Report No. 41

Statistical Process Control and Automatic Process Control – A Discussion

George Box and Tim Kramer*

January 1990

This research was sponsored by the National Science Foundation under Grant No. DDM-8808138, and by the Vilas Trust of the University of Wisconsin-Madison.

*Applied Statistician, Hewlett Packard, Avondale, Pennsylvania
Controversy is often found in industry concerning the appropriate roles for Statistical Process Control and Automatic Process Control (Feedback, Feedforward Control). In this paper we discuss some of the relevant questions for example: Is the object to eliminate assignable causes of variation? Is the object to regulate the process? What is the nature of the noise? Are the effects of the adjustment immediate or does the process contain dynamics (inertia)? What is the cost of being off target? What is the cost of making an adjustment? Is there a cost of observing the process?

Simple models are postulated for the noise, the dynamics, and various costs. It is shown that many standard schemes such as Shewhart control charts, cusum charts, exponential charts, proportional-integral feedback control, and various feedforward schemes are appropriate or inappropriate depending on the answers to the above questions and the choice of parameters in the models. This leads to a much better fundamental understanding of the different circumstances which favor various schemes.

Key words: Statistical Process Control, Automatic Process Control, feedback, noise, cost of adjustment, non-stationarity, minimum cost regulation.
Rationales for Statistical Process Control (SPC) and for Automatic Process Control (APC) are explored and issues that sometimes arise are discussed. For the particular problem of process regulation the importance of, often unstated, assumptions is illustrated. Some implications of non-stationarity are presented and minimum cost schemes obtained for some simple, but practically interesting, models for the disturbance, the dynamics, and for various costs. Some examination of the implications of non-stationarity is presented.

1. Introduction

Control schemes can take many forms. In particular, Shewhart and other control charts are frequently employed for what is called Statistical Process Control (SPC). By contrast, various forms of feedback, feedforward, and feedforward-feedback regulation are used for Automatic Process Control (APC). The people responsible for SPC and those responsible for APC are usually from different departments and have different technical backgrounds so it is hardly surprising that recently there has been some controversy and misunderstanding. Some of these issues have been addressed by Deming (1950, 1986), Box and Jenkins (1963), Box, Jenkins and MacGregor (1974), MacGregor (1976, 1987), Taguchi (1981), Alwan and Roberts (1986), Crowder (1986), Hunter (1986), Lorenzen and Vance (1986), MacGregor (1987), Adams (1988), Harris (1989), Hahn, Faltin, Tucker, Richards and Vander Wiel (1988), Adams and Woodall (1989), and Vander Wiel and Tucker (1989). We are grateful for valuable exchanges of ideas and information.

One reason for misunderstanding is the different purposes for which control schemes are used and the different meanings associated with the words process control. SPC and APC originated in different industries - the "parts" industry and the "process" industry respectively. The control objectives of these two industries were often very different.

The parts industry was typically attempting to reproduce individual items as accurately as possible; for example, to manufacture steel rods with diameters having smallest possible variation about the target value T. The process industries on the other hand often employed continuous systems and were typically concerned with yields of
product, percentage conversion of chemicals, measures of purity, and so forth. Their objective was usually to control various aspects of the process so as to produce consistently the highest possible mean values for these measures.

In the parts industry, it was likely that the properties of feed materials such as steel sheet were reasonably well controlled. In the process industries changes in external variables, such as ambient temperature and the properties of feed stocks, could produce undesirable changes and some form of regulation using continuous feedback or feedforward control was frequently employed to compensate for this.

In the parts industry, the cost of adjustment of the process was frequently substantial. It could involve, for example, the stopping of the machine and the replacement of a tool. Also, the frequency of monitoring the process might be an appreciable cost factor. In the process industries frequently, but not invariably, the only non-capital cost involved in automatic process regulation was the cost of being off-target.

In the past, industries of the two kinds were often segregated into more or less watertight compartments. Recently, however, these sharply drawn lines have begun to disappear.* One reason is that some new processes, such as the manufacturing of computer chips, are hybrids with certain aspects of manufacturing like those of the parts industries and other aspects like those of the chemical industry. Another reason is that conglomerate companies, where both kinds of manufacture occur, are now much more common. A third reason is that greater awareness of the importance of control has led each industry to experiment with the control technology of the other.

Thus, the previous isolation is now less common and the initial reaction of the different kinds of control specialists when their specialities have overlapped has sometimes been predictably acrimonious. This has been largely due, we believe, to misunderstandings. One objective of this paper is to try to remove some of these misunderstandings by more careful examination of the objectives and assumptions of different kinds of control techniques and by simple illustration. We begin by discussing examples of SPC and of APC.

---

*This is illustrated by the recent interest taken by the U.S. parts industry in experimental design. Large portions of this industry seem to have been unaware of the decades of experience with these techniques which is available from the process industries in their own country.
1.1 An Example of SPC Using a Shewhart Chart

To clarify some issues, consider the following example of the use of a Shewhart chart. A company has a small employee health clinic where cuts, bruises and other minor injuries are treated. To monitor the situation, a count is kept of the number of new patients attending the clinic each week. Over an extended period of time the average number is 20.6. Some data is plotted on the control chart shown in Figure 1. If the process were in a state of control the data would be a random sequence from the Poisson distribution with mean 20.6 drawn on the right of the figure. If the data does not look like a random sequence from this reference distribution, we have reason to suppose that something else is going on.

![Control Chart](image)

Figure 1. A control chart for weekly numbers of new patients at a clinic in relation to a Poisson reference distribution.

In practice, the whole reference distribution is not usually shown on the chart. Its function is served by control lines drawn at $\pm 3\sigma$ and sometimes, in addition, by warning lines at $\pm 2\sigma$. In the above example these lines are at $20.6 \pm k\sqrt{20.6}$. Control charts of this kind are of great value in helping to monitor and to distinguish between what Dr. W. Edwards Deming calls common causes and special (or assignable) causes. Common causes are those associated with the overall behavior of the system when it is in a state of control. For example, if management decided to install new machinery with special guards, making injury less likely, and a new stable system was established with a mean of
only 10, then a new common cause system would have been set up characterized by a new Poisson reference distribution with a mean of 10. Special causes on the other hand, are those producing temporary deviations from the stable state. Thus, a single data value falling above the upper control limit might turn out to be associated with a temporary increase in the number reporting sick as an excuse to attend a rarely available sporting event.

In statistical terms, the common cause system refers to generation of data by a stable model (random drawings from a Poisson distribution) while special causes correspond to temporary deviations from the model signalled by outliers or unusual patterns of points.

1.2 An Example of APC With Feedback Control

When a system is affected by a source of non-stationary disturbance such as a drifting change, caused, for example, by changes in the quality of feedstock or in ambient temperature. We can try to eliminate this cause of variation but, if elimination is impossible or is too expensive, we may need to compensate for it by feedback or feedforward control.

\[
e_t = Z_t + Y_t
\]

Figure 2. A feedback control scheme.

Figure 2 shows a simple feedback control loop. For illustration, suppose we need to control the viscosity of a polymer as close as possible to some target value \( T \) by manipulating the catalyst formulation \( X \), and that the need for control arises because of non-stationary disturbances whose overall effect on viscosity is represented at the output by \( Z \). Thus \( Z_t \) represents what would have happened to viscosity if no control action had been
taken. At time $t$ the effect of present and past adjustments to the catalyst formulation is experienced at the output as a compensation of $Y_t$ units of viscosity. If, as is convenient, we consider the disturbance $Z_t$ and the attempted compensation $Y_t$ as deviations from the target value $T$, then the output error, that is the deviation from the target, will be

$$e_t = Z_t + Y_t.$$  \hspace{1cm} (1)

The feedback control equation determines how the catalyst formulation $X$ should be adjusted and calls for action which is some function of the current and past errors. In many automatic controllers, measurements and control actions are taken continuously. In particular, for the commonly used Proportional-Integral or PI controller, the control action $X$ is a linear function of the present error and the integral over time of past errors.

For simplicity in what follows we will consider a *discrete system* where the process is monitored, and it is possible to take control action, at equally spaced time intervals. The appropriate analog of PI control is then

$$-X_t = k_0 + k_p e_t + k_i \sum_{i=1}^{t} e_i$$  \hspace{1cm} (2)

where $k_p$ and $k_i$ are positive constants.

2. Models for Disturbances and Dynamics

Proportional-integral feedback control schemes of this kind have been used for a very long time (Mahr, 1970) and were originally developed empirically. They can, however, be shown to provide minimum mean square error control about the target if certain specific assumptions are true. These we now discuss.

2.1 A Non-Stationary Disturbance Model

The need for process regulation arises when the system is afflicted with disturbances which cause it to drift off target if no action is taken. In developing a model for feedback control, therefore, it is essential to incorporate a reasonably realistic representation of the non-stationary disturbance which must be compensated.

The familiar *stationary* disturbance model with "white noise" errors independently and identically distributed about the target

$$Z_t = T + a_t.$$  \hspace{1cm} (3)
represents the situation where there is no drift and the process is in a perfect state of control. An extreme form of non-stationary drift could be represented by the random walk model \( Z_i = Z_{i-1} + a_i \) in which the first difference of the disturbance \( Z \) is represented by white noise. Equivalently, the random walk model representing drift after some initial time \( i = 0 \) at which the mean was on target is

\[
Z_t = T + \sum_{i=1}^{t} a_i .
\]  

(4)

In what follows we adopt for our disturbance model an interpolation between these two extremes

\[
Z_t = T + a_t + \gamma \sum_{i=1}^{t-1} a_i , \quad 0 \leq \gamma \leq 1
\]

(5)

in which \( \gamma \) is a measure of non-stationarity of the model. When \( \gamma = 0 \) we have the stationary model (3); when \( \gamma = 1 \) the random walk (4); intermediate values of \( \gamma \) can represent slight, moderate, or severe degrees of non-stationarity. A fuller discussion of this model can be found, for example, in Box & Jenkins (1976). For convenience we here set out some of its main properties.

A Measure of Non-Stationarity:

Consider the interval variance \( V(Z_{t+m} - Z_t) \) of the difference of observations taken \( m \) steps apart in a time series. The inflation of this variance for an interval of length \( m \) relative to that for a unit interval is measured by

\[
G(m) = \frac{V(Z_{t+m} - Z_t)}{V(Z_{t+1} - Z_t)}
\]

(6)

which is a natural measure of non-stationarity. For a white noise process, \( G(m) \) is independent of \( m \) and is equal to 1. For the non-stationary model of equation (5) the interval variance increases linearly with \( m \) at the rate \( \beta \) which depends only on \( \gamma \), thus:

\[
G(m) = 1 + \beta (m-1) \quad \text{with} \quad \beta = \frac{\gamma^2}{1 + (1-\gamma)^2} .
\]

(7)

In Appendix A it is shown that, on reasonable assumptions, such a linear increase in the variance uniquely implies the non-stationary model (5). This linear property may account for the model's central usefulness in representing non-stationary behavior. Insofar
as smooth functions can be approximated locally by linear ones, the model might, for moderate ranges of $m$, approximate any situation where the interval variance increases monotonically with $m$. Some idea of the behavior of the non-stationary disturbance model (5) may be obtained from Table 1 which shows, for various values of $\gamma$, the number of intervals $m$ needed for the interval variance to double (for $G(m)$ to take the value 2).

<table>
<thead>
<tr>
<th>Random Walk</th>
<th>White Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>1.00 0.80 0.50 0.30 0.20 0.10 0.05 0.01</td>
</tr>
<tr>
<td>$m$</td>
<td>2 2.6 6 18 42 182 762 19,802</td>
</tr>
</tbody>
</table>

Table 1. Number $m$ of intervals needed for interval variance to double.

For the parts industry it is easy to imagine mildly non-stationary situations where $\gamma$ might be between, say, 0.1 and 0.3.

The model (5) may also be written

$$Z_t = \hat{Z}_t + a_t$$

where $\hat{Z}_t$ is the exponentially weighted moving average (EWMA) of past data

$$\hat{Z}_t = w_0 Z_{t-1} + w_1 Z_{t-2} + \ldots \text{ with } w_i = \gamma^i$$

where $\theta = 1 - \gamma$ is often called the smoothing constant.

The EWMA $\hat{Z}_t$ can be shown to be the minimal mean square error forecast of the next observation $Z_t$ and then the forecast error $Z_t - \hat{Z}_t$ is $a_t$. The forecast can be conveniently updated as each new observation comes to hand by using the relation

$$\hat{Z}_{t+1} = \gamma Z_t + \theta \hat{Z}_t \quad (\theta = 1 - \gamma) .$$

By differencing equation (5) we obtain the model in the form of a stationary moving average time series

$$Z_t - Z_{t-1} = a_t - \theta a_{t-1} .$$

This form provides a check on the appropriateness of this non-stationary disturbance model, since it implies that the autocorrelations $\rho_k$ of the differences $W_t = Z_t - Z_{t-1}$ are all zero except for that at lag $k=1$. 

7
Suppose a record of, say, 50 or preferably more, successive values of a disturbance $Z_t$ affecting an industrial process is available (or can be reconstructed), then an estimate of the autocorrelation of the differences $W_t$ at lag $k$ is

$$r_k = \frac{\Sigma W_t W_{t-k}}{\Sigma W_t^2}.$$

(12)

Bearing in mind that the standard errors of these estimated autocovariances are approximated by $n^{-1/2}$, inspection of the sample autocorrelation function ($r_k$) can indicate the plausibility of the hypothesis that all $\rho_k$ except $\rho_1$ are zero.

Also, an estimate of the non-stationarity parameter $\gamma$ can be obtained from

$$\hat{\gamma} = 1 + \frac{1}{2r_1} \{ (\frac{1}{2r_1})^2 - 1 \}^{1/2}.$$

(13)

A table of this function is provided in the Box & Jenkins book, where also estimation of $\gamma$ using the more efficient likelihood method is presented. A fuller discussion of diagnostic procedures is also given there.

2.2 A Dynamic Model Describing the Inertia of the Process

In the feedback scheme shown in Figure 2, control is achieved by changing catalyst formulation $X_t$ which in turn changes viscosity $Y_t$. In modeling such a relation between the input and output of a process, we must allow for its inertial characteristics. A simple "first order" dynamic model which can approximate the behavior of many processes is characterized by the first order difference equation

$$Y_t = \text{constant} + \delta Y_{t-1} + g(1-\delta) X_{t-1} \quad 0 < \delta < 1.$$

(14)

The inertial properties of the equation can be appreciated from the consideration that, t time periods after a unit step change is made in $X$, the change in $Y$ will be $g(1-\delta^t)$. So that for this dynamic model the output change asymptotically approaches $g$ units and $g$ is called the system gain.

2.3 The Optimal Feedback Controller

If now we suppose that the disturbance is represented by equation (5) and that the process dynamics are represented by equation (14) then a PI control equation of the form
of equation (2) produces minimum mean square error* about the target value provided that
the proportional and integral constants $k_p$ and $k_i$ are set to the values

$$k_p = \frac{\gamma \delta}{g(1-\delta)} \quad \text{and} \quad k_i = \frac{\gamma}{g}.$$  \hspace{1cm} (15)

Equivalently, if we suppose that the cost of being off target is a quadratic function $k_T e_t^2$ of the deviations from the target, then this scheme would minimize overall cost provided we assume that this is the only additional cost beyond the initial capital cost of the control system.

2.4 A Further Simplification

For the purpose of the discussion that follows, we will adopt a simpler dynamic model where it can be assumed that essentially all the change induced by a step change in $X$ occurs in a single time interval. This corresponds to setting $\delta = 0$ in equation (14) and the dynamic model becomes

$$Y_t = \text{constant} + g X_{t-1}.$$  \hspace{1cm} (16)

The optimal feedback equation is then

$$-X_t = k_0 + k_1 \sum_{i=1}^{t} e_i$$  \hspace{1cm} (17)

where $k_1 = \gamma/g$ as before.

Notice that all the feedback schemes so far discussed have the feature that to achieve optimal control, changes must be made after each observation.

3. Criticisms of SPC and APC

Having presented simple examples of SPC and APC schemes and the assumptions that underlie them, we will discuss some aspects of controversy that sometimes arise.

SPC practitioners have sometimes criticized APC

i) for overcompensating disturbances; in particular, in the parts industry, stories of improvement when automatic controllers are disconnected are common,

*If $\delta$ approaches unity, a minimum mean square error scheme may require excessive control action. In that case modified schemes are often employed see, for example Åström (1970), Box & Jenkins (1970), MacGregor (1972). However, for the present purpose, we will not further discuss that complication.
ii) for compensating disturbances rather than removing them,

iii) for concealing information that might be used for quality improvement.

It is useful to discuss these questions and to employ the models introduced above for illustration. We begin by considering criticisms of automatic control by SPC practitioners. On the other hand, those responsible for APC have sometimes argued that Shewhart control charts are inefficient for regulating a process.

3.1 Overcompensation?

For illustration we consider a feedback scheme of the kind discussed above in the simplest case where all the change in the output induced by a change in \( X \) is realized in one time interval as in equation (16) so that the appropriate feedback equation (17) is simply

\[
-X_t = k_0 + k_1 \sum_{i=0}^{t} e_i.
\]

Although minimum mean square error control would be produced if we set \( k_1 = \gamma/g \), the control could be poor or even unstable if the constant \( k_1 \) were wrongly chosen. For illustration consider an extreme example where feedback is applied to a process for which the "disturbance" is white noise and so that \( \gamma \) is zero, and hence \( k_1 \) should be set equal to 0. Nevertheless suppose that \( k_1 \) was set equal to some positive value \( \hat{k}_1 \). Then if \( e_t \) is the error at the output having variance \( \sigma_e^2 \), it is shown in Appendix B that the inflation of the output variance would be

\[
\sigma_e^2/\sigma_a^2 = \begin{cases} 
(1 - \frac{1}{2g} \hat{k}_1)^{-1} & 0 \leq g \hat{k}_1 < 2 \\
\infty & 2 < g \hat{k}_1 < \infty
\end{cases} .
\]

(18)

For illustration with \( g = 1 \)

\[
\begin{array}{cccccc}
\hat{k}_1 & 0 & 0.50 & 1 & 1.5 & \geq 2 \\
\sigma_e^2/\sigma_a^2 & 1 & 1.33 & 2 & 4.0 & \infty
\end{array}
\]

Thus, for a process already in a state of control the application of feedback could increase the variance or even render the process unstable. Notice however that if the system had been correctly identified, then a value of \( \gamma = 0 \) would be substituted in \( k_1 = \gamma/g \) and the appropriate control action called for would be \( -X_t = k_0 \). That is, no control.
This concept was discussed earlier by Deming (1950, 1986) in relation to an experiment in which a marble is dropped through a funnel. Since, in this experiment the error is purely random, attempts to compensate by moving the funnel will merely increase the dispersion. In particular, Deming’s rule (2) in which an adjustment is made to “compensate” the last error corresponds, in our notation, to making an adjustment \( X_t - X_{t-1} = -e_t \) or equivalently to using the control equation \(-X_t = k_0 + \sum_{i=1}^{t} e_i\) with \( k_0 = X_0 \). This then is equivalent to setting \( k_t = 1 \) and from the table we see that the variance is doubled, as was found earlier by Deming.

Notice, however, that if successive errors had not been white noise but had followed the non-stationary model (5), then minimum mean square error control would be obtained with \(-X_t = k_0 + \gamma \sum_{i=1}^{t} e_i\) or \( X_t - X_{t-1} = -\gamma e_t \). In that case, the adjustment \( X_t - X_{t-1} \) should go partway towards compensating \( e_t \). Specifically, rather than discounting all of \( e_t \), only a proportion of \( \theta = 1 - \gamma \) should be discounted.

Unfortunately, in practice, a standard controller has often been hooked up to a system, the characteristics of which had not been identified. Furthermore, in recent years, controllers of the type developed for the process industries have sometimes been transferred to the parts industries without sufficient thought being given to their applicability. This is undoubtedly the basis for stories about disconnection of a controller producing smaller variation.

Formal identification of the disturbance and of the dynamics by the fitting of appropriate models (see for example Box & Jenkins 1976) could avoid these problems, but might be too tedious for routine use. A practical, but less than perfect, alternative is to use a more careful experimental approach for tuning a standard controller. For example using the output mean square error, or preferably its logarithm, as the objective function, the optimal setting for the control constants in a PI controller could be found using response surface methods - that is steepest descent using fractional factorial designs, plus in some cases surface exploration near the minimum (see for example Box (1951)). To avoid upsetting the system during the period of experimentation, runs might be made in the evolutionary operation mode (Box 1957, Box and Draper 1969).

Such an approach would, in particular, lead to no feedback control action when the process was already in a state of control and so would avoid overcompensation. This
experimental approach could be particularly valuable for systems which included many controllers whose effects might interact with each other. Different criteria, of course, such as immediacy of response could, if desired, be optimized in this way.

3.2 Some Disturbances Cannot be Removed

When a process is affected by a non-stationary disturbance we have two alternatives: we can try to eliminate it or we can compensate for it. Thus, a switch to an alternative supplier may result in a more uniform feedstock. But if a feedstock is a naturally occurring one such as a metallic ore, crude oil or lumber from a forest, achievement of uniformity may not be possible or too expensive and we must instead try to compensate for differences by feedback or feedforward regulation. Again, if we find that the temperature variation in Wisconsin from winter to summer is too extreme, some of us may move to California, but if we stay we can compensate for the cold weather by using a furnace controlled by a thermostat supplying appropriate feedback control. However, the fact that a process is regulated by feedback control need not prevent the tracking down of a source of disturbance and could sometimes lead to fundamental modification of the system, as we see in the next section.

3.3 Information Need Not Be Concealed by Automatic Control

As often practiced, automatic control conceals the nature of compensated disturbances. But this need not happen. Figure 3 illustrates the point. This shows the behavior of a simulated feedback scheme in which the non-stationarity parameter in (5) was $\gamma = 0.2$ and the dynamic parameter in (14) was $\delta = 0.5$. The calculations were made assuming the system to be controlled by the PI controller

$$-X_t = \text{constant} + 0.20 \, e_t + 0.20 \sum_{i=1}^{t} e_i$$  \hspace{1cm} (19)$$

which produces minimum mean square error for these parameter values.

For such a scheme $X_t$ and $e_t$ would be, or could be made, immediately available and if the dynamics were known, the exact compensation $Y_t = \hat{Z}_t$ could be computed and the original disturbance $Z_t$ reconstructed. Although this is not usually done, all of these qualities could be displayed on charts like those in Figure 3 and could be subject to routine examination.
Such an examination can be conducted from the standpoint of a generalized concept of common and special causes, where as before common causes are associated with the modelled process changeable only by management action, and special causes are associated with temporary deviations from the modelled process, as indicated by outliers.

(a) Disturbance: $Z_t$

(b) Control Action: $X_t$

(c) Compensation: $Y_t$

(d) Deviation from Target: $e_t$

Figure 3 (a) Disturbance $Z_t$ for a simulated process subject to feedback control
(b) Control action $X_t$
(c) Compensation of the disturbance $Y_t$
(d) Deviations from the target value $e_t$
For a feedback system like that illustrated above, the disturbance model and the
dynamic model would define the *common cause* system which could be changed by
management. For example, suppose the pattern over time of the reconstructed disturbance
was carefully recorded and studied and led to the discovery that a substantial portion of it
could be accounted for by variation of a particular impurity in the feedstock. Then
management might decide to change the system either by removing the impurity from a
feedstock before it reached the process, or if this was impossible or too expensive, by
measuring it and compensating for it by appropriate feedforward control.

In addition, a *special cause* producing a temporary deviation from the underlying
system model, induced perhaps by misoperation, could be evidenced by an outlier in the
sequence \( \{ e_t \} \) and could lead to remedial action. To illustrate this, we added a deviation of
size \( 3\sigma_4 \) to the 29th value of the disturbance \( Z_t \) in Figure 3(a). After the disturbance has
been subjected to feedback control, this outlier is clearly visible in the record of the
deviations \( e_t \) from target. Thus, for a properly tuned feedback system it could be useful to
run a Shewhart chart like Figure 3(d) on the deviations from target. However, for a
process that is improperly tuned, deviations from the model will be blurred and special
causes more difficult to detect.*

Thus, our conclusion is that, although automatic control could conceal features of a
process which might be used for its improvement, this need not happen.

### 3.4 Shewhart Control Charts Inefficient for Regulating a Process?

As we have said, the main purpose of the Shewhart chart is *not* to regulate the
process, but to monitor stable operation due to common causes, and to reveal special
causes. For *this purpose* it makes sense to react to apparent process changes only if they
are established as statistically significant. Not to do so could result in continually hunting
for phantom phenomena.

Granted that this is true, some critics argue from a different base. They say that even
for the purpose of *detecting* changes, the Shewhart chart is not as efficient as, for example,
cumulative sum (Cusum) charts or exponentially weighted moving average charts (EWMA)
charts. Now the Cusum chart is based on a likelihood test which is optimally sensitive to
small changes in the mean and other charts of this nature can be devised for optimal detec-
tion of deviations of other specific kinds, for example, see Ramírez (1989); Box and
Ramírez (1990). Thus *if we know what we are looking for* we can devise a very sensitive

*Also, for more complicated optimal schemes the error signal may need to be appropriately
filtered to ensure good detectability.
test to seek out that particular discrepancy. However, we would be unwise to rely on such specialized tests exclusively since they might be useless to detect unexpected forms of disturbances. The great virtue of the Shewhart chart is that it is a plot of the actual data and so can expose types of non-randomness of a totally unexpected kind. The problem has been likened by Box (1980) to that faced by a small country wishing to install an early warning radar system against air attack. Very sensitive radars could be used to monitor certain directions known to be likely sources of aggression but, for safety, they should be adjuncts to a multidirectional screen.

Returning to the problem of regulation, it is true that for compensating a non-stationary disturbance when the only cost is that of being off target, a Shewhart chart could be sluggish and inefficient compared with a correctly tuned feedback system. However, as we see below, whenever there is a substantial cost \( C_a \) for adjustment of the process, schemes for minimum cost regulation can look, superficially at least, much more like Shewhart schemes.

4. Minimum Cost Regulation: Some Simple Schemes

From the above we see that some of the factors which determine the appropriateness of a control scheme are:

a) the purpose of the scheme
b) the nature of the disturbances which affect the system and whether or not they can be eliminated
c) the dynamics and delays in adjusting the system
d) whether or not there is a particular input \( X \) that can be manipulated to compensate disturbances in a feedback control scheme
e) whether or not the disturbance at the output is partly determined by an input which can be measured and used in a feedforward scheme
f) the cost of being off-target
g) the cost of making adjustments
h) the cost of observing the process.

To illustrate how the choice of an optimal feedback control scheme depends on some of these factors, we now consider the different schemes that would result if we assume that our only purpose was to regulate a process with minimum cost and that:

i) the disturbance \( Z_t \) is modelled by equation (5) so that with \( \gamma = 0 \) the disturbance is white noise and with \( 0 < \gamma \leq 1 \) we can represent varying degrees of non-stationary drift,
ii) the dynamics are given by equation (16) in which all the change $Y_t$ at the output
induced by a change $X_t$ at the input occurs in one time interval,

iii) the cost of being off target $C_T$ is a quadratic function of the deviation from target,

iv) there is a fixed cost $C_a$ incurred by adjusting the process,

v) there is a cost $C_m/m$ incurred by monitoring the process every $m$ units of time.

We suppose an initial pilot run is made, or is already available from routine
observation of the disturbance, using a short monitoring interval which we refer to as the
*unit interval* $m=1$. This run is used to obtain estimates of $\gamma$ and $\sigma$. In what follows we
shall need the result (Box & Jenkins (1976)) that if this disturbance is sampled at intervals of
$m$ units then the sampled process is of the same form as the original process but with
non-stationary parameter $\Theta$ and smoothing constant $\Theta = 1-\Gamma$ given by

$$\Theta = A_m - \sqrt{A_m^2 - 1}$ where $A_m = 1 + \frac{m(1-\theta)^2}{2\theta}$ \hspace{1cm} (20)$$

Then from equation (10) the exponentially smoothed value may be calculated from the
recursive relation with the index $(u+1, u, u-1, \ldots)$ applied to observations made $u$ steps
apart:

$$\hat{Z}_{u+1} = \Gamma Z_u + \Theta \hat{Z}_u \hspace{1cm} (21)$$

In the special case where the monitoring interval was regarded as fixed, Box and Jenkins
(1963) showed that the minimal cost scheme could be represented by a chart in which
appropriately smoothed values of the disturbance were plotted between two limit lines
parallel to the target axis.

More recently, Kramer (1989) has shown that when the cost of monitoring the
process is included, the optimal scheme is still of this form. Then with the cost of adjust-
ment, the monitoring cost and the cost of being off target all included the overall cost $C$
per unit of time is

$$C = \frac{C_a}{ARL} + \frac{C_m}{m} + k_T \frac{E(\sum_{i=1}^{n} c_i^2)}{ARL} \hspace{1cm} (22)$$

where the ARL is the average run length between adjustments.
By minimizing this overall cost, it is possible to determine
i) when the process should be adjusted,
ii) what size of adjustment should be made, and
iii) how often should the process be monitored and data collected.

An outline of the argument whereby this minimization is achieved is presented in
Appendix C and is given in more detail in Kramer (1989).

![Diagram](image)

Figure 4. General form for the feedback control scheme to minimize overall cost.
The values of the disturbance $Z$ are indicated by dots, the smoothed values
$\hat{Z}$ are indicated by dashes.

The optimal scheme may be represented by a chart of the kind shown in Figure 4
where the process is monitored at intervals of $m$ units of time and the disturbance $Z$
measured as a deviation from the target $T$ is indicated by dots. The appropriately smoothed $\hat{Z}$
values are indicated by dashes. No compensatory action is taken until some smoothed
value $\hat{Z}_u$ falls outside the limits $T \pm L$, when an adjustment $Y_u = -\hat{Z}_u$ is made. In Figure 4
we suppose that the chart has been zeroed at time $u = 1$ and it has again been zeroed at time
$u = 10$ by an adjustment $Y_{10} = -\hat{Z}_{10}$.

In general, suppose successive adjustments are made at times $u$ and $v$ and that the
adjustment variable $X$ is linked to $Y$ by a dynamic equation of the form (16) then $X$
will need to be changed by an amount $X_v - X_u = -\hat{Z}_{uv}/g$. The chart is seen to be similar in
appearance to an EWMA chart but the limit lines are determined by relative costs and by the
degree of non-stationarity of the process and not by questions of significance. The constant determining the limit lines is \( L = \lambda \gamma \sigma \) where \( \gamma \) and \( \sigma \) are parameters of the original disturbance process monitored at unit intervals.

To an adequate approximation values of \( \lambda \) and \( m \) can be found using Charts A and B of Figure 5. The charts require values of three parameters - the non-stationarity measure \( \gamma \) and two cost ratios \( R_a \) and \( R_m \) given by

\[
R_a = \frac{C_o/C_T}{\gamma^2} \quad R_m = \frac{C_m/C_T}{\gamma^2}
\]

where \( C_T = k_T \sigma^2 \).

Strictly speaking, both \( \lambda \) and \( m \) are functions of all three parameters \( \gamma \), \( R_a \) and \( R_m \). But the chosen parameterization has the property that \( \lambda \) is nearly independent of \( \gamma \), and \( m \) is nearly independent of \( R_a \). Thus, for most practical purposes, the simpler two-dimensional representations of Charts A and B may be used. Notice that this by no means implies that \( \lambda \) and \( m \) are independent of the non-stationary measure \( \gamma \). They are highly dependent on \( \gamma \), but this dependence is very nearly taken account of by the inclusion of \( \gamma^2 \) in the denominators of both \( R_m \) and \( R_a \).

Figure 5. Chart A shows contours of the standardized action limit \( \lambda \) for various values of \( R_m \) and \( R_a \). Chart B shows contours of the monitoring interval \( m \) for various values of \( \gamma \) and \( R_m \).
4.1 An Example of the Use of the Charts A and B

Suppose we need to control a process subject to a non-stationary disturbance which has not been possible to remove nor to cancel by feedforward control. Suppose that for an important quality characteristic the target value is 340 and the specification limits are 325 and 355. The present operating procedure is to take a sample every hour and make a determination which costs $C_m = 200 from the result of which the operator may decide to make an adjustment. From past hourly records, estimates of the disturbance parameters are obtained of \( \gamma = 1 - \theta = 0.3 \) and \( \sigma = 3 \). Material which falls just outside of a specification limit would need to be reprocessed at a cost of $1,350 per hour's worth of material. Hence, we can estimate \( k_T \) to be \( k_T = \frac{1350}{225} = 6 \). Suppose finally that each adjustment of the process costs \( C_a = 600 \). Then \( \gamma^2 = 0.3^2 = 0.09 \), the off-target cost is \( C_T = k_T \sigma^2 = 6 \times 9 = 54 \) and the relative adjustment cost is \( R_a = \frac{C_a}{C_T} = \frac{600}{54} = 12.35 \). Finally, the relative monitoring cost is \( R_m = \frac{C_m}{C_T} = \frac{200}{54} = 41.2 \). Thus, from Chart A with \( R_a = 123.5 \) and \( R_m = 41.2 \) we see that \( \lambda \) is about 3.2, so that \( L = \lambda \gamma \sigma = 3.2 \times 0.3 \times 3 = 2.9 \) and the limit lines should be placed at 340 ± 2.9. Also, from Chart B with \( \gamma = 0.3 \) and \( R_m = 41.2 \), \( m \) is about 10 units so the monitoring interval should be about ten hours.

For this example using equation (20) with \( \theta = 0.7 \) and \( m = 10 \), we find that \( \Theta = 0.34 \). Thus, the smoothed value of the disturbance monitored at the interval \( m = 10 \) which is plotted on the chart, is computed from \( \hat{Z}_{u+1} = 0.66 Z_u + 0.34 \hat{Z}_u \). The process is adjusted as soon as this smoothed value crosses either limit line. The recursion can be initiated by setting the smoothed value \( \hat{Z}_2 \) equal to \( Z_1 \). But notice that re-initiation is not needed after every adjustment, since the relevant values of \( Z \) and \( \hat{Z} \) are both adjusted by the same amount.

4.2 Special Cases

(a) \( \gamma = 0 \): In this case, the disturbance is a white noise stationary process. The standardized limit \( \lambda \) and the monitoring interval \( m \) are then both infinite. This says the control action to be taken for a process "known to be" in a perfect state of control is no action.

(b) \( C_a \) negligible: In this case for any fixed \( m \) (say \( m = 1 \)) \( \lambda \) tends to zero and the limits \( T \pm L \) converge on the target value. Adjustments must therefore be made as each new value becomes available. Each adjustment is such as will just cancel the deviation of the exponentially smoothed value from the target value. Hence, after the u\textsuperscript{th} observation,
the effect of the adjustment is \( \hat{Z}_u - \hat{Z}_{u-1} \). These adjustments are made by manipulating \( X \) whence the total adjustment at time \( u \) is

\[
-X_u = k_0 + \frac{1}{g} \sum_{i=1}^{u} (\hat{Z}_i - \hat{Z}_{i-1})
\]

But \( \hat{Z}_i - \hat{Z}_{i-1} = \gamma e_i \). After setting \( \gamma/g = k_1 \) this gives

\[
-X_u = k_0 + k_1 \sum_{i=1}^{u} e_i
\]

which is the minimum mean square error feedback equation (17). Notice that it provides minimum cost of regulation only if the cost of adjustment \( C_a \) is negligible. Control action depends on the EWMA of the disturbance \( Z \) or equivalently the cumulative sum of the output errors \( e \). This is not, of course, the same as regulation based on a Cusum of the disturbance \( Z \) itself, as has sometimes been proposed.

(c) \( C_a \) is negligible but \( m \) is not fixed: In this case a feedback scheme is obtained in which limit lines are on the target so that adjustments are made after each observation. But the optimal monitoring interval may be greater than \( m = 1 \). Consider again the numerical example we used earlier where it was supposed that the cost of adjustment \( C_a = 600 \). If instead, this cost had been negligible, the value of \( \lambda \) would have been zero but the value of \( m \) obtained from Chart B would still be approximately equal to ten so that adjustments would be made after each observation taken ten unit intervals apart.

(d) \( C_a \) is not negligible, \( m \) is fixed: In this case action limits, \( T \pm \lambda \gamma \sigma \) are determined directly by \( R_a \). This was the special case considered by Box and Jenkins (1963). Adjustments, when needed, are made to bring the predicted value to the target value.

(e) \( \gamma = 1 \): This case could theoretically occur and indeed is the assumption made by Taguchi (1981) in his derivation of control schemes. However, this degree of non-stationarity is so extreme (meaning, for example, that the variance doubles after only two intervals of time) that it can hardly be regarded as describing any control situation likely to be met in reality. Even with this random walk assumption Adams and Woodall (1989) have shown that Taguchi's approximations are inadequate and can lead to suboptimal schemes.
5. Conclusions

Statistical process control is a tool whose primary uses are to continuously monitor the common cause system and to detect significant deviations possibly pointing to special (assignable) causes. It appropriately employs considerations of statistical significance to trigger action and so reduces the chance of fruitless pursuit of phenomena produced by chance alone.

Automatic process control using feedforward and feedback systems is a tool primarily for process regulation to compensate for the effect of non-stationary disturbance. Such schemes should be designed to optimize some desirable criterion, for example, to minimize mean square error about the target value or better still to minimize some measure of cost. For the purpose of pure regulation questions of statistical significance are irrelevant, but when automatic control systems are poorly designed or poorly tuned, it is possible for them to increase rather than reduce variation about the target value.

An optimal scheme can be designed by modelling the process dynamics and the disturbance at the output, but this approach can be tedious. Alternatively a standard control system, such as a proportional-integral feedback system can be optimized by using experimental design. This can be particularly useful when there are many controllers operating on a process.

APC is often operated so that the nature of the disturbance which is being compensated is concealed and unusual deviations from the target cannot be taken into account. It is pointed out that this need not happen. Records of applied adjustments and deviations from target can be displayed. Also, provided that the dynamics of the system are known, the actual compensation applied, and the disturbance which is being compensated, can be reconstructed and examined.

It is useful to extend the idea of common causes and special causes to automatic control systems. In particular it is shown how study of the nature of the disturbance in a feedback control scheme could lead to the (common cause) system being changed by management. Furthermore, study of deviations from target for an optimal operating system could point to special causes.

The choice of an optimal system for regulation depends on costs as well as on the nature of the disturbance and the dynamics. For illustration, simple models are proposed for off-target cost, the cost of adjustment and the monitoring cost and charts are derived that allow optimal schemes to be chosen for a variety of circumstances. In particular it is shown that feedback control in which adjustments are applied at every opportunity is only optimal if the monitoring cost and the cost of adjustment are negligible. When the cost of adjustment is not negligible, schemes which superficially look much more like a Shewhart
chart or an EWMA chart are obtained but with the limit lines decided by relative costs and not by questions of statistical significance. When the cost of monitoring the process is appreciable the optimal procedure may require the process to be looked at less frequently.
Appendix A: **Linear Increase in the Interval Variance Implies the Non-stationary Model of Equation (5)**

Suppose* \( E(Z_{t+m} - Z_t) = 0 \) and the interval variance \( I_m = V(Z_{t+m} - Z_t) \) exists for all \( t \) and \( m \). Write \( Z_{t+1} - Z_t = W_t \), \( C_i = E(W_{t+i} W_t) \), \( \rho_i = \frac{C_i}{C_0} \).

Then \( G(m) = \frac{I_m}{I_1} = m + 2(m-1) \rho_1 + \ldots + 2 \rho_{m-1} \)

which is of the form \( G(m) = 1 + \beta (m-1) \) only if

\[
1 + 2 \rho_1 = \beta \quad \text{and} \quad \rho_i = 0 \quad \text{for} \quad i = 2, 3, \ldots \infty.
\]

This implies the model is of the form of (11) and hence of (5).

In particular, if \( Z_t \) is white noise \( \rho_1 = -\frac{1}{2} \quad \beta = 0 \quad G(m) = 1 \)

if \( Z_t \) is a random walk \( \rho_1 = 0 \quad \beta = 1 \quad G(m) = m \)

---

*The argument is readily extended to a model with a linear trend so that \( E(Z_{t+m} - Z_t) = \mu m \).
Appendix B: Inflation of Variance Due to Mistuning of a Simple Feedback Controller

For the feedback scheme of equation (17) with the disturbance represented by equation (5) and the process dynamics by equation (16), to achieve minimum mean square error, we need to set the controller constant \( k_I = \gamma / g \), in which case \( \sigma_c^2 / \sigma_a^2 = 1 \). But suppose that \( k_I \) is set equal to some value \( k'_I \) which is not necessarily equal to \( k_I \), then

\[
e_t = Z_t + Y_t = a_t + \gamma \sum_{i=1}^{t-1} a_i - g k_I \sum_{i=1}^{t-1} e_i
\]  
(B.1)

After setting \( \phi = 1 - g k_I \) and differencing, equation (B.1) implies that

\[
e_t - \phi e_{t-1} = a_t - \theta a_{t-1} \quad \text{where } \theta = 1 - \gamma \quad \text{and } 0 \leq \theta \leq 1
\]  
(B.2)

For this autoregressive - moving average process

\[
\sigma_c^2 / \sigma_a^2 = \begin{cases} 
1 + \theta^2 - 2\phi \theta & -1 < \phi \leq 1 \\
-\phi^2 & \text{otherwise}
\end{cases}
\]

(B.3)

This shows the degree of inflation of the variance, due to mistuning, for any desired values of \( \gamma = 1 - \theta \), \( g \), and \( k_I = (1 - \phi) / g \).

In the special case discussed in Section 3.1 of the paper, \( Z_t \) is white noise so that \( \gamma = 0 \), \( \theta = 1 \) and

\[
\sigma_c^2 / \sigma_a^2 = \begin{cases} 
\frac{2}{1 + \phi} & -1 < \phi \leq 1 \\
\infty & \text{otherwise}
\end{cases}
\]

(B.4)

whence equation (18) follows.
Appendix C: Minimal Cost Regulation: An Outline of the Argument

Suppose that a process which is subject to the disturbance model (5) is observed at regular intervals. Suppose further that adjustments can be made at a cost $C_a$ which can bring the observed process to a desired target, $T$, within one interval. Finally, suppose that there is a cost associated with being off-target that is proportional to the square of the deviation with proportionality constant $k_T$.

Now, if the process is going to make just one additional unit of material, then the expected cost when an adjustment is made is

$$k_T E(e_{t+1}^2) + C_a = k_T E(a_{t+1}^2) + k_T(\hat{Z}_{t+1} + Y_{t+1})^2 + C_a$$

which is minimized by setting the compensating variable, $Y_{t+1}$, so that it negates the predicted disturbance, $\hat{Z}_{t+1}$. When this is done, the expected cost for adjustment is $k_T \sigma_a^2 + C_a$. If no adjustment is made, then the expected cost is

$$k_T E(Z_{t+1} + Y_{t+1})^2 = k_T \sigma_a^2 + k_T(\hat{Z}_{t+1} + Y_{t+1})^2$$

Hence, an adjustment should be made if

$$|\hat{Z}_{t+1} + Y_{t+1}| \geq \left(\frac{C_a}{k_T}\right)^{1/2} \equiv L_1$$

One can show that if $N$ additional items are to be made, then the process should be adjusted if

$$|\hat{Z}_{t+1} + Y_{t+1}| \geq L_N$$

where $L_N$ is a function of the adjustment cost, the off-target proportionality constant, and the disturbance parameters. The values $L_1, L_2, \ldots$ form a sequence which has a limiting value $L$. Therefore, the optimal strategy for an infinite horizon is to adjust the process when the forecasted observation deviates from target by an amount $L$.

Suppose now that the process is observed at a multiple $m$ of the unit interval and associated with each observation is a cost $C_m$. The above logic holds once again and the minimal cost strategy is to again adjust when the forecasted observation deviates from target by an amount $L(m)$ where the notation indicates that the limit will depend on the
monitoring interval as well as the costs and disturbance parameters. It can be shown that
the expected cost per run (interval between adjustments) is then

\[ C_a + \text{ARL} \frac{C_m}{m} + k_T \sum_{i=1}^{n} e_i^2 \]  \hspace{1cm} (B.5) 

whence the cost per interval is

\[ \frac{C_a}{\text{ARL}} + \frac{C_m}{m} + k_T \sum_{i=1}^{n} \frac{e_i^2}{\text{ARL}} \]  \hspace{1cm} (B.6) 

For a disturbance with parameters \((\gamma, \sigma_a^2)\), monitoring interval \(m\), and limits at
\(\pm L = \lambda \gamma \sigma_a\), this cost can be expressed as

\[ \frac{C_a}{m \cdot h(\lambda)} + \frac{C_m}{m} + k_T \frac{\theta}{\theta_m} \sigma_a^2 + m k_T \gamma^2 \sigma_a^2 g(\lambda) \]  \hspace{1cm} (B.7) 

where \(g(\lambda)\) and \(h(\lambda)\) are both functions solely of \(\lambda\) which relate to the expected
squared deviation from target and ARL, respectively. The values \(m\) and \(\lambda\) are
determined which minimize this equation.

This last equation may be scaled by dividing by \(k_T \gamma^2 \sigma_a^2\) to give

\[ \frac{R_a}{m \cdot h(\lambda)} + \frac{R_m}{m} + \frac{\theta}{\gamma^2 \theta_m} + mg(\lambda) \]  \hspace{1cm} (B.8) 

This form shows that the choice of \(\lambda\) and \(m\) which minimizes the total cost can be
obtained as functions of \(R_a\), \(R_m\), and \(\gamma\).
References


