The most widely employed procedure for interrupted time series analysis consists of a two-step procedure: (1) determining the ARIMA model by examining the pattern of autocorrelations and partial autocorrelations; and (2) employing a general linear model solution after the effect of dependency has been removed. In order to determine the reliability and accuracy of model identification, 12 extensively trained subjects were each asked to identify 32 different computer generated time series. Six commonly occurring models were employed with different levels of dependency (high, medium, or low) and different numbers of data points (N=40 and N=100). The overall accuracy, 28%, was affected by the number of data points, the type of model, and the degree of dependency.

THE RELIABILITY AND ACCURACY OF TIME SERIES MODEL IDENTIFICATION

WAYNE F. VELICER
JOHN HARROP
University of Rhode Island

Methods of statistical analysis have been developed recently for the situations in which multiple observations are made on a single unit (or subject) across time. In the simplest form, a series of observations are gathered, an intervention occurs, and a second series of observations are gathered. The purpose of the analysis is to determine if the intervention resulted in a significant change in the level and/or slope of the series. Variations on this basic design permit the testing of a variety of alternative designs and potential intervention effects (Glass et al., 1975; Cook and Campbell, 1979). These procedures are especially appropriate for applied research in the social and behavioral sciences.

The most widely employed procedure is the Glass et al. (1975) approach that involves first identifying the appropriate Autoregressive

AUTHOR'S NOTE: *This work was partially supported by Grant CA 27821 from the National Cancer Institute. Send requests for reprints or further information to Wayne F. Velicer, Psychology Department, University of Rhode Island, Kingston, Rhode Island 02881.*

EVALUATION REVIEW, Vol. 7 No. 4, August 1983 551-560

Integrated Moving Averages (ARIMA) model from inspection of the autocorrelations and partial autocorrelations, and then employing a general linear model solution that removes the dependency. The general ARIMA model includes three parameters: p, the order of the autoregressive component; d, the degree of differencing required to produce a stable series; and q, the order of the moving averages component. The model identification step is also sometimes described as "model building" (Granger and Newbold, 1977; McCleary et al., 1980) or "formulation" (Nerlove et al., 1979). Glass et al. (1975) emphasize the necessity of correctly identifying the model and suggest a minimum of 50 observations before and after an intervention for accurate identification. Requiring a minimum of 100 observations, if necessary, severly limits the application of this technique in applied settings in which gathering that many data points is often prohibitive. Prior to this study, however, no one had systematically investigated the reliability and accuracy of model identification for either the recommended number or a more practical lower number.

The recent debate on analysis of the effect of the Massachusetts Gun Control Law (Deutsch and Alt, 1977; Hay and McCleary, 1979; and Deutsch, 1979) demonstrates the potential controversy in this area. Three different variables were analyzed by Deutsch and Alt (1977) and subsequently by Hay and McCleary (1979): homicides, gun assaults, and armed robberies. For each of the variables, Deutsch and Alt include a difference parameter—either in the seasonal component of the model or the regular model when no seasonal component was needed—while Hay and McCleary argue that this parameter is inappropriate. The choice of model had no impact on the analysis of homicides or gun assaults. However, the choice of model leads to different conclusions with respect to armed robberies: Deutsch and Alt found a significant change, but Hay and McCleary did not.

The purpose of the present study was to systematically (using computer generated series) vary the number of observations and type of model in order to determine the effects this would have on the reliability and accuracy of model identification. A comparison between the recommended number of observations (n=100) and a lesser number (n=40) was of primary interest. If the reliability and accuracy are comparable, then the potential for application of this procedure is

increased. Of secondary interest was determining whether the type of model would affect the results.

METHOD

Twelve graduate students who were taking or had taken a course in time series analysis using Glass et al. (1975) as a text were divided into two groups of six each. They were each given 32 ARIMA series to identify from the autocorrelations and parital autocorrelations computed at various lags. The series were generated by a computer program in accordance with predetermined ARIMA models and parameters. The program employed in the data generation was adapted from Padia (1975). The two groups had the same series to identify except that they were generated using a different set of random numbers. All subjects had covered the text book material dealing with model identification, had attended class lectures and demonstrations, and had also attempted at least two in-class examples selected from the text book. The extent of the subjects' training was, therefore, greater than typically would be found for an applied researcher relying only on published materials.

The ARIMA models that were simulated were the (0,0,0), (0,0,1), (1,0,0), (0,1,0), (0,1,1), and (2,0,0) with ϕ or θ values as appropriate to the model of .8, .4, .2, –.4, and –.8. The first five models were chosen because they were found to be the most commonly encountered in the social and behavioral sciences (Glass et al., 1975: 115-118) and the last was included so that the raters were not always dealing with first order models. The values of ϕ and θ were chosen to be representative of the range of dependency that might be encountered. For each of these 16 models a series was generated with either 40 points or 100 points assuming an intervention takes the place at midpoint ($n_1 = n_2 = 20$ or $n_1 = n_2 = 50$). The following population parameters were employed for all series generated: preintervention level = 0.0, postintervention change in level = 1.0, error mean = 0.0, and error variance = 1.00. Table 1 presents the design of the study.

Each subject was instructed to identify the model, that is (p,d,q), in accordance with Glass et al. (1975) using the autocorrelations and partial autocorrelations provided for the preintervention and postinter-

TABLE 1
Design of Study

Model	Dependency		Rating Group[1]	
	Lag 1	*Lag 2*	*1*	*2*
(0,0,0)	0	0	40,100	40,100
(0,0,1)	.8	0	40,100	40,100
(0,0,1)	.4	0	40,100	40,100
(0,0,1)	−.4	0	40,100	40,100
(0,0,1)	−.8	0	40,100	40,100
(0,1,0)	0	0	40,100	40,100
(1,0,0)	.8	0	40,100	40,100
(1,0,0)	.4	0	40,100	40,100
(1,0,0)	−.4	0	40,100	40,100
(1,0,0)	−.8	0	40,100	40,100
(0,1,1)	.8	0	40,100	40,100
(0,1,1)	.4	0	40,100	40,100
(0,1,1)	−.4	0	40,100	40,100
(0,1,1)	−.8	0	40,100	40,100
(2,0,0)	.8	.2	40	100
(2,0,0)	.4	.4	100	40
(2,0,0)	−.4	−.4	40	100
(2,0,0)	.2	.8	100	40

1. Entries indicate sample size of series.

vention data points. The CORREL program (Glass and Bower, 1974) was employed to calculate autocorrelations and partial autocorrelations. They were not provided with knowledge of what models and degrees of dependency they might expect.

RESULTS

Table 2 presents the accuracy (percentage correct) for the six models and for the two levels of number of points, averaged across the 12 subjects and different degrees of dependency. The overall accuracy of correct identification was 28% (109 of 384). Accuracy was generally higher for the series with more observations (36% for N = 100 versus 20% for N = 40). Accuracy was generally higher for simple models [(0,0,0), (0,0,1), (1,0,0)] than for more complex models [(0,1,0), (0,1,1), (2,0,0)].

Table 3 presents the accuracy (percentage correct) for the six models and for the three levels of dependency averaged across the number of data

TABLE 2
Accuracy (Percentage Correct) of Model Identification Classified by Type of Model and Number of Observations[1]

Type of Model	*Number of Observations*		
	N = 40	*N = 100*	*Total*
(0,0,0)	42 (12)	42 (12)	42
(0,1,0)	0 (12)	8 (12)	4
(0,0,1)	46 (48)	67 (48)	56
(0,1,1)	6 (48)	19 (48)	13
(1,0,0)	19 (48)	46 (48)	32
(2,0,0)	0 (24)	4 (24)	2
Total	20 (192)	36 (192)	28 (384)

1. Total possible number of identifications given in parentheses.

points and the 12 subjects. Accuracy was generally higher for models with high dependency (44%) as compared to models with medium or zero dependency (24% and 23%, respectively).

Table 4 presents the cross-classification of model identification (frequencies) for the correct model and model identification by the subjects. In general, the errors did not consistently involve a single model but were dispersed across a variety of models. Several observations, however, may be made. The (0,0,1) examples were never misidentified as (1,0,0) models. The (1,0,0) examples were, however, misidentified as (0,0,1) models as frequently as they were correctly identified. The (0,1,0) models were frequently misidentified as (1,0,0) models. The category (0,1,0) was seldom employed in any identification while the (0,0,0) and (0,0,1) classifications were the only ones overemployed.

The accuracy rate of the subjects was also investigated to determine if significant individual differences existed. The range of correct identification between the 12 subjects was from 12.5% to 44%. A test of

TABLE 3
Accuracy (Percentage Correct) of Model Identification Classified by Degree of Dependency and Type of Model[1]

Type of Model	*Degree of Dependency*			
	None	*Medium (±.40)*	*High (±.80)*	*Total*
(0,0,0)	42 (24)	–	–	42
(0,1,0)	4 (24)	–	–	4
(0,0,1)	–	52 (48)	60 (48)	56
(0,1,1)	–	4 (48)	21 (48)	13
(1,0,0)	–	15 (48)	50 (48)	32
(2,0,0)[2]	–	–	–	–
Total	23 (48)	24 (144)	44 (144)	32 (336)

1. Total possible number of identifications given in parentheses.
2. ARIMA (2,0,0) models could not be classified by degree of dependency because of the presence of two different dependency parameters.

proportions found no significant differences ($p < .05$) between the 12 subjects, $\chi(11) = 15.82$.

DISCUSSION

The most surprising result of this study was the disappointingly low overall accuracy rate, 28%. Even when we restrict our sample to the recommended number of data points (N=100), the accuracy rate was still only 36%.

One possible explanation for these results is that the training received by the raters was inadequate. We believe, however, that the training received by our raters was more extensive than typically would be the case for an independent researcher who is seeking to master and employ

TABLE 4
Cross Classification of Correct Model and Model Identified by Subjects (Frequencies)

Correct Model	*Model Identified by Subject*						
	(0,0,0)	*(0,1,0)*	*(0,0,1)*	*(0,1,1)*	*(1,0,0)*	*(2,0,0)*	*Other*
(0,0,0)	10	–	1	7	–	–	6
(0,1,0)	1	1	6	–	11	–	5
(0,0,1)	14	–	54	12	–	4	12
(0,1,1)	17	–	18	12	23	12	14
(1,0,0)	9	–	31	5	31	2	18
(2,0,0)	6	–	6	6	11	1	18
Totals	57	1	116	42	76	19	73

these procedures without classroom instruction and practice. The absence of individual differences between the raters can also be viewed as evidence against this explanation. A second possible explanation of the results is that even the recommended number of data points (N=100) is inadequate. However, since the requirement of 100 observations is considered prohibitive for many applied research problems, an increase in the required number of observations is not feasible. Therefore, we conclude that the model identification step in time series analysis must be viewed as problematic at best.

The other findings of the present study were more consistent with the expected pattern of results. First, the increase from 40 to 100 observations did improve the accuracy of identification. Second, model identification was less accurate for zero dependency or medium dependency models than for high dependency models. It should be noted, however, that poor model identification is less of a concern for low dependency levels than for high dependency levels since the mathematical distinctions between the models are minor for near-zero dependency. Third, accuracy was generally higher for simple models compared to more complex models. Since Glass et al. (1975) report a relatively low frequency of models with higher order terms, the relatively poor accuracy for these models may not be a severe problem.

Two possible patterns of model misidentification are of theoretical interest. First, it can be demonstrated algebraically that a low order series of one type can be represented by a high order series of another type—that is, a (1, 0, 0) model could be represented by a (0, 0, K) model

in which the value of K needed for adequate representation is determined by the degree of dependency. Therefore, the identification of (1, 0, 0) models as (0, 0, K) models and the identification of (0, 0, 1) models as (K, 0, 0) models would result in no error in the subsequent analysis. An examination of Table 4 indicates that this was not the pattern of misidentification. Second, Glass et al. (1975) indicate that the most critical parameter in model identification is the d parameter. Therefore, if the misidentification involves only the p and q parameters, the problem is not as severe. An examination of Table 4 shows that the subjects had severe problems both in detecting differencing where it was needed and in indicating that differencing was needed when it was not.

On the basis of the present study, there are two alternatives that need further investigation. Either the method of model identification must be improved, or alternative methods of time series analysis that bypass the model identification step deserve further consideration. McCleary et al. (1980) and Nerlove et al. (1979) have recently described more extensive procedures for model identification. An evaluation of this procedures accuracy is not available. Three different approaches to time series analysis that do not require model identification have been proposed. Simonton (1977) proposed that a (1,0,0) model be assumed for all applications. The reasonability of this assumption and the effect of violations need to be explored. Algina and Swaminathan (1979) have proposed an alternative to the Simonton (1977) approach that employs the sample variance-covariance matrix as an estimator in a modified least-squares solution. This approach, however, requires, that the number of subjects be greater than the number of observations per subject, a requirement that precludes the application of this method in a large number of potential situations. Velicer and McDonald (forthcoming) have proposed employing a general transformation matrix to provide an approximate solution for all commonly occurring models. This procedure essentially involves a higher order solution for all analysis, followed by a test of the residuals. The accuracy of this approximation has not been systematically evaluated. These three approaches do not require model identification as a first step in time series analysis.

The most basic question at the present time is whether model misidentification will have a major impact on the subsequent estimation and testing of intervention effects. This question is important for the

Glass et al. (1975) approach, since the accuracy of model identification can no longer be assumed, and—in light of the results of this study—is in fact unlikely. It is also an important issue for two of the proposed alternatives, Simonton (1977) and Velicer and McDonald (forthcoming). Glass et al. (1975) suggest that misidentification is not a severe problem, but this critical issue clearly deserves extensive investigation.

REFERENCES

ALGINA, J. and H. A. SWAMINATHAN (1979) "Alternatives to Simonton's analysis of the interrupted and multiple-group time series designs." Psych. Bull. 86: 919-926.

COOK, T. D. and D. T. CAMPBELL (1979) Quasi-Experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally.

DEUTSCH, S. J. and F. B. ALT (1977) "The effect of Massachusetts' gun control law on gun-related crimes in the city of Boston." Evaluation Q. 1: 543-568.

DEUTSCH, S. J. (1979) "Lies, damn lies and statistics: a rejoinder to the comment by Hay and McCleary." Evaluation Q. 3: 315-328.

GLASS, G. V and C. BOWER (1974) Computer program CORREL. Boulder: University of Colorado.

GLASS, G. V, V. L. WILLSON, and J. M. GOTTMAN (1975) Design and Analysis of Time Series Experiments. Boulder: Colorado Assoc. Univ. Press.

GRANGER, C.W.J. and P. NEWBOLD (1977) Forecasting Economic Time Series. New York: Academic Press.

HAY, R. A., Jr. and R. McCLEARY (1979) "Box-Tiao time series model for impact assessment: a comment on the recent work of Deutsch and Alt." Evaluation Q. 3: 277-314.

McCLEARY, R., R. A. HAY, Jr., E. E. MEIDINGER, and D. McDOWALL (1980) Applied Time Series Analysis for the Social Sciences. Beverly Hills, CA: Sage.

NERLOVE, M., D. M. GRETHER, and J. L. CARVALHO (1979) Analysis of Economic Time Series. New York: Academic Press.

PADIA, W. L. (1975) "The consequences of model misidentification in the interrupted time-series experiment." Ph.D. dissertation, University of Colorado.

SIMONTON, D. K. (1977) "Cross-sectional time-series experiments: some suggested statistical analysis." Psych. Bull. 84: 489-502.

VELICER, W. F. and R. P. McDONALD (forthcoming) "Time series analysis without model identification." Multivariate Behavioral Research.

Wayne F. Velicer is a Professor of Psychology at the University of Rhode Island. His previous research has involved principal component and factor analysis, personality assessment, prediction models, and school psychology.

John Harrop is currently working on his dissertation for the Ph.D. degree in Clinical Psychology. Formerly he worked as an electrical engineer for the Naval Underwater Systems Center in Newport, Rhode Island.