Division of Research
Graduate School of Business Administration
The University of Michigan

April 1976

BOX-JENKINS SEASONAL FORECASTING USING TRANSFORMED DATA: A CASE STUDY

Working Paper No. 129

by

Craig F. Ansley

The University of Michigan

FOR DISCUSSION PURPOSES ONLY

ABSTRACT

In this paper, sales data for a line of office equipment are analyzed by Box-Jenkins methods using the Box-Cox transformation, given by

$$y\lambda(t) = \frac{y(t)^{\lambda} - 1}{\lambda} \qquad \lambda \neq 0$$

$$y_0(t) = \ln y(t)$$

where λ is the transformation parameter. It uses likelihood equations and a numerical algorithm for their solution which have been developed by Ansley, Spivey, and Wrobleski.

A step-by-step account of the analysis and forecasting of the data is presented, together with a discussion of some of the special problems encountered. Forecasting performance of the model is shown to be superior to the best ARIMA model based on the logarithmic transformation.

Introduction

The Box-Jenkins approach to time series analysis and forecasting is becoming widely used in many business and economic applications. Examples of seasonal forecasting, however, have not appeared often in the literature. Amongst the few published examples are analyses of airline passenger data, Box and Jenkins [4, Section 9.2]; automobile registration data, Nelson [10, Chapter 7]; and a detailed case study of sales data by Chatfield and Prothero [7].

The sales data study is somewhat disturbing to users of Box-Jenkins seasonal forecasting methods--the authors were unable to obtain a satisfactory forecasting model. Box and Jenkins [5], Harrison [8], and Tunnicliffe Wilson [14] suggested that this occurred because Chatfield and Prothero had improperly transformed their data initially by taking logarithms. A more flexible family of transformations, suggested originally by Box and Cox [3], was put forward by these discussants as a more suitable approach to choosing an initial transformation, and two, [5] and [14], were able to obtain much better results in this way. The Box-Cox family of transformations is given by

$$\begin{aligned} y_\lambda(t) &= \frac{(y(t))^\lambda - 1}{\lambda} \qquad \lambda \neq 0 \\ y_0(t) &= \ln y(t) \end{aligned} \tag{1}$$

where y(t), assumed positive, is a nonstationary time series and λ is the transformation parameter.

A formal development of the likelihood function for joint estimation of the transformation parameter λ and the other parameters of a seasonal ARIMA model is given by Ansley, Spivey, and Wrobleski [1],

[2]. These authors also have developed an algorithm for approximate solution of the likelihood equations. This algorithm is easily implemented and requires only a modest modification of existing Box-Jenkins computer programs.

This paper outlines the effects of the transformation on data structure and summarizes estimation and identification procedures. It then provides a step-by-step account of the application of the transformation to the analysis and forecasting of sales data for a line of office equipment. The results demonstrate a significant improvement in forecasting performance over the log transformation approach.

Nature of the Transformation

For a fixed value of λ the transformation (1) is simply a linear transformation of the power transformation

$$y_\lambda(t) = \{y(t)\}^\lambda \ . \tag{2}$$

We can therefore use the transformation (2) to gain some insight into the effects of transformations on the data structure.[1]

The first step in the Box-Jenkins methodology is to reduce a nonstationary series to a stationary series by means of differencing operations. The differenced series should have

(i) constant mean,

(ii) constant variance,

and we often assume that

(iii) the differenced series is normal.[2]

The first property, constant mean, requires that the original series have both trend and seasonal components generated by a polynomial. The most common example is a series with a straight-line trend plus seasonal fluctuations of constant amplitude. A very simple example is given in Figure 1.

Often, however, a series has some other trend pattern, such as an exponential trend. The usual procedure in such a case is to transform the data initially; for an exponential trend, one would choose a logarithmic transformation.

Unfortunately, a logarithmic transformation can "overtransform" the data. In Figure 2 we show a series with an increasing trend and its logarithms. Note that the original series has a trend that could easily be mistaken for an exponential trend and has seasonal fluctuations of increasing magnitude. The logarithmic transformation, on the other hand, has a decreasing trend, with decreasing seasonal amplitude. This is an example of overtransformation.

Overtransformation can and does arise in practice. For example, the data plotted by Chatfield and Prothero [7, p. 298] shows an increasing trend with increasing seasonal amplitude, while the logarithms of the data show a decreasing trend with decreasing seasonal amplitude. With the benefit of hindsight, one could point to this as the source of their problems in developing a forecasting model based on a logarithmic transformation.

The series in Figure 2 was generated by raising that in Figure 1 to the power of 2.5. If we were to transform the data back using $\lambda = .4$ in equation (2), we would obtain the original series. A constant mean could then be obtained by simple differencing operations.

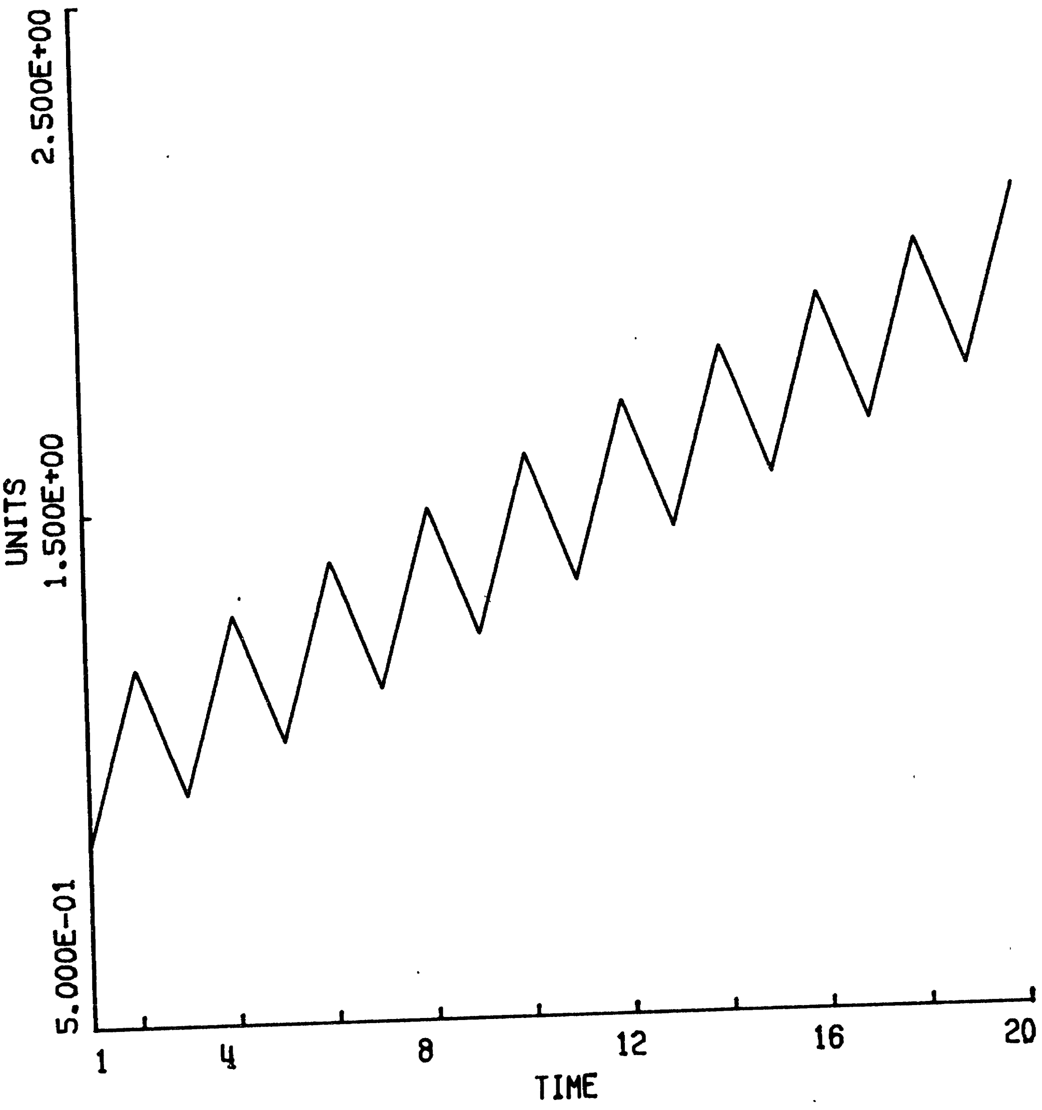

Fig. 1. Series with a linear trend and stable seasonal component.

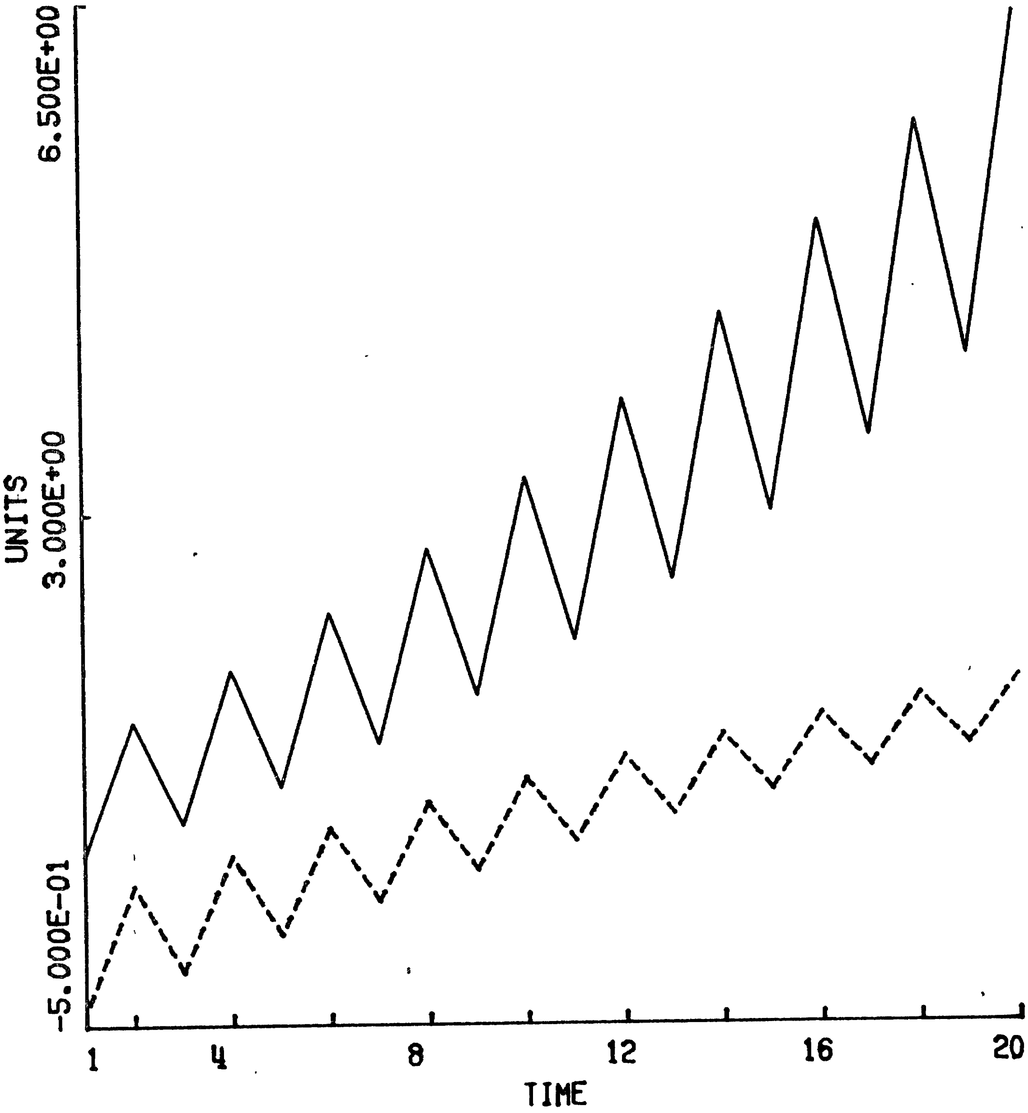

Fig. 2. . Example of overtransformation by logarithms.

—— Original series

---- Logarithms

The second requirement of the differenced series is that it have constant variance. With real data, we often take one difference and one seasonal difference. These two operations remove parabolic trends, and given the random fluctuations in real data, it can be very difficult to detect a varying mean in the differenced series. If the data are improperly transformed originally, however, the variance of the differenced series will not be constant.

Finally, the choice of the initial transformation will affect the distribution of the individual values of the series. A carefully chosen initial transformation can lead to a more nearly normal distribution for the differenced series.

It is clear that a family of transformations indexed by a single parameter cannot simultaneously satisfy (i), (ii), and (iii) in every case. The maximum likelihood procedures outlined below in effect provide a rational procedure for weighting the importance of these requirements in model estimation.

Outline of the Transformation Estimation Procedure

We assume that (1) we have observations y(t) from a time series and that (2) for some value λ of the transformation parameter, the transformed observations follow a seasonal ARIMA(p,d,q)·(P,D,Q) process. Using the notation of Box and Jenkins [4, p. 205] we have

$$\phi(B)\Phi(B^S)\nabla^d\nabla_S^D y_\lambda(t) = \delta + \theta(B)\textcircled{H}(B^S)a(t) \qquad (3)$$

where B is the backshift operator (see Box and Jenkins [4, p. 8]), and where

s = length of seasonal cycle;

d = degree of differencing;

D = degree of seasonal differencing;

δ = constant term;

$$\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p;$$

$$\theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q;$$

$$\Phi(B^s) = 1 - \Phi_1 B^s - \dots - \Phi_P B^{sP};$$

$$\textcircled{H}(B^s) = 1 - \textcircled{H}_1 B^s - \dots - \textcircled{H}_Q B^{sQ} .$$

We write $\underline{\phi}$ to represent the vector $\phi_1,\dots,\phi_p$ of autoregressive parameters and similarly $\underline{\theta}$, $\underline{\Phi}$ and $\underline{\textcircled{H}}$ to represent the vectors of moving average, seasonal autoregressive, and seasonal moving average parameters, respectively.

Suppose we observe n+d+sD values of the time series $y(-d-sD+1),\dots,y(0),y(1),\dots,y(n)$. Assuming that the first d+sD values $y(-d-sD+1),\dots,y(-1),y(0)$ are fixed, it has been shown by Ansley, Spivey, and Wrobleski that the log likelihood is given by

$$L = \text{const.} - \frac{n}{2} \ln \sigma^2 + \frac{1}{2} \ln |M_n| - \frac{S}{2\sigma^2} + \ln J \qquad (4)$$

where

$$\sigma^2 = \text{variance of } a(t)$$

$$\sigma^2 M_n^{-1} = \text{covariance matrix of } w_\lambda(1),\dots,w_\lambda(n)$$

$$J = \prod_{t=1}^{n} \{y(t)\}^{\lambda-1}$$

$$S = \sum_{-\infty}^{n} [a(t)]^2$$

and where

$$w_\lambda(t) = \nabla^d \nabla_s^D y_\lambda(t)$$

and

$$[a(t)] = E[a(t)|w_\lambda(1),\ldots,w_\lambda(n)] \ .$$

We can shed some light on this complicated expression by the following comments. First, the likelihood function is conditional on the first d+sD observations because the two differencing operations effectively reduce the data set by this number. For example, if we had 72 observations on monthly data (s=12) and we took one backward difference and one seasonal backward difference to reduce the data to stationarity, the differenced series would contain 59 = 72 - 12 - 1 differences.

Second, the formula for J, which is, in fact the Jacobian for the transformation (1), excludes the first d+sD observations. For the example above, we would calculate J using only the last 59 observations.

The expected values [a(t)] appearing in the sum of squares S are calculated from the difference equations

$$\phi(B)\Phi(B^s)[w_\lambda(t)] = \delta + \theta(B)\Theta(B^s)[a(t)] \quad . \tag{5}$$

An algorithm for the numerical solution of these equations is given by Box and Jenkins [4, Section 7.2].

Maximization of L is simplified by a numerical algorithm developed by Ansley, Spivey, and Wrobleski [1], [2]. They show that maximum likelihood estimates can be obtained by minimizing

$$S_z = \sum_{-\infty}^{n} \{\hat{a}_z(t)\}^2 \tag{6}$$

where $\hat{a}_z(t)$ is obtained from the difference equations

$$\phi(B)\Phi(B^s)z_\lambda(t) = \delta_z + \theta(B)\text{\textcircled{H}}(B^s)\hat{a}_z(t) \ , \qquad (7)$$

where

$$z_\lambda(t) = w_\lambda(t)/J^{1/n} \ ,$$

and

$$\delta_z = \delta/J^{1/n} \ .$$

Because the equations (7) are in the same form as equation (5), the same algorithm can be used for their solution. In these equations δ is replaced by δ_z, and after S_z is minimized, an estimate of δ is obtained from the value $\hat{\delta}_z$ by

$$\hat{\delta} = \hat{\delta}_z J^{1/n}(\hat{\lambda}) \ .$$

Any computer program for Box-Jenkins estimation will contain:

(i) an algorithm for solving difference equations such as (7), and

(ii) a nonlinear least-squares algorithm that can be used to minimize the sum of squares S_z in equation (6).

This method for estimating the parameter λ simultaneously with the other ARMA model parameters can therefore be incorporated into an existing computer package with minimum modification. A flow chart for the major steps involved is given in Figure 3. The main difference between the standard and modified estimation procedures is that in estimating λ, the series must be transformed and differenced prior to each evaluation of the sum of squares within the nonlinear least-squares algorithm.

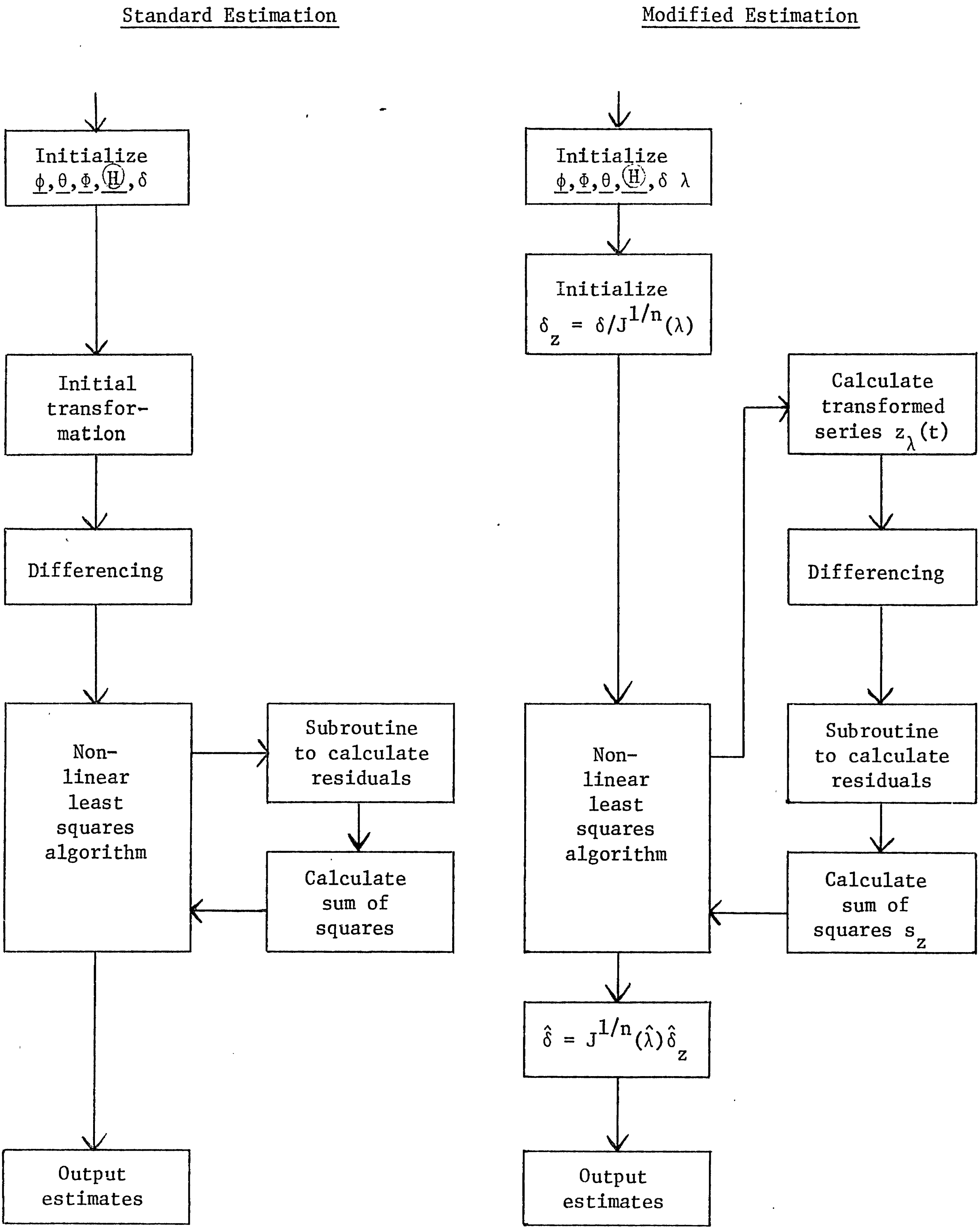
Standard Estimation
Modified Estimation
Initialize $\underline{\phi},\underline{\theta},\underline{\Phi},\underline{\textcircled{H}},\delta$
Initial transformation
Differencing
Non-linear least squares algorithm
Subroutine to calculate residuals
Calculate sum of squares
Output estimates
Initialize $\underline{\phi},\underline{\Phi},\underline{\theta},\underline{\textcircled{H}},\delta\ \lambda$
Initialize $\delta_z = \delta/J^{1/n}(\lambda)$
Calculate transformed series $z_\lambda(t)$
Differencing
Non-linear least squares algorithm
Subroutine to calculate residuals
Calculate sum of squares s_z
$\hat{\delta} = J^{1/n}(\hat{\lambda})\hat{\delta}_z$
Output estimates

Fig. 3. Flowchart of standard and modified estimation routines.

The Box-Jenkins program employed here uses the Marquardt non-linear least-squares algorithm [9]. The program was adapted as outlined above to minimize S_z simultaneously over all the parameters including the transformation parameter λ, and it has proven efficient over a wide range of ARIMA models. On the University of Michigan's Amdahl 470 computer, the modification requires approximately 10 percent more CPU time than does standard estimation.

Model Identification

The estimation procedures described above assume that the order of the ARIMA(p,d,q)·(P,D,Q) model is known. A problem arises immediately in that the autocorrelation function used in identification will be a function of the initial transformation parameter and will change as the parameter is changed.

Experience has shown the following strategy to be successful:

(i) Choose an initial transformation which seems reasonable from an examination of the raw data.

(ii) Using this initial transformation, carry out the usual Box-Jenkins identification procedure to find p,d,q,P,D,Q and initial estimates of the parameters.

This simple solution needs some justification. Experience with a large number of business and economic time series has shown that the identification of (p,d,q) and (P,D,Q) is not affected by the choice of the transformation parameter λ over a wide range of values. Moreover, initial estimates of other model parameters are relatively insensitive to the choice of λ. However, as λ is a scale parameter, any constant term will be sensitive to the choice of λ, and the initial choice of this constant term must be compatible with the initial choice of λ.

In addition, the likelihood function has been found to be well behaved and unimodal over a wide range of ARIMA model configurations. This suggests that optimization algorithms should converge efficiently from any reasonable initial value of λ.

Box and Jenkins [5] suggest a heuristic method for rapid approximate evaluation of λ. The value thus obtained could be used here to provide an initial value for λ, but we have found that the additional computer time and effort are not worthwhile.

The Data and an Initial Transformation

Data were obtained for monthly sales of a line of office equipment January 1969 through December 1975. The data are given in Table 1 and plotted in Figure 4. For the sake of example, the last twelve data points, plotted with a dashed line in Figure 4, were omitted from the identification and estimation phases of the analysis so that they could be used to test forecasting performance.

The series has a marked downward trend resembling an exponential decay. There is also a seasonal pattern with decreasing amplitude. These two factors suggest that a reasonable initial transformation would be the logarithmic transformation, i.e., $\lambda=0$ in equation (1).

The logarithms of the data, plotted in Figure 5, show a downward trend that is more nearly linear than the original series, although some appearance of decay remains. This suggests that the final transformation has a parameter $\lambda<0$. The amplitude of seasonal fluctuations is more nearly constant over the series, although there is some decrease, again suggesting that the final transformation parameter may be negative. Although it appeared that a logarithmic transformation may, in this case, be an undertransformation, it was an adequate initial choice.

TABLE 1

SALES OF OFFICE EQUIPMENT JANUARY 1969--DECEMBER 1975

	1969	1970	1971	1972	1973	1974	1975
Jan	1564	1061	857	757	691	636	553
Feb	1586	1116	917	823	746	652	586
Mar	1475	1083	890	779	724	636	564
Apr	1459	1006	846	768	702	619	553
May	1343	973	823	741	691	591	519
June	1221	934	796	702	647	553	497
July	1155	890	741	652	613	531	481
Aug	1111	868	729	641	591	508	470
Sept	1155	906	790	663	624	536	492
Oct	1105	879	774	691	624	547	492
Nov	1083	868	774	685	613	531	481
Dec	1050	873	713	663	569	503	453

Differencing

Let the observed series at time t be $y(t)$, and the initial transformed series be $y_0(y) = \ln y(t)$. The first step is to reduce the transformed series to approximate stationarity by differencing operations. As the series $y_0(t)$ has both a trend and a seasonal pattern, plots and autocorrelations of the series $\nabla^d\nabla_{12}^D y_0(t)$ were examined for various integer values of d and D (see Box and Jenkins [4, Chapter 9]). It has been shown (see Appendix to Chatfield and Prothero [7]) that the operator $\nabla\nabla_{12}$ will eliminate both a linear trend and a stable seasonal pattern.

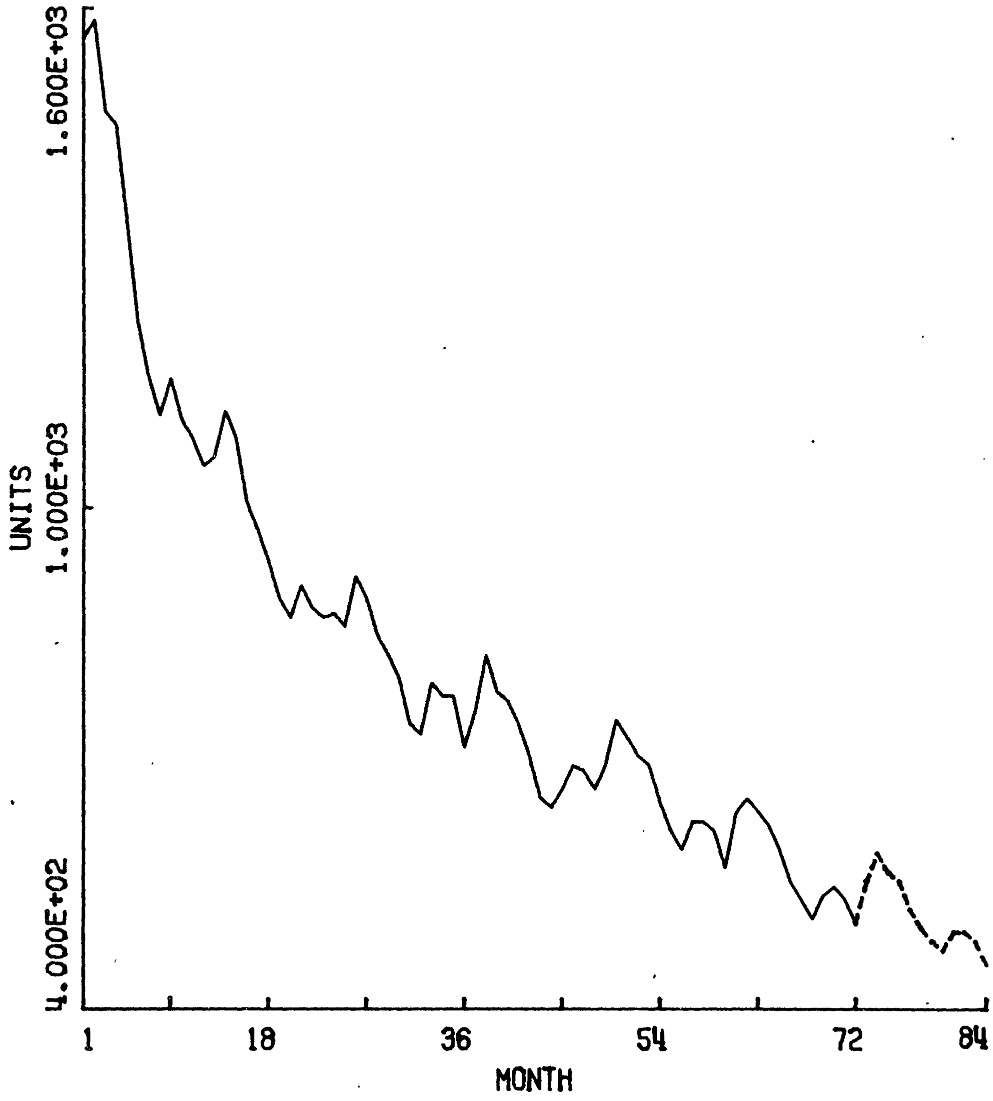

Fig. 4. Sales of office equipment from January 1969 to December 1975.

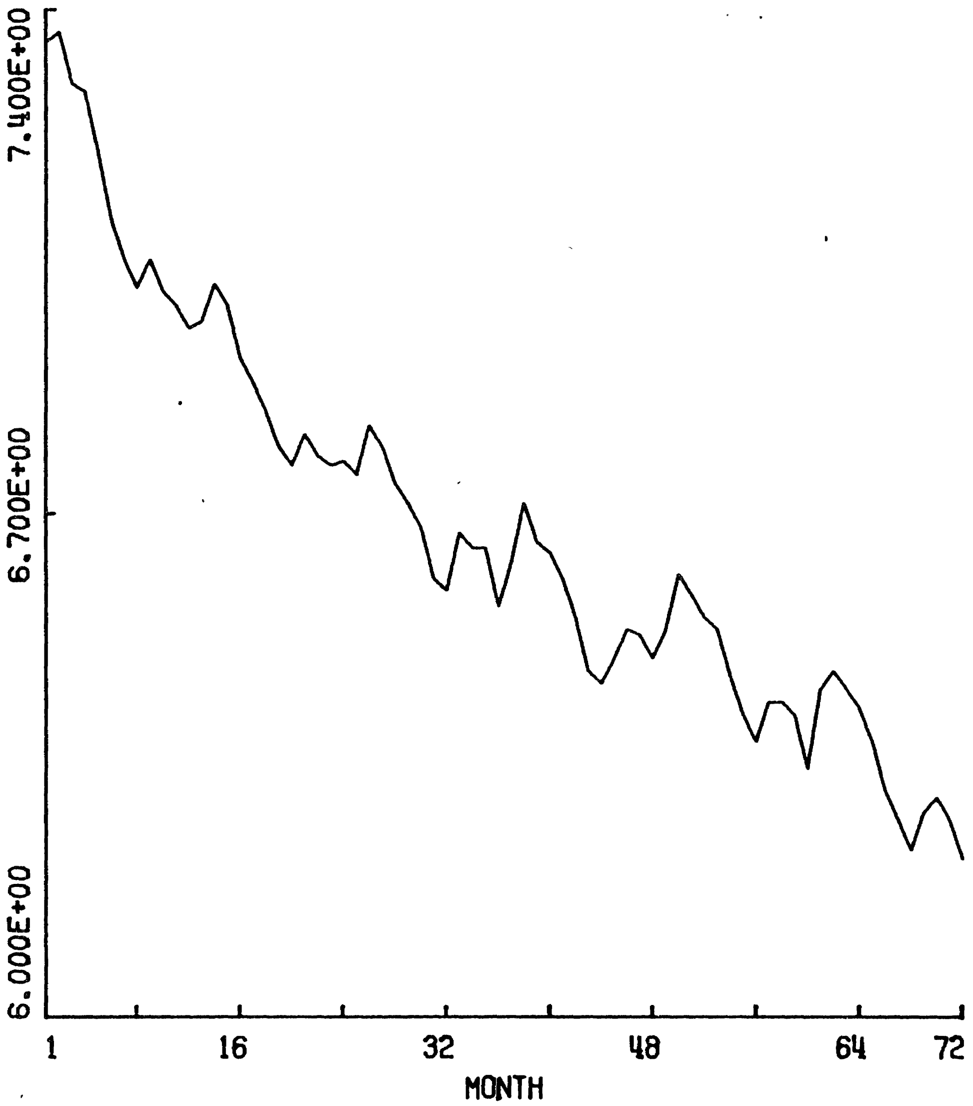

Fig. 5. Logarithms of sales data for January 1969 to December 1974.

The sample autocorrelations for the series $\{y_0(t)\}$,$\{\nabla y_0(t)\}$, $\{\nabla_{12}y_0(t)\}$, and $\{\nabla\nabla_{12}y_0(t)\}$ are given in Table 2. Note that the series contain 72,71,60, and 59 terms, respectively.

For both $\{y_0(t)\}$ and $\{\nabla_{12}y_o(t)\}$ the autocorrelations begin large and positive then fall steadily to large negative values. Clearly, further differencing is required--a necessity that also can be easily verified in this case from plots of the series. The autocorrelations of $\{\nabla y_0(t)\}$ have a seasonal cycle with peaks at the lags of 12,24, and 36 and troughs at 6,18, and 30. Examination of the plot of $\{\nabla y_0(t)\}$ shows a marked seasonal pattern but little trend. These findings strongly indicate that a seasonal difference is required. The autocorrelations of $\{\nabla\nabla_{12}y_0(t)\}$ diminish quickly after lag 12 and there is no marked trend or seasonal pattern. It is thus reasonable to assume that $\{\nabla\nabla_{12}y_0(t)\}$ is approximately stationary. The plot in Figure 6 shows no obvious trend or seasonality, confirming this choice of differencing operator. Note, however, that the variance of the series tends to decrease over time, suggesting once again that a negative value of the transformation parameter may ultimately be required.

Identification of the Differenced Series

We have obtained an approximately stationary series $w_0(t)$ given by

$$w_0(t) = \nabla\nabla_{12}y_0(t) \quad .$$

Next we must find a model of the ARMA class which will adequately describe the autocorrelation structure of $w_0(t)$. The class of seasonal ARMA models can be written

TABLE 2

SAMPLE AUTOCORRELATION OF VARIOUS DIFFERENCES OF $\{y_o(t)\}$

Series	Lags	Autocorrelations											
$y_o(t)$	1-12	.93	.85	.78	.72	.65	.60	.57	.54	.51	.49	.47	.45
	13-24	.40	.35	.30	.26	.22	.19	.17	.15	.13	.12	.12	.11
	25-36	.07	.03	.00	-.04	-.07	-.10	-.11	-.11	-.13	-.14	-.14	-.14
$\nabla y_o(t)$	1-12	.12	-.16	-.13	.00	-.03	-.22	-.05	.05	-.07	-.25	.18	.62
	13-24	.15	-.12	-.12	-.03	.00	-.24	-.02	.06	-.13	-.23	.14	.50
	25-36	.07	-.04	-.19	-.02	.02	-.21	.02	.00	-.09	-.19	.08	.29
$\nabla_{12}y_o(t)$	1-12	.84	.77	.71	.61	.53	.49	.42	.38	.34	.28	.24	.19
	13-24	.18	.15	.12	.07	.04	.02	-.02	-.05	-.09	-.10	-.09	-.16
	25-36	-.20	-.19	-.21	-.20	-.19	-.22	-.23	-.26	-.30	-.32	-.35	-.34
$\nabla\nabla_{12}y_o(t)$	1-12	-.43	.06	.03	.05	.01	.12	-.04	-.11	.25	-.18	.23	-.36
	13-24	.22	-.06	.13	-.04	-.05	.00	-.08	.27	-.23	.01	.05	.14
	25-36	-.26	.13	-.05	-.06	.16	-.14	.04	-.09	.13	-.06	.03	-.09

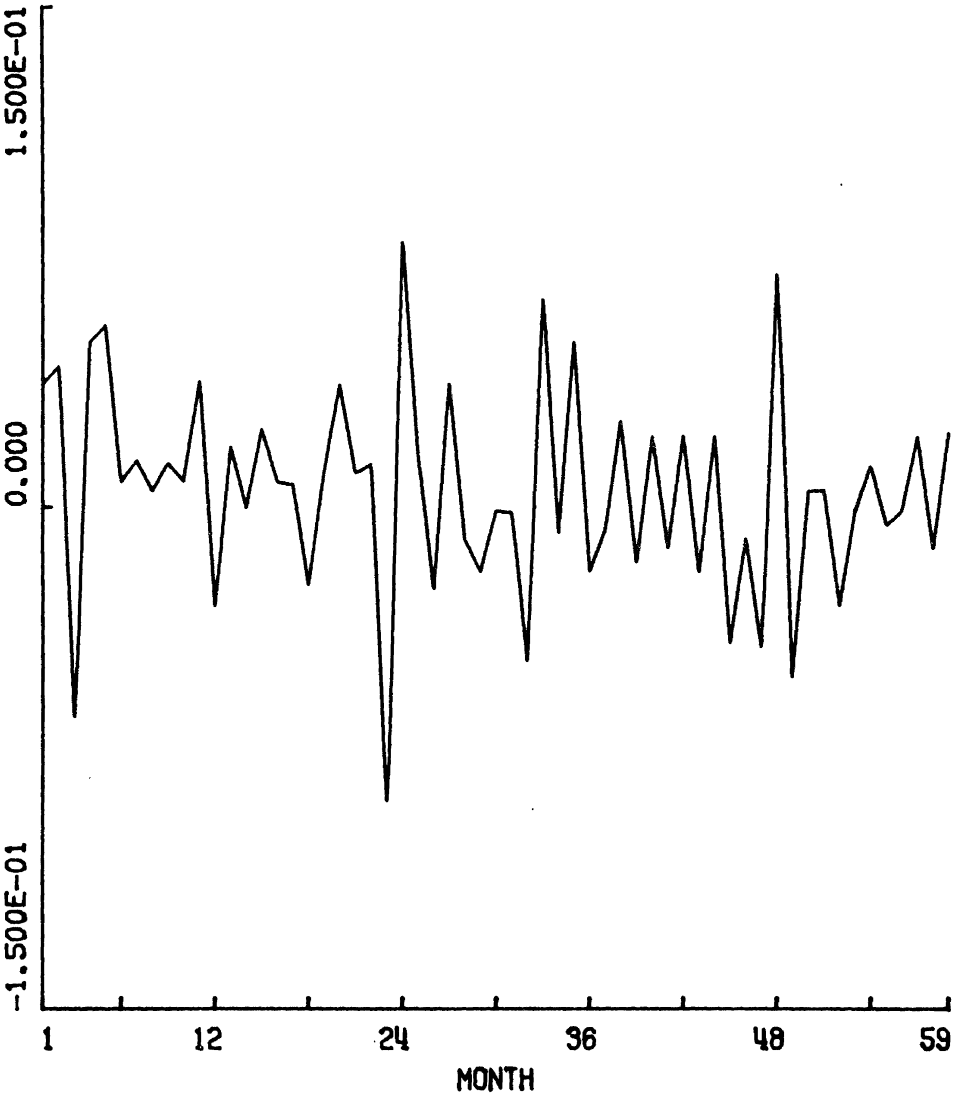

Fig. 6. Differences $\{\nabla\nabla_{12} \ln y(t)\}$.

$$\phi(B)\Phi(B^{12})w_0(t) = \delta + \theta(B)\,Ⓗ\,(B^{12})a(t)$$

as in equation (3) in which the polynomials $\phi(B),\theta(B),\Phi(B^{12})$ and Ⓗ (B^{12}) have degrees p,q,P, and Q, respectively. To find suitable values of p,q,P, and Q, we begin by examining the sample autocorrelations and partial autocorrelations. These statistics and their approximate standard errors are given in Table 3. (The autocorrelations are, of course, the same as those given previously in Table 2.)

Approximate standard errors for the sample autocorrelations r_k assuming zero autocorrelations at lags $\geq$ k were obtained from the expression

$$\hat{\sigma}(r_k) = \frac{1}{\sqrt{n}}\{1+2(r_1^2+....+r_{k-1}^2)\}^{1/2}$$

(see Box and Jenkins [4, p. 177]). The partial autocorrelations are calculated from the autocorrelations r_k via the Yule-Walker equations [4, p. 55]. The standard error for the estimated partial autocorrelation $\hat{\phi}_{kk}$ at lag k given that the underlying true partial autocorrelations are zero at lags $k \geq 0$, is approximately $1/\sqrt{n}$ [4, p. 178].

From Table 3 we see that the only significant (i.e., greater than two standard errors) autocorrelations are at lags 1 and 12. At lag 1 the autocorrelation r_1 is large and negative, followed by a number of autocorrelations close to zero. The first partial autocorrelation $\hat{\phi}_{11}$ is large and negative followed by values which decay more slowly. This suggests that the series $\{w_0(t)\}$ is of the form

$$w_0(t) = a_t - \theta a_{t-1} + \text{other terms} . \qquad (8)$$

Turning now to lag 12, we see that there is a large, negative autocorrelation, $r_{12} = -.36$, flanked by positive autocorrelations,

TABLE 3

AUTOCORRELATIONS AND PARTIAL AUTOCORRELATIONS OF $\{w_o(t)\}$

Lag	Autocorrelations						Approximate Standard Error
1-6	-.43	.06	.03	.05	.01	.12	.13
7-12	-.04	-.11	.25	-.18	.23	-.36	.15
13-18	.22	-.06	.13	-.04	-.05	.00	.18
19-24	-.08	.27	-.23	.01	.05	.14	.19
25-30	-.26	.13	-.05	-.06	.16	-.14	.20
31-36	.04	-.09	.13	-.06	.03	-.09	.21
	Partial Autocorrelations						
1-6	-.43	-.15	-.01	.08	.08	.20	.13
7-12	.12	-.11	.15	-.05	.18	-.32	.13
13-18	-.07	-.06	.14	.14	.03	.01	.13
19-24	-.16	.14	.04	-.18	.10	.04	.13
25-30	-.13	-.12	.06	.06	.01	-.05	.13
31-36	-.03	.01	.04	.05	.02	-.04	.13

$r_{11} = .23$ and $r_{13} = .22$. There is no significant autocorrelation at lags 24 or 36 suggesting that a single seasonal MA parameter could adequately explain the observed autocorrelation at lag 12. The following tentative model was chosen for $w_0(t)$

$$w_0(t) = \delta + (1-\theta B)(1-\Theta B^{12})a(t) \quad . \tag{9}$$

If r^2 = VAR(a(t) the autocovariances are given by

$$\begin{aligned}
\gamma_0 &= (1+\theta^2)(1+\Theta^2)\sigma^2, \\
\gamma_1 &= -\theta(1+\Theta^2)\sigma^2, \\
\gamma_{11} &= \theta\Theta\sigma^2, \\
\gamma_{12} &= -\Theta(1+\theta^2)\sigma^2, \\
\gamma_{13} &= \theta\Theta\sigma^2
\end{aligned}$$

[4, p. 329]. Note that for such a model $\gamma_{11} = \gamma_{13} \neq 0$; this is consistent with the observed sample autocorrelations. Now it can be seen that

$$\begin{aligned}
\rho_1 &= -\theta/(1+\theta^2) \\
\rho_{12} &= -\Theta/(1+\Theta^2) \quad .
\end{aligned}$$

Thus, we can obtain initial estimates of θ and Θ by solving the equations

$$-\frac{\theta}{1+\theta^2} = r_1 = -.43$$

$$-\frac{\Theta}{1+\Theta^2} = r_{12} = -.36 \quad .$$

Discarding roots having an absolute value exceeding 1, we obtain $\theta_0 = .57$ and $\Theta_0 = .43$.

The series $\{w_0(t)\}$ has 59 points and ranges from -.088 through .078. Its mean is .0045. This value is sufficiently large relative to the range to justify including a constant term in the model for a first estimation run. For the model (9) $Ew_0(t) = \delta$, so that an appropriate initial estimate for the constant term is $\delta_0 = .0045$.

To summarize, thus far we have tentatively identified the series as an ARIMA(0,1,1)·(0,1,1) model with initial values $\lambda_0 = 0$, $\theta_0 = .57$, $\textcircled{H}_0 = .43$, and $\delta_0 = .0045$.

Finally, it is interesting to compare Table 3 with the sample autocorrelations and partial autocorrelations for the differences $(\nabla\nabla_{12})$ for the raw series, as shown in Table 4. It can be seen that there is very little variation between the statistics for the logarithms $(\lambda=0)$ and the raw data $(\lambda=1)$, especially at lags where values are significant. This confirms our earlier comment that autocorrelations and partial autocorrelations are insensitive to the initial choice of the transformation parameter λ.

Estimation

Approximate maximum likelihood estimates were obtained using the least-squares algorithm described earlier. An important consideration here is that most nonlinear least-squares computer packages will not accept zero initial values; our Marquardt routine is no exception. It was therefore necessary to perturb the initial value of $\lambda = 0$. Because we suspected that the final value of λ would be negative, an initial value of $\lambda_0 = -.05$ was chosen. Note that this perturbation must be small because the initial estimate of the constant term was chosen to

TABLE 4

AUTOCORRELATIONS AND PARTIAL AUTOCORRELATIONS OF $\nabla\nabla_{12}y(t)$

(Raw Data)

Lag	Autocorrelations						Approximate Standard Error
1-6	-.26	.08	.13	.08	.11	.10	.13
7-12	.00	-.13	.23	-.06	.17	-.25	.14
13-18	.17	-.03	.11	-.03	-.03	-.01	.16
19-24	-.07	.27	-.26	.05	.04	.03	.17
25-30	-.15	.08	-.09	-.06	.15	-.15	.18
31-36	.07	-.14	.13	-.05	-.04	-.05	.19
	Partial Autocorrelations						
1-6	-.26	.02	.17	.16	.17	.15	.13
7-12	.01	-.24	.06	.00	.21	-.20	.13
13-18	.06	-.04	.13	-.03	.02	-.10	.13
19-24	-.12	.15	-.04	-.05	.12	-.01	.13
25-30	-.16	-.08	.03	-.01	.08	-.03	.13
31-36	.05	-.06	.00	.03	-.06	-.04	.13

be consistent with $\lambda = 0$. A large perturbation could make the initial values of δ and λ incompatible and possibly cause the least squares routine to diverge.

The estimated values and their standard errors[3] are shown in Table 5.

TABLE 5

PARAMETER ESTIMATES AND STANDARD ERRORS

		Standard Errors	
Parameter	Estimate	Unconditional	Conditional
θ	.423	.117	.114
$\textcircled{H}$	.891	.054	.054
δ	.663E-3	.117E-2	.232E-3
λ	-.212	.219	
σ^2	.351E-4		

There are two ways of viewing the distributions of the estimators $\hat{\theta}$, $\hat{\textcircled{H}}$, and $\hat{\delta}$. We can consider their distributions conditional on a fixed value of λ. Or we can admit the possibility of random variation in $\hat{\lambda}$ and thus allow additional random variation in $\hat{\theta}$, $\hat{\textcircled{H}}$, and $\hat{\delta}$.

The procedure we are using is designed to choose the best metric, or transformation, for our analysis. Given that we have selected $\lambda = -.212$, we should examine the adequacy of our model for the data thus transformed. In other words, we should examine the estimates using statistics conditioned on our choice of λ. From Table 5 we see that from this point of view, each of the estimates $\hat{\theta}$, $\hat{\textcircled{H}}$, and $\hat{\delta}$ is more than two standard errors away from zero and thus can be considered significant.

It is worth noting that if we had first transformed our data using $\lambda = .212$ and then estimated the other parameters using standard Box-Jenkins techniques, we would have arrived at identical estimates with the same estimated standard errors as the conditional standard errors in Table 5.

We can reconcile the two points of view by examining the correlation matrix[4] of the estimators $\hat{\theta}$, $\hat{\text{Ⓗ}}$, $\hat{\delta}$, and $\hat{\lambda}$ given in Table 6. This matrix allows random variation in all of the estimators jointly and therefore corresponds to the unconditional case above.

TABLE 6

ESTIMATED CORRELATION MATRIX

	$\underline{\theta}$	$\underline{\text{Ⓗ}}$	$\underline{\delta}$	$\underline{\lambda}$
θ	1.00			
Ⓗ	-.11	1.00		
δ	-.20	.12	1.00	
λ	-.22	.10	.98	1.00

There is very little correlation between either $\hat{\theta}$ and $\hat{\lambda}$ or $\hat{\text{Ⓗ}}$ and $\hat{\lambda}$, implying that the standard deviations of $\hat{\theta}$ and $\hat{\text{Ⓗ}}$ will be little affected by conditioning on λ. This confirms the figures in Table 5. These results are also consistent with our observation that the autocorrelations and partial autocorrelations, which are functions of θ and Ⓗ, are insensitive to the transformation parameter λ.

On the other hand, there is extremely high correlation between $\hat{\lambda}$ and $\hat{\delta}$. As pointed out earlier, this is expected because λ is a scale parameter directly affecting the measurement units of $\hat{\delta}$. Consequently, if we allow λ to vary, we must also expect $\hat{\delta}$ to vary, not only because of the random errors in the model, but also because its units of measurement are changing. Thus the large, unconditional standard error for $\hat{\delta}$ (shown in Table 5) also is expected. This standard error is of little use in determining the significance of $\hat{\delta}$; however it can be interpreted only as a measure of the effect of changes in λ on the units of measurement of the transformed data.

We also see in Table 5 that the transformation parameter $\hat{\lambda}$ is only one standard error away from zero and could, therefore, be regarded as insignificant. But our purpose is not to test the hypothesis that the logarithm is the correct transformation: it is to find the transformation that best explains the observed data. We therefore prefer $\lambda = -.212$ to a logarithmic (λ=0) transformation and expect to obtain improved forecasting performance through use of this transformation. The only relevance of the standard error of $\hat{\lambda}$ is that perhaps the forecasting improvement might be expected to be smaller than if $\hat{\lambda}$ were, say, two standard errors removed from zero.

Finally, it is useful to examine the nature of the likelihood surface. In Figure 7 we have plotted the variation of the likelihood L with θ and λ. (In this instance L is maximized over Ⓗ and δ.) Note that the surface is regular and unimodal, with nearly elliptical contours, especially near the maximum, supporting our assertion that standard, nonlinear least-squares algorithms will be efficient in this application.

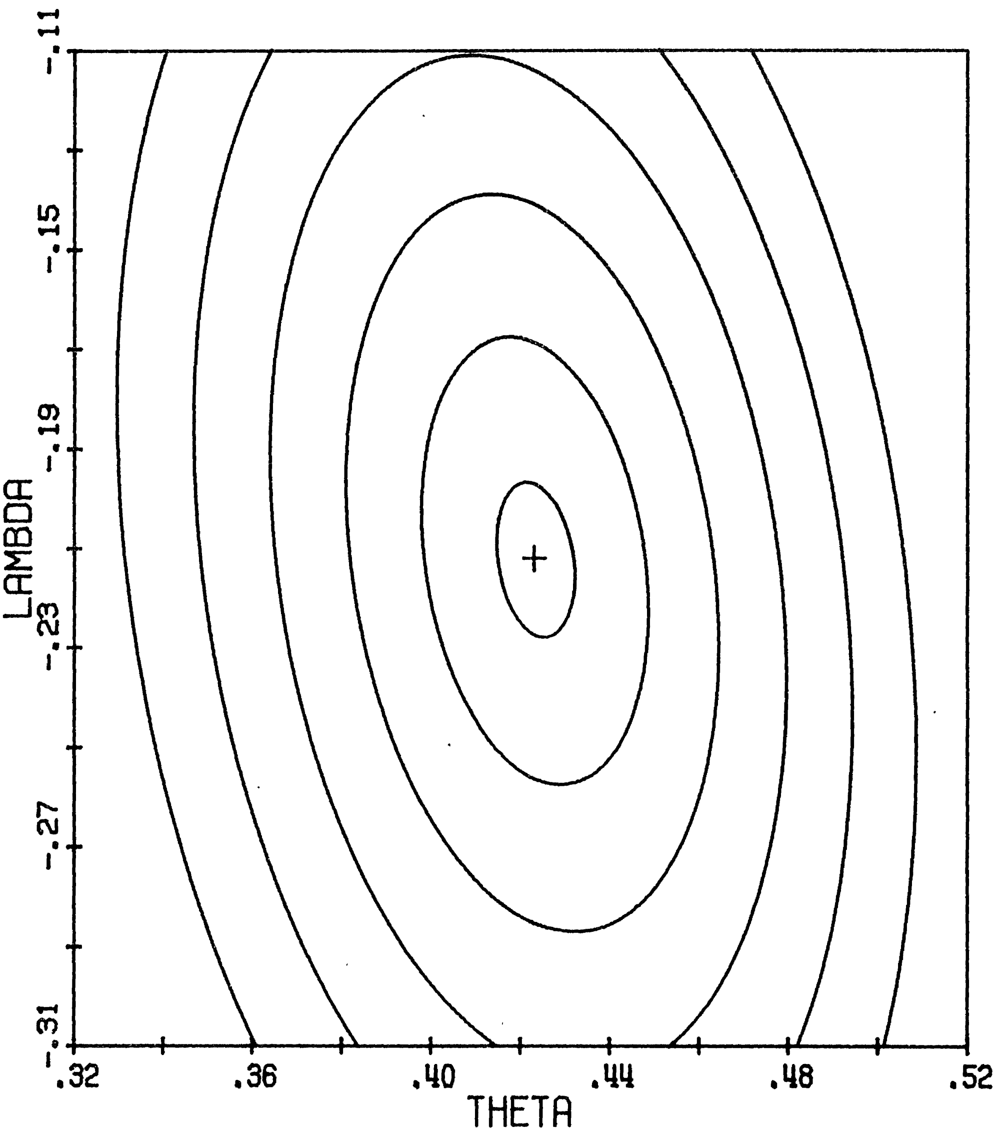

Fig. 7. Log likelihood contours over θ and λ.

The contours are only slightly inclined to the axes, indicating only slight correlation between θ and λ. Similar comments apply to the variation of the likelihood L with Ⓗ and λ, as shown in Figure 8.

Figure 9 gives the variation of the likelihood L, maximized over θ, Ⓗ, and δ, with λ. This illustration shows L is approximately quadratic in λ near the maximum and that the curve is well behaved and unimodal.

Diagnostics

We have now estimated jointly the transformation parameter and the parameters of the ARIMA(0,1,1)·(0,1,1) model based on the transformed data. The next step is to examine the model for adequacy of fit to the transformed data.

The adequacy of fit was tested by examining the residuals, their autocorrelations, and their cross-correlations with the differenced transformed series. As a first step, we plotted the residuals (see Figure 10). Examination of this plot reveals no evidence of systematic change in variance, nor of cyclical patterns. There is some suggestion of "stickiness" in one or two places, but the significance of this was easily checked through the autocorrelation function.

The autocorrelation function of the estimated residuals is given in Table 7. Nowhere do the autocorrelations exceed two standard errors in magnitude. At small lags and at multiples of 12, however, the standard errors can be somewhat smaller than those shown [6], and care must be taken to allow for this possibility. Even so, the only doubtful value is at lag 36, but with no other suspicious values

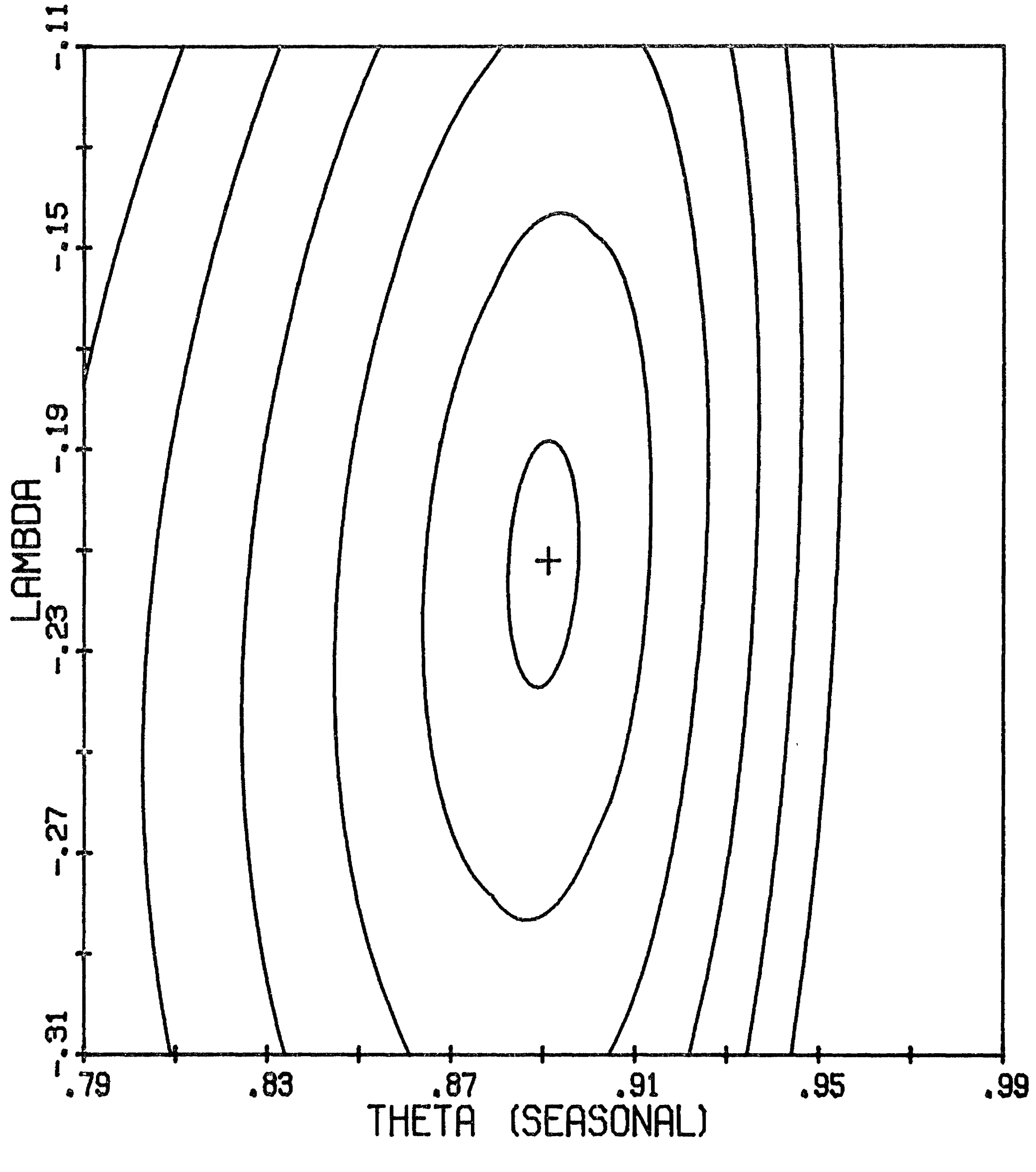

Fig. 8. Log likelihood contours over Ⓗ and λ.

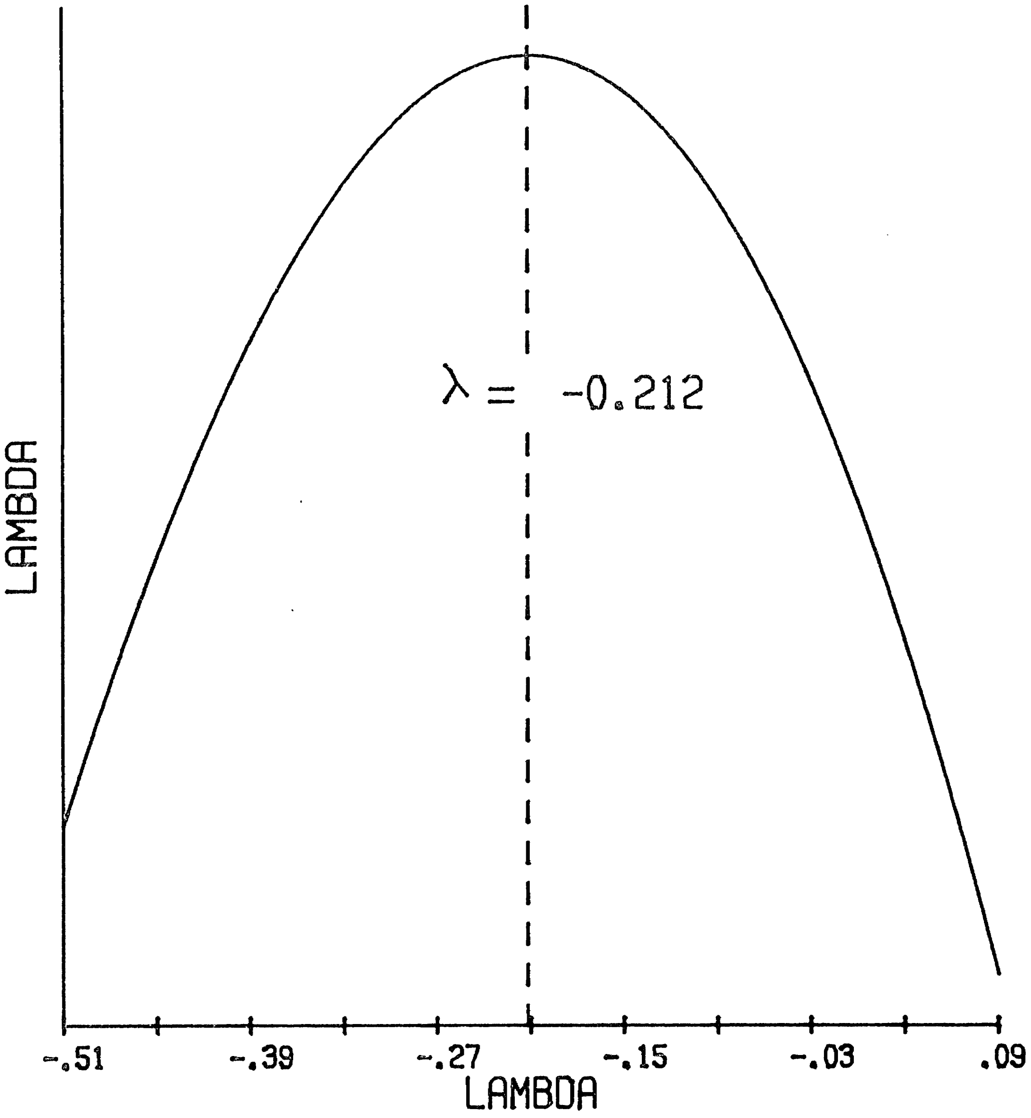

Fig. 9. Variations of log likelihood L with X.

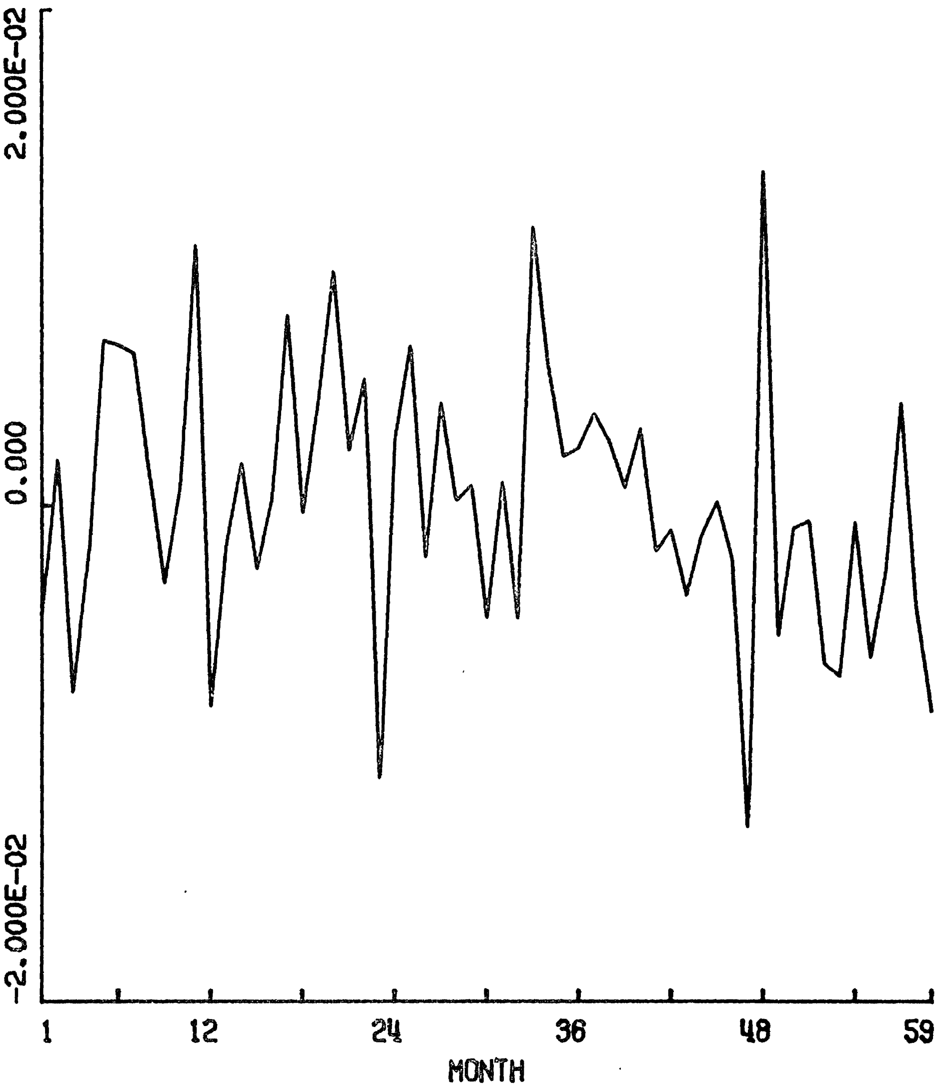

Fig. 10. Residuals of ARIMA (0,1,1)·(0,1,1) model for $\{y_\lambda(t)\}$.

there is insufficient evidence to conclude that the residuals are not white noise.

TABLE 7

RESIDUAL AUTOCORRELATIONS

Lag							Approximate Standard Error
1-6	-.07	.03	.15	-.03	-.01	.18	.13
7-12	-.08	.01	.20	-.16	.11	.05	.14
13-18	.01	.09	.12	-.09	.03	-.09	.15
19-24	-.09	.07	-.10	-.10	.08	-.01	.16
25-30	-.25	.01	-.01	.06	.13	-.11	.16
31-36	.03	-.04	-.11	-.07	-.02	-.20	.17

It should be noted that the distributions derived by Box and Pierce do not allow for the effect of estimating the transformation parameter λ. Thus we have assumed that these distributions are not greatly affected by the estimation process and that the standard errors are approximately the same as in a fixed transformation. This assumption is supported by our observation that estimates of θ and Θ are insensitive to λ, although an analytical investigation must eventually be undertaken.

The second test calculated the Q statistic [6], [4, p. 291]. This is given by

$$Q = 59 \sum_{1}^{36} r_k^2(\hat{a}) = 26.5$$

where $r_k(\hat{a})$ is the residual autocorrelation at lag k. The factor of 59 is the number of values in the differenced series, and the upper limit of 36 is chosen as the value beyond which theoretical autocorrelation becomes negligible. Box and Pierce have shown that Q is asymptotically χ^2 with 36-3 = 33 degrees of freedom (there being 3 parameters, θ, Θ, and δ, estimated in the model). We have assumed that the Q statistic will, in this case, be approximately χ^2 with 32 degrees of freedom, a reduction of one degree of freedom to allow for estimation of the additional parameter λ. The 5 percent critical value of χ^2 for 32 degrees of freedom is 46.2. This test therefore provides no evidence rejecting the estimated model.

Finally, we can examine the cross correlations between the differenced series $w_\lambda(t) = \nabla\nabla_{12}y_\lambda(t)$ and the estimated residuals $\hat{a}(t)$. For positive values of k, we would expect the cross-correlations of $w_\lambda(t)$ and $\hat{a}(t+k)$ to be close to zero. The cross-correlations between $w_\lambda(t+k)$ and $\hat{a}(t)$ should be somewhat removed from zero at small lags and multiples of 12 but should diminish as k increases. This general pattern is apparent in Table 8.

We conclude from this and the preceding section that the data is closely approximated by the model

$$\begin{aligned} y_\lambda(t) &= \frac{\{y(t)\}^\lambda - 1}{\lambda} \qquad \lambda = -.212 \\ w_\lambda(t) &= \nabla\nabla_{12}y_\lambda(t) \\ w_\lambda(t) &= .000663 + (1-.423B)(1-.891B^{12})a(t) \qquad (10) \\ \mathrm{Var}\{a(t)\} &= .0000351 \ . \end{aligned}$$

TABLE 8

CROSS-CORRELATIONS OF RESIDUALS AND DIFFERENCED SERIES

Lag	Cross-Correlations $\hat{a}(t)$, $w_\lambda(t+k)$					
1-6	-.46	.20	-.03	-.05	.05	.02
7-12	-.09	.04	.07	-.16	.16	-.46
13-18	.16	.06	.02	-.07	.03	-.10
19-24	.02	.08	-.17	.09	-.03	.02
25-30	-.19	.04	-.04	.06	.05	-.11
31-36	.07	-.12	.03	-.04	.00	-.12

	Cross-Correlations $\hat{a}(t+k)$, $w_\lambda(t)$					
1-6	-.02	-.05	.01	.05	-.06	.23
7-12	.04	-.17	.27	.00	-.07	.01
13-18	.19	-.07	.18	.00	-.05	.07
19-24	-.11	.13	.03	-.13	.05	.25
25-30	-.21	.03	.06	-.02	.15	-.01
31-36	-.05	.04	.10	-.11	.04	-.01

Forecasting

Forecasts were calculated up to twelve months ahead and compared with actual values.

A formula for forecasts of the transformed data can be easily obtained for the model (10). We use $\hat{w}_{\lambda t}(\ell)$ to represent the forecast of $w_\lambda(t+\ell)$ made from the origin date t. Noting that the forecast of $a(t+\ell)$ is zero for positive values of ℓ, we have

$$w_{\lambda t}(\ell) = .000663 - .423\hat{a}(t) - .891\hat{a}(t-12) + .377\hat{a}(t-13) ,$$

and for $1 < \ell \leq 12$ we have

$$\hat{w}_{\lambda t}(\ell) = .000663 - .891\hat{a}(t-12) + .377\hat{a}(t-13) .$$

Now

$$w_\lambda(t) = y_\lambda(t) - y_\lambda(t-1) - y_\lambda(t-12) + y_\lambda(t-13),$$

and thus

$$y_\lambda(t) = w_\lambda(t) + y_\lambda(t-1) + y_\lambda(t-12) - y_\lambda(t-13) .$$

We therefore can obtain forecasts $\hat{y}_{\lambda t}(\ell)$ of $y_\lambda(t+\ell)$ from

$$\hat{y}_{\lambda t}(1) = \hat{w}_{\lambda t}(1) + y_\lambda(t) + y_\lambda(t-11) - y_\lambda(t-12) ,$$

and for $1 < \ell \leq 12$

$$y_{\lambda t}(\ell) = \hat{w}_{\lambda t}(\ell) + \hat{y}_{\lambda t}(\ell-1) + y_\lambda(t+\ell-12) - y_\lambda(t+\ell-13) .$$

Approximate tolerance limits for $\hat{y}_{\lambda t}(\ell)$ were calculated using the methods described by Box and Jenkins [4, Chapter 5]. Forecasts and tolerance limits for the transformed data are given in Table 9. Note that the limits are tolerance (or probability) limits, not confidence intervals. They are only approximate because they are based on estimated values of θ, $\textcircled{H}$, and σ^2, rather than true values.

TABLE 9

TRANSFORMED SERIES--FORECASTS FROM ORIGIN t = 72

Lead Time	Forecast $\hat{y}_{\lambda t}(\ell)$	95 Percent Tolerance Limits
1	3.4712	± .0118
2	3.4933	± .0137
3	3.4879	± .0153
4	3.4804	± .0168
5	3.4747	± .0181
6	3.4554	± .0193
7	3.4416	± .0205
8	3.4365	± .0216
9	3.4521	± .0227
10	3.4516	± .0237
11	3.4483	± .0247
12	3.4376	± .0256

Forecasts $\{\hat{y}_t(\ell)\}$ of the original series $\{y(t+\ell)\}$ were then obtained by means of the inverse transformation

$$\hat{y}_t(\ell) = (-.212\hat{y}_{\lambda t}(\ell)+1)^{(-1/.212)} .$$

Approximate 95 percent tolerance limits were obtained by similarly inverting the limits for $\hat{y}_{\lambda t}(\ell)$. The final results are shown in Table 10 and plotted in Figure 11.

There is a special problem with calculating forecasts as in Table 10. Box-Jenkins forecasts are minimum, mean-square error forecasts, and, under the assumption of independent normal residuals, these are conditional expectations. Because the normal distribution is symmetrical, the mode and mean coincide and the forecasts become most probable forecasts.

TABLE 10

FORECASTS OF ORIGINAL SERIES FROM ORIGIN DATE t=72

Lead Time	Lower 95 Percent Limit	Forecast	Upper 95 Percent Limit	Actual	Error
1	511	534	559	553	19
2	551	581	613	586	5
3	537	569	604	564	-5
4	519	553	590	553	0
5	505	541	580	519	-22
6	468	503	541	497	-6
7	443	478	516	481	3
8	433	469	508	470	1
9	456	497	542	492	-5
10	455	496	541	492	-4
11	447	490	538	481	-9
12	429	471	518	453	-18

However, if we obtain such forecasts for transformed data with a value of $\lambda \neq 1$, then use the inverse transformation to obtain forecasts of the original series, the distribution of the forecast error will be skewed and the mode and mean will not be the same. The forecasts will still correspond to the mode of the distribution, and therefore will be most probable forecasts, but they will not be conditional expectations nor minimum mean square error forecasts. Some discussion of the possibility of adjusting forecasts to allow for a quadratic loss function is given by Chatfield and Prothero [7], Harrison [8], and Box and Jenkins [5].

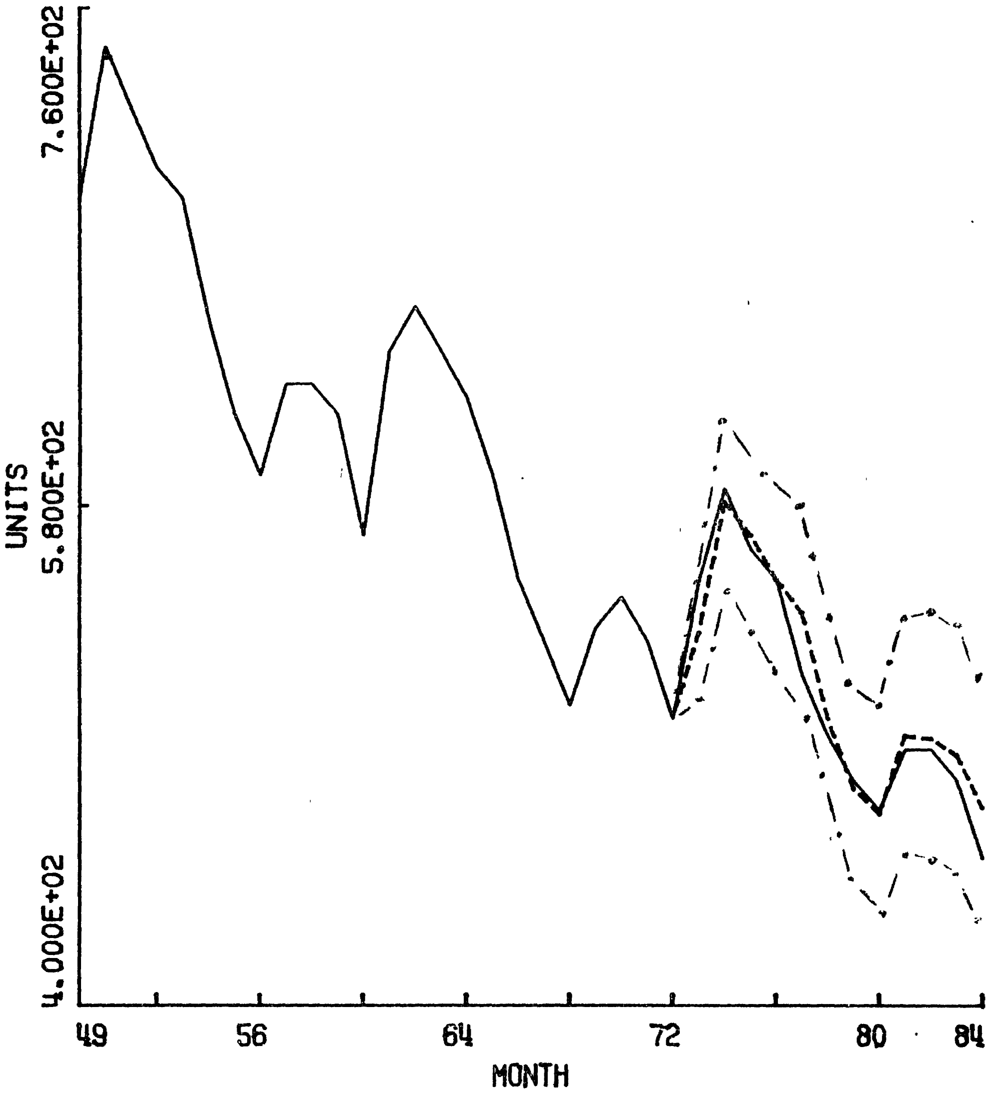

Fig. 11. Original Series with forecasts and approximate 95 percent tolerance limits from origin date 72 (December 1974).

—— Original series

---- Forecasts

_._. 95 percent tolerance limits

Comparison with Log Model

To compare the forecasting performance of the transformed model with that of the more conventional log model, we fit the best model to the logarithms and calculated a set of forecasts. The log model used was

$$w_0(t) = \nabla\nabla_{12}\ell n\ y(t)$$

$$w_0(t) = .00349 + (1-.389B)(1-.904B^{12})a(t)\ .$$

All parameters in this model were significant, and no serious irregularities were found in the diagnostics.

The following sets of forecasts were generated for the log model. First we found one-step-ahead forecasts for origin dates 72,73,....,83. Then we calculated two-steps-ahead forecasts for origin dates 72,73,....,82, and so on, until finally we obtained a single 12-steps-ahead forecast for the origin date 72. The same model parameters were used at each origin date; no attempt was made to re-estimate the parameters as additional data points became available with the advancing origin date. A similar set of forecasts was obtained for the transformed data model (λ=-.212). In both cases, forecasts of the original data were obtained by simple inversion of forecasts of the transformed data, and no adjustment was made for skewness.

Table 11 summarizes the forecasting errors for the transformed (λ=-.212) model and the log model. These results speak for themselves: the model for λ=-.212 was clearly superior.

TABLE 11

COMPARISON OF FORECASTING PERFORMANCE

	λ = -.212		λ = 0(log)		Percentage Improvements		Individual Improvements	
Steps Ahead	Mean Absolute Error	Mean Square Error	Mean Absolute Error	Mean Square Error	Mean Absolute Error	Mean Square Error	Number	Percentage
1	8.1	116	8.7	125	7	7	9	75
2	10.0	168	11.7	194	15	13	9	82
3	11.5	213	13.2	269	13	21	7	70
4	14.6	264	16.2	354	10	25	6	67
5	15.1	263	18.8	400	20	34	7	88
6	16.4	298	20.0	431	18	31	7	100
7	18.8	446	23.0	620	18	28	6	100
8	22.6	665	28.2	954	20	30	5	100
9	25.0	772	32.3	1169	23	34	4	100
10	27.7	947	36.0	1469	23	36	3	100
11	29.5	1003	39.5	1693	25	41	2	100
12	32.0	1024	42.0	1764	24	42	1	100
Overall	15.2	345	18.4	495	17	30	66	85

Discussion

Use of the Box-Cox transformation in this example has been shown to give greatly improved forecasting performance. This approach is new and relatively untried, however, and there are several issues yet to be resolved.

First, identification depends largely on the near orthogonality of the estimator for λ and the estimators of the other structural parameters. A group at the University of Michigan is working on a formal proof of this property, but until they are finished, we have only our rather limited experience on which to rely. Similarly, we believe that asymptotic distributions of parameter estimators and residual autocorrelations remain unchanged except for the loss of a degree of freedom, but, again, a rigorous analytical proof is required.

Forecasting performance has not yet been fully investigated. Experience to date has disclosed substantial improvements in forecasts, and we have found no case in which inferior forecasts were generated. However, a detailed empirical study covering a wide range of time series is being planned by researchers at The University of Michigan to substantiate this experience. It will adopt the same general approach as used by Reid [13] and Newbold and Granger [4] to compare different forecasting methods. Despite these reservations, all indications are that the modified Box-Jenkins method is easy to use and gives generally better forecasts. In short, it seems to work.

Footnotes

1. The reason for working with the transformation (1) rather than the power transformation (2) is that the former is continuous at $\lambda = 0$ (the log transformation).

2. Pierce [12] has shown that the Box-Jenkins least-squares estimators are consistent and have the same asymptotic distributions for both normal and non-normal series.

3. Standard errors were obtained by inverting the estimated information matrix (see Box and Jenkins [4, p. 227]).

4. This matrix was obtained by inverting the estimated information matrix.

References

1. Ansley, C. F., Spivey, W. A., and Wrobleski, W. J. "An Analysis of Transformations for Time Series." Proceedings of the American Statistical Association, Business and Economics Section, 1975, pp. 211-16.

2. _______. "A Class of Transformations for Box-Jenkins Seasonal Models" paper submitted for publication to Applied Statistics, February 1976.

3. Box, G. E. P., and Cox, D. R. "An Analysis of Transformations." Journal of the Royal Statistical Society, B, No. 26 (1964): 211-43.

4. Box, G. E. P., and Jenkins, G. M. Time Series Analysis, Forecasting and Control. San Francisco, Calif.: Holden Day, 1970.

5. _______. "Some Comments on a Paper by Chatfield and Prothero and on a Review by Kendall." Journal of the Royal Statistical Society A, No. 135 (1973): 337-45.

6. Box, G. E. P., and Pierce, D. A. "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models." Journal of the American Statistical Association 64 (1970): 1509-26.

7. Chatfield, C., and Prothero, D. L. "Box-Jenkins Seasonal Forecasting: Problems in a Case Study." Journal of the Royal Statistical Society A, No. 136 (1973): 295-315.

8. Harrison, P. J. "Discussion of a Paper by Dr. Chatfield and Dr. Prothero." Journal of the Royal Statistical Society A, No. 136 (1973): 319-24.

9. Marquardt, D. W. "An Algorithm for Least Squares Estimation of Nonlinear Parameters." Journal of the Society for Industrial and Applied Mathematics 2 (1963): 431-41.

10. Nelson, C. R. Applied Time Series Analysis for Managerial Forecasting. San Francisco, Calif.: Holden Day, 1973.

11. Newbold, P., and Granger, C. W. J. "Experience with Forecasting Univariate Time Series and the Combination of Forecasts." Journal of the Royal Statistical Society A, No. 137 (1974): 131-64.

12. Pierce, D. A. "Least Squares Estimation in the Regression Model with Autoregressive-Moving Average Errors." Biometrika 58 (1971): 299-312.

13. Reid, D. J. "A Comparative Study of Time Series Prediction Techniques on Economic Data." Unpublished Ph.D. Thesis, University of Nottingham, 1969.

14. Tunnicliffe Wilson, G. "Discussion of a Paper by Dr. Chatfield and Dr. Prothero." Journal of the Royal Statistical Society A, No. 135 (1973): 315-19.