QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL
Qual. Reliab. Engng. Int. (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/qre.852

Research

The Funnel Experiment: The Markov-based SPC Approach

Gonen Singer and Irad Ben-Gal*,†
Department of Industrial Engineering, Tel Aviv University, Ramat-Aviv 69978, Israel

The classical funnel experiment was used by Deming to promote the idea of statistical process control (SPC). The popular example illustrates that the implementation of simple feedback rules to stationary processes violates the independence assumption and prevents the implementation of conventional SPC. However, Deming did not indicate how to implement SPC in the presence of such feedback rules. This pedagogical gap is addressed here by introducing a simple feedback rule to the funnel example that results in a nonlinear process to which the traditional SPC methods cannot be applied. The proposed method of Markov-based SPC, which is a simplified version of the context-based SPC method, is shown to monitor the modified process well. Copyright © 2007 John Wiley & Sons, Ltd.

Received 16 August 2006; Accepted 15 January 2007

KEY WORDS: process monitoring; change point; stochastic processes

1. *INTRODUCTION AND MOTIVATION*

The *funnel experiment*[1] is one of the most popular examples explaining the main principles in statistical process control (SPC). The example emphasizes that one should not interfere with a stable 'in-control' process while it fluctuates around its mean, but rather wait until a statistically significant 'out-of-control' signal is obtained. In the example, a 'no intervention' ('no compensation') rule is compared against simple feedback rules that are shown to increase the variance of the process.

The use of feedback rules to continuously regulate a running process is quite common in practice. In many industrial environments, selected process parameters are adjusted by feedback control policies that are based on past observations. As an example, the recipe settings of some wafers in semiconductor processes are adjusted by using the measurements of previously produced wafers. Such a feedback control scheme is called 'run to run' control. Gerold *et al.*[2] describe some implementations of run to run control at a Motorola microprocessor manufacturing facility. Studies of the run to run control were performed for chemical-mechanical polishing[3] and for silicon epitaxy[4]. Examples for other feedback control policies that are based on past observations include color adjustments between fabric batches[5] and management tampering of non-manufacturing environments[6].

It is well known that feedback policies, as well as more complicated control theory techniques that are not considered in this paper, often create nonlinear dynamics of the controlled observations[6–11]. Sometimes the structure of these dynamics may be established by the observed trajectories of the output. However, in general, such identification is not simple in noisy environments[12]. In any case, one cannot assume that the observations

*Correspondence to: Irad Ben-Gal, Department of Industrial Engineering, Tel Aviv University, Ramat-Aviv 69978, Israel.
†E-mail: bengal@eng.tau.ac.il

are independent once feedback control rules are used. As a result, it is improper to implement conventional SPC charts, which rely on the independence assumption, to monitor autocorrelated processes.

In recent years, considerable effort has been spent on developing SPC methods for autocorrelated data, such as variants of the ARIMA and residual charts[13–16] EWMA charts[17–21], and CUSUM charts[22–25]. These control charts have been developed based on the assumption that either the controlled observations follow a time-series model or that certain characteristics of the underlying process, such as the autocorrelation structure, are known. Using such *a priori* information, the control charts can be constructed from observations or residuals using traditional hypothesis tests. For other control charts, such as the ARMA chart[20], the model or the correlation structure is needed for determining the control limits and the chart parameters.

Within the above-mentioned framework of methods, Ben-Gal *et al.*[26] proposed a context-based SPC (CSPC) method to monitor a state-dependent process of varying dependence order. The suggested method is based on information theory principles, and uses a variable-order Markov (VOM) model to represent the monitored process, without requiring prior knowledge of the model parameters. Moreover, CSPC does not assume a closed-form time-series model, which is often required by conventional SPC approaches for autocorrelated processes. One disadvantage of the CSPC method is the involvement of a handful technical aspects in the construction and the monitoring of the VOM models. These aspects are partially based on information theory principles and require relatively complex computations with respect to simple SPC schemes.

In this paper, we propose a simplified version of the CSPC method. We call it Markov-based SPC (MSPC), because it is based on conventional (fixed-order) Markov models rather than the VOM models. Moreover, MSPC relies on well-known statistical concepts, such as Pearson's Chi-squared statistics and contingency tables, rather than the information theory concepts, such as relative entropy (Kullback Leibler distance), that are used by the CSPC method[26].

Similarly to previous publications[6,11,27,28], we use the funnel experiment as a convenient pedagogical framework to analyze the proposed MSPC. For this purpose, we extend the funnel experiment by considering a simple feedback rule, which is based on averaged discretized compensations of the funnel. The implementation of the MSPC to the known funnel experiment provides us with two advantages. First, it allows us to analyze the effects of feedback rules (even simple ones) on the dynamics of the monitored process. Second, it enables a nice framework to compare the proposed MSPC method with conventional SPC methods for autocorrelated processes. We find that the MSPC clearly outperforms the considered conventional methods when monitoring the manipulated process.

The rest of this paper is organized as follows. In Section 2, we briefly describe the funnel experiment and review some related literature. In Section 3, we introduce a simple feedback rule and analyze the resulting controlled process, which is found to be nonlinear and state-dependent. In Section 4, we illustrate that some conventional SPC methods fail to monitor the process generated by the introduced feedback rule. In Section 5, we use the proposed MSPC method to successfully monitor this process. Section 6 concludes the paper.

2. THE FUNNEL EXPERIMENT AND DEPENDENT PROCESSES

One of the earliest and most popular examples in the quality control literature is the *funnel experiment*. Deming[1] introduced this example to promote the idea of SPC. The experiment considers a funnel pointed downward and centered above the target, as shown in Figure 1. Marbles that are smaller than the diameter of the funnel's opening are dropped into it in succession. As they fall, they hit near the target, whereas the exact hit locations are random. Henceforth, the consequential falls of marbles around the target is called the 'hit process'. Following Shewhart[29], one can say that the hit process is considered *stable* or *in statistical control* with respect to the distances from the hits to the target.

Deming's example showed that the manipulation of a stable hit process by feedback (compensation) rules often results in a less stable process. For the purpose of illustration, Deming considered the following four compensation rules.

DOI: 10.1002/qre

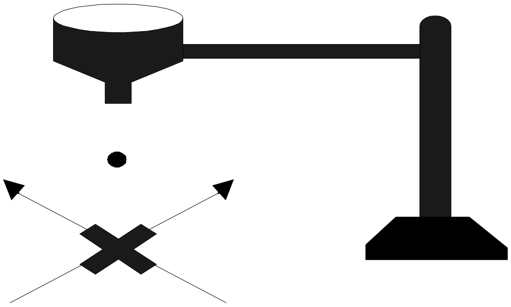

Figure 1. The funnel experiment

1. *No compensation*, leaving the funnel without an adjustment even if consecutive drops failed to hit the target.
2. *Exact compensation*, moving the funnel by a vector negative to the error in the last drop.
3. *Over-compensation*, setting the funnel equal and opposite to the *accumulated* deviation from the target.
4. *Consistency*, centering the funnel over the last observation or drop.

Following Deming[1], the funnel experiment has become a popular subject of several papers that deal with SPC strategies[6,11,27,28,30,31]. Boardman and Boardman[6] analyzed the variance of the *hit processes* subject to Deming's compensation rules. They showed that the hit process of rule 2 is twice as variable as the hit process of rule 1, that rule 4 results in a random walk of the hit process and that the hit process of rule 3 is about three times more spread than the random walk of rule 4. Based on these results, the authors re-emphasized Deming's conclusion that frequent adjustments of a stable process results in an increased variance in comparison to the variance of the unadjusted process. However, these works did not indicate how to implement a SPC scheme for the hit process that is governed by compensation feedback rules.

A complementary recommendation was proposed by MacGregor[32], indicating that the no compensation rule holds only when the process is stationary. MacGregor pointed out that for non-stationary processes, one could profitably use control theory techniques. A subsequent development of the idea of integrating SPC and control theory was proposed by Box and Kramer[7]. Their basic idea was that the process should be adjusted only when an out-of-control signal is observed. Although this idea was studied more than three decades before[33], the work of Box and Kramer inspired many new papers in this area[8–11,34].

It is well known that the integration of SPC with control rules, even simpler feedback rules, often creates statistical dependencies within the process observations. In order to address such cases, many publications assumed that the controlled process follows a known time-series model. This model could then be used to filter the data and obtain state-independent dynamics of the process residuals. MacGregor[32] and Montgomery *et al.*[11] showed that under this assumption one can use conventional SPC charts to identify changes in the process. Note, however, that this method cannot be used if the process does not follow a known time-series model. Moreover, many of the implemented feedback rules generate nonlinear and state-dependent dynamics of the manipulated process. Under these dynamics, it is much harder to identify change points in the process, because both types of statistical errors increase[26,35]. In the next section, an example of such a scenario is considered when using a simple feedback rule to manipulate the funnel experiment. The resulting hit process, which is nonlinear and state dependent, is used as a benchmark process in later sections to compare the MSPC with conventional SPC approaches for autocorrelated processes.

3. A MARKOVIAN HIT PROCESS

Consider an non-intervened stochastic hit process whose hit locations can be modeled by a probability distribution which is centered around the target. Appendix A presents such a discrete probability distribution

DOI: 10.1002/qre

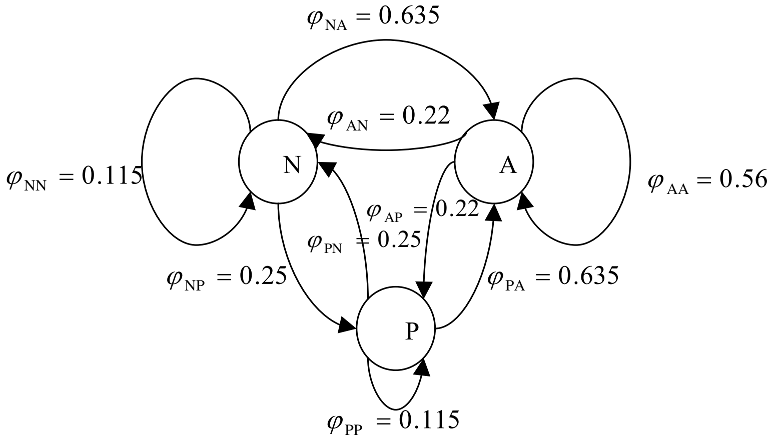

Figure 2. The Markov chain representing a discretized hit process for $q = 0.5$

with a single parameter q. Thus, by selecting different values for q, one can generate different hit processes on which various SPC procedures can be tested.

To this process let us apply a simple feedback rule that can be regarded as a simple extension of rule 2 that was proposed by Deming and uses two observations for compensation instead of a single one. The rule is called the average feedback (AF) rule, because it is based on the average compensation of the last two hits (instead of a single hit). In particular, assume that the errors on both axes are independent, and let us analyze only the x-axis process. Denote by z_t the distance from the target (the errors) on the x-axis at time t for the non-intervened process. Accordingly, the AF rule is implemented by compensating for each observation by the average of the last two hits, resulting in the following controlled hit process:

$$z'_t = \begin{cases} z_t & t = 1, 2 \\ z_t - \dfrac{(z_{t-1} + z_{t-2})}{2} & t = 3, \ldots, n \end{cases} \tag{1}$$

Note that although the AF rule introduces nonlinear dynamics, the mean of the controlled hit process remains centered on the target, i.e. $E(z'_t) = 0$.

As the proposed MSPC monitors discrete processes over a finite alphabet, we discretize the z'_t values to obtain a final symbol set $x_t \in \{\text{P, A, N}\}$, $t = 1, \ldots, n$, denoting, respectively, a Positive deviation from the target ($z'_t > 0.5$), an Accurate hit on the target ($-0.5 \le z'_t \le 0.5$) or a Negative deviation from the target ($z'_t < -0.5$). Appendix A shows that the probabilities for the discretized controlled x_t values depend only on q, and can be easily obtained analytically. Moreover, because the AF rule introduces dependencies between consecutive hits, the resulting discretized hit process can be modeled by a simple Markov chain.

Figure 2 presents the obtained Markov model for the discretized hit process with $q = 0.5$. The model parameters are the transition probabilities $P(x_{t+1}|x_t)$ between consecutive hits $x_t, x_{t+1} \in \{\text{N, A, P}\}$ that are obtained by maximum likelihood estimation methods.

For the purpose of illustration the obtained Markov models in Figure 2 resulted from a somewhat artificial example based on the funnel experiment. Nonetheless, there are many real-world processes that can be represented by Markov models, such as the growth of a population, queuing systems, genetic text, traffic systems, service of machines, chemical reactions, radio-active transformation and manufacturing processes[12,36]. Thus, although we focus here on the funnel experiment for pedagogical reasons, it is noted that the proposed MSPC method can be implemented for many of these Markovian processes.

In the following sections, we use the procedure in Appendix A to generate both 'in-control' and 'out-of-control' sequences. The 'in-control' string, x_t^{in}, $t = 1, \ldots, 5000$ is based on the initial generating process with $q = 0.5$, while the 'out-of-control' string x_t^{out}, $t = 1, \ldots, 5000$, is generated by a modified process

DOI: 10.1002/qre

Table I. Analysis of different ARIMA-based SPC methods for ‘in-control’ data

ARIMA model	n	MSE	UCL $\alpha = 5\%$	Beyond UCL	$\hat{\alpha}$	$\widehat{\text{ARL}}$	Estimated model parameters
ARIMA(1,0,0)	5	0.03	0.151	40	0.4	2.037	$\hat{z}_t = 2.77 - 0.397 z_{t-1}$
ARIMA(0,0,2)	5	0.03	0.152	43	0.43	1.869	$\hat{z}_t = 1.987 + 0.4\varepsilon_{t-1} - 0.077\varepsilon_{t-2}$
ARIMA(1,0,4)	5	0.029	0.149	41	0.41	2.037	$\hat{z}_t = 1.199 + 0.397 z_{t-1} + 0.851\varepsilon_{t-1}$ $- 0.314\varepsilon_{t-2} + 0.096\varepsilon_{t-3} - 0.259\varepsilon_{t-4}$
ARIMA(1,0,0)	10	0.012	0.067	68	0.68	1.441	$\hat{z}_t = 2.159 - 0.085 z_{t-1}$
ARIMA(2,0,0)	10	0.012	0.068	68	0.68	1.441	$\hat{z}_t = 2.345 - 0.091 z_{t-1} - 0.087 z_{t-2}$
ARIMA(1,0,4)	10	0.012	0.068	65	0.65	1.534	$\hat{z}_t = 2.063 - 0.037 z_{t-1} + 0.046\varepsilon_{t-1}$ $+ 0.081\varepsilon_{t-2} - 0.055\varepsilon_{t-3} + 0.082\varepsilon_{t-4}$
ARIMA(1,0,0)	20	0.006	0.033	64	0.64	1.544	$\hat{z}_t = 2.485 - 0.248 z_{t-1}$
ARIMA(1,0,1)	20	0.006	0.033	59	0.59	1.67	$\hat{z}_t = 2.072 - 0.041 z_{t-1} + 0.242\varepsilon_{t-1}$

with $q = 0.8$. These two sequences are used to analyze and compare a battery of conventional SPC methods (Section 4) with the proposed MSPC method (Section 5). Once again, our goal is to identify the exact change point between the ‘in-control’ and the ‘out-of-control’ processes.

4. *USING CONVENTIONAL SPC METHODS*

In this section we apply known SPC methods for both ‘in-control’ and ‘out-of-control’ hit processes that resulted from the applied AF rule. We follow previous studies[12,26] and show that the conventional ARIMA, CUSUM and the EWMA methods are inadequate for monitoring the nonlinear Markovian hit process.

4.1. ARIMA-based SPC charts

The ARIMA family of models is widely applied for the representation and filterization of autocorrelated processes. If an autocorrelated process is well described by an ARIMA model, then a model-based filtering yields independent and approximate Gaussian residuals, to which traditional SPC can be applied. Apley and Shi[15] proposed the generalized likelihood ratio test (GLRT) method that takes advantage of the residual transient dynamics in the ARIMA model when a mean shift is introduced. Friedlander and Porat[37] proposed an algorithm for estimating the moving average and ARMA parameters of non-Gaussian processes from sample high-order moments. Cox[38], Lawrance and Lewis[39] and Benjamin *et al.*[40] extended the Gaussian ARMA time-series models to a non-Gaussian framework. In our case study, the resulting time-series from the AF rule can only attain a finite number of outcomes, therefore all of the control charts that are based on normal distributed data are not expected to perform well. On the other hand, it has been shown[13,15], that simple ARIMA models, such as AR(1) or IMA(1,1), can effectively filter a wide range of autocorrelated processes even if they do not fit the model exactly. Next, we check the applicability of various ARIMA models to the monitoring of the Markovian hit process.

We start by simulating the ‘in-control’ process. Table I presents the best-found ARIMA models, in-terms of their Type I statistical errors and their ‘in-control’ average run length (ARL), as analyzed by the *Statgraphics* software package. The columns of the table are, respectively: the ARIMA model; the subgroup size n (i.e. fitting a model and then applying an $\bar{X}$ chart to a subgroup of n residuals); the mean squared error (MSE); the upper control limit (UCL), which is determined as $Z_{0.975}\sqrt{\text{MSE}/n}$, where $Z_{0.975} = 1.96$ is the 0.975 quantile of the standard normal distribution, appropriate for a type I error of $\alpha = 5\%$; the actual number of runs that fell beyond the UCL; the estimated Type I statistical error, calculated by dividing the number of runs that fell beyond the UCL by 100, the number of analyzed subgroups; the estimated ARL which was computed directly from the control charts; and the estimated ARIMA model. In general, we obtained high values of Type I statistical errors and low ARL values in all the ARIMA charts, including those that are not presented here.

Figure 3 presents, as an example, one control chart for the AR(1) residuals with $n = 5$, i.e. applying the Shewhart $\bar{X}$ chart to a subgroup of n =5 residuals of the AR(1) model. Note that AR(1) is the best found model

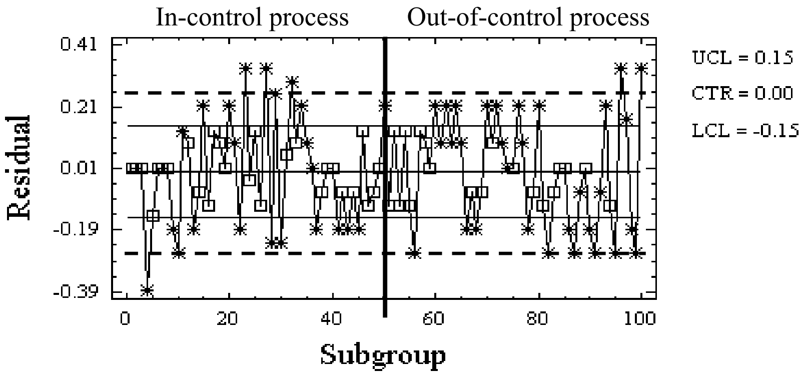

Figure 3. The AR(1) residual control chart (with a sample size of $n = 5$) applied to 250 'in-control' data points and 250 'out-of-control' data points. The limits are based on the 0.975 quantile of the standard normal distribution. The asterisks '*' indicate an 'out-of-control' data point, the square '□' indicate an 'in-control' data point and the dashed line indicates the 99.74 percentile (three standard deviations) control limits

in terms of the estimated ARL and Type I statistical error. The first 50 points in the chart are based on 250 data points generated by the 'in-control' process x_t^{in}, $t = 1, \ldots, 250$, while the last 50 points are based on 250 data points generated by the 'out-of-control' process x_t^{out}, $t = 251, \ldots, 500$. A data point is indicated as 'out-of-control' by an asterisk '*' when it falls beyond the control limits or if it exhibits some non-random pattern of behavior. As can be seen, not only are more than 40% of the first 50 points marked erroneously as 'out-of-control', but also it is impossible to indicate a change point to distinguish between the two processes. The lack of a clear change point is apparent even if we rely on the dashed control limits that are based on three standard deviations (0.9973 quantile). For these limits we obtain a Type I error of 23% and an estimated ARL of 4.32.

High values of statistical errors are also obtained when applying different ARIMA models with various values for the subgroup size and the model parameters. A justified question is whether the performance of the ARIMA models can be improved by increasing the subgroup size. Note from Table II that regardless of the model, an increase of the subgroup size results in a poorer ARL (this was indicated also for much larger sample sizes that are not presented here). The explanation is that a 'better' estimation of the model parameters by a larger sample cannot improve, and in fact often worsens, the ARL performance if the model is 'wrong'.

4.2. *CUSUM- and EWMA-based SPC charts*

The EWMA and CUSUM control charts are widely applied for the detection of small shifts in the process mean. In general, these charts are often much faster than the Shewhart charts at detecting shifts that are equal to or less than two standard deviations with the same sample size[41,42]. Several papers do not recommend the use of the CUSUM and the EWMA charts for general autocorrelated data[43,44]. In contrast, other papers recommend the use of these charts to autocorrelated series, if the data can be properly filtered based on known generating model[21,45]. The latter approach is infeasible when the correlation structure is unknown *a priori*, as in our case study.

Figures 4 and 5 present, as an example, the CUSUM(H–K) and EWMA($\lambda = 0.2$) control charts with a sample size of $n = 5$. Again, the first 50 points in the chart are generated by the 'in-control' process x_t^{in}, $t = 1, \ldots, 250$, while the last 50 points are generated by the 'out-of-control' process x_t^{out}, $t = 251, \ldots, 500$. Not only are all 500 points marked as 'in-control' (with an ARL $\rightarrow \infty$), but it is also impossible to distinguish between the two processes and indicate a change point. The same phenomena are also obtained when applying different EWMA, CUSUM(V-mask) and CUSUM(H–K) models with various values for the subgroup size and the model parameters. The poor results are not surprising because the process mean remains fixed for both the 'in-control' and 'out-of control' stages of the process.

DOI: 10.1002/qre

Table II. The estimated transition probabilities base on a training dataset x_t^{in}, $t = 1, \ldots, 5000$

	x_{t+1}			
x_t	N	A	P	Marginal distribution
N	113 (0.110)	644 (0.625)	273 (0.265)	1030 (0.206)
A	656 (0.223)	1634 (0.557)	644 (0.220)	2934 (0.587)
P	261 (0.252)	656 (0.633)	119 (0.115)	1036 (0.207)

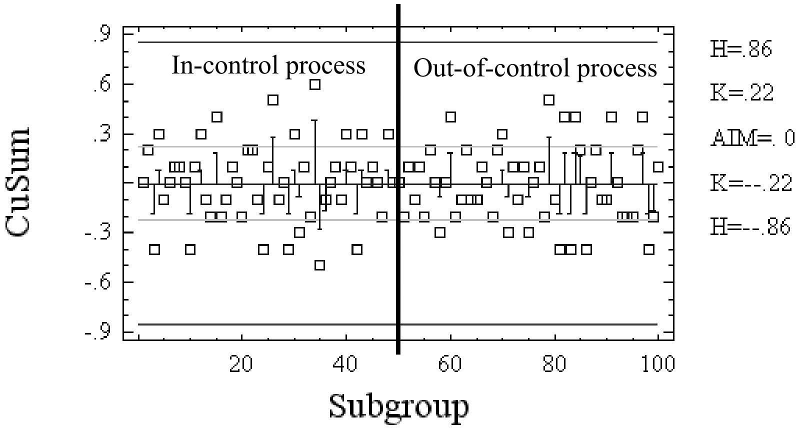

Figure 4. The CUSUM(H–K) control chart (with a sample size of $n = 5$) applied to 250 'in-control' data points and 250 'out-of-control' data points. The limits (H values) are based on the 0.975 quantile of the standard normal distribution. As seen, all points fall within the control limits

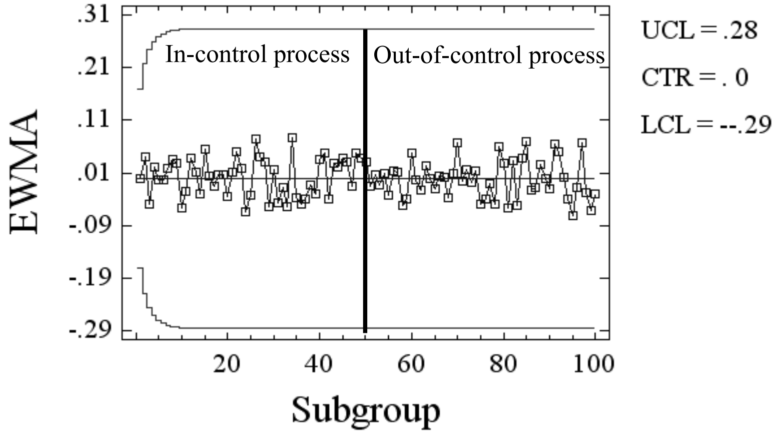

Figure 5. The EWMA control chart with $\lambda = 0.2$ and sample size $n = 5$ applied to 250 'in-control' data points and 250 'out-of-control' data points. The limits are based on 0.975 quantile of the standard normal distribution. As can be seen, all points fall within the control limits

DOI: 10.1002/qre

5. THE MSPC APPROACH

In this section we present the MSPC approach through a running example based on the funnel experiment. In particular, we consider the *hit process* of the funnel experiment that results from the AF rule. The proposed approach assumes a Markovian property of the process; however, it does not require prior knowledge of the order of the Markov model or the values of its transition parameters, which can be estimated numerically by maximum-likelihood estimation methods. In Section 5.1, we assume that the generating hit process is unknown *a priori* and numerically estimate the transition probabilities of the reference Markov model. This numerical procedure is practically important because the analytical generating model is often unknown in reality. In Section 5.2, the MSPC control limits are obtained by simplifying the approach proposed by Ben-Gal *et al.*[26] and Ben-Gal and Singer[12]. Finally, in Section 5.3 we analyze the performance of the MSPC method. In particular, we check that the monitored statistics remain within the control limits as long as the data are generated by the 'in-control' process z_t^{in} $(q = 0.5)$ and, alternatively, that it deviates from these limits once the data are generated by an 'out-of-control' process z_t^{out} $(q = 0.8)$. The procedure for obtaining the dependency structures and the corresponding transition probabilities of these processes is given in Appendix A.

5.1. Estimation of the reference Markov model

In the first stage of the MSPC approach, we construct the reference Markov model that represents the hit process based on a string of 5000 'in-control' data points x_t^{in}, $t = 1, \ldots, 5000$. Apart from assuming a Markovian property of the process, which can be validate by using proper χ^2 tables, the method does not require any prior knowledge such as the type of the AF rule, the transition parameters or the specific order of the model (see Appendix A). A Markov model of order 1 is estimated as the best-found model for the 'in-control' process.

The reference Markov model is constructed numerically using a straightforward maximum-likelihood estimation method. In particular, we count the observed frequencies of symbols in the sequence, denoted by $\hat{n}(x_{t+1}|x_t)$, and use these counts to estimate the relative transition probabilities. For example, the estimated conditional probability of a symbol, say A given N (i.e. the transition probability φ_{NA}), is calculated by the ratio of the number of negative hits (1030 in this case) that are immediately followed by accurate hits in the sequence (644 in this case). Accordingly, the maximum-likelihood estimate $\hat{P}(\text{A}|\text{N}) = 644/1030 = 0.625$. In a similar manner, we estimate the rest of the parameters of the transition probability matrix, which is given in Table II. Columns 2–4 in the table show the counts and the values in the parentheses are the estimated transition probabilities $\hat{P}(x_{t+1}|x_t)$. The last column shows the counts and the estimated steady-state probabilities (marginal distributions) denoted by $\hat{P}(x)$.

The estimated reference model in Table II can now be compared with the unknown analytic Markov model in Figure 2. Note the small dissimilarities in the transition probabilities between the analytical and the numerical models, e.g. $P(\text{A}|\text{N}) = 0.635$ versus $\hat{P}(\text{A}|\text{N}) = 0.625$. However, the estimated model reveals well the symmetric process. For example, the probability of obtaining the sequence $\text{P} \rightarrow \text{N} \rightarrow \text{A}$ (positive error, negative error and an accurate hit) is approximately equal to the probability of obtaining the sequence $\text{N} \rightarrow \text{P} \rightarrow \text{A}$, which is the symmetric sequence replacing symbol P with symbol N and *vice versa*. This result is derived from the respective transition and transient probabilities, as $P(\text{P}) \approx P(\text{N})$ and $P(\text{N}|\text{A}) \approx P(\text{P}|\text{A})$.

Evidently, the nonlinear dependencies in the process that are well represented by the transition probability matrix in Table II cannot be captured by linear correlation measures. The correlation coefficient, which is often implemented in practice to identify autocorrelation in the data, does not reveal any such relation in this case. In fact, had a practitioner used the correlation coefficient as a measure of association, he would have concluded that the data are independent. For lag 1 (dependency between two consecutive observations) the empirical correlation coefficient is 0.0164, for lag 2 the empirical correlation coefficient is 0.0153 and for lag 3 the empirical correlation coefficient is 0.0152. It should not be surprising that the correlation coefficient does not reveal the dependencies that exist in the string, because these are not of a linear form.

Once the reference model has been estimated the SPC procedure can be implemented. The remaining task is to distinguish between insignificant changes that result from the process stochasticity versus significant changes that imply a change in the generating process. The next section addresses this task by deriving the appropriate MSPC control limits.

DOI: 10.1002/qre

5.2. *Calculation of the MSPC control limits*

In this section, we used the well-known Chi-squared statistic to compare the observed frequencies with their respective model-based expected frequencies. The observed frequencies represent the transition probabilities at different monitoring periods. The expected frequencies are either based on a known reference Markov model that represents the 'in-control' behavior of the system or on an estimated reference model that is constructed by using a string of 'in-control' data as seen in Section 5.1. Practically, it is recommended that the string length adheres to the Chi-squared sampling principle as suggested by Cochran[46]. This principle requires that at least 80% of the sampling bins, corresponding in this case to the cells in the transition probabilities matrix, contain at least four data points (further principles for estimating the 'in-control' model can be found in Ben-Gal *et al.*[26]).

Under the null hypothesis that the same underlying Markov process generates the data at different periods, the Pearson's Chi-squared statistic is approximately Chi-squared distributed, i.e.

$$\chi^2 = \sum_{x \in \{\mathrm{N,A,P}\}} \frac{(\hat{n}(x_{t+1}|x_t) - n(x_{t+1}|x_t))^2}{n(x_{t+1}|x_t)} \leqslant \chi^2_{d(d-1),\alpha} \tag{2}$$

where $\hat{n}(x_{t+1}|x_t)$ are the observed frequencies, $n(x_{t+1}|x_t)$ are the expected frequencies, d is the size of the symbol set (in this case $d=3$) and α is the Type-I error value determining the confidence level, which is commonly equal to 0.95 ($\alpha=0.05$). Thus, the UCL is obtained by the 95th percentile of the Chi-squared distribution and depends only on the alphabet size. In this case UCL is equal to $\chi^2_{d(d-1),\alpha} = \chi^2_{6,0.05} = 12.59$, while the LCL, which represents a total identity between the observed and the expected frequencies, is equal to zero (a single-sided control chart).

The value of the Chi-squared statistic between the observed frequencies in Table II and the expected frequencies (calculated by multiplying the transitions probability parameters in Figure 2 by a sample size of 5000) is equal to $\chi^2 = 3.31$. As it is much smaller than the UCL it ensures that there is no significant difference between the observed and the expected frequencies or, in other words, that the observed frequencies in Table II represents the reference model in Figure 2 well.

Once the upper bound is established, the SPC scheme is obtained by plotting the Pearson's Chi-square statistics on the control chart with respect to the UCL. Following the minimum discrimination information principle[14,47], at each monitoring period we use a new sequence to construct a frequency table in a similar manner to Table II and computes the new Chi-squared statistics.

5.3. *Identifying a change point by the MSPC method*

In order to complete the numerical study, a set of 100 sequences of the reference *hit process* is generated. Each sequence consists of 5000 data points that are generated by implementing the numerical procedure described in Appendix A based on $q=0.5$. For the purposes of illustration, another set of 100 'out-of-control' sequences is generated by repeating the same numerical procedure based on a modified generating parameter $q=0.8$. The Chi-squared statistics between the monitored models and the reference model are computed by using Equation (2) for both the 'in-control' and 'out-of-control' samples. All of the Chi-squared statistics are then plotted against the UCL.

Figure 6 presents the Chi-squared statistics of the 'in-control' samples, while Figure 7 presents Chi-squared statistics of the 'out-of-control' samples (the processes are not plotted in the same graph due to their large-scale differences). Note that most of the Chi-squared statistics in the first 100 samples fall within the control limits. As expected from the 95% confidence level, 4 out of 100 runs are erroneously indicated as 'out-of-control' runs, resulting in an estimated ARL of 17. The Chi-squared statistics of the second set of sequences are distinctively confirmed to be generated from an 'out-of-control' process. In summary, note that the proposed monitoring approach performs well with respect to both statistical errors: 96% of the 'in-control' runs adhere to the control limit rule, while 100% of the 'out-of-control' runs are clearly identified to be generated from a modified process. The change point is clearly evident by a sharp jump in the Chi-squared values approximately from 4 to 1900.

DOI: 10.1002/qre

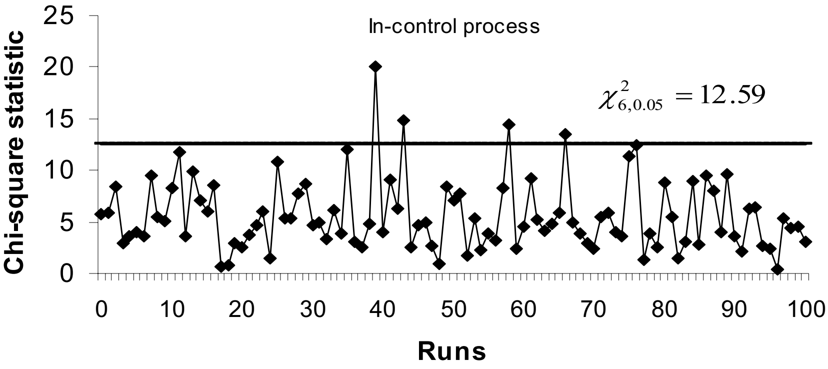

Figure 6. Control chart of the CSPC for 'in-control' sample data

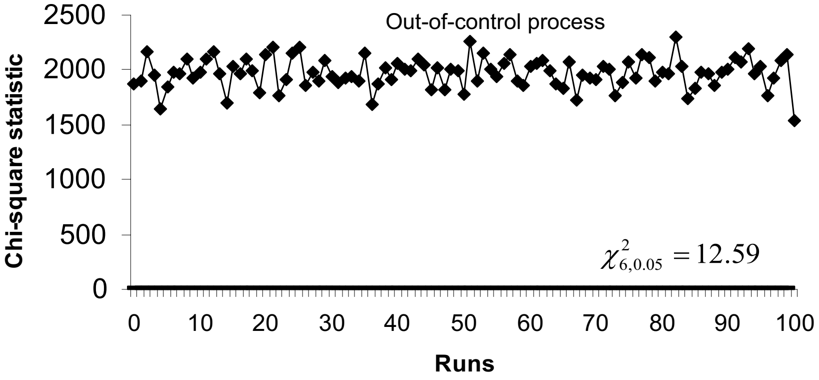

Figure 7. Control chart of the CSPC for 'out-of-control' sample data

6. CONCLUSIONS

In his classical funnel experiment, Deming[1] used four feedback rules to illustrate the difficulties in implementing SPC when the data points are dependent. However, the funnel example did not show how to handle the autocorrelated data that result from the used feedback rules. We find the classic framework of the funnel experiment to be practically convenient to illustrate the effects of complex dynamics that are created by feedback rules, and to promote new monitoring methodologies for such cases. In particular, we have applied a simple 'fifth' feedback rule to the funnel example, called the AF rule, which extends Deming's second rule by using two observations for compensation instead of a single one. The discretized values of the AF rule follow a nonlinear Markovian hit process to which traditional SPC methods cannot be applied successfully. We have shown that the proposed MSPC procedure enables to model the dependence in the controlled hit process and accurately monitor it. Moreover, the principles of the MSPC are simpler with respect to the previously proposed CSPC method and rely on commonly used statistical principles.

A main disadvantage of the MSPC method is the relatively large amount of discrete data that it requires for the monitoring, restricting its online analysis capabilities. Further research is required to optimize the MSPC procedure to various types of data. As suggested by the anonymous referee, one possible research direction is to look at the literature of wavelet-based control charts, which also involve hidden trees[48].

DOI: 10.1002/qre

Acknowledgements

We thank the anonymous referee for his careful reading of this paper, and for making a number of useful suggestions that have led to an improved manuscript.

REFERENCES

1. Deming WE. *Out of the Crisis*. MIT Center for Advanced Engineering Study: Cambridge, MA, 1986; 327–331.
2. Gerold D, Hershey R, McBrayer K, Sturtevant J. Run-to-run control benefits to photolithography. *Proceedings of the 9th Sematech AEC/APC Workshop*, Incline Village, NV, September 1997.
3. Bonning D, Moyne W, Smith T. Run by run control of chemical-mechanical polishing. *Proceedings of the IEEE CPMT International Electronics Manufacturing Technology Symposium*, 1995.
4. Rosenthal R, Solomon P, Charpenay S, Bonanno A, Zhang W, Eiklebarry W. Run-to-run control of a single wafer epitaxial silicon fabrication process. *Proceedings of the Sematech AEC/APC Symposium*, Incline Village, NV, September 1997.
5. Shore H. *Total Quality, Quality Control and Quality by Design*, 1992 (in Hebrew).
6. Boardman TJ, Boardman EC. Don't touch that funnel. *Quality Progress* 1990; **23**:65–69.
7. Box GEP, Kramer T. Statistical process monitoring and feedback adjustment—a discussion. *Technometrics* 1992; **34**:251–267.
8. English JR, Martin T, Yaz E, Elsayed E. Change point detection and control using statistical process control and automatic process control. *Proceedings of IIE Annual Conference*, Dallas, TX, 2001.
9. Del Castillo E. *Statistical Process Adjustment for Quality Control*. Wiley: New York, 2002.
10. Box GEP, Luceno A. *Statistical Control by Monitoring and Feedback Adjustment*. Wiley: New York, 1997.
11. Montgomery D, Keats JB, Runger GC, Massina WS. Integrating statistical process control and engineering process control. *Journal of Quality Technology* 1994; **26**:79–87.
12. Ben-Gal I, Singer G. SPC via context modeling of finite state processes: An application to production monitoring. *IIE Transactions on Quality and Reliability* 2004; **36**:401–415.
13. Box GEP, Jenkins GM. *Times Series Analysis, Forecasting and Control*. Holden Day: Oakland, CA, 1976.
14. Alwan LC, Roberts HV. Time-series modeling for statistical process control. *Journal of Business and Economic Statistics* 1988; **6**:87–95.
15. Apley DW, Shi J. The GRLT for statistical process control of autocorrelated processes. *IIE Transactions* 1999; **31**:1123–1134.
16. Jiang W, Wu H, Tsung F, Nair VN, Tsui KL. Proportional integral derivative control charts for process monitoring. *Technometrics* 2002; **44**:205–214.
17. Montgomery DC, Mastrangelo CM. Some statistical process control methods for autocorrelated data. *Journal of Quality Technology* 1991; **23**:179–204.
18. Tseng S, Adams BM. Monitoring autocorrelated processes with an exponentially weighted moving average forecast. *Journal of Statistical Computation and Simulation* 1994; **50**:187–195.
19. Lu CW, Reynolds MR. EWMA control charts for monitoring the mean of autocorrelated processes. *Journal of Quality Technology* 1999; **31**:166–188.
20. Jiang W, Tsui KL, Woodall W. A new SPC monitoring method: The ARMA chart. *Technometrics* 2000; **42**:399–410.
21. Testik MC. Model inadequacy and residuals control charts for autocorrelated processes. *Quality and Reliability Engineering International* 2005; **21**:115–130. DOI: 10.1002/qre.611.
22. Runger GC, Willemain TR, Prabhu S. Average run lengths for CUSUM control charts applied to residuals. *Communication in Statistics—Theory and Methods* 1995; **24**:273–282.
23. VanBrackle LN, Reynolds MR. EWMA and CUSUM control charts in the presence of correlation. *Commnications in Statistics—Simulation and Computation* 1997; **26**:979–1008.
24. Timmer DH, Pignatiello J, Longnecker M. The development and evaluation of CUSUM-based control charts for an AR(1) process. *IIE Transactions* 1998; **30**:525–534.
25. Lu CW, Reynolds MR. CuSum charts for monitoring an autocorrelated process. *Journal of Quality Technology* 2001; **33**:316–334.
26. Ben-Gal I, Morag G, Shmilovici A. CSPC: A monitoring procedure for state dependent processes. *Technometrics* 2003; **45**:293–311.
27. Coleman DW. Adapting Deming's funnel experiment to a content-specific area. *Simulation and Gaming* 1999; **30**(1):8–19.

DOI: 10.1002/qre

28. Stepanovich PL. Using system dynamics to illustrate Deming's system of profound knowledge. *Total Quality Management and Business Excellence* 2004; **15**(3):379–389.
29. Shewhart WA. Finding causes of quality variations. *Manufacturing Industries* 1926; **11**:125–128.
30. Del Castillo E. A note on two process adjustment models. *Quality and Reliability Engineering International* 1986; **14**(1):23–28.
31. Henderson R. EWMA and industrial applications to feedback adjustment and control. *Journal of Applied Statistics* 2001; **28**(3-4):399–407.
32. MacGregor JF. A different view of the funnel experiment. *Journal of Quality Technology* 1990; **22**:255–259.
33. Barnard GA. Control charts and stochastic processes. *Journal of the Royal Statistical Society. Series B* 1959; **21**(2):239–271.
34. Tsung F, Zhao Y, Xiang L, Jiang W. Improved design of proportional integral derivative charts. *Journal of Quality Technology* 2006; **38**(1):31–44.
35. Thomas WN, Lloyd PP. Understanding variation. *Quality Progress* 1990; **23**:70–78.
36. Bharucha-Reid AT. *Elements of the Theory of Markov Processes and Their Applications*. Dover: Mineola, NY, 1997.
37. Friedlander B, Porat B. Asymptotically optimal estimation of MA and ARMA parameters of non-Gaussian processes from higher-order moments. *IEEE Transactions on Automatic Control* 1990; **35**:27–35.
38. Cox DR. Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics* 1981; **8**:93–115.
39. Lawrance AJ, Lewis PAW. A new autoregressive time series model in exponential variables [NEAR(1)]. *Advances in Applied Probabability* 1981; **13**:826–845.
40. Benjamin M, Rigby RA, Stasinopoulos DM. Generalized autoregressive moving average models. *Journal of the American Statistical Association* 2003; **98**:214–223.
41. Hinkley DV. Inference about the change-point from cumulative sum tests. *Biometrika* 1971; **58**(3):509–523.
42. Pettitt AN. A simple cumulative sum type statistic for the change-point problem with zero-one observations. *Biometrika* 1980; **67**(1):79–84.
43. Johnson RA, Bagshaw M. The effect of serial correlation on the performance of CUSUM test. *Technometrics* 1974; **16**:103–112.
44. Winkel P, Zhang NF. Serial correlation of quality control data—on the use of proper control charts. *Scandinavian Journal of Clinical and Laboratory Investigation* 2004; **64**(3):194–204.
45. Yashchin E. Performance of CUSUM control schemes for serially correlated observations. *Technometrics* 1993; **35**(1):37–52.
46. Cochran WG. The chi-square test of goodness of fit. *The Annals of Mathematical Statistic* 1952; **23**:315–345.
47. Kullback S. *Information Theory and Statistics*. Wiley: New York, 1959.
48. Chen J, Chang WJ. Applying wavelet-based hidden Markov tree to enhancing performance of process monitoring. *Chemical Engineering Science*, 2005; **60**(18):5129–5143.

APPENDIX A

In the following we present the procedure for the generation of the hit process resulting from the AF rule. Then we present the procedure to construct the analytical reference Markov model of this process.

I. Generating a sequence of uniformly distributed values

Generate a string of n data points from a uniform distribution $u_t \in [0, 1]$, $t = 1, 2, \ldots, n$.

II. Discretizing the string

Obtain a string of n data points $z_t \in \{-1, 0, 1\}$, $t = 1, \ldots, n$, by using an inverse transform method based on a discrete and symmetric distribution, which is fully determined by a single parameter q,

$$z_t(q) = \begin{cases} -1 & \text{if } u_t \leqslant q/2 \\ 0 & \text{if } q/2 < u_t \leqslant 1 - q/2 \\ +1 & \text{if } 1 - q/2 < u_t \leqslant 1 \end{cases} \tag{A1}$$

DOI: 10.1002/qre

Table AI. Numerical generation of a controlled hit process with 10 data points

Number	Step I: u_t Uniformly random	Step II: $z_t\,(q=0.5)$ Inverse transform	Step III: z_t' Apply the AF rule	Step IV: x_t Discretization
1	0.620	0	0	A
2	0.828	1	1	P
3	0.716	0	−0.5	A
4	0.218	−1	−1.5	N
5	0.202	−1	−0.5	A
6	0.725	0	1	P
7	0.530	0	0.5	A
8	0.244	−1	−1	N
9	0.269	0	0.5	A
10	0.749	0	0.5	A

where $0 \leqslant q \leqslant 1$ and $E(z_t)=0$. The z_t points represent the distance from the target (the errors) on the x-axis at time t for the ‘no compensation’ rule (rule 1). Assume that the errors on both axes are independent, and thus focus on the x-axis process.

III. Implementing the AF rule

Implement the AF rule by reducing from each observation the *average* of the last two hits. The controlled hit process is given by

$$z_t' = \begin{cases} z_t & t = 1, 2 \\ z_t - \dfrac{(z_{t-1}+z_{t-2})}{2} & t = 3, \ldots, n \end{cases} \tag{A2}$$

Note that although certain dynamics are introduced to the z_t process by the AF rule, the mean of the controlled hit process remains centered on the target, i.e. $E(z_t')=0$.

IV. Discretizing the controlled hit process

Discretize the z_t' values to obtain a final symbol set $x_t \in \{\text{P, A, N}\},\ t=1,\ldots,n$, representing either a Positive deviation from the target, an Accurate hit on the target or a Negative deviation from the target, respectively, where

$$x_t = \begin{cases} \text{N} & \text{if } z_t' < -0.5 \text{ with probability} \frac{1}{8}(q^3-2q^2+4q) \\ \text{A} & \text{if } -0.5 \leqslant z_t' \leqslant 0.5 \text{ with probability } -\frac{1}{4}(q^3-2q^2+4q-4) \\ \text{P} & \text{if } z_t' > 0.5 \text{ with probability } \frac{1}{8}(q^3-2q^2+4q) \end{cases} \tag{A3}$$

The probabilities for the discretized controlled x_t values depend only on q (the parameter of the generating process) and can be obtained by summing up the relevant manipulated outputs. For example, the probability of N is equal to $P(\text{N}) = P(z_t'=-2)+P(z_t'=-1.5)+P(z_t'=-1)$. Table AI exemplifies the generation of n= 10 data points for a selected value of $q=0.5$.

Generating a representative Markov chain model

Let us illustrate a three-step procedure for the construction of the analytic Markov model which is given in Figure 2 for the process defined by $q=0.5$. First, we obtain the transition probabilities between tuples of three uncontrolled hits, $(z_{t-2}, z_{t-1}, z_t) \rightarrow (z_{t-1}, z_t, z_{t+1})$. These tuples are defined as the states of the uncontrolled hit process. Second, we calculate the steady-state probabilities of the controlled and discretized hit process, i.e. $P(x_t = \mathrm{P})$, $P(x_t = \mathrm{A})$ and $P(x_t = \mathrm{N})$. Finally, we calculate the transition probabilities between consecutive discretized hits (N, A, P) that are shown in Figure 2.

Let us consider the feasible states of the uncontrolled hit process. Each state is characterized by three consequent observations, (z_{t-2}, z_{t-1}, z_t). This uncontrolled hit process has $3^3 = 27$ different states that are denoted, respectively, by s_i, $i = 1, \ldots, 27$, where $s_1 = (-1, -1, -1)$, $s_2 = (-1, -1, 0)$, $s_3 = (-1, -1, 1)$, $s_4 = (-1, 0, -1)$, $s_5 = (-1, 0, 0)$, $s_6 = (-1, 0, 1)$, $s_7 = (-1, 1, -1)$, $s_8 = (-1, 1, 0)$, $s_9 = (-1, 1, 1)$, $s_{10} = (0, -1, -1), \ldots, s_{27} = (1, 1, 1)$. An example for a feasible transition between the tuples $(z_{t-2}, z_{t-1}, z_t) \rightarrow (z_{t-1}, z_t, z_{t+1})$ is $(-1, -1, -1) \rightarrow (-1, -1, 1)$, i.e. observing a positive error at time $t+1$. The transition probability equations are as follows:

$$\begin{aligned} 0.25P_i + 0.5P_{i+9} + 0.25P_{i+18} &= P_{3i-2} \quad i = 1, \ldots, 9 \\ 0.5P_i + 0.5P_{i+9} + 0.25P_{i+18} &= P_{3i-1} \quad i = 1, \ldots, 9 \\ 0.25P_i + 0.5P_{i+9} + 0.25P_{i+18} &= P_{3i} \quad i = 1, \ldots, 9 \end{aligned} \tag{A4}$$

where P_i denotes the stationary probability of the state s_i.

Solving the above set of transition probability equations, one can calculate the stationary probabilities of the uncontrolled hit process: $P_1 = P_3 = P_7 = P_9 = P_{19} = P_{21} = P_{25} = P_{27} = \frac{1}{64}$; $P_2 = P_4 = P_6 = P_8 = P_{10} = P_{12} = P_{16} = P_{18} = P_{20} = P_{22} = P_{24} = P_{26} = \frac{1}{32}$; $P_5 = P_{11} = P_{13} = P_{15} = P_{17} = P_{23} = \frac{1}{16}$; $P_{14} = \frac{1}{8}$. In order to obtain the stationary probabilities of Negative, Accurate and Positive hits, one has to sum up the appropriate probabilities based on (A4):

$$\begin{aligned} P(\mathrm{N}) &= P(z'_t = -2) + P(z'_t = -1.5) + P(z'_t = -1) = P_{25} + (P_{16} + P_{22}) + (P_7 + P_{13} + P_{19} + P_{26}) = 0.2 \\ P(\mathrm{P}) &= P(z'_t = 2) + P(z'_t = 1.5) + P(z'_t = 1) = P_{19} + (P_{12} + P_6) + (P_9 + P_2 + P_{15} + P_{21}) = 0.2 \\ P(\mathrm{A}) &= P(z'_t = -0.5) + P(z'_t = 0) + P(z'_t = 0.5) = (P_4 + P_{10} + P_{17} + P_{23}) \\ &\quad + (P_1 + P_8 + P_{14} + P_{20} + P_{27}) + (P_5 + P_{11} + P_{18} + P_{23}) = 0.6 \end{aligned}$$

Now one can estimate the expected frequencies of each symbol, denoted by $n(\cdot)$, in a sequence of 5000 observations:

$$(n(\mathrm{N}),\ n(\mathrm{A}),\ n(\mathrm{P})) = (5000 \times 0.2,\ 5000 \times 0.6,\ 5000 \times 0.2) = (1016,\ 2968,\ 1016)$$

Transition probabilities between hits (N, A, P)

Following the above steps, one can calculate the transition probabilities between the controlled and the discretized hit process. For example, the probability of observing two consequent Negative hits is calculated as follows:

$$\begin{aligned} P(\mathrm{N}|\mathrm{N}) = {}& \frac{P_7}{P(\mathrm{N})} \cdot P(\mathrm{N}|s_7) + \frac{P_{13}}{P(\mathrm{N})} \cdot P(\mathrm{N}|s_{13}) + \frac{P_{16}}{P(\mathrm{N})} \cdot P(\mathrm{N}|s_{16}) + \frac{P_{19}}{P(\mathrm{N})} \cdot P(\mathrm{N}|s_{19}) + \frac{P_{22}}{P(\mathrm{N})} \cdot P(\mathrm{N}|s_{22}) \\ & + \frac{P_{25}}{P(\mathrm{N})} \cdot P(\mathrm{N}|s_{25}) + \frac{P_{26}}{P(\mathrm{N})} \cdot P(\mathrm{N}|s_{26}) = 0.115 \end{aligned}$$

In a similar manner, one can find the other transition probabilities, for example

$$P(\mathrm{A}|\mathrm{N}) = 0.635, \quad P(\mathrm{P}|\mathrm{N}) = 0.25, \quad P(\mathrm{N}|\mathrm{P}) = 0.25, \quad P(\mathrm{A}|\mathrm{P}) = 0.635, \quad P(\mathrm{P}|\mathrm{P}) = 0.115$$

The corresponding expected frequencies in a sequence with 5000 observations are

$$n(\mathrm{N}|\mathrm{N}) = 254, \quad n(\mathrm{A}|\mathrm{N}) = 645, \quad n(\mathrm{P}|\mathrm{N}) = 117, \quad n(\mathrm{N}|\mathrm{P}) = 117, \quad n(\mathrm{A}|\mathrm{P}) = 645, \quad n(\mathrm{P}|\mathrm{P}) = 254$$

Authors' biographies

Gonen Singer holds PhD, MSc and BSc degrees in Industrial Engineering from Tel-Aviv University. For several years he has taught in the Open University and in Tel-Aviv University. His research interests include simulation, quality control, stochastic processes and graphical user interfaces. Nowadays he is working as a project manager and Information Systems Analyst in the Israeli Air Force.

Irad Ben-Gal is a Senior Lecturer at Tel-Aviv University. He holds a BSc (1992) degree from Tel-Aviv University, and MSc (1996) and PhD (1998) degrees from Boston University. He is a member of the Institute for Operations Research and Management Sciences (INFORMS) and the Institute of Industrial Engineers (IIE), and on the editorial board of Trends in Applied Sciences Research. His papers have been published in *IIE Transactions, International Journal of Production Research, Technometrics, IEEE Transaction, Bioinformatics* and other journals. He has received several research grants, among them from General Motors, IEEE, the Israeli Ministry of Science and the European Community. He has worked for several years in industrial organizations. His research interests include statistical methods for control and analysis of stochastic processes, applications of information theory to industrial problems, and automation and computer integrated manufacturing systems.

DOI: 10.1002/qre