Survey Methodology, June 1985
Vol. 11, No. 1, pp. 51–64
Statistics Canada

Performance of ARIMA Models in Time Series[1]

KIM CHIU, JOHN HIGGINSON, and GUY HUOT[2]

ABSTRACT

This study is mainly concerned with an evaluation of the forecasting performance of a set of the most often applied ARIMA models. These models were fitted to a sample of two hundred seasonal time series chosen from eleven sectors of the Canadian economy. The performance of the models was judged according to eight variable criteria, namely: average forecast error for the last three years, the chi-square statistic for the randomness of the residuals, the presence of small parameters, overdifferencing, underdifferencing, correlation between the parameters, stationarity and invertibility. Overall and conditional rankings of the models are obtained and graphs are presented.

KEY WORDS: X11-ARIMA; Ranking; Priority; Criteria

1. INTRODUCTION

Our socio-economic environment is unstable and uncertain; inflation, recessions, and increasing pollution are among the factors contributing to increasing instability. We try to resolve the problem by using a method of forecasting that permits us to evaluate the impact of the frequent changes. ARIMA models (Box - Jenkins, 1970) are flexible enough to deal with such frequent changes in time series.

The purpose of this paper is to study a set of eight criteria which when applied to the Box-Jenkins method permit an evaluation of the fitting and forecasting performance of a set of the most often applied ARIMA models to Canadian economic time series. The question of which models perform well is important for programs like the X-11-ARIMA (Dagum 1980) which automatically fits a fixed small set of models (three models in the case of the X-11-ARIMA) to the series.

Section 2 introduces eight criteria: the average forecast error for the last three years, the chi-square statistic for the randomness of the residuals, the presence of small parameters, overdifferencing, underdifferencing, correlation between the parameters, stationarity and invertibility. Section 3 discusses the criteria and summarizes the results. Section 4 ranks the models conditionally and unconditionally. Section 5 compares within-sample and out-of-sample extrapolated values for the last three years.

2. THE CRITERIA

In this section we give a brief discussion of the eight criteria used in ranking the models.

[1] Presented at (1) Business and Economic Forecasting Session of the Canadian Operational Research Symposium, Ottawa, May 1984 and (2) Business and Economic Statistics Section of the American Statistical Association Meetings, Philadelphia, August 1984.

[2] K. Chiu, J. Higginson, and G. Huot, Time Series Research and Analysis Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

Stability

The stability condition of a process Z_t is either "stationary" or "non-stationary". It indicates how well the system remembers the shocks a_{t-j}, $j = 1, 2, \ldots$, and how fast or slowly the response of the system to any particular shock decays. For a process

$$\begin{aligned} Z_t &= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots \\ &= \psi(B) a_t, \end{aligned}$$

where $a_t \sim NID(0, \sigma_a^2)$, the filter is said to be stable if the sequence $\{\psi_i\}$ is convergent. For a general ARIMA model (p, d, q),

$$\phi(B)\,(1 - B)^d Z_t = \theta(B) a_t,$$

the stability condition is that all the λ_i of the characteristic equation

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p = (1 - \lambda_1 B)\,(1 - \lambda_2 B) \ldots (1 - \lambda_p B) = 0$$

for the process are strictly inside the unit circle, i.e. $|\lambda_j| < 1$.

Invertibility

The process Z_t may be expressed as:

$$Z_t = a_t + \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \ldots$$

The system is said to be invertible if the sequence $\{\pi_i\}$ is convergent. The criterion is considered to be of primary importance because if the invertibility condition fails, the generating function $\pi(B)$ of the π's increases without bound. This means the current event of the system depends more on events in the distant past than in the recent past, and the process is physically meaningless.

The invertibility condition for a general ARIMA model (p, d, q), is that the ν_i of the characteristic equation

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q = (1 - \nu_1 B)\,(1 - \nu_2 B) \ldots (1 - \nu_q B) = 0$$

for the process are strictly within the unit circle, i.e. $|\nu_i| < 1$.

Underdifferencing

In the AR(p) model, when one or more of the λ_i, say λ_k approaches 1; then from

$$\begin{aligned} \phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \\ &= (1 - \lambda_1 B) \ldots (1 - \lambda_{k-1} B)\,(1 - \lambda_k B) \ldots (1 - \lambda_p B) \\ &= (1 - \lambda_1 B) \ldots (1 - \lambda_{k-1} B)\,(1 - \lambda_{k+1} B) \ldots (1 - \lambda_p B)\,(1 - \lambda_k B), \end{aligned}$$

we have $\phi(B)$ approaching

$$(1 - \phi_1' B - \phi_2' B^2 - \ldots - \phi_{p-1}' B^{p-1})\,(1 - B).$$

Therefore, a differencing operator may be needed for this system, and the AR(p) model becomes an ARI(p − 1, 1) model. Furthermore, when λ_k approaches 1, we may have non-stationarity.

Overdifferencing

Consider the general ARIMA model (p, d, q) (P, D, Q)$_s$,

$$\phi(B)\Phi(B)\,(1 - B)^d(1 - B^s)^D Z_t = \theta(B)\Theta(B)a_t.$$

If any ν_i of the characteristic equation $\theta(B) = 0$ approach 1, i.e. if any $(1 - \nu_i B)$ approach $(1 - B)$, we can eliminate $(1 - B)$ from both sides.

Test of randomness for the a_t's

Correlation in the residuals is not desirable since we want an unbiased estimate of the parameters for the process.

The statistic

$$Q = n(n + 2) \sum_{k=1}^{m} (n - k)^{-1} \varrho_k^2$$

as modified by Prothero and Wallis (1976) and Ljung and Box (1978) from the Chi-square test of Box and Pierce is used.

Here n is the sample size, $k = 1, 2, \ldots, m$ are the various lags, and ϱ_k are the autocorrelations. Q is used for the testing of the randomness of the residuals.

Small Parameters

Generally speaking, when the number of parameters of a given model is increased, the mean sum of squares σ_a^2 is reduced. However, only large parameters, or those parameters significantly different from 0 can contribute to a significant reduction of σ_a^2. To check for a small parameter, we may need an F-test (Pandit and Wu 1983):

$$F = \frac{A_1 - A_0}{s} \div \frac{A_0}{N - r} \sim F(s, N - r)$$

where r is the number of parameters of the model and s is the number of parameters which are restricted to zero. N is the number of observations, A_0 is the smaller sum of squares of the restricted model, and A_1 is the larger sum of squares of the restricted model.

But in our study here, we choose two constants, 0.05 and 0.10, as our indicator of the presence of a small parameter.

Correlation of the Parameters

High positive or negative correlation between parameters reflects ambiguity in the estimated values since a range of parameter values results in models with equally good fit. Therefore, if some of the elements in the correlation matrix of estimated parameters are large in absolute value, say greater than or equal to 0.9, the model may be reduced by deleting some of the smaller parameters.

Forecasting Error

No matter how we define a good model or bad model, we still have a primary interest in the forecasting error of the model. In this paper we use the mean absolute percentage forecasting error of one-year-ahead forecast

$$MAPE = \frac{1}{N}\sum_{\ell=1}^{N} \left| \frac{Z_{t+\ell} - \hat{Z}_t(\ell)}{Z_{t+\ell}} \right| \times 100\%$$

where ℓ is 12 or 4, and $\hat{Z}_t(\ell)$ is the forecast with lead time ℓ.

3. EVALUATION OF THE ARIMA MODELS

The eight criteria have been put into two groups. The first group considers good fitting of parsimonious models while the second considers the quality of the forecasts. This distinction between fitting and forecasting is important; good fitting and good forecasting are not equivalent.

These criteria have been used to evaluate and rank seven of the most often applied ARIMA models, namely:

1. $(0, 1, 1)\,(0, 1, 1)_s$
2. $(0, 1, 2)\,(0, 1, 1)_s$
3. $(0, 2, 2)\,(0, 1, 1)_s$
4. $(2, 1, 2)\,(0, 1, 1)_s$
5. $(1, 1, 0)\,(0, 1, 1)_s$
6. $(2, 1, 0)\,(0, 1, 1)_s$
7. $(2, 1, 0)\,(0, 1, 2)_s$

where "*s*" is 12 if the series is monthly and 4 if it is quarterly.

These models were fitted to a sample of 167 monthly seasonal time series chosen randomly from eleven sectors of the Canadian economy: national accounts; labour; prices; manufacturing; fuel, power and mining; construction; food and agriculture; domestic trade; external trade; transportation; and finance. About 40 quarterly time series from national accounts and finance were also tested.

The series are mostly multiplicative, according to the Bell Canada model test (Higginson 1976). That is, the different components (trend-cycle, seasonal, and irregular) are multiplied together to produce the raw series. Therefore, the amplitudes of the seasonal component frequently increase with increasing levels of the trend. The multiplicative series received a logarithmic transformation before the first three and last three models were fitted. The fourth model was fitted to the untransformed series in all cases.

Looking at the non-seasonal part of an ARIMA model which is associated with the trend-cycle and extremes, we see that the models can be grouped into three classes. Class I is models 1, 2 and 3 whose ordinary part includes only one or two first differences and one or two moving average parameters. Class III includes models 5, 6 and 7 whose ordinary part includes only one first difference and some autoregressive parameters. Model 4 (Class II) forms a class by itself; its non-seasonal part is mixed. We see that the seasonal part of all models is the same except for model 7.

Although the eight criteria are analysed separately in this section, several of them are dependent. For example, we shall see that the excess of parameters in model 4 generates problems of nonstationarity, noninvertibility, under- and overdifferencing, and correlation.

In Sections 3 and 4, we test within-sample extrapolated values for the seven ARIMA models. That is, the models are fitted to the whole series thus providing the parameters to be used for calculating the forecasts for the last three years. This is the way ARIMA forecasts are evaluated in the X-11-ARIMA program.

3.1 Criteria for Fitting Parsimonious ARIMA Models

The stationarity condition requires that all the roots of the autoregressive characteristic equation be inside the unit circle. We see in Table 1 that non-stationarity occurs only for model 4, in three cases. These appear to be due to overparametrization of the model.

In order for the model to be invertible, it is necessary that the roots of the moving average characteristic equation be inside the unit circle. Only model 4 has many cases of noninvertibility, 20%, as we see in Table 2. Two explanations are possible. There is first of all the case of straightforward noninvertibility. In some other cases noninvertibility was accompanied by nonstationarity. The fact that the autoregressive part may have roots near unity might have caused autocorrelation in the residuals. The moving average parameters would then take higher values to compensate.

An important criterion in judging the appropriateness of the ARIMA models for the series is the chi-square test of Box and Pierce (1970) (modified by Prothero and Wallis in 1976, and by Ljung and Box in 1978), applied to the autocorrelation of the residuals. Table 3 shows for each of the seven models the number and the percentage of series that fail the chi-square test at different levels. We see from this table first, that within a given class of models the simpler models have higher failure rates and second, that the failure rate depends to a large degree on the class of the model. The first point is illustrated by models 2 and 6 which having one more parameter than models 1 and 5, have a higher number of series passing this test. The evidence for the second point is that moving average models appear to satisfy the

Table 1

Failure in Stationarity

	CLASS I			CLASS II	CLASS III		
CRITICAL VALUE	Model 1 (0, 1, 1) (0, 1, 1)	Model 2 (0, 1, 2) (0, 1, 1)	Model 3 (0, 2, 2) (0, 1, 1)	Model 4 (2, 1, 2) (0, 1, 1)	Model 5 (1, 1, 0) (0, 1, 1)	Model 6 (2, 1, 0) (0, 1, 1)	Model 7 (2, 1, 0) (0, 1, 2)
--	-- --	-- --	-- --	3 2%	-- --	-- --	-- --

Table 2

Failure in Invertibility

	CLASS I			CLASS II	CLASS III		
CRITICAL VALUE	Model 1 (0, 1, 1) (0, 1, 1)	Model 2 (0, 1, 2) (0, 1, 1)	Model 3 (0, 2, 2) (0, 1, 1)	Model 4 (2, 1, 2) (0, 1, 1)	Model 5 (1, 1, 0) (0, 1, 1)	Model 6 (2, 1, 0) (0, 1, 1)	Model 7 (2, 1, 0) (0, 1, 2)
--	1 1%	2 1%	3 2%	33 20%	2 1%	2 1%	1 1%

Table 3

Failure in Chi-Square

	CLASS I						CLASS II		CLASS III					
CRITICAL VALUE	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
1%	31	19%	18	11%	29	17%	26	16%	62	37%	21	13%	20	12%
5%	45	27%	36	22%	46	28%	41	25%	82	49%	49	29%	42	25%
10%	61	37%	48	29%	56	34%	55	33%	89	53%	60	36%	56	34%
15%	72	43%	57	34%	69	41%	66	40%	101	60%	71	43%	64	38%
20%	83	50%	62	37%	80	48%	76	46%	106	64%	80	48%	73	44%
30%	100	60%	77	46%	94	56%	88	53%	119	71%	95	57%	89	53%
40%	111	66%	97	58%	107	64%	99	59%	127	76%	104	62%	100	60%
50%	121	72%	106	63%	118	71%	113	68%	135	81%	117	70%	116	69%
60%	131	78%	121	72%	128	77%	129	77%	141	84%	127	76%	121	72%

chi-square test better than autoregressive models. This may be due to the presence of extremes in the series. At the 5% level for example, model 1 fails for 27% of the series compared with 49% for its autoregressive counterpart model 5. As well as all models of class III, the mixed model, class II, is inferior to the second model of class I.

Underdifferencing occurs when a root of the characteristic equation of the autoregression polynomial is close to unity, say a distance ξ from unity. Here ξ is set equal to 0.1. We see in Table 4 that only model 4 is underdifferenced. This may be attributed to overparametrization. Model 4 has two autoregressive parameters and two moving average parameters in its non-seasonal part. Just through the estimation, there is a moderate chance that at least one of the autoregressive parameters will be greater than or equal to 0.9.

In this discussion the critical levels chosen for overdifferencing are 0.90 and 0.95. Table 5 shows that models 3 and 4 are most often overdifferenced. Model 3 has two first differences and two non-seasonal moving average parameters. If the second first difference is not necessary, autocorrelation is created in the series that has been differenced once already. The moving average polynomial will model this introduced autocorrelation by having one of its roots close to unity. We can therefore simplify the model by eliminating one moving average parameter and one difference. As to model 4, this may be due to overparametrization.

Table 4

Failure in Underdifferencing

	CLASS I						CLASS II		CLASS III					
CRITICAL VALUE	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
.90	--	--	--	--	--	--	14	8%	--	--	--	--	--	--

In ARIMA modelling of a stochastic process, it is enough to consider the first two moments, that is, the mean and autocovariance. The test on the size of the parameters serves only to eliminate those that contribute very little or nothing to the explanation of the autocovariance.

Table 6 illustrates two things. First, the simplest models pass this test better than more complicated models. After a logarithmic transformation, most of the multiplicative series in the sample will follow a straight line fairly closely (except for seasonal variation), so a "first difference" model will fit them using few parameters. Adding an extra unnecessary parameter to the model will often result in its receiving a small estimate from the estimation. Second, the estimated values of the moving average parameters are small (less than .05 or .10) more often than the estimated values of the autoregressive parameters. For example at the level of 0.05, the second autoregressive parameter in model 6 is judged unnecessary 13% of the time compared with 29% of the time for the second moving average parameter in model 2. Similarly, the addition of a second seasonal moving average parameter increased the failure rate from 13% in model 6 to 43% in model 7.

Table 5

Failure in Overdifferencing

	CLASS I						CLASS II		CLASS III					
CRITICAL VALUE	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
.90	8	5%	11	7%	43	26%	50	30%	7	4%	9	5%	14	8%
.95	3	2%	6	4%	19	11%	37	22%	3	2%	3	2%	6	4%

Table 6

Failure in Small Parameter

	CLASS I						CLASS II		CLASS III					
CRITICAL VALUE	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
.05	15	9%	49	29%	21	13%	42	25%	12	7%	22	13%	72	43%
.10	26	16%	88	53%	43	26%	73	44%	31	19%	45	28%	114	68%

Table 7

Failure in Correlation

	CLASS I						CLASS II		CLASS III					
CRITICAL VALUE	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
--	--	--	3	2%	86	51%	124	74%	--	--	--	--	--	--

High positive or negative correlations between parameter estimates are undesirable and reflect ambiguity in the estimation situation since a range of parameter combinations result in models with equally good fits. Table 7 shows that only models 2, 3 and 4 fail the correlation test, i.e. the absolute value of at least one of the correlations is ≥ 0.90. The problem is minimal for model 2, and serious for models 3 and 4 where 51% and 74% of the fits had highly correlated parameters. This may be due to overdifferencing in model 3 and the presence of too many parameters in model 4.

3.2 Criterion for Extrapolation of ARIMA Models

This criterion attempts to ensure the quality of the forecasts of the ARIMA models. We require that the average percentage forecast error of the fitted error be below a certain level.

Table 8 shows that six of the seven models are equivalent from the point of view of forecasts, i.e. the number of autoregressive and moving average parameters does not affect the forecast error of the model averaged over all the series. Of course, some models perform better for certain series.

Table 9 shows the average forecast error and standard deviation of the error under two possible outcomes: passing and failing the forecast error criterion. Not only is the failure rate of model 3 higher than that of the other models, but the table shows that when it fails,

Table 8

Failure in Forecast Error

	CLASS I						CLASS II		CLASS III					
CRITICAL VALUE	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
%		%		%		%		%		%		%		%
10	89	53	84	50	101	60	80	48	84	50	85	51	85	51
15	57	34	58	35	69	41	53	32	57	34	56	34	55	33
20	39	23	40	24	51	31	40	24	40	24	40	24	40	24
25	32	19	33	20	43	26	32	19	36	22	14	20	34	20
30	24	14	26	16	35	21	24	14	27	16	27	16	27	16

Table 9

Conditional Mean (M) and Standard Deviation (SD) of the Average Forecast Error

		CLASS I						CLASS II		CLASS III					
Critical Value	Out-come	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
		(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
15%	Pass	7%	4.0	6%	3.9	7%	4.1	6%	3.8	7%	3.9	7%	4.0	7%	3.9
	Fail	35%	22.3	36%	22.5	41%	26.4	36%	21.4	38%	24.5	37%	23.4	37%	23.0

its average forecast error is bigger. The forecast errors of model 3 are increased by its over-differencing. However, when the forecast errors of model 3 pass the criterion, their average is as small as that of the other models.

4. RANKING OF THE MODELS

To rank the models, the eight criteria are used at different acceptance levels. Tables 10 and 11 present the overall and conditional rankings of the models. Table 10 gives the total

Table 10

Overall Ranking of the Models

	2 criteria		8 criteria*		8 criteria*		8 criteria*
	FE ≤ 15%		FE ≤ 15%		FE ≤ 15%		FE ≤ 15%
	$\chi^2 \geq$ 5%		$\chi^2 \geq$ 5%		$\chi^2 \geq$ 5%		$\chi^2 \geq$ 5%
			SP ≤ .10		SP ≤ .05		SP ≤ .05
			OD ≥ .90		OD ≥ .90		OD ≥ .95
Models	% of series that passed	Models	% of series that passed	Models	% of series that passed	Models	% of series that passed
4	52%	1	34%	6	38%	6	39%
7	51%	6	31%	1	37%	1	38%
6	49%	5	23%	2	29%	2	29%
2	48%	2	20%	5	26%	5	28%
1	44%	3	13%	7	25%	7	27%
3	41%	7	11%	3	17%	3	19%
5	32%	4	2%	4	4%	4	5%

*As well as the four criteria listed, the four other criteria mentioned in the text were imposed.

Table 11

Conditional Ranking of the Models

	2 criteria		8 criteria*		8 criteria*		8 criteria*
	FE ≤ 15%		FE ≤ 15%		FE ≤ 15%		FE ≤ 15%
	$\chi^2 \geq$ 5%		$\chi^2 \geq$ 5%		$\chi^2 \geq$ 5%		$\chi^2 \geq$ 5%
			SP ≤ .10		SP ≤ .05		SP ≤ .05
			OD ≥ .90		OD ≥ .90		OD ≥ .95
Models	% of series that passed	Models	% of series that passed	Models	% of series that passed	Models	% of series that passed
4	52%	1	34%	6	38%	6	39%
7	9%	3	6%	3	9%	3	9%
2	1%	6	4%	7	4%	1	4%
3	1%	5	2%	2	3%	4	2%

*As well as the four criteria listed, the four other criteria mentioned in the text were imposed.

success rate of the models. Table 11 gives first the total success rate of the best model; the following models are chosen according to their success with series with which all higher models have failed.

Table 10 shows that:

- when only the chi-square statistic (χ^2) and average forecast error (FE) are used as criteria, models 4 and 7, which have the most parameters, rank at the top.
- on the other hand, the use of all criteria favour the simplest models (models 1 and 6), at all levels of small parameter (SP) and overdifferencing (OD) criteria.
- models 1 and 6 usually rank close together, although model 1 has one less parameter than model 6.
- when model 6 is not first it is a close second.
- the more the criteria are relaxed, the higher the pass ratio is, although the ranking of the models remains about the same.

In table 11 we see that:

- when all criteria are used, models 1 and 6 which ranked first and second in table 10 now rank only first and third.
- second place belongs to model 3. This model, which in table 10 ranked third, fifth and sixth with total success rates of 41%, 13%, 17%, and 19%, here ranks fourth once and second three times. This is because model 3 fits well an important family of series (series with a steep trend) that all other models fit poorly.
- moving average and autoregressive models are not mutually exclusive. These two families of models are complementary and necessary in fitting and forecasting series.
- when we require only that the average forecast error be less than 15% and the chi-square statistic be greater than 5% and nothing else, the combined success rate of models 4, 7, 2 and 3 together is 63%.
- when all the criteria are used, the models chosen are simple and their combined success rate varies between 46% and 54% using the levels of 15% and 5% described just above. The success rate depends on the levels of small parameter and overdifferencing used.

Even though model 1 does not appear in the third column of table 11, it would appear there if the level of forecast error permitted were raised to 20%.

The criteria and levels used in selecting models in figures 1 and 2 are the same as are used in the second column of tables 10 and 11, except that in figure 1 the average forecast error permitted varies between 10% and 99% while in figure 2 the chi-square criteria varies between 10% and 60%.

Figure 1 shows that:

- models 1, 3 and 6 perform the best.
- the ranking of the models tends to remain the same.
- the performance of the first model increases more rapidly than that of the others, going from 23% to 59% compared with an increase from 13% to 17% for model 3. This point needs clarification. Model 1 is chosen according to its unconditional performance, while the other models are chosen according to their conditional ranking.
- the increase in performance of the models according to unconditional ranking is greater than the increase when using conditional ranking.

We see in figure 2 that

- models 1, 3 and 6 are generally the best models for any level of chi-square.
- models 1 and 6 trade places but are not mutually exclusive.

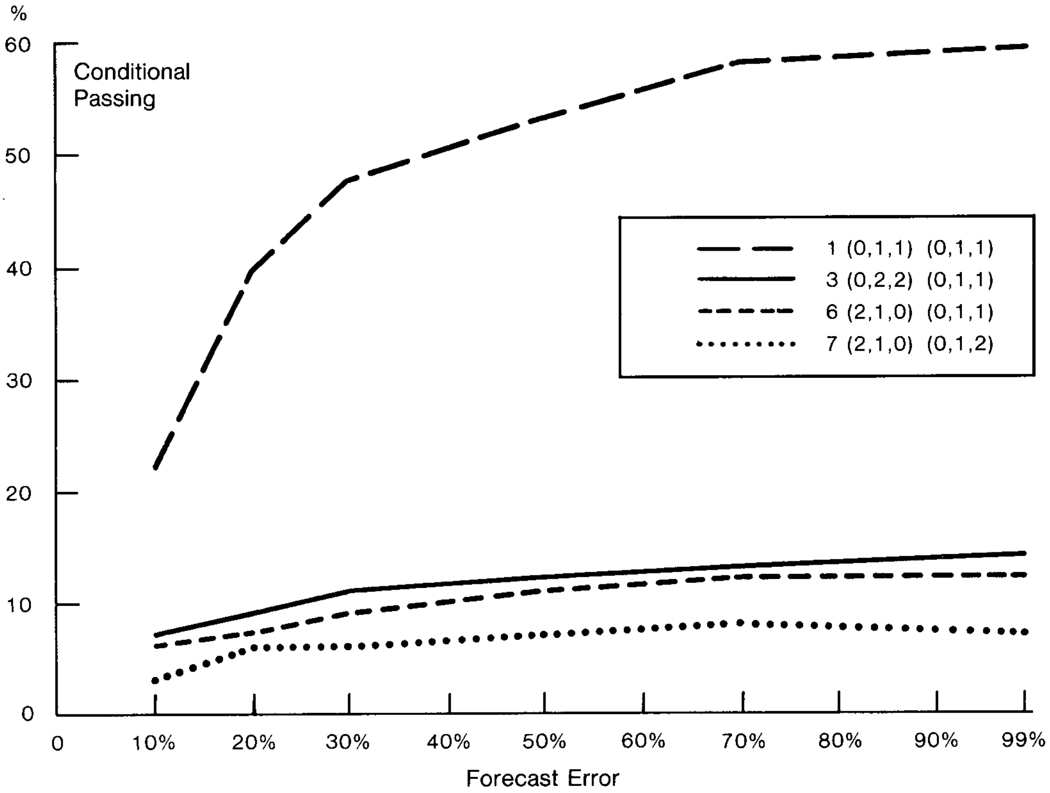

Figure 1. Model Priority Chart for Different Levels of the Forecast Criterion

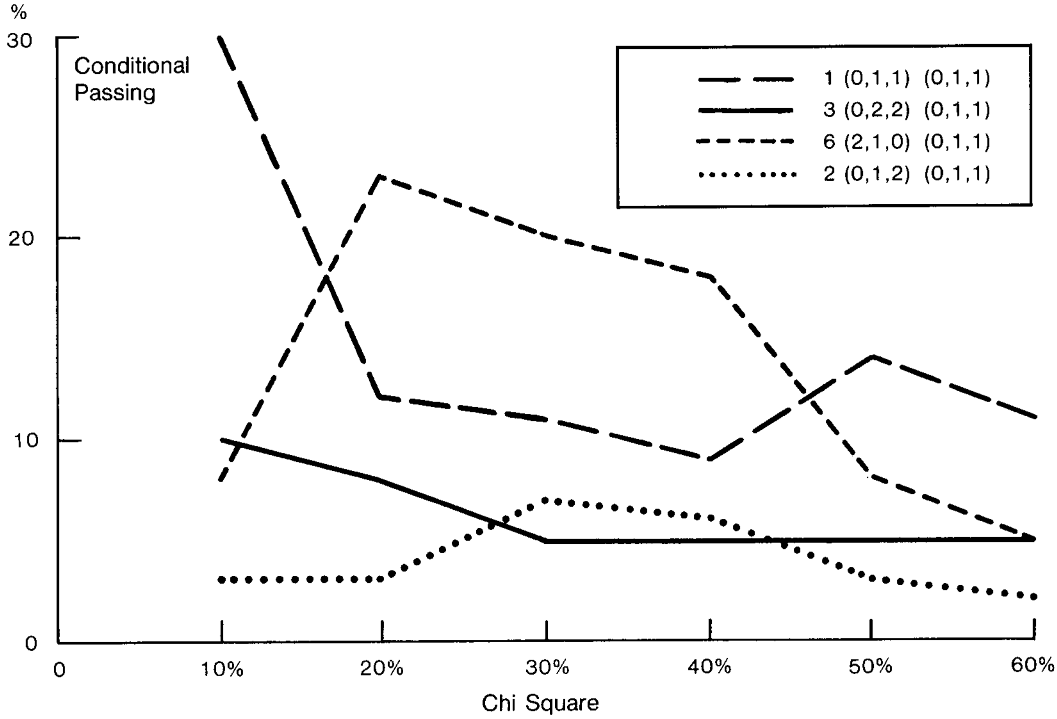

Figure 2. Model Priority Chart for Different Levels of the Chi-Square Criterion

Table 12

Conditional ranking of the ARIMA models for the sectors of the Canadian economy

Sectors	Models ranking and % of series that passed							
	first model	%	second model	%	third model	%	fourth model	%
Labour	1	79	3	14	–	0	–	0
Prices	5	50	7	17	2	8	–	0
Manufacturing	3	19	6	14	1	5	2	5
Fuel, Power and Mining	1	46	6	4	–	0	–	0
Domestic Trade	1	53	6	7	7	7	–	0
External Trade	6	21	–	0	–	0	–	0
Transportation	1	54	5	8	–	0	–	0
Finance	1	32	3	11	–	0	–	0

Table 12 presents the conditional ranking of the ARIMA models for those sectors of the Canadian economy for which we fitted twelve or more series. The criteria and levels used in ranking the models are the same as those used in the second column of tables 10 and 11. We see that

- models 1 and 6 are generally the best performers.
- the combined success rate of the models varies considerably from one sector to another, from 93% in the labour sector to only 21% in external trade.
- this success rate is at least 50% for five sectors. The rate depends on the structure of the series, changes in the structure, and the amount of irregular in the series. The rate is good considering that for two of the last three years Canada suffered a severe recession which strongly affected the structure of the series. The success rate for external trade is always low because those series are very irregular.

5. WITHIN-SAMPLE AND OUT-OF-SAMPLE FORECASTS

The within-sample forecasts are obtained by fitting the models to the entire series in order to estimate the parameters and calculate the forecasts for the last three years. The out-of-sample forecasts do not use information from after the forecast time origin. For each forecast origin, the parameters are re-estimated.

Table 13

Failure Rate in Forecast Error for Within-Sample and Out-of-Sample Forecasts

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
	(0, 1, 1) (0, 1, 1)	(0, 1, 2) (0, 1, 1)	(0, 2, 2) (0, 1, 1)	(2, 1, 2) (0, 1, 1)	(1, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 2)
	%	%	%	%	%	%	%
Within-sample	34	35	41	32	34	34	33
Out-of-sample	31	32	42	33	31	32	31

Table 14

Conditional and Unconditional Ranking of the Models

Unconditional ranking		Conditional Ranking	
Models	% of series that passed	Models	% of series that passed
1	40%	1	40%
6	28%	2	5%
5	27%	7	4%
2	20%	3	3%
3	14%		
7	10%		
4	2%		

Table 13 shows the rate of failure in forecast error at the 15% level for within-sample and out-of-sample forecasts. The difference between the two is small and is well within one standard deviation for each model. The X-11-ARIMA seasonal adjustment program uses within-sample forecasts because they cost less.

Table 14 has been prepared using the same criteria and levels as were used in the second columns of tables 10 and 11. The unconditional ranking is exactly the same as that in the second column of table 10. Only the success rates of the first three models differ, and in table 14, model 1 is clearly superior to the other models. However, the conditional ranking is different from that appearing in the second column of table 11.

The conditional rankings in tables 11 and 14 differ for two reasons. First, of course, table 14 uses out-of-sample forecasts. Another important reason is that the calculation of the seven other criteria was based on one year less data, and the missing year contained a severe recession. Thus the structure of the series and the choice of models is markedly different.

It appears therefore that the conditional ranking of the models for both within-sample and out-of-sample forecasts depends on the phase of the business or economic cycle in which the series ends.

6. CONCLUSION

Our objective was to rank a set of seven ARIMA models according to their fitting and forecasting of a large sample of time series.

- when only the chi-square statistic and the average forecast error are used as criteria, models 4 and 7 rank at the top.
- The use of all eight criteria favours the simplest models (1 and 6) and model 3.
- Models 1 (moving average model) and 6 (autoregressive model) rank close together in unconditional ranking, although model 1 has one less parameter than model 6.
- In conditional ranking, these two both rank highly but are not mutually exclusive. That is, moving average and autoregressive models are complementary and both are necessary in fitting and forecasting series.
- Although Model 3 ranks near the bottom, it fits well an important family of series (series with a steep trend) that all other models fit poorly.
- The nonparsimonious models (numbers 4 and 7) have a combined success rate of 61% compared to a success rate that varies between 44% and 52% for parsimonious models 1, 6 and 3.

- The combined success rate of the models varies considerably from one economic sector to another, from 93% in the labour sector to only 21% in external trade. This rate depends on the structure of the series, changes in the structure, and the amount of irregular in the series.
- It appears that the conditional ranking of the models for both within-sample and out-of-sample forecasts depends on the phase of the business or economic cycle in which the series ends.

ACKNOWLEDGEMENT

We acknowledge helpful discussions with Mr. Normand Laniel and we are very grateful to Ms Helen Lim and Mr. Alfred Papineau for their valuable computational assistance, and to Ms B. Cohen for her typing assistance.

REFERENCES

BOX, G.E.P., and JENKINS, G.M. (1970). *Times Series Analysis Forecasting and Control.* Holden Day: San Francisco.

BOX, G.E.P., and PIERCE, D.A. (1970). Distribution of residual autocorrelations in autoregressive integrated moving average time series models. *Journal of the American Statistical Association,* 65, 1509-1526.

DAGUM, E.B. (1980). *The X-11-ARIMA Seasonal Adjustment Method.* Catalogue No. 12-564E, Statistics Canada, Ottawa.

DRAPER, N.R., and SMITH, H. (1981). *Applied Regression Analysis.* John Wiley & Sons, Inc.

HIGGINSON, J. (1976). A test for the presence of seasonality and a model test. Research Paper, Time Series Research and Analysis Division. Statistics Canada, Ottawa.

LJUNG, G.M., and BOX, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika,* 65, 297-307.

PANDIT, S.M., and WU, S.M. (1983). *Time Series and System Analysis with Applications.* John Wiley & Sons, Inc.

PLOSSER, C.I., and SCHWERT, G.W. (1977). Estimation of a non-invertible moving average process. *Journal of Econometrics,* 6, 199-224.

PROTHERO, D.L., and WALLIS, K.F. (1976). Modelling macroeconomic time series (with discussion). *Journal of the Royal Statistical Society,* AL39, 468-500.