Simulation Procedures for Box-Jenkins Models

ANGUS IAN MCLEOD

Statistics and Actuarial Science Group, University of Western Ontario London, Ontario, Canada N6A 5B9

KEITH WILLIAM HIPEL

Department of Systems Design, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

New simulation procedures are presented for generating synthetic data from either a nonseasonal or a seasonal Box-Jenkins model. The simulation techniques are designed so that random realizations of the underlying stochastic process are employed for starting values. Because fixed beginning values are not utilized, systematic bias is not introduced into the synthetic trace. When the data have been transformed by a Box-Cox transformation, the inverse transformation can be conveniently incorporated into the simulation process. Also a new algorithm is presented for simulating integrated models. A method is developed for incorporating parameter uncertainty into simulation studies, and it is explained how this technique can be used in the design of reservoirs. Practical applications are presented to demonstrate the efficacy of the aforesaid simulation methods. In addition, the Fortran computer subroutines for the simulation procedures are included in an appendix with the microfiche edition of this paper.

INTRODUCTION

The Box-Jenkins family of stochastic models [*Box and Jenkins*, 1970] constitutes a flexible class of models for describing natural, economic, and other types of time series. In fitting a Box-Jenkins model to a given series, adherence to the identification, estimation, and diagnostic check stages of model construction is recommended. Recently, *Hipel et al.* [1977] have presented some new procedures to simplify and also substantiate the three stages of model construction. *McLeod et al.* [1977] have demonstrated the utility of the contemporary modeling techniques given by *Hipel et al.* [1977] by applying these procedures to both nonseasonal and seasonal time series.

Once a Box-Jenkins model has been properly fit to a data set, the chosen model can be used for applications such as forecasting and simulation. The purpose of this paper is to present proper simulation procedures for both nonseasonal and seasonal Box-Jenkins models.

Simulation is now a widely accepted technique to aid in both the design and the operation of water resources systems. Although synthetic data generation is now extensively utilized, there are still some chronic problems that require proper resolution. For instance, many current simulation methods that are widely accepted do not use correct initial values. Although the effect of starting values is transitory, it could cause systematic bias in a simulation study, and therefore, as was pointed out by *Moran* [1959, chapter 5] and *Copas* [1966], the choice of initial values is important. To attempt to overcome this problem, some researchers discard the first section of a synthetic time series supposedly to get rid of the effects of initial values. However, exactly how many values of the generated series should be rejected, and how much computer time is wasted by generating data that are not used?

As an example of a conservative approach to the effect of starting values, consider the simulation study of *Brown and Hardin* [1973]. These authors used deterministic starting values for an autoregressive process of order 2 and then generated a series with a length of 30,000 values. The first 15,000 values of the synthetic trace were discarded supposedly to nullify the effects of using nonrandom initial values.

Paper number 8W0458.
0043-1397/78/058W-0458$01.00

The simulation procedures given in this paper do not require fixed starting values. They are designed in a manner such that random realizations of the underlying stochastic process are used as initial values. Therefore the results of a simulation study are not significantly biased, and it is not necessary to disregard any of the generated data.

Often it is necessary to generate k' time series of length k. Some researchers resort to producing a single synthetic series of length $k'k$ and then splitting this long series into k' series of length k. If any serial correlation is present, then the results of any simulation study will be biased by this rather crude procedure. To overcome this problem, the authors recommend generating k' separate time series of length k. If the generating procedures given in this paper are adopted, then each time another series of length k is obtained, new random realizations of the stochastic process are used as starting values.

Previously, researchers had attempted to devise methods for incorporating uncertainty of the model parameters into a simulation study. It is shown that the procedure suggested by *Vicens et al.* [1975] is not a satisfactory method to employ. Consequently, in this paper a new algorithm is developed which properly handles parameter uncertainty in a simulation experiment. In addition, comments are made regarding the problem of selecting a proper model to fit to a given data set. After the appropriate model has been determined, the model can be utilized for simulating synthetic data.

In the next section, simulation procedures are presented for generating synthetic traces from Box-Jenkins models. It is clearly explained in which situations a specific simulation method should be employed. Because it is a straightforward undertaking to extend the simulation techniques to the seasonal case once the nonseasonal methods are known, only the simulation results for nonseasonal Box-Jenkins models are given. Furthermore, the techniques of simulation are also presented for models containing both nonseasonal and seasonal differencing operators and for data that have been transformed by a Box-Cox transformation. Methods for tackling the problems of both parameter and model uncertainty are discussed. Following this, three practical applications are given to portray the effectiveness of the new simulation techniques. The computer programs for the simulation methods

are listed in the appendix.[1] (*Farebrother and Berry* [1974], *Hannan* [1970], *Healy* [1968], *Knuth* [1969]; *Nicholls* [1972], and *Pagano* [1973] are cited in the appendix.)

Simulation Techniques

Waterloo Simulation Procedure 1 (Wasim 1)

A stationary nonseasonal Box-Jenkins model with a mean of zero can be written in the form

$$\phi(B)z_t = \theta(B)a_t \tag{1}$$

where t is discrete time that is spaced at equal time intervals; z_t is the value of the process at time t; B is the backward shift operator defined by $Bz_t = z_{t-1}$ and $B^s z_t = z_{t-s}$, where s is a positive integer; $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is the nonseasonal autoregressive (AR) operator or polynomial of order p such that the roots of the characteristic equation $\phi(B) = 0$ lie outside the unit circle for nonseasonal stationarity and the ϕ_i, $i = 1, 2, \cdots, p$, are the nonseasonal AR parameters; $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ is the nonseasonal moving average (MA) operator or polynomial of order q such that the roots of $\theta(B) = 0$ lie outside the unit circle for invertibility and θ_i, $i = 1, 2, \cdots, q$, are the nonseasonal MA parameters; and a_t are identically independently distributed residuals with mean 0 and variance σ_a^2 (IID $(0, \sigma_a^2)$); often the residuals are assumed to be normally independently distributed (NID $(0, \sigma_a^2)$). The process given in (1) is referred to as an autoregressive moving average (Arma) model. The notation (p, q) is used to indicate the orders of the AR and MA operators.

For simulation purposes the zero mean stationary seasonal Arma model can be considered as a natural extension of the nonseasonal process given in (1). Models with a nonzero mean (or any other type of deterministic component) are simulated by first generating the corresponding zero mean process and then adding on the mean component.

Suppose that the z_t are expanded in terms of a pure MA process. This is termed the random shock form of an Arma process and is written as

$$z_t = \frac{\theta(B)}{\phi(B)} a_t = \psi(B)a_t = (1 + \psi_1 B + \psi_2 B^2 + \cdots)a_t \tag{2}$$

where $\psi_0 = 1$. The coefficients ψ_i $(i = 1, 2, \cdots)$ of the random shock operator $\psi(B)$ are obtained by equating coefficients in the operator identity [*Box and Jenkins*, 1970, chapter 4]:

$$\phi(B)\psi(B) = \theta(B) \tag{3}$$

If an AR operator is present, $\psi(B)$ forms an infinite series and therefore must be approximated by the finite series

$$\psi(B) \simeq 1 + \psi_1 B + \psi_2 B^2 + \cdots + \psi_{q'} B^{q'} \tag{4}$$

It is necessary to choose q' such that $\psi_{q'+1}, \psi_{q'+2}, \cdots$ are all negligible. Since the model is stationary, this can be accomplished by selecting q' sufficiently large that the error given below is kept as small as desired:

$$\gamma_0 - \sum_{i=0}^{q'} \psi_i^2 < \text{error} \tag{5}$$

where γ_0 is the theoretical variance of a given Arma process with $\sigma_a^2 = 1$ and is calculated by using the algorithm of *McLeod* [1975] and error is the chosen error level (for example, error $= 10^{-5}$).

To obtain a synthetic series of k observations, first generate $k + q'$ white noise terms $a_{-q+1'}, a_{-q'+2'}, \cdots, a_0, a_1, a_2, \cdots, a_k$. Next, calculate

$$z_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots + \psi_{q'} a_{t-q'} \tag{6}$$

where $t = 1, 2, \cdots, r$ and $r = \max(p, q)$. The remaining z_t are easily determined from the equation

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{7}$$

where $t = r + 1, r + 2, \cdots, k$.

The use of (7) avoids the truncation error present in (4). Nevertheless, if an AR operator is present (i.e., $p > 0$), there will be some systematic error in the simulated data due to the approximation involved in (6). However, this bias can be kept to a tolerable level by selecting the 'error' term in (5) to have a specified minimum value. Of course, if the model is pure MA $(0, q)$, then set $q' = q$, and (6) will be exact and can be utilized to generate all of the synthetic data.

An inherent advantage of the Wasim 1 simulation technique is that the only restriction on the white noise terms is that they are IID $(0, \sigma_a^2)$. Although in many situations it is often appropriate to employ NID $(0, \sigma_a^2)$ innovations, this simulation method does not preclude considering other types of distributions. For instance, after a relatively long hydrological time series has been modeled, the residuals from the historical data could be used to form an empirical distribution function for generating the white noise. This approach is illustrated in the applications section. In other situations it may be warranted to simulate the white noise by employing Johnson variates [*Hill et al.*, 1976; *Hill*, 1976] or perhaps one of the distributions suggested by *Delleur et al.* [1976, pp. 961–963]. *Atkinson and Pearce* [1976] discuss the computer generation of beta, gamma, and normal random variables. For generating normally distributed disturbances it is recommended that the method of *Marsaglia and Bray* [1964] be employed.

Waterloo Simulation Procedure 2 (Wasim 2)

Suppose that it is necessary to generate k terms of an Arma (p, q) model with innovations that are NID $(0, \sigma_a^2)$. The following simulation procedure is exact to simulate $z_1, z_2, \cdots, z_k$ for all stationary Arma (p, q) processes.

1. Obtain the theoretical autocovariance function γ_j for $j = 0, 1, \cdots, p - 1$ by using the algorithm of *McLeod* [1975] with $\sigma_a^2 = 1$.
2. Utilize (3) to determine the coefficients ψ_j for $j = 1, 2, \cdots, p - 1$.
3. Form the covariance matrix $\Delta\sigma_a^2$ of $z_p, z_{p-1}, \cdots, z_1, a_p, a_{p-1}, \cdots, a_{p-q+1}$.

$$\Delta = \begin{vmatrix} (\gamma_{i-j})_{p\times p} & (\psi_{j-i})_{p\times q} \\ (\psi_{i-j})_{q\times p} & (\delta_{i,j})_{q\times q} \end{vmatrix}_{(p+q)\times(p+q)} \tag{8}$$

In (8) the (i, j) element and dimension of each partitioned matrix are indicated. The values of $\delta_{i,j}$ are 1 or 0 according to whether $i = j$ or $i \neq j$, respectively. When $i - j < 0$, then $\gamma_{i-j} = \gamma_{j-i}$ and $\psi_{i-j} = 0$.

[1] Appendix is available with entire article on microfiche. Order from American Geophysical Union, 1909 K Street, N.W., Washington, D. C. 20006. Document W78-006; $1.00. Payment must accompany order.

4. Determine the lower triangular matrix **M** by Cholesky decomposition [*Ralston*, 1965, p. 410] such that

$$\mathbf{\Delta} = \mathbf{MM}' \tag{9}$$

5. Generate $e_1, e_2, \cdots, e_{p+q}$ and $a_{p+1}, a_{p+2}, \cdots, a_k$, where the e_t and a_t sequences are NID $(0, \sigma_a^2)$.

6. Calculate $z_1, z_2, \cdots, z_p$ from

$$z_{p+1-t} = \sum_{j=1}^{t} m_{t,j} e_j \qquad t = 1, 2, \cdots, p \tag{10}$$

7. Determine $a_{p-q+1}, a_{p-q+2}, \cdots, a_p$ from

$$a_{p+1-t} = \sum_{j=1}^{p+t} m_{t+p,j} e_j \qquad t = 1, 2, \cdots, q \tag{11}$$

8. Obtain $z_{p+1}, z_{p+2}, \cdots, z_k$ by using

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \qquad t = p+1, p+2, \cdots, k \tag{12}$$

9. If another series of length k is required, then return to step 5.

For a particular Arma model it is only necessary to calculate the matrix **M** once, no matter how many simulated series are synthesized. Therefore Wasim 2 is economical with respect to computer time required, especially when many time series of the same length are generated.

Often the white noise disturbances can be assumed to be NID $(0, \sigma_a^2)$, and it is desirable to have as much accuracy as possible in order to eliminate bias. For this situation the authors recommend using Wasim 2 for a pure AR model or an Arma process. When a pure MA process with innovations that are NID $(0, \sigma_a^2)$ is being simulated, the Wasim 1 and Wasim 2 procedures are identical.

Simulation of Integrated Models

For annual geophysical time series of a moderate length (perhaps a few hundred years) it is often reasonable to assume that a stationary model can adequately model the data. *Hipel and McLeod* [1978], for example, fit stationary Arma models to 23 time series which are measured from six different geophysical phenomena. Nevertheless, certain types of time series that are used in water resources could be nonstationary. The average annual cost of hydroelectric power and the total annual usage of water-related recreational facilities constitute two types of measurable processes which possess mean levels and variances that could change significantly over time. In general, time series that reflect the socioeconomic aspects of water resources planning may often be nonstationary even over a short time span. Consequently, in certain situations it may be appropriate to incorporate a nonseasonal differencing operator into the nonseasonal model in order to account for the nonstationarity.

If a Box-Jenkins seasonal model is fit to seasonal data, usually both nonseasonal differencing and seasonal differencing are required to account for the nonstationarity [*Hipel*, 1975]. Consider the case of average monthly observations. If the monthly mean and perhaps variance change from one year to the next for each specific month, then fitting a nonstationary seasonal Box-Jenkins to the data may prove to be reasonable. For example, the average monthly water demand for large cities tends to increase from year to year for each month. For the aforementioned situation the simulation procedures for integrated models could be useful.

When seasonal geophysical data, such as average monthly river flows, are being considered, the individual monthly averages may have constant mean values, but the means vary from month to month. Consequently, the time series of all the given data is by definition nonstationary, but it still may not be appropriate to employ a nonstationary seasonal Box-Jenkins process to model the data. Rather, the given natural time series is deseasonalized to produce a stationary nonseasonal data set, and subsequently, a nonseasonal model is fit to the deseasonalized data. For example, prior to fitting a nonseasonal Arma model to the data it is a common procedure to standardize average monthly river flow time series to eliminate seasonality (see, for example, *Hipel* [1975, chapter 8], *Tao and Delleur* [1976], and *McKerchar and Delleur* [1974]). *Clarke* [1973] and *Croley and Rao* [1977] present extensive descriptions of deseasonalization procedures for daily, weekly, and monthly data.

Although caution should be exercised when nonstationary data are being modeled, it is evident that situations may arise during which it is suitable to invoke differencing. Any Box-Jenkins model that contains a differencing operator is termed an integrated model. Suppose that it is necessary to simulate k values of z_t by using an integrated process. A stationary w_t series is related to the nonstationary z_t series by the equation

$$w_t = \nabla^d \nabla_s^D z_t \qquad t = d' + 1, d' + 2, \cdots, k \tag{13}$$

where s is the seasonal length ($s = 12$ for monthly data); $\nabla^d = (1 - B)^d$ is the nonseasonal differencing operator of order d to produce nonseasonal stationarity of the dth differences, usually $d = 0$, 1, or 2; $\nabla_s^D = (1 - B^s)^D$ is the seasonal differencing operator of order D to produce seasonal stationarity of the Dth differenced data, usually $D = 0$, 1, or 2 (for nonseasonal data, $D = 0$); and $d' = d + sD$.

Because of the differencing in (13) the d' initial values $z_1, z_2, \cdots, z_{d'}$, which determine the 'current level' of the process, are assumed known. Given the d' initial values, the time series integration algorithm forms the integrated series z_t for $t = d' + 1, d' + 2, \cdots, k$. The integrated series is derived theoretically from the relationship

$$z_t = S^d S_s^D w_t \tag{14}$$

where $S = \nabla^{-1} = 1 + B + B^2 + \cdots$ is the nonseasonal summation operator and $S_s = \nabla_s^{-1} = 1 + B^s + B^{2s} + \cdots$ is the seasonal summation operator.

When (14) is employed to obtain an integrated series, the methods of the previous sections are utilized to determine the w_t sequence. Then the integration algorithm that is developed presently in this section is used to evaluate (14). The situation in which it is necessary to simulate data from a nonseasonal model containing a differencing operator is first considered. This is followed by a discussion of the generation of synthetic data from a general seasonal model that possesses a seasonal differencing operator and perhaps also a nonseasonal differencing operator.

Nonseasonal model: The integration algorithm for a nonseasonal Box-Jenkins model (i.e., $s = D = 0$) is as follows. For $i = 1, 2, \cdots, d$, (1) determine the starting value $\nabla^{d-i} z_t$ by differencing the given initial values $z_1, z_2, \cdots, z_d$, and (2) calculate $\nabla^{d-i} z_t$ for $t = d + 1, d + 2, \cdots, k$ by employing the identity

$$\nabla^{d-i} z_t = \nabla^{d+1-i} z_t + \nabla^{d-i} z_{t-1} \tag{15}$$

Seasonal model: For a seasonal model the integration algorithm is subdivided into two parts. The first stage consists of performing the nonseasonal integration. For $i = 1, 2, \cdots, d$, (1) determine the starting value $\nabla^{d-i}\nabla_s{}^D z_t$ by differencing the given initial values $z_1, z_2, \cdots, z_{d'}$, and (2) calculate $\nabla^{d-i}\nabla_s{}^D z_t$ for $t = d' + 1, d' + 2, \cdots, k$ by using the equation

$$\nabla^{d-i}\nabla_s{}^D z_t = \nabla^{d+1-i}\nabla_s{}^D z_t - \nabla^{d-i}\nabla_s{}^D z_{t-1} \tag{16}$$

In the second step the seasonal integration is performed. For $i = 1, 2, \cdots, D$, (1) determine the starting values $\nabla_s{}^D z_t$ for $t = d', d' - 1, \cdots, d' = s$, by differencing the given initial values $z_1, z_2, \cdots, z_{d'}$, and (2) calculate $\nabla_s{}^{D-i} z_t$ for $t = d' + 1, d' + 2, \cdots, k$ by using the equation

$$\nabla_s{}^{D-i} z_t = \nabla_s{}^{D-i+1} z_t + \nabla_s{}^{D-i} z_{t-s} \tag{17}$$

Models With Power Transformations

In Box-Jenkins modeling the model residuals are assumed to be independent, homoscedastic, and usually normally distributed. The most critical supposition is the independence assumption, and its violation can cause drastic consequences [*Box and Tiao,* 1973, p. 522]. However, if the homoscedastic and normality assumptions are not fulfilled, they are often reasonably well satisfied when the observations are transformed by a Box-Cox transformation [*Hipel et al.,* 1977; *McLeod et al.,* 1977].

Consider a Box-Cox transformation of the form [*Box and Cox,* 1964]

$$\begin{aligned} z_t{}^{(\lambda)} &= [(z_t + \text{const})^\lambda - 1]/\lambda \qquad \lambda \neq 0 \\ z_t{}^{(\lambda)} &= \ln (z_t + \text{const}) \qquad \lambda = 0 \end{aligned} \tag{18}$$

where const is a constant. The $z_t{}^{(\lambda)}$ are obtained by the methods of the preceding sections, and the synthetic data z_t are then calculated by the inverse transformation

$$\begin{aligned} z_t &= (\lambda z_t{}^{(\lambda)} + 1)^{1/\lambda} - \text{const} \qquad \lambda \neq 0 \\ z_t &= \exp z_t{}^{(\lambda)} - \text{const} \qquad \lambda = 0 \end{aligned} \tag{19}$$

Computer Algorithms

The simulation procedures labeled Wasim 1 and Wasim 2 have been coded in American National Standards Institute standard Fortran. The computer programs for these simulation methods, along with other subroutines such as the integration and inverse Box-Cox transformation procedures, are listed in the appendix in the microfiche edition of this paper. The appendix is divided into three sections. First, the various types of numerical methods that are employed in the subroutines are presented; second, a detailed description of the structure of the subroutines is given; and third, the actual programs are listed.

Waterloo Simulation Procedure 3 (Wasim 3)

The Wasim 3 algorithm can be used in simulation studies in which it is necessary to incorporate parameter uncertainty into the analysis. Suppose that it is necessary to generate k' synthetic traces of length k. When each series of length k is being generated, different values of the model parameters are randomly selected if Wasim 3 is employed. The Wasim 3 procedure is explained only for a nonseasonal Arma model, since extension to the seasonal case is straightforward.

Suppose that the historical time series containing N values is modeled as an Arma (p, q) process that has an estimated mean level of $\hat{\mu}$. The Gaussian white noise residuals have an estimated variance denoted by $\hat{\sigma}_a{}^2$. Let the vector of the estimated Arma parameters be given by

$$\hat{\beta} = (\hat{\phi}_1, \hat{\phi}_2, \cdots, \hat{\phi}_p, \hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_q) \tag{20}$$

The vector of the true model parameters is denoted by

$$\beta = (\phi_1, \phi_2, \cdots, \phi_p, \theta_1, \theta_2, \cdots, \theta_q) \tag{21}$$

The mean level of the true model is μ, while the variance of the white noise is $\sigma_a{}^2$.

If a noninformative prior distribution is used for the model parameters, then β, μ, and $\sigma_a{}^2$ are approximately independent with posterior distributions given by

$$\beta \sim N(\hat{\beta}, \mathbf{V}_{\hat{\beta}}) \tag{22}$$

where $\mathbf{V}_{\hat{\beta}}$ is the estimated covariance matrix of β which is usually calculated at the estimation stage of model development and $\sim$ means 'is distributed as.'

$$\mu \sim N\left[\hat{\mu}, \left(\frac{1 - \hat{\phi}_1 - \hat{\phi}_2 - \cdots - \hat{\phi}_p}{1 - \hat{\theta}_1 - \hat{\theta}_2 - \cdots - \hat{\theta}_q}\right)^{-2} \frac{\hat{\sigma}_a{}^2}{N}\right] \tag{23}$$

$$\sigma_a{}^2 \sim N[\hat{\sigma}_a{}^2, 2\hat{\sigma}_a{}^4/4] \tag{24}$$

The results given in (22), (23), and (24) are based upon large-sample theory. Nevertheless, these results can be used to obtain some idea of the importance, if any, of parameter uncertainty in a particular situation. It should be noted that if an informative prior were used, the variances of the parameters would be less and hence the parameter uncertainty would also decrease.

The following algorithm for Wasim 3 can be used to allow for parameter uncertainty when k' series of length k are to be generated from an Arma (p, q) model.

1. Set $i = 1$.
2. Randomly generate values for β, μ, and $\sigma_a{}^2$ using the posterior distributions given in (22), (23), and (24), respectively. Denote the generated parameter values as β_i, μ_i, and $\sigma_{a,i}{}^2$. Refer to the book by *Janson* [1966] for a method for obtaining random values from a multivariate normal distribution.
3. Use Wasim 2 (or Wasim 1) for an Arma (p, q) process with parameters β_i, μ_i, and $\sigma_{a,i}{}^2$ to simulate a synthetic series of length k that is represented by $z_1{}^{(i)}, z_2{}^{(i)}, \cdots, z_k{}^{(i)}$. If the model contains a Box-Cox transformation, the inverse transformation in (19) is required.
4. Set $i = i + 1$. If $i \leq k'$, then repeat steps 2 and 3 to obtain another possible realization of the time series. When $i > k'$, the Wasim 3 procedure is terminated.

Model Uncertainty

In the synthetic hydrology approach to reservoir design and operation an Arma model may be fit to a historical river flow time series and then used to simulate other possible realizations of the river flows. Two sources of possible error may arise. The model selected may be inappropriate, or the estimated parameters may be inaccurate. The procedures of *Box and Jenkins* [1970] and extensions developed by *McLeod* [1977*a*, *b*, 1978], *Hipel et al.* [1977], and *McLeod et al.* [1977] emphasize techniques for selecting an appropriate model followed by efficient parameter estimation and diagnostic checking for possible model inadequacies. It is thus reasonable to suppose that the selected model is at least approximately valid.

On the other hand, if a possibly inappropriate model, such as the fractional Gaussian noise model, is fit to the data and no checks of model adequacy are done, a seriously inadequate model may arise. It is demonstrated by *McLeod and Hipel* [1978] that the use of fractional Gaussian noise models may give very poor fits to annual river flow time series. If the methods advocated in the papers of *Hipel et al.* [1977] and *McLeod et al.* [1977] are used with a hydrologic time series of at least 50 observations, the selection of an inappropriate model is not likely to occur.

Parameter Uncertainty in Reservoir Design

In this section an algorithm is presented for estimating the expected utility of a reservoir design given the specified Arma (p, q) model for the river flow data and a posterior distribution $p(\beta, \mu, \sigma_a^2)$ for the parameters. Furthermore, it is pointed out that the method of *Vicens et al.* [1975] is inappropriate.

For a given river flow time series $z_1, \cdots, z_k$ and a particular reservoir design D the (vector valued) net benefit function is given by

$$NB = NB(z_1, \cdots, z_k; D) \tag{25}$$

and the utility is

$$U = U(NB) \tag{26}$$

The expected utility of D is then given by

$$\begin{aligned} u(D) &= E\{U[NB(z_1, \cdots, z_k; D)]\} \\ &= \int_{z_1} \cdots \int_{z_k} \int_{\beta_1} \cdots \int_{\beta_{p+q}} \int_{\mu} \int_{\sigma_a^2} \\ &\quad \cdot U[NB(z_1, \cdots, z_k; D)] \\ &\quad \cdot p(z_1, \cdots, z_k \mid \beta, \mu, \sigma_a^2) p(\beta, \mu, \sigma_a^2) \\ &\quad \cdot dz_1 \cdots dz_k \, d\beta_1 \cdots d\beta_{p+q} \, d\mu \, d\sigma_a^2 \end{aligned} \tag{27}$$

The best design D_0 maximizes the value of $u(D)$.

After a Box-Jenkins model is fit to the given time series of historical river flows, the following algorithm may be used to estimate $u(D)$ and a confidence interval (or Bayesian probability interval) for $u(D)$.

1. Set $i = 1$, $T_1 = 0$, and $T_2 = 0$. Let k' be the number of series of length k that are to be generated. For example, k' may have a value of 10,000.
2. Generate a synthetic time series $z_1^{(i)}, \cdots, z_k^{(i)}$ using the Wasim 3 algorithm.
3. Calculate $u_i = U[NB(z_1^{(i)}, \cdots, z_k^{(i)}; D)]$, set $T_1 = T_1 + u_i$, and set $T_2 = T_2 + u_i^2$.
4. Set $i = i + 1$, and go to step 2 if $i \leq k'$. Go to step 5 if $i > k'$.
5. Set

$$\bar{u} = (1/k')T_1 \tag{28}$$

and let

$$S_{\bar{u}} = \{[(1/k')T_2 - \bar{u}^2]/k'\}^{1/2} \tag{29}$$

The calculated $\bar{u}$ provides an estimate of $u(D)$, and a 95% confidence interval (or Bayesian probability interval) for $u(D)$ is given by $\bar{u} \pm 1.96 S_{\bar{u}}$. Although the aforesaid algorithm has been explained for a nonseasonal Arma (p, q) algorithm, the same approach is valid for seasonal models. The number of generated synthetic traces (i.e., k') can be increased if more accuracy is required or decreased if less accuracy is required.

In simulation studies it is essential that a synthetic data sequence generated from a stochastic model resemble statistically the historical observations. However, if the technique suggested by *Vicens et al.* [1975] is implemented, this criterion cannot be fulfilled. For the case of a Markov process these authors derive a Bayesian posterior distribution for z_{k+1} given the previous values $z_1, z_2, \cdots, z_k$ and an appropriate prior distribution. The posterior distribution is used to simulate a value for z_{k+1}. Repeating the procedure k times yields a new simulated series $z_{k+1}, z_{k+2}, \cdots, z_{2k}$. However, it is assumed that the given time series $z_1, z_2, \cdots, z_k$ was generated from a Markov model. Unfortunately, this is not true for the synthetic series $z_{k+1}, z_{k+2}, \cdots, z_{2k}$. That is, the generated time series $z_{k+1}, z_{k+2}, \cdots z_{2k}$ is not a possible realization of an Arma (1, 0) model. Thus the synthetic data that are generated by the algorithm of *Vicens et al.* [1975] are not a possible realization of the underlying stochastic process even if the assumed type of model is correct. If such a synthetic trace for streamflows were used to evaluate net benefits for a reservoir design, spurious results could be produced.

Applications

Three applications are presented to illustrate the advantages and usefulness of the simulation procedures presented in this paper. The first example demonstrates that the employment of Wasim 2 in simulation studies avoids bias that is due to fixed starting values. The second example shows how the model residuals from the historical data can be used in conjunction with Wasim 1 for generating synthetic data. Finally, the third example demonstrates how parameter uncertainty can be incorporated into a simulation study by using Wasim 3.

Avoidance of Bias in Simulation Studies

The rescaled adjusted range (RAR) and the Hurst coefficient K are two statistics that are important in problems related to the Hurst phenomenon. *McLeod and Hipel* [1978] have reassessed the controversies surrounding the Hurst phenomenon and have demonstrated that Arma processes are superior to fractional Gaussian noise models. In addition, *Hipel and McLeod* [1978] have shown that Arma models preserve the historical RAR or equivalently K. Accordingly, Arma models are important tools for utilization in hydrologic studies and should be used in preference to the so-called long-memory processes.

If the underlying process is a Box-Jenkins model, it can be shown theoretically that the RAR is a function only of the sample size and the AR and MA parameters [*Hipel*, 1975, Appendix B]. *Hipel and McLeod* [1978] demonstrate how to obtain the empirical cumulative distribution function (ECDF) for the RAR when the generating process is a specified stochastic model. In particular, consider the ECDF for a Markov model (i.e., an Arma (1, 0) process) with an AR parameter having a value of 0.7. When the Wasim 2 technique is employed to generate 10,000 sequences of length 30, the value of the 0.95 quantile for the ECDF of the RAR is 12.15. The 95% confidence interval for this value is calculated to be from 12.09 to 12.19 (see *Conover* [1971, p. 111] for the method for calculating the confidence interval for a quantile).

If random realizations of the stochastic process are not utilized as starting values, systematic bias can be introduced into a simulation study such as the development of the ECDF for the RAR. For the Markov model with an AR parameter

TABLE 1. Parameter Estimates for an Arma (2, 0) Model Fit to the Gota River Data

Parameter	Estimate	Standard Error
ϕ_1	0.591	0.079
ϕ_2	−0.274	0.078

having a magnitude of 0.7, 10,000 sequences of length 30 were generated, and for each sequence the mean value of zero was used as a starting value. In addition, exactly the same disturbances that were utilized in the simulation study using Wasim 2 were employed for the biased study. The value of the 0.95 quantile for the biased ECDF of the RAR is 12.01. The 95% confidence interval for this quantile value is from 11.97 to 12.05. Notice that the confidence interval for the biased result does not intersect with the corresponding interval for the unbiased study. Consequently, fixed initial values should not be used in the development of the ECDF for a specified statistic and generating mechanism.

Simulation Studies Using the Historical Disturbances

When Wasim 1 is used, it is not necessary to assume that the model residuals are NID $(0, \sigma_a^2)$. In fact, it is not necessary to determine any theoretical distribution for the disturbances to follow. Rather, in certain situations it may be advantageous to use the residuals from the historical data to form an empirical distribution for generating the innovations. For example, when a relatively large sample is available, it may be desirable to use the empirical distribution of the residuals for simulation studies, no matter what theoretical distribution the empirical results may most closely resemble. In other instances it may be difficult to determine which theoretical distribution to fit to the disturbances, and consequently, it may be profitable to employ the empirical distribution of the residuals. However, it should be pointed out that when the historical disturbances are employed, it is not possible to have a generated disturbance that is more extreme than the calculated residuals. Nevertheless, because of the form of the difference equation for a Box-Jenkins model that is fit to correlated data, it is possible that values of the generated data may be more extreme than those in the given time series.

A river flow time series is considered to demonstrate how the empirical distribution for the residuals can be used in practice. The average annual flows of the Gota River in Sweden from 1807 to 1957 are available in a paper by *Yevyevich* [1963]. A model is fit to these data by following the identification, estimation, and diagnostic check stages of model construction. The identification stage reveals that it may be appropriate to estimate an AR process of order 2. By using the method of *McLeod* [1977*a*, *b*], efficient parameter estimates and corresponding standard errors are calculated as listed in Table 1. At the estimation stage the white noise residuals $\{\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_{150}\}$ are determined by using the back forecasting technique of *Box and Jenkins* [1970, chapter 7]. Diagnostic checks performed on the residuals confirm that the modeling assumptions are satisfied. In particular, by calculating confidence limits for the residual autocorrelation function using the technique of *McLeod* [1978], the residuals are shown to be white noise.

To obtain synthetic data by using the Gota model, the Wasim 1 method is employed, and the white noise terms are chosen by selecting at random an element of the set $\{\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_{150}\}$. After one of the historical innovations is utilized, it is put back into the set of historical disturbances. Therefore selection is done with replacement, and this method is equivalent to using the empirical distribution of the residuals for the generation of the white noise terms.

As an example of a simulation study using the Gota model, consider the development of the ECDF for the Hurst coefficient K. The historical disturbances and the Wasim 1 technique are used to generate 10,000 sequences, each sequence containing 150 values. By calculating K for each of the 10,000 traces the ECDF for K can be obtained as shown in Table 2 for a series length which is the same as that of the historical time series.

The historical value of K for the Gota River is calculated to be 0.689. Notice that the observed K value does not lie in the tails of the ECDF for K in Table 2. The probability that K for the Gota model is greater than the historical K is 0.281. *Hipel and McLeod* [1978] apply this procedure to 23 geophysical time series and, by invoking a particular statistical test, demonstrate that Box-Jenkins models do preserve the Hurst coefficient K or equivalently the RAR.

Parameter Uncertainty in Simulation Experiments

An average annual river flow series having a length of 96 years is modeled to show how parameter uncertainty can be brought into a practical simulation study. The yearly river flows of the Mississippi River at St. Louis from 1861 to 1957 are available in an article by *Yevyevich* [1963]. By following the three stages of model development the best process for modeling the Mississippi flows is found to be an Arma (0, 1) model. The maximum likelihood estimate for the MA parameter θ_1 is −0.306 with a standard error of 0.097.

By using Wasim 1 (or equivalently Wasim 2) the Mississippi model is employed to generate 10,000 series of length 96. The RAR is calculated for each of the 10,000 traces. The expected value or mean of the RAR for the 10,000 series is 13.439 with a standard deviation of 0.030.

The Mississippi model is used with Wasim 3 to generate another 10,000 series of length 96. The innovations are different from those used for the simulation study with Wasim 1. For each trace of length 96 the value of the MA parameter used in Wasim 3 is determined by the equation

$$\theta_1 = -0.306 + 0.097\epsilon_t \tag{30}$$

where $t = 1, 2, 3, \cdots, 10{,}000$ and $\epsilon_t \sim$ NID (0, 1). Because the RAR is not a function of the mean level of the process or the

TABLE 2. Distribution of K for the Gota Model

Quantile	Value of K for Empirical White Noise
0.025	0.556
0.050	0.571
0.100	0.590
0.200	0.613
0.300	0.630
0.400	0.645
0.500	0.658
0.600	0.671
0.700	0.686
0.800	0.703
0.900	0.725
0.950	0.744
0.975	0.757

variance of the model residuals, it is only necessary to vary the MA parameter randomly for this particular simulation study. The expected value of the RAR for the 10,000 synthetic data sets is 13.443 with a standard deviation of 0.031. A comparison of the results for the simulation experiment using a constant MA parameter with those for the simulation experiment utilizing a varying model parameter reveals that there is no significant difference between the two expected values of the RAR. Hence for this particular study, parameter uncertainty is not a crucial factor.

Conclusions

Improved simulation procedures are now available for generating synthetic traces from Box-Jenkins models. Because random realizations of the underlying stochastic process are used as starting values, bias is not introduced into the simulated sequences. Furthermore, these techniques can be used in conjunction with models containing differencing operators or data that have been transformed by a Box-Cox transformation.

If the Wasim 1 method is utilized, it is not necessary that the distribution of the residuals be Gaussian. As is shown by an example, the empirical distribution of the residuals can be used for generation purposes. In addition, Wasim 1 is exact for a pure MA process. On the other hand, Wasim 2 is an exact simulation procedure for any Arma model. The only restriction with Wasim 2 is that the residuals are NID $(0, \sigma_a^2)$.

When parameter uncertainty is incorporated into a simulation study, the Wasim 3 procedure is the proper method to implement. If it is deemed necessary to consider parameter uncertainty in reservoir design, an algorithm is suggested in the paper that links Wasim 3 with the design problem. To circumvent difficulties with model uncertainty, it is recommended that a proper Arma model be fit to the given data set by following three stages of model development [*Hipel et al.*, 1977; *McLeod et al.*, 1977]. The use of the fractional Gaussian noise model should be avoided.

The Fortran computer algorithms for the simulation procedures are given in the appendix of the microfiche edition of the paper. Consequently, these techniques can be implemented immediately by the practitioner. The authors of this paper recommend that researchers involved in Monte Carlo studies with Box-Jenkins models employ the contemporary simulation procedures of this paper in their research endeavors.

Acknowledgments. The authors wish to thank Paul Newbold of the University of Nottingham for suggesting improvements to the Wasim 2 algorithm.

References

Atkinson, A. C., and M. C. Pearce, The computer generation of beta, gamma and normal random variables, *J. Roy. Statist. Soc., Ser. A*, *139*(4), 431–448, 1976.

Box, G. E. P., and D. R. Cox, An analysis of transformations, *J. Roy. Statist. Soc., Ser. B*, *26*, 211–252, 1964.

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

Box, G. E. P., and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Mass., 1973.

Brown, T. J., and J. C. Hardin, A note on Kendall's autoregressive series, *J. Appl. Probab.*, *10*, 475–478, 1973.

Clarke, R. T., Mathematical models in hydrology, Irrigation and Drainage Paper, 282 pp., Food and Agr. Organ. of the U. N., Rome, 1973.

Conover, W. J., *Practical Nonparametric Statistics*, John Wiley, New York, 1971.

Copas, J. B., Monte Carlo results for estimation of a stable Markov time series, *J. Roy. Statist. Soc., Ser. A*, *129*, 110–116, 1966.

Croley, T. E., II, and K. N. R. Rao, A manual for hydrologic time series deseasonalization and serial independence reduction, *Rep. 199*, 151 pp., Iowa Inst. of Hydraul. Res., Univ. of Iowa, Iowa City, 1977.

Delleur, J. W., P. C. Tao, and M. L. Kavvas, An evaluation of the practicality and complexity of some rainfall and runoff time series models, *Water Resour. Res.*, *12*(5), 953–970, 1976.

Farebrother, R. W., and G. Berry, A remark on algorithm AS 6, *J. Roy. Statist. Soc., Ser. C*, *23*, 479, 1974.

Hannan, E. J., *Multiple Time Series Analysis*, John Wiley, New York, 1970.

Healy, M. J. R., Algorithm AS 6, triangular decomposition of a symmetric matrix, *J. Roy. Statist. Soc., Ser. C*, *17*, 195–197, 1968.

Hill, I. D., Algorithm AS 100, normal-Johnson and Johnson-normal transformations, *J. Roy. Statist. Soc., Ser. C*, *25*(2), 190–192, 1976.

Hill, I. D., R. Hill, and R. L. Holder, Algorithm AS 99, fitting Johnson curves by moments, *J. Roy. Statist. Soc., Ser. C*, *25*(2), 180–189, 1976.

Hipel, K. W., Contemporary Box-Jenkins modelling in hydrology, Ph.D. thesis, Univ. of Waterloo, Waterloo, Ont., 1975.

Hipel, K. W., and A. I. McLeod, Preservation of the rescaled adjusted range, 2, Simulation studies using Box-Jenkins models, *Water Resour. Res.*, *14*(3), 509–516, 1978.

Hipel, K. W., A. I. McLeod, and W. C. Lennox, Advances in Box-Jenkins modeling, 1, Model construction, *Water Resour. Res.*, *13*(3), 567–575, 1977.

Janson, B., *Random Number Generators*, Victor Pettersons, Bokindustri Akiebolag, Stockholm, 1966.

Knuth, D. E., *The Art of Computer Programming*, vol. 2, Addison-Wesley, Reading, Mass., 1969.

Marsaglia, G., and T. A. Bray, A convenient method for generating normal variables, *SIAM Rev.*, *6*, 260–264, 1964.

McKerchar, A. I., and J. W. Delleur, Application of seasonal parametric linear stochastic models to monthly flow data, *Water Resour. Res.*, *10*(2), 246–255, 1974.

McLeod, A. I., Derivation of the theoretical autocovariance function of autoregressive-moving average time series, *J. Roy. Statist. Soc., Ser. C*, *24*(2), 255–256, 1975.

McLeod, A. I., Topics in time series and econometrics, Ph.D. thesis, Dep. of Statist., Univ. of Waterloo, Waterloo, Ont., 1977*a*.

McLeod, A. I., Improved Box-Jenkins estimators, *Biometrika*, *64*(3), 531–534, 1977*b*.

McLeod, A. I., On the distribution of residual autocorrelations in Box-Jenkins models, *J. Roy. Statist. Soc., Ser. B*, *40*, in press, 1978.

McLeod, A. I., and K. W. Hipel, Preservation of the rescaled adjusted range, 1, A reassessment of the Hurst phenomenon, *Water Resour. Res.*, *14*(3), 491–508, 1978.

McLeod, A. I., K. W. Hipel, and W. C. Lennox, Advances in Box-Jenkins modeling, 2, Applications, *Water Resour. Res.*, *13*(3), 577–586, 1977.

Moran, P. A. P., *The Theory of Storage*, Methuen, London, 1959.

Nicholls, D. F., On Hannan's estimator of Arma models, *Aust. J. Statist.*, *14*, 262–269, 1972.

Pagano, M., When is an autoregressive process stationary?, *Commun. Statist.*, *1*, 533–544, 1973.

Ralston, A., *A First Course in Numerical Analysis*, McGraw-Hill, New York, 1965.

Tao, P. C., and J. W. Delleur, Seasonal and nonseasonal Arma models in hydrology, *J. Hydraul. Div. Amer. Soc. Civil Eng.*, *102*(HY10), 1541–1559, 1976.

Vicens, G. J., I. Rodriguez-Iturbe, and J. C. Schaake, Jr., Bayesian generation of synthetic streamflows, *Water Resour. Res.*, *11*(6), 827–838, 1975.

Yevyevich, V. M., Fluctuation of wet and dry years: Part 1—Research data assembly and mathematical models, *Hydrol. Pap. 1*, Colo. State Univ., Fort Collins, Colo., 1963.

(Received March 2, 1977;
revised February 27, 1978;
accepted April 27, 1978.)