Quality Engineering
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lqen20

The CUSUM and the EWMA Head-to-Head

Douglas M. Hawkins [a] & Qifan Wu [a]
[a] School of Statistics , University of Minnesota , Minneapolis , Minnesota
Published online: 17 Mar 2014.

To cite this article: Douglas M. Hawkins & Qifan Wu (2014) The CUSUM and the EWMA Head-to-Head, Quality Engineering, 26:2, 215-222

To link to this article: http://dx.doi.org/10.1080/08982112.2013.817014

Quality Engineering, 26:215–222, 2014

ISSN: 0898-2112 print/1532-4222 online
DOI: 10.1080/08982112.2013.817014

Taylor & Francis
Taylor & Francis Group

The CUSUM and the EWMA Head-to-Head

Douglas M. Hawkins, Qifan Wu
School of Statistics, University of Minnesota, Minneapolis, Minnesota

ABSTRACT The cumulative sum (CUSUM) and the exponentially weighted moving average (EWMA) control charts are alternatives to the Xbar chart. The CUSUM's theoretical optimality suggests that it should outperform the EWMA for detecting persistent shifts, but practitioners have long thought that the two perform about equally. Each also involves design decisions on the likely shift in the process. This article quantifies the effect of these choices and concludes that, though the CUSUM outperforms the EWMA at the shift for which each was designed, if the actual shift is smaller than that used in the design, the EWMA may respond faster.

KEYWORDS cumulative sum control chart, exponential weighted moving average control chart, statistical process control, inertia, process mean, steady state, sustained shift

INTRODUCTION

Since Walter Shewhart introduced the first control chart for statistical process control in the 1920s, the Shewhart Xbar chart has remained popular. The major reasons for this are its ease of implementation without intensive statistical training and its low cost in time and resources. However, as asserted by Stoumbos et al. (2000, p. 993), "such simple charts are usually far from optimal and may even be inappropriate." In particular, Shewhart charts are not competitive for detecting small but sustained shifts in the process (Hawkins and Olwell 1998; Reynolds and Stoumbos 2005). This failing of Shewhart charts comes from their limitation to the information from only the most recent rational group, ignoring any information contained earlier in the data sequence (Montgomery 2013). The Western Electric supplementary runs rules were developed as a way to carry information forward from successive rational groups, but these rules have been found to be not particularly effective. Champ and Woodall (1987) gave a compendium of the rules and showed that the improvement in out-of-control behavior came at the cost of a disproportionate increase in false alarms, leading to the conclusion that the rules are of dubious value for hastening the response to moderate-sized persistent shifts. As Hawkins and Zamba (2003) pointed out, it is exactly these moderate-sized sustained shifts that may be most damaging, supporting a need for methods more sensitive to shifts between one-half and one standard deviation.

The key to improved performance is accumulating evidence from sequences of process readings. Two major charts that do this are the

Address correspondence to Douglas M. Hawkins, School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis, MN 55455. E-mail: dhawkins@umn.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lqen.

cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) control charts. They both accumulate information from successive readings and signal a change when a shift occurs, even if the change is relatively small so that a Shewhart Xbar chart fails to detect it. According to Montgomery (2013), the general consensus is that the practical performances of the CUSUM and EWMA are quite similar and neither of them has a clear advantage over the other. Thus, users only need to implement one or the other to monitor their process. Some statistical process control practitioners recommend that the CUSUM or the EWMA charts be used in combination with the Shewhart charts (Hawkins and Olwell 1998; Lucas 1982; Ryan 2011; Woodall 2000), to gain the power of the former to detect small sustained shifts and the latter to detect large, possibly intermittent, shifts. Reynolds and Stoumbos (2005) provided a deeper discussion of this approach. However, this is not our concern in this article. We only address the question of which is more effective in detecting sustained shifts in mean, the CUSUM or the EWMA.

DEFINITIONS AND FORMULAS

Suppose that the process produces readings that, in control, have mean μ_0 and standard deviation σ, both of which are assumed known. Unknown special (or assignable) causes can intervene at any time and disrupt this in-control behavior. When a special cause event happens, it will lead to a change in mean, or in variability, or both, or even in distribution. We will only concentrate on the change in process mean. Some special causes are intermittent and resolve themselves even without corrective action; others are sustained, or persistent, and continue until the process owner takes corrective action. This is the type of shift we consider.

The CUSUM and the EWMA each have two design parameters, one of which tunes the chart for a particular magnitude of shift in the mean and the other of which sets the in-control (IC) average run length (ARL). But the actual shift that occurs when the process goes out of control will not necessarily be the shift to which the chart was tuned, raising the question of how sensitive chart performance is to variation in the size of the actual out-of-control shift compared to the design value.

Another issue is the time at which the process goes out of control. This could be right at, or close to, the time at which the charting begins—this situation is called an initial-state shift. Or the shift could happen after the chart has been in operation for some time—this is called a steady-state shift. The CUSUM chart has a well-known optimality property: if a shift occurs in steady state, the CUSUM to which it is tuned has a faster average response than does any other chart (Hawkins and Olwell 1998), but little formal work has been done on the size of this benefit, the impact of initial-state shifts, or the occurrence of a different-sized shift.

Turning to definitions, the CUSUM charting statistic is applied to an unending stream of readings $X_1, X_2, \ldots, X_n$. CUSUMs are parametric procedures. Each distributional model for the data defines a particular CUSUM, based on the likelihood ratio test, for a shift in its parameter. The most familiar CUSUM is that for a change in the mean of normal data. It requires a reference value k, which tunes the CUSUM to be particularly sensitive to a specific anticipated shift. The CUSUM to detect an upward shift in process mean is initialized to

$$C_0^+ = c. \tag{1}$$

Traditionally c was zero, but this makes the CUSUM insensitive to shifts that occur soon after startup. For this reason, it is common to give the CUSUM a head start by initializing it to some c greater than zero. Lucas and Crosier (1982) motivated this approach, calling it the fast initial response CUSUM. After this initialization, the CUSUM is updated with each new process reading,

$$C_n^+ = \max\ (0, C_{n-1}^+ + (X_n - \mu_0) - k), \tag{2}$$

and signals if $C_n^+ > H$, where H is a second design constant called the decision interval that is used to fix the IC ARL at some target level.

To tune the CUSUM for an upward shift from μ_0 to μ_1, set $k = (\mu_1 - \mu_0)/2$. In other words, k is half the anticipated shift size. It is known that if the process does indeed have a shift in the steady state from μ_0 to μ_1, then there is no other procedure with the same IC ARL that will match the performance of the CUSUM. This is the theoretical optimality of the CUSUM. CUSUMs for other parameters and other distributions share this same optimality property. Hawkins and Olwell (1998) described the general methodology for setting up CUSUMs in chapter 6, and other chapters list some of the other distributions for which CUSUMS are known.

An attraction of the CUSUM is that, following a signal, it provides simple and quite effective estimates of the time and the magnitude of the shift. The CUSUM resets to zero when the data stream indicates that the process is in control, so when the CUSUM does exceed its control limit, an estimate of the end of the in-control period is given by the time of the most recent reset. Further, if the mean shifts from μ_0 to some value $\mu_1>\mu_0$, the successive values of C_n^+ will drift up with an average slope of $\mu_1-\mu_0-k$, and so the slope of the line from the axis up to the point that signals gives an estimate of $\mu_1-\mu_0-k$, and hence (since μ_0 and k are known) of the new mean μ_1u.

As the shift may, in most settings, be either upward or downward, the upward CUSUM is implemented along with a downward CUSUM. This downward CUSUM usually mirrors the upward shift and uses the same reference value k. The equation for downward CUSUM is

$$\begin{aligned} C_0^- &= -c \\ C_n^- &= \min\ (0, C_{n-1}^- + (X_n - \mu_0) + k). \end{aligned} \quad [3]$$

The pair of charts signals if either $C_n^+>H$ or $C_n^-<-H$.

The other control chart, the EWMA, also requires two parameters: λ and L. It is defined by

$$\begin{aligned} Z_0 &= \mu_0 \\ Z_n &= \lambda X_n + (1-\lambda)Z_{n-1} \end{aligned}, \quad [4]$$

where λ is a parameter that determines the weight assigned to the current sample value. The EWMA's control limits are defined by

$$\begin{aligned} LCL &= \mu_0 - L\sigma\sqrt{\tfrac{\lambda}{2-\lambda}[1-(1-\lambda)^{2n}]} \\ UCL &= \mu_0 + L\sigma\sqrt{\tfrac{\lambda}{2-\lambda}[1-(1-\lambda)^{2n}]} \end{aligned} \quad [5]$$

and the EWMA signals if Z_n is either above the UCL or below the LCL. The constant L sets the IC ARL.

If the mean shifts from μ_0 to μ_1, then the EWMA drifts in expectation from its in-control value μ_0 up (or down) toward an asymptote at μ_1. Though this is not as tidy as the estimates from the CUSUM, the latest point on the EWMA provides an evolving estimate of the new process mean.

The control limits of the EWMA vary with n, going from $\mu_0 \pm L\sigma\lambda$ for $n = 1$ to the asymptotic value $\mu_0 \pm L\sigma\sqrt{\lambda/(2-\lambda)}$. Some users prefer using constant control limits and use the asymptotic value for all n. The price paid for this greater simplicity is that the wider control limits for small n reduce the ability of the EWMA to react quickly to early shifts. As we consider maximal performance more important than the modest simplicity gained by using a constant limit rather than one that varies, we favor using the exact control limits of Eq. [5] rather than the asymptotic control limits.

The EWMA is also tuned to the size of shift through the weight λ—small λ if you anticipate a small shift and large λ if you anticipate a larger shift (Lucas and Saccucci 1990), but there is no explicit formula to find the most appropriate λ for a particular anticipated shift.

A unique phenomenon of the EWMA is the so-called inertia effect. If an upward shift happens when the EWMA statistic is in the lower part of its range, it first has to come back up to the center line before it can continue up to the UCL. This delays the detection of the shift in μ (Montgomery 2013; Reynolds and Stoumbos 2006; Woodall and Mahmoud 2005). Various adaptions have been proposed such as the adaptive EWMA of Cappizi and Masarotto (2003) and one-sided EWMAs with reflecting barriers, but these refinements add complexity to what is otherwise a very simple procedure, and it is not clear that they are generally used.

Not all control charts suffer from inertia—Shewhart charts do not, and the CUSUM may actually be more effective in steady state than in initial state because in the steady state it is likely to already be on its way to the control limit.

Lucas and Saccucci (1990) provided a detailed comparison of the performance of the CUSUM and the EWMA, including a number of enhancements to the basic EWMA incorporating a head-start capability paralleling that of the CUSUM. In their comparisons, however, the EWMA was tested using the asymptotic control limit rather than the exact limit, which to our mind needlessly sacrifices some initial-state performance, and so their results apply to a different problem than our calculations using the exact limit.

SIMULATION METHODS

The performance of the charts then involves several factors—the chart itself (CUSUM or EWMA), the shift to which the chart is tuned, the time of the shift, and the magnitude of the shift. We explored the impact of these factors with simulation. Streams of random variables following a normal distribution with $\mu = 0$ and $\sigma = 1$ were generated. From the instant of going out of control, a constant δ was added to each subsequent reading in the sequence. We took 121 values of δ, from 0 to 3, with a step size 0.025, to represent the actual shifts of the process mean.

The CUSUM and EWMA were designed for small ($\delta = 0.5$), medium-sized ($\delta = 1$), and large ($\delta = 2$) shifts in mean. In our simulation, only upward shifts were studied because a downward shift gives the same behavior as an upward shift by the same absolute amount.

To explore the impact of the time of the shift, the artificial shifts were applied to the sample sequence from either the beginning of the sequence (simulating an initial-state shift) or after 50 terms of the sequence (simulating a steady-state shift.) The value 50 seems to be generally accepted as long enough to all but erase the effects of initial conditions, and this expectation was confirmed by some runs using 100 initial in-control readings that gave essentially identical results to those with 50.

The IC ARL was set to 500, which is a widely used choice. We have no reason to doubt that our substantive conclusions would carry over to other sensible choices for the IC ARL.

The reference value k of the three CUSUMs was set to the optimal level for that size shift, and the corresponding decision interval H was calculated using the software provided by Hawkins and Olwell (1998). This led to the design parameters in Table 1.

In all runs, the default head-start $c = 0.5H$ was used.

TABLE 1 **CUSUM Chart Parameters**

Shift	δ	k	H
Small	0.5	0.25	8.585
Medium	1	0.5	5.071
Large	2	1	2.665

TABLE 2 **EWMA Chart Parameters**

Shift	δ	λ	L
Small	0.5	0.047	2.595
Medium	1	0.134	2.883
Large	2	0.364	3.045

Turning to the design of the EWMA, there is no explicit formula for the best λ for a specific shift δ. The parameters for EWMA were selected from a table of steady-state optimal parameter settings kindly provided by Edgard M. Maboudou-Tchao of the University of Central Florida. He generated this table by a grid search of λ and L values yielding an in-control ARL of 500 and having the minimum out-of-control ARL at the design shift. The resulting parameters are shown in Table 2.

For each combination of chart, design shift size, and δ, 10 million data sequences were simulated and the run length from the time of shift to the signal time computed. Any data set in which a signal occurred before the actual time of the shift was discarded. From these, the out-of-control ARLs were computed.

RESULTS

Figures 1 to 4 show the ARL of each of these control charts. The ARL is on a logarithmic scale to better show its main features. Each figure shows the ARL of the chart tuned for small, medium, or large shifts. Dotted vertical lines show the three design shifts.

As expected, in Figures 2, 3, and 4, each of the three designs is best in some settings. The chart designed for $\delta = 0.5$ has the lowest ARL for all small δ values; that designed for $\delta = 2$ is best for all large δg values; and that designed for $\delta = 1$ is best in the middle range.

Figure 1 may be surprising—it shows that the EWMA with the smallest λ was the lowest ARL right across the range of δ values; the same observation

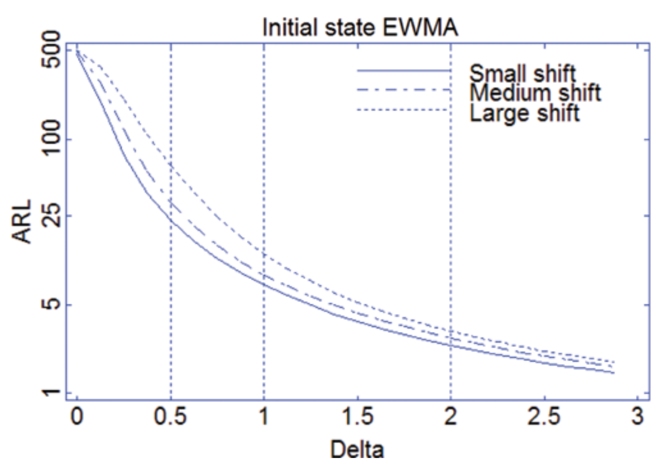

FIGURE 1 **OOC ARL of EWMA with initial-state shifts.**

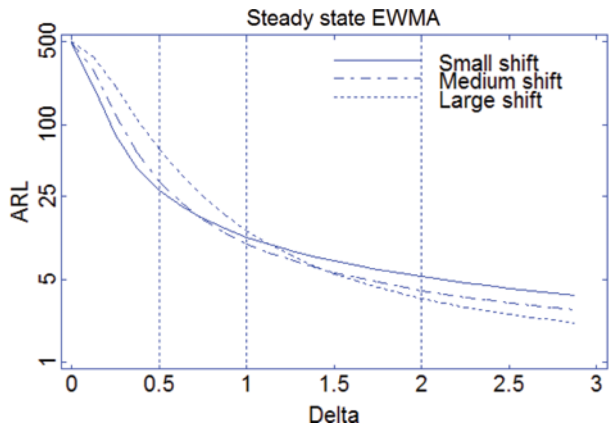

FIGURE 2 OOC ARL of EWMA with steady-state shifts.

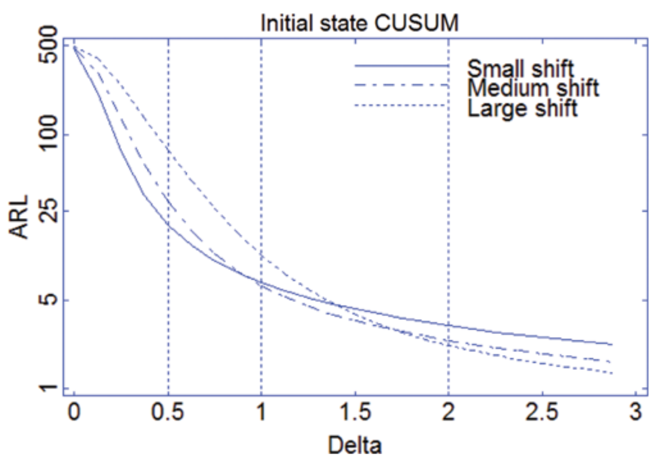

FIGURE 3 OOC ARL of CUSUM with initial-state shifts.

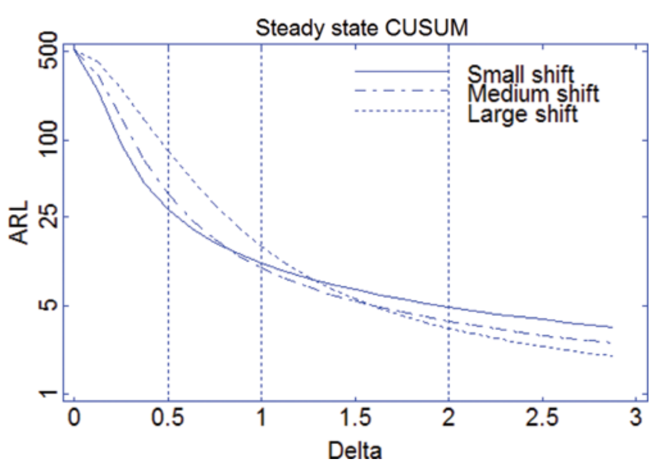

FIGURE 4 OOC ARL of CUSUM with steady-state shifts.

was, however, made by Frisén (2003) and Frisén and Sonesson (2006), so this is not a novel discovery.

Comparing Figures 1 and 2, the expected inertia of the EWMA is clear, with each line in Figure 2 being well above its Figure 1 counterpart. Perhaps more surprising, the CUSUM panels also show clear evidence of inertia. This comes about because of the head start, which turns out to initialize the CUSUM to an even more favorable starting value than its gets from being in steady state.

Coming back to the main topic, we are interested in comparing the performance of the EWMA and CUSUM with a view to routinely using one or the other. This choice is helped by looking at the ratio

$$\text{ARL Ratio} = \frac{\text{Cusum ARL}}{\text{EWMA ARL}} \qquad [6]$$

as a function of the shift δ. Where this ratio exceeds 1, the EWMA outperforms the CUSUM; where it is less than 1, the CUSUM is better.

Figure 5 shows the ratio for shifts occurring in the initial state. Looking at where the vertical markers cross the curves of the ratios, we can see that if the shift is about what was expected, the CUSUM outperforms the EWMA. More generally, the CUSUM designed for any particular shift mainly outperforms the EWMA designed for the same shift, with two striking exceptions:

- If you designed for a small shift ($\delta = 0.5$) and a large shift ($\delta > 1$) occurs, the misdesigned EWMA is much better than the misdesigned CUSUM.
- If you designed for a large shift ($\delta = 2$) and a small shift ($\delta < 0.8$) occurs, the misdesigned EWMA is much better than the misdesigned CUSUM.

Figure 6 shows the corresponding plot for the steady state. Here the EWMA outperforms the

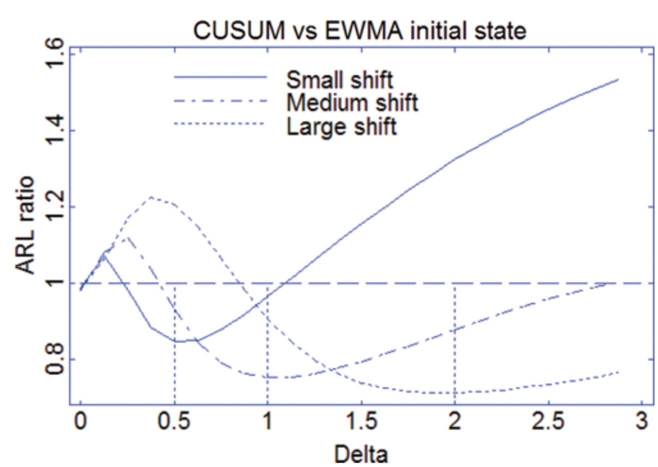

FIGURE 5 Comparison of EWMA and CUSUM for initial-state shifts.

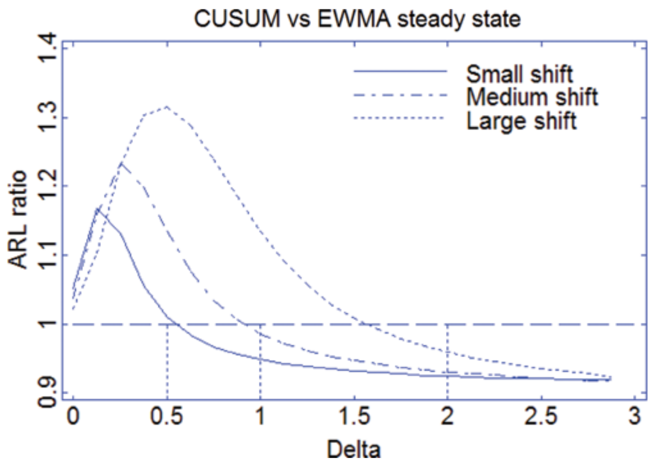

FIGURE 6 **Comparison of EWMA and CUSUM for steady-state shifts.**

CUSUM if the actual shift is smaller than the design shift (much smaller for $\delta = 2$), but above this threshold the CUSUM is uniformly better.

EXAMPLE

We illustrate the six charts with some simulated data. Montgomery (2013, p. 240) gives an example of monitoring the flow width of wafers produced by a hard-bake process. His Phase I data set gives the estimates μ_0 = 1.51 μm, σ =0.14 μm for the in-control setting. We simulated a data set with 25 readings coming from this normal distribution, following which the mean increased by 0.14 μm (i.e., one standard deviation) to 1.65 μm. This sequence was then run through the three EWMAs and the three CUSUMs discussed earlier. A time plot of the data is given in Figure 7 and the EWMA and CUSUM charts are shown in Figures 8–13.

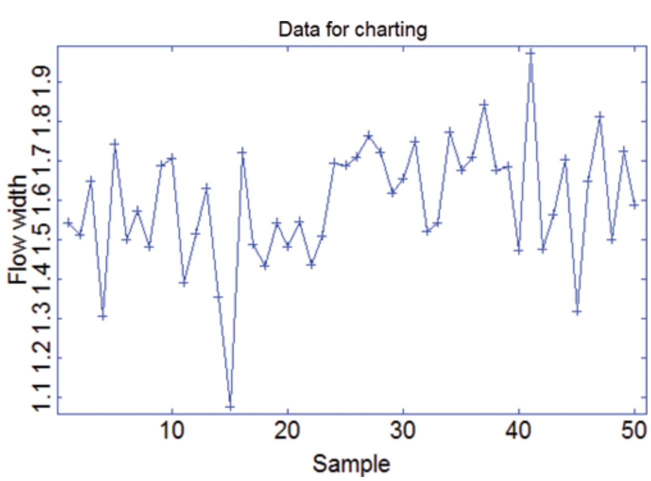

FIGURE 7 **Illustrative simulated data.**

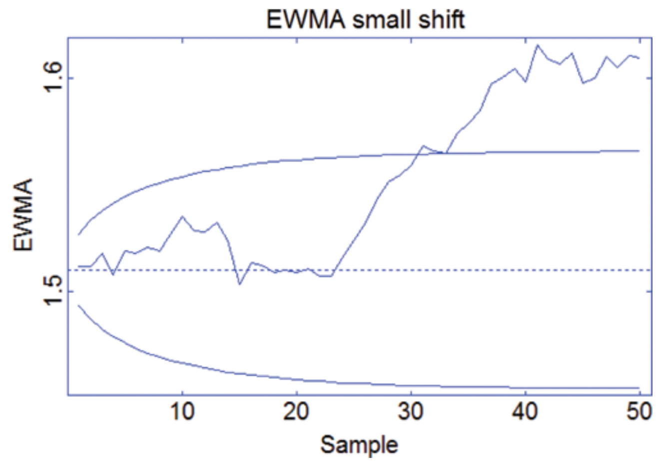

FIGURE 8 **EWMA tuned to small shift applied to example data.**

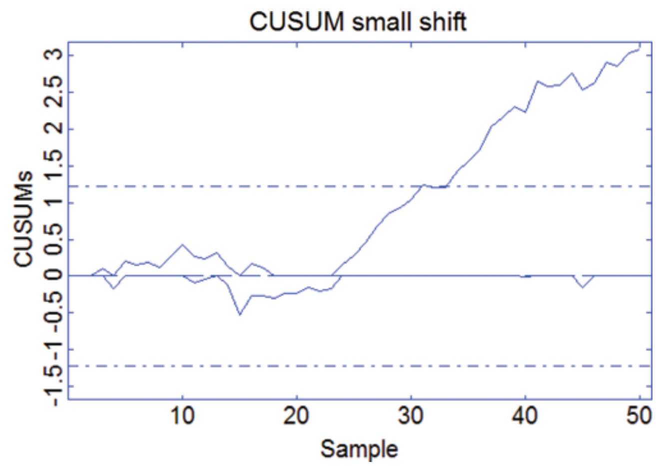

FIGURE 9 **CUSUM tuned to small shift applied to example data.**

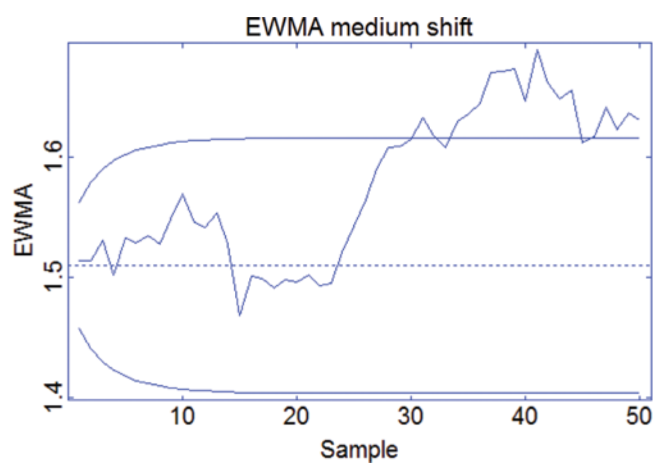

FIGURE 10 **EWMA tuned to medium-sized shift applied to example data.**

Following a signal by C^+ in the CUSUM, tracking backward, the last i for which C_i^+ was zero is an estimate of the last time the process was in control before the shift. Taking the slope of the segment from

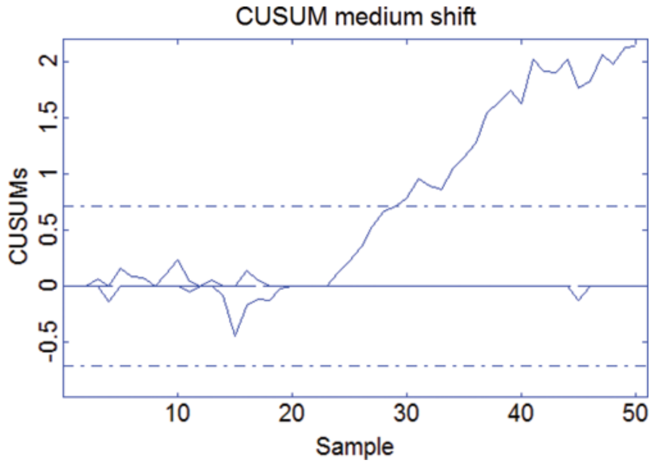

FIGURE 11 **CUSUM tuned to medium-sized shift applied to example data.**

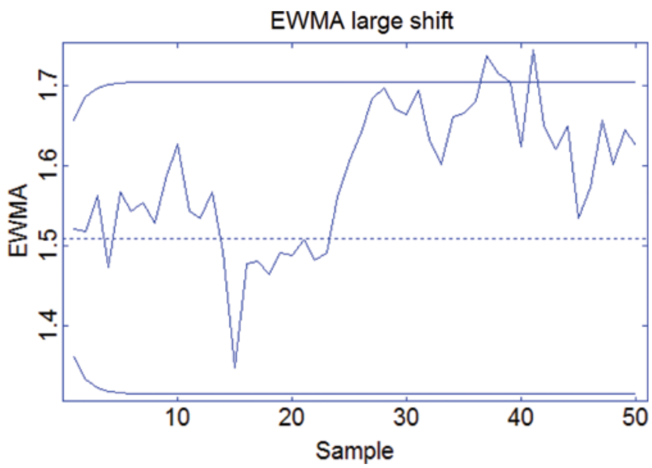

FIGURE 12 **EWMA tuned to large shift applied to example data.**

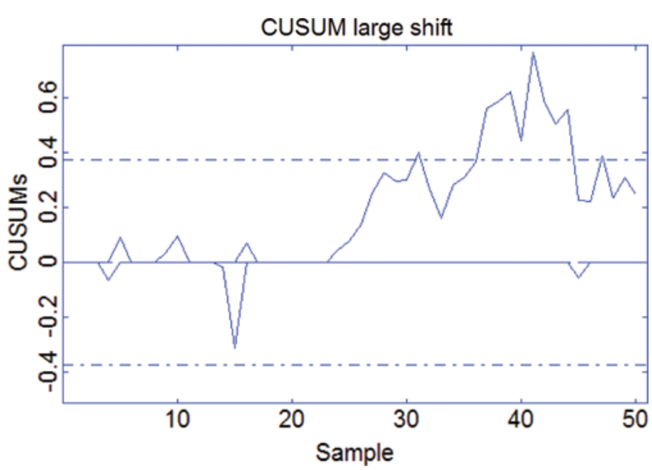

FIGURE 13 **CUSUM tuned to large shift applied to example data.**

that point to the point giving the signal and adding $\mu_0 + k$ gives an estimate of the new mean.

TABLE 3 **Follow-up Estimation From Example**

	EWMA		CUSUM		
Design	Signal time	Current mean	Signal time	Last IC time	New mean
Small	31	1.57	31	23	1.70
Medium	31	1.63	29	23	1.70
Large	37	1.74	31	23	1.70

The EWMA does not give an estimate of the time of change, but its value at the time of the signal gives an estimate of the new mean. Table 3 gives the values for the simulated data:

With regard to performance, the actual shift is one standard deviation, which our earlier discussion called a medium-sized shift, and, as expected, the CUSUM designed for a medium-size shift signaled sooner than any of the other five charts. The EWMA designed for a large shift had a much slower reaction, but the other four charts all signaled at the 31st reading, six readings after the actual shift occurred.

All three CUSUMs estimated the last in-control time as being observation 23, a bit earlier than when the shift was actually introduced. The postshift mean was simulated to be 1.65. All three CUSUMs gave the estimate 1.70, reflecting a known moderate upward bias in the estimate. The current mean of the EWMA designed for the medium-sized shift was almost exactly on target, but the other two were much further from the truth.

CONCLUSION

The inadequacy of the Shewhart chart for detecting small but persistent shifts in the process mean has led to the exploration of alternative tools. The relative merits of the CUSUM and the EWMA control charts have been debated for over half a century, but neither of them has supplanted the other. This is partly due to the way they are used—the EWMA is more convenient for estimating where the process mean is following a signal; the CUSUM is better for estimating when the shift occurred—but relative performance is an important consideration. Our work has evaluated the performances of both methodologies in different settings and situations.

The simulations agree with theory that if the actual shift is close to what was expected, the CUSUM outperforms the EWMA. The advantage is substantial for large shifts (25% for an initial-state shift with $\delta = 2$) and for initial-state shifts.

Misspecification matters. Both the CUSUM and the EWMA lose performance if the shift is much different than was anticipated, and it is worse to specify a too-large than a too-small shift.

Combining the two factors, if the user is wrong about the true shift and designs for a large δ but a much smaller shift occurs, then the EWMA suffers less than the CUSUM.

The decision of what shift to design for does not get as much attention as it perhaps should. Discussion in the literature often suggests that the user is expected to know that the mean is either its in-control value of (say) 1.51 μm or the out-of-control value of (say) 1.65 μm. The reality may be that users design for a shift large enough to matter but small enough not to be obvious. The work of Hawkins and Zamba (2003) suggested that the costliest sustained shifts are those somewhat less than a standard deviation, because they are small enough to escape detection but large enough to incur substantial costs. Motivated by guarding against them, one might favor CUSUMs with k values of 0.25 to 0.5 and EWMAs with λ values of 0.05 to 0.15. These values are in line with, though perhaps somewhat smaller than, textbook recommendations.

ACKNOWLEDGMENT

The authors are grateful for the referee's careful reading of the manuscript and numerous valuable suggestions for improvement.

ABOUT THE AUTHORS

Douglas M. Hawkins is a Professor, School of Statistics, University of Minnesota. He is a Fellow of ASQ.

Qifan Wu is a graduate student at Northwestern University.

REFERENCES

Capizzi, G., Masarotto, G. (2003). An adaptive exponentially weighted moving average control chart. *Technometrics*, 45:199–207.

Champ, C. W., Woodall, W. H. (1987). Exact results for Shewhart control charts with supplementary runs rules. *Technometrics*, 29: 393–399.

Frisén, M. (2003). Statistical surveillance. Optimality and methods. *International Statistical Review*, 71:403–434.

Frisén, M., Sonesson, C. (2006). Optimal surveillance based on exponentially weighted moving averages. *Sequential Analysis*, 25:379–403.

Hawkins, D. M., Olwell, D. H. (1998). *Cumulative Sum Charts and Charting for Quality Improvement*. New York: Springer.

Hawkins, D. M., Zamba, K. D. (2003). On small shifts in quality control. *Quality Engineering*, 16:143–149.

Lucas, J. M. (1982). Combined Shewhart–CUSUM quality control schemes. *Journal of Quality Technology*, 14: 51–59.

Lucas, J. M., Crosier, R. B. (1982). Fast initial response for CUSUM quality-control schemes: Give your CUSUM a head start. *Technometrics*, 25:295–301

Lucas, J. M., Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. *Technometrics*, 32(1):1–12.

Montgomery, D. C. (2013). *Introduction to Statistical Quality Control*. 7th ed. New York: Wiley.

Reynolds, M. R., Jr., Stoumbos, Z. G. (2005). Should exponentially weighted moving average and cumulative sum charts be used with Shewhart limits? *Technometrics*, 47(4):409–424.

Reynolds, M. R., Jr., Stoumbos, Z. G. (2006). Comparisons of some exponentially weighted moving average control charts for monitoring the process mean and variance. *Technometrics*, 48(4): 550–567.

Ryan, T. P. (2011). *Statistical Methods for Quality Improvement*. 3rd ed. New York: Wiley.

Stoumbos, Z. G., Reynolds, M. R., Jr., Ryan, T. P., Woodall, W. H. (2000). The state of statistical process control as we proceed into the 21st century. *Journal of the American Statistical Association*, 95(451):992–998.

Woodall, W. H. (2000). Controversies and contradictions in statistical process control. *Journal of Quality Technology*, 32:341–378.

Woodall, W. H., Mahmoud, M. A. (2005). The inertial properties of quality control charts. *Technometrics*, 47(4):425–436.