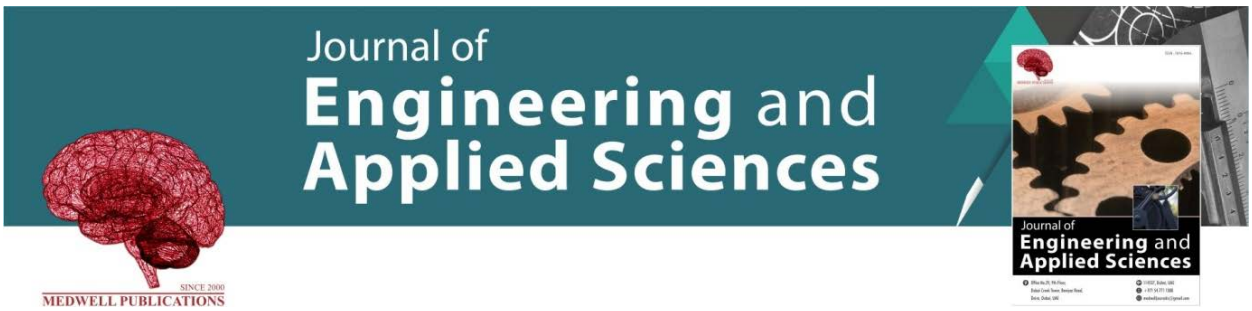

The Impact of Data Anomaly on EWMA Phase II Performance

[1]Ayu Abdul Rahman, [1]Sharipah Soaad Syed Yahaya, [2]Abdu Mohammed Ali Atta, [1]Nor Aishah Ahad and [1]Hashibah Hamid
[1]School of Quantitative Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
[2]Department of Mathematics, Statistics and Physics, College of Arts and Sciences, Qatar University, P.O. Box 2713 Qatar, Doha

Key words: ARL, EWMA control chart, process location, contamination, robust, trimmed mean, MADn

Corresponding Author:
Ayu Abdul Rahman
School of Quantitative Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia

Page No.: 3010-3015
Volume: 15, Issue 15, 2020
ISSN: 1816-949x
Journal of Engineering and Applied Sciences

Abstract: In applying control chart with estimated parameters for monitoring changes in a process, Phase I samples are typically assumed to be free of outliers or any other data anomaly. Naturally, the sample mean and the sample standard deviations are used as estimators, yielding efficient estimates for the chart. Nonetheless, when Phase I may be contaminated, this regular practice is no longer suitable as classical estimators are susceptible to the effect of outliers which in turn may affect control chart performance. This study shows that the effect is not trivial via. the application of EWMA control chart. Moreover, this study focuses on the effect using alternative and robust Phase I estimators on the EWMA when the chart is used to monitor changes in the process mean. In this study, an automatic trimmed mean estimator is used to provide estimate for the process mean. Meanwhile, for the standard deviation of the process, this study employs three different estimators including the corresponding robust scale estimator used in the trimming process of the location measure. Simulated data were used to test the performance of the EWMA control charts. The finding based on mean and percentiles of the run-length distribution shows quicker detection of out-of-control status when robust statistics were used to compute parameter estimates in Phase I of the EWMA chart upon contamination in the data set.

INTRODUCTION

In practice, control charts are widely used to monitor changes in a process from an in-control state to an out-of-control state. Generally, implemented in two phases, the monitoring of prospectively collected observations from the process occurs in the second phase, Phase II which requires that information of the process parameters readily available. Or if not, the process parameters shall be estimated in the first phase, Phase I. Common approach is to compute them via the usual sample mean and sample standard deviation. Nonetheless, these classical estimators are easily perturbed by outliers or any other data anomalies. Unfortunately, it has long

been recognized that in practical situations, Phase I data often contain these undesirable traits[1]. As such those estimates could be misleading and no longer a true representative of the in-control model. The downfall may be reflected in the inability of the chart to produce the desired in-control run length properties[2]. Jensen *et al*.[3] managed an extensive literature survey germane to the present discussion and enumerated some ideas for future research. Among their recommendations were to study robust or alternative Phase I estimators and their impact on Phase II chart performance.

To date, numerous robust works appear in SPC literature. On Shewhart-type charts, robust point estimators were considered by Langenberg and Iglewicz[4], Rocke[5, 6], Abu-Shawiesh and Abdullah[7] and Schoonhoven *et al*.[8], among many others researchers. Indeed as shown in the works of the aforementioned researchers, the effect of outliers can be attenuated via applications of robust statistics in establishing efficient control limits for the Shewhart chart. However, it is well noted that the Shewhart chart is it is not competitive for identifying small and moderate shifts in the process[9, 10]. Due to this limitation, community in SPC has started to focus on robustifying memory-type control charting structures namely Cumulative Sum (CUSUM) and Exponentially Weighted Moving Average (EWMA) charts. Both CUSUM and EWMA charts perform comparably as shown in the work by Hawkins and Wu[11] and considered to be very effective particularly in Phase II[12]. As memory-type charts, both CUSUM and EWMA use information from sequences of sample data and therefore, more sensitive to small shifts in the process[13].

Some notable robust work on CUSUM chart can be referred by Hawkins[14] who considered obtaining robust Phase I estimates via. wins orization a procedure in which an outlier in the data is replaced by some choice of threshold value. The researcher established the effectiveness of the proposed robust approach in keeping the false alarm rates acceptably close to the pre-specified value without much loss in the chart's sensitivity to detect mean shifts. Similarly, Rahman *et al*.[15] who were working on robustifying CUSUM chart via. highest breakdown point scale estimator namely median absolute deviation about the median (MADn) claimed similar good performances under contaminated data scenario. Meanwhile, through the application of EWMA control chart, Zwetsloot *et al*.[16] verified the advantageous of using trimmed mean and median, two robust statistics that are easily computed but very effective in mitigating the effect of outliers which in turn improve control chart performance considerably. Their research, however, left some room for improvement as we observe biased EWMA chart in terms of Average Run Length (ARL); a metric commonly used to measure control chart performance[17, 18]. Further explanation on the ARL will be given in Section 2. For now, it is worth to mention that the ARL-biased chart is undesirable in SPC, since, it implies that the chart signals frequently when the process is actually in-control and yet, fails to do so, under small changes in the process data. This study is intended to fill the gap and contribute to the existing literature by studying alternative Phase I estimators for mean (μ) and standard deviation (σ) of the process when EWMA control chart is used to monitor mean shifts.

In this study, a robust location estimator based on a trimming procedure is considered for estimating μ. It is an automatic trimmed mean estimator which has been used frequently in hypothesis testing, resulting in good results over type I error (i.e., false alarm rate) and power[19-21]. It is not to be confused with the usual trimmed mean estimator which assumes symmetric trimming and a fixed selection of trimming proportion before outliers can be discarded from the data. Conversely, in using a robust automatic trimmed mean estimator, no trimming proportion has to be specified prior to the application. This is because the estimator takes into consideration on the distributional shape of the data and as such the trimming can be asymmetric too if needed.

Under relaxation of the known process parameters assumption, standard deviation of the process is also estimated in this study using three different scale estimators. The performance of the EWMA control chart based on these three different scale estimators, paired with the aforementioned automatic trimmed mean estimator is the focus of this study.

DESCRIPTION OF EWMA CONTROL CHART

This study involves the EWMA control chart for location where data are prospectively collected in a rational subgroup concept defined by their sample of size n. The observations are assumed to be independent and normally distributed (iid) with parameters μ and σ, denoted as $Y_{ij} \sim$ iid $N(\mu, \sigma)$ where i = 1, 2, ..., n, j = 1, 2, ... when the process is in-control, let $\mu = \mu_0$ and $\sigma = \sigma_0$.

Introduced by Roberts[22], the EWMA chart involves plotting the statistic based on weighted average of all sample means defined as $E_j = (1-\lambda) E_{j-1} + \lambda \bar{Y}_j$ where, $\bar{Y}_j$ is the mean of sample j and $\lambda \in (0, 1]$ which is the weighting factor. The starting value E_0 is set to μ in this study. The chart signals a change in the process whenever E_j falls outside the control limits defined as:

$$\left.\begin{array}{l} LL_E = \mu - L\frac{\sigma}{\sqrt{n}}\sqrt{\frac{\lambda}{2-\lambda}}, \\ UCL_E = \mu + L\frac{\sigma}{\sqrt{n}}\sqrt{\frac{\lambda}{2-\lambda}} \end{array}\right\} \quad (1)$$

where, LCL_E and UCL_E denote the lower and upper limits for the EWMA, respectively. The width of these so-called asymptotic control limits is set by a positive constant, L. The distinction in EWMA performance based on a symptotic limits and time-varying limits can be referred by Steiner[23].

The parameters μ and σ in Eq. 1 usually need to be estimated in Phase I. For such purpose, choices for the estimators are discussed in Subsection 2.1. In regard to the construction and evaluation of the Phase II EWMA chart, this study considers the Average Run Length (ARL) defined as the expected number of plotted chart statistics before a signal occurs. Applicable in both states of the process, i.e., in-control and out-control, the ARL sends false alarms when there is no change in the process mean but as soon the mean shifted, the ARL signifies the shift detection capability. Known as in-control and out-of-control ARL, hence, forth denoted by ARL_0 and ARL_1, respectively, a good chart is viewed with a significantly large ARL_0 while its ARL_1 is as small as possible.

Using the ARL, the following setting for the EWMA design charting structure is adhered: $ARL_0 \approx 370$ (which is an in-control performance of the Shewhart chart with generic 3-sigma limits) when the process is in-control and normally distributed. For $\lambda = 0.13$, L is set at 2.92 for n = 10 as advised by Jones[24].

PHASE I ESTIMATORS

Consider X_{ij}, i = 1, 2, ..., n and j = 1, 2, ..., m denote the Phase I observations in which the data are used to obtain $\hat{\mu}$ and $\hat{\sigma}$ estimates μ of and σ, respectively. Define $X_{(i)j}$ as the ith-order statistic within sample j, M_j as the median of sample j, S_j as the standard deviation of the sample j, R_j as the range of sample j and MAD as the median of the values $|X_1-M|, ..., |X_n-M|$. For estimating μ, we consider average automatic trimmed mean $\bar{X}_{at}$ defined as:

$$\hat{\mu} = \frac{1}{m}\sum_{j=1}^{m} \bar{X}_{at_j} \quad (2)$$

Introduced by Wilcox and Keselman[21], the $\bar{X}_{at}$ averages the values remaining in sample j after outliers are discarded:

$$\bar{X}_{at} = \frac{\sum_{i=i_1}^{n-i_2} X_{(i)}}{n-i_1-i_2} \quad (3)$$

Where:

i_1 = The number of observations X_i such that $(X_i-M)<-2.24\ (MAD_n)$

i_2 = The number of observations X_i such that $(X_i-M)->2.24\ (MAD_n)$ and $MAD_n = MAD/0.6745$

Table 1: Phase I estimators and the investigated EWMA charts

Estimators		
Location	Scale	Charts
$\bar{X}_{at}$	MAD_n	E_{atM}
	$\bar{S}/c_4$	E_{ats}
	$\bar{R}/d_2$	E_{atR}

The use of M and MAD_n in detecting outliers results in the highest possible breakdown point (50%) for the automatic trimmed mean estimator[21]. It means that the estimate remains bounded when less than half of the data are contaminated. Moreover, the trimming procedure provided by this estimator takes into consideration the distributional shape of data which excludes any unnecessary loss information. In practice, σ can estimated by one of the following estimators:

$$\hat{\sigma} = \bar{S}c_{4,}\hat{\sigma} = \bar{R}/d_2; \hat{\sigma} = \overline{MAD_n} \quad (4)$$

where:

$$\bar{S} = \frac{\sum_{j=1}^{m} S_j}{m} \text{ and } \bar{R} = \frac{\sum_{j=1}^{m} R_j}{m}$$

Mahmoud *et al.*[25] and $\overline{MAD_n} = \sum_{j=1}^{m} MAD_{n_j}/m$. It should be noted that the constants c_4 and d_2 depend only on the sample size n and are tabulated for normal distribution in various statistical textbooks, for example, Montgomery[10]. Both sample standard deviation and sample range have breakdown point of 0 which means that the estimates based on this statistic are unreliable in the presence of outliers. The MAD_n on the other hand has the highest possible breakdown point, i.e., 50%. This high breakdown property is particularly useful when degree of contamination in Phase I is quite serious as demonstrated later in this study. The estimators considered are listed in Table 1.

DATA SCENARIOS AND SIMULATION PROCEDURES

Table 1 is the pairing of the chosen robust location estimator, $\bar{X}_{at}$ with three different scale estimators for obtaining Phase I estimates. In total, three EWMA control charts with various robust limits were constructed in this study. Their performances were investigated when Phase I data may or may not be contaminated. For such purpose, several data scenarios were considered in this study. First is the in-control environment in which Phase I data are study. Their performances were investigated when Phase I data may or may not be contaminated. For such purpose,

Table 2: ARL of the EWMA control chart with estimated parameters for m = 50 and n = 10

Phase I		ARL and percentiles															
		δ = 0				δ = 0.1				δ = 0.3				δ = 0.5			
Contaminations	Charts	ARL	25th	50th	90th	ARL	25th	50th	90th	ARL	25th	50th	90th	ARL	25th	50th	90th
In-control	E_{atM}	198	20	50	355	86	14	45	430	10	13	53	7	5	2	6	12
	E_{ats}	367	28	106	936	145	20	84	445	12	15	55	7	6	2	6	12
	E_{atR}	362	28	106	940	145	22	90	448	12	15	55	7	6	2	6	11
CN1	E_{atM}	181	12	50	276	80	16	143	682	10	8	45	345	5	4	8	20
	E_{ats}	324	30	121	354	127	48	170	767	12	4	23	115	5	4	8	34
	E_{atR}	324	32	121	349	127	54	187	698	12	4	21	118	5	4	8	37
CN2	E_{atM}	239	26	75	486	107	52	101	542	11	3	45	345	5	5	11	65
	E_{ats}	450	60	78	652	176	65	189	980	13	2	23	115	6	6	11	78
	E_{atR}	446	62	100	582	175	69	187	1000	13	1	21	118	6	9	11	76
CN3	E_{atM}	1872	34	109	620	882	60	289	1027	24	20	60	321	8	5	10	100
	E_{ats}	3133	77	342	980	1538	98	543	998	36	25	56	198	9	5	7	128
	E_{atR}	3183	76	456	1027	1521	87	525	879	37	21	76	187	9	7	7	101

several data scenarios were considered in this study. First is the in-control environment in which Phase I data are described by $N(\mu_0, \sigma_0)$. Without loss of generality, μ_0 is set at 0 and σ_0 takes value of 1. Meanwhile, indication of data anomaly in Phase I is captured using Contaminated Normal (CN) distribution: a mixture distribution frequently used in SPC literature when issues of robustness and/or outliers need to be addressed, for examples, Nazir *et al*.[26] and Human *et al*.[27].

In CN distribution, (100-p)% of the observations come from N(0, 1) and the rest of the observations, i.e., p% come from N(0, w) where p denotes the proportion of contamination and w is the standard deviation of an outlier. It is assumed that these outliers appear occasionally in a subgroup. Thus, rather than affecting the subgroup as a whole, their appearance may be indicated by a single unusual value in the samples. To examine how sensitive the proposed EWMA control charts to these occasional outliers, these setting for CN distributions are adopted: (p, w) = (0.05, 5), (0.10, 5) and (0.10, 10). Accordingly, they are named as CN1, CN2 and CN3.

Monte Carlo simulation study was performed to obtain the ARL of the investigated EWMA charts. Several percentile points of the run length distribution, i.e., 25th, 50th and 90th are also reported to support the ARL finding. To obtain those numerical values reported in Table 2, the following simulation study was conducted. First, m = 50 and n = 10 were drawn from the in-control N(0, 1) and three contaminated normal distributions: CN1, CN2 and CN3. Next, the mean was computed with the automatic trimmed mean estimator, $\bar{X}_{at}$, for $\hat{u}$ and three scale estimators (Table 1) for $\hat{\sigma}$. Based on these estimates, the EWMA charts were constructed according to Equation 1. Next, 15,000 new samples of size 10 were drawn from $N(\delta, 1)$ until the associated E_j falls outside the control limits. This gives the corresponding run length of j-1. The calculations were made for $\delta = \{0, 01, 0.3, 0.5\}$ where, δ is the shift size in standard deviation units. Iterations were completed for 10,000 runs. By averaging the sum of the run lengths over 10,000 runs, we have the ARL. The results are presented in Table 2.

SIMULATION OUTCOMES

Table 2 presents the ARL and percentiles of the length distribution of the investigated EWMA control chartwhen Phase I data follow, CN1, CN2 and CN3. It should be noted that all three charts share the same estimate for μ. Therefore, the difference in the performance of the EWMA control chart in Phase II is discussed with respect to the choice for σ.

First, consider the situation where Phase I data is uncontaminated, N(0, 1). This is illustrated by the upper part of the table under in-control scenario. The EWMA control charts based on $\bar{S}/c_4$ and $\bar{R}/d_2$ have ARL_0 of approximately 370. On the other hand, the EWMA control chart based on the robust estimator $\overline{MAD_n}$ falls short as it yields an extremely low ARL_0 and low 50th percentile of the in-control ARL. Thus, the chart is expected to give more false alarms than expected if the process is actually in-control. In this situation, it is undesirable to use extremely robust estimators based on median, i.e., $\bar{X}_{at}$ and the $\overline{MAD_n}$ to construct the control limits for the EWMA chart. Despite the drawback, the EWMA chart based on $\bar{X}_{at}$ and the $\overline{MAD_n}$ has the quickest declining of ARL_1 values among all. This is evident as soon as δ shifts from 0 to some value.

Next, consider situation where Phase I data may be contaminated. Finding based on CN1, CN2 and CN3 for the three charts indicate that the ARL_0 levels can be much lower than the nominal ARL or sometimes much higher than the expected 370. On CNN which is the worst-case contamination scenario prescribed in this study, the ARL_0 can be as high as 8.6 times than 370. It happens when the EWMA control chart is designed with $\bar{S}/c_4$. A high value of ARL_0 is desirable. However, an ARL_0 value as large as

3183 (from $\bar{S}/c_4$) or 3133 (from $\bar{R}/d_2$) is definitely not useful in practice since it alludes to detection delay of a process change. Indeed, the corresponding ARL_1 for these two EWMA charts, i.e., E_{at_S} and E_{at_R} are extremely high when the process mean starts to shift to some out-of-control mean value.

General observation shows that is getting harder to rein the ARL_0 as level of contamination increases in Phase I data. The situation worsened with inflation in the variance of the outliers. This is particularly true when is estimated via. non-robust estimators. The variation in ARL_0 is slightly less severe under this circumstance if $\overline{MAD_n}$ was employed in designing the chart. More importantly using $\overline{MAD_n}$ in Phase I yields the smallest ARL_1 for the EWMA chart in all scenarios for all magnitude of shifts, δ. It is also equally important to note that all charts are ARL-unbiased, i.e., $ARL_0 > ARL_1$ for all δ. Clearly, the use of robust automatic trimmed mean estimator, $\bar{X}_{at}$ for estimating the location parameter is advantageous. With the highest possible breakdown point, the $\bar{X}_{at}$ certainly deserve serious consideration in the implementation of any control charting structure.

CONCLUSION

This study has studied the effect of parameter estimation on the Phase II performance of the EWMA chart when Phase I may or may not contain contaminated observations. The finding shows that practical use of control chart could be undermined when Phase I consist of some data anomalies. Comparison between the EWMA control chart has been instigated when two non-robust dispersion estimators, i.e., sample standard deviation and sample range are paired with a robust point location estimator, namely the automatic trimmed mean. Apart from that, the analysis has been extended to cover the performance of the EWMA control chart when MADn which is a highly robust measure of scale estimate is paired with the previously employed robust point location estimator. Under Phase I normal and contaminated normal distributions, the resulting Phase II EWMA charts based on the non-robust scale estimates are almost identical. Best performance in the out-of-control situation, however, is offered by the MADn. Note that even though this estimator is preferred among the discussed alternative measure of process standard deviation, the resulting in-control ARL of the chart under normality is extremely low. Thus, prudent care ought to be exercised if one wish to employ both robust point location and dispersion measures in Phase I as they may yield a very narrow Phase II limits.

ACKNOWLEDGEMENT

The researchers would like to acknowledge the work that has led to this study which is supported by Universiti Utara Malaysia, Fundamental Research Grant Scheme (S/0 Code 13578) of the Ministry of Higher Education, Malaysia.

REFERENCES

01. Janacek, G.J. and S.E. Meikle, 1997. Control charts based on medians. J. R. Stat. Soc. Ser., 46: 19-31.
02. Hawkins, D.M., P. Qiu and C.W. Kang, 2003. The changepoint model for statistical process control. J. Qual. Technol., 35: 355-366.
03. Jensen, W.A., L.A. Jones-Farmer, C.W. Champ and W.H. Woodall, 2006. Effects of parameter estimation on control chart properties: A literature review. J. Qual. Technol., 38: 349-364.
04. Langenberg, P. and B. Iglewicz, 1986. Trimmed mean X and R charts. J. Qual. Technol., 18: 152-161.
05. Rocke, D.M., 1989. Robust control charts. Technometrics, 31: 173-184.
06. Rocke, D.M., 1992. XQ and RQcharts: Robust control charts. Statistician, 41: 97-104.
07. Abu-Shawiesh, M.O. and M.B. Abdullah, 1999. New robust statistical process control chart for location. Qual. Eng., 12: 149-159.
08. Schoonhoven, M., H.Z. Nazir, M. Riaz and R.J. Does, 2013. Robust location estimators for the X control chart. Qual. Control Appl. Stat., 58: 25-26.
09. Reynolds Jr., M.R. and Z.G. Stoumbos, 2005. Should exponentially weighted moving average and cumulative sum charts be used with Shewhart limits?. Technometrics, 47: 409-424.
10. Montgomery, D.C., 2013. Introduction to Statistical Quality Control. 7th Edn., John Wiley&Sons, Hoboken, New Jersey, USA.,.
11. Hawkins, D.M. and Q. Wu, 2014. The CUSUM and the EWMA head-to-head. Qual. Eng., 26: 215-222.
12. Woodall, W.H., 2017. Bridging the gap between theory and practice in basic statistical process monitoring. Qual. Eng., 29: 2-15.
13. Zwetsloot, I.M. and W.H. Woodall, 2017. A head-to-head comparative study of the conditional performance of control charts based on estimated parameters. Qual. Eng., 29: 244-253.
14. Hawkins, D.M., 1993. Robustification of cumulative sum charts by winsorization. J. Qual. Technol., 25: 248-261.
15. Rahman, A.A., S.S.S. Yahaya and A.M.A. Atta, 2018. A robust CUSUM control cumulative sum control chart for monitoring the process mean based on a high breakdown point scale estimator. J. Eng. Applied Sci., 13: 3423-3429.

16. Zwetsloot, I.M., M. Schoonhoven and R.J. Does, 2016. Robust point location estimators for the EWMA control chart. Qual. Technol. Quant. Manage., 13: 29-38.
17. Sunthornwat, R., Y. Areepong and S. Sukparungsee, 2017. Average run length of the long-memory autoregressive fractionally integrated moving average process of the exponential weighted moving average control chart. Cogent Math., Vol. 4,
18. Sunthornwat, R., Y. Areepong and S. Sukparungsee, 2018. Analytical and numerical solutions of average run length integral equations for an EWMA control chart over a long memory SARFIMA process. Songklanakarin J. Sci. Technol., 40: 885-895.
19. Ochuko, T.K., S. Abdullah, Z. Zain and S.S.S. Yahaya, 2015. Winsorized modified one step m-estimator in alexander-govern test. Mod. Applied Sci., 9: 51-67.
20. Wilcox, R.R., 2003. Multiple comparisons among dependent groups based on a modified one-step M-estimator. J. Applied Stat., 30: 1231-1241.
21. Wilcox, R.R. and H.J. Keselman, 2003. Repeated measures one-way ANOVA based on a modified one-step M-estimator. Br. J. Math. Stat. Psychol., 56: 15-25.
22. Roberts, S.W., 1959. Control chart tests based on geometric moving averages. Technometrics, 1: 239-250.
23. Steiner, S.H., 1999. EWMA control charts with time-varying control limits and fast initial response. J. Qual. Technol., 31: 75-86.
24. Jones, L.A., 2002. The statistical design of EWMA control charts with estimated parameters. J. Qual. Technol., 34: 277-288.
25. Mahmoud, M.A., G.R. Henderson, E.K. Epprecht and W.H. Woodall, 2010. Estimating the standard deviation in quality-control applications. J. Qual. Technol., 42: 348-357.
26. Nazir, H.Z., N. Abbas, M. Riaz and R.J. Does, 2016. A comparative study of memory-type control charts under normal and contaminated normal environments. Qual. Reliab. Eng. Int., 32: 1347-1356.
27. Human, S.W., P. Kritzinger and S. Chakraborti, 2011. Robustness of the EWMA control chart for individual observations. J. Applied Stat., 38: 2071-2087.