.27	5442.69
0.10	5206.90	10,638.48	5948.82	8550.87	5682.24	6336.62
	UCL = 15.16		UCL = 14.43		UCL = 14.22	

Note: ω and θ are the magnitude and percentage of outliers, respectively; m is the phase-I sample; ARL is the average run length; and SDRL is the standard deviation run length.

Furthermore, the rate of improvements appreciated as the percentage and magnitude of outliers increased. Figures 2–5 depict the results in Tables 3 and 4 (outliers without screening) side-by-side with Tables 5 and 6 (SDRE outliers screening) to closely observe the improvements. Tables 3 and 4 depict the IC ARL_0 and $SDRL_0$ of the multivariate Shewhart chart from the mixed distribution in (5), with some percentages of outliers, $\theta = [0\%, 10\%]$ and the magnitude $\omega = 1, 2$. Here, all values should approximate to the nominal ARL of 370, since the environment was IC. When $\theta = 0\%$, it implies the absence of outliers in the process, so the ARL values for all the different m samples clustered around 370, although their SDRL values did not. It can be easily observed from the two tables that the outliers' effect on the chart worsened as the percentage and magnitude of the outliers increased. Also, the effect on the ARL values was more obvious as the m sample increased, and vice-versa for the SDRL values. In general, there was more than a 600% increment in the ARL and SDRL values when $\omega = 1$ and a more than 3000% increment when $\omega = 2$. All of these were due to less than 10% outliers in the data.

The standard errors of the run length properties results reported in Tables 1–6 were between 0.066% and 0.506%. These values validate the precision of the ARL and SDRL values. In addition, in Tables 1 and 2, the results of the known cases are the best and the ideal results. However, the unknown case results improved and converged to those of the known cases as the m phase-I sample increased. For the results of the outlier cases in Tables 3–6, the outliers' effect was more pronounced as the percentage, θ, and magnitude, ω, of outliers increased. These points further justify and validate the precision and consistency reported results.

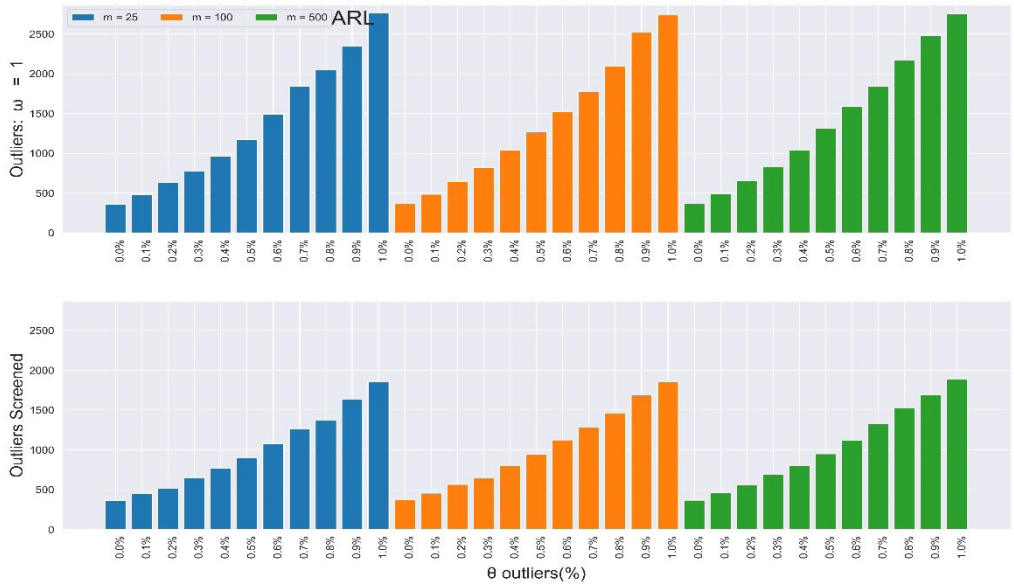

Figure 2. In-control ARL values of the multivariate Shewhart chart from mixed distribution with and without SDRE multivariate outliers screening ($\omega = 1$).

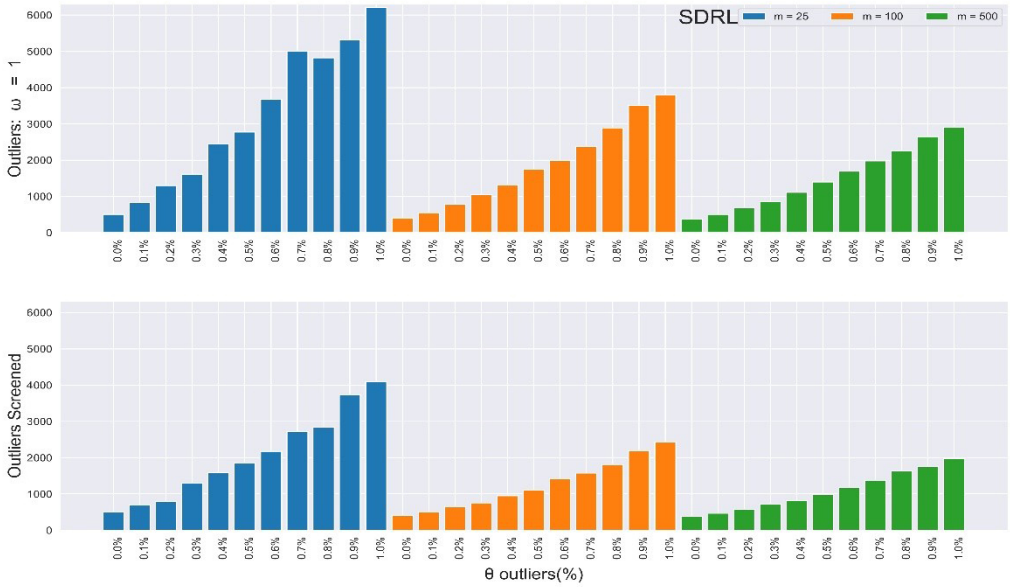

Figure 3. In-control SDRL values of the multivariate Shewhart chart from a mixed distribution with and without SDRE multivariate outliers screening ($\omega = 1$).

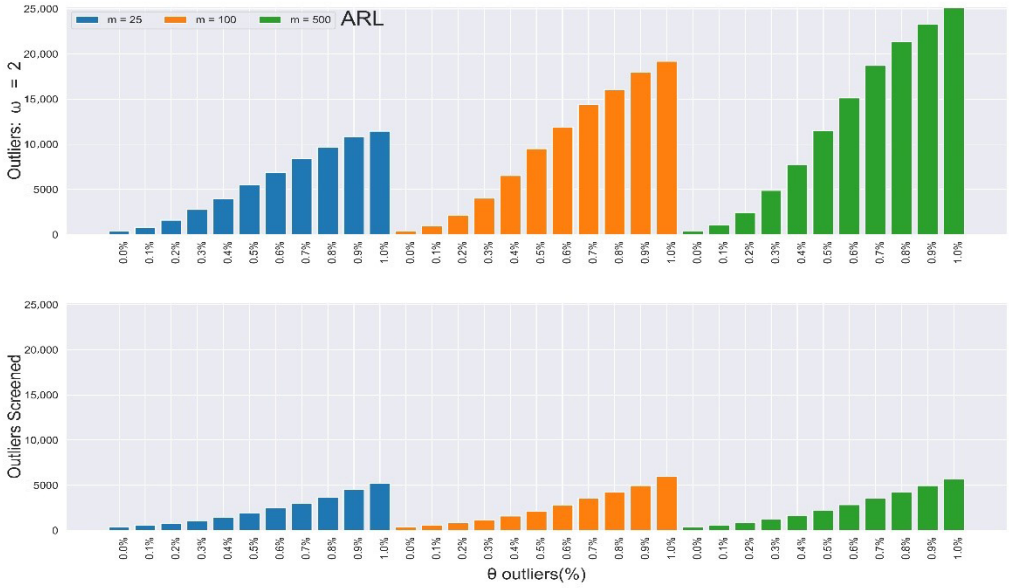

Figure 4. In-control ARL values of the multivariate Shewhart chart from a mixed distribution with and without SDRE multivariate outliers screening ($\omega = 2$).

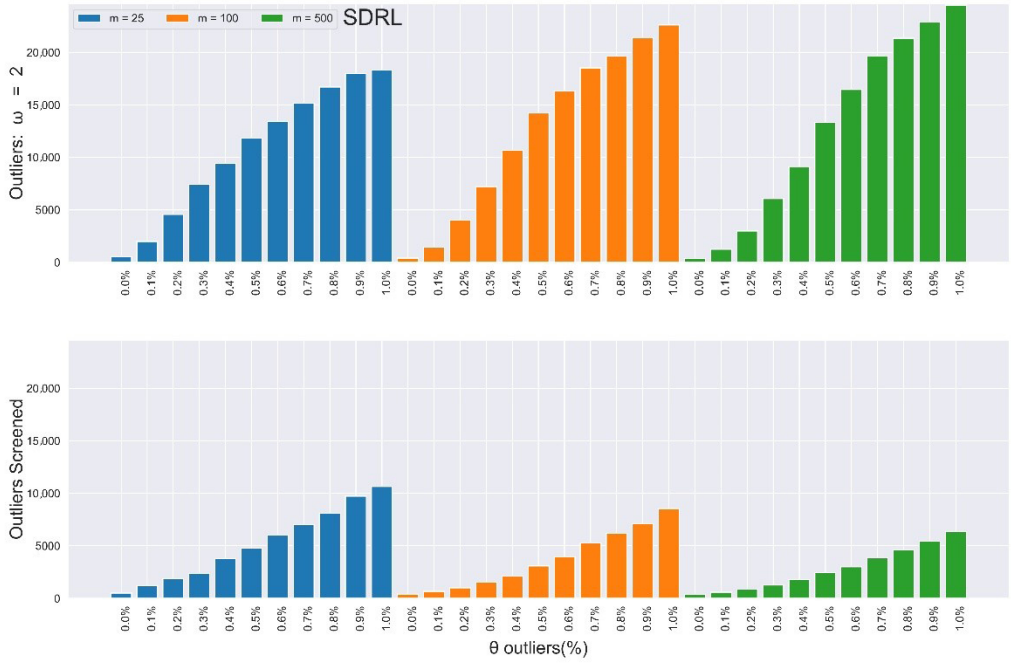

Figure 5. In-control SDRL values of the multivariate Shewhart chart from a mixed distribution with and without SDRE multivariate outliers screening ($\omega = 2$).

4. Illustrative Example with Real-Life Dataset

In the manufacturing industry, carbon fiber tubes are a crucial and widely used material in numerous applications. They are preferred over many traditional materials such as aluminum, titanium, and steel, because of their unique features: resistance to fatigue, high strength and fitness to weight, resistance to corrosion, dimensional stability, and many more. This has resulted in carbon fiber gaining vast application in the manufacturing industry. The manufacturing process of carbon fibers is partly chemical and mechanical. They are mostly made of carbon atoms which bound together in microscopic crystals. The manufacturing process goes through spinning, stabilizing, carbonizing, surface treating, and sizing. The tubes are thin strands of material which are long in diameter. The minute size of carbon fibers requires close monitoring of the manufacturing process. In this study, we monitored three quality characteristics in the manufacturing process of a specific carbon fiber tubing. The characteristics are the inner diameter, thickness, and length of the tubes in inches.

The data were of two stages: in phase-I, each quality characteristic consisted of $m = 25$ sample points each with a size of $n = 5$. Phase-II consisted of 20 observations each of size $n = 5$ for every quality characteristic. Without any loss of generality, and for conformity with the aim of the study, the illustrative example was categorized into three cases:

- *Case 1*-Parameter estimation: Here, we employed the phase-I data to compute the default mean vector and covariance matrix ($\hat{\boldsymbol{\mu}}$ and $\boldsymbol{S}$), assuming the process parameters were unknown, and then used the estimates to compute the plotting statistics T_i^2 for monitoring the phase-II data as explained in (3) and plotted it against the UCL.
- *Case 2*-Parameter estimation with outliers: We infused $\theta = 7\%$ of outliers with a magnitude $\omega = 3$ and degrees of freedom $v = 5$ in the phase-I data, to simulate the mixed distribution described in (5), obtained the default parameter estimates ($\hat{\boldsymbol{\mu}}$ and $\boldsymbol{S}$), and then used the estimates to compute the plotting statistics T_i^2 for monitoring the phase-II dataset and plotted it against the UCL
- *Case 3*-Parameter estimation with outliers and screening: The third case was similar to the second case, but we used the SDRE ($\boldsymbol{t}, \boldsymbol{V}$) as in (6) to estimate the process parameters from 25 phase-I samples, and to employ the SDRE in the Mahalanobis distance to detect

and screen out the outliers. Then, we computed the default parameter estimates ($\hat{\boldsymbol{\mu}}$ and $\boldsymbol{S}$) from the remaining screened data, and then computed the plotting T_i^2 for monitoring the phase-II dataset and plotted it against the same UCL.

The summaries of the estimations of the parameters for these three cases are given in Table 7, while their plotting statistics T_i^2 and corresponding decisions are given in Table 8. In case 1, all observations are IC as they are all below the UCL = 15.16, except the fourth observation 19.4183, which plots beyond the UCL. This case represents when the process parameters are estimated from some preliminary samples without outliers. For case 2, all the plotting T_i^2s were below the control limit despite the presence of outliers. The fourth observation that was plotted beyond the UCL in case 1 was masked due to the outliers' effect. Case 2 reveals the effect of outliers in the preliminary samples. It also shows the inferiority of using the default mean vector and covariance matrix for estimating parameters, especially when the samples are prone to outliers. In case 3, the OoC fourth observation in case 1 is was detected OoC in this case. With the same magnitude and percentage of outliers as in case 2, case 3 was as efficient as case 1 when there were no outliers. This substantiates the improvement of the proposed SDRE and Mahalanobis distance's procedures of estimating parameters and detecting outliers as claimed by the simulation results. Figure 6 depicts a visual representation of Table 8.

Table 7. $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{S}$ estimates from the phase-I sample for the three cases under study.

	Case 1			Case 2			Case 3		
$\hat{\boldsymbol{\mu}}$	0.9927	1.0357	50.0120	1.0000	1.0412	50.0844	0.9946	1.0406	50.0172
	0.0022	0.0026	0.0040	0.0034	0.0029	0.0045	0.0027	0.0040	0.0053
$\boldsymbol{S}$	0.0026	0.0128	0.0038	0.0029	0.0140	0.0023	0.0040	0.0165	0.0079
	0.0040	0.0038	0.0495	0.0045	0.0023	0.2391	0.0053	0.0079	0.0507

Table 8. T_i^2 values and decisions of the three cases with $\theta = 0.07$, $\omega = 3$, and $v = 5$.

i	Case 1		Case 2		Case 3	
	T_i^2	Decision	T_i^2	Decision	T_i^2	Decision
1	4.6350	IC	3.1616	IC	4.4034	IC
2	2.7626	IC	2.8659	IC	2.6736	OoC
3	6.5246	IC	3.2198	IC	6.0763	IC
4	19.4183	OoC	12.2758	IC	15.9676	OoC
5	2.8439	IC	0.8071	IC	2.6362	IC
6	2.7068	IC	0.6549	IC	2.4079	IC
7	4.8002	IC	0.9782	IC	4.1318	IC
8	0.8486	IC	0.5265	IC	0.8318	IC
9	1.0873	IC	1.1708	IC	1.1421	IC
10	1.1025	IC	1.5938	IC	1.0298	IC
11	0.3968	IC	0.2688	IC	0.2967	IC
12	1.9768	IC	0.9525	IC	1.9403	IC
13	7.4164	IC	2.8188	IC	6.8404	IC
14	12.0136	IC	5.5007	IC	9.4973	IC
15	3.7087	IC	1.9715	IC	2.8400	IC
16	2.7188	IC	1.3448	IC	2.2457	IC
17	5.7081	IC	1.4230	IC	4.6036	IC
18	3.4934	IC	3.5898	IC	3.6257	IC
19	10.6969	IC	10.1956	IC	10.5667	OoC
20	8.1595	IC	2.7431	IC	7.4650	IC

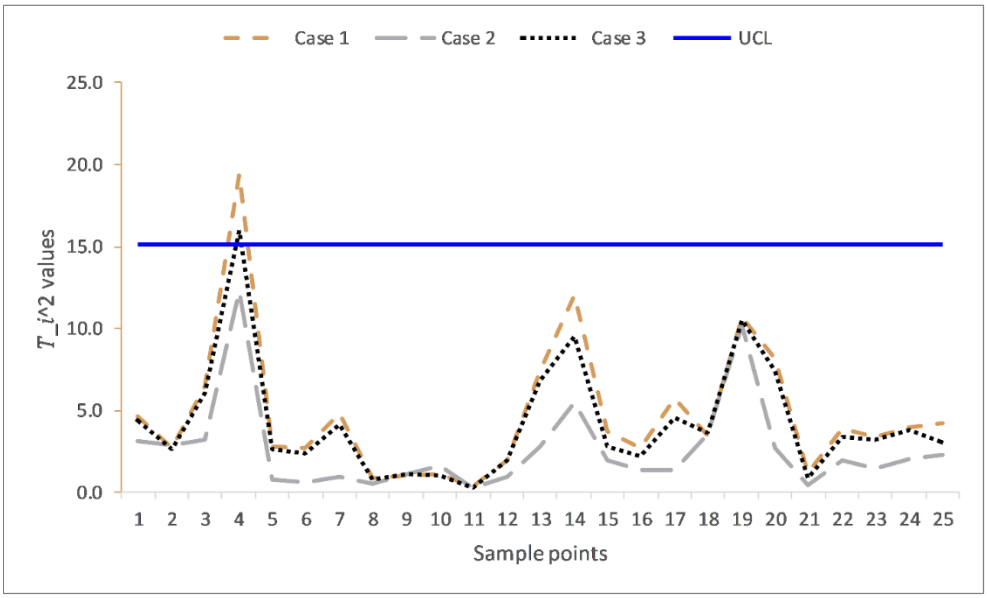

Figure 6. The multivariate Shewhart charts from real life data extracted from carbon fiber tubes.

5. Conclusions

This research paper evaluated the in-control performance of the multivariate Shewhart control chart when the parameters were estimated from phase-I samples that were prone to outliers. The study observed the negative effect of estimation and outliers on the chart's performance. Hence, we proposed a more efficient and robust multivariate Shewhart chart based on the Stahel-Donoho robust estimators and Mahalanobis distance to detect and screen outliers from the phase-I samples. Through the Monte-Carlo simulation approach, the ARL and SDRL for a different number of phase-I samples from small to medium to large were computed. The findings show that with the presence of outliers, even with large phase-I samples, the effect on the chart's performance was severe. The results further show that the proposed chart based on SDRE and Mahalanobis distance restored the efficiency of the multivariate Shewhart chart with smaller phase-I samples. Therefore, it is rational to incorporate the SDRE and Mahalanobis distance in default multivariate Shewhart structures, especially when the process parameters are estimated from phase-I samples prone to outliers. The findings of this study were substantiated with real-life application in the manufacturing industry, where three qualities of carbon fiber tubes were monitored. The scope of this study was limited to monitoring the location parameter in a multivariate Shewhart chart. However, the study can be extended to monitoring dispersion parameters in multivariate Shewhart charts and other charting schemes, such as multivariate cumulative sum (MCUSM) and exponentially weighted moving average (MEWMA).

Author Contributions: Conceptualization, I.A.R. and N.A.; methodology, I.A.R., N.A. and M.R.; software, I.A.R. and M.R.A.; validation, N.A. and M.R.; formal analysis, I.A.R., N.A., M.R.A. and M.R.; investigation, I.A.R. and N.A.; resources, I.A.R.; data curation, I.A.R. and N.A.; writing—original draft preparation, I.A.R.; writing—review and editing, I.A.R., N.A., M.R.A. and M.R.; visualization, N.A., M.R.A. and M.R.; supervision, N.A., M.R.A. and M.R.; project administration, N.A. and M.R.; funding acquisition, N.A. and M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Deanship of Scientific Research (DRS) at King Fahd University of Petroleum and Minerals (KFUPM) under Project No. IN191047.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The simulated data used in this article may be generated in R, using parameter values and the algorithm in Section 2.5. The real data set is available from the book referenced in [2].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hawkins, D.M. *Identification of Outliers*; Chapman and Hall: London, UK, 1980; doi:10.1007/978-94-015-3994-4.
2. Montgomery, D.C. *Introduction to Statistical Quality Control*, 6th ed.; John Wiley & Sons, Inc.: New York, NY, USA, 2009.
3. Psarakis, S.; Vyniou, A.K.; Castagliola, P. Some recent developments on the effects of parameter estimation on control charts. *Qual. Reliab. Eng. Int.* **2014**, *30*, 1113–1129, doi:10.1002/qre.1556.
4. Saleh, N.A.; Mahmoud, M.A.; Jones-Farmer, L.A.; Zwetsloot, I.; Woodall, W.H. Another look at the ewma control chart with estimated parameters. *J. Qual. Technol.* **2015**, *47*, 363–382, doi:10.1080/00224065.2015.11918140.
5. Jones, L.A. The statistical design of ewma control charts with estimated parameters. *J. Qual. Technol.* **2018**, *34*, 277–288, doi:10.1080/00224065.2002.11980158.
6. Abbas, N.; Abujiya, M.R.; Riaz, M.; Mahmood, T. Cumulative sum chart modeled under the presence of outliers. *Mathematics* **2020**, *8*, 269, doi:10.3390/math8020269.
7. Raji, I.A.; Lee, M.H.; Riaz, M.; Abujiya, M.R.; Abbas, N. Outliers detection models in shewhart control charts; An application in photolithography: A semiconductor manufacturing industry. *Mathematics* **2020**, *8*, 857, doi:10.3390/math8050857.
8. Abbas, N. A robust S2 control chart with Tukey's and MAD outlier detectors. *Qual. Reliab. Eng. Int.* **2020**, *36*, 403–413, doi:10.1002/qre.2588.
9. Guarnieri, J.P.; Souza, A.M.; Jacobi, L.F.; Reichert, B.; da Veiga, C.P. Control chart based on residues: Is a good methodology to detect outliers? *J. Ind. Eng. Int.* **2019**, *15*, 119–130, doi:10.1007/s40092-019-00324-0.
10. Bakar, Z.A.; Mohemad, R.; Ahmad, A.; Deris, M.M. A comparative study for outlier detection techniques in data mining. In Proceedings of the 2006 IEEE Conference on Cybernetics and Intelligent Systems, Bangkok, Thailand, 7–9 June 2006.
11. Vidmar, G.; Blagus, R. Outlier detection for healthcare quality monitoring-A comparison of four approaches to over-dispersed proportions. In *Quality and Reliability Engineering International*; John Wiley and Sons Ltd.: Hoboken, NJ, USA, 2014; Volume 30, pp. 347–362.
12. Zhang, H.; Albin, S. Detecting outliers in complex profiles using a χ^2 control chart method. *IIE Trans. (Inst. Ind. Eng.)* **2009**, *41*, 335–345, doi:10.1080/07408170802323000.
13. Manenti, F.; Buzzi-Ferraris, G. Criteria for outliers detection in nonlinear regression problems. *Comput. Aided Chem. Eng.* **2009**, *26*, 913–917, doi:10.1016/S1570-7946(09)70152-X.
14. Militino, A.F.; Palacios, M.B.; Ugarte, M.D. Outliers detection in multivariate spatial linear models. *J. Stat. Plan. Inference* **2006**, *136*, 125–146, doi:10.1016/j.jspi.2004.06.033.
15. Fan, S.-K.S.; Huang, H.-K.; Chang, Y.-J. Robust multivariate control chart for outlier detection using hierarchical cluster tree in SW2. *Qual. Reliab. Eng. Int.* **2013**, *29*, 971–985, doi:10.1002/qre.1448.
16. Gnanadesikan, R.; Kettenring, J.R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **1972**, *28*, 81, doi:10.2307/2528963.
17. Pranata, A.; Sadik, K. Comparison of Hotelling, MVE and WD for Detecting Outlier in Robust Multivariate Control Chart. *Int. J. Sci. Eng. Res.* **2016**, *7*, 1138–1142.
18. Hubert, M.; Debruyne, M.; Rousseeuw, P.J. Minimum covariance determinant and extensions. *Wiley Interdiscip. Rev. Comput. Stat.* **2018**, *10*, e1421.
19. Stahel, W.A. *Breakdown of Covariance Estimators*; Eidgenössische Technische Hochschule: Zürich, Switzerland, 1981.
20. Maronna, R.A.; Yohai, V.J. The behavior of the Stahel-Donoho robust multivariate estimator. *J. Am. Stat. Assoc.* **1995**, *90*, 330–341, doi:10.1080/01621459.1995.10476517.
21. Rousseeuw, P.; Hubert, M. High-breakdown estimators of multivariate location and scatter. In *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 49–66. ISBN 9783642354946.
22. Donoho, D.L. *Breakdown Properties of Multivariate Location Estimator*; Harvard University: Cambridge, MA, USA, 1982.
23. Ghorbani, H. Mahalanobis distance and its application for detecting multivariate outliers. *Math. Inf.* **2019**, *34*, 583–595, doi:10.22190/FUMI1903583G.
24. Zuo, Y.; Cui, H.; He, X. On the Stahel-Donoho estimator and depth-weighted means of multivariate data. *Ann. Stat.* **2004**, *32*, 167–188, doi:10.1214/aos/1079120132.
25. Abid, M.; Nazir, H.Z.; Riaz, M.; Lin, Z. In-control robustness comparison of different control charts. *Trans. Inst. Meas. Control* **2017**, *40*, 3860–3871, doi:10.1177/0142331217734302.
26. Abid, M.; Nazir, H.Z.; Tahir, M.; Riaz, M.; Abbas, T. A Comparative Analysis of Robust Dispersion Control Charts with Application Related to Health Care Data. *J. Test. Eval.* **2019**, *48*, 247–259, doi:10.1520/JTE20180572.
27. Zwetsloot, I.M.; Schoonhoven, M.; Does, R.J.M.M. Robust point location estimators for the EWMA control chart. *Qual. Technol. Quant. Manag.* **2016**, *13*, 29–38, doi:10.1080/16843703.2016.1139845.