PAPADOPOULOS, PHAEDON PANAGIOTIS

A STUDY OF COMPARATIVE FORECASTING

The University of Oklahoma PH.D. 1980

THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A STUDY OF COMPARATIVE FORECASTING

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

BY

PHAEDON PANAGIOTIS PAPADOPOULOS

Norman, Oklahoma

1980

A STUDY OF COMPARATIVE FORECASTING

APPROVED BY

James F. Horrell

Rollin J. Watson

Raymond Dacey

C.B. Schwartzkopf

DISSERTATION COMMITTEE

ACKNOWLEDGEMENTS

I wish to express my appreciation to all the people who contributed toward the completion of my doctoral studies.

I am very grateful to Dr. Albert B. Schwarzkopf for helping me in many ways from the time I first came to the United States for graduate studies at The University of Oklahoma. My research is especially influenced by his approach to applied mathematics.

I am also grateful to Dr. Raymond Dacey for all the help and the support he provided while I worked and studied in the College of Business Administration. Because of his personality and academic excellence, I was persuaded to seek an interdisciplinary degree combining computer science, mathematics and business administration.

Dr. Collin Watson was always helpful and knowledgeable in applied statistics and computer approaches.

I am deeply indebted to Dr. James Horrell for his invaluable guidance and advice. Dr. Horrell, untiringly and meticulously, read the various drafts of my dissertation making numerous suggestions. His kindness, patience, friendliness and expertise made possible the success of my program.

I am extremely thankful to Ms. Julia Rojas, Ms. Patricia Wickham

and Mr. Warren Dickson for their time, patience and persistence in typing this manuscript.

I am also thankful to Ms. Alice Watkins for supervising as well as assisting in the compilation of this dissertation.

I wish to thank Paul Lewis for his precise and artistic rendering of the graphs in my dissertation.

I am grateful to my wife Isabel for her help, patience, understanding and moral support.

Finally, I would like to dedicate this dissertation to my parents, Panagiotis and Eugenia Papadopoulos, for everything they have done and continue to do for me. I would never have reached this point in my career without their help. I shall always be grateful to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS USED IN THIS STUDY

AEP	Adaptive Estimation Procedure
AF	Adaptive Filtering
AIC	Akaike Information Criterion
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
ARRSES	Adaptive Response Rate Single Exponential Smoothing
BJ	Box-Jenkins
LMS	Least Mean Square
MA	Moving Average
MAD	Mean Absolute Deviation
MAPE	Mean Absolute Percentage Error
MPE	Mean Percent Error
MSE	Mean Square Error
OLS	Ordinary Least Squares
SAFT	Self Adaptive Forecasting Technique
SS	State Space

INTRODUCTION

During the past two decades, there has been an increasing number of comparative forecasting studies. The objective of these studies is to compare different forecasting methodologies with the hope of finding the best methodology. These studies have led to conflicting reports and controversies.

This dissertation examines almost all published comparative studies and delineates a list of fallacies occurring in comparative forecasting studies. These fallacies most commonly give rise to the existing controversies. Since the controversies in forecasting stem from comparisons of the various approaches, a brief synopsis of the most currently employed univariate and multivariate methodologies are presented.

A STUDY OF COMPARATIVE FORECASTING

CHAPTER ONE

BOX-JENKINS METHODOLOGY

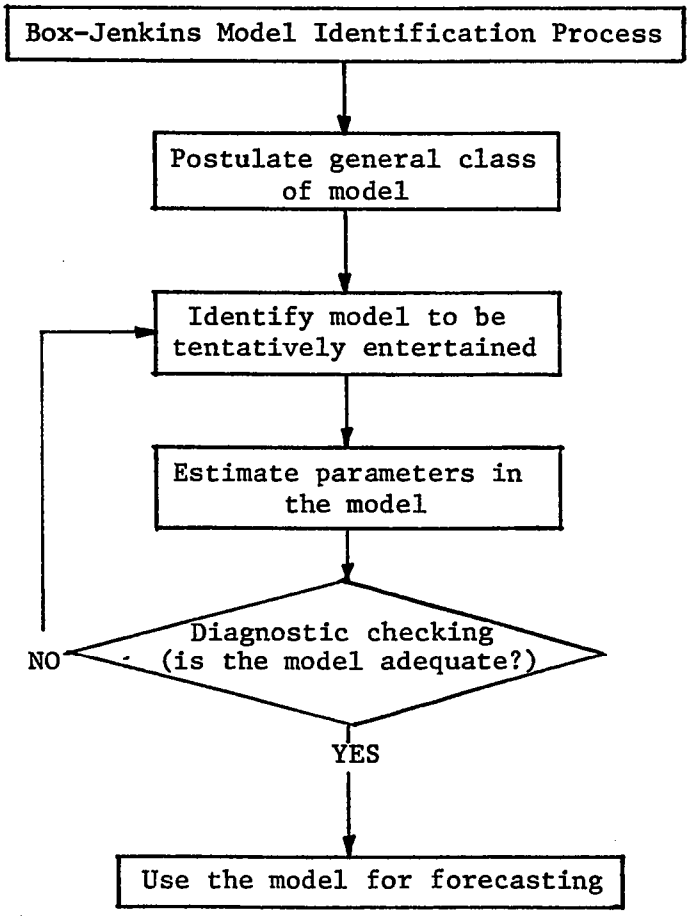
> The best qualification of a prophet is
> to have a good memory
> -Marquis of Halifax

1.1 Introduction

The Box-Jenkins methodology assumes that the reality giving rise to a historical time series pattern can be adequately and parsimoniously represented by a member of a large and rich class of models. Through an iterative approach, a possible model from the general class of models is identified and then checked against the data and assumptions underlying the general class of models for proper fit. If the specified model is satisfactory, it is used for forecasting; if not, the process is repeated using a modified tentative model suggested by residual analysis. The process is shown in Figure 1.0.

Gwilym M. Jenkins, in his book, "Practical Experiences with Modelling and Forecasting Time Series," gives an excellent presentation of the BJ methodology. This book together with other works by Box and Jenkins [67,11, 96, 83 and 46] constitutes the basis for the discussion to be presented in the following sections.

FIGURE 1.0



From: G.P. Box and G.M. Jenkins, Time Series Analysis Forecasting and Control, Holden Day, Inc., 1970, p. 19.

Section 1.2 describes the usefulness of univariate models for forecasting a time series from its own past history along with a brief mathematical description of the mathematical model.

Section 1.3 discusses the role of transfer function for relating an output time series which is to be forecast to a set of related input variables. These models, whose mathematical description is also found in section 1.3, enable a time series to be forecasted not only

from its own past history but also from the past history of other related variables.

Section 1.4 describes the intervention models, a class of models which can be used to represent unusual events such as a strike, a holiday or a change in definition of a variable.

Section 1.5 describes the objectives of a class of models called multivariate stochastic models which can represent several output series with mutual interactions or feedback. A mathematical description of these models is given also in section 1.5.

1.2 Univariate Stochastic (single output) Models

The simplest forecasting equation occurs when one is asked to forecast the future of a time series from a knowledge of its past history alone. Such a simple-minded approach could be questionable in accuracy when very accurate forecasts over the long term were expected. Jenkins [66] claims that univariate stochastic models, despite their simplicity, are important for the following reasons:

i. In some situations, it may be the only feasible practical approach to adopt because of the sheer magnitude of the problem.

ii. In other situations, it may be impossible to find variables which are related to the variable being forecast leaving the univariate model as the only means of forecasting.

iii. In any case, the development of a univariate model provides a "yardstick" with which more sophisticated models can be compared.

iv. Univariate models can be used for 'screening' data during the early stages of an analysis. The presence of large residuals, for example, in a univariate model, may correspond to abnormal events such as a strike, or to faulty data.

A mathematical description of univariate models will be given below as they apply to stationary, non-stationary and seasonal time series.

Stationary Models. A stationary time series can be represented by a wide class of models, linear in the transformed variable, called autoregressive-moving average (ARMA) models, that is:

$$(Z^*_t - c) = \phi_1(Z^*_{t-1} - c) + \ldots + \phi_p(Z^*_{t-p} - c) + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \qquad (1.2.1)$$

where $Z^*_t = Z_t^{(\lambda)}$

c is the constant mean of the series, and

$Z_t^{(\lambda)}$ represents the class of power transformations defined by

$$Z_t^{(\lambda)} = \begin{cases} Z_t^{\lambda}, & \lambda \neq 0 \\ \ln Z_t, & \lambda = 0 \end{cases} \qquad (1.2.2)$$

where λ is a vector or parameters defining the transformation. The main objective of the transformation is to produce residuals in the

fitted model that have a constant variance. It is required that the transformed residuals have a common probability distribution [66, p. 95].

Model (1.2.1) represents the current value of the transformed series as a linear function of:

(a) past values of the transformed series $Z_t^{(\lambda)}$,

(b) current and past values of the residuals a_t and can be written in operator form as:

$$Z_t^{(\lambda)} - c = \frac{\theta(B)}{\phi(B)} a_t = \frac{1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q}{1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p} a_t \quad (1.2.3)$$

where the backward shift operator B is defined by:

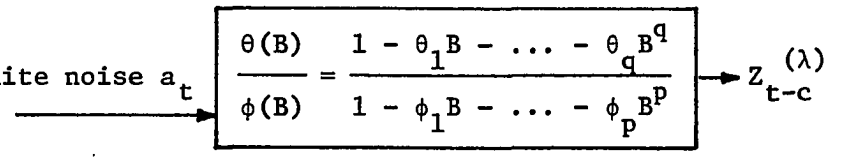
$$B^j a_t = a_{t-j}$$

The parameters in equation (1.2.3) need to satisfy the following conditions:

i. The MA(q) process is stationary regardless of the values of the the weights $\{\theta_i\}$, but it is invertible only if the roots of $\theta_q(B) = 0$ outside the unit circle where $\theta_q(B)$ is defined as $\theta_q(B) = (1 - \theta_1 B^1 - \theta_2 B^2 - \ldots - \theta_q B^q)$.

ii. The AR(p) process is stationary only if the roots of $\phi(B) = 0$ lie outside the unit circle, but it is invertible for all values of the weights $\{\phi_i\}$ where $\phi_p(B)$ is defined as $\phi_p(B) = (1 - \phi_1 B^1 - \phi_2 B^2 - \ldots - \phi_p B^p)$

Pictorially we can represent the transformed series $Z_t^{(\lambda)}$ as the output from a linear filter whose input is a random series with zero mean and constant variance ("white noise") and whose filter function is a ratio of $\theta(B)$ and $\phi(B)$.



To achieve parsimony, that is a representation which economizes in the use of parameters, it is necessary to include, in general, both AR and MA components in the model. In contrast, "the use of an AR model to represent series which is described by an MA model, or vice versa, will result in the prodigal use of parameters" [66, p. 98].

Non-Stationary Models. Many time series behave as if they have no constant mean. Such time series are called non-stationary in the mean.

Successive differencing may reduce a non-stationary stochastic time series to a stationary time series. If a non-stationary time series can be reduced to a stationary series by applying a suitable degree of differencing, we say the original series is homogeneously non-stationary. Thus, a class of models, useful for representing a wide range of practical situations, can be obtained by first differencing the transformed series d times to induce stationarity, that is,

$$W_t = \nabla^d Z_t^{(\lambda)} \tag{1.2.4}$$

The stationary series W_t can then be represented by an ARMA model

$$W_t - c = \frac{\theta(B)}{\phi(B)} a_t = \frac{1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q}{1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p} a_t \qquad (1.2.5)$$

Equations (1.2.4) and 1.2.5) define a model known as the Autoregressive Integrated Moving Average model or ARIMA (p,d,q) model. Yaglom [66, p. 99] has described models involving differencing as accumulated processes. Such models are capable of describing a wide class of stochastic trends, whose coefficients adapt as each observation comes to hand. Thus:

> . . . models involving single differencing ∇ can be used to describe series whose level is continuously updated by random shocks; models involving double differencing ∇^2 can describe series whose level and slope are continuously updated by random shocks, and so on [66, p. 99].

In most cases, however, single differencing is adequate to describe most non-stationary time series.

Seasonal Models. Jenkins and Watts [68] and Box and Jenkins [11] proposed models capable of dealing with seasonal series. In order to describe series containing seasonal patterns with period s, whether stationary or not, they developed a new class of models: the seasonal $(p,d,q)X(P,D,Q)_s$ model defined by

$$W_t = \nabla^d \nabla_s^D Z_t^{(\lambda)} \qquad (1.2.6)$$

$$W_t - c = \frac{\theta(B)\ \Theta(B^s)}{\phi(B)\ \Phi(B^s)} a_t \qquad (1.2.7)$$

where $\phi(B)$, $\theta(B)$ are non-seasonal AR and MA operators as defined in

(1.2.3), and

$$\nabla_s Z_t^{(\lambda)} = Z_t^{(\lambda)} - Z_{t-s}^{(\lambda)} \tag{1.2.8}$$

is the seasonal differencing operator, and

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_p B^{Ps} \tag{1.2.9}$$

$$\Theta(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \ldots - \Theta_Q B^{Qs} \tag{1.2.10}$$

are seasonal AR and MA average operators, and constant c measures the mean of the appropriately transformed and differenced series W_t.

Figure 1.1 shows the filter of the seasonal model defined by equation (1.2.6) and 1.2.7). Also Figure 1.2 shows a flow diagram for building univariate models.

1.3 Transfer Function (single output - multiple input) Noise Models

The objective of transfer function models is to describe methods for estimating dynamic relationships between

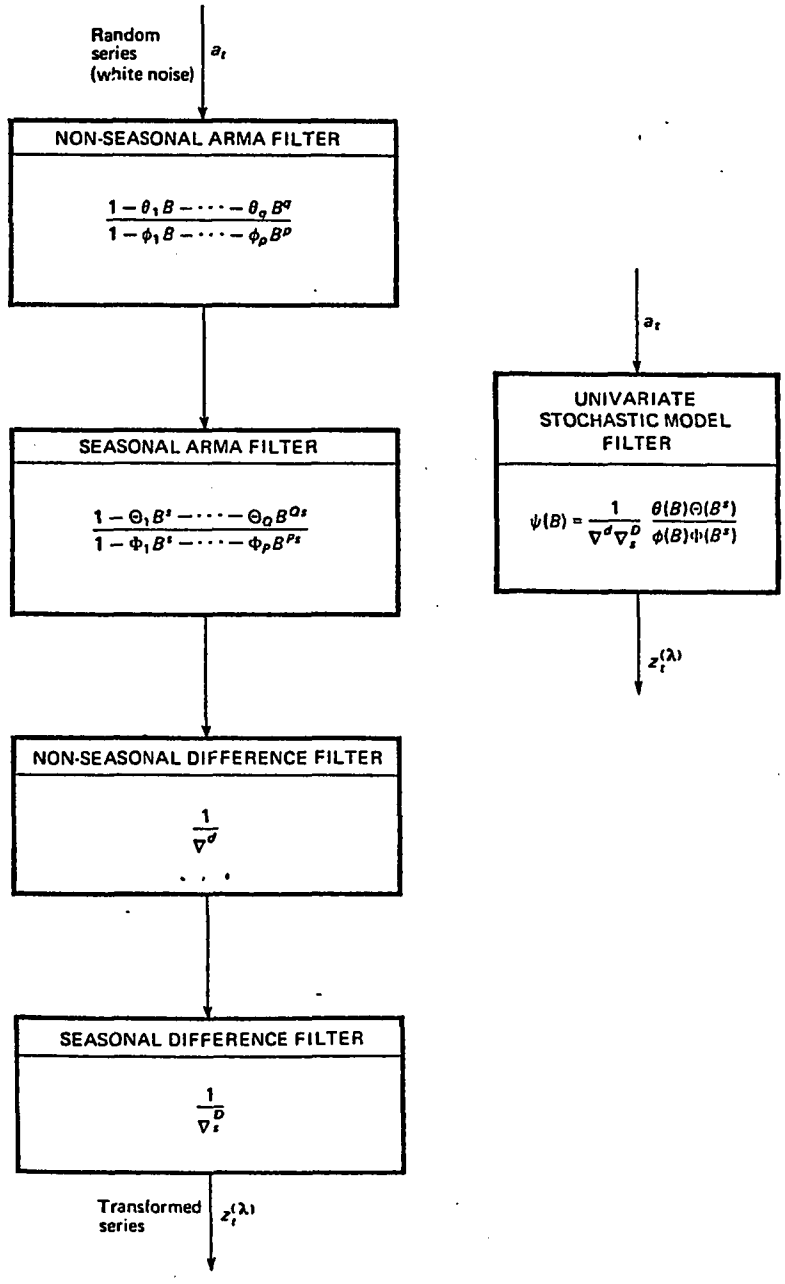
i. output variable Y_t and

ii. input variables $X_{1t}, X_{2t}, \ldots X_{kt}$

When a system is operating in open loop, there is no feedback between output and inputs. In other words, there is a univariable flow from $X_{1t}, X_{2t}, \ldots, X_{kt}$ to Y_t. Graphically the Transfer Function Model for the single output - single input case is illustrated in Figure 1.3. Y_t can be split into two components, U_t and N_t such that:

FIGURE 1.1

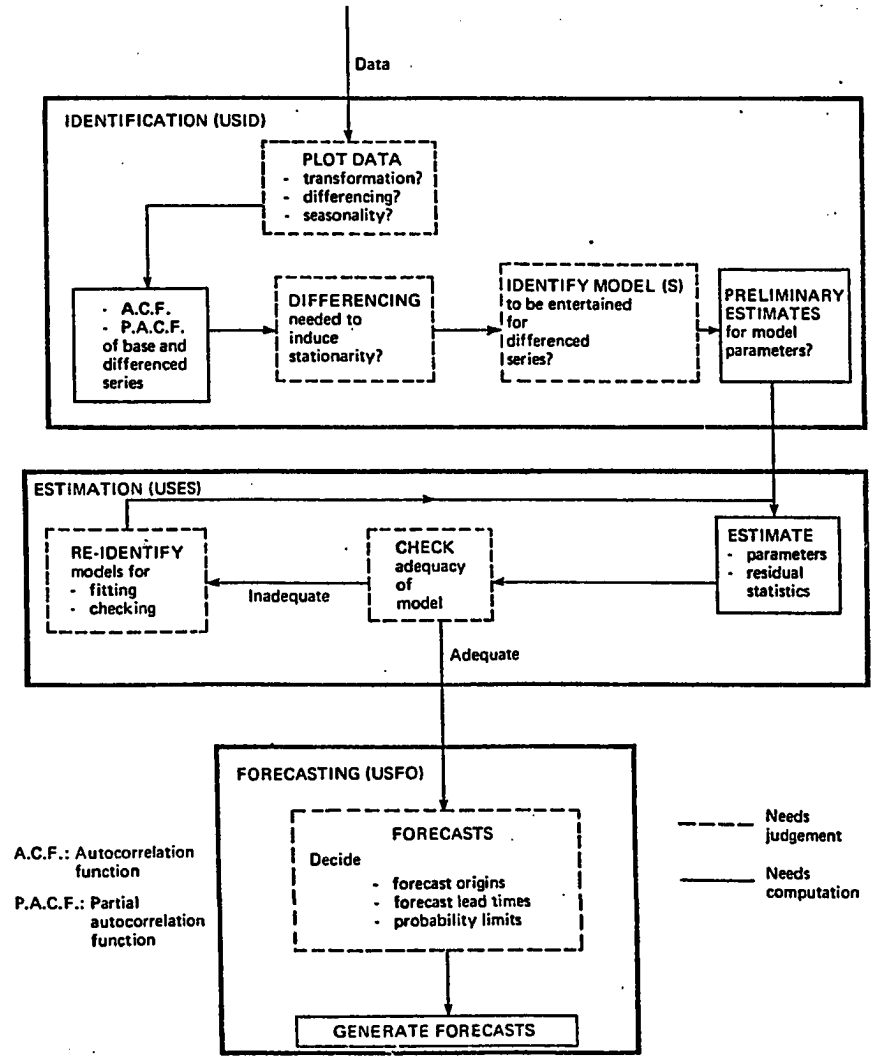
FILTER REPRESENTATION OF
SEASONAL UNIVARIATE STOCHASTIC MODEL



From: Jenkins, Practical Experiences with Modelling and Forecasting Time Series

FIGURE 1.2

FLOW DIAGRAM FOR UNIVARIATE STOCHASTIC MODEL BUILDING AND FORECASTING, BASED ON THREE COMPUTER PROGRAMS: USID (UNIVARIATE STOCHASTIC IDENTIFICATION PROGRAM), USES (UNIVARIATE STOCHASTIC ESTIMATION PROGRAM), USFO (UNIVARIATE STOCHASTIC FORECASTING PROGRAM).



From: Jenkins, Practical Experiences with Modelling and Forecasting Time Series

$$Y_t = U_t + N_t \qquad (1.3.1)$$

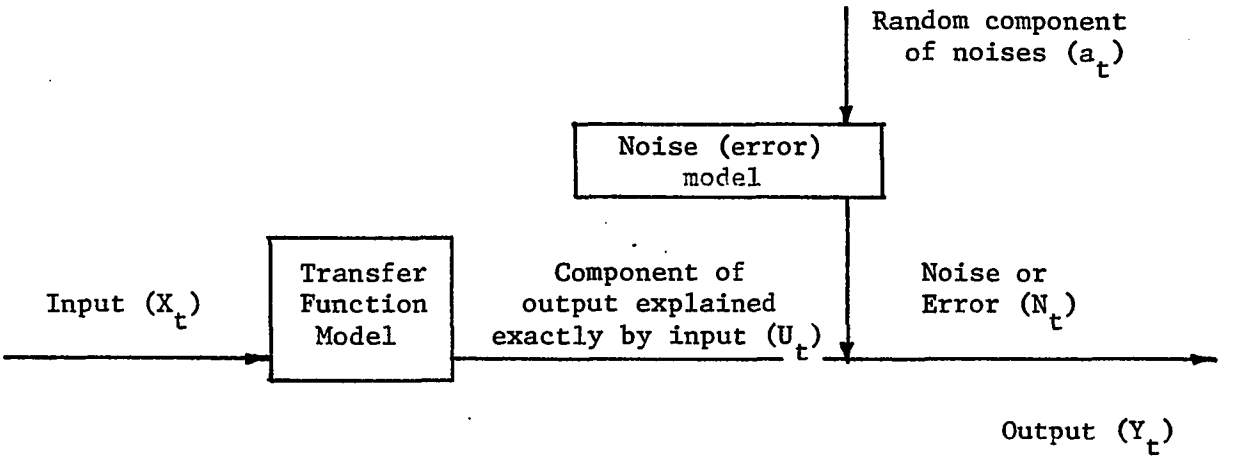
where Y_t is the observed output

U_t is that part of Y_t which can be explained exactly in terms of X_t

N_t is that part of Y_t that cannot be explained in terms of X_t and is called the noise (or disturbance).

FIGURE 1.3

SCHEMATIC REPRESENTATION OF SINGLE OUTPUT, SINGLE INPUT TRANSFER FUNCTION-NOISE MODEL



From: Jenkins, Practical Experiences with Modelling and Forecasting Time Series

A general way of representing a linear dynamic relationship between U_t and X_t is

$$U_t - \delta_1 U_{t-1} - \cdots - \delta_r U_{t-r} = W_o X_{t-b} - W_1 X_{t-b-1} - \cdots - W_s X_{t-b-s}$$

that is

$$U_t = \frac{W_o - W_1 B - \ldots - W_s B^s}{1 - \delta_1 B - \ldots - \delta_r B^r} X_{t-b} = \frac{W(B)}{\delta(B)} X_{t-b} = U(B)X_t$$

where $U(B) = \frac{W(B)}{\delta(B)} B^b$ [66, p. 100]

The transfer function U(B) contains

i. a moving average operation W(B)

ii. an autoregressive operation $\delta(B)$

iii. a pure delay parameter b, which represents the number of complete time intervals before a change in X_t begins to have an effect on Y_t.

If in equation (1.3.1) we allow for the need to transform the variables $X_{1t}, X_{2t}, \ldots, X_{kt}, Y_t$, as well as differencing X_t different than Y_t, we come up with the general form

$$\nabla^{d_{Y_t}}({}^{\lambda}Y) = U_t + N_t = \frac{W(B)}{\delta(B)} \nabla^{dl_{X_{t-b}}}({}^{\lambda}X) + N_t \qquad (1.3.2)$$

If we assume that N_t is non-stationary, then we can represent N_t by an ARIMA (p,d,q) model

$$\nabla^{dN} N_t = c + \frac{\theta(B)}{\phi(B)} a_t \qquad (1.3.3)$$

Multiplying (1.3.2) by ∇^{dN}, we get

$$\nabla^{dN}\nabla^{d_{Y_t}}({}^{\lambda}Y) = \frac{W(B)}{\phi(B)} \nabla^{dN}\nabla^{dl_{X_{t-b}}}({}^{\lambda}X) + \nabla^{dN} N_t \text{ or}$$

$$\nabla^{dN+d}{}_{Y_t}(^{\lambda}Y) = \frac{W(B)}{\phi(B)} \nabla^{dN+d1}{}_{X_{t-b}}(^{\lambda}X) + \nabla^{dN}{}_{N_t}$$

if we set $dY = dN+d$

$dX = dN+d_1$

$y_t = \nabla^{dY}{}_{Y_t}(^{\lambda}Y)$

$x_t = \nabla^{dX}{}_{X_t}(^{\lambda}X)$

and $\nabla^{dN}{}_{N_t}$ with its equal from equation (1.3.3), we get

$$y_t = c + \frac{W(B)}{\phi(B)} x_{t-b} + \frac{\theta(B)}{\phi(B)} a_t \qquad (1.3.4)$$

Equation (1.3.4) may be generalized for several input variables $X_{1t}, X_{2t}, \ldots, X_{kt}$ to

$$y_t = c + \sum_{j=1}^{k} \frac{W_j(B)}{\delta_j(B)} (x_{j,t-b_j} - \bar{x}_j) + \frac{\theta(B)}{\phi(B)} a_t \qquad (1.3.5)$$

each X-variable having a transfer function with its own moving average operator $W_j(B)$, AR operator $\delta_j(B)$ and pure delay b_j.

Seasonal Transfer Function Noise Models. The model may be seasonal because:

i. Input is seasonal

ii. Noise is seasonal

iii. Transfer function is seasonal

iv. Combination of the above.

If we denote the seasonal period with s, the general seasonal transfer

function model can be written as [67]:

$$\nabla^{d}Y_{t}^{(\lambda_Y)} = \sum_{j=1}^{k} \frac{W_j(B)}{\delta_j(B)} \nabla^{d_j} X_{jt}^{(\lambda_{X_j},t)} + N_t$$

where $$\nabla^{dN}\nabla_{s}^{DN}N_t = \frac{\theta(B)\,\Theta(B^s)}{\phi(B)\,\Phi(B^s)}\,a_t \qquad (1.3.6)$$

Figure 1.4 shows the non-seasonal transfer function-noise model for one input. Figure 1.5 shows a filter representation of the seasonal multiple input transfer function model (1.3.6) and Figure 1.6 shows a flow diagram for building transfer function models.

1.4 Intervention Models

The objectives of Interventional analysis makes allowances in time series models for large external events such as

i. a strike

ii. a sales promotion

iii. a change in a law (introduction of a new law)

In order to quantify such external events, we introduce to the model the following "dummy variables."

i. "pulse" variable which is set equal to 1 when an anomalous event occurs and 0 otherwise.

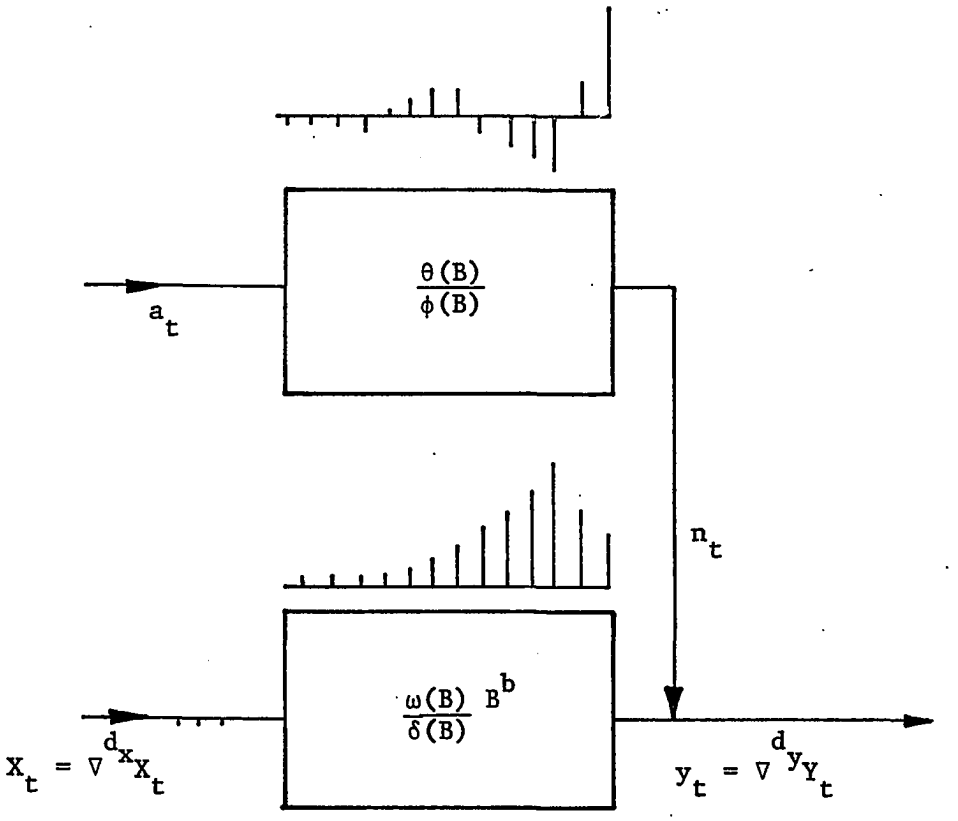
ii. "step" variables, set to 0 before a change (such as a policy change, or a new law or a change in definition in an economic variable) and to 1 after such a change [66].

The above events, if left out, would cause large residuals or distortion of model structure and parameter estimates.

FIGURE 1.4

NON-SEASONAL TRANSFER FUNCTION - NOISE MODEL FOR ONE INPUT

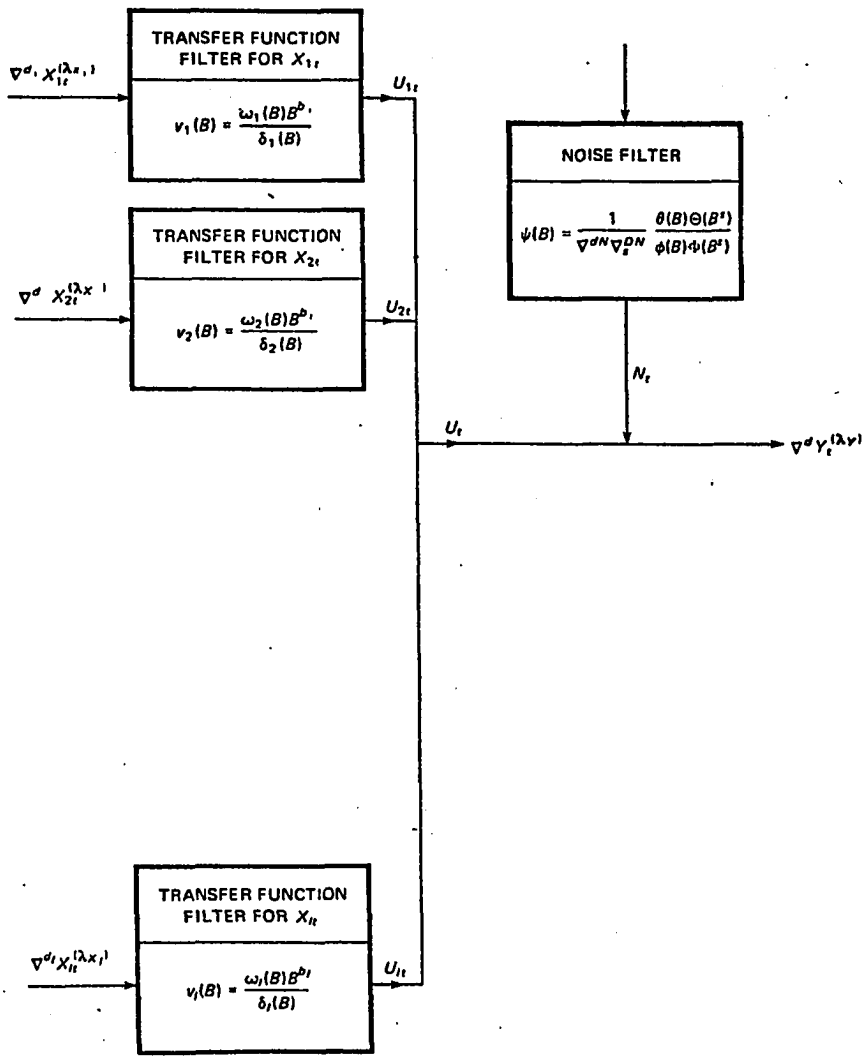
$$y_t = c + \frac{\omega(B)}{\delta(B)} X_{t-b} + \frac{\theta(B)}{\phi(B)} a_t$$



From: Jenkins, The Theory and Practical Application of Univariate and Transfer Function Analysis

FIGURE 1.5

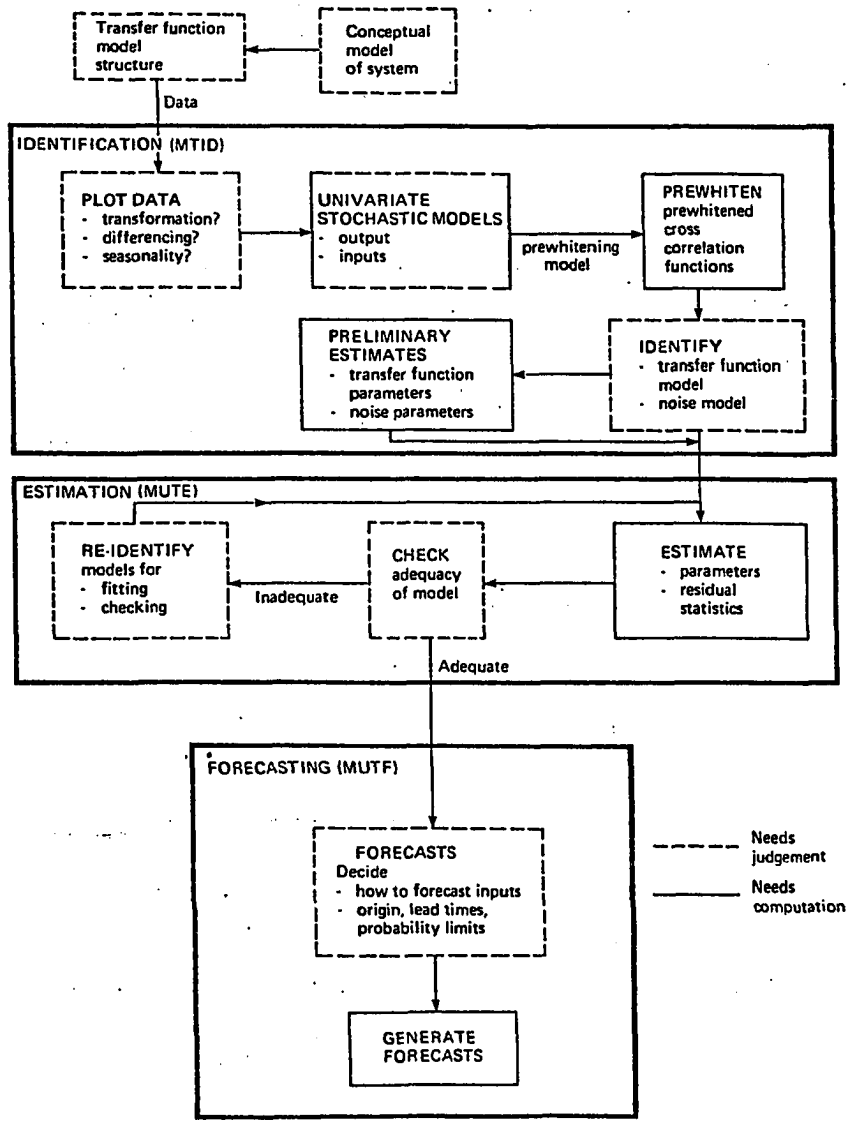
FILTER REPRESENTATION OF SEASONAL MULTIPLE INPUT TRANSFER FUNCTION MODEL



From: Jenkins, Practical Experiences with Modelling and Forecasting Time Series.

FIGURE 1.6

FLOW DIAGRAM FOR TRANSFER FUNCTION MODEL BUILDING AND FORECASTING BASED ON THREE COMPUTER PROGRAMS: MTID (MULTIPLE INPUT TRANSFER FUNCTION IDENTIFICATION), MUTE (MULTIPLE INPUT TRANSFER FUNCTION ESTIMATION), MUTF (MULTIPLE INPUT TRANSFER FUNCTION FORECASTING).



From: Jenkins, Practical Experiences with Modelling and Forecasting Time Series.

To investigate the effect of such an intervention variable ξ_t on the variable being modelled, we may postulate a lag structure

$$Y_\tau^{(\lambda)} = \frac{W(B)}{\delta(B)} \xi_{t-b} \tag{1.4.1}$$

whose parameters can be estimated as in a transfer function model.

In order to identify intervention models, we inspect the data and the residuals. As a result of a known external event, inspection of the data may suggest ways in which that event has changed the course of the series [66, p. 107].

> As an example, inspection of many consumer price indices may indicate that the dramatic oil price increase in the last quarter of 1973 was responsible for the consumer price indices rate of change increase. The intervention model
>
> $$\nabla Y_t^{(\lambda)} = W_o \xi_t \tag{1.4.2}$$
>
> can depict that effect, where ξ_t is a step function between 0 and 1 and the point of the so-called 'oil crisis.'

Examination of the residuals from the model fitted before an intervention variable is introduced can also indicate need for possible intervention model. For example, a large negative residual followed by a large positive residual in a univariate model may be due to a loss of sales during the period of a "strike" and a catching-up in deliveries in the period following the strike [66, p. 107]. Such an effect may be described by the model:

$$Y_t^{(\lambda)} = (W_o - W_1 B)\xi_t \tag{1.4.3}$$

where ξ_t is a pulse of unit height at the point where the strike

occurred. Figure (1.7) shows examples of the effect of the Y_t variables in equation (1.4.1) which can be modelled by simple transfer function models when the intervention variable ξ_t is a step or pulse.

Introducing the Noise into an Intervention Model. Suppose that the univariate model

$$\nabla\nabla_{12} Y^{(\lambda)} = (1-\theta B)\ (1-\theta B^{12})\ a_t \qquad (1.4.4)$$

has been fitted to a series, excluding the period when the abnormal event occurred. If the intervention mechanism was defined by (1.4.3), then we could postulate a model

$$Y_t^{(\lambda)} = (W_o - W_1 B)\ \xi_t + N_t \qquad (1.4.5)$$

where N_t is a noise term describing the behavior of the series in the absence of the abnormal event. Assuming that $W_o = W_1 = 0$ and combining (1.4.4) and (1.4.5) we get

$$\nabla\nabla_{12} N_t = (1-\theta B)\ (1-\theta B^{12})\ a_t \qquad (1.4.6)$$

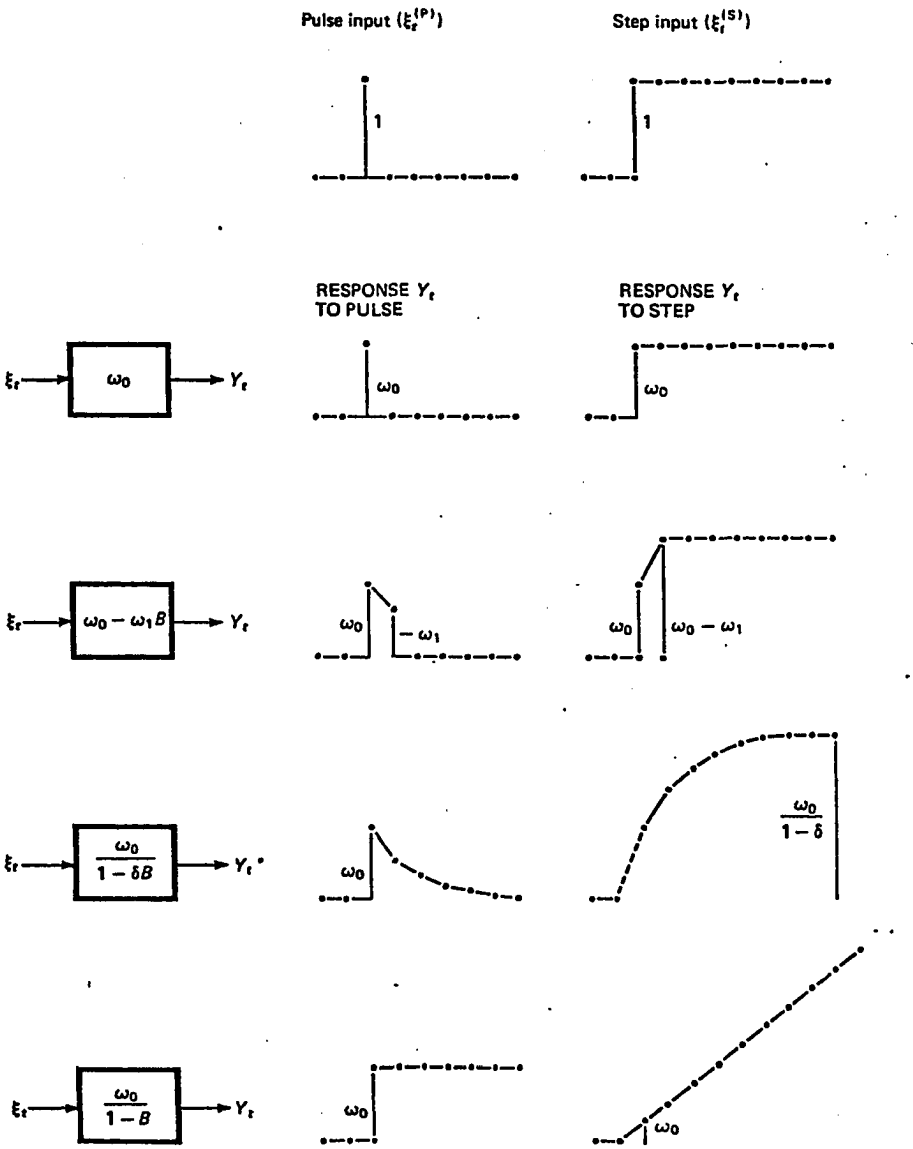
Combining again (1.4.6) with (1.4.5), we obtain the overall intervention model

$$\nabla\nabla_{12} Y_t^{(\lambda)} = (W_o - W_1 B)\ \nabla\nabla_{12}\xi_t + (1-\theta B)(1-\theta B^{12})\ a_t$$

Similarly, equation (1.4.2) can be formulated as

FIGURE 1.7

EXAMPLES OF DYNAMIC EFFECTS WHICH CAN BE SIMULATED IN INTERVENTION ANALYSIS USING A 'PULSE' INPUT AND STEP INPUT



From: Jenkins, Practical Experiences with Modelling and Forecasting Time Series

$$\nabla Y_t^{(\lambda)} = W_o \xi_t + N_t' \tag{1.4.7}$$

Setting $W_o=0$ and solving for N_t' equation (1.4.7) becomes

$$N_t' = \nabla Y_t^{(\lambda)} \tag{1.4.8}$$

Substituting equation (1.4.8) into equation (1.4.4), we get

$$\nabla_{12} N_t' = (1-\theta B)(1-\theta B^{12})\, a_t \tag{1.4.9}$$

Finally, combining (1.4.8) and (1.4.9), we get the overall intervention model

$$\nabla\nabla_{12} Y_t^{(\lambda)} = W_o \nabla_{12} \xi_t + (1-\theta B)(1-\theta B^{12})\, a_t \tag{1.4.10}$$

1.5 Multivariate Stochastic Models

The assumptions underlying the transfer function model, a completely unidirectional model, may not always be justified in practice due to feedback between the output and the inputs. As an example, Jenkins [66, p. 21] provides the well known pair of time series consisting of the annual number of hogs sold in the United States and the corresponding price of hogs on January 1st of each year. An increase in the number of hogs at a particular year may bring down the price of hogs the following year due to an excess of supply. Conversely, however, if the price of hogs falls at a particular year, the farmers, due to a lack of incentive, will probably let the number of hogs go down also. Here the input variable X_t (number of hogs) affects the output variable

Y_t (price of hogs) and vice-versa.

In situations of this kind, it is important to treat both variables (in general, the several variables involved) on an equal or reciprocal basis so that the two way feedback between each pair of variables can be disentangled. This requires the building of multivariate stochastic models (or multiple output models) to describe the mutual dependence between the variables [66, p. 22].

M.S. Bartlett [5, 7] and M.H. Quenouille [66, p. 109) were the pioneers in the area of multivariate stochastic models. Quenouille generalized the ARMA model of

$$Z_t^{(\lambda)} - c = \frac{\theta(B)}{\phi(B)} = \frac{1-\theta_1 B - \ldots - \theta_q B^q}{1-\phi_1 B - \ldots - \phi_p B^p} a_t$$

from its univariate form to

$$\underline{\phi_o Z_t} = \underline{\phi_1 Z_{t-1}} - \ldots + \underline{\phi_p Z_{t-p}} + \underline{\theta_o a_t} - \underline{\theta_1 a_{t-1}} \ldots$$

$$- \underline{\theta_q a_{t-q}} \qquad (1.5.1)$$

where Z_t is a column vector whose transpose $\underline{Z_t'} = (Z_{it}, Z_{2t}, \ldots, Z_{nt})$ is a low vector of n series; $\underline{\phi_i}$, $\underline{\theta_j}$ are nxn autoregressive and moving average matrices respectively; $\underline{a_t}$ is a vector whose elements a_{it} are mutually uncorrelated at all times.

Model (1.5.1) provides a useful starting point but is inadequate for two basic reasons. Firstly, model (1.5.1) constrains the univariate models for the individual time series Z_{it} to have AR operators

which have the same order and same parameter values [66, p. 110]. Secondly, model (1.5.1) assumes stationarity; that is, the n time series are in statistical equilibrium about fixed means.

A. Alavi [1] in his Ph.D. thesis, "Some Multivariate Extensions of Box-Jenkins Forecasting," suggested the following way to remove the former highly undesirable constraint:

Write model (1.5.1) in the form

$$\begin{bmatrix} \phi_{11}(B) & \phi_{12}(B) & \cdots & \phi_{1n}(B) \\ \phi_{21}(B) & \phi_{22}(B) & \cdots & \phi_{2n}(B) \\ \vdots & \vdots & & \vdots \\ \phi_{n1}(B) & \phi_{n2}(B) & \cdots & \phi_{nn}(B) \end{bmatrix} \begin{bmatrix} Z_{1t}-c_1 \\ Z_{2t}-c_2 \\ \vdots \\ Z_{nt}-c_n \end{bmatrix} =$$

$$\begin{bmatrix} \theta_{11}(B) & \theta_{12}(B) & \cdots & \theta_{1n}(B) \\ \theta_{21}(B) & \theta_{22}(B) & \cdots & \theta_{2n}(B) \\ \vdots & \vdots & & \vdots \\ \theta_{n1}(B) & \theta_{n2}(B) & \cdots & \theta_{nn}(B) \end{bmatrix} \begin{bmatrix} a_{1t} \\ a_{2t} \\ \vdots \\ a_{nt} \end{bmatrix}$$

or $$\underline{\phi}(B)(\underline{Z}_t-\underline{c}) = \underline{\theta}(B)\underline{a}_t \qquad (1.5.2)$$

where the AR operator $\phi_{ij}(B)$ is a polynomial of degree p_{ij} in the backward shift operator B and the MA operator $\theta_{ij}(B)$ is a polynomial of degree q_{ij} in (B). In (1.5.2), which will be referred to as a multivariate ARMA model or ARMA $(\underline{P},\underline{Q})$, the polynomials in the diagonal positions start with unity while the polynomials in the off-diagonal

positions start with a power of B thus making the a_{it} the one-step-ahead forecast errors. Those errors then must be allowed to have a covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 \cdots & \sigma_{2n} \\ \sigma_{n1} & \sigma_{n2} & \sigma_n^2 \end{bmatrix} \tag{1.5.3}$$

with $\sigma_{ij}=\sigma_{ji}$, but otherwise mutually uncorrelated at non-simultaneous times [66, p. 111].

As far as the second disadvantage of model (1.5.1) is concerned, namely the fact that model (1.5.1) assumes stationarity, we can generalize the model (1.5.2) to

$$\underline{\phi}(B)\ (\underline{W}_t - \underline{c}) = \underline{\theta}\ (B)\ \underline{a}_t \tag{1.5.4}$$

where $$\underline{W}_t' = (\nabla^{d_1} Z_{it}^{(\lambda_1)}) \ \ldots, \ \nabla^{dn} Z_{nt}^{(\lambda_n)})$$

with $\underline{\phi}(B)$ and $\underline{\theta}(B)$ as defined in (1.5.2) and $\underline{c}$ is a vector of constants. Such a model will be referred to as a multivariate ARIMA $(\underline{P},\underline{d},\underline{Q})$ model where the matrices $\underline{P} = (p_{ij})$, $\underline{Q} = (q_{ij})$ determine the degrees of the polynomials in the AR and MA matrices, and the row vector $\underline{d}' = (d_1, d_2, \ldots, d_n)$ has elements corresponding to the degrees of differencing required to induce stationarity in each of the individual time series.

Seasonal Multivariate Stochastic Models. As in the univariate

and transfer function models it is possible to extend the seasonal and non-seasonal differencing to multivariate stochastic models to induce stationary. The element W_{it} of the vector W_t in (1.5.4) needs to be defined as

$$W_{it} = \nabla^{di} \nabla_s^{Di} Z_{it}^{(\underline{\lambda}_i)} \tag{1.5.5}$$

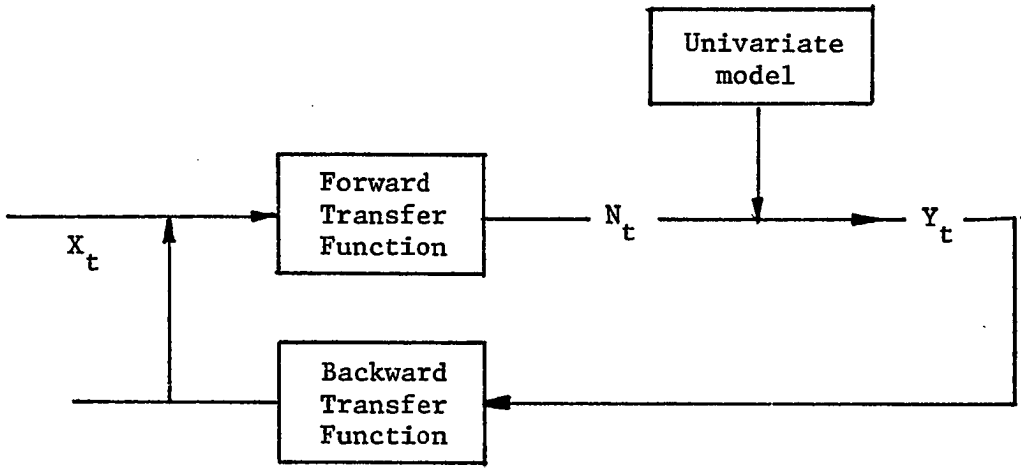
Stationarity and Invertibility Conditions. The conditions that the multivariate stochastic models need to satisfy are similar to the ones satisfied by the univariate models. Namely, the parameters in the AR matrix $\underline{\phi}(B)$ in (1.5.4) must satisfy the condition that the roots of $|\underline{\phi}(B)|=0$ lie outside the unit circle. Also, the parameters in the MA matrix $\underline{\theta}(B)$ in (1.5.4) must satisfy the condition that the roots of $|\underline{\theta}(B)|=0$ lie outside the unit circle.

The stationarity condition of the multivariate stochastic models ensures that the statistical properties of the differenced time series are time invariant; whereas, the invertibility condition ensures that, if the model (1.5.4) is used to generate simultaneous forecasts of the transformed series $Z_{it}^{(\underline{\lambda})}$, the weights applied to previous observations will die out as we stretch further into the past [66, p 112].

Figure 1.8 illustrates the Two-Way Feedback multivariate stochastic model. Figure 1.9 shows a multivariate stochastic model with multiple output, whereas Figure 1.10 shows a multivariate transfer function model with multiple output - multiple input, and Figure 1.11 shows the filter representation of seasonal multivariate stochastic model.

FIGURE 1.8

TWO-WAY FEEDBACK



From "The Theory and Practical Application of Multivariate and Multivariate Transfer Function Analysis" by Gwilym Jenkins and Partners Ltd.

FIGURE 1.9

MULTIPLE OUTPUT

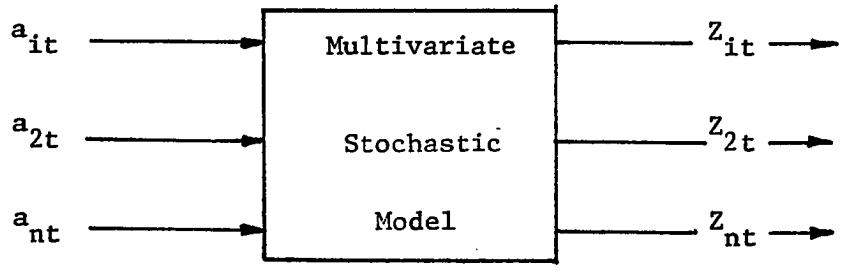
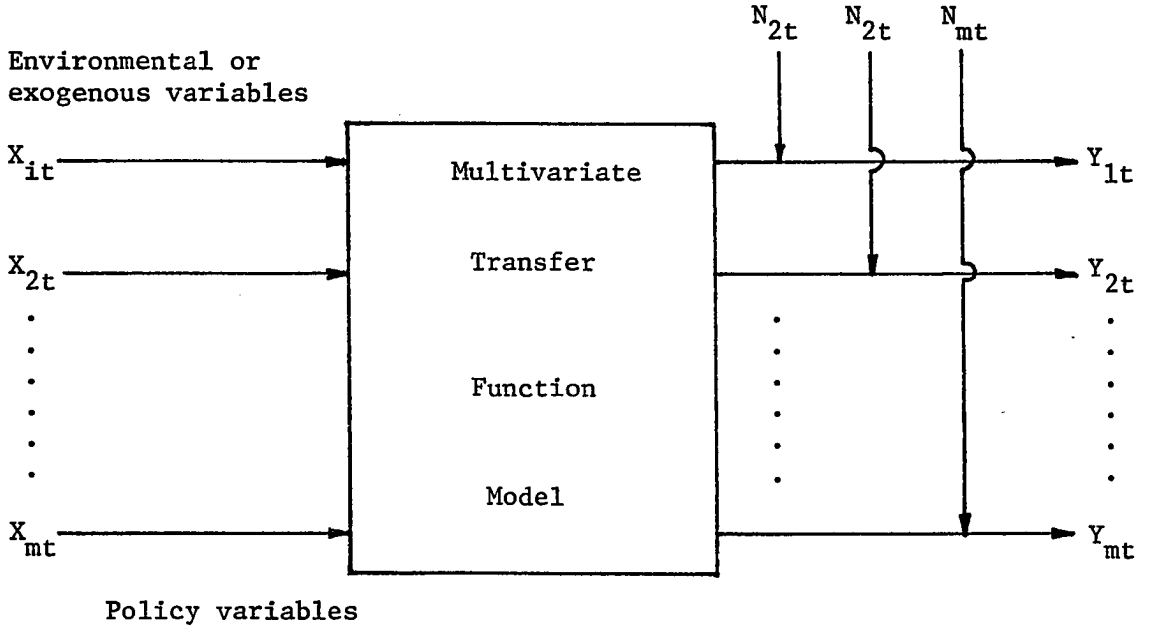


FIGURE 1.10

MULTIVARIATE TRANSFER FUNCTION MODEL

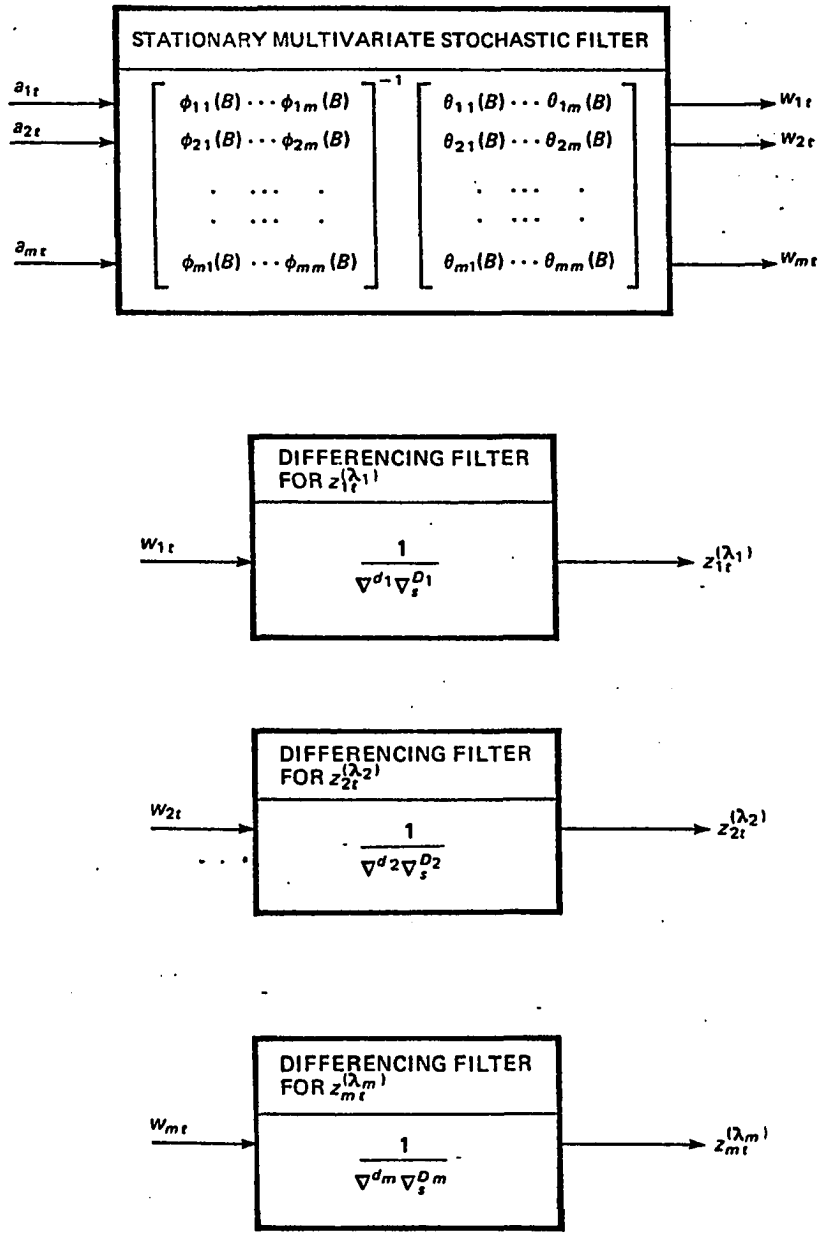


From "The Theory and Practical Application of Multivariate and Multivariate Transfer Function Analysis" by Gwilym Jenkins and Partners Ltd.

Model Building. Figure 1.12 shows a flow diagram for building multivariate stochastic models. The model building is carried out in three steps: identification, estimation and checking. The alignment stage, which is included in the identification step, is simply shifting the time series forward or backwards relative to each other until the cross correlation functions are approximately centered at zero [66, p. 116].

FIGURE 1.11

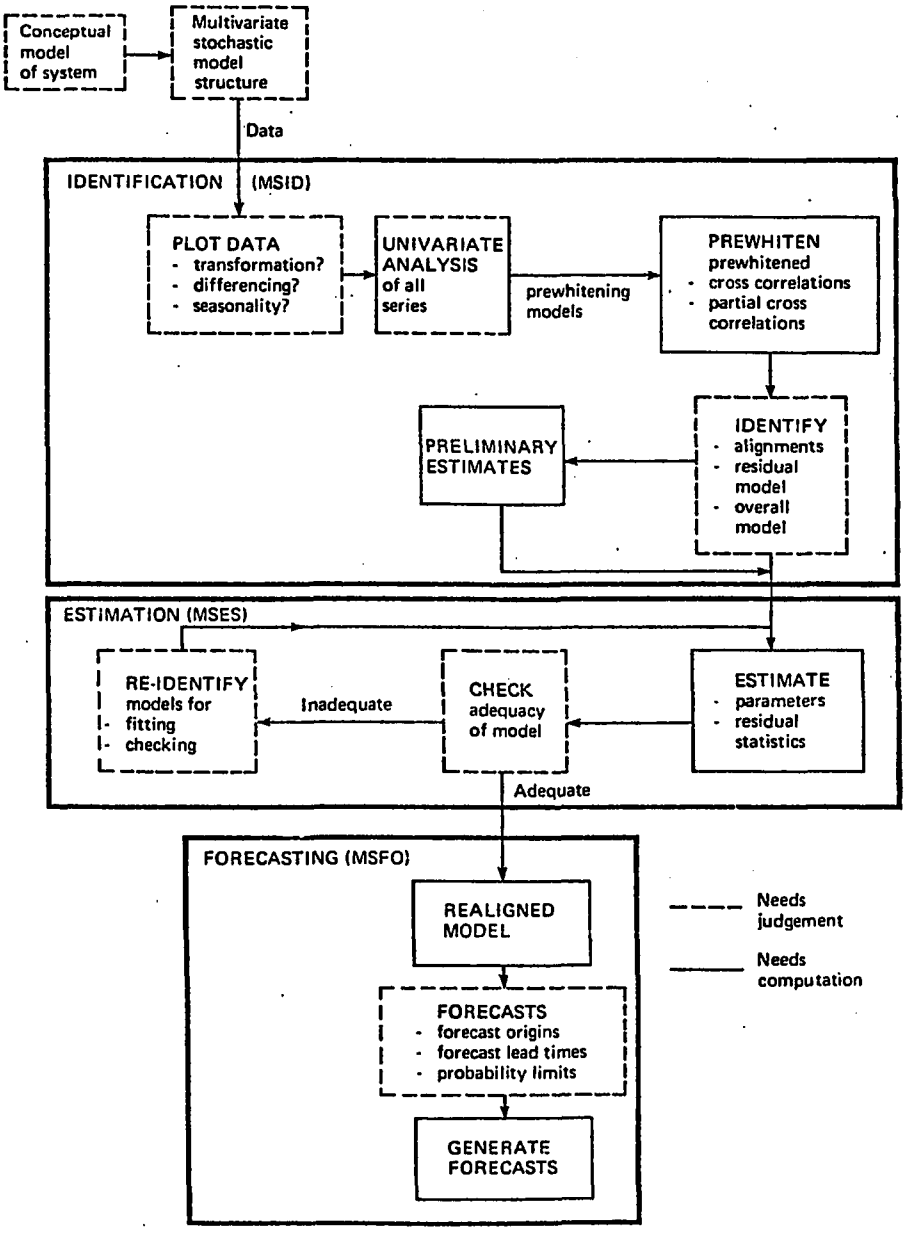
FILTER REPRESENTATION OF SEASONAL MULTIVARIATE STOCHASTIC MODEL



From: Jenkins, Practical Experiences with Modelling and Forecasting Time Series

FIGURE 1.12

FLOW DIAGRAM FOR MULTIVARIATE STOCHASTIC MODEL BUILDING BASED ON THREE COMPUTER PROGRAMS: MSID (MULTIVARIATE STOCHASTIC IDENTIFICATION), MSES (MULTIVARIATE STOCHASTIC ESTIMATION), MSFO (MULTIVARIATE STOCHASTIC FORECASTING).



From: Jenkins, Practical Experiences with Modelling and Forecasting Time Series

CHAPTER TWO

EXPONENTIAL SMOOTHING

> A religious seer is predicting that the world will end at 10 P.M. tonight. For more details, watch the news at eleven.
> -TV News Flash

2.1 Introduction

Exponential smoothing is an approach that produces forecasts of sufficient accuracy but which, at the same time, is quick and inexpensive to operate. It is a fully automated procedure thus reducing the need for skilled manpower in its operation. The exponential smoothing method is established as a short term forecast (perhaps up to a year or so ahead) and retains a good deal of popularity in industrial forecasting in spite of its theoretical limitations when compared with the more sophisticated Box-Jenkins procedure. A great danger in employing a fully automatic predictor is that rather poor forecasts might result unless one exercises some control over the quality of forecasts produced. To meet this need, a number of automatic monitoring [135] or tracking systems have been developed and are used in conjunction with exponential smoothing procedures. This serves as a forecast quality check and can be useful in the modification of inadequate forecasts.

In this chapter, a few commonly used exponential smoothing procedures will be briefly examined. A wide range of exponential smoothing models can be derived as special cases of Kalman filtering and Box-Jenkins procedures. And this being the case, more emphasis will be given to Kalman filtering and Box-Jenkins procedures than exponential smoothing. W.J. Granger and Paul Newbold have an excellent presentation on exponential smoothing in their book Forecasting Economic Time Series. I am indebted to them for the major part of the following discussion.

Forecasting Methods Based on Exponential Smoothing

2.2 Exponential Smoothing for a Constant Process

This model was developed by Holt [65] and Brown [19] and is also called single smoothing model or exponential smoothing model. In principle, exponential smoothing "smoothes" historical observations to eliminate randomness. At each time period, the forecasts are recursively updated using the most current observations. This is a weighing scheme that would apply the most weight to the most recent observed values and decreasing weights to the older values.

Basically, in the case of forecasting with moving averages, we have

$$S_{t+1} = \frac{Z_t + Z_{t-1} + Z_{t-2} + \cdots Z_{t-N+1}}{N} \qquad (2.2.1)$$

where S_t = forecast at time t

Z_t = observation at time t

N = number of values included in the forecast.

Equation (2.21) can be written as:

$$S_{t+1} = \frac{Z_t}{N} + \frac{Z_{t-1} + Z_{t-2} + \cdots Z_{t-N+1}}{N} + \frac{Z_{t-N}}{N} - \frac{Z_{t-N}}{N}$$

$$= \frac{Z_t}{N} + S_t - \frac{Z_{t-N}}{N} \qquad (2.2.2)$$

Equation (2.2.2) states that each new forecast S_{t+1} is based on the preceding forecast S_t adjusted by $\frac{Z_t}{N} - \frac{Z_{t-N}}{N}$. Assume now that only the most recent observation Z_t and the forecast made for the same period S_t were available. A good approximation for Z_{t-N} would be S_t, the seasonal forecast value of the previous period. Then (2.2.2) would become $S_{t+1} = \frac{Z_t}{N} + S_t - \frac{S_t}{N}$, i.e., $S_{t+1} = (\frac{1}{N})\ Z_t + (1- \frac{1}{N})\ S_t$ (2.2.3)

Form (2.2.3) indicates that the forecast for period t+1 consists of weighing the most recent observation Z_t with the weight $\frac{1}{N}$ and the most recent forecast with the value of $1- \frac{1}{N}$. Using the Greek lowercase letter alpha as a substitution for $\frac{1}{N}$ (i.e., smoothing constant),

$$S_{t+1} = \alpha Z_t + (1-\alpha)S_t \qquad (2.2.4)$$

Equation (2.2.4) is the general form used in computing a forecast by the method of single exponential smoothing. Compared with the method of moving averages, it solves the problem of storing the last N observed values, each assigned equal weight, $\frac{1}{N}$, to each of the last N observations and 0 weight to all observation before period t-N. Expanding (2.2.4) by substituting in

the value for S_t, we get

$$S_{t+1} = \alpha Z_t + \alpha(1-\alpha)Z_{t-1} + \alpha(1-\alpha)^2 Z_{t-2} + \alpha(1-\alpha)Z_{t-3} + \ldots \quad (2.2.5)$$

From (2.2.5), we can see that decreasing weights are being given to the older observations since $0<\alpha<1$. The older the observation, the smaller weight it gets assigned.

An equivalent simple exponential model [53, 106] which replaces an original series $X_1, X_2, \ldots X_4$ by a smoothed series $\overline{X}_t$, is given by

$$\overline{X}_t = \alpha X_t + (1-\alpha)\overline{X}_{t-1} \qquad 0<\alpha<1 \qquad (2.2.6)$$

Here the forecast of X_{n+h}, denoted by $F_{n,h}$, for some positive integer h is given by

$$F_{n,h} = \overline{X}_n \qquad (2.2.7)$$

According to (2.2.6) the forecasts of all future values of the series are given by the latest available smooth value. As soon as new observations become available, i.e. actual values for X_t are available, the updating mechanism of (2.2.6) updates the previous estimate of $\overline{X}_{t-1}$ at time t and produces the new estimate of level $\overline{X}_t$ which is a weighted estimate average of X_t and $\overline{X}_{t-1}$.

A starting value for equation (2.2.6) is suggested by setting $\overline{X}_1 = X_1$. Equation (2.2.6) can then be used recursively for t = 2, 3, . . ., n.

Holt-Winters Approach

2.3.1 Non-Seasonal

Simple exponential smoothing is not applicable when there is a seasonal pattern in the data. Holt [65] and Winters [146] extend the simple exponential smoothing algorithm so that it accounts for times series consisting of level, trend and, possibly, a seasonal factor in addition to the unpredictable residual element.

Treating the nonseasonal time series which is made up locally of the sum of level, linear trend and residual, we denote the estimate of level at time t by $\overline{X}_t$ and of trend by T_t where

$$\overline{X}_t = AX_t + (1-A)\ (\overline{X}_{t-1} + T_{t-1}) \qquad 0<A<1 \qquad (2.3.1.1)$$

$$T_t = C\ (\overline{X}_t - \overline{X}_{t-1}) + (1-C)T_{t-1} \qquad 0<C<1 \qquad (2.3.1.2)$$

Formulas (2.3.1.1) and (2.3.1.2) modify previous estimates when new observations are available. The simplest approach to the "starting up" value is to set $T_2 = X_2 - X_1$, $\overline{X}_2 = X_2$ and solve the above formulas recursively for $t = 3, 4, \ldots, n$. Forecasts of future values of the series are given by

$$f_{n,h} = \overline{X}_n + hT_n \qquad (2.3.1.3)$$

The choice of assigning values for A, C will be brought up at the end of Section 2.3.2.

2.3.2 Seasonal Holt-Winters Approach

The most commonly employed variant of the Holt-Winters method assumes that the seasonal factor F_t is multiplicative while the trend remains

additive. In this case, for a seasonal series with $X_t>0$, for every $t\in N$ and period s, the seasonal factor F_t is given by

$$F_t = D(X_t/\bar{X}_t) + (1-D)F_{t-s} \qquad 0<D<1 \qquad (2.3.2.1)$$

The level $\bar{X}_t$ which can be thought as level with the seasonality out is estimated now by

$$\bar{X}_t = A(X_t/F_{t-s}) + (1-A)(\bar{X}_{t-1} + T_{t-1}), \quad 0<A<1 \qquad (2.3.2.2)$$

The trend component is given by equation (2.3.1.2) of the previous section.

$$T_t = C\ (\bar{X}_t - \bar{X}_{t-1}) + (1-C)T_{t-1} \qquad 0<C<1 \qquad (2.3.1.2)$$

"Starting up" values are given by

$$F_j = Xj/\bar{X}_s \text{ where } \bar{X}_s = \frac{1}{s}\sum_{k=1}^{s} X_k \quad j = 1, \ldots, s,\ T_s = 0$$

Equations (2.3.2.1), (2.3.2.2) and (2.3.1.2) can be used recursively for t = s+1, s+2, . . ., n. The forecasts for future values, for additive and seasonally multiplicative trend, are given by

$$f_{n,h} = (\bar{X}_n + hT_n)F_{n+h-s}, \qquad h = 1, 2, 3, \ldots, s$$

$$= (\bar{X}_n + hT_n)\ F_{n+h-2s}, \qquad h = s+1, s+2, \ldots 2s \qquad (2.3.2.3)$$

When we deal with situations where the seasonal factor is additive rather than multiplicative, equations (2.3.2.1) and (2.3.2.2) can be replaced by

$$F_t = D(X_t-\bar{X}_t) + (1-D)F_{t-s} \qquad 0<D<1 \qquad (2.3.2.4)$$

and $$\bar{X}_t = A(X_t - F_{t-s}) + (1-A)(\bar{X}_{t-1} + T_{t-1}), \qquad 0<A<1 \qquad (2.3.2.5))$$

The forecast equation (2.3.2.3) is now replaced by

$$f_{n,h} = \bar{X}_n + hT_n + F_{n+h-s} \quad , \quad h = 1, 2, \ldots, 2$$

$$= \bar{X}_n + hT_n + F_{n+h-2s} \quad , \quad h = s+1, s+2, \ldots, 2s \qquad (2.3.2.6)$$

A major drawback to exponential smoothing is that there is no easy way to determine an appropriate value for α in (1.2.6). Determining appropriate values for A, C and D employed in the Holt-Winters algorithms is also problematic. In general, the lower the values of these constants the more steady the final forecasts will be since the use of low values gives considerably more weight to past observations. Consequently, any random fluctuation in the future will not have any major effect in the determination of the forecast. At the same time, however, the model will be quite insensitive to any sudden changes of behavior of the series.

Holt and Winters propose to select those values for A, C, and D that would have best "forecast" the given situation. The element of arbitrariness is not absent here, however, since a decision is made on the criterion of accuracy (cost or error function) and on the magnitude of projection of the forecast ahead. The most common procedure is to posit a quadratic cost function and to seek the smoothing constants that provide the best one-step ahead forecasts. The procedure is to choose a grid of possible values of A, C, and D and to calculate the one-step-ahead forecasts, $f_{t,1}, t = m, m+1, \ldots, n-1$, for each set of the smoothing constants.

The set for which the sum of squared errors

$$S = \sum_{t=m+1}^{n} (X_t - f_{t-1,1})^2$$

is smallest is then used to calculate actual forecasts of all future values of the series. The starting point m is an integer large enough to allow the effects of the choice of initial "starting up" values to have died down.

One advantage of the above procedure is that it is easily implementated as a computer program that will automatically produce the best choices for A, C and D. One disadvantage is that this approach demands storage of all past observations and, hence, greatly increases computer storage requirements.

A major weakness of Holt-Winters method, besides the requiring of three smoothing parameters and the considerable work which is done to generate the optimal set of smoothing constants A, C and D, is that once the optimal values are found, there is no easy way to modify them when a basic change in the data takes place. An alternative way [83, p. 79] to worrying about optimal values is to find good initial estimates for equations (2.3.2.1), (2.3.2.2) and (2.3.1.2), then, specify small values for A, C, and D (around .1 to .2). The forecasting system then reacts slowly but steadily to changes in the data. The disadvantage of this strategy is that it gives a low response system. This is a general, low cost method for forecasting all types of data which, despite its low response, achieves long-term stability.

2.4 General Exponential Smoothing - Brown's Approach

Brown is credited with creating an alternative procedure, gener-

al exponential smoothing [20]. Define the time series X_i to be

$$X_i = \sum_{j=1}^{k} a_j f_j (i-t) + e_i, \quad i = 1, 2, \ldots t \qquad (2.4.1)$$

where:

i. We assume that the time series X_i is the linear combination of k known deterministic functions of time plus a residual

ii. The functions f_j, j=1, . . . k, are generally taken to be polynomials, exponentials and mixtures of sine and cosine terms.

iii. The model (2.4.1) is assumed to hold only locally; so, we can always minimize the discounted sum of squared errors:

$$S = \sum_{i=1}^{t} \beta^{t-i} e_i^2 \qquad 0 < \beta < 1 \qquad (2.4.2)$$

In matrix form, let $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_t \end{bmatrix}$, $a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$

$$H = \begin{bmatrix} f_1(1-t) & f_2(1-t) & \ldots & f_k(1-t) \\ f_1(2-t) & f_2(2-t) & \ldots & f_k(2-t) \\ \vdots & \vdots & & \vdots \\ f_1(0) & f_2(0) & \ldots & f_k(0) \end{bmatrix} = \begin{bmatrix} f'(1-t) \\ f'(2-t) \\ \vdots \\ f'(0) \end{bmatrix}$$

$$W = \begin{bmatrix} \beta^{(t-1)/2} & & & \\ & \beta^{(t-2)/2} & & 0 \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix}$$

Equation (2.4.2) then becomes

$$S = (WX-WHa)'(WX-WHa) \tag{2.4.3}$$

Differentiating (2.4.3) with respect to a, we obtain a minimum for

$$\hat{a}(t) = (H'W'WH)^{-1}H'W'WX \tag{2.4.4}$$

where $\hat{a}(t)$ is now a function of t.

If we define $F(t) = H'W'WH = \sum_{i=1}^{t} \beta^{t-i}\underline{f}(i-t)\underline{f}'(i-t)$ (2.4.5)

and $$g(t) = H'W'WX = \sum_{i=1}^{t} \beta^{t-i}X_i\underline{f}(i-t), \tag{2.4.6}$$

then we can write (2.4.4) as

$$\hat{a}(t) = F^{-1}(t)g(t) \tag{2.4.7}$$

Equation (2.4.5) implies that

$$F(t) = F(t-1) + \beta^{t-1}\underline{f}(1-t)\underline{f}'(1-t) \tag{2.4.8}$$

Suppose now that there exists a non-singular matrix L such that

$$\underline{f}(t) = L\underline{f}(t-1), \ \forall T\varepsilon N, \tag{2.4.9}$$

then from (2.4.6) we get

$$g(t) = \beta L^{-1}g(t-1) + \underline{f}\,(0)X_t \tag{2.4.10}$$

Further, for functions that do not die out too quickly, it follows from

(2.4.9) that for moderately large t, F(t) will converge to some steady state matrix F, so that a(t) can be written as:

$$a(t) = F^{-1}g(t) \qquad (2.4.11)$$

Thus, the forecast of X_{n+h} is given by

$$f_{n,h} = \hat{a}'(n)\underline{f}(h) \quad \text{or} \quad f_{n,h} = g'(n)F^{-1}\underline{f}(h) \qquad (2.4.12)$$

where $g(n)$ is calculated recursively from (2.4.10)

F is obtained from (2.4.8)

$\underline{f}(h)$ is a vector of known constants.

With a procedure similar to the one used to find the Holt-Winters predictor (i.e. by optimizing over a grid of possible values) or by visual inspection of the characteristics of the series under consideration, one can assign a value to the discount parameter β. Brown suggests choosing β so that $.75 \leq \beta \leq .95$, the actual value used depending on the stability of the series, while Harrison [61] proposes that a value in the neighborhood of $\beta^k = .8$ would frequently be appropriate.

The general exponential smoothing approach poses a few problems that curtail its applicability and its usefulness. First, D.J. Reid [117] notes that the errors from the fitted model are very often serially correlated suggesting that suboptimal forecasts will be produced. He notes that substantial improvement in forecast performance may be achieved by fitting a first-order autoregressive model to these residuals. The second difficulty concerns seasonal time series and deals in particular with the question of how many harmonics to fit in the model. Reid suggests starting with a few

harmonic terms and, every time we add additional terms, testing for an improvement in fit. However, when a large number of coefficients are to be fitted, estimation becomes less efficient, and Reid concludes that "it is normally desirable to keep the number of fitting functions as low as possible provided they still adequately describe the times series" [117].

Brown's method uses only one smoothing parameter, β. Precision estimate of a model such as:

$$X_t = a_1 + a_2 t + a_3 \cos(\frac{2\pi t}{12}) + a_4 \sin(\frac{2\pi t}{12}) + a_5 \cos(\frac{2\pi t}{6}) + a_6 \sin(\frac{2\pi t}{6}) + a_7 \cos(\frac{2\pi t}{4}) + a_8 \sin(\frac{2\pi t}{4}) \qquad (2.4.13)$$

requires β to be fairly large so as to allow a fair number of data points to have appreciable weight; otherwise, one is estimating with very few degrees of freedom. On the other hand, one often requires an exponential smoothing procedure to adapt very quickly so as to put the most weight on the few most recent observations. These two requirements are mutually incompatible, and one wonders of the performance of Brown's method for seasonal time series. This questionable performance prompted C.W.J. Granger and Paul Newbold to rather bluntly ask, "Is it reasonable to expect the Brown predictor to do with a single parameter what requires three parameters for Holt-Winters" [54, p. 169]?

2.5 Relationships Between Box-Jenkins Models and Exponential Smoothing

In this section, the relationships between exponential smoothing and the Box-Jenkins procedures will be apparent as a result of an attempt to find the stochastic processes for which exponential smoothing predictors are

optimal.

Considering the Holt-Winters seasonal predictor in its additive form, we denote with $\overline{X}_t$ the level of series at time t. We have

$$\overline{X}_t = A(X_t - F_{t-s}) + (1-A)(\overline{X}_{t-1} + T_{t-1}) \tag{2.5.1}$$

$$T_t = C(\overline{X}_t - \overline{X}_{t-1}) + (1-C)T_{t-1} \tag{2.5.2}$$

$$F_t = D(X_t - \overline{X}_t) + (1-D)F_{t-s} \tag{2.5.3}$$

If we denote the one-step ahead forecast of X_t by $f_t = f_{t-1,1}$, then

$$f_t = \overline{X}_{t-1} + T_{t-1} + F_{t-s} \tag{2.5.4}$$

and the forecast error is

$$e_t = e_{t-1,1} = X_t - f_t = X_t - (\overline{X}_{t-1} + T_{t-1} + F_{t-s}) \tag{2.5.5}$$

Solving (2.5.1) for $\overline{X}_t - \overline{X}_{t-1}$, we have

$$\overline{X}_t - \overline{X}_{t-1} = T_{t-1} + A(\overline{X}_t - \overline{X}_{t-1} - T_{t-1} - F_{t-s}) \tag{2.5.6}$$

Substituting (2.5.5) in (2.5.6), we get

$$\overline{X}_t - \overline{X}_{t-1} = T_{t-1} + Ae_t \tag{2.5.7}$$

Solving (2.5.2) for $T_t - T_{t-1}$, we get

$$T_t - T_{t-1} = C(\overline{X}_t - \overline{X}_{t-1}) - CT_{t-1} \tag{2.5.8}$$

Substituting (2.5.7) into (2.5.8), we get

$$T_t - T_{t-1} = C(T_{t-1} + Ae_t) - CT_{t-1} =$$

$$= CT_{t-1} + CAe_t - CT_{t-1} = CAe_t \qquad (2.5.9)$$

Similarly solving (2.5.3) for $F_t - F_{t-s}$, we get

$$F_t - F_{t-s} = D(X_t - \overline{X}_t - F_{t-s}) \qquad (2.5.10)$$

or by (2.5.7)

$$F_t - F_{t-s} = D(X_t - \overline{X}_{t-1} - T_{t-1} - F_{t-s} - Ae_t) \qquad (2.5.11)$$

The quantity $X_t - \overline{X}_{t-1} - T_{t-1} - F_{t-s} = e_t$, according to (2.5.5), so

$$F_t - F_{t-s} = D(e_t - Ae_t) = (1-A)De_t \qquad (2.5.12)$$

Now we introduce the back-shift operator as:

$$B^j e_t = e_{t-j}$$

Using the back-shift operator notation, we have:

$$\overline{X}_t - \overline{X}_{t-1} = T_{t-1} + Ae_t \qquad \text{or}$$

$$T_{t-1} = \overline{X}_t - \overline{X}_{t-1} - Ae_t \qquad (2.5.13)$$

We write (2.5.7) as:

$$\overline{X}_{t-1} - \overline{X}_{t-2} = T_{t-2} + Ae_{t-1} \qquad (2.5.14)$$

and (2.5.9) as:

$$T_{t-1} - T_{t-2} = ACe_{t-1} \qquad (2.5.15)$$

We substitute in (2.5.15) the value of T_{t-1} from (2.5.13)

$$\bar{X}_t - \bar{X}_{t-1} - Ae_t - T_{t-2} = ACe_{t-1} \qquad (2.5.16$$

We subtract (2.5.14) from (2.5.16), and we have

$$\bar{X}_t - \bar{X}_{t-1} - \bar{X}_{t-1} + \bar{X}_{t-2} = ACe_{t-1} - T_{t-2} - Ae_{t-1} + Ae_t + T_{t-2}$$

or $\bar{X}_t - 2\bar{X}_{t-1} + \bar{X}_{t-2} = Ae_t - (1-C)Ae_{t-1}$ or, using the back-shift operator B, we have

$$(1-B)^2\bar{X}_t = A[1-(1-C)B]e_t \qquad (2.5.17)$$

Similarly, from (2.5.9) and (2.5.12) we have

$$(1-B)T_t = ACe_t \qquad (2.5.18)$$

and $$(1-B^s)F_t = (1-A)De_t \qquad (2.5.19)$$

Combining (2.5.17), (2.5.18) and (2.5.19), we get

$$(1-B)^2(1-B^s)(\bar{X}_{t-1} + T_{t-1} + F_{t-s}) =$$

$$[AB(1-B^s)[1-(1-C)B] + ACB[(1-B)(1-B^s) +$$

$$(1-A)DB^s(1-B)^2]e_t \qquad (2.5.20)$$

From (2.5.5), we get $\bar{X}_{t-1} + T_{t-1} + F_{t-s} = X_t - e_t$, and, thus, (1.5.20) becomes

$$(1-B)^2(1-B^s)X_t = [(1-B)^2(1-B^s) + AB(1-B^s)[1-(1-C)B]$$

$$+ ACB(1-B)(1-B^s) + (1-A)DB^s(1-B)^2]e_t$$

Now for an optimal forecast, the errors e_t will constitute a white noise process ε_t. Thus, if the Holt-Winters additive seasonal predictor is to produce optimal forecasts, then the series X_t must be generated by a process of the form

$$(1-B)^2(1-B^s)X_t = (1+b_1B + b_2B^2 + b_sB^s + b_{s+1}B^{s+1} +$$

$$b_{s+2}B^{s+2})\varepsilon_t$$

where the five coefficients b_1, b_2, b_s, b_{s+1} and b_{s+2} are functions of the three smoothing constants A, C and D. Hence, this exponential smoothing predictor is optimal for a process generated by a particular member of the class of seasonal models of the Box-Jenkins family.

In a similar way it can be shown that the simple exponential smoothing predictor derived from

$$\bar{X}_t = \alpha X_t + (1-\alpha)\bar{X}_{t-1}, \qquad 0<\alpha<1 \qquad (2.5.21)$$

is optimal iff X_t is generated by the ARIMA (0, 1, 1) process $(1-B)X_t = [1-(1-\alpha)B]\varepsilon_t$ as originally shown by Muth [99].

Harrison [60] shows that the Holt-Winters nonseasonal predictor gen-

erated by

$$\bar{X}_t = AX_t + (1-A)(\bar{X}_{t-1} + T_{t-1}),\ T_t = C(\bar{X}_t - \bar{X}_{t-1}) + (1-C)T_{t-1} \tag{2.5.22}$$

gives optimal forecasts if X_t is generated by an ARIMA (0, 2, 2) process.

More recently Goodman [52] and Cogger [39] show that forecasts produced by multiple exponential smoothing of order k are optimal in a minimum MSE for a restricted class of the ARIMA (0, k, k) processes. McKenzie [88] extends these results to direct smoothing models with transcendental terms.

Cogger [39] considers the nonseasonal Brown predictor, obtained by estimating a polynomial in time of degree m, showing that optimality of this predictor implies that the underlying process is generated by an ARIMA (0, m+1, m+1) model.

D.J. Reid [117] notes that the k-type exponential smoothing forecasts at time t for lead time t are of the form

$$\hat{X}_t(\tau) = \alpha_o(t) + \alpha_1(t)\tau + \ldots\ldots\ldots + \alpha_k(t)\tau^k \tag{2.5.23}$$

where $\alpha_i(t)$ is a linear function of the first k smoothed values of the series; a proof of this statement is given by Brown and Meyer [41]. The result (2.5.23) shows that the forecasts produced by any degree of multiple exponential smoothing are equivalent to those produced by a Box-Jenkins type of predictor of degree (k-1, 1, k) for k-type smoothing. To quote Reid:

> smoothing imposes constraints in the form of relationships between the coefficients of the auto-regressive and moving average operators, so that in general it will be sub-optimal. Only in the rare case where

> these same constraints are present in the true generating model will we expect multiple smoothing to do as well as, or better than, the Box-Jenkins model of equivalent degree [131].

This last statement will be of extreme importance when we consider the "Controversies among Forecasting Literature" in Chapter VII.

2.6 Monitoring Forecast Performance

Lynwood A. Johnson and Douglas C. Montgomery [69] summarize some of the most important reasons for the popularity of the exponential smoothing methods. They are as follows:

(1) The selection of the form of the time series model can be done in a rational manner based either on objective historical data or subjective consideration of the future. A modest amount of historical data is usually sufficient for determining initial parameter values.

(2) Determination of initial values for model parameters is usually easily done.

(3) Model parameters often have intuitive meaning to the forecaster.

(4) Only limited data storage is required.

(5) The same model form may be used for a large number of time series.

(6) Periodic revision of model parameters is easily accomplished by means of simple algebraic expressions.

(7) Forecast generation based on the model is straightforward extrapolation over the lead time of interest.

(8) Forecasts can often be stated in terms of prediction intervals with little additional effort.

(9) Cumulative forecasts can usually be expressed as closed form expressions involving the model's parameters and the lead time length.

(10) The relation between stability and responsiveness of the forecasting procedure can be adjusted easily by changing the rate of smoothing.

(11) Tracking signal tests for forecast control are easy to apply with corrective action on out-of-control situations being possible either automatically through programmed logic or through external intervention.

(12) The cost of developing the model and operating it is less than that of more sophisticated time series methods and causal models while the accuracy obtained in the forecasting stage, as opposed to the parameter estimation or fitting stage, is often comparable.

However, as has been shown, particular exponential smoothing methods are optimal only for corresponding underlying stochastic processes which in many cases are subsets of the general class of the Box-Jenkins models. If somehow the assumed generating process is different than the true underlying process, then, the forecasts produced could be very far from optimal. It would be extremely valuable to have a fully automatic check on forecast performance so forecasts that violated a built-in safeguard would be signaled out. The forecaster may then devote his attention to those few time series that present a problem while forecasts of the remaining series could be produced routinely. In this section, we will examine two proposals in this context. The first one is due to Harrison and Davies [62] and the second one is due to Trigg [135].

2.6.1 Harrison and Davies "Cusum" Technique

P.J. Harrison and O.L. Davies [62] suggest the use of cummulative sum (cusum) techniques for the control of routine forecasts. Briefly, their methodology is as follows:

If we denote f_1 to be a one-step-ahead forecast of X_t with errors e_t and the system begins to produce forecasts for time t=1, then the cummulative sums of the forecasts errors are

$$C_1 = e_1,\ Cj = C_{j-1} + e_j,\ j = 2, 3, \ldots \qquad (2.6.1.1)$$

The cusum chart is a chart where these cummulative sums are plotted. Inspection of the cusum chart may indicate any tendency toward bias in the forecasts. Harrison and Davies suggest a backward sequential test procedure so that we define at time t

$$S_1 = e_t = C - C_{t-1}$$

$$S_2 = e_t + e_{t-1} = C_t - C_{t-2}$$

$$\vdots$$

$$S_k = e_t + e_{t-1} + \ldots + e_{t-k+1} = C_t - C_{t-k}$$

For every new observation that occurs, a new S_i is calculated and tested against corresponding control limits $\pm L_i$. If the calculated value of S_i is not within the (-L, L), lack of control is signaled. The procedure, Harrison and Davies note, only provides a vague indication of the way in which the forecasting scheme needs to be adjusted. Harrison and Davies show that, if the magnitude of the limits L_i is a linear function of the

number of observations comprising the sum to be tested, considerable implications are achieved as far as storage of information is concerned. The authors suggest that, if the forecast errors are independent, appropriate limits can be found through a "monogram" of Ewan and Kemp [44].

2.6.2 Trigg's "Tracking Signal"

Brown [20] presents a tracking signal to be the sum of the forecasting errors divided by the Mean Absolute Deviation (M.A.D.). M.A.D. is a convenient measure of the noise in the system and is obtained by a simple smoothing process upon the absolute forecasting errors. The updating equations for any new, available data are: Sum of errors - previous sum of errors + latest error.

M.A.D. = $(1-\alpha)$ previous M.A.D. + α latest absolute error where α is the smoothing constant.

$$\text{Tracking signal} = \frac{\text{Sum of Errors}}{\text{M.A.D}}$$

Brown computes significance levels for the tracking signal which if exceeded should prompt investigation. Brown's tracking signals have the following two disadvantages. To quote Trigg:

1. Once the tracking signal has gone out of limits, it will not necessarily return within limits even though the forecasting system itself comes back in control. Consequently, intervention is necessary to set the sum of the errors back to zero if future false alarms are to be avoided. Such interventions can be tedious and may tend to be neglected when several hundred items are being forecast.

2. Ironically, if the system starts to give exceptionally accurate forecasts, the tracking signal may go out of limits. For example, if perfect forecasts begin to occur, the M.A.D. will tend to zero whilst the sum of the errors will remain unaltered. The tracking signal, thus, clearly tends to infinity [135].

If instead of the sum of errors we use a smoothed error, we can reunite the

updating equations as:

$$\text{Smoothed error} = (1-\alpha)\text{ previous smoothed error} + \alpha\text{ latest error} \tag{2.6.2.1}$$

$$\text{M.A.D.} = (1-\alpha)\text{ previous M.A.D.} + \alpha\text{ latest absolute error} \tag{2.6.2.2}$$

$$\text{Tracking signal} = \frac{\text{Smoothed Error}}{\text{M.A.D.}} \tag{2.6.2.3}$$

If we denote the smoothed error at time t by E_t and the M.A.D. by D_t, then we have

$$E_t = (1-\alpha)E_{t-1} + \alpha e_t \tag{2.6.2.4}$$

$$D_t = (1-\alpha)D_{t-1} + \alpha|e_t| \tag{2.6.2.5}$$

$$\text{Tracking signal} = \frac{E_t}{D_t} \tag{2.6.2.6}$$

Equation (2.6.2.4) can be expanded into the form

$$E_t = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i e_{t-i}$$

and on the assumption that the errors are uncorrelated

$$\text{Var}(E_t) = \sigma_e^2 \alpha^2 \sum_{i=0}^{\infty} (1-\alpha)^{2i}$$

Since $1-\alpha$ is always less than 1, the series is convergent and sums to

$$\frac{\alpha^2\sigma_e^2}{1-(1-\alpha)^2}$$

Two sigma limits for the smoothed error are

$$\pm \frac{2\alpha\sigma_e}{\sqrt{2_a - a^2}} \quad \text{or} \pm 2\sigma_e \sqrt{\frac{\alpha}{2-\alpha}}$$

There is a linear relation between σ_e and M.A.D. For a wide range of distributions, $\sigma_e \approx 1.2$ M.A.D. or $\sigma_e \approx 1.2D_t$ where σ_e is the standard deviation of the error series.

Two standard error limits are thus given by

$$\pm 2.4\alpha / \sqrt{2\alpha - \alpha^2} \qquad (2.6.2.7)$$

Trigg derives the cummulative distribution of the tracking signal by simulation (assuming in addition a normal distribution for the errors) and concludes that for $\alpha = .1$ equation (2.6.2.7) provides a good approximation for a 5% level test.

Trigg's assumption that the errors are uncorrelated is frequently violated by exponential smoothing predictors. It is apparent that the validity test rests crucially on the above assumption. Batty [139] demonstrates this point in a particular case.

CHAPTER THREE

STATE SPACE
KALMAN FILTERS AND THEIR APPLICATION
TO FORECASTING

> If you can look into the seeds of time, and
> say which grain will grow and which will not,
> speak then unto me
>
> -Shakespeare

3.1 Introduction

A good introductory presentation on Kalman Filters has been given by Spyros Makridakis and Steven C. Wheelwright [83 and 84]. They define Kalman filtering to be a combination of two independent estimates to form a weighed estimate or prediction. One estimate can be based on prior knowledge and the other estimate can be based on new information. Kalman filtering combines the two estimates to obtain an improved estimate.

Harrison and Stevens [63] use the name Bayesian forecasting and Kalman filters interchangeably. An obvious parallelism is found between Bayesian's "prior" and "posterior" and Kalman filtering's "prior" and "new estimate."

In section 3.2, a brief historical overview of the Kalman filtering will be given. Section 3.3 will briefly cover the Bayesian

approach in forecasting and the state space models. Section 3.4 will deal with the properties of Kalman filtering and the applications of state space models in forecasting and model design. Section 3.5 will deal with an example of multidimensional identification and forecasting using state space models and, in particular, using the automated - software package PROJECT. Finally, in section 3.6, some conclusions will be reached on state space forecasting and Kalman filtering. Their properties, advantages and limitations will be briefly examined.

3.2 Historical Overview of the Kalman Filtering Approach

Kolmogoroff's work, written in 1941, is a pioneering effort on discrete time stationary stochastic process [90]. His and the work of Wiener [140] for the continuous case provide the starting point basis of modern filtering theory. Both Wiener and Kolmogoroff center their work on estimating the white noise, e_t in

$$Z_y = Z_t' + e_t \tag{3.2.1}$$

where Z_t is expressed as deviation from the overall mean of the time series and Z_t' is the original message, or pattern, generated by the real process represented through the time series. Kolmogoroff uses a presentation suggested by Wold [148] and Wiener reduces the prediction problem to the solution of a Weiner-Hopf integral equation:

$$\gamma_{xx}'(K) = \int_{-\infty}^{\infty} W(v)\gamma_{xx}(K-v)dv$$

where W(v) is the weighing function or impulse-response function and γ

denotes the out covariance. Recapturing Z_t' can be achieved by solving (3.2.1). However, such a solution was difficult to obtain, thus, limiting the applicability of the Wiener-Kolmogoroff filter. Originally, the solution was obtained by using spectral factorization.

Later, in 1947, Levinson provides some discrete approximations in order to obtain W(v) in equation (3.2.1) [90]. When new data becomes available, he devises recurrence equations to obtain estimates of W(v). At this point, however, without the help of a computer, the matrix inversion is still posing a problem when the number of observations involved is large. Zadeh and Ragazzini in 1950 simplify the problem of solution for the finite memory case by introducing the idea of "shaping filters" which provides a simplified approach for the solution of the Hopt-Wiener equation. Bootom in 1952 deals with the problem on the non-stationary time series approach. Stratonovich, in 1957, works on the scalar continuous case and Swerling [127], in 1959, extends the Kolmogoroff-Wiener results for the no process noise. Overall, except for a few insignificant aspects, his work is identical to the later work of Kalman and Bucy.

In 1960, R.E. Kalman [71] points out that the methods already developed for solving the Wiener problem are subject to the following limitations which curtail their practical usefulness:

i. The optimal filter is specified by its impulse response. It is not an easy task to synthesize the filter from such data.

ii. Numerical determination of the optimal impulse

response is often quite involved and poorly suited to machine computation.

iii. Important generalizations (e.g. growing-memory filters, non-stationary prediction) require new derivations frequently of considerable difficulty to the non-specialist.

iv. The mathematics of the derivations are not transparent. Fundamental assumptions and their consequences tend to be obscured.

Kalman approaches the Wiener problem from the point of conditional distribution and expectations. In this way, basic facts of the Wiener theory are quickly obtained; the scope of the results and the fundamental assumptions appear clearly. All statistical calculations and are based on first and second order averages; thus, the (iv.) limitation is eliminated.

Following particularly Bode and Shannon [10], Kalman represents arbitrary random signals as the output of a linear dynamic system excited by independent or uncorrelated random signals ("white noise"). His approach differs from the conventional one only in the way linear dynamic systems are described; he emphasizes the concept of "state" and "state transition"; in other words, he specifies linear systems in terms of first-order difference equations.

With the state-transition method, a single derivation covers a large variety of problems: growing and infinite memory filters, stationary and nonstationary statistics. In other words, difficulty

(iii) above disappears. Having guessed the "state" of the estimation (i.e., filtering or prediction) problem correctly, one is led to a non-linear difference equation for the covariance matrix of the optimal estimation error. From the solution of this equation the coefficients of the difference (or differential) equation of the optimal linear filter are obtained without further calculations. He also shows that the filtering problem is the dual of the noise-free regulator problem.

In 1961, R.E. Kalman and R.S. Bucy [73] in their article "New Results in Linear Filtering and Prediction Theory" derives a non-linear differential equation of the Riccati type for the covariance matrix of the optimal filtering error. The solution of this "variance equation" completely specifies the optimal filter for either finite or infinite smoothing intervals and stationary or nonstationary statistics. The variance equation is closely related to the Hamiltonian (canonical) differential equations of the calculus of variations. In some cases, the authors, Kalman and Bucy, provide analytical solutions. They also make an extensive use of the Duality Principle in relating stochastic estimation and deterministic control problems and in proving theoretical results.

Kalman and Bucy, unlike Wiener and Kolmogoroff who work in the frequency domain, involve their system description in the time domain and introduce state-space notation which offers considerable conceptual and computational advantages.

R.S. Bucy [107] claims that when the stationary stochastic processes are Markovian with continuous paths, the Wiener-Kolmogoroff

theory is simply a special case of the Kalman-Bucy theory of linear, finite time filtering for nonstationary processes.

In 1969, R.A. Singer and Paul A. Frost [121] derive upper and lower bounds on the error covariance matrices of the Kalman and Wiener filters for linear finite state time invariant system. These bounds yield a measure of the relative estimation accuracy of these filters and provide a practical tool for determining when the implementational complexity of a Kalman filter can be justified. The calculation, though, of these bounds requires little more than the determination of the corresponding Wiener filter.

In 1977, Raman K. Mehra [121] assumes that the required knowledge of all the systems and noise parameters for the Kalman filter is unknown (i.e. all these parameters have an unknown value) and, therefore, must be identified before use in the Kalman filter. He then presents a correlation technique which identifies a system in its canonical form. The estimates are shown to be asympototically normal, unbiased and consistent. The scheme is capable of being implemented on-line and can be used in conjunction with the Kalman filter. A technique for more efficient estimation by using higher order correlations is also given. He suggests a recursive technique to determine the order of the system when the dimension is unknown. The results are first derived for stationary processes and are then extended to nonstationary processes.

Emanuel Parzen, in 1979, in his article "Forecasting and Whitening Filter Estimation," [112] suggests a different approach to time series modeling in which the identification stage is not

accomplished by graphical inspection of the time series and of computed auxiliary sample functions such as autocorrelation function, partial autocorrelation function and spectrum. Rather, the transfer function g_∞ of the whitening filter is estimated directly and parameterized parsimoniously by using a criterion function called (CAT) for determining the order of approximating autoregressive schemes. His reason for diverging from the popular Box-Jenkins approach of time-series model identification is that ARMA schemes are useful and desirable only for modeling nonstationary time-series. According to Parzen, the advantage of his approach is that the identification process currently used in ARMA modelling, which is based on visual inspection of autocorrelation and partial autocorrelation functions, can be replaced by objective criteria based on estimating the transfer function in the frequency domain of the infinite AR whitening filter.

Raman K. Mehra [120] and Jazwinski [65A] give overviews of Kalman filters and their application to forecasting. The discussion presented here of the Bayesian approach and that of state space and of Kalman filters follows Mehra.

3.3 Bayesian Approach and State Space Models

3.3.1 There are four basic elements in a decision theory formulation for forecasting future values of a given random process:

i. unknown state of the world denoted by $\{X(t)\}$ where $x(t)$ is a state vector, for $t = o, 1, 2, 3 \ldots$

ii. Prior knowledge of the process $\{X(t)\}$ in the form of prior probabilities and evolution equations that

result in a prior probability specification $p(\{X(t)\})$

iii. A vector of observations $\{Y(t)\}$ associated with the true state $\{X(t)\}$ according to a known probability law $P(\{Y(t)\}/\{X(t)\},\theta)$ where θ is set of parameters.

iv. A loss function $\ell[\{X(t)\}, \{\hat{X}(t)\}]$ that expresses the loss to the decision maker and whose expected value should be minimized, for an optimal decision strategy, by making an appropriate choice of $\{\hat{X}(t)\}$. The forecast $\hat{X}(t)$ is a function of all the observations at time t.

Raiffa [114] shows that an efficient way to obtain an optimal decision is to compute recursively in time the posterior distributions of the state $\{X(t)\}$ and perform expectations of the loss function with respect to these distributions. Calculations of the posterior probabilities, nevertheless is not an easy task, and that is exactly where a Kalman filter becomes useful.

3.3.2 State Space Models. State space models are based on the Markov property that the present state of a random process depends probabilistically on past states of the process only through the state observed in the most immediate past. A general state space vector model of finite dimension is typically specified in terms of the following five quantities:

i. Input vector $u(t)$, output $y(t)$ and internal state variable $x(t)$.

ii. A transformation rule of the state vector from one time to the next.

iii. A relationship between u(t), y(t) and x(t).

iv. Initial state x(0).

v. Joint statistics of all random variables.

Mathematically $x(t+1) = f[x(t), u(t), \theta, t] + w(t)$ (3.3.2.1)

$y(t) = h[x(t), u(t), \theta, t] + v(t)$ (3.3.2.2)

$t = 0, 1, 2, \ldots$

where x(t) is an nx1 state vector

u(t) u(t) is an rx1 input vector

w(t) is qx1 process noise vector

θ is mx1 parameter vector

y(t) is px1 output vector

w(t) and v(t) are uncorrelated white noise sequences with known distributions. Also the distribution of x(0) is assumed known.

As a mathematical model, the state space model is only an approximation to reality. There are a number of popular forecasting models which can be reduced in the form of equations (3.3.2.1) and (3.3.2.2). When (3.3.2.1) and (3.3.2.2) are linear the state space model is known as the Gauss-Markov model. The linearity reduces equations (3.3.2.1) and (3.3.2.2) to

$$x(t+1) = \phi x(t) + Gu(t) + \Gamma w(t) \quad (3.3.2.3)$$

$$y(t) = Hx(t) + v(t), \; t = 0, 1, 2, \ldots \quad (3.3.2.4)$$

where w(t) and v(t) are assumed to be Gausian white noise (GWN) sequences with zero mean and covariances Q and R respectively. The matrices ϕ,

G, H, Γ, Q, R and P_o are deterministic where P_o is defined as the covariance of the initial state x(o) which is normally distributed with mean $\hat{x}_o$ and covariance P_o.

What we gain with representation (3.3.2.3) and (3.3.2.4) over (3.3.2.1) and (3.3.2.2) is that, by solving the set of first-order vector difference equations, we can compute the mean, covariance and correlation functions for x(t) and y(t) [117]. Furthermore, the posterior distribution $p[x(t)|y(t), y(t-1), \ldots y(1)]$ turns out to be Gaussian, and its first two moments are computed recursively by the Kalman filter.

One example that shows the modeling flexibility of the state-space vector is the derivation of the equation of a first order autoregressive moving average model. In equations (3.3.2.3) and (3.3.2.4) process noise w(t) and measurement noise v(t) have quite different interpretations and effects; v(t) represents the error inherent in observing the true state of the system x(t) and w(t) represents the random shocks during the evolution of x(t). is we assume

i. All matrices are time invariant.

ii. Neglect v(t) i.e. v(t) = 0

iii. All state variables can be observed, i.e. y(t) = x(t),

then we can write equation (3.3.2.3) as:

$$y(t+1) = \phi y(t) + Gu(t) + \Gamma w(t) \qquad (3.3.2.5)$$

R.S. Bucy and P.D. Joseph [25] show that equation (3.3.2.5) resents an AR(1) process with observed input u(t) and random errors w(t). Chow [34] and Mehra [90] show how econometric simultaneous equation

models may be written in the form of equation (3.3.2.5).

Now, we assume that $w(t) = 0$, i.e., there is no process noise and that also the initial state $x(0)$ is known. Then, given $\{u(t)\}$, the $\{x(t)\}$ process is deterministic, and by substituting $x(t) = y(t) - v(t)$ into equation (3.3.2.3), we have $y(t+1) - v(t+1) = \phi[y(t) - v(t)] + Gu(t)$ or

$$y(t+1) = \phi y(t) + Gu(t) + [v(t+1) - v(t)] \qquad (3.3.2.6)$$

P. Whittle [72] shows that equation (3.3.2.6) corresponds to a vector first-order autoregressive moving average (ARMA) model. The same type of result is obtained even if the $w(t)$ term is kept in the model.

3.4 Kalman Filter Properties and Model Design

R.K. Mehra in "Kalman Filters and their Applications to Forecasting" TIMS Studies in the Mgt Sciences 12 (1979), p. 75-94 derives the one-step-ahead prediction equation for the state space model using the state space equations, and by using Bayes's rule, he shows how to do measurement updates. The prediction and update Kalman equations are then derived and the innovation and stability properties of the Kalman filters are discussed in detail. These properties are mentioned here briefly.

Innovation Property. The one-step-ahead prediction error sequence $v(t) = y(t) - H\hat{x}(t|t-1)$ is known as the innovation sequence since it represents new information brought by observations $y(t)$ in addition to the information contained in the past observation history y_{t-1}. The innovation property is shown to be used to test the

optimality of Kalman filters to detect changes in process model and to build adaptive, robust Kalman filters.

Stability. Kalman [72] and others show that the Kalman filter possesses the property of global asymptotic stability for a completely controllable and observable system.

Raman K. Mehra also gives some guidelines when it comes to Kalman filter design and testing. These are characterized as model selection, parameter specification, algorithm selection, sensitivity analysis, validation and testing:

Model Selection is perhaps the most important step in Kalman filter design. When there is little a priori modeling information available, as is the case in several socioeconomic and business applications, it would be better to identify the state vector directly from the historical data. Mehra and Cameron [93] discuss such a technique. The use of multiple models is also a good strategy when models can be developed on theoretical as well as on empirical bases. Multiple models are extremely useful when the model of the system may be expected to change suddenly in time without knowing the exact time of the change.

Parameter Specification. Once a model is selected, one has to specify the matrices ϕ, G, Γ, H, Q, R, $\hat{X}_o$ and P_o. Physical understanding of the process can provide adequate information for theoretically based models. For black-box models, the parameters can be identified from the historical input-output data once special canonical forms are assumed. Other techniques such as Maximum Likelihood or Bayesian approach have also been developed for estimating unknown

parameters in theoretical models by using past historical data. In case of little or no historical information, Kalman filtering may be started with very little objective information and adapted as data become available.

Sensitivity Analysis. In general, the effect on modeling and parameter errors on the performance of the Kalman filter can be determined through sensitivity analysis. It is shown [91] that the Kalman filter approach has extreme sensitivity to the underestimation of the measurement noise variance.

Validation and Testing. Statistical tests for checking the whiteness property are:

i. Correlation tests for testing local dependence [92]

ii. Integrated spectrum test for periodic linear dependence [68]

iii. Run tests for linear and nonlinear dependence [45]

3.5 Multidimensional Identification and Forecasting Using State Space Models

The following descriptions of the identification and forecasting processes for state space models is summarized from a paper presented at the ORSA/TIMS Conference in Miami, Florida November 3, 1976 by Alan V. Cameron and Raman K. Mehra [28]. According to Cameron and Mehra, the need for multidimensional or multiple time series modeling and forecasting stems from the lack of a technique that adequately describes the behavior of a multiple time series. The Box-Jenkins methodology is inadequate since, according to the authors,

the multiple time series case requires certain new concepts and techniques not present in the Box-Jenkins approach.

Computer runs of various data, using the State Space program PROJECT, were obtained and analyzed in a seminar on SS during the month of November 20-23, 1980 given by Raman Mehra in Boston, Massachusetts. Some of the concepts to be presented in this section were introduced and discussed during the above seminar.

The PROJECT program has been developed based on the theory of Systems Identification and Control Theory. The program can perform differencing and choose the order of the model automatically with minimal user intervention. The PROJECT program is based on State Space Models, Stochastic Realization Theory, Statistical Decision Theory, Canonical Correlation Analysis and Kalman Filtering combined in a unique way for Multiple Time Series Analysis. A brief discussion of the technical aspects is given below:

State Vector. The state of a system is defined as a collection of all information from the present and past history of the process sufficient to predict its future behavior. Sufficient means that the State Vector contains only uncorrelated elements.

A more mathematical definition of the State Vector of a process is: the basis of a linear vector space spanned by the predictors of the present and future observations of the process which has been derived from the present and past observations of the system. For a stationary multidimensional process $y(t)$, the State Space will be spanned by the components of the predictors $y(t+k|t)$ $k= 0, 1, 2, \ldots$

where $y(t+k|t)$ is the best k step ahead linear predictor of $y(t+k)$ given $y(t)$, $y(t-1)$, $y(t-2)$ . . . . The basis for this predictor State Space at time t is called the State Vector $x(t)$.

Since in practice there are many processes and systems for which theoretical models are not available and statistical models have to be developed directly from the data, the PROJECT program develops the statistical model directly from the data identifying the state of the system from a canonical correlation analysis of the observed data and by using State Vector Canonical Models whose parameters are uniquely defined from the input-output properties of the system.

The AR Mathematical Model. Consider the AR model:

$$y_t = A_1 y_{t-1} + y_{t-2} + \ldots + A_p y_{t-p} = u_{t+1}$$

where y_t is the p dimensional vector of the observed series, A1, . . ., A_p are pxp matrices and u_{t+1} is the one step ahead forecast error and p is the order of the AR model.

The Akaike Information Criterion (AIC). The AIC is defined as:

AIC = -2 (maximum log likelihood)

+2 (number of free parameters within the model).

To quote Cameron and Mehra in "User Manual for State Space Forecasting & Modeling Building Program."

> This criterion is used in the identification of both State Vector and Autoregressive Models. When several competing models are being fitted to a single or multiple time series process, the model with the smallest value of the AIC is chosen as the best model. The AIC provides a consistant criterion for comparing increasing model goodness of fit (measured by the likelihood function) versus increasing model complexity (measured by the number of parameters). It quantifies the idea of parsimony in statistical model building.

State Space Models. Let y_t be a pxl vector of an observed time series. Then a state vector model of the process is defined as:

$$x_{t+1} = Fx_t + Gu_{t+1} \tag{3.5.1}$$

$$y_t = Hx_t \tag{3.5.2}$$

where x_t is an nxl vector of state variables

H is a pxn matrix

F is a nxn matrix

G is a nxp matrix

u_{t+1} is a pxl vector of one-step ahead prediction errors or "innovations" which is a zero mean white noise process. u_{t+1} is defined as $u_{t+1} = y_{t+1} - y_{t+1/t}$. The notation $y_{t+1/t}$ is the prediction of y at time t for one step ahead (t+1). Matrices F, G and H depend on the statistical properties of the process. The covariance matrix of u_{t+1} is denoted by Σ. The system of equations (3.5.1) and (3.5.2) are completely specified in terms of the quantities (n, F, G, H, Σ).

For a Canonical State Space Model, a unique determination of the State Vector x(t) and the matrices F, G and H can be obtained by performing a canonical correlation analysis between the two sets of random variables:

i. the present and past values of the process

y(t), y(t-1)y(t-2)...,y(t-m)

and

ii. the present and future predictors for the process

y(t), y(t+1/t), y(t+2/t), . . .

The AIC (Akaike Information Criterion) is used to determine both m

and the components of the State Vector x(t). The parameters of F, G and H are obtained from the Canonical Correlation Analysis.

It was shown in this chapter that the State Vector Model of equation (3.5.1) and (3.5.2) is equivalent to an ARMA model in the sense that both models will output a series y_t with identical statistical properties. Cameron and Mehra, nevertheless, point out that there are several advantages of State Vector Models over ARMA. These advantages are

i. Once the State Vector Model is identified, prediction or forecasting is done trivially by setting $u_t = 0$ for all future values of it. The prediction of some of the series when the future values for the rest are known is also done easily by using a Kalman Filter.

ii. A multidimensional ARMA model of the type $Y_t + B_1 y_t + \ldots + B_m y_{t-m} = u_t + A_1 u_{t-1} + \ldots + A_L u_{t-1}$ of matrices ($B_1 \ldots B_m$, $A_1 \ldots A_L$) may produce a y_t series with identical properties. (This problem does not arise for univariate series.) The same problem exists for the State Vector but it is easily solved by restricting (F, G, H) to the so-called 'Canonical Forms'. The restrictions on ($B_1 \ldots B_m$, $A_1 \ldots A_L$) are more complicated and are difficult to incorporate in an identification program [28].

Stochastic Realization Theory: The Stochastic Realization problem is that of determining the internal structure of a State Vector Model given its external behavior: determine (n, F, G, H, Σ) given the correlation function (C_o, C_1, C_2. . .) of the output y_t. Future conditions such as the minimality of n and the uniqueness of (F, F, H, Σ) are imposed to develop a parsimonious representation whose parameters can be identified uniquely from the data. The correlation function (C_o, C_1,...) is not exactly known and must be estimated from the observed time series (y_1, y_2, . . . y_n). In PROJECT, this problem is solved by using tools

of Statistical Decision Theory and Canonical Correlation Analysis.

Canonical Correlation Analysis. Canonical Correlation Analysis can be considered as a generalization of regression analysis: a regression problem is concerned with finding a linear combination of components of a vector random variable which has maximum correlation with a scalar variable whilst Canonical Correlation Analysis is concerned with finding the maximum correlation between linear combinations of components from two vector random variables. For each of the two vector random variables, the canonical correlation analysis finds linear combinations of the original vectors which are independent of each other and which have zero means and unit variances. In the State Space program, the two vector random variables considered are the predictor and data vector spaces, and the basis of each of these is found through the Canonical Correlation Analysis.

ARMA Models. Autoregressive Moving Average models for a multiple series y(t) have the form:

$$y(t) = A_1 y(t-1) + \ldots + A_p y(t-p) +$$
$$u(t) - B_1 u(t-1) - \ldots - B_q u(t-q) \text{ where}$$

u(t) is a p dimensional vector,

A_i and B_i are pxp dimensional matrices and u(t) are the residual errors.

There is one-to-one transformation between State Space models and ARMA models. Some of the equivalences of AR, MA and ARMA to State Space models are given below. The materials are from the seminar given by Raman Mehra on State Space forecasting on November 20-23, 1980 in Boston, Massachusetts.

EQUIVALENCE OF MA AND STATE VECTOR MODEL

$$y(t) = u(t) - B_1 u(t-1) - \ldots - Bqu(t-q)$$

let

$$x(t) = \begin{bmatrix} u(t) \\ u(t-1) \\ \cdot \\ \cdot \\ \cdot \\ u(t-q) \end{bmatrix}$$

State Vector Model:

$$\begin{bmatrix} u(t) \\ u(t-1) \\ \cdot \\ \cdot \\ \cdot \\ u(t-q) \end{bmatrix} = \begin{bmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ I & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & I & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ & & & \cdot & & & & & \\ & & & \cdot & & & & & \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & I & 0 \end{bmatrix} \begin{bmatrix} u(t-1) \\ u(t-2) \\ \cdot \\ \cdot \\ \cdot \\ u(t-q-1) \end{bmatrix} + \begin{bmatrix} I \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} u(t)$$

$$x(t) = \Phi x(t-1) + \Gamma u(t)$$

$$y(t) = [I - B_1 - B_2 \ldots - B_q]x(t)$$

$$= Hx(t)$$

EQUIVALENCE OF AR AND STATE SPACE MODEL

$$y(t) = A_1 y(t-1) + \ldots + A_p y(t-p) + u(t)$$

let

$$x(t) = \begin{bmatrix} y(t) \\ y(t-1) \\ \vdots \\ y(t-p+1) \end{bmatrix}$$

Then

$$y(t) = [I \;\; 0 \;\; 0 \ldots 0]x(t)$$

$$= Hx(t)$$

$$\begin{bmatrix} y(t) \\ y(t-1) \\ \vdots \\ y(t-p+1) \end{bmatrix} = \begin{bmatrix} A_1 & A_2 & \cdots & A_p \\ I & 0 & \cdots & 0 \\ 0 & I & \ddots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix} \begin{bmatrix} y(t-1) \\ y(t-2) \\ \vdots \\ y(t-p) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} u(t)$$

or

$$x(t) = \Phi X(t-1) + \Gamma u(t)$$

$\uparrow$
white
noise

EQUIVALENCE OF ARMA AND STATE VECTOR MODEL

$$y(t) = A_1 y(t-1) + \ldots + A_p y(t-p) + u(t) - B_1 u(t-1) - \ldots - B_q u(t-q)$$

LET

$$\left[\begin{array}{c} y(t) \\ y(t-1) \\ \vdots \\ y(t-p+1) \\ \hline u(t) \\ u(t-1) \\ \vdots \\ u(t-q+1) \end{array}\right]$$

THEN

$$x(t) = \left[\begin{array}{cccccc|ccccc} A_1 & A_2 & \cdot & \cdot & \cdot & A_p & -B_1 & \cdot & \cdot & \cdot & -B_q \\ I & 0 & \cdot & \cdot & \cdot & \cdot & & & & & 0 \\ 0 & I & 0 & \cdot & \cdot & \cdot & & & & & \cdot \\ \cdot & & & & & & & & & & \cdot \\ \cdot & & & & & & & & & & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & & & & & 0 \\ \hline & & & & & & 0 & \cdot & \cdot & \cdot & 0 \\ & & & & & & I & 0 & 0 & \cdot\cdot & 0 \\ & & & 0 & & & \cdot & & \cdot & \cdot & \cdot \\ & & & & & & \cdot & & & \cdot & \cdot \\ & & & & & & \cdot & & & & \cdot \\ & & & & & & 0 & \cdot & \cdot & \cdot & 0 \end{array}\right] x(t-1) + \left[\begin{array}{c} I \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ \hline I \\ 0 \\ \cdot \\ \cdot \\ 0 \end{array}\right] u(t)$$

$$y(t) = [I\ 0\ \ldots\ 0]x(t)$$

In general, Raman Mehra claims that if n is the dimension of the State Space Vector x(t), then the equivalent ARMA model has n AR and n-1 MA parameters, i.e., the State Space model is equivalent to ARMA (n,n-1).

Forecast Generation. Given the State Space model:

$$x(t+1) = F \cdot x(t) + G \cdot u(t+1)$$

$$y(t) = H \cdot x(t)$$

then k forecasts for future values are recursively generated by assuming $u(t+k) = 0$ for $k>0$,

$$\text{i.e. } x(t+k) = F \cdot x(t+k-1)$$

$$= F \cdot F \cdot x(t+k-2)$$

$$= F^{**}k \cdot x(t)$$

and

$$y(t+k) = H \cdot x(t+k)$$

$$= H \cdot F^{**}k \; x(t)$$

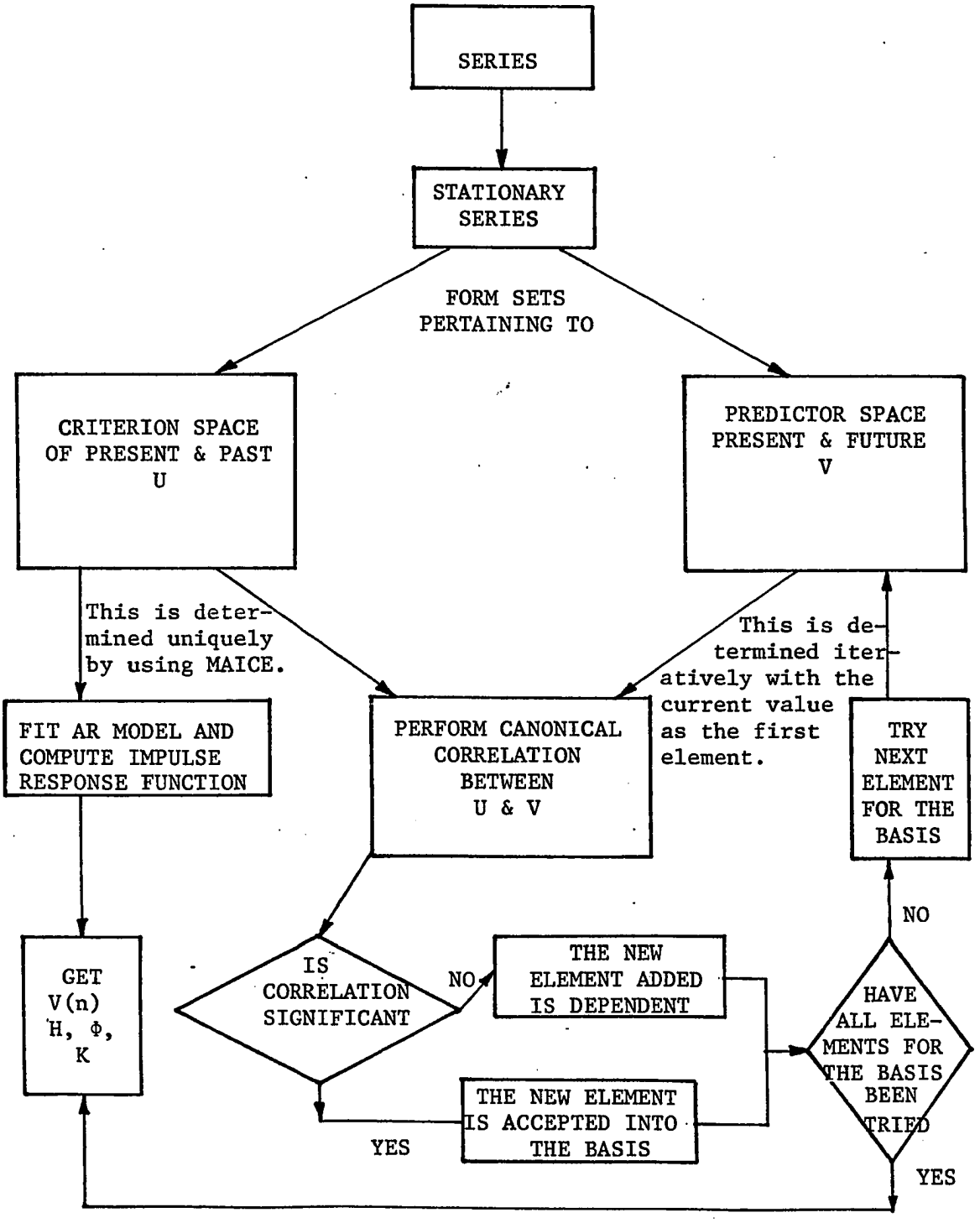
It is in the forecast generations that the Kalman filtering is used. Figure 3.2 shows the model of the process and optimal filter for the discrete case.

State Space Model Development Program Structure. The major functions performed in the State Space model building and forecasting program are:

1. Input of the time series to be forecast and the optional model and forecast control parameters.
2. Creation of a stationary series by regular or seasonal differencing or transformation.
3. Creation of the data space from present and past values of the series.

FIGURE 3.1

FLOW DIAGRAM OF STATE SPACE
FORECASTING PROGRAM "PROJECT"



$$x(t+1) = Fx(t) + \Gamma w(t)$$

$$y(t) = Hx(t) + v(t)$$

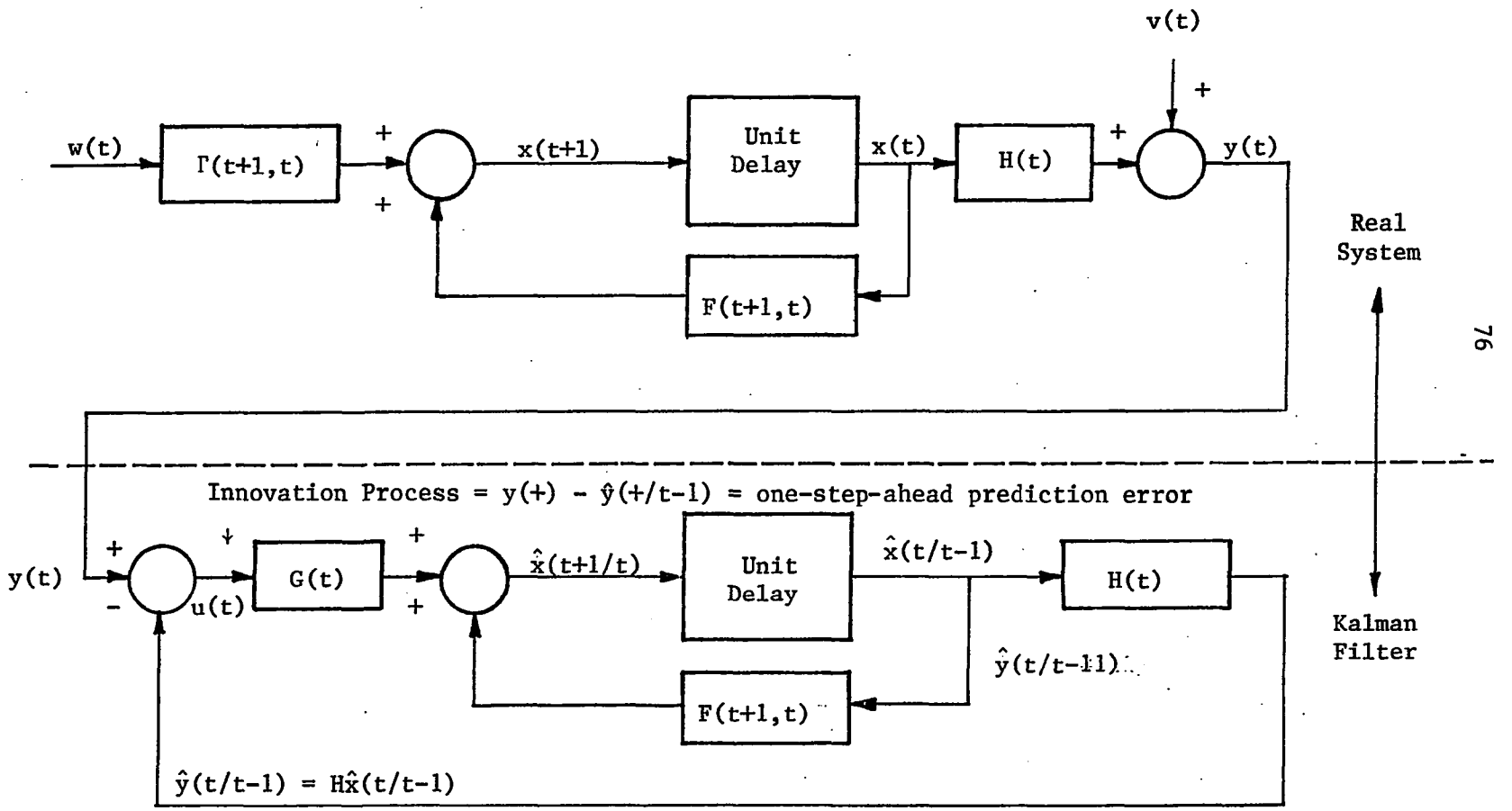


FIGURE 3.2

MODEL OF THE MESSAGE AND OPTIMAL FILTER: DISCRETE CASE

$$\hat{x}(t+1/t) = F\hat{x}(t/t-1) + Gu(t)$$

4. Creation of the predictor space from forecasts of present and future values of the series.
5. Iterative performance of Canonical Correlation Analyses between the data and predictor spaces.
6. If the canonical correlation is significant, addition of a new element to the predictor space basis vector.
7. When all independent elements of the predictor space have been tried, compute the transition matrix from the canonical correlation results and compute the impulse response matrix from the autoregressive model.
8. Generate the forecasts and the residual one-step-ahead forecast errors for the fitted model over the historical data.
9. Perform diagnostic checks on the residuals and calculate goodness of fit statistics.
10. Generate forecasts and confidence limits from the model and un-difference and/or un-transform the forecasts back to the original data levels.

PROJECT does not use nonlinear searches in developing estimates of the parameters. The Canonical Correlation and AR modelling techniques provide asymptotically efficient estimates of all system parameters.

The direct development, testing and use of State Space model for a two dimensional process is now illustrated. Cameron and Mehra make the claim that minimal user interaction is required in using the PROJECT program. The numbered paragraphs refer to the corresponding

numbers on the sample computer printout from PROJECT.

A.1 First the data for both series are entered; the data vector at time t is Y_t =(Y1DAT(t), Y2DAT(y))^T.

A.2 The development of a State Space program requires each given data series in the multidimensional process to be stationary or to be induced to a stationary process. Options are available for automatic and specified differencing.

A.3 The dimension of the State Vector X_t is increased one element at a time until further additions would be correlated to previous elements of the basis as judged by a suitable test criterion. The first element added to the State Vector is Y1DAT(t). This element is the current value at time t of the first time series. The second element added to X_t is Y2DAT(t) as being the current value at time t of the second time series. The third element added to X_t is Y1DAT(t+1). This element is the best one-step-ahead predictor at time t for the first time series. The fourth element added to the State Vector X_t is Y2DAT(t+1). This element is the best one-step-ahead predictor at time t for the second time series: the best predictor for Y2DAT(t+1) given the data Y_t, Y_{t-1}, Y_{t-2}. . . Similarly, the fifth element added to the State Vector is Y2DAT(t+2). This element is the best two-steps-ahead predictor at time t for the second time series. No further uncorrelated elements can be added to the State Vector. Hence, the dimension of the State Vector X_t is 5x1. X_t has the form:

$$x_t = \begin{bmatrix} Y1DAT(t) \\ Y2DAT(t) \\ Y1DAT(t+1) \\ Y2DAT(t+1) \\ Y2DAT(t+2) \end{bmatrix}$$

A.4 The transition matrix F is calculated automatically by the program according to

$$x_{t+1} = Fx_t + Gu_{t+1}$$

F has a dimension of nxn or in this case, 5x5.

A.5 The program automatically calculates the matrix H which relates the State Vector x_t to the observation time series vector Y_t, i.e., $Y_t = Hx_t$ where H has dimension pxn or, in our case, 2x5.

A.6 G is also calculated automatically from the equation $x_{t+1} = Fx_t + Gu_{t+1}$ where u_{t+1} has dimension pxl or in our example 2xl and is defined as $u_{t+1} = Y_{t+1} - Y_{t+1/t}$ and x_{t+1} is the State Vector at time t+1. G and H are the matrices calculated above.

A.7 The innovation vector u_t for the State Space model is the vector of residuals errors for one-step-ahead forecasts. The residual errors for a well-fitted model are uncorrelated and identically distributed random variables with zero mean.

The PROJECT program calculates the mean vector for the residual errors.

A.8 The program calculates the covariance matrix Σ of the residual errors. For a well-fitted model, the diagonal elements of the

covariance matrix (i.e. the residual variances) should be significantly less than those of the original data. The off-diagonal elements of the residual covariance matrix representing the cross covariance of the residuals should be close to zero since one of the assumptions for the residuals is to be uncorrelated.

A 9. The normalized correlations for the residuals are calculated. The correlations of the residuals should be small for a well-fitted model. A measure of their significance is provided by their standard deviation given after the last correlation. No residual correlations should be greater than two standard deviations away from zero for a well-fitting model.

A10. An overall measure of goodness of fit of the State Space model is provided by a Relative Goodness of Fit Criterion. The higher the value of the Relative Goodness of Fit, the better the model fits the data.

A11. The program calculates an R^2 test statistic equal to one minus the ratio of the residual error variance to the beginning variance of each of the original series. In our example, the first element is the R-squared test of .829 indicating that 82.9% of the variance in the first time series has been explained by the State Space model.

Conclusions

2.6.1 Kalman Filters. Kalman Filtering based on State Space models is emerging as a general approach to forecasting. Some of its

advantages are:

i. It is a flexible approach to obtain optimal forecasts for a large number of different models for linear, non-linear and time varying processes.

ii. Kalman Filtering is the most general approach to statistical estimation and prediction. It has been shown by Harrison and Stevens [63] that all forecasting methods are special cases of Kalman Filters.

iii. Kalman Filters provide complete probability distribution on forecasts so that confidence limits and expected values of loss functions can be evaluated explicitly.

iv. These filters can deal with changes in the model, the parameters and the variances. The filters not only can detect significant changes in the time series but also can adapt to these changes.

Limitations: The difficulty with Kalman Filters is that a lot of technical questions are yet to be answered. The approach itself has grown out of engineering. Many statisticians and operations researchers know little about the approach or have difficulty understanding the State Space notation. Practical difficulties such as initial estimates for parameters, variances, covariances and the transition matrix still exist.

2.6.2 State Space Forecasting. When comparing State Space forecasting to the Box-Jenkins approach, Drs. Alan Cameron and Raman Mehra, the creators of State Space forecasting, find seasonal advantages

of the next method over BJ. These advantages are:

i. No sharp discontinuity in complexity when going from univariate to multivariate case

ii. Order selection done automatically

iii. Generally nonlinear search not required

iv. One computer command performs all three steps of Identification, Estimation and Forecasting

v. Computationally faster procedure.

In respect to the superiority of the State Space models over the BJ approach, the International Airline Passenger data of Table 4.1 were trained via the PROJECT program of the State Space forecasting. State Space forecasting does not make any provision for transformation of the data in order to induce homoscedasticity. Apparently homoscedasticity is expected to be induced by differencing the time series. The model that the State Space program identified has the form:

$$(1-\phi B)\ (1-B)\ (1-B^{12})\ Z_t = a_t$$

This is very different from the model which Box and Jenkins [11, p. 306] identified for the classic International Airlines data:

$$(1-B)\ (1-B^{12})\ Z_t = (1-\theta B)\ (1-\theta B^{12})a_t$$

and which has been generally acknowledged to be the definite model for description of this well-known series. Clearly, the State Space program showed no MA components.

When State Space forecasting is compared against the Bayesian Approach, Dr. Cameron and Mehra claim

i. Bayesian Forecasting requires the user to specify a model structure. State Space Forecasting identifies the model structure from the data.

ii. All the desirable features of Bayesian Forecasting (short data lengths, a priori information, adaptability) can be incorporated in State Space Forecasting. Different State Space models can be related to each other.

iii. State Space Forecasting generates and bridges the gap between the Box-Jenkins and Bayesian Forecasting.

In chapter seven the controversies in literature on forecasting will be discussed; a part will be dedicated to the State Space approach and its competative edge over the other forecasting methods as claimed by Dr. Mehra and Alan Cameron.

EXAMPLE 1

A STATE SPACE MODEL FOR A
TWO DIMENSIONAL PROCESS

```
RUNNH:  PROBE FIN

COMMAND ?SEL Y1DAT, Y2DAT

COMMAND ?PROJECT Y1DAT, Y2DAT PRINT RES, MAT, COR                    A.1
```

STATE SPACE FORECAST

500 OBSERVATIONS, 2 SERIES

RANGE = 001 -500

NO REGULAR OR SEASONAL DIFFERENCING PERFORMED A.2

THE FOLLOWING ARE THE ELEMENTS OF THE STATE VECTOR A.3
Y1DAT(T)
Y2DAT(T)
Y1DAT(T+1)
Y2DAT(T+1)
Y2DAT(T+2)

F MATRIX
5 ROWS 5 COLUMNS A.4

ROW	1	0.0000	0.0000	1.0000	0.0000	0.0000
ROW	2	0.0000	0.0000	0.0000	1.0000	0.0000
ROW	3	-.4411	0.0445	0.9088	0.0058	0.0000
ROW	4	0.0000	0.0000	0.0000	0.0000	1.0000
ROW	5	-.0006	0.4808	0.0213	-1.2196	1.5656

H MATRIX
2 ROWS 5 COLUMNS A.5

ROW	1	1.0000	0.0000	0.0000	0.0000	0.0000
ROW	2	0.0000	1.0000	0.0000	0.0000	0.0000

G MATRIX
5 ROWS 2 COLUMNS A.6

ROW	1	1.0000	0.0000

5 ROWS

ROW	2	0.0000	1.0000
ROW	3	1.6910	-.0692
ROW	4	0.0144	1.4841
ROW	5	-.0265	1.0198

RESIDUAL MEAN VECTOR A.7

2 ROWS 1 COLUMNS

ROW	1	-.0007
ROW	2	0.0001

RESIDUAL COVARIANCE MATRIX A.8

2 ROWS 2 COLUMNS

ROW	1	0.9579	-.0216
ROW	2	-.0216	1.0669

NORMALIZED CORRELATIONS FOR RESIDUALS A.9

2 ROWS 2 COLUMNS 19 LAGS

LAG	1		
ROW	1	0.072	0.013
ROW	2	0.017	0.008
LAG	2		
ROW	1	-.008	-.005
ROW	2	-.040	-.002
LAG	3		
ROW	1	-.017	0.021
ROW	2	0.042	-.033
LAG	4		
ROW	1	0.033	-.010
ROW	2	-.027	0.004
LAG	5		
ROW	1	0.019	-.008
ROW	2	-.037	0.013
LAG	6		
ROW	1	0.049	0.004
ROW	2	-.028	-.013
LAG	7		
ROW	1	0.016	0.056

ROW	2	0.039	-.030
LAG	8		
ROW	1	0.007	-.015
ROW	2	0.016	0.045
LAG	9		
ROW	1	-.041	-.036
ROW	2	0.012	0.050
LAG	10		
ROW	1	0.029	0.013
ROW	2	0.009	0.039

STANDARD DEVIATION IN RESIDUAL CORRELATIONS 0.04

RELATIVE GOODNESS OF FIT (16 D.F.) 1615.7239 A.10

R SQUARED TEST	ORIGINAL DATA	DIFFERENCED DATA	A.11
Y1DAT	0.82936	0.82936	
Y2DAT	0.78328	0.78328	

COMMAND ?

CHAPTER FOUR

ADAPTIVE FILTERING

> Any astronomer can predict just where every star will be at half past eleven tonight; he can make no such prediction about his daughter
> -James Truslow Adams

4.1 Introduction

This chapter will deal with techniques developed to modify forecasting parameters in a dynamic manner. In this sense, the parameters may be altered at each time period depending on the flow of forecast errors or the additional information that has built up and concerns the behavior of the time series.

The adaptive filtering approach of the general ARMA model of the form

$$\hat{X}_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + e_t$$

$$- \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} \qquad (4.1.1)$$

and the adaptive-response-rate-single exponential smoothing (ARRSES) will be examined in detail. Adaptive filtering is to the general ARMA model of (4.1.1) what ARRSES is to single exponential smoothing.

There are several adaptive methods of forecasting available. The author feels that the ones proposed by Trigg and Leach [136] for the case of single exponential smoothing and the one proposed by Spyros Matridakis and Steven Wheelwright [82] for the case of integrated ARMA model respectively are representative, probably most widely applied and perform reasonably well.

4.2 Adaptive Response Rate Single Exponential Smoothing (ARRSES)

Adaptive-response-rate-single exponential smoothing (ARRSES) has an advantage over single exponential smoothing: it does not require specification of a value for α. This feature is particularly attractive when an extremely large number of items needs to be forecasted. Even more important is the ability of ARRSES to change the value of α on an ongoing basis when changes in the pattern of the data occur.

It is a common procedure to use different values of the smoothing constant at different times in the analysis of a common series. A large value of α, which seemed to be appropriate at the start of an exponential smoothing procedure when the procedure is based on only a few observations, is no longer appropriate in light of more information concerning the time series. Instead, a smaller value of α should be introduced since the forecast should not any more rely heavily on early forecasts. ARRSES is adaptive in the sense that the value of α will change automatically when there is a basic pattern requiring a different alpha.

ARRSES is based on the same equations as a single exponential smoothing. Formula (2.2.4) of chapter II gives us the equation for

single exponential smoothing.

$$S_{t+1} = \alpha Z_t + (1-\alpha)S_t \tag{4.2.1}$$

which in the case of α changing over time becomes

$$S_{t+1} = \alpha_t Z_t + (1-\alpha_t)S_t \tag{4.2.2}$$

The parameter α varies and is calculated according to

$$\alpha_{t+1} = \left| \frac{E_t}{M_t} \right| \tag{4.2.3}$$

where

$$E_t = \beta e_t + (1-\beta)E_{t-1} \tag{4.2.4}$$

$$M_t = \beta|e_t| + (1-\beta)M_{t-1} \tag{4.2.5}$$

and

$$e_t = Z_t - S_t \tag{4.2.6}$$

β is usually set to .1 or .2

From (4.2.3) we can see that α will vary with the ratio of actual to absolute error values. Equation (4.2.4) smoothes the actual errors while (4.2.5) smoothes the absolute error values. Again, the objective of this procedure is to allow the forecasting model to react faster to sudden changes in the level or a shift in the demand pattern.

Several other techniques that are beyond the scope of this dissertation have been developed to automatically control the values of one or more smoothing constants. The most widespread and established

ones are found in W.M. Chow [35] whose adaptive control technique monitors several exponential smoothing constants, Roberts and Reed [119] whose adaptive-control technique, called SAFT, an acronym for Self Adaptive Forecasting Technique, is an extension of Chow's method, and Montgomery [95] whose scheme is similar to the Roberts and Reed method. Montgomery and Johnson [96], pp. 175-188] have an excellent presentation of all the above adaptive control forecasting methods, their advantages and disadvantages, and differentiates these methods according to their accuracy and computational simplicity.

4.3 An Integrated ARMA Adaptive Filtering for Time Series Forecasting

4.3.1 Historical Overview. Spyros Makridakis and Steven Wheelwright [83, p. 676] define adaptive filtering to be a method of time-series forecasting which determines the optimal set of parameters (weights) to be applied in such a way that the square of the errors will be minimized. It is a time series method belonging to the group of ARMA techniques.

Forecasting with adaptive filtering is first reported by the above authors in two articles in 1973. The authors extend the concept of adaptive filtering from its limited AR form to sequential ARMA models [82]. Several other people contribute to the popularity of this method. Among them are C.D. Lewis with two papers in 1973 and 1975 and A.J. Long who in 1975 extended the concept of adaptive filtering to include mixed ARMA models which in turn led to the generalization of the method so it could deal with all types of

processes and data [83, p. 309]. Since then, the technique of adaptive filtering has become widely used, and its theoretical development has evolved in several directions. Use of adaptive filtering was reported to have been extended to multivariate processes [83, p. 309].

A theoretical description of adaptive filtering for an ARMA model will be given in the next sections based primarily on the pioneer work of Spyros Makridakis and Steven C. Wheelwright [82, 83 and 79].

4.3.2 The Adaptive Filter. The use of heuristic optimization procedure - the method of deepest descent - in identifying the filter parameters and obtaining values for the impulse response function is credited to Widrow, Makridakis and Wheelwright [82, p. 426]. Most other filtering techniques use analytical optimization procedures to determine filter parameter values.

As a methodology for time series forecasting, adaptive filtering bases its forecast on a weighted sum of past observations. The complete expression of the adaptive filtering AR model is, therefore,

$$X_t = \phi_{1t}X_{t-1} + \phi_{2t}X_{t-2} + \phi_3 X_{t-3} + \ldots + \phi_{pt}X_{t-p} + e_t$$

$$= \sum_{k=1}^{p} \phi_{kt}X_{t-k} + e_t \qquad t = p+1, p+2, \ldots, n. \quad (4.3.2.1)$$

Equation (4.3.2.1) is one of the two major components of Box-Jenkins methodology for time series analysis. (The other component is the MA process and consists of a weighing of past error terms.)

The fact that equation (4.3.2.1) is part of the ARMA models of the Box-Jenkins methodology has caused some people to overlook the major differences between the AR model of the Box-Jenkins methodology and equation (4.3.2.1). The difference is that in the ARMA models of the Box-Jenkins methodology the parameters ϕ_i of the AR model are fixed while the adaptive filtering approach has the particularly attractive ability to adjust the parameters ϕ_i of (4.3.2.1) as new data become available. However, this may be an advantage with some data series and a disadvantage with others [79, p. 227].

Some of the advantages of adaptive filtering are that it is conceptually and computationally simple; it requires only a small number of data points; it has few constraints connected with its use, and, above all, is a truly self-adaptive method that can adjust automatically to changing data patterns. The formula for modifying the filtering parameters of (4.3.2.1) according to the method of steepest descent is

$$\phi'_{it} = \phi'_{it-1} + 2ke_t X_{t-i} \qquad (4.3.2.2)$$

$$i = 1, 2, \ldots, p$$

$$t = p+2, p+2, \ldots, n$$

where ϕ'_{it} is the new adapted parameter

ϕ'_{it-1} is the old parameter

k is a learning constant that determines the speed of adaptation

e_t is the residual error

X_{t-i} is the time-series value at period t-i

Expression (4.3.2.2) will approach the optimal parameter values (the ones that minimize MSE) as long as k lies within certain limits. It has been shown by B. Widrow that by repeatedly using equation (4.3.2.2) under the necessary conditions parameter values that result in successively smaller MSE can be easily attained [83, p. 287].

The subjective feelings of the forecaster can be incorporated in the value of k. For example, when the forecaster expects a continuation of the basic data pattern, a "normal" value of k, such as 1/p, can be used. If change in the data pattern is expected, then the value of k might be increased; whereas a change in the amount of randomness, but no basic change in the pattern, might lead the forecaster to decrease the value of k [82, p. 427].

To speed convergence to the optimum values, a value of k as close to 1 as possible will usually require fewer iterations of (4.3.2.2) to achieve a minimum MSE. Such a value of k, nevertheless, can result in divergence: instead of a smaller MSE on each iteration, the MSE increases. To avoid this problem, k must be set equal to or smaller than 1/p [83, p. 287]. When this is done and the data are standardized, convergence towards a minimum MSE is assured. The sufficient conditions for convergence of the adaptive filtering algorithm (AFA) are given in [83, pp. 310-313] where for an AR process this condition is assured as long as the learning constant k is within the bounds

$$0 < K < \frac{1}{\sum_{i=t-p}^{t-1} X_i^2}, \quad t = p+1, p+2, \ldots n \qquad (4.3.2.3)$$

For MA models, e_i^2 should be substituted for X_i^2 to obtain

$$0 < K < \frac{1}{\sum_{t=i-q}^{t-1} e_i^2} \qquad (4.3.2.4)$$

Finally, for mixed ARMA models, both X_i^2 and e_i^2 should be included in the denominator giving

$$0 < K < \frac{1}{\sum_{i=t-p}^{t-1} X_i^2 + \sum_{i=t-q}^{t-1} e_i^2} \qquad (4.3.2.5)$$

Adaptive filtering had three specific weaknesses in the past, that have recently been overcome. These weaknesses were:

i. Failure to make efficient use of prior information in arriving at initial estimates of the parameter values.

When adaptive filtering is being applied and no prior information is available, then a practical procedure for initializing the parameters is to set

$$\phi_1 = \phi_2 = \phi_3 = \ldots = \phi_p = \frac{1}{p} \qquad (4.3.2.6)$$

Using then the method of deepest descent formula (4.3.2.2), we can obtain an improved set of weights as more information becomes available. This process consists of calculating X_t and e_t from equation (4.3.2.1) using the first p data points. We substitute these values and the appropriate X_i into (4.2.2.2) to determine an improved set of weights Φ'.

> This sequence of steps can then be repeated by dropping the first data point in the set of pX values and adding the next data point (p+1). Thus, a new set of data points can be combined with the revised weights to obtain values for $\hat{X}_{t+1}$ and e_{t+1}. This procedure can be applied until all existing observations have been used. [82, p. 429]

One now can return to the first p data points, retrain the most recent set of weight values and repeat the process. If the data series is stationary and if k satisfies the necessary conditions [83, pp. 310-313], then this process will converge toward the optimal set of weights Φ_i^*.

This iterative procedure is computationally time consuming. The use of Yule-Walker equations (4.3.2.7) rather than equation (4.3.2.6) can eliminate this drawback. The Yule-Walker equations express the autocorrelations of p time lags as a linear function of the autocorrelations of other time lags in the following manner.

$$\begin{aligned}
\rho_1 &= \phi_1 + \phi_2\rho_1 + \ldots + \Phi_p\rho_{p-1} \\
\rho_2 &= \phi_1\rho_1 + \phi_2 + \ldots + \phi_p\rho_{p-2} \\
&\vdots \\
\rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \ldots + \phi_p
\end{aligned} \qquad (4.3.2.7)$$

In matrix form, equation (5.3.2.7) can be written as

$$\rho_p = \Phi_p P_p \qquad (4.3.2.8)$$

Solving (4.3.2.8) for Φ_p, we get

$$\Phi_p = P_p^{-1}\rho_p \qquad (4.3.2.9)$$

By replacing the theoretical autocorrelations ρ_p by the estimated autocorrelations r_p, we can determine, using (4.3.2.9), a starting set of weights for applying the iterative procedure described above (i.e., the Yule-Walker equations). Doing this, we substantially reduce the number of computations required to approach the optimal set of weights, Φ_p^*.

ii. Difficulty of comparison of applications among different time series if data are not standardized.

Certain benefits are obtained by standardizing the X_i values so they fall between 0 and 1; although, it is possible to apply adaptive filtering to a time series without standardizing the data values.

When the time series is stationary, standardization assures that equation (4.3.2.2) will give convergence providing learning constant k is less tha 1/p and tends to speed up the process. One satisfactory and easy procedure for standardization consists of dividing each X_i in the set of such p values by the square root of the sum of the squared values. The result of the division are the standardize values

$$X_t^* = \frac{X_t}{\sqrt{\sum_{i=1}^{p} X_i^2}} \qquad (4.3.2.10)$$

iii. Required stationarity in the data.

The concept of stationarity that so far has been assured is not expected to hold in most business time series. We can overcome this

problem of not requiring stationarity through repeated use of adaptive filters which is equivalent to taking successive differences of the time series. Indeed, if the time series X_t (t=1,2,....n), is not stationary in the first degree, a filter with weight equal to 1(i.e. $\phi_{it} = 1$) and k=0 can transform the series into one that is stationary at the first level. Equation (4.3.2.1) becomes

$$\hat{X}_t = X_{t-1} \qquad (4.3.2.11)$$

and the error becomes

$$e_t = X_t - \hat{X}_t = X_t - X_{t-1} = (1-B)\, X_t \qquad t=2,3,\ldots,n$$

where B is the backward shift operator. The above equation (4.3.2.11) is exactly equivalent to taking the first difference (1-B) of the series. If non-stationarity exists at the second level, a filter with $\phi_{1t}=2$, $\phi_{2t}=-1$, k=0 can be applied to give

$$\hat{X}_t = 2X_{t-1} - X_{t-2} \qquad (4.3.2.12)$$

and the error

$$e_t = X_t - \hat{X}_t = X_t - 2X_{t-1} + X_{t-2} = (1-B)^2 X_t.$$

This filter is equivalent to taking the second difference $(1-B)^2$ of the series.

In practice, examination of the autocorrelation function for

the time series is used to determine stationarity.

Whether or not the aim is filtering (i.e., separate the noise from the signal or pattern) or forecasting, elimination of non-stationarity through the use of a sequence of filters will eventually give a differenced series that is stationary. Once stationarity has been achieved, the output of the filters can be defined as:

$$X'_t = e_t$$

and equation (4.3.2.1) becomes

$$X'_t = \phi_1 X'_{t-1} + \phi_2 X'_{t-2} + \ldots + \phi_p X'_{t-p} + e'_t$$

where X'_t is now a stationary series.

Choosing the appropriate degree of the AR process requires examination of the autocorrelation coefficients and partial autocorrelations. Makridakis and Wheelwright [79] give an empirical rule of thumb that can serve as a guideline when using adaptive filtering.

> As a practical rule, one can choose the degree of the AR process - the number of weights - to equal the time lag corresponding to the largest positive autocorrelations coefficient after a time lag of 3. This is equivalent to setting p equal to the length of seasonality. If the data are not seasonal, p can be set equal to 1, 2, or 3. Setting p equal to, say, 12 if monthly seasonality is present may result in specifying too many parameters (a non-parsimonious model) but this is a disadvantage to be compared with the advantage of considerably less complexity in using the method. Similarly, for non-seasonal data, p can be set equal to 3, thus providing a simple yet general approach for dealing with seasonal and non-seasonal data. Again, doing so will mean that too many parameters may be used. When this is the case, their values will be close to zero. For forecasting purposes, therefore, there will be little harm. [79, pp. 231-232]

The choice of the degree for the general MA model

$$\hat{X}_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} \qquad (4.3.2.13)$$

is similar to the AR choice. The degree should be equal to the time lag of the largest autocorrelation coefficient. Following this rule will usually result in an adequate model and in random residuals [79, p. 232].

A much more general approach in using adaptive filtering can be achieved with the standard ARMA model of the form:

$$\hat{X}_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + e_t \qquad (4.3.2.14)$$

$$- \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q}$$

Equation (4.3.2.2) can be used to modify the ϕ_i parameters using adaptive filtering, and its equivalent form (4.3.2.15) can be used to modify the θ_i parameters:

$$\theta'_{it} = \theta_{it-1} - 2ke^*_t e^*_{t-i} \qquad (4.3.2.15)$$

Inclusion of the MA term in the general ARMA model adds more complexity than is justified by modest gains in accuracy. Thus, restricting the approach of repeated applications of adaptive filtering to models only of the form given by (4.3.2.1) is recommended [82, p. 432].

4.3.3 Comparison of ARMA Methodologies and Adaptive Filtering. There are some similarities between ARMA forecasting methodologies and

adaptive filtering - mainly in the representation of the model. But there are also some very unique differences. The most important difference is that the parameters ϕ_i are updated in the adaptive filtering approach as new observations become available; whereas, in the Box-Jenkins methodology, these parameters remain fixed. Also equation (4.3.2.2) involves use of the error terms which typically are used only in the MA component of the ARMA models.

Adaptive filtering combining characteristics of both AR and MA processes requires very little knowledge to identify the appropriate AR or MA model. In the case of an AR model, for example, the order of this model is generally set equal to the length of seasonality which can be determined from the autocorrelation coefficients. If the data are not seasonal, then p is set equal to 2 or 3 [82, p. 427]. Makridakis and Wheelwright see no reason why adaptive filtering can not be used in a completely parsimonious model as is the Box-Jenkins methodology. They just believe that whatever is lost in parsimony is gained in simplicity; so, in their work on adaptive filtering, parsimony has intentionally been considered secondary.

Adaptive-response-rate exponential smoothing and adaptive filtering deal much better with step changes and transient situations than methods based on classical statistics because they update their parameters to account for changes in pattern. Furthermore, they can deal with changes in trend much better than fixed model/fixed parameters methods. However, even these two methods cannot do as well as the Kalman filters which deal with variable models, variable parameters

and variable variances simultaneously [83, p. 424].

4.3.4 Example of Adaptive Filtering and Comments. The set of time series data for international airline passengers taken from Box-Jenkins, Table 4.1, was used for forecasting with both the adaptive filtering and Box-Jenkins approach. A program written in BASIC was used to apply the adaptive filtering approach to the first 102 observations, and also, some program was used to obtain a forecast applying the Box-Jenkins methodology. (Makridakis and Wheelwright [82] do not mention which program was used for the BJ methodology or who did the forecasting.) The BJ methodology resulted in a smaller MSE over the data values for which the ARMA model was fitted, but adaptive filtering gave a smaller MSE for the forecast values as shown in Table 4.2. In other words,

> ...it appears that the Box-Jenkins approach 'overestimates' the series by including some of the noise as part of the pattern. It is perhaps somewhat ironic that Box-Jenkins is used largely for forecasting and adaptive filtering is used primarily for filtering when Table 4.2 suggests that their relative advantages indicate that just the opposite might be more appropriate....

Makridakis and Wheelwright imply that classical statistical estimation procedures attempt to minimize the MSE of the fitted model which is appropriate for the past but may be inappropriate for the future. If the criterion is to minimize the MSE of the model fitted to historical data, classical estimation procedures can provide a minimum MSE by assuming a fixed model with fixed parameters and variance. However, when the MSE of the future is to be minimized, adaptive-response-rate-exponential smoothing, adaptive filtering or the Kalman filters can do as well for future periods as classical estimation

TABLE 4.1

INTERNATIONAL AIRLINE PASSENGERS: MONTHLY TOTALS
(THOUSANDS OF PASSENGERS)
JANUARY 1949 - DECEMBER 1960*

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	183	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	281	277	317	313	318	174	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	505	508	461	390	432

*144 observations

procedures even when the data pattern does not change. When there are changes in the data, since these changes can be identified and incorporated in forecasting, these methods may do better than the classical estimation procedures [83, p. 424].

TABLE 4.2

COMPARISON OF FORECASTING RESULTS
(CUMULATIVE DECOMPOSED FORECASTING RESULTS)

	Mean percentage error	Mean absolute percentage error	Mean squared error
1-12 months ahead			
Adaptive filtering	0.923	3.430	208.8
Box-Jenkins	-6.44	6.799	989.4
13-24 months ahead			
Adaptive filtering	2.644	4.130	364.3
Box-Jenkins	-10.141	10.323	2,042.4
25-36 months			
Adaptive filtering	6.663	7.640	1,752.8
Box-Jenkins	-11.656	11.777	2,928.1

CHAPTER FIVE

ADAPTIVE ESTIMATION PROCEDURES

> I always avoid prophesying beforehand because it is much better to prophesy after the event has already taken place.
>
> -Winston Churchill

5.1 Introduction

Adaptive estimation procedure (AEP) is another approach that has been developed to modify forecasting parameters in a dynamic manner. It is a technique similar to the approach of adaptive filtering in the sense that it, too, is a fully automated procedure for estimating and updating time-varying parameters of linear and nonlinear models in real time [18].

AEP was originally developed for multivariate time series causal modeling by Carbone in 1975 and Carbone and Longini in 1975 and 1976 [30, 141]. In 1979, Stuart Bretschneider, Robert Carbone and Richard Longini [18] extended the applicability of the Carbone-Longini adaptive estimation procedure and adapted it to univariate time series forecasting through the use of the distributed lag forecast model. In "An Adaptive Approach to Time-Series Forecasting" a description of AEP is given; the similarities and differences between the adaptive filtering (AP) and adaptive estimation procedure (AEP) are pointed out; a comparison with adaptive filtering, the Box Jenkins methodology and multiple regression analysis, as it applies to

time series analysis, is provided. The authors of this paper [18] also provide the background for the discussion presented in the next sections.

5.2 Carbone-Longini AEP Approach

Consider the following model for time series forecasting:

$$Y_i = \sum_{k=1}^{p} \phi_{ik} X_{ik} + \epsilon_i \qquad i = 1, 2, \ldots, T \qquad (5.2.1)$$

where Y_i is the p+i th element of the series

p is the number of lagged variables specified

X_{ik} is the p+i-k element of the time series

ϕ_{ik} is the parameter associated with the kth lagged variable at observation i

ϵ_i is the random noise or disturbance term

To ensure stationarity of the time series, first or second order differences are taken, and after selecting an appropriate value for p, the time-varying parameters ϕ_{ik} of equation (5.2.1) are left to be estimated.

Estimates of the ϕ_{ik} parameters are obtained in the case of the adaptive filtering approach (AF) by applying the Widrow-Hoff least mean square (LMS) algorithm [142, 141]. As was pointed out in Chapter IV of this dissertation, close approximate solutions to the Wiener Hopf equation are obtained via the method of steepest descent. This equation defines, for stationary data, optimal estimates of fixed parameters of linear models through minimization of mean square error (MSE) [18, p.233].

The LMS adaptive algorithm can be written for model (5.2.1) as

$$\hat{\phi}_{ik} = \hat{\phi}_{i-1,k} + 2k\, X_{ik}(Y_i - \hat{Y}_i) \qquad k = 1, 2, \ldots N \qquad (5.2.2)$$

where k is the damping factor to control stability such that $0<k<1$ and

$$\hat{Y}_i = \sum_{k=1}^{p} \hat{\phi}_{i-1,k}\, X_{ik} \qquad (5.2.3)$$

is a predicted value of Y_i computed by applying the estimates for observation i-1 to the data at i.

The original derivation of equation (5.2.2) employed a standard transformation to the data. This transformation introduced an intercept term into the forecast and required transformation of parameter estimates between each forecasting and estimation step [18, p. 234]. As was pointed out in Chapter IV, Wheelwright and Makridakis modified LMS by introducing the standardization of the input values as

$$X_{ij} = \frac{X_{ij}}{\sqrt{\sum_{k=1}^{p} X_{ik}^2}} \quad , \qquad j=1, 2, \ldots p+1 \qquad (5.2.4)$$

The transformation of (5.2.4) avoids the creation of the intercept term but does not eliminate the need for transformation of parameter estimates between each forecasting and estimation step. To overcome the problem, adaptive filtering was implemented according to the following procedures:

i) To determine whether or not difference transformation of the data is necessary, autocorrelations of p or more lags are estimated and analyzed.

ii) Initial estimates for the parameters are specified in order to apply equation (5.2.2). The Yule-Walker approach is one among several approaches suggested to achieve this goal.

iii) Once initial estimates for the parameters are specified, a value for the k is selected and equation (5.2.2) is applied iteratively starting with the first observation and continuing through the final observation. Taking the ending parameter estimates from the first iteration, the process is repeated several times until reduction between iterations in the mean square error (MSE) becomes negligible. The resulting parameters from the above process are then used to forecast the time series for time T+1. Equation (5.2.2) provides the basis for revising values of the parameters as new observations become available. At this point, though, many believe that it is uncertain that the application of (5.2.2) to new observations will allow the parameters to accurately adapt to new conditions [18, p. 234]. Widrow, McCool, Larimore and Johnson [144] claim that the LMS algorithm has been developed to provide approximate estimates of fixed parameters that minimize mean squared error - not to capture time variation in parameters. Preliminary results indicate difficulties with LMS to track time variation in parameters [18, p. 234].

Here is where Robert Carbone and Richard Longini introduced an alternative form of updating estimates of the parameters in (5.2.1) by using the following equation:

$$\hat{\phi}_{ij} = \hat{\phi}_{i-1,j} + \left| \hat{\phi}_{i-1,j} \right| \left[\frac{(Y_i - \hat{Y}_i)}{\hat{Y}_i} \cdot \frac{X_{ij}}{\bar{X}_{ij}} \cdot k \right] \quad \text{for all } j \qquad (5.2.5)$$

$$\text{where } \bar{X}_{ij} = \alpha X_{ij} + (1-\alpha)\bar{X}_{i-1,j}, \quad 0 < \alpha < 1$$

and, thus, forming another methodology of adaptive filtering, the Carbone-Longini Adaptive Estimation Procedure (AEP) approach. In contrast to (5.2.2), "the AEP algorithm was established in a heuristic manner through experimentation and logical considerations, rather than via a pure deductive process. It is based upon engineering concepts in feedback theory and pattern recognition and is designed to capture without use of a priori knowledge the time variation that may arise in parameters." [18, p. 235]

AEP is very similar to AF in the sense that autocorrelations are first examined, initial estimates of the parameters and a value for k are selected followed by implementation of a training process. In AF, the training process is given by (5.2.2) whereas, in AEP, it is given by (5.2.5). Also, in AEP no standardization of the data is necessary since it is built into (5.2.5). There are two principal distinctions within the training process. One complete training iteration consists of initially moving through the data from observation one to T followed by moving backwards beginning with T to the first observation. This comprises a complete training cycle. The forward-backward mode in the training process eliminates potential problems in phase shift [18, p. 235]. Also, in equation (5.2.5) a correction limit may be imposed to detect unwarranted overreactions to possible outliers or bad data. Once a final set of parameters is estimated, forecasts are made on the basis of the terminal values of parameters obtained from the forward portion of the last training cycle. As new observations become available, application of equation (5.2.5) allows the parameters of equation (5.2.1) to be updated

to new conditions that may arise.

In both the AF and the AEP algorithms, the selection of the damping factor k plays an important role. A large value of k will result in faster adaptation but, at the same time, may cause oscillations. Under stationary conditions, a small value for k will cause slow adaptation, and since overreactions to transient errors are reduced, we will end up with best fit performance. In contrast, under dynamic environments, a compromise between fast adaptation, necessary to track variations in the parameters and slow adaptation for containing unwarranted oscillations, will result in best performance on historical data.

Fitting historical data and forecasting future observations, however, are different tasks that require different prerequisites [18, p. 235]. For forecasting purposes, it may be that better performance is achieved with a faster rate of adaptation (i.e., larger value for k) irrespective of conditions experienced in the past. This will lead though, in most cases, to a poorer fit of the historical data due to the increased attention (and weight for this matter) given to the more recent observations of the time series. As a result of a faster rate of adaptation, final parameter estimates would be more reflective of prevalent conditions at that time since the algorithms are then set for tracking parameter variations over the time series [18, p. 235].

5.3 A Comparative Study of Adaptive Filtering (AF), Box-Jenkins (BJ), Adaptive Estimation Procedure (AEP) and Ordinary Least Squares (OLS)

The "International airlines passengers data" found in Time Series Analysis Forecasting and Control were used again by Stuart Bretschneider,

Robert Carbone and Richard Longini in applying the AEP and OLS procedures. Box and Jenkins originally examined this time series, shown in Table 5.1, in applying their technique to modeling seasonal data [11, p. 300-321]. Later, Makridakis and Wheelwright, as mentioned in Chapter IV of this manuscript, using the above set of data, demonstrated AF and also made comparisons between the AF and BJ approaches.

The comparative study done by Bretschneider, Carbone and Longini involved training the first 102 observations of the time series for developing the forecasting model. The remaining 36 observations were used for comparison with the forecasted values so as to assess forecast performance. Forecasts were made on the basis of one month ahead, two months ahead, and so on, up to thirty-six months ahead. Each month-ahead forecast was used to project successive values of the series (a technique also known as bootstrapping). In the context of a twelve-lag distributive model, for example, the forecast for period 138 is obtained by using projected values of the series for periods 126 through 137 as though they were actual values. Tables 5.2, 5.3, 5.4, 5.6, and 5.8 first appeared in [18, pp. 236-239].

In Table 5.2 for the AF approach, prior to the estimation process, the first differences of the data were taken to eliminate non-stationarity revealed by the examination of the autocorrelation function. Makridakis and Wheelwright initialized parameters for (5.2.2) with solutions of the Yule-Walker equations and specified $k = .025$. More than twenty-five training iterations resulted in an MSE of 91.46 for the final fit.

For the AEP approach, Stuart Bretschneider, Robert Carbone and Richard Longini also used a twelve-lag formulation with a first difference

TABLE 5.1

INTERNATIONAL AIRLINE PASSENGERS: MONTHLY TOTALS
(THOUSANDS OF PASSENGERS)

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	317	313	318	374	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	606	508	461	390	432

*144 observations

TABLE 5.2

MEAN SQUARE ERROR OF FIT BY FORECASTING TECHNIQUE

Technique	Mean Square Error (MSE) of Fit
OLS	88.46
BJ	71.85
AF(LMS)	91.46
AEP	100.18

TABLE 5.3

COMPARISON OF PARAMETER ESTIMATES FOR THE TWELVE-LAG MODELS
(By Technique)

Variables	OLS	AF(LMS)	EAP	
			Begin	End
Lag 1	.0124	-.014	.0106	.0102
Lag 2	-.0831	-.101	-.0023	-.0023
Lag 3	.0140	-.014	-.0119	-.0124
Lag 4	.1083	-.110	-.0405	-.0426
Lag 5	.0092	-.039	-.0002	-.0002
Lag 6	-.0721	-.075	-.0021	-.0022
Lag 7	-.0268	-.049	-.0391	-.0415
Lag 8	-.1104	-.153	-.0147	-.0155
Lag 9	.0149	-.011	-.0114	-.0120
Lag 10	-.1065	-.140	-.0029	-.0030
Lag 11	.0865	.086	.0854	.0828
Lag 12	.9562	.925	1.0513	1.0236

on the first 102 observations to develop AEP and OLS model. In applying AEP, equation (5.2.5) was initialized with each weight equal to 1/12, k=.06, α=.01, and the initial mean of each lag variable equal to ten [14, p. 237]. The training data were processed through thirty training cycles with a correction limit of .1. Since equation (5.2.5) requires all data entries to be positive, a scaling factor needed to be added to all values when negative data entries were present. A simple method used for the above data in selecting a scale involved adding one to the absolute value of the largest negative number.

5.4 Results of the Comparative Study

Table 5.3 presents parameter estimates for OLS, LMS (AF) AND AEP twelve-lag models. Each technique identified the twelfth lag as the most significant factor. Applying these results to the forecasting process, three indicators of forecast performance were calculated over a variety of forecast horizons. Mean percentage error (MPE) provides a measure of bias by indicating the tendency of a model to over or under predict while the mean absolute percentage error (MAPE) and (MSE) are measurements of forecast performance.

Table 5.4 presents decomposed forecast statistics by separate twelve-month periods. The twelve-lag model produced better forecast results than the Box-Jenkins approach irrespective of the parameter estimation technique applied. When considering the third year out, AEP exhibits the most robust results. While the indicators of performance for AEP remained relatively constant over the time horizon, those for all other models produced exponentially increasing errors.

TABLE 5.4

COMPARISON OF DECOMPOSED FORECASTING RESULTS BY TWELVE-MONTH PERIOD

	Mean Percent Error (MPE)	Mean Absolute Percent Error (MAPE)	Mean Squared Error (MSE)
Months 1-12			
OLS	-.308	3.703	247.2
Adaptive filtering	.923	3.430	208.8
Box-Jenkins	-6.44	6.799	989.4
AEP	-1.86	4.68	448.6
Months 13-24			
OLS	1.464	3.137	206.3
Adaptive filtering	4.428	4.830	519.4
Box-Jenkins	-13.842	13.847	3095.4
AEP	-3.740	4.42	454.6
Months 25-36			
OLS	10.599	10.559	2454.1
Adaptive filtering	14.66	14.66	4529.8
Box-Jenkins	-14.685	14.685	4699.5
AEP	-1.25	4.52	546.7

Table 5.5 shows the rankings of the four methods in the various forecast horizons using both the MSE and MAPE criteria.

The cumulative results in Table 5.6 illustrate the change in performance characteristics at the time the horizon is extended. The effectiveness of the method appears to be related to the forecast horizon as shown in Table 5.7: over the first twelve months, LMS has minimum MSE; over the twenty-four month horizon, OLS performs best; over a thirty-six month period, AEP produces a superior performance.

TABLE 5.5

RANKINGS OF THE DECOMPOSED FORECASTING RESULTS BY TWELVE-MONTH PERIOD

	MAPE				MSE			
	Months				Months			
Method	1-12	13-24	25-36	Overall	1-12	13-24	25-36	Overall
OLS	2	1	2	1	2	1	2	1
AF	1	3	3	3	1	3	3	3
FJ	4	4	4	4	4	4	4	4
AEP	3	2	1	2	3	2	1	2

TABLE 5.6

CUMULATIVE OF DECOMPOSED FORECASTING RESULTS

	Mean Percent Error (MPE)	Mean Absolute Percent Error (MAPE)	Mean Squared Error (MSE)
1-12 months ahead			
OLS	-.308	3.703	247.2
Adaptive filtering	.923	3.43	208.8
Box-Jenkins	-6.44	6.799	989.4
AEP	-1.86	4.68	448.6
1-24 months ahead			
OLS	.578	3.42	226.7
Adaptive filtering	2.644	4.13	364.3
Box-Jenkins	-10.141	10.323	2042.4
AEP	-2.80	4.55	451.6
1-36 months ahead			
OLS	3.90	5.79	969.2
Adaptive filtering	6.663	7.64	1752.8
Box-Jenkins	-11.65	11.77	2928.1
AEP	-1.45	4.54	483.3

TABLE 5.7

RANKINGS OF CUMULATIVE DECOMPOSED FORECASTING RESULTS

	MAPE				MSE			
	Months				Months			
Method	1-12	1-24	1-36	Overall	1-12	1-24	1-36	Overall
OLS	2	1	2	1	2	1	2	1
AF	1	2	3	2	1	2	3	2
BJ	4	4	4	4	4	4	4	4
AEP	3	3	1	3	3	3	1	3

Table 5.8 considers the very short run by using a forecast horizon of six months.

TABLE 5.8

COMPARISON OF SHORT-RUN FORECASTING RESULTS
(1-6 months ahead)

Method	MPE (mean percent error)	MAPE	MSE	Rankings MAPE	Rankings MSE
OLS	3.196	3.196	184.6	3	3
AF	3.902	3.902	259.5	4	4
BJ	1.703	2.041	92.5	1	1
AEP	2.820	2.820	153.6	2	2

In Table 5.8, BJ produced the best performance with AEP providing the second best. This is understandable in light of BJ's superior fit for this data [18, p. 239].

5.5 Implications of the Results and Conclusions

There seems to be computationally very little difference between LMS or AF and AEP. AF and AEP produced more accurate results when compared with the BJ methodology. There also appears to be a surprising relationship between fit performance and overall forecast performance. Bretschneider, Carbone and Longini state:

> The AEP approach which yielded the largest MSE of fit for the historical data, provided the most stable and accurate forecast over the longest time horizon. The question that arises is whether this last observation is a result of the particular data examined, the methodology was applied to the data. The answer appears to be in a combination of the last two reasons. As stated before, when applying AEP or adaptive filtering, the specification of a larger learning constant during the training process allows parameter estimates to be more reactive to changes in the data. This should result in an improved forecasting capability even if a poorer fit performance is exhibited due to tracking errors incorporated in the training process [18, p. 240].

The poor performance of the BJ procedure lead Stuart Bretschneider, Robert Carbone and Richard Longini to wonder about its empirical usefulness in terms of improving forecast capability through its complex identification process. They consider AF and AEP an alternative approach to forecasting that does not depend upon a criterion of fit but rather one of tracking or reacting to changes in the time series as they occur.

In chapter VII where the controversies in forecasting will be examined, some comments will be made and some explanations will be given for the deceiving poor performance of the BJ methodology when compared with the flashy performance of the AF and AEP procedures.

CHAPTER SIX

COMBINATION OF FORECASTS

> Time present and time past are both perhaps present in time future and time future contained in time past.
>
> T.S. Eliot

6.1 Introduction

There are many cases in which two or more forecasts have been made of the same event. Most forecasters, when this occurs, tend to accept the better forecast and discard any other. J. M. Bates and C. W. J. Granger [8] note that

> Whilst this may have some merit where analysis is the principal objective of the exercise, this is not a wise procedure if the objective is to make as good a forecast as possible, since the discarded forecast nearly always contains some useful independent information. This independent information may be of two kinds:
>
> i. One forecast is based on variables or information that the other forecast has not considered.
> ii. The forecast makes a different assumption about the form of the relationship between the variables.

Of the two kinds of information stated above, the first kind leads to situations in which a combined forecast improves upon the better individual forecast more often than does the second.

The first step that a forecaster should take before combining individual forecasts is to check that the individual sets of forecasts are unbiased. Once this assumption is met, the forecaster should choose

a method for determining the weights to be assigned to each of the individual forecasts before combining.

In the next few sections, a brief discussion of the theoretical approach in combining forecasts will be made, and a few examples will be given which illustrate the superiority of the combined forecast over the individual ones. Some conclusions will also be made.

The main source of discussion to be made is based on the work and findings of Reid [117], C. W. J. Granger [53], J. M. Bates and C. W. J. Granger [8] and C. W. J. Granger and Paul Newbold [54 and 105].

6.2 Theoretical Approach of a Combined Forecast

Table 1 shows one-step forecast errors of total airline passenger-miles for the months of 1953 by the Box-Jenkins methodology and Brown's Exponential Smoothing approach. This example first appeared in Operational Research Quarterly, Vol. 20, No. 4, page 452. This is the simplest case of combining forecasts:

(Combined forecast) = ½(Brown + Box Jenkins).

For 1953, the MSE for the combined is less than for the two original forecasts. This was also done for the whole period 1951-1960 when an enumeration of these and other forecasts was made at a later stage and the combined had MSE of 130.2 versus 177.7 for the Brown and 148.6 for the Box-Jenkins approach. Thus, a very simple method of combining has provided a set of forecasts that on average is superior to both constituent forecasts.

In the above example, equal weights were given to each of the individual forecasts. In many cases, though, one would wish to give

TABLE 6.1

FORECAST ERRORS OF PASSENGER MILES FLOWN
1953

Month	Brown's Exponential Smoothing Forecast errors	Box-Jenkins Adaptive Forecasting errors	Combined Forecast (½ Brown + ½ Box-Jenkins) errors
January	1	-3	-1
February	6	-10	-2
March	18	24	21
April	18	22	20
May	3	-9	-3
June	-17	-22	-19.5
July	-24	10	-7
August	-16	2	-7
September	-12	-11	-11.5
October	-9	-10	-9.5
November	-12	-12	-12
December	-13	-7	-10
Variance of errors	196	188	150

greater weight to the better set of forecasts (i.e., forecasts that seemed to contain the lower MSE). We are faced again with the problem of determining these weights, and the intention is to choose a method which is likely to yield low errors for the combined forecasts.

In the search for more flexible weighting schemes, we assume that a linear combined forecast is obtained by giving weights k and (1-k) respectively to the two forecasts sets. The problem is one of choosing k. We also assume that both forecasts were unbiased (either naturally or after correction for the average percentage - or absolute - bias has

been applied). If we denote the error variances of the two forecasts by σ_1^2, σ_2^2 for all values of time t and by ρ their correlation coefficient, then the variance (or mean squared error since the forecasts were assumed unbiased) of the combined forecast σ_c^2 is given by

$$\sigma_c^2 = k^2\sigma_1^2 + (1-k)^2\sigma_2^2 + 2\rho k\sigma_1(1-k)\sigma_2 \qquad (6.2.1)$$

The combined forecast is also unbiased since it is the linear combination of the two sets of forecasts with weights k and (1-k). We choose to minimize the overall variance σ_c^2 with respect to the parameter k since the choice of k should be made so that the errors of the combined forecasts are as small as possible. Differentiating with respect to k and equating to zero, we obtain a solution:

$$k = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \qquad (6.2.2)$$

In the case where $\rho = 0$, equation (6.2.2) reduces to

$$k = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \qquad (6.2.3)$$

Assuming second order stationarity, it can be shown [8, p. 463] that the MSE of (6.2.1) is at least as small as that of the best individual forecast method. Since this requires that the second order moments are known ex ante ([117], p. 229), which is not the case in forecasting, J. M. Bates and C. W. J. Granger [8, p. 453] consider ways of

estimating k that neither require the knowledge of the second moments nor need quite such a prohibitive assumption as stationarity.

Equations (6.2.2) and (6.2.3) are used as a basis for some of the methods that follow below. Thought was given to the possibility that the performance of one of the forecasts might be changing over time (perhaps improving) and that a method based on an estimate of the error variance since the beginning of the forecasts might not therefore be appropriate [8, p. 453]. Consequently, methods (iii) and (iv) below give more weight to recent errors than those of the past.

Five methods are considered and outlined below. We denote by $e_{i,t}$ the error in the forecast from method i for time period t.

$$E_i = \sum_{t=T-v}^{T-1} (e_{i,t})^2$$

$$S_i^2 = \sum_{t=1}^{T-1} w^t (e_{i,t})^2$$

$$C = \sum_{t=1}^{T-1} w^t e_{1,t} e_{2,t} \qquad (6.2.4)$$

The weights k_T have in all cases been determined from past (known) errors except for k_1 which has been arbitrarily chosen as .5 for all methods. The methods are:

$$\text{(i)} \quad k_T = \frac{E_2}{E_1 + E_2}$$

$$\text{(ii)}\quad k_T = xk_{T-1} + (1-x)\frac{E_2}{E_1 + E_2} \qquad \text{for } 0 < x < 1$$

$$\text{(iii)}\quad k_T = \frac{S_2^2}{S_1^2 + S_2^2}$$

$$\text{(iv)}\quad k_T = \frac{S_2^2 - C}{S_1^2 + S_2^2 - 2C}$$

$$\text{(v)}\quad k_T = xk_{T-1} + (1-x)\left\{\frac{|e_{2,T-1}|}{|e_{1,T-1}| + |e_{2,T-1}|}\right\}$$

where u,x and w are parameters to be chosen.

Bates and Granger apply the various combining formulas to the international airline passenger data, looking at several pairs of univariate one-step-ahead forecasts, with generally successful results. The rather poor estimate of C in method (iv) led this method to give worse results than the others on the above study and the suggestion was made that (iv) ought to be modified to

$$k_T = \frac{S_2^2 - zC}{S_1^2 + S_2^2 - 2zC} \qquad 0 < z < 1.$$

Newbold and Granger []05] consider the combination of one-step-ahead Box-Jenkins, Holt-Winters, and stepwise autoregressive

forecasts for the 80 monthly series collection of both seasonal and nonseasonal macroeconomic series and micro sales data. It appears that procedures that ignore correlation between the forecast errors (methods i,ii and iii) are considerably more successful than those that attempt to take account of it. It is also true that the combined forecast of Box-Jenkins with one of the fully automatic procedures outperforms both individual forecasts on about 40 percent of all occasions for the more successful combining methods.

In another study, Granger and Newbold [56] give an example which uses 20 one-step-ahead quarterly forecasts of real inventory investment generated by the Wharton-EFU econometric model [54, p. 276]. Forecasts of the same quantity are generated from a univariate Box-Jenkins procedure and combined with the model forecasts. Although the Box-Jenkins forecast outperforms the econometric forecast, the combined forecast considerably outperforms the better individual forecast. Apparently, the econometric forecast contains very useful information absent in the Box-Jenkins forecast.

Many of the results discussed in this section can be extended to the combination of more than two forecasts as discussed in Granger [53], Reid [117] and Newbold and Granger [105]. Again, similar results, as in the case of combining two forecasts, can be obtained when we extend the combination to more than two forecasts.

6.3 Conclusion

There is a considerable gain when an improved forecast is obtained using a combination of forecasting approaches. Combining

forecasts is relatively easy to do, is often successful and eliminates the necessity of deciding which of a pair of forecasts is preferred. The reason a combined forecast may be preferable is that very often neither of the constituent forecasts is using all the data in the available information set in an optimal fashion. Granger [53, p. 160] points out that "success of combination suggests that a more general model should be attempted, including the better features of the models underlying the constituent forecasts." As a result, combining forecasts will be particularly successful when the constituent forecasts are based on quite different philosophies such as using a Box-Jenkins model and a regression model. Granger and Newbold [56] have shown that combining statistical and econometric forecasts can be highly rewarding.

CHAPTER SEVEN

CONTROVERSIES IN FORECASTING

> An unsophisticated forecaster uses statistics as a drunken man uses lamp posts - for support rather than for illumination.
> After Andrew Lang

Given the availability of so many forecasting techniques, many forecasters attempt to assess a technique's relative worth and try to reach some conclusion as to which method performs better than the rest. Many studies involving various forecasting procedures do little more than cause debates among researchers. The disconcerting aspect of these debates stems from the fact that many researchers use the various forecasting methodologies inaccurately.

One classic example of such a debate is a study by Reid [117] and a similar study by Newbold and Granger [105]. Newbold and Granger discuss the results of their work as well as Reid's work. In this study, Reid, according to Newbold and Granger in Forecasting Economic Time Series:

> ...allowed for the first time a comparison of the performance of Box-Jenkins and various exponential smoothing predictors over a large set of real data. Reid assembled a collection of 113 macroeconomic time series and generated forecasts, employing methods he deemed reasonable, to each series. Both the Box-Jenkins and Brown's generalized exponential smoothing predictors were applied to every series in the collection. Holt Winters was applied to 69 series and Harrison's to 47. The Brown predictor modified to take account of the possibility of first-order autoregression in the one-step forecast error was also evaluated....

Reid found for the one-step-ahead prediction that Box-Jenkins outperformed Brown's method on 88 percent of the trials, it outperformed Holt-Winters 70 percent and Harrison 77 percent of the time:

> In a further larger study, Newbold and Granger analyzed a collection of 106 time series, 80 of which were monthly and 26 quarterly. The collection included both seasonal and non-seasonal macroeconomic series and micro sales data. The Box-Jenkins, Holt-Winters and stepwise autoregression methods were applied to every series in the collection...[105].

The results of the comparison of the Box-Jenkins versus Holt-Winters for one-step-ahead forecasts is very similar to that obtained by Reid.

Other researchers disagree with the conclusion that the Box-Jenkins approach is superior to the one by Brown, Holt-Winters and Harrison. A study by Groff [57], for example, concludes that "the forecasting errors of the best of the Box-Jenkins models that were tested are either approximately equal to or greater than the errors of the corresponding exponentially smoothed models for most series." Similarly, Guerts and Ibrahim [51 , p 18], although they examined only a single series, find that the "exponentially smoothed models patterned on Brown's model and the Box-Jenkins approach seem to perform equally well."

The findings of Groff and Guerts and Ibrahim are interesting, but erroneously supported. Groff uses the Box-Jenkins methodology incorrectly. C. Chatfield [80 , p 129] claims that "Groff's version of Box-Jenkins bears little resemblance to that originally proposed by Box and Jenkins. Groff's results should be disregarded since the Box-Jenkins procedure was misinterpreted." Anderson [80] is more specific about Groff's use of Box-Jenkins methodology out of its context. "All

he did," Anderson claims, "was to try out a specific selection of ARIMA models." Guerts and Ibrahim, on the other hand, overgeneralized from their examination of only one time series.

Not only do forecasters disagree as to the preeminence of a single forecasting approach over another, but they also disagree over the superiority of various combinations of forecasts. Once again Newbold and Granger contributed to this field. Newbold and Granger [105] find that a "slight improvement in accuracy is obtained when the Box-Jenkins is combined with one of the fully automatic procedures of Holt-Winters and stepwise autoregressive approach." Newbold and Granger also claim that, even though the Box-Jenkins forecast on the average is better than the econometric forecast, the latter contains very useful information absent in the former. This information is brought to light when the two forecasts are combined. "After all," they claim, "it would be reasonable to expect a combination to be most profitable when the individual forecasts are very dissimilar in nature" [105]. But if one can improve forecasting accuracy by combining the forecast of a univariate Box-Jenkins model with an econometric model, such as regression, why handicap the Box-Jenkins model with an inappropriate univariate stochastic model, where a transfer function model would seem more appropriate?

As far as combining a Box-Jenkins forecast with one from the class of models in the stepwise autoregressive or the Holt-Winters method, Newbold and Granger [105, p 136] point out that "proponents of the Box-Jenkins approach may find the attempt to combine Box-Jenkins forecasts with those derived from less sophisticated univariate methods

intuitively unpromising." They acknowledge that

> After all, the class of models considered in the stepwise autoregression procedure and the non-seasonal variant of the Holt-Winters method are merely subsets of the general ARIMA class of models. It is of course true that if one knew that a given time series was generated by a particular process of the [ARIMA (p,d,q)] class, optimal forecasts could be derived from that model alone. However, much as the Box-Jenkins model-building process tells us about the underlying generating mechanism, we can never be absolutely certain that a particular model is appropriate...

To this statement of Newbold and Granger, Professor G. M. Jenkins [see 105, p. 148] replies

> Since the Holt-Winters and stepwise autoregression methods correspond to special cases of the ARIMA models proposed by Professor Box and myself, it follows that a linear combination of these forecasts with a Box-Jenkins forecast results in a forecast which corresponds to another more elaborate model in the class of ARIMA processes. If the later model gives better forecasts than the forecast from the original Box-Jenkins model, the question then arises as to why a close approximation to the more elaborate model, if needed, was not arrived at during the model building stage.

Professor M. B. Priestly agrees with Professor Jenkins' point of view on the combination of forecasts. Professor Priestly [see 105, p. 153] comments that Newbold and Granger's suggestion about combining forecasts is an interesting one, but its validity seems to depend on the assumption that the model used in the Box-Jenkins approach is inadequate; otherwise, the Box-Jenkins forecast alone would be optimal.

Still other researchers also find conflicting results when studying the combination of forecasts. J. M. Craddock [see 105, p. 156] reports success obtained from the combination of forecasts in long-range weather forecasting while D. J. Reid reports an example in which the combined forecast is outperformed by both individual forecasts. These are conflicting reports which basically reflect the decision maker's

philosophy toward forecasting situations. The element which makes these philosophies differ is mainly the forecaster's degree of expertise.

Because of these conflicting results some researchers attempt to explain how various factors affect the accuracy of a forecasting technique. Two such researchers are Chatfield and Prothero.

C. Chatfield and D. L. Prothero [33] make a critical appraisal of the Box-Jenkins procedure which is used to forecast sales of an engineered product for a lead time of up to 12 months. A total of 77 monthly observations are given. Based on the findings of their analysis Chatfield and Prothero conclude that the Box-Jenkins procedure involves a subjective element which allows one to choose from a wide class of models. They see this greater flexibility as both the strength and weakness of the Box-Jenkins approach. There is an advantage to being able to choose from a wide class of models rather than being restricted to one particular model, but at the same time, the subjective assessment involved in choosing a model means that considerable experience is required in interpreting sample correlation functions. For their set of data, Chatfield and Prothero found that the Box-Jenkins procedure was less successful since the data exhibit high multiplicative seasonal variation. They cited as one possible reason for the unsatisfactory Box-Jenkins performance the use of an erroneous logarithmic transformation of the data.

G. Tunnicliff Wilson [see 33, p. 315] pinpoints the reason for the poor performance of the Box-Jenkins methodology in the study by Chatfield and Prothero. He reports that the primary culprit of the above study

is the logarithmic transformation applied to the data in an effort to stabilize the amplitude of the seasonal variation. According to Tunnicliff Wilson, "visual examination of the transformed series showed evidence of overtransformation." As far as the inappropriateness of the Box-Jenkins methodology on multiplicative seasonal data is concerned, he adds that the models proposed by Box and Jenkins are extremely flexible and in practice have been found capable of adequately representing a wide variety of series including seasonal sales data. In an attempt to explain how different forecasters arrive at conflicting conclusions, Tunnicliff Wilson points out that "many statisticians are more familiar with the use of curve fitting methods rather than difference equations for representing time series." He believes that the unfamiliarity of some forecasters with the difference equation models can very well pose a problem and account for differing results.

Professor P. J. Harrison [see 33, p. 319] also endeavored to explain the discrepancy in the results of Chatfield's and Prothero's study. Harrison claims that the Box-Jenkins approach was used out of context. The professor explains with the following:

> If I were to visit the Amazon Jungle; meet a native; tell him about cars; give him a book on how to drive; then his chief lets him visit civilization for the first time; he sees a Rolls-Royce; gets in it and crashes it; who or what is to blame?
>
> I think that we can be pretty sure that we would not blame the Rolls. Either I, the chief or the native or all are at fault. Why? Because of course the Rolls has been tested outside its design context.

Harrison concurs with Wilson who believes that the differing results can be accounted for by evaluating the role of the forecaster and not just the forecast methodology chosen.

Newbold [see 33, p. 324] also discusses the weakness of the Chatfield and Prothero study. He questions whether forecasts made of just one series from one base point can be expected to say very much about the general merits of the forecasting method employed. As for Chatfield and Prothero's regarding of the flexibility in the Box-Jenkins approach as a potential defect, "presumably on the grounds that given enough rope a man might hang himself," Newbold also adds, "Is it not also the case that given very little rope one can do very little with it?" Just as Wilson and Harrison, Newbold too, believes that the key to the weak performance of the Box-Jenkins methodology lies in the forecaster and not in the method itself.

The debate stemming from the Chatfield and Prothero study is similar to that debate arising from the Makridakis and Hibon study of 1979. Once again the forecasters and not the methodologies examined account for the questionable results.

The purpose of Makridakis and Hibon's study is to assess forecasting performance by evaluating the accuracy of various forecasting methodologies. In their paper, "Accuracy of Forecasting," they examine 111 time series collected from a variety of sources including several countries, industries and companies. In this study, taking n_j as the number of data points in the jth series, n_j-12 points were used to develop a forecasting model and, subsequently 12 forecasts were obtained. The error e_{ij} is defined as

$$e_{ij} = X_t - \hat{X}_{tj} , \quad t = n_j-11, n_j-10, \ldots, n_j$$

where X_t is the actual data value at period t,

$\hat{X}_{tj}$ is the value forecast by the jth method and

e_{ij} is the forecast error.

To assess forecast accuracy, forecasters, and Makridakis and Hibon in their study, measure the forecast error e_t bearing in mind that different accuracy criteria may produce different rankings over a set of forecasting methodologies. The most common measures of accuracy also used by Makridakis and Hibon, are the mean square error (MSE), Theil's U coefficient and the mean absolute precentage error (MAPE) for the above data as

(a) The mean absolute percentage error (MAPE):

$$\text{MAPE} = \frac{1}{K} \sum \frac{|e_t|}{X_t} \times 100$$

(b) The mean square error (MSE) = $\sum e_t^2/K$

(c) Theil's U-statistic = $[\{\sum e_t^2/\sum (X_t - X_{t+1})^2\}]^{\frac{1}{2}}$

where $e_t = X_t - \hat{X}_t$, X_t is the actual value, $\hat{X}_t$ is the one-period-ahead forecasted value. The MSE involves a quadratic loss function and is preferred when more weight is to be given to big errors. Its disadvantage is that it does not allow for comparisons across methodologies since it is an absolute measure related to a specific series. The U-coefficient is a relative measure; it assumes a quadratic loss function

and has several other properties that make its use attractive [129, pp 21-36]. Its disadvantage is that its interpretation is more difficult than the MAPE; moreover, the U-statistic has no upper bound so a few very large values can easily distort the comparisons.

Makridakis and Hibon concluded that a decision maker who would like to apply a single forecasting method to the 111 time series would have obtained very different results depending upon which loss function he wanted to minimize and whether he wanted to minimize the errors in the model fitting or the errors in the forecasting phase. Overall, however, "he would have done as well by using simpler rather than more sophisticated methods" [80, p 116].

M. B. Priestley [see 80, p. 127] and Guerts and Ibrahim [51, p. 187] agree with Makridakis and Hibon's conclusion that different results would have been obtained in the study had a different criterion measuring forecasting accuracy been chosen. This criterion, though, according to Priestly, must be specified before one can start to consider the optimal forecast derived from a particular model. Priestley believes that Makridakis and Hibon "apparently follow the reverse procedure, i.e., they first calculate the forecast, and consider afterward how best to assess its accuracy." In any case, Priestley disagrees with the Makridakis and Hibon attempt to classify series according to which forecasting approach performs best - something that was done extensively throughout their study. As Preistley explains

> I do not believe that it is very fruitful to attempt to classify series according to which forecasting techniques perform 'best.' The performance of any particular technique when applied to a particular series depends essentially on (a) the model which the

series obeys; (b) our ability to identify and fit this model correctly and (c) the criterion chosen to measure forecasting accuracy.

I agree with Professor Priestley. Studies that attempt to classify forecasting techniques and assess forecast performance reveal mostly the degree of sophistication of the forecaster and his ability to justify applying one forecasting technique over another. This also seems to be the opinion of other researchers who participated in the discussion of Makridakis and Hibon's study. C. Chatfield, for example, believes that the reason that empirical studies sometime give different results may depend on the selected sample of time series and more likely on the skill of the analysts and on their individual interpretation of a particular forecasting method. Dr. Chatfield [see 80, p. 130] argues that "empirical studies say more about the respective analysts than they do about the methods."

So far it has been clear that irrespective of the forecasting methodologies evaluated or the particular study conducted, the reason for arriving at conflicting results can be traced to the forecaster's ability to properly set up a comparative study and use the competing forecasting techniques in their design context. Other studies whose controversial results can be traced mainly to the forecasters are the Makridakis and Wheelwright [82] for the Adaptive Filtering (AF) and the Carbone, Longini and Bretschneider [18] study for the Adaptive Estimation Procedure (AEP).

The above researchers claim that AF and AEP procedures give better forecasts than the Box-Jenkins methodology when applied to the

famous airline passenger data. (These data are shown in Table 5.2. A complete description of their findings is presented in Chapter Four for the AF and Chapter Five for the AEP procedure, respectively.) The strength of the AF and AEP, according to their respective authors, relies on the fact that the parameters of their model are adjusted every time a new observation becomes available. Let us assume that we forecast 36 months ahead. If no new observation becomes available during these 36 months and the forecast is done through bootstrapping, i.e., use the one period ahead forecast as if it were an actual value, there is no way for the AF and AEP procedures to change their parameters since no new observation became available. This being the case, it is extremely unlikely that an AF or AEP model, which yielded larger M.S.E. of fit for the historical data, can provide the most stable and accurate forecast over the 36-month time horizon. The only way the AF and AEP procedures can update their parameters is by the inclusion of a new observation in their data from a month to month basis forecasting. In the latter case, a 36-month-ahead forecast using the AF and AEP procedures will require 35 updates of the parameters of the model. Comparison, in this case, between AF, AEP and BJ, although meaningless from a theoretical point of view, would have been more appropriate and fair, if every time the AF and AEP models profited from the additional information provided by the inclusion in the training data set of a newcomer data point, the same point had been added to the training data set for the BJ model. In any case, Chatfield [see 80, p. 129] reports that Makridakis and Wheelwright made their comparisons for a single base

period. Paul Newbold [see 33, p. 234] believes that forecasts made of just one series from one base point "can be expected to say very little about the general merits of the forecasting method employed." Chatfield goes on to say that "when forecasts were made for different base periods in the forecast period, it turned out that AF was worse on average than Box-Jenkins and Holt-Winters." Incidentally, the forecast comparisons made in the study by Makridakis and Hibon [80] were also done for a single base period at time (n-12). So in the AF and AEP versus BJ controversy, again we have a comparative study which tries to compare non-comparable methodologies (automatic procedures such as AF and AEP whose parameters are continuously updated versus human intervention procedures whose model parameters are fixed) and also makes its forecast comparisions for a single base period in time. Once again, it is clearly the forecaster who must take the blame.

There are many automatic forecasting methodologies which are described by many as "black boxes" and whose intentions are to simplify the forecaster's role in the model building and forecasting process. Of course these automated procedures are as good as the human that built them. It is very unlikely that a machine can be designed with the magnitude of artificial intelligence required to perform sophisticated tasks such as identification of the proper model among the general class of models in the BJ methodologies.

This point is particularly germane since the unfortunate and unfair choice of having the BJ methodology represented by an automatic BJ procedure has already been implemented at a comparison contest in

conjunction with an ORSA/TIMS meeting meeting in Los Angeles in 1979. The competing forecasting techniques were AF, AEP, BJ, State Space and Combination of Forecasts. Allan V. Cameron and Raman K. Mehra [29] are the researchers who developed and automated the State Space methodology. State Space, they claim, is important to business and economic forecasting applications because

> It provides a proven new forecasting methodology. At the recent National ORSA/TIMS contest, it was judged as 'best indivi-ual forecasting technique.'

Even if one accepts the results of the contest at face value, all that can be said is, in this empirical study, for the particular set of time series examined, for the particular base period used, and for the specific forecast performance criterion chosen to assess the forecast performance, the State Space method seems to give, on the average, better results than the other forecasting methodologies. Mehra and Cameron, however, seem to use the ORSA/TIMS contest judgement ("best forecast technique and a proven new methodology") as justification for the very existence of their methodology. In the following paragraphs, it will be proven that the results were erroneous - mainly because of the unfit choice of an automatic procedure to represent the BJ methodology.

In the comparative study of the 1979 ORSA/TIMS Forecasting Tournament, four series were examined:

Series 1: Lay-off rates, "seasonally adjusted" monthly series 1952-1968

Series 2: Index of new business formation, monthly series 1949-1965

Series 3: Non durable inventories, 1958-1970

Series 4: Housing starts, 1959-1971.

The data was for the United States; the first two series were taken from the Business Conditions Digest data tape, the second two were unseasonally adjusted series provided by the Bureau of the Census. The data are shown in Tables 7.1, 7.2, 7.3 and 7.4, respectively.

TABLE 7.1

LAY-OFF RATE 'SEASONALLY ADJUSTED,' MONTHLY, 1952-1968

(The first 168 observations are used as training data and the last 36 observations are used to assess the model's forecasting performance.)

1.50	1.50	1.40	1.60	1.30	1.50	3.00	1.30	0.90	0.80	0.80	1.00
0.90	1.00	1.00	1.10	1.20	1.20	1.40	1.70	1.90	2.20	2.50	2.60
2.90	2.70	2.80	2.70	2.40	2.30	2.10	2.20	2.10	1.90	1.80	1.80
1.50	1.40	1.50	1.40	1.40	1.70	1.80	1.70	1.40	1.50	1.30	1.50
1.60	2.20	1.80	1.60	2.20	1.80	1.70	1.50	1.80	1.60	1.60	1.50
1.50	1.70	1.50	1.70	2.00	1.70	1.80	2.10	2.30	2.70	2.90	2.80
3.30	3.20	3.50	3.30	3.10	2.40	2.50	2.30	2.20	2.00	1.90	2.00
1.80	1.70	1.70	1.70	1.70	1.80	1.80	2.00	2.10	2.90	2.40	1.90
1.50	1.90	2.40	2.30	2.30	2.50	2.40	2.60	2.40	2.60	2.60	2.80
2.70	3.00	2.50	2.10	2.30	2.20	2.30	1.90	2.20	1.80	1.90	1.90
1.90	2.00	1.80	1.80	2.00	2.00	2.00	2.20	2.00	2.10	1.90	1.90
2.00	1.80	1.90	1.90	1.80	1.80	1.70	1.90	1.90	1.80	1.80	1.70
1.70	1.90	1.80	1.60	1.70	1.60	1.60	1.50	1.60	1.70	1.50	1.60
1.40	1.40	1.40	1.50	1.40	1.40	1.40	1.60	1.40	1.40	1.40	1.40
-	-	-	-	-	-	-	-	-	-	-	-
1.20	1.10	1.10	1.20	1.10	1.30	1.40	1.20	1.00	1.10	1.20	1.30
1.40	1.50	1.60	1.50	1.40	1.30	1.40	1.30	1.30	1.30	1.20	1.20
1.40	1.30	1.20	1.20	1.20	1.20	1.30	1.40	1.20	1.20	1.10	1.10

TABLE 7.2

INDEX OF NEW BUSINESS FORMATION, MONTHLY, 1949-1965

(The first 168 observations are used as training data and the last 36 observations are used to assess the model's forecasting performance.)

96.2	91.7	88.2	88.3	85.6	85.5	83.4	84.3	86.2	86.1	88.2	90.0
88.9	91.5	93.1	95.1	94.2	95.8	94.5	93.5	92.8	92.6	93.2	92.2
93.1	93.4	94.8	91.8	92.1	91.7	92.2	91.9	93.7	94.1	95.7	94.9
96.0	96.6	97.2	96.5	98.4	99.4	97.2	99.9	100.1	99.8	99.0	98.7
99.0	98.9	98.0	98.2	95.7	94.0	94.4	94.0	90.6	90.7	89.2	90.0
88.7	88.1	87.8	89.8	90.1	90.2	91.0	92.4	92.9	94.5	95.3	95.1
98.5	100.0	100.1	99.4	99.5	100.2	100.0	99.1	99.2	97.9	97.8	97.4
97.4	97.8	97.6	96.3	96.4	95.0	94.7	94.2	93.2	94.4	92.6	93.0
91.7	91.5	91.9	91.7	91.1	91.8	91.3	90.3	89.7	88.9	88.1	86.6
86.3	85.4	84.9	84.8	87.7	88.3	89.8	91.9	92.9	93.0	93.9	94.6
96.5	97.0	98.3	98.6	97.8	96.4	96.3	96.1	96.1	95.3	96.8	97.0
97.6	96.1	94.7	94.8	93.0	93.2	92.6	91.0	90.6	90.3	87.9	87.3
85.3	87.1	88.0	88.4	88.7	89.0	88.5	87.7	87.8	89.3	90.2	90.1
90.0	90.7	90.9	90.6	90.6	90.4	90.5	91.1	91.2	91.2	90.7	90.9
-	-	-	-	-	-	-	-	-	-	-	-
91.6	92.8	93.1	91.9	92.4	92.8	93.4	94.3	94.0	94.3	94.0	94.5
95.2	95.8	95.7	96.8	98.0	96.4	96.3	96.6	99.0	99.8	98.2	98.7
99.0	99.1	98.6	97.3	97.9	98.7	99.1	98.3	98.7	98.2	98.7	99.5

TABLE 7.3

NON-DURABLE INVENTORIES, MONTHLY, 1958-1970

(The first 120 observations are used as training data and the last 36 observations are used to assess the model's forecasting performance.)

20317	20182	20042	19850	19742	19773	19565	19632	19588	19773	19942	20120
20141	20719	20010	20038	20131	20318	20290	20400	20463	20675	20791	21002
21169	21124	20954	21029	21216	21343	21295	21420	21464	21502	21496	21559
21823	21868	21822	21957	21937	21994	21916	22062	22016	22186	22374	22545
22701	22728	22645	22565	22742	22917	22880	23031	23250	23448	23588	23735
23859	23831	23675	23583	23602	23647	23497	23609	23703	24035	24281	24343
24399	24507	24469	24408	24399	24136	24044	24089	24151	24576	24833	25048
25263	25232	25126	25044	25041	24944	24970	25066	25065	25303	25659	26061
25452	26626	26704	26815	27001	27167	27286	27309	27330	27571	27882	28207
28762	28921	29010	29190	29236	29088	29015	29003	29052	29169	29410	29798
-	-	-	-	-	-	-	-	-	-	-	-
30201	30207	30213	30370	30616	30553	30738	30913	30840	31212	31490	31823
32070	32337	32400	32882	32942	32651	32634	32852	32964	33208	33509	33758
33920	34253	34182	34540	34604	34462	34212	34097	34032	34521	34876	34976

TABLE 7.4

HOUSING STARTS, MONTHLY, 1959-1971

(The first 120 observations are used as training data and the last 36 observations are used to assess the model's forecasting performance.)

99.2	100.0	130.7	156.0	156.0	153.3	149.7	142.5	140.1	123.4	106.5	96.4
87.4	93.5	93.8	124.8	133.8	128.1	118.3	135.1	102.6	113.2	94.5	70.9
73.1	79.2	109.3	117.1	131.6	140.6	129.9	130.3	131.2	129.9	106.1	86.6
83.6	78.5	118.1	152.4	157.6	140.2	140.1	149.5	117.0	138.0	122.5	95.0
81.8	90.7	128.5	166.4	176.4	158.1	153.5	147.6	145.7	168.3	121.1	96.8
99.3	102.2	132.2	150.9	155.9	162.4	143.0	141.3	123.2	142.9	113.1	94.5
85.8	83.0	124.1	151.3	157.7	158.6	141.5	131.6	126.2	135.2	112.7	101.9
81.9	79.0	122.4	143.0	133.9	123.5	100.0	103.7	91.9	79.1	75.1	62.3
61.7	63.2	92.9	115.9	134.2	131.6	126.1	130.2	125.8	137.0	120.2	83.1
82.7	87.2	128.6	164.9	144.5	142.5	142.5	141.0	139.5	143.3	129.5	99.3
-	-	-	-	-	-	-	-	-	-	-	-
105.8	94.6	135.6	159.9	157.7	150.5	126.5	127.5	132.9	125.8	97.4	85.3
69.2	77.2	117.8	130.6	127.3	141.9	143.5	131.5	133.8	143.8	128.3	124.1
114.8	104.6	169.3	203.6	203.5	196.8	197.0	205.9	175.6	181.7	176.4	155.3

The methods that competed for the "best forecasting technique" were:

A: Box-Jenkins methodology provided by David Reilly, Automatic Forecasting Systems Incorporated;

B: AEP by Robert Carbone and Stuart Bretschneider, Ohio State University;

C: AF also by Robert Carbone and Stuart Bretschneider;

D: State Space by Alan Cameron, University of California, Irvine;

E: Combining forecasts by C. W. J. Granger, University of California, San Diego.

Table 7.5 is taken from [29, Lecture 5, p 5] and shows the mean squared errors for the various cases and the following table (Table 7.6)

TABLE 7.5

MEAN SQUARED ERROR
(ONE-STEP FORECASTS)

Method	Series 1	Series 2	Series 3 (E05)	Series 4
Box-Jenkins	.24000	.88300	.31200	345.5
AEP	.01350	.90800	.32300	215.2
AF	.01500	.87600	.30200	213.8
State Space	.01300	.80200	.24800	114.3
Combined	.01350	.84900	.31200	160.1

[29, Lecture 5, p 6] shows the rankings of the five methods using both the mean squared error (MSE) and mean absolute error (MAE) criteria.

TABLE 7.6

RANKINGS OF FORECASTING METHODS USING MSE AND MAE CRITERIA
(ONE-STEP FORECASTS)

Method	Mean Squared Error Rank By Series					Mean Absolute Error Rank By Series				
	1	2	3	4	Overall	1	2	3	4	Overall
Box-Jenkins	5	4	4	5	5	5	5	5	5	5
AEP	2	5	5	4	4	3	3	3	3	3
AF	4	3	2	3	3	4	4	4	4	4
State Space	1	1	1	1	1	1	1	1	1	1
Combined	3	2	3	2	2	2	2	2	2	2

The overall rankings are found by ranking the totals of the individual ranks for the four series. It is of interest to notice that the ranking of methods may vary depending on which performance criterion is used, i.e., depending on whether MSE or MAE is chosen to evaluate the model's performance. What is more interesting is that the BJ methodology is ranked last overall, irrespective of the error criterion. The BJ methodology continued to perform poorly for the six-step and twelve-step-ahead forecasts. Many researchers, including Stern [see 105, p. 150], Priestley [see 105, p. 152], believe that no automatic procedure can outperform one which permits intervention by a knowledgeable researcher.

When the same data were used to apply the BJ methodology in the original form proposed by Professors Box and Jenkins (i.e., a human intervention or forecaster-cooperative form) a much better performance

was achieved. The structure of the automatic BJ model and that of the forecaster-cooperative BJ model are compared below.

Series	Automatic BJ Model	Forecaster-Cooperative BJ Model
1	$(1-.45B-.35B^2)(z_t-1.65)=a_t$	$(1+.1912B)(1-B)z_t=a_t$
2	$(1-.25B^2)(1-B)z_t=a_t$	$(1-.8224B)(1-B)z_t=(1-.6341B)a_t$
3	$(1-.18B^{12})(1-B)z_t=.077+a_t$	$(1+.2164B-.3068B^{11})(1-B)(1-B^{12})z_t=(1-.7963B^{12})a_t$
4	$(1-.75B)(1-.7B^{12})\times(z_t-122)=a_t$	$(1+.4216B+.2243B^2+.3342B^{16})(1-B)\times(1-B^{12})z_t=(1-.7149B^{12})a_t$

The rankings for the one-, six-, and twelve-step-ahead forecasts for the automatic versus forecaster-cooperative BJ models are shown in Table 7.7, 7.8 and 7.9 respectively. The new rankings indicate that for the four time series examined and using the MSE as the error criterion in evaluating the models' performances, the BJ methodology achieved results as good as or better than the results of the other methodologies.

David Reilly reports that for the automated BJ methodology, "all four series were treated identically with respect to model building

TABLE 7.7

OVERALL RANKINGS BY MSE FOR ONE-STEP-AHEAD FORECAST
(36 OBSERVATIONS)

Method	Series 1	Series 2	Series 3	Series 4	New Rank	Old Rank
BJ (Cooperative)	.013	.793	.295	117.1	1	--
BJ (Automatic)	.024	.883	.312	345.5	--	5
AEP	.0135	.908	.323	215.2	5	4
AF	.015	.876	.302	213.8	4	3
State Space	.013	.802	.248	114.3	3	2
Combined	.0135	.849	.312	160.1	1	1

TABLE 7.8

OVERALL RANKINGS BY MSE FOR SIX-STEP-AHEAD FORECAST
(31 OBSERVATIONS)

Method	Series 1	Series 2	Series 3 (E06)	Series 4	New Rank	Old Rank
BJ(Cooperative)	.033	3.76	.122	549	1	--
BJ (Automatic)	.068	3.55	.232	1206	--	4
AEP	.078	4.82	.228	1010	5	5
AF	.035	3.81	186	972	4	3
State Space	.033	3.23	.172	622	2	1
Combined	.044	3.51	.107	714	3	2

TABLE 7.9

OVERALL RANKINGS BY MSE FOR TWELVE-STEP-AHEAD FORECAST
(25 FORECASTS)

Method	Series 1	Series 2	Series 3 (E06)	Series 4	New Rank	Old Rank
BJ (Cooperative)	.048	9.45	.299	1574	1	--
BJ (Automatic)	.101	10.51	.465	1683	--	3
AEP	.265	16.78	.791	2444	5	5
AF	.032	11.81	.505	1864	4	3
State Space	.048	8.89	.433	2004	3	2
Combined	.085	8.21	.174	1866	1	1

even though two were labelled non-seasonal and two were labelled seasonal. Box-Jenkins procedures allow the data to speak for themselves and, in this case, the Box-Jenkins procedures correctly selected seasonal coefficients for the seasonal series."

Apparently, through the automatic version of the BJ methodology provided by David Reilly, the data could not speak for themselves; they otherwise would have spoken out that the supposedly seasonally adjusted data of Series One, for example, still showed seasonality of period 12 and that also a logarithmic transformation was necessary for the series. Using the non-automatic version of the BJ methodology, as originally prepared by Box and Jenkins, it was found that one particular residual at time t=7 seemed to be an anomaly in the data. Indeed the seventh observation of Series One in Table 7.1 has a value of three, by far the

largest value in the sample! If an intervention model is fitted to the data with a pulse at lag seven we get the model

$$\nabla \ln Y_t = \underset{\pm.09}{-.296}\nabla \xi_t + (1-.32B^{12})a_t$$

Or if we distribute the pulse at lags six, seven, eight we get

$$\nabla \ln Y_t = (\underset{\pm.1}{.206} + \underset{\pm.1}{1.00B} + \underset{\pm.1}{.28B^2})\nabla \xi_t + (1-\underset{\pm.07}{.26B^{12}})a_t$$

The latter is a much better model for Series One since the ln transformation of the data removes the heteroscedasticity and the intervention model takes care of the anomaly of the data at time t=7.

This is exactly why it is so important for the researcher to spend a lot of time trying to find out as much about his data as he possibly can. Harrison [see 33, p. 319] claims that "to treat a data series in isolation is to confess to ignorance." Every time series is different and requires careful consideration. An automatic procedure that indiscriminantly treats every time series in the same way is bound to miss some information which may be significant to the model building process. Throughout the model building process, the researcher should consult the experts who, knowing more about the data and the particular situation, should intuitively be able to determine whether it is reasonable to expect cyclical, seasonal or stationary behavior in the data. An outlier in the residual may reveal an event which can only be explained by someone who has worked closely with the data. Even a thoroughly checked

and otherwise acceptable model should be rejected if it implies results absurd or internally inconsistent.

A good forecaster starts with his data. The forecaster is frequently not an expert in forecasting techniques applying a method to data he knows nothing about, but rather an expert in the data using mathematical methods and judgement to make a prediction [see 80, p. 120].

Professor Jenkins epitomized the constant effort required by the forecaster who strives for excellence with the following: "People should fall in love with the data, but hate their model" [69].

A researcher's failure to carefully study his data is just one cause of incorrect results and, eventually, controversy among researchers. Controversy is also caused by forecasting methodologies which have been applied in ways that are inconsistent with the assumptions used to develop the methodologies, by comparing fixed model fixed parameters methodologies with methodologies whose models have variable parameters, by forecasts which are made of just one series from one base point, by inappropriate comparison criterion, by overgeneralization of the findings of a single forecasting study, by the researcher's potential bias, and finally, controversy may also be caused due to program variation and sophistication indebted in the forecasting software.

All the above reasons for conflicting findings and results have one thing in common: the forecaster behind them. A forecaster must simply formulate the problem, define the objectives and use or develop the analytic method that believes is the most appropriate for the

situation. A forecasting technique is only as good as the forecaster who uses it. There is no such thing as a "naive forecast;" there are, however, "naive forecasters!"

CHAPTER EIGHT

CONCLUDING REMARKS

> My interest is in the future because I am going to spend the rest of my life there
>
> C.F. Kettering

As was set forth in the introduction, all comparative forecasting studies have been examined and a perspective of what has been done collectively in this area has been developed.

Brief synopses of the most currently considered and used methodologies in comparative studies of univariate techniques have been developed in a single place.

The ORSA-TIMS data have been carefully scrutinized and reanalyzed. New BJ models have been built. As a result, it was shown that a forecaster's skill is definitely an issue in comparative studies.

The validity and merit of comparative studies has been called into question. If they are going to be done, then they should have some appropriate statistical experimental design. However, D.J. Reid has indicated that

> It is not feasible to select a random sample of time series for analysis. This implies that it is not possible to make inference about the population of all time series using conventional statistical techniques. What we can do, is get just an impression of the worth of particular techniques [117].

A list of fallacies that have occurred in comparative studies has been delineated. These fallacies are the basis of the major controversies and inconsistencies that have emerged from past comparative studies.

A State Space model has been developed for the analysis of the International Airlines data. This model was used to compare forecasting results with those that Box and Jenkins produced for the same data set. The same training data and same forecast lead times were used. State Space could not compete with the originally Box-Jenkins derived model. This further study illustrates how dependent comparative studies are on the data sets selected.

It has been demonstrated that there is considerable variability in software packages put forth for the analysis of time series. Part of the difficulty that forecasters have may be due to the software they use for their analysis.

Being familiar now with the plethora of fallacies that researchers commit in their comparative studies, one might very appropriately ask the question: why do we have comparative studies? Comparative studies, the way they are done, misleadingly report rankings of forecasting methods. Perhaps, they should instead report rankings of the forecaster's sophistication behind the reported methodologies. Even in the case where competent and conscientious forecasters participate in a comparative study, the results of their findings bear little impact. As noted above, it is not feasible to select a random sample of time series for analysis. This implies that it is not possible to make

inference about the population of all time series using conventional statistical techniques. From an empirical point of view, we are not certain that, when we compare particular methodologies, the same rankings will hold when a different training data set from the same series will be used.

Researchers are attempting the impossible when they attempt to find a forecast methodology which remains best regardless of which data is analyzed.

In comparative studies, not only are the forecast methodologies being compared but also implicitly, the forecaster's abilities are being compared. The forecaster is judged because he decides how to use a methodology.

Because of the role of the subjective forecaster in forecasting, forecasting is both a science and an art. As Professor Jenkins elucidates,

> A forecaster must exercise creativity combined with solid knowledge when identification of the forecasting model is made. If. . . [a forecaster]. . . could possibly make this a mechanical process, then science would cease to exist altogether because there would be no room for creativity. In this sense, forecasting and time series modelling is no different then any other science. There are always anxious that have to be satisfied and rules that have to be obeyed, but there is also room for creative thinking.*

It is from this "creative thinking" that controversy arises: without a "solid knowledge" a forecaster cannot hope to accurately use a forecasting methodology.

*Professor Gwilym Jenkins made this statement in a conversation he and I had during the Univariate and Multivariate Box-Jenkins Seminar on November 17, 1980.

Future comparative studies should consider Multivariate approaches. If improvement is reported with the use of Multivariate as compared to Univariate methodologies, then all univariate comparative studies would be obsolete.

Forecasters also need to establish more appropriate criteria to assess forecasting performance. For example, researchers should develop a criterion with a loss function incorporated in it.

Further, researchers should attempt to set up experimental designs which would allow comparative studies to be of some practice use to forecasters.

BIBLIOGRAPHY

1. Alavi, A. "Some Multivariate Extensions of Box-Jenkins Forecasting." Dissertation, University of Lancaster, 1973.

2. Anderson, Von O. "The Elimination of Spurious Correlation Due to Position in Time or Space." Biometrika, 10 (1914-15), 269-79.

3. Batty, M. "Monitoring an Exponential Smoothing Forecasting System." Operational Research Quarterly, 20 (1969), 319-25.

4. Bartlett, M.S. "On the Theoretical Specification and Sampling Properties of Autocorrelated Time Series." Journal of the Royal Statistical Society, 8, No. 1 (1946), 27-41.

5. _____. "Some Aspects of the Time Correlation in Regard to Tests of Significance." Journal of the Royal Statistical Society, 98 (1935), 536.

6. Bartlett, M.S., and Dianandra, P.H. "Extensions of Quenouille's Test for Autoregressive Schemes." Journal of the Royal Statistical Society, B15 (1950), 107.

7. Bartlett, M.S., and Rajalkashman, D.U. "Goodness-of-Fit Tests for Simultaneous Autoregressive Series." Journal of the Royal Statistical Society, B15 (1953), 107.

8. Bates, J.M., and Granger, C.W.J. "The Combination of Forecasts." Operational Research Quarterly, 20, No. 4 (1969), 451-68.

9. Blin, J.M.; Stohr, E.A.; and Bagamery, B. "Input-Output Methods in Forecasting." Management Science, TIMS Studies, 12 (1979), 113-139.

10. Boode, H.W., and Shannon, C.E. "A Simplified Derivation of Linear Least Squares Smoothing and Prediction Theory." Proceedings of the IRE, 38 (1950), 417-25.

11. Box, George E.P., and Jenkins, Gwilym M. Time Series Analysis Forecasting and Control. San Francisco: Holdenday, 1976.

12. Box, George E.P., and Jenkins, Gwilym M. "Some Statistical Aspects of Adaptive Optimization and Control." Journal of the Royal Statistical Society, B24 (1962), 297-343.

13. Box, George E.P., and Jenkins, Gwilym M. "Some Recent Advances in Forecasting and Control I." Applied Statistics, 17 (1968), 91.

14. Box, George E.P., and Jenkins, Gwilym M. "Intervention Analysis with Applications to Economic and Environmental Problems." Journal of the American Statistical Association, 70, No. 349 (1975), 70-79.

15. Box, George E.P., and Tiao, G.C. "Comparison of Forecast and Actuality." Applied Statistics, 25, No. 3 (1976), 195-200.

16. Box, George E.P., and Tiao, G.C. "A Change in Level of a Non-Stationary Time Series." Biometrika, 52, Nos. 1 and 2 (1965), 181-92.

17. Box, George E.P., and Tiao, G.C. "A Canonical Analysis of Multiple Time Series." Biometrika, 64 (1977), 355.

18. Bretschneider, Stuart; Carbone, Robert; and Longini, Richard L. "An Adaptive Approach to Time Series Forecasting." Decision Sciences, 10 (1979), 232-44.

19. Brown, R.B. Statistical Forecasting for Inventory Control. New York: McGraw-Hill, 1959.

20. _____. Smoothing, Forecasting and Prediction of Discrete Time Series. Englewood Cliffs, N.J.: Prentice-Hall, 1962.

21. Brown, R.G., and Meyer, Richard F. "The Fundamental Theorem of Exponential Smoothing." Operational Research Quarterly, 9 (1961), 673-85.

22. Bryson, A.E., and Ho, V.C. Optimal Programming, Estimation and Control. New York: Blasdell, 1968.

23. Bucy, R.S. "Non-Linear Filtering." IEEE Transactions on Automatic Control, (1965), 198.

24. _____. "Recent Results in Linear and Non-Linear Filtering." Stochastic Problems in Control, June, 1968, 87-105.

25. Bucy, R.S., and Joseph, P.D. Filtering for Stochastic Processes with Application to Guidance. New York: Wiley, 1968.

26. Burman, J.P. "Moving Seasonal Adjustment of Economic Time Series." Journal of the Royal Statistical Society, 128A (1965), 534-58.

27. _____. "Seasonal Adjustment - A Survey." Management Sciences, TIMS Studies, 12 (1979), 45-57.

28. Cameron, Alan V., and Mehra, Raman K. "A Multidimensional Identification and Forecasting Technique Using State Space Models." ORSA/TIMS Conference, November 3, 1976.

29. Cameron, Alan V., and Mehra, Raman K. Handbook on Business and Economic Forecasting for Single and Multiple Time Series., 1980.

30. Carbone, Robert, and Longini, R. "A Feedback Model for Automated Real Estate Assessment." Management Science, 24 (1977), 241-48.

31. Carbone, Robert, and Gorr, W.L. "An Adaptive Diagnostic Model for Air Quality Management." Atmospheric Environment, 12 (1978), 1785-91.

32. Chambers, John C.; Mullick, S.K.; and Smith, D.D. "How to Choose the Right Forecasting Technique." Harvard Business Review, (July-August 1974), 45-74.

33. Chatfield, C., and Prothiero, D.L. "Box-Jenkins Seasonal Forecasting: Problems in a Case Study." Journal of the Royal Statistical Society, 136 (1973), 295-352.

34. Chow, G. Analysis and Control of Dynamic Economic Systems. New York: Wiley, 1975.

35. Chow, W.M. "Adaptive Control of the Exponential Smoothing Constant." Journal of Industrial Engineering, 16, No. 5 (1968), 314-17.

36. Crane, D.B., and Crotty, J.R. "A Two Stage Forecasting Model: Exponential Smoothing and Multiple Regression." Management Science, 13, No. 18 (1967), B501-07.

37. Christ, C.F. "Judging the Performance of Econometric Models of the U.S. Economy." International Economic Review, 11, No. 6, (1965), B119-35.

38. Cogger, K.O. "Time Series Analysis and Forecasting with an Absolute Error Criterion." Management Science, TIMS Studies, 12 (1979), 189-201.

39. _____. "The Optimality of General-Order Exponential Smoothing." Operational Research Quarterly, 22, No. 4 (1974), 858-67.

40. Carbo, V., and Pindyck, R.S. "An Econometric Approach to Forecasting Demand and Firm Behavior: Canadian Telecommunications." Management Science, TIMS Studies, 12 (1979), 95-111.

41. Dalrymple, Douglas J. "Sales Forecasting Methods and Accuracy." Business Horizons, 18 (December 1975), 69-73.

42. Daniels, Dr. "Comments on Bartlett's 'On the Theoretical Specification and Sampling Properties of Autocorrelated Time Series.'" Journal of the Royal Statistical Society, 8, No. 1 (1946), 86-90.

43. Elton, E.J., and Gruber, M.J. "Earnings Estimates and the Accuracy of Expectational Data." Management Science, (April 1972), B409-24.

44. Ewan, W.D., and Kemp, K.W. "Sampling Inspection of Continuous Processes with No Autocorrelation between Successive Results." Biometrika, 47 (1960), 239-71.

45. Fama, E. "The Behavior of Stock Market Prices." Journal of Business of the University of Chicago, January 1965.

46. Ferratt, T.W., and Mabert, V.A. [illegible] and Survey: A Description and Application of the Box-Jenkins Methodology." Decision Sciences, 3 (1972), 83-119.

47. Fildes, R., and Howell, S. "On Selecting a Forecasting Model." Management Science, TIMS Studies, 12 (1979), 297-312.

48. Frame, R.; Jedamus, P.; and Taylor, R. Statistical Analysis for Business Decisions. New York: McGraw-Hill, 1976.

49. Fronza, G., and Rinaldi, S. "An Introduction to Prediction and Filtering Problems." Conferenze de Seminario di Matematica, July 1977, 1-18.

50. Gerstenfeld, A. "Technological Forecasting." The Journal of Business, , 10-18.

51. Guerts, M.D., and Ibrahim, I.B. "Comparing the Box-Jenkins Approach with the Exponentially Smoothed Forecasting Model Application to Hawaii Tourists." Journal of Marketing Research, 12 (May 1975), 182-88.

52. Goodman, M.L. "A New Look at Higher-Order Exponential Smoothing for Forecasting." Operations Research, 22, No. 4 (1974), 880-88.

53. Granger, G.W.J. Forecasting in Business and Economics. New York: Academic Press, 1980.

54. Granger, G.W.J., and Newbold, Paul. Forecasting Economic Time Series. New York: Academic Press, 1977.

55. Granger, G.W.J., and Newbold, Paul. "Spurious Regressions in Econometrics." Journal of Econometrics, 2 (1974), 111-20.

56. Granger, G.W.J., and Newbold, Paul. "Economic Forecasting: The Atheist's Viewpoint." In Modeling the Economy. Ed. G.A. Renton. London: Heinemann, 1975.

57. Groff, Gene K. "Empirical Comparison of Models for Short Range Forecasting." Management Science, 20, No. 1 (September 1973), 22-31.

58. Gross, D., and Ray, J.L. "A General Purpose Forecast Simulator." Management Science, 11, No. 6 (April 1965), B119-35.

59. Halmer, O. "The Utility of Long-Term Forecasting." Management Science, TIMS Studies, 12 (1979), 141-47.

60. Harrison, P.J. "Exponential Smoothing and Short-Term Sales Forecasting." Management Science, 13, No. 11 (1967), 821-42.

61. _____. "Short-Term Sales Forecasting." Applied Statistics, 14, (1965), 102-39.

62. Harrison, P.J., and Davies, O.L. "The Use of Cumulative Sum (cusum) Techniques for the Control of Routine Forecasts of Product Demand." Operations Research, 12 (1964), 325-33.

63. Harrison, P.J., and Stevens, C.F. "A Bayesian Approach to Short-Term Forecasting." Operational Research Quarterly, 22, No. 4 (1971), 341-62.

64. Helmer, O.; Richard, M.; and Hansson, J.K. "An Exposition of the Box-Jenkins Transfer Function Analysis with an Application to the Advertising-Sales Relationship." Journal of Marketing Research, 14 (May 1977), 227-39.

65. Holt, C.C. Forecasting Seasonal and Trends by Exponentially Weighted Moving Averages. Pittsburg: Carnegie Institute of Technology, 1957.

65A. Jazwinski, A.H. Stochastic Processes and Filtering Theory. New York: Academic Press, 1970.

66. Jenkins, Gwilym M. Practical Experiences with Modeling and Forecasting Time Series. Jersey, J.K.: G.J.P., 1979.

67. _____. The Theory and Practical Application of Univariate and Transfer Function Analysis. Washington, D.C.: G.J.P., 1980.

68. Jenkins, Gwilym M., and Watts, D.G. Spectral Analysis and Its Applications. San Francisco: Holdenday, 1968.

69. Johnson, L.A., and Montgomery, D.C. "Forecasting with Exponential Smoothing and Related Methods." Management Science, TIMS Studies, 12 (1979), 31-44.

70. Kahneman, D., and Tuersky, A. "Intuitive Prediction: Biases and Corrective Procedures." Management Science, TIMS Studies, 12 (1979), 313-27.

71. Kalman, R.E. "A New Approach to Linear Filtering and Prediction Problems." Journal of Basic Engineering, 82 (1960), 340-48.

72. _____. "New Methods in Wiener Filtering Theory." In Proceedings of The First Symposium on Engineering Application of Random Function Theory and Probability. Ed. G.L. Bogdanoff and F. Kozin. New York: Wiley, 1963.

73. Kalman, R.E., and Bucy, R.S. "New Results in Linear Filtering and Prediction Theory." Journal of Basic Engineering, 83 (March 1961), 95-107.

74. Kirby, R.M. "A Comparison of Short and Medium Range Statistical Forecasting Methods." Management Science, 13, No. 4 (December 1966), B202-14.

75. Kushner, H.J. "On Differential Equations Satisfied by Conditional Problem Densities of Markov Processes." SIAM Journal of Control, 2 (1964), 106-119.

76. Levenbach, H.; Cleary, J.P.; and Fryk, D.A. "A Comparison of Arima and Econometric Models for Telephone Demand." Proceedings of the American Statistical Association, (1974), 448-50.

77. Levine, A.H. "Forecasting Technique." Management Accounting, (January 1967), 31-36.

78. Makridakis, Spyros. "A Survey of Time Series." International Statistical Review, 44, No. 1 (1976), 29-70.

79. Makridakis, Spyros; Hodgsdon, Anne; and Wheelwright, Steven C. "An Interactive Forecasting System." American Statistician, 28, No. 4 (November 1974), 153-58.

80. Makridakis, Spyros, and Hibon, M. "Accuracy of Forecasting: An Empirical Investigation." Journal of the Royal Statistical Society, 142A, part 2 (1979), 97-145.

81. Makridakis, Spyros, and Wheelwright, Steven C. "Forecasting: Issues and Challenges for Marketing Management." Journal of Marketing, (October 1977), 24-38.

82. Makridakis, Spyros, and Wheelwright, Steven C. "Adaptive Filtering: An Integrated Autoregressive/Moving Average Filter for Time Series Forecasting." Operational Research Quarterly, 28, No. 2 (1977), 425-37.

83. Makridakis, Spyros, and Wheelwright, Steven C. Forecasting: Methods and Applications. Santa Barbara: Wiley, 1978.

84. Makridakis, Spyros, and Wheelwright, Steven C. Forecasting Methods for Management. New York: Wiley, 1973.

85. Makridakis, Spyros, and Wheelwright, Steven C. "Forecasting: Framework and Overview." Management Science, TIMS Studies, 12 (1979), 1-15.

86. Makridakis, Spyros, and Wheelwright, Steven C. "Forecasting the Future and the Future of Forecasting." Management Science, TIMS Studies, 12 (1979), 329-52.

87. McClain, J.O., and Thomas, L.J. "Response-Variance Tradeoffs in Adaptive Forecasting." Operations Research, 21, No. 2 (March-April 1973), 554-568.

88. McKenzie, E. "An Analysis of General Exponential Smoothing." Operations Research, 23, No. 11 (1976), 131-40.

89. Mehra, R.K. "On Line Identification of Linear Dynamic Systems with Applications to Kalman Filtering." IEEE Transactions on Automatic Control, AC-16, No. 1 (February 1971), 12-21.

90. _____. "Kalman Filters and their Applications to Forecasting." Management Science, TIMS Studies, 12 (1979), 75-94.

91. _____. "On Optimal and Suboptimal Linear Smoothing." Proceedings of the National Electronics Conference, (December 1968).

92. _____. "Identification of Variances and Adaptive Kalman Filtering." IEEE Transaction of Automatic Control, AC-15, No. 12 (April 1970), 173-84.

93. Mehra, R.K., and Cameron, A.V. "State Space Forecasting for Single and Multiple Series." Meeting of ORSA, (Fall 1976).

94. Miller, R.B., and Wichern, D.W. Intermediate Business Statistics. New York: Holt, 1977.

95. Montgomery, D.C. "Adaptive Control of Exponential Smoothing Parameters by Evolutionary Operation." AIIE Transactions, 2, No. 3 (1970), 268-69.

96. Montgomery, D.C., and Johnson, L.A. Forecasting and Time Series Analysis. New York: McGraw-Hill, 1976.

97. Moriarty, M., and Adams, Arthur. "Issues in Sales Territory Modeling and Forecasting Using Box-Jenkins Analysis." Journal of Marketing Research, 16 (May 1979), 221-32.

98. Mott, C.H. "Forecast Disclosure." Management Accounting, (July 1973), 17-18.

99. Muth, J.F. "Optimal Properties of Exponentially Weighted Forecasts of Time Series with Permanent and Transitory Components." Journal of the American Statistical Association, 55, No. 290 (1960), 299-306.

100. Narasimham, G.U.L.; Castellono, V.F.; and Singpurwalla, N.D. "On the Predictive Performance of the BEA Quarterly Economic Model and a Box-Jenkins Type Arima Model." Proceedings of the American Statistical Association, (1974), 501-04.

101. Naylor, T.H., and Seaks, T.G. "Box-Jenkins Methods: An Alternative to Econometric Models." International Statistical Review, 40, No. 2 (1972), 123-37.

102. Nelson, C.R. "The Prediction Performance of the FRB-MIT-PENN Model of the U.S. Economy." American Economic Review, 62, (December 1972), 902-17.

103. _____. Applied Time Series Analysis for Managerial Forecasting. San Francisco: Holdenday, 1973).

104. Newbold, Paul. "Time Series Model Building and Forecasting: A Survey." Management Science, TIMS Studies, 12 (1979), 59-73.

105. Newbold, Paul, and Granger, C.W.J. "Experience with Forecasting Time Series and the Combination of Forecasts." Journal of the Royal Statistical Society, 137, part 2 (1974), 131-64.

106. O'Connell, R.T., and Bowerman, B.L. Forecasting and Time Series. Belmont, California: Duxbury, 1979.

107. Olkin, I.; Glesser, L.; and Derman, C. Probability Models and Applications. New York: Macmillan, 1980.

108. Pack, D.J. "Concepts, Theories and Techniques: Revealing Time Series Interrelationships." Decision Sciences, 8 (1977), 377-402.

109. _____. A Computer Program for the Analysis of Time Series Models Using the Box-Jenkins Philosophy. Colombus: Ohio State University Press, 1977.

110. Page, E.S. "On Problems in Which a Change in a Parameter Occurs at an Unknown Point." Biometrika, 44 (1957), 248-57.

111. Parzen, Emanuel. Modern Probability Theory and Its Applications. New York: Wiley, 1960.

112. _____. "Forecasting and Whitening Filter Estimation." Management Science, TIMS Studies, 12 (1979), 149-65.

113. Prasad, V.K., and Poetzel, R.W. "An Application of Canonical Analysis to Distribution Decisions." Proceedings of the American Statistical Association, (1974), 525-29.

114. Raiffa, Howard. Decision Analysis. Reading, N.J.: Adison-Wesley, 1970.

115. Raiffa, Howard, and Schlaifer, R. Applied Statistical Decision Theory. Cambridge: Massachusetts Institute of Technology, 1961.

116. Raine, J.R. "Self-Adaptive Forecasting Reconsidered." Decision Sciences, 2 (April 1971), 181-91.

117. Reid, D.J. "A Comparative Study of Time Series Prediction Techniques on Economic Data." Dissertation, University of Nottingham, 1969.

118. Rippe, R.D., and Wilkinson, M. "Forecasting Accuracy of the McGraw-Hill Anticipatory Data." Journal of the American Statistical Association, 69, No. 348 (December 1974), 849-54.

119. Roberts, S.D., and Reid, R. "The Development of a Self-Adaptive Forecasting Technique." AIIE Transactions, 1, No. 4 (1969), 314-22.

120. Ross, S. Introduction to Probability Models. New York: Academic Press, 1972.

121. Singer, R.A., and Frost, P.A. "On the Relative Performance of the Kalman and Wiener Filters." IEEE Transactions on Automatic Control, (August 1969), 390-94.

122. Spivey, W.A., and Wrobleski, W.J. "Analyzing and Forecasting Time Series." Proceedings of the American Statistical Association, (1974), 92-101.

123. Staelin, R., and Turner, R.E. "Error in Judgmental Sales Forecasts: Theory and Results." Journal of Marketing Research, 10 (February 1973), 10-16.

124. Steece, B.M., and Wood, S.D. "An Arima-Based Methodology for Forecasting in a Multi-Item Environment." Management Science, TIMS Studies, 12 (1979), 167-87.

125. Stekler, H.O. "Forecasting with Econometric Models: An Evaluation." Econometrica, 36, Nos. 3-4 (July-October 1968), 437-63.

126. Stratonovich, R.L. Conditional Markov Processes and Their Application to the Theory of Optimal Control. New York: Elsevier, 1968.

127. Swerling, P. "First Order Error Propagation in a Stagewise Smoothing Procedure for Satellite Observations." Journal of Astronautic Science, 6 (1959), 46-52.

128. Taylor, C.J. "A Simple Graphical Method of Exponential Smoothing with a Linear Trend." Operational Research Quarterly, 18, No. 1 (1967), 61-63.

129. Theil, H. Applied Economic Forecasting. Amsterdam: North Holland, 1966.

130. Theil, H.; Boot, J.C.; and Kloek, T. Operations Research and Quantitative Economics. New York: McGraw-Hill, 1965.

131. Theil, H., and Kosobud, R.F. "How Informative are Consumer Buying Intention Surveys." Review of Economics and Statistics, (February 1968), 50-59.

132. Theil, H., and Wage, S. "Some Observations on Adaptive Forecasting." Management Science, 10, No. 2 (January 1964), 198-206.

133. Thomopoulos, Nick. Applied Forecasting Methods. Englewood Cliffs, N.J.: Prentice-Hall, 1980.

134. Thompson, H.E., and Krajewski, L.J. "A Behavioral Test of Adaptive Forecasting." Decision Sciences, 3, Sec. 3 (1972), 108-19.

135. Trigg, D.W. "Monitoring a Forecasting System." Operational Research Quarterly, 15 (1964), 271-74.

136. Trigg, D.W., and Leach, A.G. "Exponential Smoothing with an Adaptive Response Rate." Operational Research Quarterly, 18, No. 1 (1967), 53-60.

137. Trimble, D.B. "Choosing a Forecasting Technique." Thesis, Washington State University, 1973.

138. Waller, A.M. "Large-Sample Estimation of Parameters for Autoregressive Processes with Moving-Average Residuals." Biometrika, 49, Nos. 1 and 2 (1962), 117-31.

139. Wheelwright, Steven C., and Clarke, D.G. "Corporate Forecasting: Promise and Reality." Harvard Business Review, (November-December 1976), 40-52.

140. Wiener, N. The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications. New York: Wiley, 1949.

141. Widrow, B.P., and Glover, J.R. "Adaptive Noise Cancelling: Principles and Applications." Proceedings of the IEEE, 63 (December 1975), 1692-1716.

142. Widrow, B.P.; Manley, P.; Griffins, L.; and Goode, B. "Adaptive Antenna Systems." Proceedings of the IEEE, 55 (December 1967), 2143-59.

143. Widrow, B.P.; McCool, J.M.; Larimore, M.G.; and Johnson, C.R. "Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter." Proceedings of the IEEE, 64 (August 1976), 1151-62.

144. Wilkinson, M.; Rippe, R.; and Morrison, D. "Industrial Market Forecasting with Anticipations Data." Management Science, 22, No. 6 (February 1976), 639-51.

145. Wilson, G.T. "The Estimation of Parameters in Multivariate Time Series Models." Journal of the Royal Statistical Society, B35 (1973), 76.

146. Winters, P.R. "Forecasting Sales by Exponentially Weighted Moving Averages." Management Science, 6 (1960), 324-42.

147. Wise, J. "The Autocorrelation Function and the Spectral Density Function." Biometrika, 42 (1955), 151-59.

148. Wold, H. A Study in the Analysis of Stationary Time Series. Uppsala, Sweden: Almquist and Wiksell, 1938.

149. Yule, G.U. "Why Do We Sometimes Get Nonsense Correlations between Time Series? A Study in Sampling and the Nature of Time Series." Journal of the Royal Statistical Society, 89, part 1 (January 1926), 1-69.