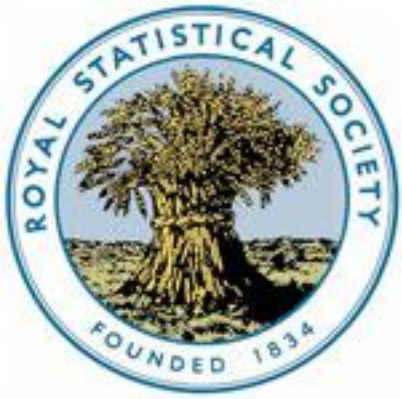

Some Recent Advances in Forecasting and Control†

Part I

G. E. P. Box
University of Wisconsin, U.S.A.
and
G. M. Jenkins
University of Lancaster, U.K.

1. Introduction

1.1. *Nature of Forecasting and Control Problems*

Until fairly recently the word "control" has been principally associated in the statistician's mind with quality control, and especially with the quality control chart techniques developed originally by Shewhart in the United States and by Dudding and Jennett in Great Britain.

During the war the development of sequential inspection methods by Wald and Barnard gave new impetus to techniques in which sequential aspects were emphasised and led eventually to the introduction of cumulative sum charts by Page (1954, 1957) and by Barnard (1959).

The need for control implies the existence of an inherent disturbance in the process of one kind or another such as might be described by a time series. Thus, in recent years we find contributions to control problems from workers in stochastic processes such as Whittle (1963) and Bather (1963). Because one approach to control would be to forecast the deviation from target which would occur if no action were taken and then to act so as to cancel out that deviation, *forecasting* and *control* problems are closely linked together. However, we can forecast a time series in an optimal manner only if we have an adequate stochastic model for that series.

In the past a great deal of attention has been given to stationary time series models which have the property of remaining in equilibrium about a constant mean. However, forecasting has been of particular importance in business and economics where many series (for example, the monthly sales of an industrial product) are *non-stationary* and have no natural mean. It is not surprising, therefore, that the economic forecasting methods which have been proposed by such workers as Holt (1957, 1960), Winters (1960) and Brown (1962) and the control chart techniques proposed by Roberts (1959), all using the exponentially weighted moving averages, are appropriate for a particular type of *non-stationary* process. The fact that such methods have been successful supplies a clue to the *kind* of non-stationary model which might be useful in these problems.

† Originally presented at the European Meeting of Statisticians, held at Imperial College, London, in September 1966.

To a control engineer the word "control" has had a different connotation. He usually thinks in terms of feedback and feed forward control loops, the dynamics and stability of the system, and often of particular types of hardware to carry out the control action. In this paper we outline a statistical approach to optimal forecasting and to the optimal design of feedback and feedforward control schemes that we have developed in previous papers (Box and Jenkins, 1962, 1963, 1965; Box *et al.* 1967) and which will be described in a forthcoming book (Box and Jenkins, 1968).

The control techniques we discuss are different from those of standard quality control procedures, but this is because they have a different purpose. We certainly do not believe that the traditional quality control chart is unimportant or outmoded. Appropriate *display* of data on such a chart (rather than the burying of it in a process record book) ensures that changes that occur are regularly brought to the attention of those in charge of the process. They are thus nudged into seeking "assignable causes" for the changes and a continuous incentive for process improvement is achieved. This device is of enormous importance because it can stimulate new thinking about the process. However, in many situations a control scheme is required which adjusts some variable, whose precise effect on the quality characteristic is *known*, so as to minimize the variation of this quality characteristic about a target value. It is with such control problems that we are concerned here.

1.2. *An Outline of the Approach*

We suppose throughout that observations are available at *discrete* equispaced intervals of time. For example, in a sales forecasting problem, figures might be available every month and we might wish to forecast sales for $1, 2, 3, \ldots, 12$ months ahead. Again, in a chemical process, observations and the opportunity to make control changes might occur every 5 minutes, every hour, or every shift depending on the rate at which the state of the system could change. In the case of a chemical process discrete observations might arise from a discrete or batch process, or a continuous record of the process characteristic might be "sampled" at equally spaced intervals. In practice, if the sampling interval is suitably chosen almost nothing will be lost by employing the discrete rather than the continuous record and there may be considerable gain in the simplicity of the analysis.

The optimal forecasts of future values of a time series are determined by the stochastic model that describes that series. Therefore the main object in statistical analysis directed to forecasting must be in obtaining a suitable stochastic model for the series in question. Therefore, we first develop a class of *stochastic models* which are capable of representing not only stationary behaviour but also non-stationary behaviour of the kind that we have encountered in practice. We show how models which satisfactorily describe a particular series may be derived and how they can be used to forecast seasonal as well as non-seasonal series. The same kind of stochastic model used in the forecasting problem may also be used to represent the disturbances which infect a system and which make control action necessary.

Now any control action which is taken will not be felt immediately but usually its effect will build up gradually because of the inertia of the system. Therefore we next describe *dynamic models* capable of representing the dynamic relationship between a controlling variable X and a controlled variable Y and we show how these dynamic models may be fitted to data obtained from the system.

An important principle in the choice of our models is that they should, whilst

adequately representing the data, contain as few parameters as possible. Following Tukey we call this the *principle of parsimony*.

In Part II of this paper we shall describe how the stochastic and dynamic models may be brought together to design optimal feed-forward and feedback control schemes and also how the parameters in the stochastic and dynamic models may be simultaneously estimated from measurements made on the operating system.

2. Time Series Models

A criterion of great importance in discussing time series is *stationarity*. A series is strictly stationary if its properties are completely unaffected by a shift in the time origin. In particular, a stationary series varies about some *fixed* mean μ. It exhibits no change in mean and no drift.

2.1. *Autoregressive and Moving Average Models for Stationary Time Series*

Suppose we denote the values of a stationary series at equally spaced times $t, t-1, t-2, \ldots$ by $w_t, w_{t-1}, w_{t-2}, \ldots$ Let $a_t, a_{t-1}, a_{t-2}, \ldots$ be a "white noise" series consisting of uncorrelated random Normal deviates all having mean zero and variance σ_a^2. It is helpful to think of these a's as a series of random "shocks".

The time series model we employ, originally developed by Yule, is essentially a device for transforming the original series w_t, the observations of which are often highly correlated, into a series of uncorrelated component shocks a_t which can be thought of as generating the series. There are basically two different ways in which this is done.

The deviation $\dot{w}_t = w_t - \mu$ from the mean μ can be made linearly dependent on previous deviations $\dot{w}_{t-1} = w_{t-1} - \mu$, $\dot{w}_{t-2} = w_{t-2} - \mu$, etc., and on a_t. We then have what is called an *autoregressive* model. Thus

$$\dot{w}_t = \phi_1 \dot{w}_{t-1} + a_t, \tag{1}$$

$$\dot{w}_t = \phi_1 \dot{w}_{t-1} + \phi_2 \dot{w}_{t-2} + a_t, \tag{2}$$

are autoregressive models of orders 1 and 2, respectively.

Alternatively, we can make $\dot{w}_t$ linearly dependent on a_t and on one or more previous a's. We then have what is called a *finite moving average model.* Thus

$$\dot{w}_t = a_t - \theta_1 a_{t-1}, \tag{3}$$

$$\dot{w}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}, \tag{4}$$

are moving average models of orders 1 and 2, respectively. One might ask: can an autoregressive model be used to represent moving average behaviour? The answer is that this can be done but an infinite number of autoregressive terms are needed to represent a finite moving average model and vice versa.

To ensure parsimony we may need terms of both kinds and we are thus led to the *general mixed autoregressive-moving average model of order* (p, q), which may be written

$$\dot{w}_t - \phi_1 \dot{w}_{t-1} - \ldots - \phi_p \dot{w}_{t-p} = a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}, \tag{5}$$

where p and q would by 0, 1 or 2 in most applications. To manipulate models of this kind it is convenient to define a backward shift operator B such that

$$Bw_t = w_{t-1}. \tag{6}$$

Using the operator B, (5) can be written

$$\phi_p(B)\dot{w}_t = \theta_q(B)a_t, \tag{7}$$

where

$$\phi_p(B) = 1-\phi_1 B-\phi_2 B^2 \ldots -\phi_p B^p,$$
$$\theta_q(B) = 1-\theta_1 B-\theta_2 B^2 \ldots -\theta_q B^q,$$

are polynomials in B of degree p and q respectively and $\phi_p(B)$ is called the *autoregressive* operator and $\theta_q(B)$ the *moving average* operator.

For example, the models of equations (1), (2), (3), and (4) could be written

A.R. 1: $(1-\phi_1 B)\dot{w}_t = a_t,$

A.R. 2: $(1-\phi_1 B-\phi_2 B^2)\dot{w}_t = a_t,$

M.A. 1: $\dot{w}_t = (1-\theta_1 B)a_t,$

M.A. 2: $\dot{w}_t = (1-\theta_1 B-\theta_2 B^2)a_t.$

Now consider the first order autoregressive model (1). The values of the series may be built up recursively as follows:

$$\begin{aligned}
\dot{w}_1 &= \phi_1\dot{w}_0+a_1,\\
\dot{w}_2 &= \phi_1\dot{w}_1+a_2 = \phi_1^2\dot{w}_0+\phi_1 a_1+a_2,\\
\dot{w}_3 &= \phi_1\dot{w}_2+a_3 = \phi_1^3\dot{w}_0+\phi_1^2 a_1+\phi_1 a_2+a_3,\\
\dot{w}_t &= \phi_1\dot{w}_{t-1}+a_t = \phi_1^t\dot{w}_0+\phi_1^{t-1}a_1+\phi_1^{t-2}a_2+\ldots+a_t.
\end{aligned} \tag{8}$$

We can ensure stationarity for this series by requiring that ϕ_1 lies between the values -1 and $+1$. If ϕ_1 lay outside these limits (if for example, ϕ_1 were equal to 2) then we can readily see from equation (8) that the deviation $\dot{w}_t$ would be dominated by remote events led by $\dot{w}_0$ and a_1 which would become more and more important as t became larger. On the other hand, if ϕ_1 lay between -1 and $+1$, as we require, the behaviour of $\dot{w}_t$ would be dominated by the most recent shock a_t, as is sensible.

A similar argument applied to the first order moving average model (3) leads to the conclusion that θ_1 must lie between -1 and $+1$ if a_t is not to be dominated by remote events. If this condition is satisfied the moving average model is said to be *invertible*.

Now one way of expressing the condition that ϕ_1 in the autoregressive operator $1-\phi_1 B$ lies between -1 and $+1$ is to say that the roots of the equation $1-\phi_1 B = 0$ (where B is regarded as a variable) lie *outside* the interval -1 to $+1$.

The corresponding condition for stationarity and invertibility of the general mixed autoregressive moving average model (5) is that the roots (which may be complex) of $\phi(B) = 0$ and $\theta(B) = 0$ must lie outside the unit circle and we shall suppose in all that follows that this condition is imposed.

With these conditions satisfied the model (5) turns out to be a valuable device for representing stationary time series. If the model is expressed in terms of the w_t's themselves, instead of deviations from the mean, the general form of the model may be written

$$\phi_p(B)w_t = \theta_0+\theta_q(B)a_t, \tag{9}$$

where

$$\theta_0 = (1-\phi_1-\phi_2-\ldots-\phi_p)\mu.$$

2.2. *A General Model which can represent Stationary and Homogeneous Non-stationary Time Series*

Time series representing economic phenomena and disturbances in processes to be controlled are often best represented by non-stationary models. There is an unlimited number of ways in which a time series may be non-stationary. We now adapt our models to take account of the kinds of non-stationarity which we have frequently met in practice. Figure 1(a) shows one type of non-stationary series of common occurrence. This series is homogeneous except in its level. By this is meant that apart from a vertical translation, one part of the series looks much like another. A series z_t which is stationary in its *first difference*

$$\nabla z_t = z_t - z_{t-1} = (1-B)z_t$$

exhibits precisely this kind of behaviour. Again Figure 1(b) shows a second kind of non-stationarity which is frequently met. This series has neither a fixed level nor a fixed slope but is homogeneous if one allows for differences in these characteristics. We can reproduce such behaviour in a series z_t by a representation in which the *second difference*

$$\nabla^2 z_t = z_t - 2z_{t-1} + z_{t-2} = (1-B)^2 z_t$$

follows a stationary model.

Finally then, if z_t is the variable whose behaviour we wish to represent, it is assumed that its dth difference $\nabla^d z_t = w_t$ can be represented by the stationary and invertible model of equation (9). Since $\nabla^d = (1-B)^d$, the model for z_t becomes

$$\phi_p(B)(1-B)^d z_t = \theta_0 + \theta_q(B)a_t, \tag{10}$$

which will be non-stationary unless $d = 0$. The model is said to be of order (p,d,q) where p, d, and q are usually 0, 1, or 2.

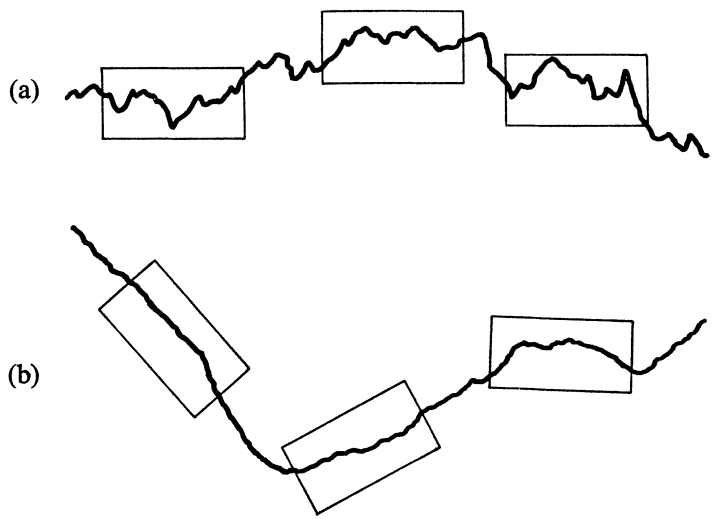

FIG. 1(a). A series showing non-stationarity in level such as can be represented by the model $\phi(B)\nabla z_t = \theta(B)a_t$.

FIG. 1(b). A series showing non-stationarity in level and in slope such as can be represented by the model $\phi(B)\nabla^2 z_t = \theta(B)a_t$.

The operator $\Phi_{p+d}(B) = \phi_p(B)(1-B)^d$ is called the *general* autoregressive operator. Since d of the roots of $\Phi_{p+d}(B) = 0$ are unity, this non-stationary operator will, of course, not satisfy the stationarity condition that all roots lie outside the unit circle. In many practical cases where differencing is needed to obtain stationarity (that is where $d \geqslant 1$), $\nabla^d z_t = w_t$ can be assumed to have a zero mean so that θ_0 in (10) can be set equal to zero.

Suppose we wish to determine a suitable model for a series for which observations $z_1, z_2, z_3, \ldots$ are available (where if possible there should be at least 50 and preferably more than 100 observations). In practice such model determination has to be done iteratively using a process of *identification, estimation, diagnostic checking, refitting and rechecking* until a satisfactory representation is found.

2.3. *Identification*

Equation (10) supplies too rich a class of models to permit immediate estimation. Therefore, using experience and the data we first identify a sub-class of models *worthy to be entertained.*

The primary data-analysis tool at this stage is the *sample autocorrelation function* of the original series and its differences. Suppose that n differences $w_1, w_2, \ldots, w_n$ are available. The sample autocorrelation coefficient at lag k for $w_t = \nabla^d z_t$ is

$$r_k(w) = c_k(w)/c_0(w),$$

where

$$c_k = \frac{1}{n}\sum_{t=1}^{n-k}(w_t - \bar{w})(w_{t+k} - \bar{w}) \quad \text{and} \quad \bar{w} = \frac{1}{n}\sum_{t=1}^{n} w_t.$$

We shall use $\rho_k(w)$ for the corresponding theoretical autocorrelation.

A suitable value for d may be inferred by finding the degree of differencing necessary to induce the sample autocorrelation function to damp out fairly quickly. For example, Table 1 shows the sample autocorrelation function of z, ∇z, and $\nabla^2 z$ for a series of IBM Common Stock Daily Closing Prices given by Brown (1962). While the sample autocorrelations for the original series are very slow to die out, indicating non-stationarity, its first and higher differences behave like those of a stationary series suggesting that we set $d = 1$.

Values to be entertained for p and q may usually be deduced by inspecting the sample autocorrelations using knowledge of the behaviour of the theoretical autocorrelation function ρ_k for various types of models. The characteristics of $\rho_k(w)$ for

TABLE 1

Sample autocorrelations for various differences of the IBM Common Stock Daily Closing Prices

Source: New York Stock Exchange, May 1961–November 1962 (369 observations)

		1	2	3	4	5	6	7	8	9	10
z	Lags 1–10	·99	·99	·98	·97	·96	·96	·95	·94	·93	·92
	11–20	·91	·91	·90	·89	·88	·87	·86	·85	·84	·83
∇z	Lags 1–10	·09	·00	−·05	−·04	−·02	·12	·07	·04	−·07	·02
	11–20	·08	·05	−·05	·07	−·07	·12	·12	·05	·05	·07
$\nabla^2 z$	Lags 1–10	−·45	−·02	−·04	·00	−·07	·11	−·01	·04	−·10	·02
	11–20	·04	·04	−·12	·13	−·17	·10	·05	−·04	−·01	·09

models of order $(1,d,0)$, $(2,d,0)$, $(0,d.1)$, $(0,d,2)$ and $(1,d,1)$ are shown in Table 2. The boundaries of the admissible parameter space are indicated by the inequalities. We see from Table 1 that the autocorrelations of ∇z are all small and appear consistent with a model of order $(0,1,0)$ or perhaps $(0,1,1)$.

Of considerable help in judging the reality of sample autocorrelations is the following approximate formula due to Bartlett for the standard error (S.E.) of r_k, namely

$$\text{S.E.}\,[r_k] \simeq \sqrt{\left\{\frac{1}{n}(1+2\rho_1^2+2\rho_2^2+\ldots)\right\}}. \tag{11}$$

Since we do not know the theoretical autocorrelations ρ_k, they have to be replaced by their sample estimates r_k.

Thus, under the assumption that the first difference of the IBM series is a moving average of order 1 (that is, the series is of order $(0,1,1)$)

$$\text{S.E.}\,[r_k] \simeq \sqrt{[\tfrac{1}{369}\{1+2(0{\cdot}09)^2\}]} = 0{\cdot}05.$$

Referring to Table 1, we see that only 3, that is 6 per cent of the sample autocorrelations of ∇z from the second onwards are greater than two standard deviations, confirming that a model of order $(0,1,1)$ is worthy to be entertained.

TABLE 2

Behaviour of theoretical autocorrelation function of dth difference of series for various simple (p,d,q) models

Order	$(1, d, 0)$	$(0, d, 1)$
Behaviour of ρ_k	$\rho_k = \phi^k$ decays exponentially	only ρ_1 non-zero
Preliminary estimates from	$\phi_1 = \rho_1$	$\rho_1 = \dfrac{-\theta_1}{1+\theta_1^2}$
Admissible region	$-1 < \phi_1 < 1$	$-1 < \theta_1 < 1$
Order	**$(2, d, 0)$**	**$(0, d, 2)$**
Behaviour of ρ_k	mixture of exponentials or damped sine wave	only ρ_1 and ρ_2 non-zero
Preliminary estimates from	$\phi_1 = \dfrac{\rho_1(1-\rho_2)}{1-\rho_1^2}$ $\phi_2 = \dfrac{\rho_2-\rho_1^2}{1-\rho_1^2}$	$\rho_1 = \dfrac{-\theta_1(1-\theta_2)}{1+\theta_1^2+\theta_2^2}$ $\rho_2 = \dfrac{-\theta_2}{1+\theta_1^2+\theta_2^2}$
Admissible region	$-1 < \phi_2 < 1$ $\phi_2+\phi_1 < 1$ $\phi_2-\phi_1 < 1$	$-1 < \theta_2 < 1$ $\theta_2+\theta_1 < 1$ $\theta_2-\theta_1 < 1$
Order	**$(1, d, 1)$**	
Behaviour of ρ_k	decays exponentially after first lag, $\rho_k = \phi\rho_{k-1}$ $(k \geqslant 2)$	
Preliminary estimates from	$\rho_1 = \dfrac{(1-\theta_1\phi_1)(\phi_1-\theta_1)}{1+\theta_1^2-2\phi_1\theta_1}$	$\rho_2 = \rho_1\phi_1$
Admissible region	$-1 < \phi_1 < 1$	$-1 < \theta_1 < 1$

By substituting sample estimates for ρ_k in Table 2, preliminary values for the model parameters (which, however, are in general not efficient estimates) may be obtained. For instance, in the case of the IBM Stock Price series suppose that we tentatively entertain the model $\nabla z_t = (1-\theta B)a_t$ of order (0, 1, 1). Then, because r_1 of ∇z_t is 0·09, a first guess for the parameter θ_1 is $-0{\cdot}09$ since this is the root of the equation $0{\cdot}09 = -\theta_1/1+\theta_1^2$ which lies within the admissible region $-1 < \theta_1 < 1$.

A complementary tool for identification called the *sample partial autocorrelation function* may also be used (see for example Box and Jenkins (1968)).

2.4. *Fitting*

Using efficient statistical methods we may now fit the tentatively identified model, or to be on the safe side, a slightly over-parameterized version of it.

On the assumption that the a's are Normally distributed, a close approximation to the maximum likelihood estimates of $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots \phi_p)$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_q)$ will be obtained by minimizing the sum of squares

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum a_t^2(\boldsymbol{\phi}, \boldsymbol{\theta}).$$

The values a_t $(\boldsymbol{\phi}, \boldsymbol{\theta})$ for any $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ may readily be calculated recursively using

$$a_t = \theta_0 + \theta_1 a_{t-1} + \ldots + \theta_q a_{t-q} + w_t - \phi_1 w_{t-1} - \ldots - \phi_p w_{t-p}$$

with $w_t = \nabla^d z_t$ and when $d \neq 0$, θ_0 would often be set equal to zero. The process can be started off by commencing with a_{p+1} and setting $a_p, a_{p-1}, \ldots a_{p-q+1}$ equal to their expected values of zero. This procedure is adequate for most purposes but a more exact calculation of the likelihood function will be described in Box and Jenkins (1968).

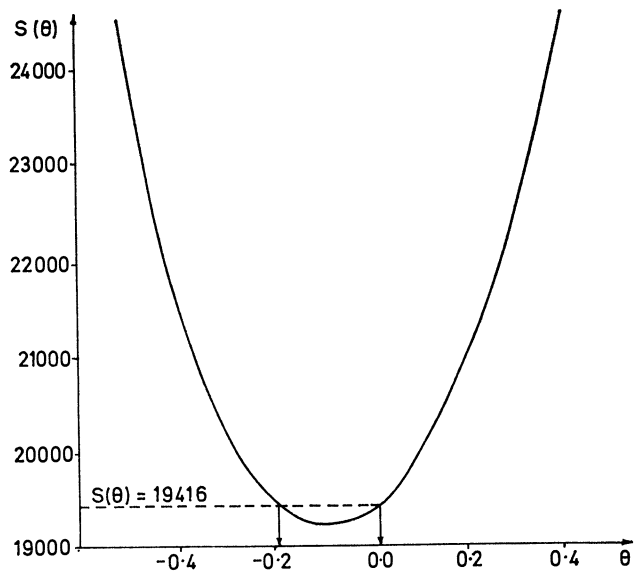

FIG. 2. Sum of squares function for I.B.M. data with approximate 95% confidence region for θ.

An approximate $1-\varepsilon$ confidence region for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ is enclosed by the contour

$$S_{1-\varepsilon}(\boldsymbol{\phi},\boldsymbol{\theta}) = S(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}})[1+\{\chi^2_{1-\varepsilon}(p+q)/\nu\}] \tag{12}$$

where $\chi^2_{1-\varepsilon}(m)$ is the upper $1-\varepsilon$ significance point of the chi-square distribution having m degrees of freedom and ν is equal to the number of a's in the sum $S(\boldsymbol{\phi},\boldsymbol{\theta})$ less the number of parameters fitted.

We illustrate again with the IBM data, using the tentatively entertained model $\nabla z_t = a_t - \theta a_{t-1}$. Figure 2 shows a plot of $S(\theta)$ against θ with a minimum at $\hat{\theta} = -0{\cdot}09$ and $S(\hat{\theta}) = 19{,}216$. The approximate 95 per cent confidence limits for θ of $-0{\cdot}19$ and $0{\cdot}03$ are those values for which

$$S(\theta) = 19{,}216\ (1+3{\cdot}84/367) = 19{,}417.$$

Complicating the model by adding an extra term on either side produced no appreciable reduction in the residual sum of squares. Hence the form of the model which was finally accepted is $\nabla z_t = a_t + 0{\cdot}1a_{t-1}$. Least squares estimates and approximate confidence limits may be obtained without the use of graphical methods using iterative non-linear least squares procedures described in Box *et al.* (1967). However, in general graphs and contour plots of the sum of squares function $S(\boldsymbol{\phi},\boldsymbol{\theta})$, or of sections of it, are of great value in illuminating the estimation situation.

2.5. *Diagnostic Checks*

If the form of the model is correct and if $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$ are close to their "true" values, then the estimated residuals $\hat{a}_t = a_t(\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\theta}})$ will be (very nearly) uncorrelated random deviates. Inadequacies of the model may be shown up for example by examining the autocorrelation function of the residuals. A fuller discussion is given in Box and Jenkins (1968).

2.6. *Seasonal Models*

One often has to analyse time series in which recurrent patterns with *known period* s occur, for example, yearly patterns in monthly sales data ($s = 12$). Here parsimony can often be achieved using multiplicative models of the type

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D z_t = \theta_q(B)\Theta_Q(B^s)a_t. \tag{13}$$

To see how this model is arrived at, suppose we are analysing a series of monthly sales data so that $s = 12$. Suppose we consider all the data at a fixed point in the period s. For example, suppose we consider the sequence of January sales figures. This series would be free of seasonality and might be described by a suitably chosen model of the general form given in equation (10). Bearing in mind that successive Januarys are $s = 12$ months apart and assuming that $\theta_0 = 0$, we would have

$$\Phi_P(B^s)\nabla_s{}^D z_t = \Theta_Q(B^s)e_t, \tag{14}$$

where

$$\nabla_s z_t = z_t - z_{t-s} \qquad \text{and} \qquad B^s z_t = z_{t-s}.$$

It could reasonably be assumed that February sales, March sales, etc. would follow precisely similar models *with the same parameters*. However, it could not be expected that the residuals e_{t+1} from February sales would be independent of the residuals e_t from January sales. To allow for this dependence a second model may be fitted to the "seasonal free" residuals e_t in the form

$$\phi_p(B)\nabla^d e_t = \theta_q(B)a_t. \tag{15}$$

On eliminating e_t between (14) and (15), we obtain (13). When $s = 12$ the model embodies parameters which describe month-to-month variation (little letters) and parameters which describe year-to-year variation (capital letters).

Procedures for identifying, fitting and checking such models closely follow those described above. For instance, it was shown in Box *et al.* (1967) that the airline passenger data of Fig. 4 was closely fitted by the model

$$(1-B)(1-B^{12})z_t = (1-0{\cdot}4B)(1-0{\cdot}6B^{12})a_t \tag{16}$$

corresponding to $p = 0$, $P = 0$, $s = 12$, $d = 1$, $D = 1$, $q = 1$, $Q = 1$, $\theta_1 = 0{\cdot}4$, and $\Theta_1 = 0{\cdot}6$. The sum of squares plot for this example is shown in Fig. 3.

2.7. *Forecasting*

Suppose now that we have determined an adequate model for a given series and we have new data $z_t, z_{t-1}, \ldots$ from the same series extending up to the present time t from which we wish to make a forecast l steps ahead. We call this an *origin t* forecast for *lead time l*.

It may be shown that the minimum mean square error forecast for any lead time is given by

$$\hat{z}_t(l) = \underset{t}{E}[z_{t+l}],$$

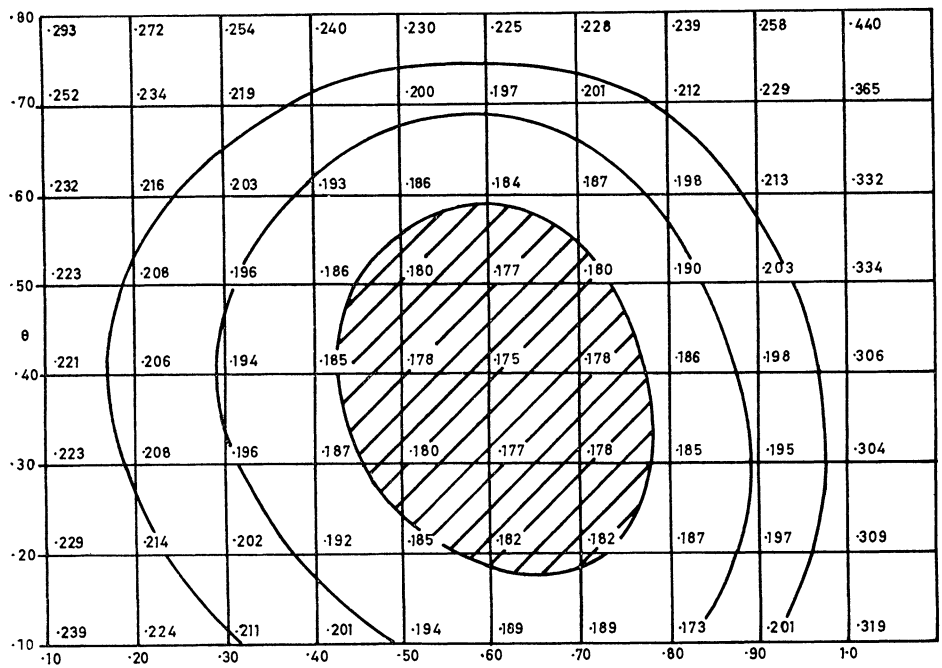

FIG. 3. Sum of squares grid and contours of $S(\theta, \Theta)$ for airline data with approximate 95% confidence region shaded

where $\underset{t}{E}$ is the conditional expectation given the z's up to time t. It follows, in particular, that

$$a_t = z_t - \hat{z}_{t-1}(1). \tag{17}$$

Thus the "shocks" a_t in the models (10) and (13) are in fact the forecast errors for unit lead time. That for an optimal forecast these "one step ahead" forecast errors ought to form an uncorrelated series is otherwise obvious. For suppose these forecast errors were autocorrelated; then it would be possible to forecast the next forecast error in which case the forecast could not be optimal.

The required expectations are easily found because

$$\underset{t}{E}[z_{t+j}] = \hat{z}_t(j), \qquad \underset{t}{E}[a_{t+j}] = 0, \qquad j = 1,2,3,...$$

$$\underset{t}{E}[z_{t-j}] = z_{t-j}, \qquad \underset{t}{E}[a_{t-j}] = a_{t-j} = z_{t-j} - \hat{z}_{t-j-1}(1), \qquad j = 0,1,2,... \quad (18)$$

For instance, to determine the 3-month ahead forecast for the airline series, we first use (16) to write down

$$z_{t+3} = z_{t+2} + z_{t-9} - z_{t-10} + a_{t+3} - 0{\cdot}4a_{t+2} - 0{\cdot}6a_{t-9} + 0{\cdot}24a_{t-10}.$$

Taking conditional expectations at time t,

$$\hat{z}_t(3) = \hat{z}_t(2) + z_{t-9} - z_{t-10} - 0{\cdot}6a_{t-9} + 0{\cdot}24a_{t-10},$$

and using (17),

$$\hat{z}_t(3) = \hat{z}_t(2) + z_{t-9} - z_{t-10} - 0{\cdot}6\{z_{t-9} - \hat{z}_{t-10}(1)\} + 0{\cdot}24\{z_{t-10} - \hat{z}_{t-11}(1)\},$$

that is

$$\hat{z}_t(3) = \hat{z}_t(2) + 0{\cdot}4z_{t-9} - 0{\cdot}76z_{t-10} + 0{\cdot}6\hat{z}_{t-10}(1) - 0{\cdot}24\hat{z}_{t-11}(1).$$

The forecast $\hat{z}_t(2)$ can be obtained in a similar way in terms of $\hat{z}_t(1)$ from $\underset{t}{E}[z_{t+2}]$.

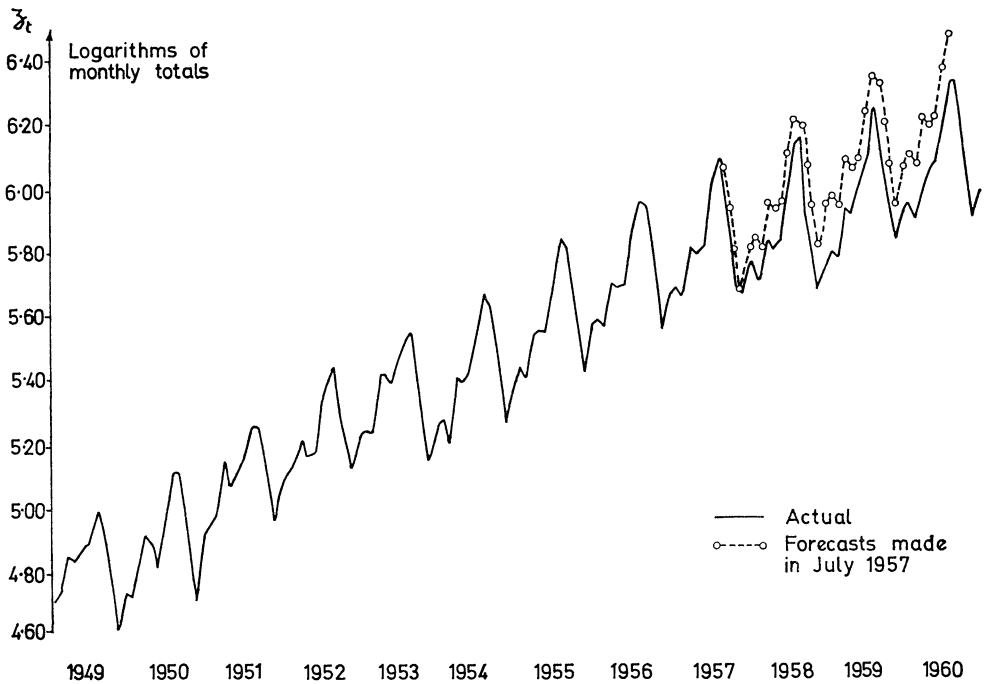

FIG. 4. Logarithms of monthly totals of international airline passengers with forecast made origin July 1957 for 1, 2, 3, . . . , 36 months ahead.

Similarly $\hat{z}_t(1)$ can be obtained from $\underset{t}{E}[z_{t+1}]$ which employs only values of the perviously occurring z's and a's. In practice then it is a very simple matter to compute the forecasts $\hat{z}_t(1)$, $\hat{z}_t(2)$, $\hat{z}_t(3)$, etc. recursively, using the forecast function

$$\underset{t}{E}[z_{t+l}] = \underset{t}{E}[z_{t+l-1}+z_{t+l-12}-z_{t+l-13}-0{\cdot}4a_{t+l-1}-0{\cdot}6a_{t+l-12}+0{\cdot}24a_{t+l-13}]$$

and (18). Note that this form of computation is ideally suited for use on an automatic computer. Using these methods, forecasts made at origin July 1957 for lead times $1,2,3,\ldots,36$ months ahead are shown in Fig. 4 where they may be compared with the values actually realised.

The procedure provides a very convenient and efficient method for industrial forecasting. In particular, it is ideally suited for forecasting sales or inventory on a large variety of products. Since only a very small amount of previous information need be stored for each product a computer with only modest storage capacity may be employed. In those cases where a past history of 50 or so observations is not available one can proceed by using experience and whatever past information *is* available to yield a preliminary model which may then be updated from time to time as more information becomes available.

3. Dynamic Models

In this section we consider the estimation of dynamic models which describe the relationship between a manipulated variable X and a controlled variable Y. Since the dynamic model describes how changes in X are transmitted into Y, it may be said to describe the *transfer function* between X and Y. Knowledge of the appropriate transfer function is essential for the design of control schemes. However, dynamic models of the type we now describe are also useful in forecasting a time series Y from past values of another time series X as well as from past values of Y.

3.1. *Linear Dynamic Models*

Suppose that in the study of the dynamic characteristics of some system, such as a chemical reactor, pairs of observations $(X_1, Y_1), (X_2, Y_2),\ldots$ are available of an input X, such as gas feed rate and an output Y, such as product viscosity. Suppose further that over the operating ranges of variation of Y and X there exists an approximately linear steady-state relationship

$$\dot{Y} = g\dot{X},$$

where $\dot{Y}, \dot{X}$ denote deviations from some average levels, and g is called the *steady state gain* of the system (or the linear regression coefficient between Y and X).

The *dynamic* characteristics of such systems can usually be represented parsimoniously by linear difference equations of the form

$$\xi(\nabla)\,\dot{Y}_{t+1} = g\eta(\nabla)\dot{X}_{t-b} \tag{19}$$

with

$$\xi(\nabla) = 1+\xi_1\nabla+\ldots+\xi_u\nabla^u,$$
$$\eta(\nabla) = 1+\eta_1\nabla+\ldots+\eta_v\nabla^v,$$

where b represents the number of whole intervals of pure dead time (delay) in the

system. Most systems occurring in practice can be represented parsimoniously with u and v at most 2. For instance, the simple model

$$(1+\xi\nabla)\,\dot{Y}_{t+1} = g(1+\eta\nabla)\dot{X}_t \tag{20}$$

or

$$\dot{Y}_{t+1} = \frac{\xi}{1+\xi}\,\dot{Y}_t + g\frac{(1+\eta)}{1+\xi}\,\dot{X}_t - \frac{g\eta}{1+\xi}\,\dot{X}_{t-1} \tag{21}$$

can represent a system whose response to a step change of X_0 in the input is to produce an eventual change gX_0 in the output which is approached exponentially at a rate depending on ξ and delayed by an amount depending on η. Fig. 5 illustrates the model (20) with $\xi = 1$, $g = 4$ and $\eta = -0{\cdot}5$.

By solving the difference equation (19) the dynamic model can be written in the alternative form

$$\begin{aligned}\dot{Y}_{t+1} &= v_0\dot{X}_t + v_1\dot{X}_{t-1} + \ldots \\ &= V(B)\dot{X}_t,\end{aligned} \tag{22}$$

where the weights v_j applied to past inputs are called the *impulse response function* of the discrete system. The form (22) is not a parsimonious way of representing the dynamic model, but is useful in identifying the model (19) as will be shown in section 3.2.

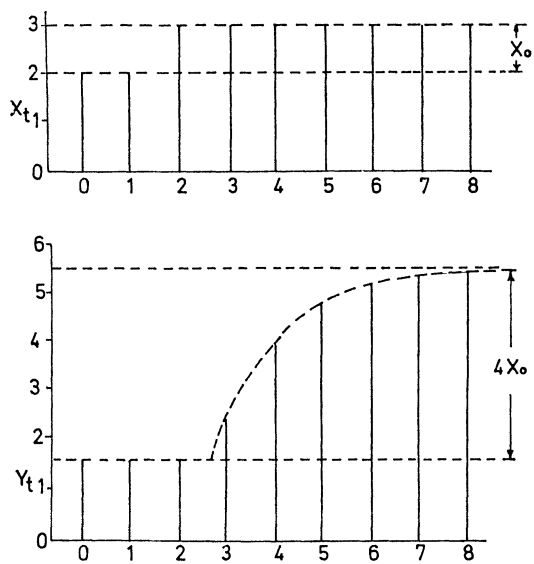

FIG. 5. Delayed exponential response to a step change produced by $(1+\nabla)(Y_{t+1} - 1{\cdot}5) = 4(1 - 0{\cdot}5\nabla)(X_t - 2)$ or $(1+\nabla)Y_{t+1} = -6{\cdot}5 + 4(1 - 0{\cdot}5\nabla)X_t$

Dynamic models with added noise

The relationship (19) between the input and output will usually be obscured by

noise due to measurement error and variation in other variables not under one's control. In this case, we can write (19) as

$$\xi(\nabla)\,\dot{Y}_{t+1} = g\eta(\nabla)\dot{X}_{t-b} + E_{t+1} \tag{23}$$

and (22) as

$$\dot{Y}_{t+1} = V(B)\dot{X}_t + F_{t+1}, \tag{24}$$

where E_{t+1} and F_{t+1} are supposed not correlated with the input process $\dot{X}_t$.

3.2. *Identification of Dynamic Models*

In the same way that the sample autocorrelation function can be used to identify univariate time series models, the basic tool in the identification of dynamic models is the *sample cross correlation function.* To describe a pair of time series by their cross correlation function it is necessary to assume that both series are stationary. Hence it is first necessary to difference both input and output d times until the resulting input and output series are stationary.

If this differencing operation is applied to both sides of (23) and (24) the differenced dynamic models become

$$\xi(\nabla)y_{t+1} = g\eta(\nabla)x_{t-b} + e_t \tag{25}$$

and

$$y_{t+1} = V(B)x_t + f_t, \tag{26}$$

where $y_t = \nabla^d \dot{Y}_t$, $x_t = \nabla^d \dot{X}_t$, $e_t = \nabla^d E_t$, $f_t = \nabla^d F_t$.

Suppose that after differencing, n pairs of differences $(x_1,y_1), (x_2,y_2) \ldots (x_n,y_n)$ are available. Then the sample cross correlation function at lag $+k$ is defined by

$$r_{xy}(k) = \frac{c_{xy}(k)}{\sqrt{\{c_{xx}(0)\,c_{yy}(0)\}}}, \quad k = 0, +1, +2, \ldots \tag{27}$$

where

$$c_{xy}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(y_{t+k} - \bar{y}), \tag{28}$$

and at lag $-k$ by

$$c_{xy}(-k) = c_{yx}(k),$$

where $\bar{x}$, $\bar{y}$ are the means of the x and y series.

Prewhitening of the input series

Suppose that it is assumed that the input x_t is uncorrelated with the noise in (26). Then, on multiplying throughout in (26) by x_{t-k+1} and taking expectations,

$$\gamma_{xy}(k) = V(B)\gamma_{xx}(k-1), \tag{29}$$

where $\gamma_{xy}(k)$, $\gamma_{xx}(k)$ are the theoretical cross covariance function and input autocovariance function respectively, and B now operates on k.

Suppose now that we carry out the usual identification and estimation methods as described in section 2 to obtain a model

$$\phi(B)\theta^{-1}(B)x_t = x'_t \tag{30}$$

which transforms the correlated input series x_t to a white noise series x'_t. Suppose also that this transformation is now applied to both sides of (26), yielding,

$$y'_{t+1} = V(B)x'_t + f'_t, \tag{31}$$

where x'_t is white noise uncorrelated with f'_t. On multiplying throughout in (31) by x'_{t-k+1} and taking expectations, we obtain

$$\gamma_{x'y'}(k) = v_k \sigma^2_{x'}. \tag{32}$$

In terms of the cross correlation function, (32) may be rewritten

$$v_k = \rho_{x'y'}(k) \frac{\sigma_{y'}}{\sigma_{x'}}. \tag{33}$$

Hence after "prewhitening", the cross correlation function is directly proportional to the impulse function.

The presence of small initial values of v_k is indicative of pure delay or dead time. Thereafter the presence of values of v_k not following a pattern indicates that terms should be introduced on the *right hand side* of the model (19) and the presence of exponential decay or damped sine wave behaviour in v_k indicates that terms should be introduced on the *left hand side* of the model (19).

3.3. *An Example of identifying a Dynamic Model*

Fig. 6 shows continuous records of the input airfeed (X) and the output carbon dioxide concentration (Y) from a gas furnace. The input airfeed was deliberately varied so as to follow an autoregressive process and the input and output records read at 9-sec intervals resulting in 226 pairs of observations.

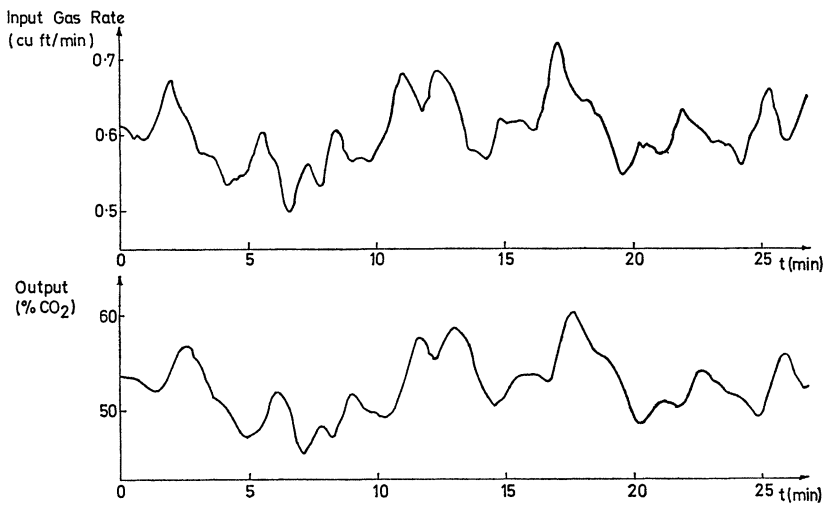

FIG. 6. Input and output records to a gas furnace.

The sample auto- and cross-correlation functions damped out fairly quickly indicating that no differencing was necessary. Hence $x_t = \dot{X}_t$, $y_t = \dot{Y}_t$. The usual identification and fitting procedure applied to the input indicates that it is a third order autoregressive process

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)x_t = x'_t$$

with $\hat{\phi}_1 = 1{\cdot}97$, $\hat{\phi}_2 = -1{\cdot}37$, $\hat{\phi}_3 = 0{\cdot}34$ and $s^2_{x'} = 0{\cdot}0353$.

Hence the transformations

$$x'_t = (1-1{\cdot}97B+1{\cdot}37B^2-0{\cdot}34B^3)x_t$$

$$y'_t = (1-1{\cdot}97B+1{\cdot}37B^2-0{\cdot}34B^3)y_t$$

were applied to the input and output series to yield the series x'_t and y'_t with $s_{x'} = 0{\cdot}188$, $s_{y'} = 0{\cdot}358$. The sample cross-correlation function between x' and y' is shown in Table 3 together with the estimate of the impulse response function obtained from (33), that is

$$v_k = \frac{0{\cdot}358}{0{\cdot}188} r_{x'y'}(k).$$

Fig. 7 shows the plot of v_k versus k and indicates that there are two whole periods of delay, then one or two preliminary values v_3 and v_4 which do not correspond to a pattern, followed by a decay pattern which could be first or second order.

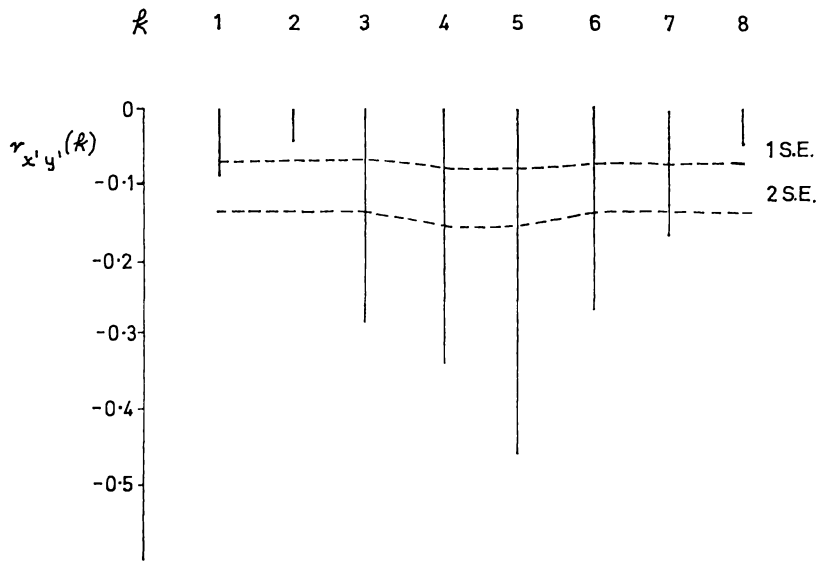

FIG. 7. Gas furnace data sample cross correlations after prewhitening.

TABLE 3

Cross correlation function and approximate impulse response function for gas furnace data

k	1	2	3	4	5	6	7	8
$r_{x'y'}(k)$	−0·05	−0·03	−0·28	−0·33	−0·46	−0·27	−0·17	−0·03
S.E. [r]	0·07	0·07	0·07	0·08	0·08	0·07	0·07	0·07
v_k	−0·09	−0·04	−0·53	−0·63	−0·88	−0·52	−0·32	−0·04

To help in the identification procedure, approximate standard errors for the sample cross correlations were computed using Bartlett's approximate formula

$$\operatorname{cov}\{r_{x'y'}(k), r_{x'y'}(l)\} \approx \frac{1}{N}\sum_{j=-\infty}^{+\infty}\{\rho_{x'x'}(j)\rho_{y'y'}(j+l-k)+\rho_{x'y'}(j+l)\rho_{y'x'}(j-k)\} \quad (34)$$

for the covariance between two values of the sample cross correlation at different lags k and l. On writing $k = l$ in (34) and making use of the fact that the x' series is white noise, the variance of a single cross correlation coefficient is given approximately by

$$\operatorname{var}\{r_{x'y'}(k)\} \approx \frac{1}{N}\left\{1+\sum_{j=-\infty}^{+\infty}\rho_{x'y'}(j+k)\rho_{y'x'}(-k)\right\}. \quad (35)$$

The standard errors given in Table 3 are based on the assumption that the cross correlations up to lag+2 and from lag+8 onwards are effectively zero. The one standard error and two standard error limits are plotted on Fig. 7 and confirm the identification of a dynamic model

$$(1+\xi_1\nabla+\xi_2\nabla^2)\,\dot{Y}_{t+1} = (1+\eta_1\nabla+\eta_2\nabla^2)\dot{X}_{t-b} \quad (36)$$

(probably with $b = 2$) or some simplification of it.

3.4. *Estimation of the Transfer Function*

In the first instance η_2 was set equal to zero in (36) and the model with added noise written as

$$\dot{Y}_{t+1} = \beta_1\dot{Y}_t+\beta_2\dot{Y}_{t-1}+\beta_3\dot{X}_{t-b}+\beta_4\dot{X}_{t-b-1}+E_{t+1}. \quad (37)$$

If the errors E_t were uncorrelated, then the parameters in (37) could be estimated by linear least squares. Under the added assumption that the E_t are Normal, these would also be maximum likelihood estimates. In practice the E_t's would rarely be uncorrelated, and hence it would be necessary to arrive at a model by iteration as is now illustrated for the gas furnace example.

Initially the model (37) was fitted by linear least squares for different values of the delay parameter b assuming that the errors E_t were uncorrelated. The minimum sum of squares was attained when $b = 1$, yielding the preliminary fitted model

$$\dot{Y}_{t+1} = 1{\cdot}39\,\dot{Y}_t-0{\cdot}55\,\dot{Y}_{t-1}-0{\cdot}14\dot{X}_{t-1}-0{\cdot}34\dot{X}_{t-2}+E_t,$$

where the dot notation is used to denote deviations of Y and X from their average values. The first ten autocorrelations of the residuals E_t from this model are given in Table 4.

TABLE 4

Autocorrelations of residuals from fitted dynamic model

k	1	2	3	4	5	6	7	8	9	10
r_k	·24	·18	·00	−·02	·01	·16	·08	·06	−·07	·00

These residuals might be explained by a first order autoregressive noise model

$$E_{t+1} = \phi E_t+a_{t+1}.$$

If so, then for an appropriate choice of ϕ we can rewrite the model (37) as

$$\tilde{Y}_{t+1} = \beta_1 \tilde{Y}_t + \beta_2 \tilde{Y}_{t-1} + \beta_3 \tilde{X}_{t-b} + \beta_4 \tilde{X}_{t-b-1} + a_{t+1} \tag{38}$$

where $\tilde{Y}_t = \dot{Y}_t - \phi \dot{Y}_{t-1}$, $\tilde{X}_t = \dot{X}_t - \phi \dot{X}_{t-1}$, and a_t is now white noise.

The model (38) was fitted to the transformed data $\tilde{Y}_{t+1} = \dot{Y}_{t+1} - \phi \dot{Y}_t$ and $\tilde{X}_{t+1} = \dot{X}_{t+1} - \phi \dot{X}_t$ for a grid of values of ϕ and b. The minimum sum of squares occurred at $b = 2$, $\phi = 0{\cdot}7$ yielding the model

$$\tilde{Y}_{t+1} = \underset{(\pm 0{\cdot}05)}{0{\cdot}90}\, \tilde{Y}_t - \underset{(\pm 0{\cdot}04)}{0{\cdot}19}\, \tilde{Y}_{t-1} - \underset{(\pm 0{\cdot}07)}{0{\cdot}48}\, \tilde{X}_{t-2} - \underset{(\pm 0{\cdot}10)}{0{\cdot}44}\, \tilde{X}_{t-3},$$

the figures in parentheses under the estimated parameters being their standard errors obtained from the usual least squares formula. The autocorrelations of the residuals a_{t+1} from this model were all small, confirming that the model is adequate.

Hence the final model is

$$(1-0{\cdot}7B)(1-0{\cdot}90B+0{\cdot}19B^2)\,\dot{Y}_{t+1} = -(1-0{\cdot}7B)(0{\cdot}48B^2+0{\cdot}44B^3)\dot{X}_t + a_{t+1}. \tag{39}$$

Rewriting (39) as

$$\dot{Y}_{t+1} = \frac{-(0{\cdot}48B^2+0{\cdot}44B^3)\dot{X}_t}{1-0{\cdot}90B+0{\cdot}19B^2} + \frac{a_{t+1}}{(1-0{\cdot}7B)(1-0{\cdot}90B+0{\cdot}19B^2)}, \tag{40}$$

we see that the fitted dynamic model is

$$(1-0{\cdot}90B+0{\cdot}19B^2)\,\dot{Y}_{t+1} = -(0{\cdot}48B^2+0{\cdot}44B^3)\dot{X}_t. \tag{41}$$

This model implies transfer function characteristics which agree very closely with those estimated in Jenkins and Watts (1968) using cross spectral analysis. In the control engineer's language it corresponds to a second order system with time constants $T_1 = 15{\cdot}8$ seconds and $T_2 = 8{\cdot}2$ seconds, and a pure delay or dead time of 22·8 seconds.

The model (40) also implies that the noise n_{t+1} at the output of the system is a third order autoregressive process

$$(1-0{\cdot}70B)(1-0{\cdot}90B+0{\cdot}19B^2)n_{t+1} = a_{t+1}. \tag{42}$$

A more direct fitting procedure which employs iterative non-linear least squares, and which is readily adapted to the analysis of multiple input data, is described in Box and Jenkins (1968).

References

Barnard, G. A. (1959). Control charts and stochastic processes. *J. R. Statist. Soc.* B, **21**, 239–271.

Bather, J. A. (1963). Control charts and the minimisation of costs. *J. R. Statist. Soc.* B, **25**, 49–80.

Box, G. E. P. and Jenkins, G. M. (1962). Some statistical aspects of adaptive optimisation and control. *J. R. Statist. Soc.* B, **24**, 297–343.

Box, G. E. P. and Jenkins, G. M. (1963). Further contributions to adaptive quality control: simultaneous estimation of dynamics: non-zero costs. *I.S.I. Bulletin*, 34th Session, Ottawa, Canada, 943–974.

Box, G. E. P. and Jenkins, G. M. (1965). Mathematical models for adaptive control and optimisation. *A.I.Ch.E.–I.Chem.E. Symposium* Series No. 4, 61–68.

Box, G. E. P., Jenkins, G. M., and Bacon, D. W. (1967). Models for forecasting seasonal and non-seasonal time series, pp. 271–311 of *Advanced Seminar on Spectral Analysis of Time Series*. Edited by B. Harris, New York: John Wiley.

BOX, G. E. P. and JENKINS, G. M. (1968). *Time Series Forecasting and Control*, San Francisco: Holden-Day. The original draft of this book has been issued in 1966 and 1967 as Technical Reports Nos. 72, 77, 79, 94, 95, 99, 103, 104, 116, 121 and 122 of the Department of Statistics, University of Wisconsin, Madison and simultaneously as Technical Reports Nos. 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 13 of the Department of Systems Engineering, University of Lancaster.

BROWN, R. G. (1962). *Smoothing, Forecasting and Prediction of Discrete Time Series*. New Jersey: Prentice-Hall.

HOLT, C. C. (1957). Forecasting trends and seasonals by exponentially weighted moving averages. O.N.R. Memorandum No. 52, Carnegie Institute of Technology.

HOLT, C. C., MODIGLIANI, F., MUTH, J. F., and SIMON, H. A. (1960). *Planning Production, Inventories and Work Force*, New Jersey: Prentice-Hall.

JENKINS, G. M. and WATTS, D. G. (1968). *Spectral Analysis and its Applications*. San Francisco: Holden-Day.

PAGE, E. S. (1954). Continuous inspection schemes. *Biometrika*, **41**, 100–114.

PAGE, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika*, **44**, 248–252.

ROBERTS, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, **1**, 239–250.

WHITTLE, P. (1963). *Prediction and Regulation*, London: English Universities' Press.

WINTERS, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, **6**, 324–342.