The Journal of National Chiao Tung University
Vol. 3, September 1977
pp. 1-12

營利事業所得稅資料之隨機模型

Stochastic Model Fitting of Business Income Tax Data

陳 稔 Zen Chen

Institute of Computer Science, N. C. T. U.

(Received March 8, 1977)

Abstract — In business income tax computation the items listed in the tax return form are mostly recorded at consecutive equally-spaced time instants. Linear stochastic models are found to be appropriate to fit these items. Firstly, the data has to be converted into a proper form suitable for fitting. Then a procedure is used to find a model of the series. It involves the model identification, parameter estimation and model diagnostic evaluation. Finally, the model obtained is used to predict the future values of the series based on a minimum mean square error criterion.

I. Introduction

The business income tax is computed based on a set of business related items, notably the sales volumes, business costs, wages, dividends. In general, these items are recorded at consecutive equally-spaced time instants. They can be classified into two types, deterministic and nondeterministic. For the nondeterministic data (e.g. sales volume), it contains maximum and minimum points which occur at about the same time instants year after year. For instance, airline companies encounter more passengers in those months during vacations. Thus, the business data reflects the seasonal characteristic. Furthermore, the company business may be affected by the individual financial strength, the national and the world-wide economics. As a result, the company may expand or shrink its business size. The component of the data showing these phenomena is called the trend of the business. Often this factor affects the data gradually. Another feature of the business data is randomness, different from the physical systems which have deterministic behaviors. The randomness at any instant may be dictated by a underlying probabilistic distribution. But the data which are close in time behave more similarly than those far apart in time, so the factor is not purely a random variables, instead they are governed by stochastic processes.

A mathematical representation of the business data has to tackle with various characteristics of the business. The precise model may be very complicate, if not impossible. The computation involved may be time costly. Some forms of approximation with simplicity in computation may be desirable. This is the purpose of this paper to propose the method of linear time series models to deal with this type of problem. To illustrate the use of time series analysis, the uniform invoice data is selected for study. The reason that this data is used is because it is a typical data used in the tax return form. Besides, the uniform invoice is easy to get and the data is believed reliable.

In Section II the monthly uniform invoice is processed such that it becomes suitable for the ensuing analysis. Section III is devoted to the determination of a time series model of the data. In Section IV, the minimum mean square error forecast based on the model obtained in Section III is used to predict the future values of monthly data. Section V is the concluding remark.

II. Transformation of Monthly Uniform Invoice Data

The time series data to be studied is the monthly uniform invoice data of a certain local company. The set of data

consists of 120 monthly invoice total (measured in thousands) from January, 1965 to December, 1974. These totals are tabulated in Table 1.

Table 1. The monthly uniform voice totals of a company (unit: NT$1,000)
$\{Z_t\}$, t = 1,2,3,...,120

Year / Month	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
Jan.	9,949	9,951	13,689	20,163	20,608	43,125	44,106	56,747	73,096	61,325
Feb.	13,981	14,282	14,860	21,294	25,697	31,438	50,038	55,640	56,064	64,981
Mar.	17,891	17,253	20,053	24,755	36,992	43,603	58,653	59,979	68,242	83,513
Apr.	14,897	15,702	18,491	24,575	36,157	48,846	63,687	61,320	68,582	77,997
May	16,346	16,485	19,282	25,222	38,404	44,079	54,656	64,679	72,085	85,361
June	16,905	18,477	19,674	25,845	42,907	47,433	66,222	69,534	71,605	84,116
July	14,946	13,080	19,795	27,067	42,543	57,972	65,660	66,236	74,471	89,748
Aug.	15,384	16,383	20,461	25,037	33,658	46,605	60,272	64,726	67,965	77,420
Sept.	16,156	15,289	18,790	24,268	33,663	44,691	55,759	61,368	66,617	74,438
Oct.	16,247	17,564	21,598	32,275	45,579	61,496	65,597	67,930	78,204	93,992
Nov.	15,521	16,426	20,868	29,748	37,324	48,359	57,303	61,994	76,080	86,706
Dec.	20,258	22,202	24,281	45,360	56,862	68,476	81,370	73,641	90,483	100,135

Later on, the first nine-year data will be used to set up the time series model and the last-year data will be used as test samples to compare with the minimum mean square error forecasts yielded by the model. The original data are different as widely as by a factor of 10. To facilitate the analysis, a smoother form is perferred so the natural logarithms of the original data are used. Table 2 contains the logarithms of the monthly data. The converted data is depicted in Fig. 1(a). This uprising curve illustrates the trend of the booming business. The data does not have an equilibrium level.

Table 2. $\{\ell n Z_t\}$

Year / Month	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
1	9.2052	9.2054	9.524	9.912	9.993	10.672	10.694	10.946	11.200	11.024
2	9.545	9.604	9.606	9.966	10.154	10.356	10.821	10.927	10.934	11.082
3	9.792	9.756	9.906	10.117	10.518	10.683	10.979	11.002	11.131	11.333
4	9.609	9.662	9.825	10.109	10.496	10.796	11.062	11.024	11.136	11.264
5	9.702	9.710	9.867	10.135	10.556	10.694	10.909	11.077	11.186	11.355
6	9.735	9.824	9.887	10.160	10.667	10.767	11.101	11.150	11.179	11.340

Year / Month	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
7	9.612	9.479	9.893	10.206	10.658	10.968	11.092	11.101	11.218	11.405
8	9.641	9.731	9.926	10.128	10.424	10.749	11.007	11.078	11.127	11.257
9	9.690	9.635	9.841	10.097	10.424	10.708	10.929	11.025	11.107	11.218
10	9.696	9.774	9.980	10.382	10.727	11.027	11.091	11.126	11.267	11.451
11	9.650	9.707	9.946	10.301	10.527	10.786	10.956	11.035	11.240	11.370
12	9.916	10.008	10.097	10.722	10.948	11.134	11.307	11.207	11.413	11.514

Meanwhile the data in Table 2 indicates a seasonal cycle with the periodicity being about one year. It would be desirable to capture this factor. This is done as follows.

Assume the original data be $\{Z_t\}$, $t = 1, 2, \ldots, 120$, as given in Table 1. Firstly it is converted into a new series $\{\ell nZ_t\}$, $t = 1, 2, \ldots, 120$. Then to remove the seasonal factor, compute

$$\nabla_{12}\, \ell nZ_t = \ell nZ_t - \ell nZ_{t-12}$$

where $t = 13, 14, \ldots, 120$.

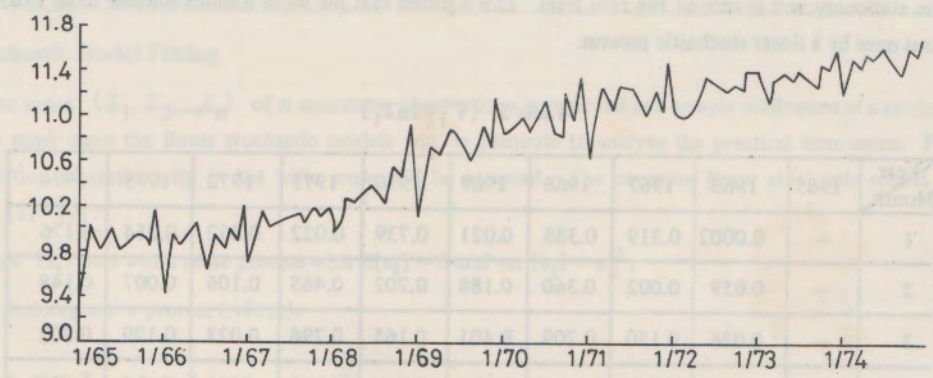


Fig. 1(a) The time series in the logarithm $\{\ell n\, Z_t\}$

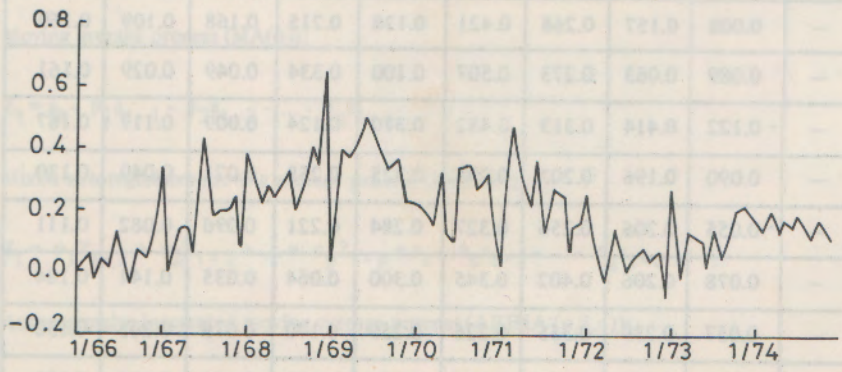


Fig. 1(b) The time series $\{\nabla_{12}\, \ell_n\, Z_t\}$

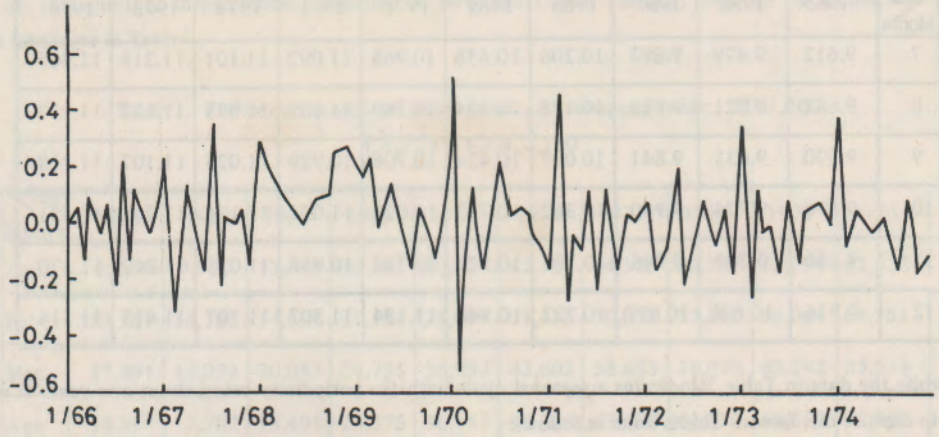


Fig. 1(c) The time series $\{\nabla\nabla_{12}\ \ell n\ Z_t\}$

The series $\{\nabla_{12}\ \ell nZ_t\}$ is tabulated in Table 3 and is depicted in Fig. 1 (b). It is found that the series $\{\nabla_{12}\ell nZ_t\}$ tends to be flatter. However, the series still fluctuates; it gradually rises to a climax and then decreases. The fluctuation can be eased up by performing an additional difference operator $\omega_t = \nabla(\nabla_{12}\ \ell nZ_t)$. The series $\{\omega_t\}$ obtained above is shown in Table 4. The plot of the series $\{\omega_t\}$ is illustrated in Fig. 1 (c); obviously the obtained series appears more stationary and is around the zero level. This signifies that the series is much suitable to be analyzed than the previous ones by a linear stochastic process.

Table 3. $\{\nabla_{12}\ \ell nZ_t\}$

Year / Month	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
1	–	0.0002	0.319	0.388	0.021	0.739	0.022	0.252	0.254	-0.176
2	–	0.059	0.002	0.360	0.188	0.202	0.465	0.106	0.007	0.148
3	–	-0.036	0.150	0.209	0.401	0.165	0.296	0.023	0.129	0.202
4	–	0.053	0.163	0.284	0.387	0.300	0.266	-0.038	0.112	0.128
5	–	0.008	0.157	0.268	0.421	0.138	0.215	0.168	0.109	0.169
6	–	0.089	0.063	0.273	0.507	0.100	0.334	0.049	0.029	0.161
7	–	-0.122	0.414	0.313	0.452	0.310	0.124	0.009	0.117	0.187
8	–	0.090	0.196	0.202	0.296	0.325	0.258	0.071	0.049	0.130
9	–	-0.055	0.206	0.256	0.327	0.284	0.221	0.096	0.082	0.111
10	–	0.078	0.206	0.402	0.345	0.300	0.064	0.035	0.141	0.184
11	–	0.057	0.239	0.355	0.226	0.259	0.170	0.079	0.205	0.130
12	–	0.092	0.089	0.615	0.226	0.186	0.173	-0.100	0.206	0.101

Table 4. $\omega_t = \{\nabla\nabla_{12} \ell n Z_t\}$

Year / Month	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
1	–	–	0.227	0.299	- 0.594	0.513	- 0.164	0.079	0.354	- 0.382
2	–	0.059	- 0.317	- 0.028	0.167	- 0.537	0.443	- 0.146	- 0.247	0.324
3	–	- 0.095	0.148	- 0.151	0.213	- 0.037	- 0.269	- 0.083	0.122	0.054
4	–	0.089	0.013	0.075	- 0.014	0.135	- 0.030	- 0.061	- 0.017	- 0.074
5	–	- 0.045	- 0.006	- 0.016	0.034	- 0.162	- 0.051	0.206	- 0.003	0.041
6	–	0.081	- 0.094	0.005	0.086	- 0.038	0.119	- 0.119	- 0.080	- 0.008
7	–	- 0.211	0.035	0.040	- 0.055	0.210	- 0.210	- 0.040	0.088	0.026
8	–	0.212	- 0.219	- 0.111	- 0.156	0.015	0.134	0.062	- 0.068	- 0.057
9	–	- 0.145	0.011	0.054	0.031	- 0.041	- 0.037	0.025	0.033	- 0.019
10	–	0.133	0.000	0.246	0.018	0.016	- 0.157	- 0.061	0.059	0.073
11	–	- 0.021	0.033	- 0.047	- 0.119	- 0.041	0.106	0.044	0.064	- 0.054
12	–	0.035	- 0.150	0.260	0.000	- 0.073	0.003	- 0.179	0.001	- 0.029

III. Stochastic Model Fitting

A time series $\{Z_1, Z_2, ..., Z_n\}$ of n successive observations is regarded as a sample realization of a stochastic process. In most cases the linear stochastic models will be adequate to analyze the practical time series. Furthmore the condition of stationarity in the wide sense will be assumed. The common linear stochastic model are given as below [1] - [2]:

Assume $\{a_t\}$ be a white noise process with $E[a_t] = 0$ and var $[a_t] = \sigma_a^2$.

(1) Autoregressive process (AR(p)):

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi_p Z_{t-p} + a_t$$

(2) Moving average process (MA(q)):

$$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - .... - \theta_q a_{t-q}$$

(3) Mixed autoregressive-moving average process (ARMA (p,q)):

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - .... - \theta_q a_{t-q}$$

(4) Autoregressive integrated moving average process (ARIMA (p,d,q)):

$$\omega_t = \phi_1 \omega_{t-1} + \phi_2 \omega_{t-2} + ... + \phi_p \omega_{t-p} + a_t - \theta_1 a_{t-1} - .... - \theta_q a_{t-q}$$

where $\omega_t = (1 - B)^d Z_t \triangleq \nabla^d Z_t$; B is a backward shift operator.

(5) Multiplicative seasonal model:

$$\Phi_p(B)\, \Psi_u(B^s) \nabla^d \nabla_s^D Z_t = \theta_q(B)\, \textcircled{H}_v(B^s)\, a_t$$

where $\Phi_p(B)$ is a polynomial in B of degree p,

$\Psi_u(B^s)$ is a polynomial in B^s of degree u,

$\theta_q(B)$ is a polynomial in B of degree q,

$\textcircled{H}_v(B^s)$ is a polynomial in B^s of degree v,

$\nabla_s^D Z_t$ is $(1 - B^s)^D Z_t$

The model building involves a tentative model identification, estimation of model parameters and model diagnostic evaluation. The entire procedure terminates provided that the model diagnostic evaluation indicates the satisfaction of the model. Otherwise, the procedure is recycled.

1. Model Identification:

For a stationary stochastic process $\{\omega_1, \omega_2, ..., \omega_n\}$, the sample autocovariance function c_k at lag k and the sample autocorrelation function r_k at lag k are computed as below.

$$c_k = \frac{1}{n-k} \sum_{t=1}^{n-k} (\omega_t - \bar{\omega})(\omega_{t+k} - \bar{\omega})$$

$$r_k = c_k / c_o$$

with $\bar{\omega} = \frac{1}{n} \sum_{t=1}^{n} \omega_t$

When the sample size n is fairly large, the distribution of r_k is approximately a normal process. It was shown [3] that the expression for the variance of r_k is given by

$$\text{var}[r_k] \cong \frac{1}{N} \sum_{v=-\infty}^{\infty} \{\rho_v^2 + \rho_{v+k}\rho_{v-k} - 4\rho_k\rho_v\rho_{v-k} + 2\rho_v^2\rho_k^2\}$$

where ρ_k is the true atuocorrelation function.

On the other hand, for an AR(k) process, the relationship between the autocorrelation functions ρ_k's can be readily shown to be

$$\rho_j = \phi_{k1}\rho_{j-1} + ... + \phi_{k(k-1)}\rho_{j-k-1} + \phi_{kk}\rho_{j-k}, j = 1, 2, ..., k.$$

The above equations are essentially the Yule-Walker equations. The entry ϕ_{kk} is called the partial autocorrelation function. A recursive formula for computing the partial autocorrelation functions is used in the computer experiment [4]. For an AR(p) process, it can be shown that ϕ_{kk} will cut off, i.e.,

$$\phi_{kk} = 0, \text{ if } k > p$$

Furthmore, the variance of $\{\phi_{kk}\}$, $k > p$, is approximately independently distributed with [5]

$$\text{var}[\phi_{kk}] \cong \frac{1}{n}, \quad k > p.$$

The model identification of a stationary time series is accomplished based on the following fact.

(a) the autocorrelation function of an AR(p) process tails off, its partial autocorrelation function has a cutoff after lag p.

(b) the autocorrelation function of an MA(q) has a cutoff after lag q, while its partial autocorrelation tails off.

(c) the autocorrelation function and the partial autocorrelation function of an ARMA(p,q) tail off. Furthermore, the autocorrelation function is dominated by a mixture of exponential and damped sine waves after (q-p) lags. Conversely, the partial autocorrelation function is dominated by a mixture of exponential and damped sine waves after (p-q) lags.

The foregoing idea is applied to the business data in Table 4. The estimated autocorrelation function and partial autocorrelation functions of the transformed data $\omega_t = \nabla \nabla_{12} \ell n Z_t$ are obtained via the computer program, as shown in Table 5.

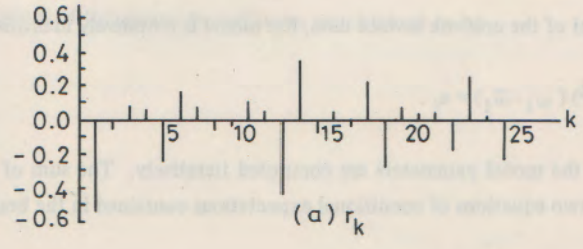
Table 5. Estimated autocorrelations $\{r_k\}$ and partial autocorrelations $\{\phi_{kk}\}$ of the invoice data

k	1	2	3	4	5	6	7	8	9	10	11	12	13
r_k	-0.527	-0.009	0.066	0.056	-0.241	0.190	0.057	-0.082	-0.005	0.105	0.045	-0.417	0.368
ϕ_{kk}	-0.53	-0.40	-0.25	-0.04	-0.32	-0.23	-0.03	0.04	0.05	0.15	0.47	-0.12	-0.10
k	14	15	16	17	18	19	20	21	22	23	24	25	
r_k	-0.084	0.026	-0.066	0.218	-0.260	0.062	0.025	0.025	-0.175	0.272	0.003	-0.230	
ϕ_{kk}	-0.16	0.02	-0.08	0.02	0.06	0.08	0.04	0.04	-0.16	0.07	0.07	-0.04	

They are also plotted in Figs 2(a) and 2(b). After careful reviewing of autocorrelation functions, they are found to tail off; the partial autocorrelation functions, after lags greater than 13, are confined in the lines about zero at $\pm 1/\sqrt{n} \cong \pm 0.1$. The above manoeuvre suggests that the tentative model is an AR(13) process. According to the principle of parsimony in the use of parameters the model of the form

$$(1 - \phi_1 B)(1 - \phi_2 B^{12})(\omega_t - \bar{\omega}_t) = a_t$$

is the proper one worthy for a close study.



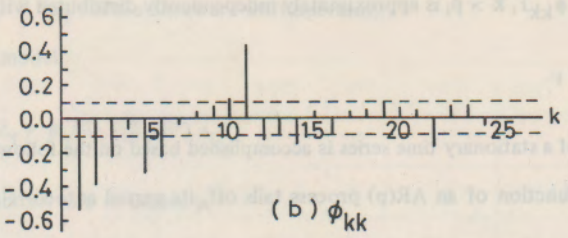


Fig. 2 (a) Estimated autocorrelations r_k of the invoice date
(b) Estimated partial autocorrelations ϕ_{kk} of the invoice data

2. Model Estimation:

In an arbitrary ARMA (p,q) model the assumption that the a_t's are normally distributed (i.e., white noise) leads to the derivation of the conditional distribution

$$f(\underset{\sim}{\omega}_n | \underset{\sim}{\phi}, \underset{\sim}{\theta}, \sigma_a) = (2\pi\theta_a^2)^{-n/2} |M|^{-1/2} \exp\{-\frac{S(\underset{\sim}{\phi}, \underset{\sim}{\theta})}{2\sigma_a^2}\}$$

where $\underset{\sim}{\omega}_n \triangleq (\omega_1, \omega_2, \ldots, \omega_n)$

$\underset{\sim}{\phi} \triangleq (\phi_1, \phi_2, \ldots, \phi_p)$

$\underset{\sim}{\theta} \triangleq (\theta_1, \theta_2, \ldots, \theta_q)$

$\sigma_a^2 \triangleq$ variance of a_t

and M is a matrix whose elements are functions of $\underset{\sim}{\phi}$ and $\underset{\sim}{\theta}$. In addition it can be shown that

$$S(\underset{\sim}{\phi}, \underset{\sim}{\theta}) = \sum_{t=-\infty}^{n} E[a_t | \underset{\sim}{\omega}_n, \underset{\sim}{\phi}, \underset{\sim}{\theta}]^2.$$

In this case the log-likelihood function is given by

$$L(\underset{\sim}{\phi}, \underset{\sim}{\theta}, \sigma_a) = -\frac{n}{2} \ln 2\pi - \tfrac{1}{2} \Pi n |M| - n \ln \sigma_a - \frac{S(\underset{\sim}{\phi}, \underset{\sim}{\theta})}{2\sigma_a^2}.$$

The maximum likelihood estimates of $\underset{\sim}{\phi}$ and $\underset{\sim}{\theta}$ are those values which maximize $L(\underset{\sim}{\phi}, \underset{\sim}{\theta}, \sigma_a)$. For a moderate and a large value of n, $L(\underset{\sim}{\phi}, \underset{\sim}{\theta}, \sigma_a)$ is dominated by $S(\underset{\sim}{\phi}, \underset{\sim}{\theta}) / 2\sigma_a^2$ which is previously given as the sum of squares of the expectation of a_t conditional on $\underset{\sim}{\omega}_n$, $\underset{\sim}{\phi}$ and $\underset{\sim}{\theta}$. Therefore the least square estimates of $\underset{\sim}{\phi}$ and $\underset{\sim}{\theta}$ are equivalent to their maximum likelihood estimates.

In the computer experiment of the uniform invoice data, the model is tentatively identified as

$$(1 - \phi_1 B)(1 - \phi_2 B^{12})(\omega_t - \bar{\omega}_t) = a_t.$$

The least square estimates of the model parameters are computed iteratively. The sum of squares $S(\underset{\sim}{\phi}, \underset{\sim}{\theta})$ is computed based on the following two equations of conditional expectations contained in the brackets.

$$[e_t] = [\tilde{\omega}_t] - \phi_1[\tilde{\omega}_{t+1}] - \phi_2[\tilde{\omega}_{t+12}] + \phi_1\phi_2[\tilde{\omega}_{t+13}]$$

$$[a_t] = [\tilde{\omega}_t] - \phi_1[\tilde{\omega}_{t-1}] - \phi_2[\tilde{\omega}_{t-12}] + \phi_1\phi_2[\tilde{\omega}_{t-13}]$$

with (i) $[\tilde{\omega}_j] = \tilde{\omega}_j$, $j = 1, 2, \ldots, n$

(ii) $[e_{-j}] = 0$, $j = 0, 1, 2, \ldots, n$

(iii) $[a_{-j}] = 0, j > J$ where $[\tilde{\omega}_{-j}]$ become zero for $j > J$.

where $\tilde{\omega}_t = \omega_t - \bar{\omega}$

The least square estimates of $\phi_1, \phi_2, \bar{\omega}_t$ are the set of values which yields the minimum value of $S(\underset{\sim}{\phi}, \underset{\sim}{\theta})$. The estimation results of the computer simulation are given in Table 6.

Table 6. Model parameter estimates and sum of squares of residuals

Est. Itex. No.	ϕ_1	ϕ_2	$\bar{\omega}_t$	$\hat{S}(\underset{\sim}{\phi}, \underset{\sim}{\theta})$
1	- 0.533	- 0.482	0.0022	1.5413
2	-0.506	- 0.439	0.0007	1.5360
3	- 0.516	- 0.445	0.0006	1.5358
4	- 0.515	- 0.442	0.0006	1.5358

The table shows the parameter estimates to be

$$\hat{\phi}_1 = -0.515$$

$$\hat{\phi}_2 = -0.442$$

$$\bar{\omega}_t = 0.0006$$

$$\hat{S}(\underset{\sim}{\phi}, \underset{\sim}{\theta}) = 1.536$$

and estimate of the residual mean $\cong 2.75 \times 10^{-4}$

estimate of the residual variance $= 1.67 \times 10^{-2}$

Also the standard deviations of the parameter estimates are obtained to be

$$\sigma(\hat{\phi}_1) = 0.131$$

$$\sigma(\hat{\phi}_2) = 0.131$$

$$\sigma(\bar{\omega}) = 0.129$$

3. Model Diagnostic Evaluation:

Suppose the form of the model is correct and the parameter values ϕ's and θ's are true, it can be shown [6] that the estimated autocorrelations $r_k(a)$ of the a's will be uncorrelated and distributed approximately normally $N(0, n^{-1})$, where n is the number of samples. In practice, the a's are estimated values due to the fact that the parameter ϕ's and θ's are estimated; the above model is only approximately correct. A better evaluation model is given in Box and Pierce [7] which states as follow.

Suppose that the first K autocorrelations $r_k(a)$, k = 1, 2,....,K, from an ARIMA(p, d, q) process is such that the weights ψ_j in the model

$$\omega_t = \phi^{-1}(B)\,\theta(B)\,a_t = \psi(B)\,a_t = (1 + \psi_1 B + \ldots + \psi_j B^j + \ldots)\,a_t$$

will be negligibly small in magnitude after j = K, then, if the model is appropriate, $Q = n \sum_{k=1}^{K} r_k^2(\hat{a})$ is approximately distributed as a chi-square distribution with a degree of freedom (K - p - q) where n = N - d, N is the number of observations of the time series. Therefore the following hypothesis testing is formed, namely, to test the hypothesis H^o : Q is from the chi-square distribution χ^2 (K - p -q) against the hypothesis H':Q is not from χ^2(K - p - q). The test used is as follow: Accept H_o if $Q < \chi_\epsilon^2$ (k - p - q) and reject otherwise, where χ_ϵ^2 (k- p -q) is the significance point such that $P_r\{Q > \chi_\epsilon^2 (k - p - q)\} = \epsilon$.

In the computer simulation of the uniform invoice data, the quantity $Q = n \sum_{k=1}^{k} r_k^2(\hat{a}) = 95 \sum_{k=1}^{48} r_k^2(a)$ has approximately a chi-square distribution with 45 degree of freedom. The observed value of Q is 52.3167. By looking up a chi-square table, it is found that

$P_r\{\chi^2$ with 45 degree of freedom $\leqslant 50.985\} = 0.75$

$P_r\{\chi^2$ with 45 degree of freedom $\leqslant 57.505\} = 0.90$

Therefore, the above check indicates the model fitting is adequate.

IV. Minimum Mean Square Error Forecast

After the stochastic model is set up, the model can be used to forecast the future values of an observed time series based on the current and previous observations. In this paper the minimum mean square error forecast is used.

Consider an ARIMA (p, d, q) model,

$$\omega_{t+\ell} = \phi_1\,\omega_{t+\ell-1} + \phi_p\,\omega_{t+\ell-p} + a_{t+\ell} - \theta_1 a_{t+\ell-1} - \cdots - \theta_q\,a_{t+\ell-q}$$

$$= (a_{t+\ell} + \psi_1 a_{t+\ell-1} + \ldots + \psi_{\ell-1} a_{t+1}) + (\psi_\ell a_t + \psi_{\ell+1} a_{t-1} + \ldots)$$

Now assume the forecast value of $\omega_{t+\ell}$ based on $\omega_t, \omega_{t-1}, \ldots, \omega_1$ be given by

$$\hat{\omega}_t(\ell) = \psi_\ell^*\,a_t + \psi_{\ell+1}^*\,a_{t-1} + \ldots\ldots$$

The the mean square error is

$$E[\omega_{t+\ell} - \hat{\omega}_t(\ell)] \equiv (1 + \psi_1^2 + \ldots + \psi_{\ell-1}^2)\,\sigma_a^2 + \sum_{j=0}^{\infty} (\psi_{\ell+j} - \psi_{\ell+j}^*)^2\,\sigma_a^2$$

which is minimized by setting $\psi_{\ell+j}^* = \psi_{\ell+j}$, j = 0, 1, 2, ...

That is, $\hat{\omega}_t(\ell) = \psi_\ell a_t + \psi_{\ell+1} a_{t-1} + \cdots\cdots$

$= E[\omega_{t+\ell} | \omega_t, \omega_{t-1}]$

The minimum mean square error forecast can be shown to be unbiased and its variance is $(1 + \psi_1^2 + \ldots + \psi_{\ell-1}^2)\sigma_a^2$.

In the multiplicative seasonal model for the uniform invoice data, the model obtained previously is used to compute the forecasts for lead times up to 12 months. The forecasts are listed in Table 7 and are found to be fairly agreeable with the actual values except the forecast for lead time $\ell = 1$.

Table 7. Forecast values $\hat{Z}_t(\ell)$ and actual values $Z_{t+\ell}$ for the period from Jan. 1974 to Dec. 1974

ℓ	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{Z}_t(\ell)$	80081.4	67113.8	78325.5	78948.1	83420.7	85763.6	85940.5	80879.6	78204.3	89522.8	84745.4	100822.1
$Z_{t+\ell}$	61325	64981	83513	77997	85361	84116	89748	77420	74438	93992	86706	100135

This may be attributed to the fact that the data point corresponding to January, 1973 behaved rather wildly. The curve of forecast values $\hat{Z}_t(\ell)$ versus actual values $Z_{t+\ell}$ is shown in Fig. 3.

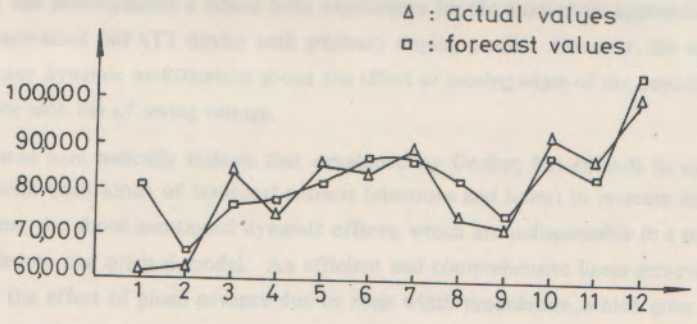


Fig. 3 Forecast values $\hat{Z}_t(\ell)$ and actual values $Z_{t+\ell}$ versus lead time ℓ.

V. Concluding Remarks

The method illustrated previously proves to be a feasible mathematical tool to analyze the business data used in the income tax computation. The data in its raw form may be distributed over a wide range and, thus, is not suitable to be fitted by a parsimonious stochastic model. Consequently it will be important to conver the original data into a flatter series.

The performance of the above analysis will be hampered by any nonnatural disturbance of the data, especially by human factors. Because this human intervention in the data destroys the natural behavior of the data. This leads to the difficulties in the model fitting. On the other hand, when the time series is long, the behavior of the latter period may deviate from that of the early stage. This is often the case when a company size is changing. Thus, it may be desired to truncate the early-stage data points as far as the estimation of the model parameters is concerned. These precautions taken will aid to the effectiveness of the time series analysis.

References

1. G. E. P. Box and G. M. Jenkins, Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco, CA. 1971.
2. T. W. Anderson, The Statistical Analysis of Time Series, John Wiley, New York, N. Y. 1971.
3. M. S. Barlett, "On the theoretical specification of sampling properties of autocorrelated time series," Jour. Royal Stat. Soc., B8, 29, (1964).
4. J. Durbin, "The fitting of time series models," Rev. Int. Inst. Stat., 28, 233, (1960).
5. M. H. Quenouille, "Approximate tests of correlation in time series," Jour. Royal Stat. Soc., B11, 68, (1949).
6. R. L. Anderson, "Distribution of the serial correlation coefficient," Ann. Math. Stat., 13, 1, (1942).
7. G. E. P. Box and D. A. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," Jour. Amer. Stat. ASSOC., 64, (1970).