3rd Research/Expert Conference with International Participations
"QUALITY" 2003, Zenica, B&H, 13 and 14 November, 2003.

EFFECT OF AUTOCORRELATION ON SPC CHART PERFORMANCE

Vesna Bucevska, Ph.D.
University "St. Cyril and Methodius", Faculty of Economics
Skopje, Republic of Macedonia

Keywords: Autocorrelation, SPC charts, stationarity

ABSTRACT

Statistical Process Control (SPC) aims at quality improvements through reduction of variances. The best known tool of SPC is the control chart. Over the years the control chart has proved to be a successful practical technique for monitoring process measurements. However, its usefulness in practice is limited to those situations where it can be assumed that successive measurements are independently distributed, whereas most data sets encountered in practice exhibit some form of serial correlation.
The question that is considered in this paper is what control chart methods should be used to monitor serially correlated data and how the signals on such charts should be interpreted.

1. STOCHASTIC PROCESS

Classical control charts assume no correlation between successive observations of the quality characteristic. In this section we define in a more precise manner what is meant by correlation between repeatedly observed measurements of a single quality characteristic. We need to know how to estimate the serial correlation, in case it exists, and how to study its effect on classical SPC charts. To achieve these goals, the concept of a stochastic process is first necessary.
A stochastic process $\{Y(t), t \in I\}$ is a family of indexed random variables, where *I* is called the index set. Sometimes we will refer to the mechanism generating the stochastic process simply as the process, which can be understood in double sense of the underlying stochastic process that the quality characteristic being modeled follows, or as the production process itself, which in turn generates the stochastic process (i.e. the quality characteristic). In applications in this paper, *t* will relate to the discrete points of time at which an observation is obtained by sampling (i.e. the index set is $I = \{..., -2, -1, 0, 1, 2, ...\}$) and the stochastic process is said to be a *discrete-time* stochastic process. If $I = \{t : -\infty < t < +\infty\}$, the process is a continuos-time stochastic process. For discrete-time stochastic processes, it is customary to denote them as $\{Y_t\}_{t \in I}$, that is, a subscript is used for discrete-time indices. Discrete-time processes can be identified by describing the behavior of its *t*th element. Thus, we can write, for example,

$$Y_t = \mu + \varepsilon_t$$

Implying that for the given process (Shewhart's model in this case) the equation holds for all discrete points in time t. In this case, the time between observation *h* equals Δ*t*.

A stochastic process can be thought of as a function of two arguments: namely, $\{Y(t,w), t \in I, w \in \varsigma\}$, where ς is the sample space of the random variables $Y(t)$. It is important to point out that for fixed $t \in I, Y(t,.)$ is a random variable; for fixed $w \in \varsigma, Y(.,w)$ is a function of time called a realization, or trajectory of the process. A time series is a set of observations, or realization, of a discrete-time stochastic process. In basic probability theory, this is analogous to the notion of a single observed value of a random variable (*y*) compared to the random variable itself (*Y*). The set of all possible realizations of a stochastic process is called the ensemble of the process.

As in many other areas of statistics, to perform valid statistical inferences from a stochastic process, we need some notion of how repeatable the underlying random experiment is under identical conditions. For example, one such statistical inference could be to predict to where a quality characteristic will move in the near future. The only information available to us is a single realization of a stochastic process, the values recorded of the quality characteristic in the past. If this process is such that its random variables differ radically at different points of time, no inference could be drawn from a single realization since the probabilistic properties of the process during the period of time when the observations were taken cannot be generalized to other periods of time. For the class of models studied in this paper, the notion that we need to make valid inferences is called stationary.

A stochastic process is strictly stationary if for any integer $k \geq 1$, the joint distribution of $\{Y_{t_1}, Y_{t_2}, ..., Y_{t_k}\}$ is identical to the distribution of $\{Y_{t_1+\tau}, Y_{t_2+\tau}, ..., Y_{t_k+\tau}\}$, where $t_i \in I, t_i + \tau \in I$ .Thus the stochastic properties of the process are unaffected by changes in the time origin. If in this definition we look at the case k=1, we note that strict stationarity implies that the distribution of Y_t is the same as the distribution for for $Y_{t+\tau}$. Therefore, strict stationary implies that the distribution of the random variables occurring at different points in time is identical.

3. MEAN AND VARIANCE OF A STATIONARY PROCESS

Since for a strictly stationary discrete-time process $\{Y_t\}_{-\infty}^{\infty}$ the probability density function $f(Y_t)$ is the same for all t, that is,

$$f(Y_t) = f(Y)$$

we have that

$$\mu = E[Y_t] = \int_{-\infty}^{\infty} yf(y)dy$$

$$\sigma^2 = Var[Y_t] = \int_{-\infty}^{\infty} y^2 f(y)dy - \mu^2 .$$

Thus μ gives the long-run or asymptotic average level of the process at every time *t* and σ^2 gives the long run dispersion of the process at every time *t*. These long-run quantities are obtained over the ensemble of the stochastic process, which is the population from which we sample. For example, $E[Y_t]$ can be thought as the level of the process at time t averaged over an infinitely large number of possible realizations. The corresponding point estimates are given by

$$\hat{\mu} = \overline{Y} = \frac{1}{N}\sum_{t=1}^{N} y_t$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{t-1}^{N} \left(y_t - \overline{Y}\right)^2$$

where N denotes the length of the time series observed.

4. AUTOCOVARIANCE AND AUTOCORRELATION

The prefix auto means a reflexive act upon oneself, thus autocovariance is the covariance that a process has "with itself". The autocovariance function $(\gamma_{t,k})$ gives a measure of linear association (covariance) between two variables of the same process, Y_t and Y_{t+k}, that are separated k period of lags, as a function of k. For any discrete-time process, we define

$$\gamma_{t,k} = Cov[Y_t, Y_{t+k}] = E[(Y_t - E[Y_t])(Y_{t+k} - E[Y_{t+k}])] \text{ for } k = 0, \pm1, \pm2, ...$$

For a strictly stationary process, the mean μ_t is constant for all times t, and the autocovariance reduces to

$$\gamma_{t,k} = \gamma_k = Cov[Y_j, Y_{j+k}] = E[(Y_t - \mu)(Y_{t+k} - \mu)] \text{ for } k = 0, \pm1, \pm2, ...$$

so the autocovariance depends only on the lag *k*. Also, note that $\gamma_0 = E[(Y_t - \mu^2)] = Var(Y_t) = \sigma^2$.

It is usually better to scale the autocovariance into a unitless quantity by dividing by the process variance. For a stationary process,

$$\rho_k = \frac{\gamma_k}{\gamma_0} \text{ for } k = 0, \pm1, \pm2, ...$$

where $-1 \le \rho_k \le 1$. Given that $\gamma_k = \gamma_{-k}$ and $\rho_k = \rho_{-k}$ (i.e., the autocorrelation and autocovariance are *even* functions), it is common to plot these two functions only for positive lags. Both the autocovariance and the autocorrelation at lag *k* give a measure of the degree of linear association between two random variables of the same process that are separated *k* periods.

When considering relations between random variables, it is useful to recall the following implications:

1. If Y_t and Y_{t+k} are independent, they are uncorrelated (i.e., $\rho_k = 0$ for all *k*)
2. If Y_t and Y_{t+k} are correlated (i.e, $\rho_k \neq 0$ for some *k*), Y_t and Y_{t+k} are dependent.

The direction of the implications is important; uncorrelated random variables
may or may not be independent. Correlation is a measure of linear association, so there might be some nonlinear association between uncorrelated variables.
It should be pointed out that the standard error of the average $(\bar{Y})$ is greatly affected by autocorrelation. Bartlett (1946) showed that

$$\sigma_{\bar{Y}} = \sqrt{\frac{\gamma_0}{N}\left[1 + 2\sum_{k-1}^{N}\left(1 - \frac{k}{N}\right)\gamma_k\right]}$$

Thus if the stochastic process is completely uncorrelated ($\gamma_k = 0$ for all *k*), the standard error of the average $\bar{Y}$ is the usual $\sigma / \sqrt{N}$. As the (positive) autocorrelation increases, so does the standard error and the point estimate $\bar{Y}$ becomes less reliable.

Point estimates of the autocovariance function are given by the sample autocovariance function, computed as

$$c_k = \hat{\gamma}_k = \frac{1}{N}\sum_{t-1}^{N-k}\left(y_t - \overline{Y}\right)\left(y_{t+k} - \overline{Y}\right) \qquad k = 0,1,2,... \qquad (1.6)$$

Similarly, the sample autocorrelation function is given by

$$r_k = \hat{\rho}_k = \frac{c_k}{c_0} \qquad k = 0,1,2,...$$

5. NEED FOR STATIONARITY

Note that the point estimates $\overline{Y}, \hat{\sigma}^2, c_k,$ and r_k are computed based on a single realization of the process (i.e., they are computed based on time-series data). Inferences based on a single realization are possible because the process is stationary.
A different form of stationarity is called weakly or covariance stationarity. A process is weakly stationary if only its first two moments (mean and covariance) are finite and independent of time (in such case the covariance depends only on the lag (. If each Y_t is normally distributed and the process is weakly stationary, the process is strictly stationary. As it turns out, the type of stochastic models we study are fully characterized by their first two moments, so when we refer to stationarity we are referring to weakly stationarity. Weak stationarity can be better understood from an engineering point of view. If the mean of the process is constant, some process engineers would say that the process is stable. As will be seen, stability is a property of a deterministic dynamic system, whereas stationary is a properly related to a stochastic process.

6. EFFECT OF AUTOCORRELATION ON SPC CHART PERFORMANCE

As mentioned before, SPC control charts assume that if in control, the process has a constant mean and is completely uncorrelated. An important practical question is to investigate what happens with the performance of SPC charts as the process exhibits more and more serial correlation. As one process engineer said "almost every production process exhibits autocorrelation."
The effect and implications of autocorrelation have been topics of frequent discussion and debate in SPC literature. For a detailed review please refer to Knoth and Schmid (2001). These issues are usually tackled numerically: take for instance the investigations by Johnson and Bagshaw (1974), Alwan (1992), Maragah and Woodall (1992), Wardell, Moskowitz and Plante (1994), Runger, Willemain and Prabhu (1995), Van Brackle III and Reynolds Jr. (1997) and Lu and Reynolds Jr. (1999). These and several other papers have tables and graphs usually referring to the ARL, to provide evidence that the performance of the appealing control schemes is severely compromised by the presence of serial correlation.
Autocorrelation often reflects increased variability. Thus, the first two options to consider should be to remove the source of the autocorrelation or to use some type of process adjustment scheme such as those discussed by Box and Luceno (1997) and Hunter (1998). Control charting can be used in conduction with process adjustment schemes, and Box and Luceno (1997) emphasized that the two types of tools should be used together. Only if these

two options prove infeasible should one consider using stand-alone control charts for process monitoring such as those discussed by Lu and Reynolds (1999), Lin and Adams (1996) and Adams and Lin (1999).
Positive autocorrelation at low lags is common because given the advances in sensor technologies, measurements are taken closer together in time. In discrete-part manufacturing, this sometimes implies that every part is measured. Observations that were generated close in time will tend to be similar, hence positive correlation at low lags will result. The following two examples illustrate the effect of autocorrelation on the performance of SPC charts.

Example: Suppose that a process is described by

$$Z_t = \varepsilon_t + \lambda Z_{t-1} + \lambda(1-\lambda)Z_{t-2} + \lambda(1-\lambda)^2 Z_{t-3} + ...$$
$$Y_t = \mu + Z_t$$

where $0 \le \lambda \le 1$ and Y_t is the quality characteristic. There are two extreme cases: if $\lambda = 0$, the process is just Shewhart's process, where the observations are completely uncorrelated. If $\lambda = 1$, the process is

$$Y_t = \mu + Z_t = \mu + \varepsilon_t + Z_{t-1}$$

which evidently is highly correlated with

$$Y_{t-1} = \mu + Z_{t-1}$$

Vander Weil considers this process as λ increases from 0 to 1.

In the realizations of Figure 1.13b and d, a sustained shift in the mean of magnitude 5σ was induced at time 100; that is, the mean of the process changes according to

$$\mu_t = \begin{cases} \mu_0 & \text{if} \quad t \le t_0 \\ \mu_0 + \delta\sigma & \text{if} \quad t > t_0 \end{cases}$$

where in this example $\delta = 5, t_0 = 100,$ and $\mu_0 = 10$. This type of shift suddenly changes the mean of the process at time t_0. As can be seen from the figures, for $\lambda = 0$, detection should be almost immediate for any SPC chart and false alarms should be infrequent. For $\lambda \ge 0.8$, the shift becomes indistinguishable from the autocorrelation structure. Such positive autocorrelation will not necessarily worsen the detection capabilities of the charts as measured by the ARL_{out} performance criterion (Goldsmith and Whitfield, 1961; Lu and Reynolds, 1999, 2001), but the number of false alarms will certainly increase. A chart that gives frequent false alarms will soon be abandoned.

Example 2: The impact of autocorrelation on SPC chart performance was studied by Mragah and Woodall (1992), who consider the process

$$Y_t = \mu + \phi Y_{t-1} + \varepsilon_t \qquad (1.8)$$

where $-1 < \phi < 1$ is a parameter that allow us to define the autocorrelation in the $\{Y_t\}$ process. If $\phi = 0$ in equation (1.8) we obtain a Shewhart's uncorrelated process. The autocorrelation function of this process is $\rho_k = \phi^k$, so the process is stationary as long as

$|\phi| < 1$. Note that for $\phi = 1$, the process turns out to be a random walk. In the case of positive autocorrelation $(0 < \phi < 1)$, the movement of the process is smoother than that of an uncorrelated process. In the case of negative autocorrelation $(-1 < \phi < 0)$, the process shows a sawtooth pattern, which compared to Shewhart's process is much more crumpled.
Maragah and Woodall (1992) computed by simulation the average number of out-of-control points that a Shewhart chart for individuals will generate it 25 observations of process (1.8) with $-0.9 \le \phi \le 0.9$ are used to set the chart limits. A chart for individuals or a Y chart is one in which no subgroups are formed. The control limits are computed using the moving-range estimator[1] $\hat{\sigma}_Y = \overline{\mathrm{MR}} / d_2$, where $\mathrm{MR}_i = |Y_i - Y_{i-1}|$. Table 1 shows some of their results.

Table 1: Number of Points Outside Limits Generated by a Shewhart Chart for Individuals Used to Monitor 25 Observations of an AR(1) Process, Obtained by Simulation

ϕ	Average	Std. Dev.
-0.9	0.001	0.035
-0.7	0.004	0.063
-0.3	0.0027	0.166
-0.1	0.058	0.241
0.0	0.094	0.316
0.1	0.135	0.377
0.3	0.312	0.610
0.7	1.807	1.887
0.9	4.081	3.331

Source: Maragah and Woodall (1992)

Note that no shift in the mean was introduced while simulating the AR(1) process. From the table it can be seen that the number of out of control signals increases with increasing positive autocorrelation. For $\phi = 0.9$, for example, an average of about four signals will occur in 25 samples. This is a much higher signal rate than the advertised average of one false alarm every 370 observations for Shewhart charts. Negative autocorrelation reduces the number if signals, but positive autocorrelation is much more common in practice as mentioned before. |Negative autocorrelation at low lags inflates the variance and hence the control limits width becomes too large, so detecting actual shifts becomes more difficult.
The evident problem in Example 2 is that the limits were computed based on a variance estimate, which assumes that the process is uncorrelated. If $\phi > 0$ (positive autocorrelation), adjacent observations will tend to be similar and the moving-range estimator will underestimate the variance of the process. This will result in limits that are too narrow, producing many alarms compare to the uncorrelated vase. This result is quite general: *The within-sample variance estimator will underestimate the process variance in a positively autocorrelated process.* If $\phi < 0$ (negative autocorrelation), the moving-range estimator will overestimate the true variance, resulting in limits that are too wide, so very few signals will be obtained as opposed to the uncorrelated case. Note that this will have an impact on the detection capabilities of real shifts (i.e., wide limits will make the chart take much longer to detect shifts in the process mean).

[1] This estimator was used because it is the most common estimator in control charts for individuals, but as mentioned below, it is not recommended for autocorrelated data.

7. CONCLUSIONS

In recent years, statistical process control (SPC) for autocorrelated processes has received a great deal of attention, due in part to the increasing prevalence of autocorrelation in process inspection data. With improvements in measurement and data collection technology, processes can be sampled at higher rates, which often leads to data autocorrelation. It is well known that the run length properties of common SPC methods like CUSUM and $\overline{X}$ charts are strongly affected by data autocorrelation, and the in-control average run length (ARL) can be much shorter than intended if the autocorrelation is positive.

The most widely researched methods of SPC for autocorrelated processes are residual-based control charts, which involve fitting some form of autoregressive moving average (ARMA) model to the data and monitoring the model residuals (i.e., the one-step-ahead prediction errors). If the model is exact, then the model residuals are independent. Consequently, standard SPC control charts can be applied to the residuals with well understood in-control run length properties.

8. REFERENCES

1. Box, G. E. P., and Luceno, A., *Statistical Control by Monitoring and Feedback Adjustment*, New York, 1997
2. Montgomery, D. C., *Introduction to Statistical Quality Control (3rd ed.),* New York, 2000
3. Rao, C. R., *Statistics and Truth: Putting Chance to Work,* Fairland, MD. International Co-Operative Publishing House, 1989
4. Ryan, T. P., *Statistical Methods for Quality Improvement (2nd ed.),* New York, 2000
5. Shewhart, W. A., *Economic Control of Quality of Manufactured Product*, New York, 1931
6. Wetherill, G.B. and Brown, D.W., *Statistical Process Control: Theory and Practice,* Chapman and Hall, London, 1991.

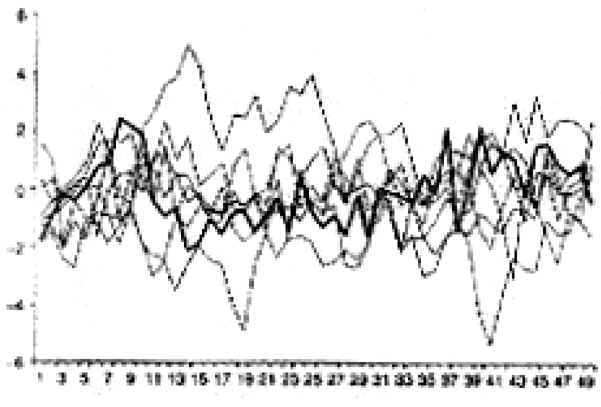

FIGURE 1. SINGLE REALIZATION (DARKER LINE) COMPARED TO OTHER POSSIBLE REALIZATIONS THAT MAKE UP THE ENSEMBLE OF A SINGLE STOCHASTIC PROCESS

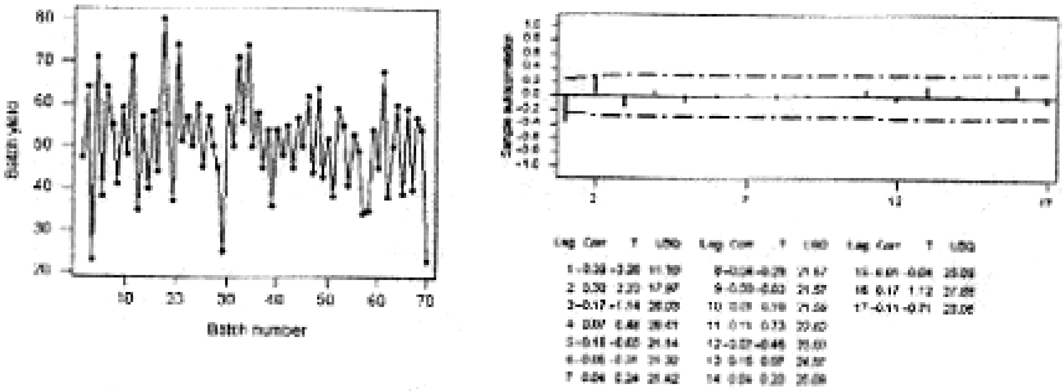

FIGURE 2. TIME-SERIES PLOT AND SAMPLE AUTOCORRELATION FUNCTION OF THE DATA IN EXAMPLE 1, CHEMICAL PROCESS DATA

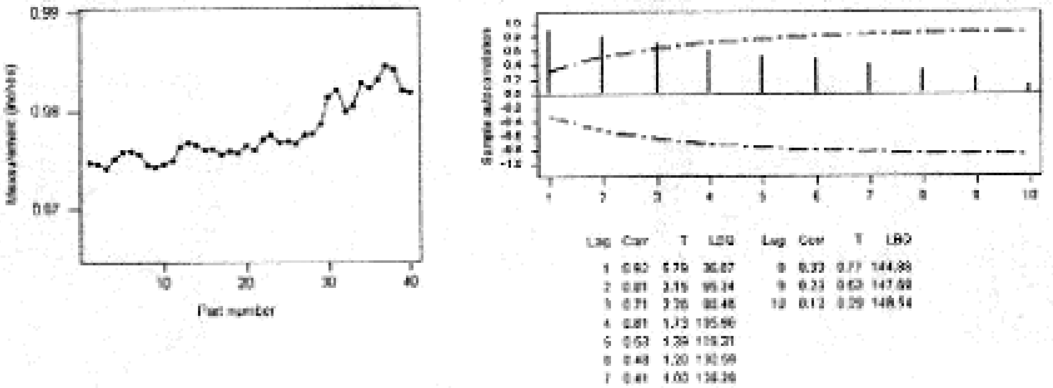

FIGURE 3. TIME-SERIES PLOT AND SAMPLE AUTOCORRELATION FUNCTION OF THE DATA IN EXAMPLE 2, MACHINED PARTS DATA

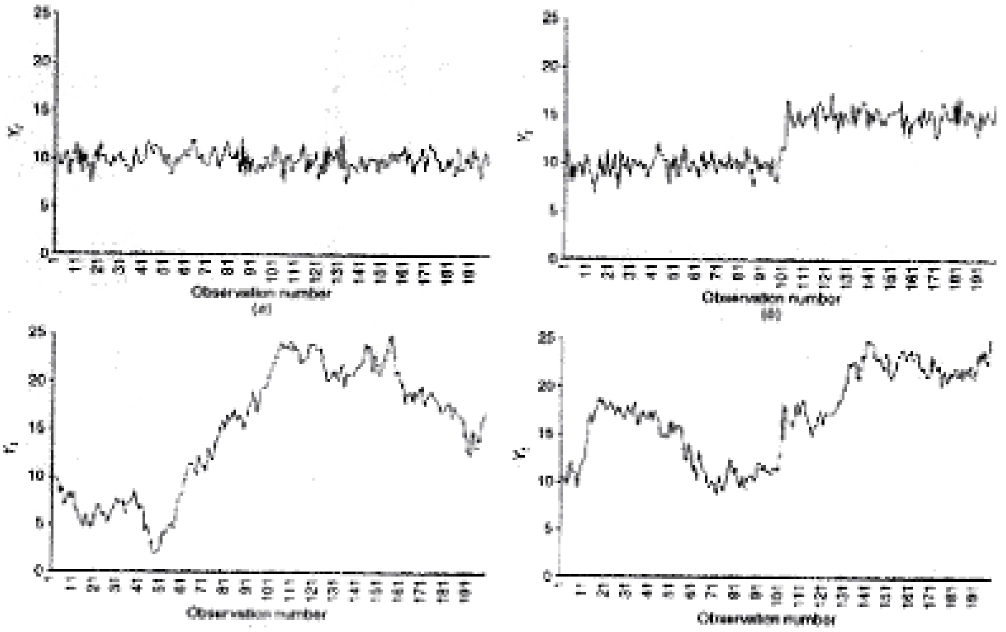

FIGURE 4. PROCESS WITH NO SUDDEN SHIFTS (a,c) AND WITH A SUDDEN SHIFT IN THE MEAN OCCURRING AT TIME t=100 (b, d), EXAMPLE 3. GRAPHS (a) AND (b) DISPLAY AN UNCORRECTED (λ=0); GRAPHS (c) AND (d) DISPLAY A HIGHLY AUTOCORRELATED (λ=0.8) SERIES