DEPARTMENT OF STATISTICS

University of Wisconsin
Madison, Wisconsin 53706

TECHNICAL REPORT NO. 310
August 1972

SOME RECENT ADVANCES IN FORECASTING AND CONTROL PART II

by

G. E. P. Box
G. M. Jenkins
J. F. MacGregor

Typist: Bernice R. Weitzel

This research was supported by the Air Force Office of Scientific Research under Grant AFOSR-72-2363.

SOME RECENT ADVANCES IN FORECASTING AND CONTROL

Part II

G. E. P. Box, G. M. Jenkins and J. F. MacGregor

1. Introduction

In Part I of this paper [6] (we apologize for the delay in presenting this final part) we presented a class of discrete time series and dynamic models together with the theory for identifying, fitting, and checking them. The principal application which was discussed there was to forecasting. In this second part we shall rely heavily on these models. We shall outline briefly an approach to control which is discussed in much more generality and detail in the papers [2, 3, 4, 5, 6] and a book [7]. Opportunity is also taken to correct a mistake which occurred in references [5,7] concerning optimal feedforward control.

In the past, the word "control" has usually meant to the statistician the quality control techniques developed originally by Shewhart [14] in the United States and by Dudding and Jennet [8] in Great Britain. Recently, the sequential aspects of quality control have been emphasised, leading to the introduction of cumulative sum charts by Page [11,12] and Barnard [1] and the geometric moving average charts by Roberts [13].

The word control has a different meaning to the control engineer. He thinks in terms of feedforward and feedback control, of the dynamics and stability of the system, and usually of particular types of hardware to carry out the control action. The control devices are automatic in the sense that information is fed to them automatically from instruments on the process and from them to adjust automatically the inputs to the process.

The control techniques discussed here are, at least from the point of view of motivation, closer to those of the control engineer than the standard quality control procedures developed by statisticians. This does not mean we believe that the traditional quality control chart is unimportant but rather that it usually performs a different function from that which we are here concerned. However, becuase there are certain analogies between the two ideas and because of some rather dubious justifications of control charts we start with a discussion of these.

2. Quality Control Charts

Suppose that observations on an industrial process are being made at equispaced intervals of time to produce a time series $\{z_t\}$. Then as Shewhart [14] pointed out, it is an excellent idea to plot the data as it comes to hand on a chart which shows the target value T. Such a plotting procedure (i) can provide timely warning of a deviation from target and of possible need for corrective action (ii) can provide clues as to possible assignable causes of variation which may subsequently be eliminated or compensated for. The first is a form of feedback control and the second a form of process improvement.

To assist judgment Shewhart introduced control lines at 2σ and 3σ limits. On the assumption that the time series $\{z_t\}$ was generated by a stochastic process in which successive deviations from the fixed mean μ were Normally and independently distributed with fixed known variance σ^2 then, when the mean μ was in fact equal to the target value T, the probability would be small that a value would lie outside the limit lines. If, however, the mean deviated from T then the probability of a point lying outside the limit lines would be larger and could readily be calculated. If the referral of a point to such limit lines is thought of as a test of the hypothesis that $\mu = T$ then, given the above assumptions, such a test would not be very powerful. Indeed, since the introduction of sequential tests by Wald and Barnard during the second World War, it has been known that, again on the assumptions outlined above, a procedure of much greater power uses the cumulative sum of the deviations $\sum_{j=1}^{t} (z_j - T)$. Following proposals by Page [11, 12] and by Barnard [1] in the 1950's charts which used this cumulative sum have been introduced with considerable success into industry.

An important assumption in the above is that the observations are independently distributed. Now industrial data occurring serially are very likely not independent but serially correlated. Suppose, for example, the observations were generated by a first order autoregressive process defined by

$$\tilde{y}_t - \phi\tilde{y}_{t-1} = a_t$$

where $\tilde{y}_t = y_t - \mu$ and $-1 < \phi < 1$ and $\{a_t\}$ is a sequence of independent random Normal variables having mean zero and variance σ^2. The appropriate sequential likelihood ratio statistics is

$$\phi\{y_t-y_0\} + (1-\phi)\{\sum_{j=1}^{t}(y_j-T)\}$$

If ϕ is not too close to one, and for moderate t, this expression is dominated by the cumulative sum in the second bracket. One might expect, therefore, that cumulative sum techniques would be robust to serial correlations of this kind as has been specifically demonstrated for instance by Goldsmith and Whitfield [9]. However, it is also true that if ϕ were equal to unity (the disturbance was a random walk), the first term would dominate, and assuming the initial value y_0 close to target one would essentially be back to plotting

$$y_t-T$$

the "Shewhart statistic."

However, an alternative approach is not to start from a procedure at all, a cusum or Shewhart chart or whatever, but start by identifying (i) a stochastic process which represents the actual behavior of the industrial series and (ii) the objectives which we desire our control or forecasting scheme to achieve. We shall then be led to whatever procedure is appropirate for the relevant circumstances.

Quality control charts have been used for a wide variety of purposes. It has been occasionally suggested [15] that they can be used to control a process in the feedback sense, and more specifically that adjustment only be made if a "significant" deviation is observed. Such a use implies important assumptions for the desired objectives. For optimal process control we should not need to be convinced of the "reality" of the difference to persuade us to action. It is enough that a policy will lead to a desirable objective such as minimization of costs or in some instances of the mean square deviation from target. To justify a "test like" procedure one normally needs the requirement that some cost is involved in making an adjustment. Which test procedure is then appropriate would depend upon our assumptions as to other costs involved and stochastic process followed by the actual industrial series. For example Roberts [13] proposed a geometric moving average chart but assumed that the observed variable y_t was a stationary white noise process centered about some mean and used this assumption to calculate the standard deviation of the geometric mean. In fact, if the process were of the kind Roberts assumes the geometric mean would be inappropriate. As we shall see later a chart like that which Roberts found intuitively sensible can be justified on a list of assumptions involving the type of stochastic process being controlled, the loss due to off-target material, and the cost of adjustment.

3. Process Control

The process control schemes we shall now discuss are appropriate for the periodic, optimal adjustment of a manipulated variable, whose effect on some quality characteristic is already known. They are designed to minimize the variation of that quality characteristic about some target value. We assume that data is available at discrete equispaced time intervals when opportunity can also be taken to make adjustments. It is assumed also that the situation commonly met, for example in the chemical and process industries, is where surveillance (by an operator or a computer) is needed in any case and so no appreciable cost is associated with corrective action.

The reason control is necessary at all is that there are inherent disturbances or noise in the system. When we can measure these disturbances directly the making of appropriate compensatory changes in some other variable to undo their effect is referred to as feedforward control. Alternatively or in addition making use of the deviation from target or "error signal" of the output (quality) characteristic itself to calculate appropriate compensatory changes is referred to as feedback control. Feedback control can be employed even when the source of the disturbances is not accurately known or their magnitude measured. More generally feedforward control can be used to compensate for those disturbances that can be measured and feedback control to compensate for the remainder.

The approach adopted is to typify the disturbances by a suitable time series or stochastic model and the inertial characteristics of the system by a suitable transfer function model. It is then possible to calculate a control equation which produces the smallest mean square error at the output possibly subject to a constraint on the variance of the manipulated variable. Execution of the control action can then be accomplished at various levels of technical sophistication - by a digital computer linked directly to the process, by a pneumatic or electronic automatic controller, or by manual manipulation by an operator using a suitable chart or nomogram.

3.1 Feedback Control

We introduce the ideas of feedback control in terms of a simple yet real example from the chemical industry. In a scheme to control the viscosity Y of a polymer employed in the manufacture of a synthetic fiber, the controlled variable, viscosity, was checked every hour and adjusted by manipulating the catalyst formulation X. The desired target value for viscosity was 47 units. It was found that the dynamic characteristics relating Y and X were such that essentially all the effect of X on Y occurred within the one hour sampling interval. The transfer function model (see Part I [6] for a discussion of discrete dynamic and stochastic models) was therefore of the form

$$\dot{Y}_t = g\dot{X}_{t-1} \qquad (1)$$

where $\dot{Y}_t$ and $\dot{X}_t$ are deviations from equilibrium values. The catalyst formulation changes were, by custom, scaled in terms of the effect they were expected to produce. Thus one unit of formulation increase was such as would decrease viscosity by one unit. Hence $g = -1.0$.

In order to design a feedback controller, it is also necessary to identify and fit by a suitable stochastic model the disturbance N_t in the output. Here N_t represents the joint effect in the viscosity measurement of all unobserved disturbances occurring in the process and is defined as the deviation from target viscosity that would occur at time t if no control action were taken. This was found to be adequately described by a time series model of order (0,1,1)

$$\nabla N_t = (1-\theta B)a_t \tag{2}$$

where $\theta = 0.53$ and a_t is a sequence of random shocks with mean zero and variance σ_a^2 (see Part I [6]). This stochastic model is non-stationary in its level (it has a tendency to drift) and its minimum mean square error forecast ℓ-steps ahead is the well known exponentially weighted moving-average of previous observations

$$\hat{N}_t(\ell) = \theta\hat{N}_{t-1}(\ell) + (1-\theta)N_t = (1-\theta)\sum_{j=0}^{\infty}\theta^j N_{t-j} \tag{3}$$

The aim of a feedback controller is to compensate for this disturbance in the output viscosity by making suitable changes in X_t. The total effect on the output viscosity at time t of the disturbance is N_t and of any compensatory action is gX_{t-1}. Thus the effect of the disturbance would be cancelled if it were possible to set $X_t = -\frac{1}{g}N_{t+1}$. Since N_{t+1} has not yet been observed, this is not possible, but we can obtain the minimum mean square control error by replacing N_{t+1} by its forecast $\hat{N}_t(1)$, that is, by taking control action

$$X_t = -\frac{1}{g}\hat{N}_t(1)$$

or in terms of the adjustment to be made $(x_t = \nabla X_t = X_t - X_{t-1})$

$$x_t = -\frac{1}{g}\{\hat{N}_t(1)-\hat{N}_{t-1}(1)\} \tag{4}$$

Using the forecasting theory in Part I [6] it can easily be shown that for this disturbance model (2) the updating expression for the forecasts is given by

$$\hat{N}_t(1) - \hat{N}_{t-1}(1) = (1-\theta)a_t \tag{5}$$

and therefore the optimal control action becomes

$$x_t = -\frac{(1-\theta)}{g} a_t$$

With this adjustment the error in the output viscosity ε_t will simply be equal to the one step ahead forecast error a_t, and so we can write the optimal adjustment as

$$x_t = -\frac{(1-\theta)}{g} \varepsilon_t = 0.47 \varepsilon_t \tag{6}$$

This controller is worthy of further discussion. As each new observation ε_t becomes available we are making an adjustment to the input which corrects for our change in the forecast of the disturbance. Recalling that a_t is the one step ahead forecast error, we can see that the updating equation (5) implies that having seen that our previous forecast $\hat{N}_{t-1}(1)$ falls short of the realized value N_t by a_t, we adjust it by an amount $(1-\theta)a_t$ which the model says from past experience is the amount of any shock which is permanently absorbed into the "level" of the process. By adding only $(1-\theta)a_t$ to the new forecast we are getting the correct balance between a control action which is too conservative and adds too little correction at each stage and one which is too quick and overcorrects by adding the full discrepancy a_t at each stage. The value of θ is indicating the correct state of conservatism between these two extremes. It can also be noted from the exponentially weighted average form for the forecast (3) that the new forecast is a linear interpolation at argument $(1-\theta)$ between the old forecast and the new observation. If $(1-\theta)$ is equal to one the evidence from past data is completely ignored and the forecast for all future time is the current value $(\hat{N}_t(\ell) = N_t)$, in which case the control action is simply $x_t = -\frac{1}{g}\varepsilon_t$. But if $(1-\theta)$ is less than one the control action (6) discounts the present shock by an amount $(1-\theta)$.

The efficiency of control action of this kind is insensitive to moderate changes in parameter values and to a sufficient approximation we can take (6) to be

$$x_t = 0.5\ \varepsilon_t$$

A convenient chart for use when, as in this example, manual control action is employed, is shown in Figure 1. On this chart the output (viscosity) scale and the action scale are arranged so that the output target is aligned with zero action, and so that one unit of output is matched by $-\frac{(1-\theta)}{g}$ units of action. To employ the chart, the plant operator simply plots the latest output (viscosity) value and reads off the appropriate adjustment on the action scale.

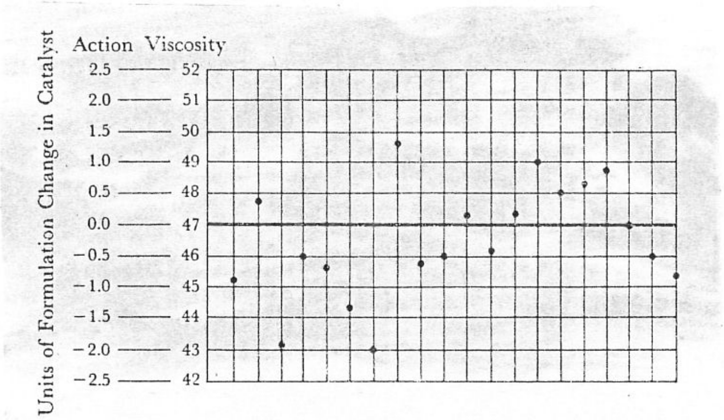


Figure 1. A feedback chart for integral control action

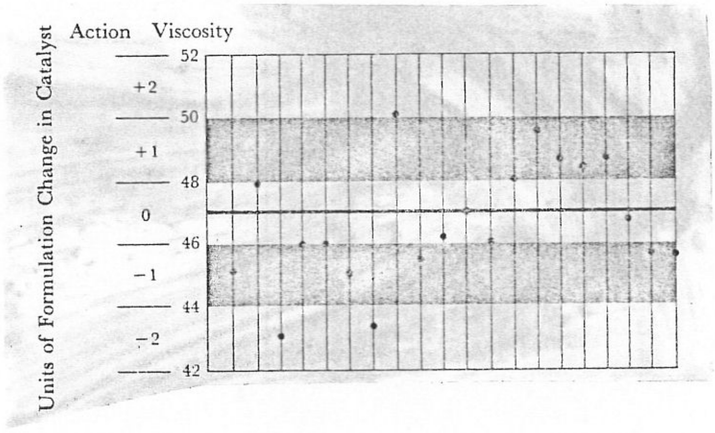


Figure 2. Feedback chart for rounded integral control action

In terms of the actual catalyst level $\dot{X}_t$ at the input the control equation (6) becomes

$$\dot{X}_t = 0.47 \sum_{j=1}^{t} \varepsilon_j \qquad (7)$$

This simple result is also worthy of further discussion. For many years continuous time controllers have been used which base the control action empirically on a mixture of proportion, integral, and derivative control. For instance if $\varepsilon(t)$ were the continuous deviation of the output from target, the level of the input $X(t)$ would be calculated by

$$X(t) = k_D \frac{d\varepsilon(t)}{dt} + k_P\varepsilon(t) + k_I\int\varepsilon(t)dt$$

where k_D, k_P, and k_I are constants. In some situations only one or two of these three modes of action are used. The discrete analogue of this continuous control equation is

$$X_t = k_D\nabla\varepsilon_t + k_P\varepsilon_t + k_I\sum\varepsilon_t$$

It will therefore be seen that the above control action (7) which we have just derived is simply the discrete analogue of integral control action. This appears similar to a cummulative sum procedure although it is important to notice that it is not accummulating the deviations from target when no control action is being taken but rather it is accummulating the deviations from target when action <u>is</u> being taken at every interval. This it can be shown is equivalent to taking an exponentially weighted moving-average of the past disturbance effects N_t, N_{t-1}, N_{t-2}, ... At the same time it should also be noted that this is not the same type of control chart recommended by Roberts [13] again because control action is being taken at every interval.

Although the feedback control chart of Figure 1 is extremely simple it is capable of further simplication. On this particular process, control had previously been carried out using a chart based somewhat arbitrarily on a sequential significance testing scheme. It had turned out in this connection that it was convenient to add or subtract from the catalyst formulation in standard steps. Possible actions were: no action, ±one step, or ±two steps of catalyst formulation.

Significance testing procedures have little relevance in the present context. However, the previous scheme did have the advantages (i) that it had not been necessary to make changes every time and (ii) when changes were called for they were of one of five definite types, making the procedure easy to apply and supervise. However, these features can easily be included in the present control scheme, with very little increase in the error, by using a "rounded" action chart.

A rounded chart is easily constructed from the original chart by dividing the action scale into bands. The adjustment made when an observation falls within the band is that appropriate to the middle point of the band on an ordinary chart. Figure 2 shows a rounded chart in which possible action is limited to -2, -1, 0, 1, or 2 catalyst formulation changes. Figures 1 and 2 have been constructed by back calculating the values of a_t from a set of operating data and reconstructing the charts that would have resulted from using an unrounded and a rounded scheme. The increase in mean square error (less than 5% for this example), which results from using the rounded scheme, is often outweighed by the convenience of working with a small number of standard adjustments.

Consider now the case of some higher order dynamic models (see Part I [6]). A first order dynamic transfer function model is of the form

$$(1 + \xi\nabla)\dot{Y}_t = g\dot{X}_{t-1}$$

where ∇ is the difference operator $(\nabla\dot{Y}_t = \dot{Y}_t-\dot{Y}_{t-1})$. If we were to again let the disturbance model at the output be represented by the integrated-moving-average (0,1,1) model

$$\nabla N_t = (1-\theta B)a_t$$

then we would find that the optimal feedback controller (that which minimizes the mean square error at the output) is

$$\dot{X}_t = - \frac{(1-\theta)}{g} \{\xi\varepsilon_t + \sum_{j=1}^{t} \varepsilon_j\}$$

which is the discrete analogue of proportional-integral control with the ratio of the amount of proportional to integral action being given by the dynamic parameter ξ. Similarly with the same disturbance model but second order dynamics we would find that the optimal controller is the discrete analogue of proportional-integral-derivative control. However, these are by no means the only kind of control that this procedure can give and a general development of the above type of minimum mean square error control theory has been given elsewhere [7].

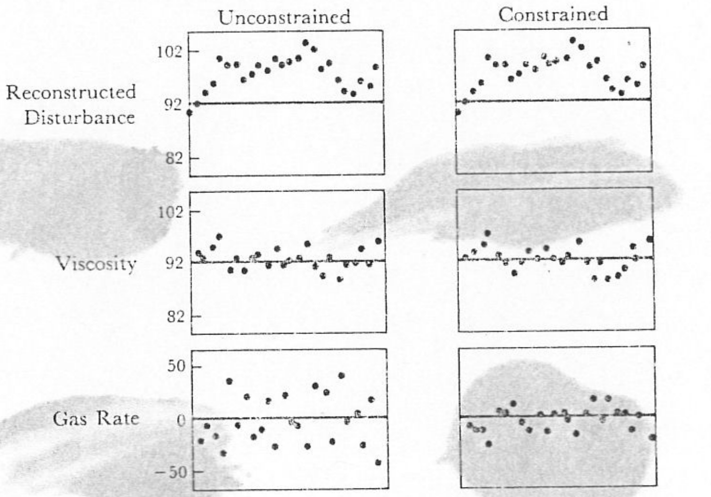
Now there are situations particularly when the dynamics are slow compared with the sampling rate where the minimum mean square error control gives impossibly large variations in the input X_t. It is then possible to introduce a constrained controller in which the mean square error of the output deviation from target ε_t is minimized subject to a constraint on the variance of the input. The remarkable feature of these constrained controllers is that by allowing only a very small increase in the mean square error of the output a very large reduction can usually be made in the variance of the input. An example of this, the details of which are described more fully in [7], is as follows. In a scheme to control the viscosity of the product of a chemical reaction by varying the input gas rate the minimum mean square error control action was found to be given by

$$x_t = -10.(\varepsilon_t - 0.5\varepsilon_{t-1}) \qquad (8)$$

If the variations in x_t as a result of this scheme were unacceptably large then a constrained scheme could be used. In particular for a 10% increase in the standard deviation of the output, the standard deviation of the input can be halved by using the control action

$$x_t = 0.15x_{t-1} - 5.5(\varepsilon_t - 0.5\varepsilon_{t-1}) \qquad (9)$$

Figure 3 illustrates this point. A set of twenty-four successive observations showing the values of inputs (gas rate) and outputs (viscosity) are reproduced in the left-hand diagrams as they were actually recorded using the optimal unrestricted scheme (8). Also shown is the reconstructed noise. Supposing the scheme to be initially on target, this reconstructed noise is the computed drift away from target that would have occurred if no control action had been taken. The right-hand diagrams show the calculated behavior that would have occurred with the same noise if the constrained scheme of equation (9) had been used. Further information on the design of these optimal constrained schemes is given elsewhere [7,10,16].



Control Equations

$x_t = -10\,(\varepsilon_t - 0.5\,\varepsilon_{t-1})$ $x_t = 0.15\,x_{t-1} - 5.5\,(\varepsilon_t - 0.5\,\varepsilon_{t-1})$

FIG. 3 Behavior of unconstrained and constrained control schemes for viscosity/gas rate example

3.2 Feedforward Control

Sometimes one or more major sources of disturbance can be located and measured. In feedforward control these measurements are used to calculate compensatory action which forestalls the effect of these disturbances on the output.

A situation arising in the manufacture of a polymer is illustrated in Figure 4.

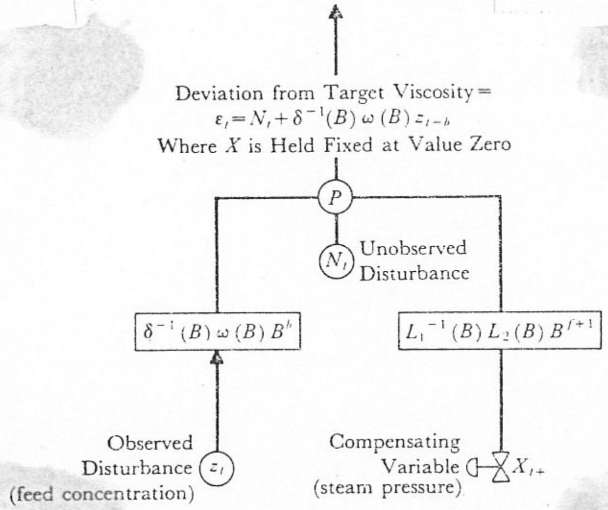

Deviation from Target Viscosity=
$\varepsilon_t = N_t + \delta^{-1}(B)\,\omega(B)\,z_{t-b}$
Where X is Held Fixed at Value Zero
P
N_t
Unobserved Disturbance
$\delta^{-1}(B)\,\omega(B)\,B^b$
$L_1^{-1}(B)\,L_2(B)\,B^{f+1}$
Observed Disturbance
z_t
(feed concentration)
Compensating Variable
X_{t+}
(steam pressure)

FIG. 4. A system at time t subject to an observed disturbance z_t and unobserved disturbance N_t, with potential compensating variable X_t held fixed at $X_t = 0$

The viscosity Y_t of product is known to vary in part due to fluctuations in the feed concentration z_t which can be observed but not adjusted. The steam pressure X_t is a control variable which is measured, can be manipulated, and is potentially available to alter the viscosity by any desired amount and hence compensate potential deviations from target. The total effect in the output of all other sources of disturbance at time t is denoted by N_t.

We shall first consider the general solution. The transfer function model which connects the observed disturbance z_t (feed concentration) and the output Y_t (viscosity) is assumed to be

$$Y_t = \delta^{-1}(B)\omega(B)B^b z_t$$

Changes will be made in X at times $t, t-1, t-2, \ldots$ immediately after the observations $z_t, z_{t-1}, z_{t-2}, \ldots$ and then held constant over the interval as in the feedback case. It is assumed that the transfer function model which connects the compensating variable X_t (steam pressure) and the output (viscosity) is

$$Y_t = L_1^{-1}(B)L_2(B)B^{f+1}X_t$$

where f is the number of whole periods of pure delay. Then if no control is exerted (X_t is held fixed at $X_t = 0$), the total error in the output viscosity will be

$$\varepsilon_t = N_t + \delta^{-1}(B)\omega(B)z_{t-b}$$

Clearly, it ought to be possible to compensate the effect of the measured parts of the overall disturbance by manipulating X_t. Now at time t and at point P in Figure 4:

(1) The total effect of the disturbance (z) is $\delta^{-1}(B)\omega(B)z_{t-b}$

(2) The total effect of the compensation (X) is $L_1^{-1}(B)L_2(B)X_{t-f-1}$

Then the effect of the observed disturbance will be cancelled if we set

$$L_1^{-1}(B)L_2(B)X_{t-f-1} = - \delta^{-1}(B)\omega(B)z_{t-b}$$

Thus the control action at time t should be such that

$$L_1^{-1}(B)L_2(B)X_t = - \delta^{-1}(B)\omega(B)z_{t-(b-f-1)} \qquad (10)$$

Case 1: $b \geq f+1$. At time t, the values $z_{t+1}, z_{t+2}\ldots$ are unknown. The control action (10) is directly realizable then only if $(b-f-1) \geq 0$ in which case the desired control action at time t is to set the manipulated variable X_t to the level

$$X_t = - \frac{L_1(B)\omega(B)}{L_2(B)\delta(B)} z_{t-(b-f-1)}$$

With this control action the component at the output (point P in Figure 4) of the deviation from target due to z_t is (theoretically at least) exactly eliminated at the observation times, and only the component N_t due to unobserved disturbances remains.

Case 2: $b < f+1$.* It can happen that $f+1 > b$. This means that an observed disturbance reaches the output before it is possible for compensating action to become effective. In this case the action of equation (10) is not realizable because at time t, when action is to be taken, the relevant value $z_{t-(b-f-1)}$ of the disturbance is not yet available. One would usually avoid this situation if one could (if for example some quicker

*The solution given for this case was incorrect in the first two printings of Box and Jenkins book Time Series Analysis, Forecasting and Control [7].

acting compensating variable could be used), but sometimes such an alternative is not available.

If the disturbance z_t can be represented by the linear stochastic model

$$\Phi(B)z_t = \theta(B)a_t \tag{11}$$

then we can express $z_t' = \delta^{-1}(B)\omega(B)z_t$ as the stochastic model

$$\Phi'(B)z_t' = \theta'(B)a_t$$

where $\Phi'(B) = \Phi(B)\delta(B)$ and $\theta'(B) = \theta(B)\omega(B)$ and $\{a_t\}$ is the same white noise sequence as in (11). This can be equivalently expressed in the form

$$z_t' = \{1 + \sum_{i=1}^{\infty} \psi_i B^i\} a_t$$

Then

$$z'_{t+f+1-b} = \hat{z}_t'(f+1-b) + e_t'(f+1-b)$$

where in this expression

$$e_t'(f+1-b) = a_{t+f+1-b} + \psi_1 a_{t+f-b} + \dots + \psi_{f-b} a_{t+1}$$

is the $(f+1-b)$-step ahead forecast error and $\hat{z}_t'(f+1-b)$ is the forecast.

Then we can write equation (10) in the form

$$L_1^{-1}(B)L_2(B)X_t = - \hat{z}_t'(f+1-b) - e_t'(f+1-b)$$

Now $e_t'(f+1-b)$ is a function of the uncorrelated random deviates a_{t+h} $(h \geq 1)$ which have not yet occurred at time t and which are uncorrelated with any variable known at time t (and are therefore unforecastable). It follows that the optimal action is achieved by setting

$$X_t = - \frac{L_1(B)}{L_2(B)} \hat{z}_t'(f+1-b) \qquad (12)$$

or by making a change in the compensating variable at time t equal to

$$x_t = - \frac{L_1(B)}{L_2(B)} \{\hat{z}_t'(f+1-b) - \hat{z}_{t-1}'(f+1-b)\}$$

The needed forecast $\hat{z}_t'(f+1-b)$, obtained as in Part I of this paper [6], can then be written conveniently in terms of the previous z_t's (viscosity measurements) and a_t's. (Recall that these a_t's are common to both the z_t' and z_t series).

This control scheme results in an additional component in the deviation ε_t from the target, which now becomes

$$\varepsilon_t = N_t + e_{t-f-1}'(f+1-b)$$

Note that in both cases of feedforward control the output deviation ε_t from target still includes the disturbance N_t representing the effect in the output at time t of all other disturbances in the system. These can often be substantial and in particular can result in uncontrolled drift and so feedback control will often have to be applied simultaneously to the output.

An example: In the manufacture of an intermediate product used for the production of a synthetic resin, the specific gravity Y_t of the product had to be maintained as close as possible to the value 1.260. The feed concentration z_t was observable but contained an uncontrollable disturbance and so was to be fed forward. The dynamic relationship between specific gravity and feed concentration over the range of normal operation is

$$(1 - 0.2B)Y_t = 0.0016z_t$$

In our general notation $\delta(B) = (1 - .2B)$, $\omega(B) = 0.0016$, and $b = 0$. Control is achieved by varying pressure X_t. The transfer model relating specific gravity and pressure was estimated as

$$(1 - 0.7B)Y_t = 0.0024X_{t-1}$$

so that $L_1(B) = (1 - 0.7B)$, $L_2(B) = 0.0024$, and $f = 0$. So far as could be ascertained the effects of pressure and feed concentration were approximately additive in the region of normal operation. Therefore equation (12) is used, since $(b\text{-}f\text{-}1) < 0$, yielding as the optimal action

$$X_t = - \frac{(1 - 0.7B)}{0.0024} \hat{z}'_t(1)$$

Study of the feed concentration showed that it could be represented by the linear stochastic model of order (0,1,1)

$$\nabla z_t = (1 - \theta B) a_t \tag{13}$$

with $\theta = 0.5$. Therefore $z'_{t+1} = \delta^{-1}(B)\omega(B) z_{t+1}$ is given by

$$z'_{t+1} = \frac{\omega_0(1-\theta B)}{(1-\delta B)(1-B)} a_{t+1}$$

This can be rearranged into the form

$$z'_{t+1} = \omega_0 a_{t+1} + \frac{\omega_0[(1+\delta-\theta) - \delta B]}{(1-\delta B)(1-B)} a_t \tag{14}$$

$$= e'_t(1) + \hat{z}'_t(1)$$

Combining equations (13) and (14) the one-step ahead forecast can be expressed in terms of the observable feed concentration measurements (z_t) as

$$\hat{z}'_t(1) = \frac{\omega_0[(1+\delta-\theta) - \delta B]}{(1-\delta B)(1-\theta B)} z_t$$

Hence the optimal feedforward control is given by

$$X_t = -\frac{(1-0.7B)}{0.0024}\ \frac{0.0016(0.3-0.2B)}{(1-0.2B)(1-0.5B)}\ z_t$$

or

$$X_t = 0.7X_{t-1} - 0.1X_{t-2} - 0.2\{z_t - 1.37z_{t-1} + .47z_{t-2}\}$$

This control action requires the storage of the three most recent measurements of feed concentration (z) and the past two settings of the manipulated pressure level (X). It is easily handled by a mini-computer if direct digital control is being used or by the use of a nomogram or a small programable desk calculator in the case of manual control.

Further aspects of this approach to feedback and feedforward control are discussed in [7,10]. Among the topics considered there are combined feedforward-feedback control, schemes in which the variance of the manipulated variable is constrained, the choice of sampling interval, and the fitting of transfer function-disturbance models from closed-loop operating data.

4. Control in the Parts Manufacturing Industry

We now would like to consider a control problem more typical of situations arising in the mass producing industries where typically a machine may be mass producing components (such as ball-bearings) and some quality characteristic (such as weight or diameter) is measured at discrete intervals of time. In problems of this kind the Shewhart chart and other control chart procedures have been traditionally used quite successfully. This leads us to wonder whether these type of charts might not be justified on more plausible assumptions than those usually given.

A typical situation seems to be where the machine which is mass producing components is subject to going out of adjustment, and that the further one is away from target the worse the quality of the component is. Although small deviations from target can be tolerated, when the deviations become sufficiently large it will be necessary to stop the machine and reset it. This situation might realistically be represented by the following set of assumptions:

(i) Rather than assuming the disturbance to be represented by random variation about a fixed mean it will be assumed that we have a situation where successive observations are dependent and have a tendency to drift. In particular we shall consider the uncontrolled quality characteristic to follow the representationally useful non-stationary (0,1,1) model considered previously:

$$\nabla z_t = (1-\theta B)a_t$$

(ii) The loss sustained through being δ units off target is proportional to the square of the deviation and is $k\delta^2$ dollars.

(iii) When the deviation is sufficiently serious, the mean level must be adjusted, but now in making each adjustment a fixed loss of C dollars is sustained.

(iv) The dynamic characteristics in making a change are of no significance, but rather the adjustment is effective at once.

A control system in which the level is periodically adjusted and the controlled observations are considered in relation to a fixed target is equivalent to a system in which the uncontrolled observations z are considered in relation to a movable set point X to which the adjustments are applied with opposite sign (see Figures 5 and 6). Therefore if X_{t+1} is the adjustable "set point" at time $t+1$, then the deviation from target at this time is $z_{t+1} - X_{t+1}$. If it costs nothing to make an adjustment $(C = 0)$, then we could minimize costs by making an adjustment $X_{t+1} = \hat{z}_t(1)$ at each stage. But we must pay $\$C$ to make an adjustment and so the predicted deviation $\hat{z}_t(1) - X_t$ at time t must be such that $k(\hat{z}_t(1) - X_t)^2$ dollars is sufficiently large to warrant paying C dollars. Because of this the set point will usually be kept at a constant level for considerable periods.

Our problem is, knowing k, C, θ, and σ_a^2, to choose an optimal policy that tells us (a) when to change and (b) by how much to change so that the over-all loss in running the control scheme is minimized.

This is a problem of sequential decision-making and can be looked at in the framework of dynamic programming. Suppose that the control procedure must terminate after one further observation z_{t+1} is taken. The expected loss viewed from time t will then be

$$\left\{\begin{array}{ll} kE(z_{t+1} - X_t)^2 = k(\hat{z}_t(1) - X_t)^2 + k\sigma_a^2 & \text{(if no change is made)} \\ kE(z_{t+1} - X_{t+1})^2 + C & \text{(if a change is made)} \end{array}\right\}$$

The loss when a change is made will be minimized by setting $X_{t+1} = \hat{z}_t(1)$

and so this becomes $k\sigma_a^2 + C$. Hence the rule which minimizes the over-all loss is to change if $|\hat{z}_t(1) - X_t| \geq (C/k)^{\frac{1}{2}} = \lambda_1$ and to continue at the current level if $|\hat{z}_t(1) - X_t| < \lambda_1$. The loss function corresponding to this best rule is

$$L_1\{\hat{z}_t(1)\} = \mathrm{Min}\{k\sigma_a^2 + k(\hat{z}_t(1) - X_t)^2 \, , \, k\sigma_a^2 + C\}$$

Denote by $L_N\{\hat{z}_t(1)\}$ the minimal expected loss if the procedure terminates after N further observations. Then the expected loss $L_N^{(0)}\{\hat{z}_t(1)\}$, if no change is made, is the expected loss one step ahead plus the minimal expected loss if we started in position $z_{t+1} - X_t$ and there were $N-1$ further observations. This will be

$$L_1^{(0)}\{\hat{z}_t(1)\} + \int_{-\infty}^{\infty} L_{N-1}\{\hat{z}_{t+1}(1)\} p\{\hat{z}_{t+1}(1)|\hat{z}_t(1)\} d\hat{z}_{t+1}(1)$$

and since $\hat{z}_{t+1}(1) - X_t = \hat{z}_t(1) - X_t + (1-\theta)a_{t+1}$, this may be rewritten as

$$L_N^{(0)}\{\hat{z}_t(1)\} = k\sigma_a^2 + k(\hat{z}_t(1) - X_t)^2 + \int_{-\infty}^{\infty} L_{N-1}\{\hat{z}_t(1) + (1-\theta)\sigma_a^2 u\} p(u) du \qquad (15)$$

where $p(u)$ is the unit normal distribution. The loss if a change is made at time $t+1$, $L_N^{(1)}\{\hat{z}_t(1)\}$, may be obtained by replacing $\hat{z}_t(1)$ by $\hat{z}_t(1) - (X_{t+1} - X_t)$ in (15) and adding C. Again this loss is minimized by setting $X_{t+1} = \hat{z}_t(1)$ and the optimal rule is to change if $L_N^{(0)}\{\hat{z}_t(1)\} \geq L_N^{(1)}\{\hat{z}_t(1)\}$ which can be expressed in the form $|\hat{z}_t(1) - X_t| \geq \lambda_N$. The sequence $\{\lambda_i\}$ decreases monotonically and tends to a limiting value λ_{opt} corresponding to the practical case where the time of operation of the control procedure is effectively infinite. This optimal control rule appears very similar to a Shewhart chart where a quantity (in this case $\hat{z}_t(1) - X_t$) is plotted and referred to control lines (in this case at $\pm \lambda_{opt}$ units above and below the target X_t). A change is made if these control lines are exceeded.

A detailed analysis of this problem has been carried out in [3]. The optimal value of the control limit λ_{opt} is summarized in Table I in terms of the general parameters θ and σ_a^2 of the stochastic model and for a range of the ratio of the cost parameters C and k. In this table we have also tabulated the expected run length $E(n)$ (the average number of observations taken before an adjustment is made) and a cost variable g_w which enables one to calculate the expected "within changes" loss $L_w = k\sigma_a^2 + k(1-\theta)^2\sigma_a^2 g_w$ due to being off target. The expected loss due to changing is given by $L_c = C/E(n)$ and the expected over-all loss L is the sum $L = L_w + L_c$.

TABLE I

$\frac{\lambda_{opt}}{(1-\theta)\sigma_a}$, E(n), and g_w as a function of $c = \frac{C/K}{(1-\theta)^2\sigma_a^2}$

$c = \frac{C/K}{(1-\theta)^2\sigma_a^2}$	$\frac{\lambda_{opt}}{(1-\theta)\sigma_a}$	E(n)	g_w
20	2.6	10.7	1.5
50	3.5	17.5	2.5
100	4.3	24.5	3.8
200	5.3	34.2	5.6
500	6.8	55.4	8.9
1000	8.2	78.2	12.7

As an example, consider the case where $k = 8$, $C = 100$, $\theta = 0.5$, $\sigma_a = 1.0$ so that $c = (C/k)/(1-\theta)^2\sigma_a^2 = 50$. From the table we find that $E(n) = 17.5$ samples and $\lambda_{opt}/(1-\theta)\sigma_a = 3.5$, so that the best rule is to change when $|\hat{z}_t(1) - X_t| \geq 3.5 \times 0.5 = 1.75$. The within-run loss is then $L_w = 8 + 2 \times 2.5 = 13$ and the loss due to changing $L_c = 100/17.5 = 5.8$, and hence the over-all loss $L = 18.8$. Figure 5 shows a series for which $\theta = 0.5$ and $\sigma_a = 1.0$ in its uncontrolled state and Figure 6 shows the same series controlled in accordance with the optimal rule. Plotted in Figure 6 are the predicted deviations from target $\hat{z}_t(1) - X$ one step ahead. Appropriate adjustment is made when a point crosses the line. This in effect periodically refers the original series to the best current prediction at new origin as shown in Figure 5. In Table II we give for this example the values of E(n) and the losses for a range of values of λ.

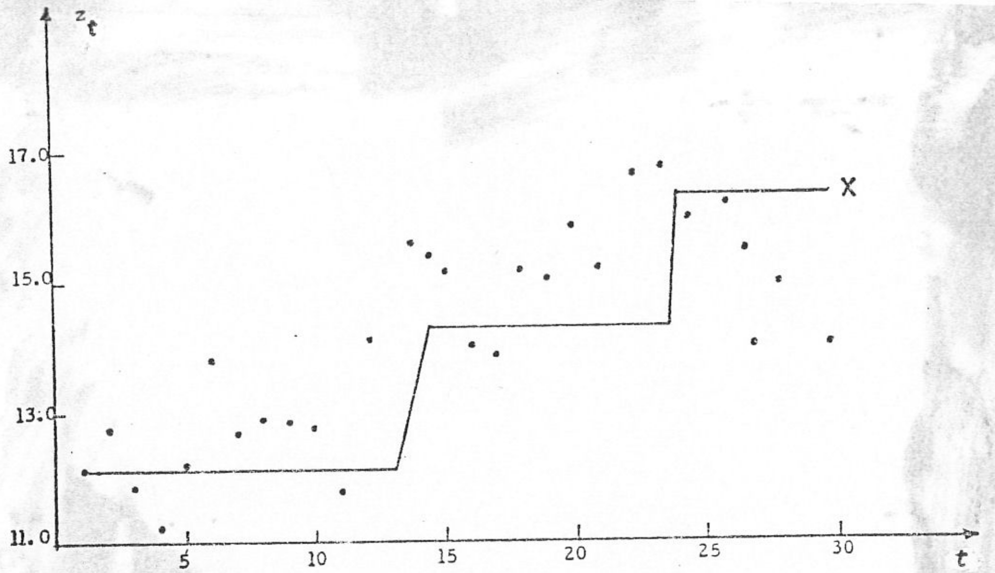


FIGURE 5. Uncontrolled series with changes in X.

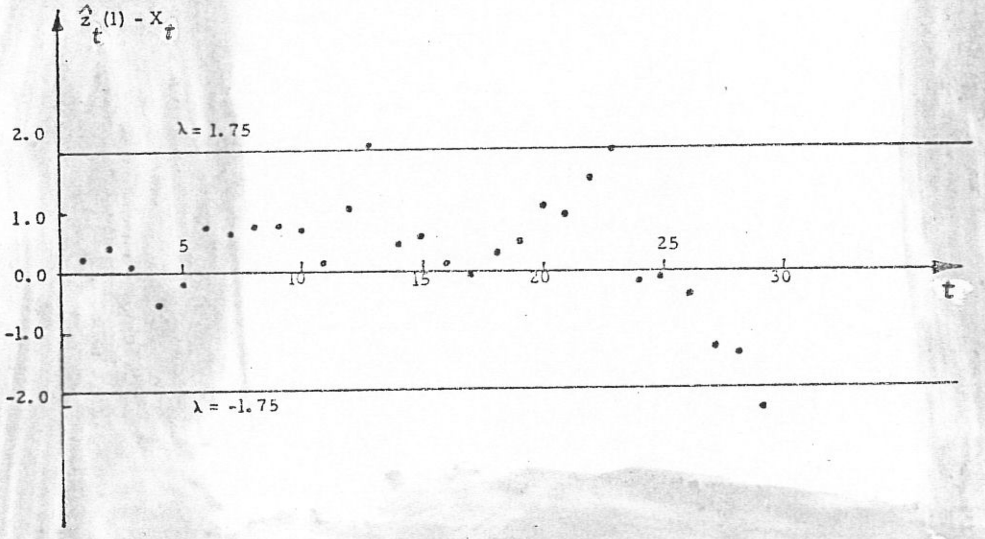


FIGURE 6 Predicted deviations from target leading to the adjustments shown in Figure 5.

TABLE II

Expected losses as a function of λ when $k=8$, $C=100$, $\theta=.5$, $\sigma_a=1$

λ	E(n)	L_w	L_c	L	Mean Square Error
0.0	1.0	8.0	100.0	108.0	1.00
1.0	7.2	9.8	13.9	23.7	1.22
1.5	13.4	11.9	7.4	19.3	1.49
1.75	17.5	13.0	5.8	18.8	1.62
2.0	21.7	14.6	4.6	19.2	1.82
2.5	31.9	18.0	3.1	21.1	2.25
3.75	58.4	26.6	1.7	28.3	3.32
4.5	92.9	38.2	1.1	39.3	4.78

It may be seen that the over-all loss L in the region of the minimum λ_{opt} is fairly flat. This suggests that the control scheme is remarkably robust to changes in the position of the "control lines".

In practice the ratio of costs C/k will not be very precisely known and the statistician is in the position which he commonly occupies where he needs to know a particular constant in order to produce an optimal answer to a problem. In this as in other instances he can usually best proceed by presenting the management with the alternative possibilities. In this example for instance the expected mean square error L_w/k of the fluctuations about target can be tabulated as we have done in Table II. This will increase as the spread of the control lines λ increases, but so will the average run length E(n) between changes, and hence the expected loss due to making changes L_c will decrease. These figures could help in the balancing of costs subjectively.

Conversely, if a particular value of λ is presently being used for the control lines, Table I would give the value of C/k to which this corresponds and enable one to examine whether this was sensible.

The form of the control scheme arrived at is worth further consideration. As illustrated in Figure 6 the one step ahead forecasts are plotted about a target X_t and referred to the control lines drawn at a distance $\pm \lambda_{opt}$ above and below the target. As long as the forecast falls within these control lines no change is made. One can see that this is similar to keeping a Shewhart chart on the predicted deviation from target one step ahead. Thus we note that a procedure very similar to one which has been found to be highly successful can be justified on assumptions which are probably more realistic than those originally adopted to justify it. By further noting that for the (0,1,1) process considered the one step ahead forecast can be written in the form

$$(\hat{z}_t(1) - X_t) = \theta(\hat{z}_{t-1}(1) - X_t) + (1-\theta)(z_t - X_t) \qquad (16)$$

we see that by plotting $(\hat{z}_t(1) - X_t)$ we have in fact the geometric moving average chart such as was suggested by Roberts [13]. However, the justification for it and the basis for choice of the "control limits" are quite different. We have not assumed as did Roberts that the series is a sequence of random independent deviates about a fixed mean but rather that it is highly correlated and has a tendency to drift. The "control limits" are seen to be related to the relative cost of making a change to that of being off target and to the parameters of the stochastic model rather than to any ideas of significance testing and probabilities of being outside control limits. In addition it should be noted that the "adjustable" parameter of the geometric moving average (16), rather than being selected in some arbitrary manner, is in our situation the θ parameter of the (0,1,1) process. As θ tends to one this stochastic process becomes a random walk and the optimal control procedure then reduces to the standard Shewhart chart, but with different "control limits". (It is worth noting that in this situation where we have introduced inertia into the system by viture of associating a cost with making a change we did not end up with a cummulative sum chart.) Thus we see that by starting off with reasonable assumptions for the problem we have been led to something which we know to have been of great value in the mass production industries.

5. Summary

An approach to discrete feedforward and feedback control is given which starts with modelling the dynamic and stochastic characteristics of the system. This description of the system together with the type of cost function involved leads to appropriate optimal control schemes. In Part I [6] a class of stochastic models was introduced which was capable of representing the kind of stationary and non-stationary behavior encountered by many practically occurring time series. A class of discrete transfer function models was also introduced which was capable of describing the dynamic relationship between a manipulated variable X and a controlled variable Y. Procedures were given for identifying, fitting, and checking these models and for using them to obtain optimal forecasts. In Part II the close link between forecasting and feedforward and feedback process control became apparent. Here the minimum mean square error controller was seen to act so as to cancel out the forecasted deviation from target which would have occurred if no control action were taken. Although usually implimented automatically this control action can often be implimented manually by the use of simple control charts not too different from the quality control charts used in the parts manufacturing industry. In the latter case the main difference in the design of a control scheme is the additional cost associated with the making of a change. In the machine tool problem sensible assumptions led naturally to a charting procedure very similar to Shewhart's which has been used very successfully in these industries.

References

[1] Barnard, G. A. (1959), "Control charts and stochastic processes", J.R.S.S., **B21**, 239.

[2] Box, G. E. P. and Jenkins, G. M. (1962), "Some statistical aspects of adaptive optimization and control", J.R.S.S., **B24**, 297.

[3] Box, G. E. P. and Jenkins, G. M. (1963), "Further contributions to adaptive quality control: simultaneous estimation of dynamics: non-zero costs", Bull. Intl. Stat. Inst., 34-th session, 943, Ottawa, Canada.

[4] Box, G. E. P. and Jenkins, G. M. (1965), "Mathematical models for adaptive control and optimization", A.I.Ch.E.-I.Ch.E. Symp. Series, **4**, 61.

[5] Box, G. E. P. and Jenkins, G. M. (1968), "Discrete models for feedback and feedforward control", *The Future of Statistics*, ed. D. G. Watts, 201, Academic Press, N.Y.

[6] Box, G. E. P. and Jenkins, G. M. (1968), "Some recent advances in forecasting and control, I", Applied Statistics, **17**, 91.

[7] Box, G. E. P. and Jenkins, G. M. (1970), *Time Series Analysis, Forecasting and Control*, Holden Day, San Francisco.

[8] Dudding, B. P. and Jennet, W. J. (1942), "Quality control charts", British Standard 600R, London.

[9] Goldsmith, P. L. and Whitfield, H. (1961), "Average run lengths in cumulative sum quality control schemes", Technometrics, **3**, 11.

[10] MacGregor, J. F. (1972), *Topics in the Control of Linear Processes with Stochastic Disturbances*, Ph.D. thesis, University of Wisconsin (also Department of Statistics Technical Reports).

[11] Page, E. S. (1957), "On problems in which a change in a parameter occurs at an unknown point", Biometrika, **44**, 249.

[12] Page, E. S. (1961), "Cumulative sum charts", Technometrics, **3**, 1.

[13] Roberts, S. W. (1959), "Control chart tests based on geometric moving averages", Technometrics, **1**, 239.

[14] Shewhart, W. A. (1931), *The Economic Control of the Quality of Manufactured Product*, MacMillan, N.Y.

[15] Truax, H. M. (1961), "Cumulative sum charts and their application to the chemical industry", Ind. Qual. Contr., **17**, 18.

[16] Wilson, G. T. (1970), "Modelling Linear Systems for Multivariate Control", Ph.D. thesis, Dept. of Syst. Engr., Univ. of Lancaster England.

DOCUMENT CONTROL DATA – R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

University of Wisconsin
Department of Statistics
Madison, Wisconsin

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

SOME RECENT ADVANCES IN FORECASTING AND CONTROLL PART II

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Scientific Interim

5. AUTHOR(S) *(First name, middle initial, last name)*

George E. P. Box, Gwilym M. Jenkins, and John F. MacGregor

6. REPORT DATE August, 1972	7a. TOTAL NO. OF PAGES 35	7b. NO. OF REFS 16

8a. CONTRACT OR GRANT NO.

AFOSR-72-2363

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

Technical Report No. 310

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Air Force Office of Scientific Research
1400 Wilson Boulevard
Arlington, Virginia

13. ABSTRACT

This paper is intended to be Part II of a two part series under the above title. Part I (Applied Statistics (1968), 17, p. 91) presented a class of discrete time series and dynamic models together with the theory for identifying, fitting and checking them. The principal application there was to forecasting.

Part II outlines an approach to discrete stochastic control which uses these models to typify the dynamic and stochastic characteristics of the system. This description of the system together with the type of cost function involved is shown to lead to appropriate optimal control schemes. Feedforward and feedback control schemes are worked out for situations typical of those occurring in the chemical and other process industries. A control problem more typical of situations arising in the parts manufacturing industries is shown to lead naturally to a control charting procedure very similar to Shewhart's which has been used very successfully in these industries.

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Discrete and stochastic control						
Cost functions						
Feedforward and feedback controllers						
Shewhart control charts						

Security Classification