A model-free Shewhart individuals control chart for autocorrelated data

Xi Ma*,***, , Laibin Zhang*, Jinqiu Hu*, Ahmet Palazoglu***

* College of Mechanical and Transportation Engineering, Laboratory of Oil and Gas Production Safety and Accident Prevention and Control, China University of Petroleum, Beijing, 102249, China.(e-mail: seven7xi@126.com).
*** Department of Chemical Engineering, UC Davis,
Davis, CA USA, (Tel: (530) 752-8774, e-mail: anpalazoglu@ucdavis.edu)

Abstract: When data are collected sequentially from a chemical process, consecutive observations are correlated resulting in serial dependence. Such dependence (or autocorrelation) would violate the assumption of sample independence when carrying out most statistical process control schemes, such as Shewhart charts for individual measurements. In this paper, a model-free Shewhart individuals control chart for autocorrelated data is proposed to reduce/eliminate the effect of autocorrelation on chart performance. The *modified* Shewhart chart, based on the true mean and variance, is used as the benchmark chart for comparison. A single skipping chart (SSC) and a combined skipping chart (CSC) are established in the proposed Shewhart control scheme. The control performances of CSC, *modified* Shewhart chart and conventional Shewhart chart are compared in terms of their mean shift detection ability using an AR (1) process. The advantages of CSC are illustrated as a model-free approach and having a performance consistent with that of the benchmark, *modified* Shewhart chart. A practical application of CSC is illustrated using data from an industrial chemical process.

Keywords: Shewhart chart; Autocorrelation; Skipping strategy; Control performance; Average run length; AR(1) process; Industrial application.

1. INTRODUCTION

In chemical process industries, Statistical Process Control (SPC) is widely used to implement online process monitoring strategies that ensure quality control over key process variables. The primary objective of SPC is to reduce process variability, which can manifest itself in the individual observations of quality characteristics (Montgomery, 2007). Traditional control charts for a single quality characteristic (i.e., the monitored process variable) such as the Shewhart, cumulative sum (CUSUM) and exponential weighted moving average (EWMA) charts, are often implemented based on the assumption that measured data are identically and independently distributed (iid). In practice, however, variable measurements are collected automatically at high sampling rates and their consecutive values are serially correlated, invalidating the independence assumption. Under these circumstances, the traditional control charts would find it difficult to distinguish between the common cause and the special cause **(**Alwan & Roberts, 1988**)**, producing a large number of false alarms and compromising effectiveness of the charts for fault detection **(**Lin et al., 2012; Reynolds & Lu, 1997**)**. The detection capability and run length distributions of traditional control charts for autocorrelated data sets have been discussed in previous studies (Alwan, 1991; Prybutok et al., 1997; Knoth & Schmid, 2004; Padgett et al., 1992; Wardell et al., 1994; Schmid & Schone, 1997; Harris & Ross, 1991; Johnson & Bagshaw, 1974). A common conclusion is that the traditional control limits established for independent data are no longer suitable when data is serially correlated.

As individual measurements are common in chemical process systems, individuals control charts are more practical than sample control charts. This paper focuses on a model-free control scheme based on the Shewhart individuals chart (Montgomery, 2007) to improve shift-detection ability when data is serially correlated. The remainder of the paper is organized as follows: We introduce the time series models, especially autoregressive (AR) models, briefly in Section 2. The conventional Shewhart and the *modified* Shewhart chart are reviewed in Section 3. A model-free Shewhart control scheme is proposed to accommodate autocorrelation in the data using a skipping strategy. The methodology is outlined in Section 4 and two control charts, i.e., combined skipping chart (CSC) and single skipping chart (SSC), are designed. In the following section, the performance of the proposed Shewhart control chart is illustrated using simulated AR (1) process data and a practical application is discussed using an industrial chemical process. Finally, conclusions are drawn and possible future directions are outlined.

2. AUTOCORRELATED PROCESS DATA

Traditional control charts operate with the assumption that data in statistical control is collected from a random process, which means that the observations of quality characteristics are serially independent. An in-control observation x_t from a random process can be represented as $x_t = \mu + \varepsilon_t$, where μ is a constant and ε_t is a sequence of random errors with mean 0 and variance σ_0^2. Time series from this model have a random variation around a constant mean, μ. This assumption, however, is rarely met in practical industrial processes and the observed process variables often exhibit serial dependence, a.k.a. autocorrelation. Due to system

dynamics and sampling frequency, it is reasonable that a real process is autoregressive, i.e., an observation of a process variable is correlated with previous observation(s) (Pan and Jarrett, 2012). An autoregressive model can then be used as a representation of a certain set of actual observed series (Box et al., 2015, Montgomery et al., 2015). If an observation at time t is only correlated with the last observation at time t-1, data can be represented as a "first-order autoregressive model" (denoted as AR (1)). An AR (1) time series can be generated from the following model: $x_t = \mu + \phi x_{t-1} + \varepsilon_t$. Here, ϕ , autocorrelation coefficient, can be interpreted as the carryover effect of the last observation on the current observation.

Autocorrelation function (ACF) (Box et al., 2015, Bisgaard and Kulahci (2011)) is used in this work to identify the "amount" of autocorrelation in a time series. The autocorrelation function, i.e., the correlation between x_t and x_{t+k} of a stationary process is defined as,

$$\rho_k = \frac{E\left(x_t - \mu\right)\left(x_{t+k} - \mu\right)}{\sigma^2} \ .$$

where μ and σ^2 are the mean and variance of the stationary time series x_t . Similarly the sample ACF is given as,

$$\hat{\rho}_k = \frac{\sum\limits_{t=k+1}^{n}\left(x_{t-k} - \overline{x}\right)\left(x_t - \overline{x}\right)}{\sum\limits_{t=1}^{n}\left(x_t - \overline{x}\right)^2}$$

where $\overline{x}$ is the sample mean of the stationary time series, x_t .

3. SHEWHART INDIVIDUALS CONTROL CHART

3.1 Conventional Shewhart Individuals Chart

The Shewhart chart for individual observations is commonly used in process industries. After estimating the process mean and standard deviation, the control limits of Shewhart individuals chart are calculated with an expected false alarm rate. The lower control limit is $LCL = \overline{x} - c \cdot S$ and upper control limit is $UCL = \overline{x} + c \cdot S$, where $\overline{x}$ and S are the estimated mean and standard deviation, respectively. When $c = 3$, 3-sigma limit is used, which corresponds to a false alarm rate of 0.0027 under the normality assumption.

As mentioned by Li (2004), the Shewhart individuals chart is robust with respect to serial dependence when the sample standard deviation is calculated from a long series of observations. This was also shown by Alwan (1991) who tested the influence of data autocorrelation on the Shewhart individuals chart and using 3-sigma control limits, he noted minimal false alarms in the chart for autocorrelated data. Apart from using the false alarm rate and the run lengths for the whole series, run rules are also often used in conjunction

with the Shewhart control chart to identify abnormality of short sequences. The commonly recommended one, the Western Electric rules (Western Electric Company, 1956) have proven to be ineffective for non-random data, with increasing false alarm rates. As a chart performance measure, run length is defined as the number of samples before the chart gives the first signal. The average run length (ARL) under normal process conditions is denoted as ARL_0, while that under abnormal process operation is represented by ARL_1. Knoth and Schmid (2004) compared the ARLs of Shewhart, EWMA and CUSUM charts for individual observations. They found that the Shewhart individuals chart is more robust to low and moderate autocorrelation in data than the other two charts in terms of ARL. As revealed in the above tests on individuals Shewhart chart, data autocorrelation does not theoretically influence the false alarm rate and the run length of signal is less likely to be highly influenced by low to moderate autocorrelation. However, in practice, one does not use a long series of measurements to estimate process mean and standard deviation if data autocorrelation is ignored. In that case, the control limits calculated may lead to considerable false alarms and make control chart misleading, when process data is highly autocorrelated.

3.2 Modified Shewhart Chart

If the process model is known or can be estimated, *modified* control limits using a model-based mean and standard deviation can be used. The *modified* chart for autocorrelated data is expected to have a larger or equal ARL_0 compared to the conventional chart for independent data, when they use the same critical value c (Knoth and Schmid, 2004).

For an AR(1) process $x_t = \mu + \phi x_{t-1} + \varepsilon_t$, the true mean $\mu / (1-\phi)$ and standard deviation $\sigma_0 / \sqrt{1-\phi^2}$ are used in the *modified* Shewhart chart, where σ_0^2 is the variance of error ε_t . It should be noted that, in reality, the *modified* chart will suffer from estimation issues caused by number of observations and also model mismatch.

4. A MODEL-FREE SHEWHART INDIVIDUALS CONTROL CHART FOR AUTOCORRELATED DATA

To guarantee the performance of control charts when autocorrelation in data cannot be ignored, a new model-free approach to running the Shewhart individuals control chart is proposed that takes advantage of an iterative skipping strategy. This method divides the data set into several subgroups by considering every other R observations, thereby assuring the assumption of independence in each subgroup. Control charts are then designed for each subgroup in Phase I. A new observation in the phase of online monitoring, i.e., Phase II, is projected onto its corresponding chart according to the skipping time lag R.

4.1 Phase I

A historical time series under normal operating conditions with a specific length is chosen to determine the control limits of the chart.

Step 1: Choice of skipping lag R. ACF limit is defined as an autocorrelation level where the influence of autocorrelation on the control performance is minimal. Time lag R is found as the first lag where the value of ACF is below the limit.

Step 2: Creation of subgroups. The Phase-I time series is skipped by R time points (lags) to create R subgroups. For example, if 100 observations are used in Phase I and R is 5, one would obtain 5 subgroups as shown in **Table 1**.

Step 3: Determination of control limits. R sets of control limits are calculated using the traditional limit formulas, after critical parameter c is predetermined. The value of c is directly related to the expected percentage of false alarms.

Table 1. Illustration of Step 2.

	Sub 1	Sub 2	Sub 3	Sub 4	Sub 5
Phase I	t-4	t-3	t-2	t-1	t
	X_1	X_2	X_3	X_4	X_5
	X_6	X_7	X_8	X_9	X_{10}
	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
	…	…	…	…	…
	X_{96}	X_{97}	X_{98}	X_{99}	X_{100}
Phase II	X_{101}	X_{102}	X_{103}	X_{104}	X_{105}
	X_{106}	X_{107}	X_{108}	X_{109}	X_{110}

4.2 Phase II

After the range of common cause is estimated in Phase I through R sets of control limits of the monitored variable, the new observations are then monitored using the control chart.

To maintain the iid assumption in each subsample, new observations are monitored in their corresponding charts. As shown in Table 1, if Phase II begins at the 101th observation, the first observation would be projected onto the control chart of the first subgroup and the second one would be onto the chart of the second subgroup and so on. The control chart of each subgroup is hereby named as the single skipping chart (SSC), and the combined skipping chart (CSC) is defined by merging the R control charts in time sequence to monitor Phase-II data points. This will be illustrated in the following section.

Two charts are established in the proposed Shewhart individuals control scheme. SSC can be considered as a conventional Shewhart individuals chart applied to observations collected at a low sampling rate. CSC is an improved Shewhart chart, monitoring every observation collected at a high sampling rate. The analyst can alternate between these two charts according to the process conditions. When the process runs at steady-state, SSC would be preferred to monitor the process less frequently. As soon as SSC detects an abnormal trend, SSC should be switched to CSC to pay close attention to the changing process

conditions. In this way, one can filter out redundant information and meaningful structural information, such as the variable trends, would be more likely to be noticed by operators.

5. CASE STUDIES

We denote the subgroup size as L and the number of subgroups as R in the skipping Shewhart control scheme. Two conventional Shewhart charts are constructed based on different length of Phase-I data. In the following case studies, conventional Shewhart chart (a) refers to the conventional chart based on a consecutive Phase-I data of length L, while conventional Shewhart chart (b) is based on a consecutive Phase-I data of length $L*R$. The *modified* Shewhart chart is constructed based on the known process model and only used in the case study of an AR (1) process. ARL is taken as the average run length of control chart in Phase II of the same simulations.

5.1 AR (1) process

In this section, four charts, i.e., CSC, the *modified* Shewhart chart, the conventional Shewhart chart (a) and the conventional Shewhart chart (b), are tested in terms of their detection performances. The ACF limit is chosen as 0.4, which is considered as a low magnitude of autocorrelation. An AR (1) process is simulated with the autocorrelation coefficient of 0.8. While the direct comparison may not be truly meaningful, a study of these various charts still provides significant insight towards their use. The critical values of c of the conventional charts (a) and (b) are adjusted as 2.987 and 2.997, respectively, to maintain their ARL_0's close to the *modified* chart which acts as the benchmark. The control limits of the CSC do not need adjustment, because its ARL_0 is approximately the same as that of the *modified* chart with the default value of $c = 3$.

Five kinds of mean-shift cases are simulated and the out-of-control ARL_1s of the four charts are averaged based on 10,000 simulations (**Table 2**). The CSC is able to signal right after the *modified* Shewhart chart signals for every simulated mean-shift case. It means that the CSC is effective in detecting faults and performs very close to the benchmark *modified* Shewhart chart. The adverse effect of autocorrelation on the conventional chart (a) is significant and the detection delay of this chart is remarkable especially for the cases of small mean shift and large data autocorrelation. After slightly adjusting the control limits, the conventional chart (b) can yield a comparable performance with the CSC. However, we must note that it may not be practical to adjust the control limits of the conventional chart and select a 'long enough' time series in Phase I.

In summary, there are two alternatives to the *modified* Shewhart chart when the process model is not available or not easy to estimate, i.e., the conventional Shewhart chart using a long enough Phase-I data and CSC. For the former it may be difficult to determine the data length in phase I that is enough for a correct estimation of process mean and standard deviation and it varies with the magnitude of autocorrelation

and underlying model. The CSC is a totally model-free and data-driven method. In its Phase I, the subgroup size can be constant and R is the only free parameter, which is based on the autocorrelation function estimated from a Phase-I time series. Both charts are able to achieve a similar performance to the *modified* chart. However, the conventional Shewhart chart is sensitive to the Phase-I data length, while the CSC performs consistently well. Considering that, we can state that the CSC can be a general method of accommodating data autocorrelation in a Shewhart chart and it is a reliable and valid alternative to the *modified* Shewhart chart.

Table 2. Comparison of ARL$_1$ in mean-shift cases.

Autocorrelation coefficient 0.8				
Mean Shift	CSC	*Modified* chart	Adjusted conventional chart (a)	Adjusted conventional chart (b)
0	556	557	550	555
0.5	275	272	296	277
1	93.6	93.2	98.9	93.9
1.5	36.8	36.3	37.3	36.6
2	15.9	16	16.2	15.9
3	4.04	4.06	4.04	4.03

It should be noted that, in industrial chemical processes, variables are measured at high sampling rates often measured in seconds. This creates serial dependence in data as system dynamics is usually slower than the measurement rate, resulting in highly serially correlated variables. If the skipping strategy is applied directly on the original observations, a large R would have to be chosen to reduce the autocorrelation to a reasonable level. A large number of SSCs would be constructed and the control scheme becomes complicated. As the observations are highly autocorrelated, it means that the measured values for adjacent few samples are nearly constant. Thus, a data preprocessing step becomes necessary before monitoring the process. The mean filter is a way of denoising and condensing data (Davies, 1992). It can reduce the complexity of control charts and also take advantage of the observations collected at a high sampling rate. The output of a mean filter is the mean value of a data window and it is then monitored in the skipping control scheme. The width of the data window is chosen as W, the maximum time lag where the autocorrelation remains at a level near 1. This adjusted control strategy seems analogous to the Shewhart chart for subgrouped data with a sample size W but the way of choosing the sample size is different.

5.2 An industrial process

In this section, CSC is applied to the monitoring of an industrial process. A depropanizing column of a gas fractionation unit in an industrial chemical plant located in China is being monitored. The process variables are collected every 5 seconds. In order to demonstrate how the CSC works in an industrial application, the column pressure and level variables are monitored.

Historical measurements of around 6 days are chosen in this case (**Fig. 1**). It is obvious in the pressure chart that the process drifts into another mode in the second half of the observations, while the level variable stays in normal condition. The autocorrelation of these two variables are estimated based on the first half of the data, plotted in **Fig. 3**. Both variables present high order autocorrelations. The width of the mean filter is chosen as 45, at which lag both autocorrelation functions start declining slowly. The data set is then condensed and the sample size is taken as 45. The first half of the data is chosen as the Phase-I data set, while the rest is assigned in Phase II. The Phase-I data set is condensed as in **Fig. 2**. The autocorrelation functions of the two variables are estimated according to the condensed data in **Fig. 3**. ACF limit is chosen as 0.4 and the CSC can be constructed with the skipping lag as 5 and 12 for the pressure and level variables, respectively. The monitored point in Phase II is updated every 3.75min in the CSC and the CSCs for these two variables in phase II are shown in **Fig. 4** using the 3-sigma control limits. In the CSC of the pressure variable, drifted points can be detected in time, while the observations of the level variable, free from the disturbance, stays within the control limits with no false alarm.
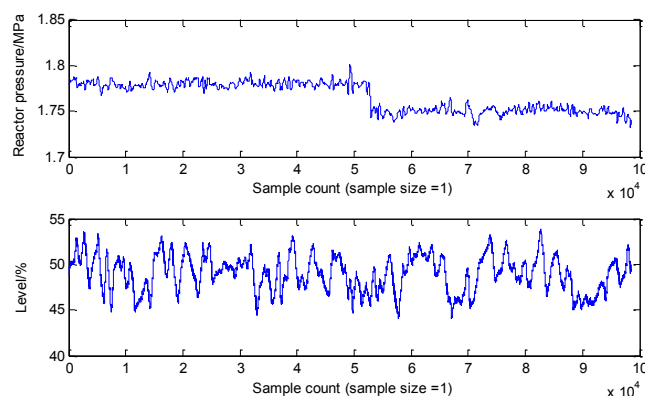

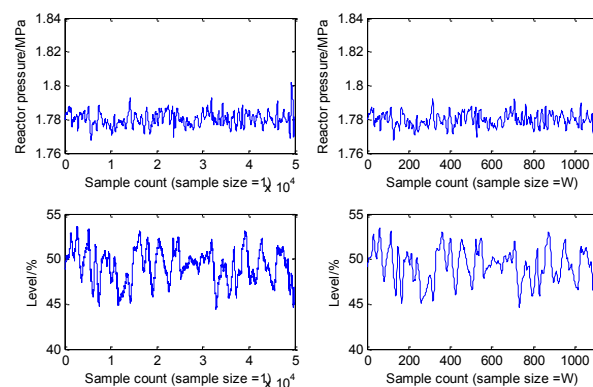Fig. 1 The original measurements for reactor pressure and level the industrial case study.

Fig. 2 Phase-I data chosen for the two variables in forms of original observations and condensed observations.
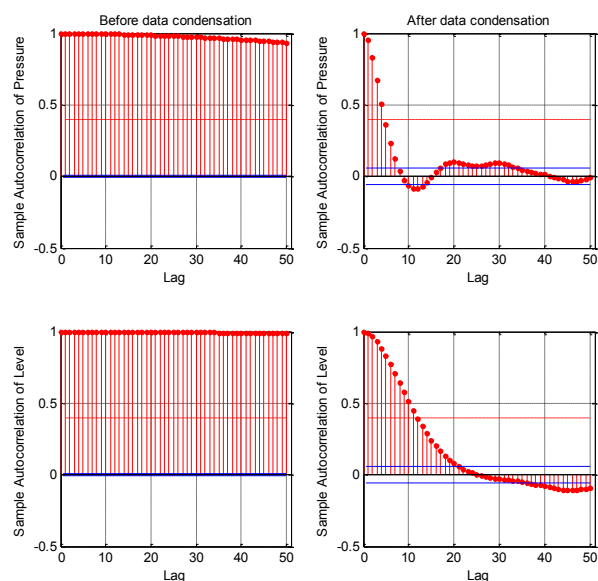
Fig. 3 Autocorrelation functions of the two variables estimated from original observations and condensed observations through a mean filter.
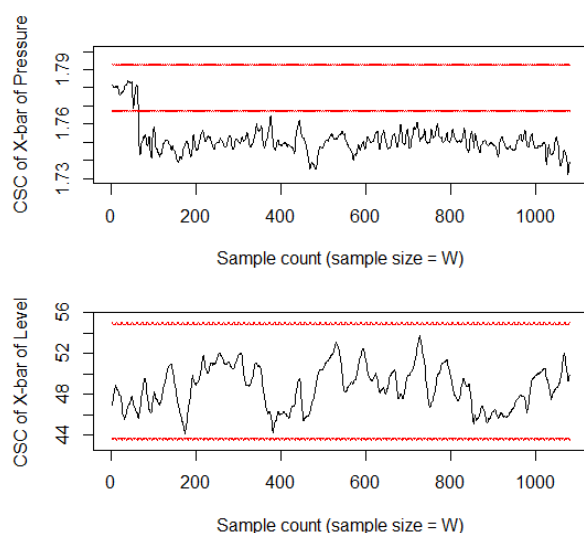

Fig. 4 CSCs of the variables of reactor pressure and level in the industrial application.

6. CONCLUSIONS

The estimation of the true mean and variance of quality characteristics in the Shewhart chart for individual observations is challenging especially when the quality characteristics are autocorrelated. A model-free Shewhart individuals control scheme for autocorrelated data is proposed. Two control charts, i.e., SSC and CSC, are designed to offer a flexible approach to monitoring individual observations. An alternative to the *modified* Shewhart chart is presented based on an iterative skipping strategy. The in-control and out-of-control performances of CSC are obtained using simulated data. Results show that CSC can achieve similar performance to the *modified* Shewhart chart. Meanwhile, conventional Shewhart charts based on different Phase-I data lengths are also carried out to monitor the same

variables. Under the assumption of independence, conventional chart (a) is least sensitive to abnormality in all faulty cases. Although CSC and the conventional chart (b) work with different control limits, they seem to perform similarly in terms of run length.

After the verification in the case study of AR (1) process that the CSC is a reasonable alternative to the *modified* chart, which is difficult to construct in practice, CSC is constructed to monitor variables in an industrial process to demonstrate its practical application.

As the ACF limit is the only free parameter in the proposed skipping Shewhart control scheme, a sensitivity analysis should be carried out to determine an appropriate value for the ACF limit.

When the Shewhart individuals chart is extended to the multivariate case as in Hotteling's T^2 chart, the effect of autocorrelation would become even more complicated. In this paper, the univariate charts provided a straightforward medium to clearly articulate the salient features of the proposed methodology. As multivariate SPC approaches exploit cross-correlation among variables, they may perform better than univariate SPC approaches by detecting the correlation abnormalities among variables. The skipping strategy proposed in this work may also be an effective way for multivariate SPC charts to reduce the effect of data autocorrelation, which will be addressed in our future work. This goal of this paper has been to lay down the groundwork to illustrate the capability of the skipping strategy by applying it to a univariate SPC chart.

REFERENCES

Alwan, L.C. (1991). Autocorrelation: Fixed versus variable control limits. *Quality Engineering*, 4(2), 167-188.

Alwan, L.C. and Roberts, H.V. (1988). Time-series modeling for statistical process control. *Journal of Business & Economic Statistics*, 6(1), 87-95.

Box, G.E., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015). *Time series analysis: forecasting and control.* John Wiley & Sons.

Davies, E.R. (1992). The relative effects of median and mean filters on noisy signals. *Journal of Modern Optics*, 39(1), 103-113.

Harris, T.J. and Ross, W.H. (1991). Statistical process control procedures for correlated observations. *The Canadian Journal of Chemical Engineering*, 69(1), 48-57.

Johnson, R.A. and Bagshaw, M. (1974). The effect of serial correlation on the performance of CUSUM tests. *Technometrics*, 16(1), 103-112.

Knoth, S. and Schmid, W. (2004). Control charts for time series: a review. *Frontiers in statistical quality control*, 7, 210-236.

Bisgaard, S. and Kulahci, M. (2011), *Time Series Analysis and Forecasting by Example,* New York: Wiley

Li, R. (2004). New Multivariate Schemes for Statistical Process Control (Doctoral dissertation, Dissertation, Department of Statistics, Univ. of Wisconsin).

Lin, S.N., Chou, C.Y., Wang, S.L. and Liu, H.R. (2012). Economic design of autoregressive moving average control chart using genetic algorithms. *Expert systems with applications*, 39(2), 1793-1798.

Montgomery, D.C. (2007). *Introduction to statistical quality control*. John Wiley & Sons.

Montgomery, D. C., Jennings, C. L. and Kulahci, M. (2015), *Introduction to Time Series Analysis and Forecasting*, 2nd Edition, New York: Wiley

Padgett, C. S., Thombs, L. A., & Padgett, W. J. (1992). On the α-risks for Shewhart control charts. *Communications in Statistics-Simulation and Computation*, 21(4), 1125-1147.

Pan, X. and Jarrett, J.E. (2012). Why and how to use vector autoregressive models for quality control: the guideline and procedures. *Quality & Quantity*, 46(3), 935-948.

Prybutok, V.R., Clayton, H.R. and Harvey, M.M. (1997). Comparison of fixed versus variable sampling interval Shewhart control charts in the presence of positively autocorrelated data. *Communications in Statistics-Simulation and Computation*, 26(1), 83-106.

Reynolds Jr, M.R. and Lu, C.W. (1997). Control charts for monitoring processes with autocorrelated data. Nonlinear Analysis: Theory, *Methods & Applications*, 30(7), pp.4059-4067.

Schmid, W. and Schone, A. (1997). Some properties of the EWMA control chart in the presence of autocorrelation. *The Annals of Statistics*, 1277-1283.

Western Electric Company. (1958). *Statistical Quality Control Handbook*, 2nd ed., AT&T technologies, Indianapolis, IN.

Wardell, D.G., Moskowitz, H. and Plante, R.D. (1994). Run-length distributions of special-cause control charts for correlated processes. *Technometrics*, 36(1), 3-17.