Innovative Systems Design and Engineering www.iiste.org
ISSN 2222-1727 (Paper) ISSN 2222-2871 (Online) DOI: 10.7176/ISDE
Vol.10, No.1, 2019

Stochastic Data Generation Technique Using Autoregressive Moving Average (ARMA) Model

Khalid A. Alkhuzai
Civil Engineering Department, Faculty of Engineering, Albaha University, Albaha, KSA

Abstract
Data generation mechanisms have been widely applied in hydrology. Models are built for generation of data having the statistical properties of the historical records. The creation of synthetic time series starts with the generation of independent normal variables with average zero and variance one, then adding the time and spatial dependence structure as well as periodic components, whichever necessary. The data generation can be accomplished via the analysis of the historical data to check its suitability for generation, Selection, identification of the form, estimation of parameters, & check of the data generation model and Application of the model & testing of the results. This paper summarizes the required work to be done as per the above steps taking the autoregressive moving average as an example of the data generation model
Keywords: Induced Voltage, – Electric Fields, HVDC Transmission, Finite Element Method, Hybrid Transmission Lines.
DOI: 10.7176/ISDE/10-1-02

1. Introduction

The designers of water resources systems have realized that evaluating their designs using past or historical data provided no guarantee that the design would perform satisfactorily in the future because future flow sequences will not be the same as past flow sequences (Haan, 1982). The last statement describes the need for data generation in order to obtain new time series simulating the possible future flows. Synthetic data series are generated by many models using autoregressive (AR) processes, Thomas-Fiering model (1967), method of fragments (Srikanthan and McMohan(1985)) and its modified version (Maheepala and Perera (1996)) , the non-parametric approach model (Sharma and O'Neil (2002))and wavelet approach (Ünal, Aksoy and Akar (2004)).

2. Time Series Modelling

Time series modelling is the process of finding a mathematical model that represents a time series. It has mainly two uses in hydrology and water resources:

A. For **generation** of synthetic hydrologic time series, and
B. For **forecasting** future hydrologic series.

Generation of synthetic time series is generally needed for:

a. Water harvesting projects.
b. Reservoir sizing,
c. Determining the risk of failure (or reliability) of water supply or irrigation systems,
d. Future planning of reservoir operation,
e. Planning capacity expansions of water supply systems,…etc,

While forecasting of hydrologic series is needed for,

a. Short term planning of reservoir operation,
b. Real time and short term operations of river basins or systems,
c. Planning operation during an on-going drought, etc.

Box and Jenkins (1976) organized the modelling in four steps:

i. The selection of the type of model
ii. The identification of the form of model
iii. The estimation of the model parameters
iv. The diagnostic check of the model.

3. Time Series and their components

Yevjevich (1972b) defined the time series as "any magnitude observed at discrete times (of equal or unequal distances), averaged over and related to interval Δt along total time T or recorded continuously with time."
Time series are considered stationary if the statistical properties such as mean and standard deviation are unaffected by a shift in the time origin. There are two basic classes for time series, deterministic series and stochastic series.
Hydrologic time series are generally divided into four components:
Over-years Trend and other deterministic changes denoted by T(t)
Cyclic or periodic changes denoted by P(t)

Almost periodic changes such as tidal effect on hydrologic time series.
Stochastic or random variation which consists of purely random component E(t) and dependant random component S(t)
Considering the second and third components as one component and summing up all components, a hydrologic time series Q(t) may be written as:
$Q(t) = T(t)+P(t)+S(t)+E(t)$
The first three components are deterministic components and the last one is a random one. Trend component refers to the upwards or downwards of the series over time. Kottegoda (1970) concluded that the hydrologic time series of less than 100 years cannot show any evidence of trend. However, it is important to note that tis statement may be true as far as the flow remains natural without being interrupted (e.g. by building a structure etc. Hence the time series can be written as:
$Q(t)= P(t)+S(t)+E(t)$
In the data generation process, the above components are separated. Therefore, the generation of new series can be considered as a reversible process of the decomposition of a time series into its various components.

4. Data Generation Models

In general, data generation can be accomplished in three steps,
Analysis of the historical data to check its suitability for generation,
Selection, identification of the form, estimation of parameters, and check of the data generation model,
Application of the model and testing of the results.

5. Analysis of the Historical Data

The historical data can be analysed as per the following sections (Naggar, 1999).

5.1 Testing the means, standard deviations, and skew coefficients for data homogeneity:

Using the historical data series, the sample mean x(j), standard deviation s(j), and skew coefficient $\hat{C}s(j)$ were computed for each calendar month (j) by the relations(Yevjevich (1972a)),

$$x(j) = \sum_{i=1}^{N} (x(i,j))/N \quad (1)$$

$$s(j) = \{\sum_{i=1}^{N} (x(i,j) - x(j))2/(N-1)\}1/2 \quad (2)$$

$$\hat{C}s(j) = N2\{\sum_{i=1}^{N} (x(i,j) - x(j))3/N\}/\{(N-1)(N-2)(s(j)3\} \quad (3)$$

Where,
X(i,j) = observation at year i and month j,
N = number of years.
Then, the series were divided into sub-series as follows,
2 series each containing half of the data,
3 series each containing third of the data, and
4 series each containing quarter of the data.
Therefore, a total number of 2+3+4= 9 sub-series were obtained.
In a similar manner, the means, standard deviations, and skew coefficients for the nine series were calculated and plotted together with (for example) the 95% confidence limits for the means and standard deviations.
The 95% confidence limits for the means and standard deviations are given by,
For the means: $x(j) \pm 1.96 \times s(j)/N$ (4)
For the standard deviations: $s(j)\ (N-1)1/2/\chi 2\alpha/2$ (5a)
: $s(j)\ (N-1)1/2/\chi 21-\alpha/2$ (5b)
Where:
$\alpha = 1 - 0.95 = .05$,
$\chi 2\alpha/2$ and $\chi 21-\alpha/2$ are the Chi-square values.
Then the results obtained should be compared with the allowable values
(Kottegoda (1970)). If they satisfy the limits then the data is considered homogenous.

5.2 Testing the stationarity of the data:

If the properties of a time series do not change with the absolute time it is called stationary. A stochastic process

is said to be strictly stationary if the probability distribution function and its associated parameters are invariant to shift in time (Al-Dabbagh (1986)). To test the stationarity, the historical data is divided into two series of equal length. Then,
The cumulative distribution for each half is computed and drawn
The serial correlation coefficients for each half is computed and drawn
If there is a good agreement in then the historical data is considered to be stationary.

5.3 Testing the normality of the data:
The normal probability density function is useful in hydrology. Yevjevich (1972b) listed a number of cases where its use is beneficial. One of these cases is the Data Generation. Testing the goodness of fit to normal probability distribution is done using the Chi-square test. Normalization can be done if needed using the Box & Cox power transformation which was suggested by Chandra et al. (1978). The transformation is given by,

$$y = (x\lambda - 1)/\lambda \qquad \lambda \neq 0 \qquad (7a)$$

$$y = \log(x) \qquad \lambda = 0 \qquad (7b)$$

Where:
x = observed (original) value,
y = transformed value,
λ = transformation coefficient.
The above transformation was selected because it proved to be an efficient one-parameter transformation and moreover, to avoid using other trial and error transformation methods.
The values of λ were obtained graphically (and checked by curve fitting) by plotting λ against skew coefficients and choosing the value of λ which gives zero coefficient. After transforming the data, Chi-square test for normality should be then done for both original and transformed data.

5.4 Testing the periodicity:
Salas et al. (1985) described the periodicity as the periodic change of statistical characteristics with time. To trace it, the serial correlogram for the historical data and transformed data should be drawn. If the periodicity is clearly noticed, it should be removed using both parametric and non-parametric methods as follows,
5.4.1 Parametric Method for Separating Periodicity:
To economize on the number of statistics needed for mathematical description of time series, the periodic component can be separated by superimposed harmonics. Fourier series can be used to describe the periodic mean M(j) and the periodic standard deviation S(j),

$$M(j) = (1/w)\sum_{j=1}^{W} Qj + \sum_{k=1}^{m} (AkCos(2\pi kj/w)\ BkSin(2\pi kj/w) \qquad (8)$$

$$M(j) = (1/w)\sum_{j=1}^{W} \sigma j + \sum_{k=1}^{m} (AskCos(2\pi kj/w)\ BskSin(2\pi kj/w) \qquad (9)$$

j= 1,2,……w (w=12 for monthly flow)
k=1,2,……m (m= number of harmonics)
Qj and σj are the mean and standard deviation for the month j respectively and,

$$Ak = (2/w)\sum_{j=1}^{W} Qj\ Cos(2\pi kj/w) \qquad (10)$$

$$Bk = (2/w)\sum_{j=1}^{W} Qj\ Sin(2\pi kj/w) \qquad (11)$$

and,

$$Ask = (2/w)\sum_{W} \sigma i\ Cos(2\pi ki/w) \qquad (12)$$

$$Bsk = (2/w)\sum_{j=1} \sigma j \, Sin(2\pi kj/w) \quad (13)$$

5.4.2 Non-Parametric Method for Separating Periodicity:

The method is given by the following simple transformation Z(i,j) of a time series Q(i,j),

$$Z(i,j) = Q(i,j) - Q(i,j)/S(j) \quad (14)$$

i = 1,2,….N

j= 1,2,…..W

where,

Z(i,j) = the standardized series with mean zero and variance one.

N= number of years,

W= number of positions inside the year (=12 for monthly data)

For an individual monthly sample, the mean Q(j) can be estimated by:

$$Q(j) = \sum_{j=1}^{N} Q(i,j)/N \quad (15)$$

And the individual monthly standard deviation:

$$S(j) = [\sum_{j=1}^{N} (Q(i,j)-Q(j))2/(N-1)]1/2 \quad (16)$$

It is noted that the number of statistics computed is equal to 2x Number of positions, i.e 24 (=2x12) for monthly data, 730 (=2x365) for daily data… and so on. This is considered as one of the disadvantages of the non-parametric method, especially for daily and or hourly data where a large number of statistics are needed to be computed. The method was used by many researchers (Young and Pisano(1968))

6. Selection of the Data Generation Model

The model selected in this paper is the Autoregressive-Moving Average Model (ARMA). The model was proposed by Carlson et al. (1970). It differs from the Autoregressive models in the sense that it has a moving average component, which adds more flexibility and makes it possible to build a model with a minimum number of parameters. Formulation of the models is given in the following sections.

7. Development of ARMA Model

An autoregressive model of order p and a moving average model of order q may be combined together to form a mixed autoregressive-moving average model called ARMA and denoted by ARMA (p,q). Considering a periodic time series yi,j , the ARMA model for yi,j can be written as,

$$y_{i,j} = \mu j + \sigma j \, z_{i,j} \quad (17)$$

where μj and σj are the periodic mean and periodic standard deviation of the season j.

zi,j may be represented by an ARMA model with either constant or time varying(periodic) coefficients.

The general form of ARMA(p,q) model with constant coefficients is,

$$zt = \varphi 1 zt-1 + \varphi 2 zt-2 + \ldots\ldots \varphi p zt-p - (\theta 1 \varepsilon t-1 + \theta 2 \varepsilon t-1 + \ldots\ldots \theta q \varepsilon t-q) + \varepsilon t \quad (18a)$$

or

$$zt = \sum_{j=1}^{p} \varphi j zt-j - \sum_{i=1}^{q} \theta i \varepsilon t-i + \varepsilon t \quad (18b)$$

or

$$zt = \sum_{j=1}^{p} \varphi j zt-j - \sum_{i=0}^{q} \theta i \varepsilon t-i \quad ,(\theta 0 = -1) \quad (18c)$$

where

φ1 , φ2 ,…..φp are the autoregressive coefficients,

θ1 , θ2 ,…..θq are the moving average coefficients,

εt is the independent normal variable,

t = w(i-1) + j (= 12(i-1) + j for monthly series).

While the general form of ARMA with periodic coefficients is,

$$z_{i,j} = \sum_{k=1}^{p} \varphi_{k,j}\, z_{i,j-k} - \sum_{m=1}^{q} \theta_{m,j}\varepsilon_{i,j-m} + \varepsilon_{i,j} \quad (19)$$

Where

$\varphi_{k,j}$ are the periodic autoregressive coefficients for k=1,…p and j=1,…w,

$\theta_{m,j}$ are the moving average periodic coefficients for m =1,..q and j=1,…w,

$\varepsilon_{i,j}$ is an independent normal random variable for year (i) ,and time interval (j).

To choose between constant or periodic coefficient models the autocorrelation function were drawn for the standardized series as shown in Fig. (9). From the figures no periodicity is noticed. Therefore, constant coefficient model is selected.

A summarized procedure for the development of the model is given in the following steps (Naggar, 1999).

Step 1: Transformation. The box and cox power transformation is utilized to get the series which is normal or approximately normal.

Step 2: Stationarization. The transformed series was stationarized using Fourier series.

Step 3: Autocorrelation and partial autocorrelation. The autocorrelation and partial autocorrelation should be calculated and drawn.

Step 4: Identification. Use the autocorrelation and partial autocorrelation functions plots in the identification of the time series which indicated the possibility of ARMA model.

Step 5: Initial estimate of the autoregressive parameter. The premier estimate of autoregressive parameter $\varphi 1$ has been obtained from the difference equation known as Yule-Walker equation given by:

$$\rho_k = \varphi 1 \rho_{k-1} + \varphi 2 \rho_{k-2} + \ldots\ldots \varphi p \rho_{k-p} \quad (20)$$

Where:

$\varphi 1, \varphi 2, \ldots\ldots \varphi p$ are the autoregressive coefficients.

Step 6: Premier estimate of the moving average parameter. The initial estimate of the moving average parameter $\theta 1$ has been obtained by:

Finding the values of lag-k autocovariance cj using,

$$c_k = \sum_{t=1}^{N-k} (z_t - \bar{z})(z_{t+k} - \bar{z})/(N-k) \quad (21)$$

Finding the value of the autocovariance cj' using the formula given by Box and Jenkins (1976),

$$c_j' = \sum_{i=0}^{p} \varphi_{j2}\, c_j + \sum_{i=1}^{p} (\varphi 0 \varphi 1 + \varphi 1 \varphi i+1 + \ldots \varphi p\text{-}i \varphi p) d_j \quad (22)$$

Where

$d_j = c_{j+1} + c_{j-1}$; $j = 1,2,\ldots q$; $\varphi 0 = -1$

$c_j = c_{-j}$

Hence,

$$c0' = \varphi 02\ c0 + \varphi 12\ c0 + \varphi 0 \varphi 1\ (c1 + c{-}1) = c0 + \varphi 12\ c0 - 2\varphi 1\ c1 \quad (23a)$$

$$c1' = \varphi 02\ c1 + \varphi 12\ c1 + \varphi 0 \varphi 1\ (c2 + c0) = c1 + \varphi 12\ c1 - \varphi 1 (c2 + c0) \quad (23b)$$

Obtaining the values of c0' and c1' from the previous equations,

Substituting in (Salas et al.(1985)),

$$\sigma\varepsilon 2 = c0'/(1 + \theta 12 + \theta 22 + \ldots \theta q2) \quad (24a) \text{ and}$$

$$\theta 1 = -(c1'/\sigma\varepsilon 2) \quad (24b)$$

Solving equations (15a) and (15b), the value of $\sigma\varepsilon 2$ and the initial estimate of $\theta 1$ were obtained.

Step 7: Maximum likelihood estimates. Refined values of the ARMA parameters $\varphi 1$ and $\theta 1$ can be obtained by minimizing the sum of squares of residuals.

Step 8: Test the goodness of fit. The Q-statistic test has been used for testing the goodness of fit as follows

$$Q = N \sum_{k=1}^{L} r_k^2(\varepsilon) \quad (25)$$

$$r_k = (N/(N-k)) \sum_{t=1}^{N-k} ((X_t - \bar{X})(X_{t+k} - \bar{X}))/\sum_{t=1}^{N} (X_t - \bar{X})^2 \quad (26)$$

Where:

r_k = lag k autocorrelation coefficient,

X_t = observation value at time t,

X_{t+k} = observation value at time t+k,

$\bar{X}$ = mean of observation values.

L may be of the order 10-30 percent of sample size N(20% is used here) . The statistic Q is approximately $\chi 2(L-p-q)$. If $Q < \chi 2(L-p-q)$ then εt is independent which implies that the model is adequate.

Step 9: Generation. The zt can be generated using,

$$zt = \varphi 1\ zt\text{-}1 + \varepsilon t - \theta 1 \varepsilon t\text{-}1 \quad (27)$$

It is necessary to give initial values for zt-1 and εt-1. These values are generally taken as the last values of the series. It is also necessary to multiply the random generated numbers εt-1 by σε2 to obtain random numbers with zero mean and variance σε2.

Step 10: Inverse standardization. Change the zt series to $z_{i,j}$ form, then $y_{i,j}$ is obtained by inverse standardization of the $z_{i,j}$ as follows,

$$y_{i,j} = \mu j + \sigma j\ z_{i,j} \quad (28)$$

Step 11: Inverse transformation. The $y_{i,j}$ was inverse-transformed by the relation,

$$x_{i,j} = (\lambda y_{i,j} + 1)1/\lambda \quad (29)$$

Where:

$x_{i,j}$ = is the generated series,

λ = is the Box and Cox transformation coefficient.

8. Generation of Data using ARMA Model

Initial and refined values of the model parameters should be obtained. The refined values are obtained by selecting values of Φ and θ in the neighbourhood of the initial estimates, then calculating the residuals and sum of the squares of the residuals. The most likelihood estimate corresponds to the minimum of the overall of squares surface. Results of the Q-statistic should also be found to ensure that the generated data has passed the test.

Finally, the historical and generated data should be drawn in one graph to enable visual inspection of the two series and to check visually whether they belong to the same population. Comparison of the Statistical properties of the historical and generated data should also be done. In order to check the adequacy of the generated sets, the means and standard deviations of the seasons should be tested at the required probability level.

Conclusions

Data generation is a time series modelling for finding a mathematical model that represents a time series. The creation of synthetic time series starts with the generation of independent normal variables with mean zero and variance one, then adding the time and spatial dependence structure as well as periodic components, whichever necessary. The generation procedure includes the analysis of the historical data to check its suitability for generation, Selection, identification of the form, estimation of parameters, & check of the data generation model and model application & testing of the results. The work to be done for testing the historical and the generated data is summarized taking the autoregressive moving average as an example of the data generation model.

Acknowledgement:

The research work is a part of a project entitled “Analysis of Rainfall Data of Albaha Region”. This project funded by the Deanship of Scientific Research, Albaha University, Kingdom of Saudi Arabia (Grant No. 117/1438). The assistance of the deanship is gratefully acknowledged.

References

Al-Dabbagh, A. R., 1986, “Stochastic Modeling of the Upper Rio Grande System,” Ph.D. Thesis, Department of Civil Engineering, New Mexico State University.

Box, G. E. P. , and Jenkins, G. M. , 1976, Time Series Analysis Forecasting and Control, Holden-Day Inc., San Francisco, California.

Carlson, R. F. , McCormick, A. J. A. , and Watts, D. G. , 1970, “Application of Linear Models to Four Annual Streamflow Series,” Water Resources Research, Vol. 6, No. 4, August, pp. 1070-1078.

Haan, C.T., 1982, Statistical Methods in Hydrology, Iowa State University Press, Iowa, USA

Kottegoda, N. T. , 1970, “Statistical Methods of River Flow Synthesis for Water Resources Assessment,” Ph.D. Thesis, University of Birmingham, England.

Maheepala, S. and Perera, C.J.C., 1996. "Monthly hydrologic data generation by disaggregation" J. Hydrology., 178, 277–291.

Naggar, O.M., 1999, “Development of Decision Support Systems in Water Resources,” Ph.D. Thesis, University of Baghdad, Iraq.

Salas, J. D. , Delleur, J. W. , Yevjevich, Y. , and Lane, W. L. , 1985, Applied Modeling of Hydrologic Time Series, Water Resource Publications, Littleton, Colorado.

Sharma, A., and R. O’Neill (2002), A nonparametric approach for representing interannual dependence in

monthly streamflow sequences, Water Resources Research., 38(7), 1100, doi:10.1029/2001WR000953.
Srikanthan, R. and T.A. McMahon. (1985), "Stochastic generation of rainfall and evaporation data." AWRC Technical Paper, No. 84, 301pp.
Thomas, H.A., and Fiering, M.B., 1967, "Mathematical Synthesis of Streamflow Sequences for the analysis of River Basin by Simulation," in Design of Water Resource Systems, Edited by: A. Maass et. al., Harvard University Press, Cambridge, Massachusetts, USA.
Ünal, N. E. , Aksoy, and Akar, H. , T., 2004, "Annual and monthly rainfall data generation schemes", Stochastic Environmental Research and Risk Assessment Journal, Volume 18, Issue 4, pp 245–257
Yevjevich, V, 1972a, Probability and Statistics in Hydrology, Water Resources Publications, Fort Collins, Colorado.
Yevjevich, V. , 1972b, Stochastic Processes in Hydrology, Water Resources Publications, Fort Collins, Colorado.

List of Symbols

E(t)	Purely random component of a time series,
P(t)	Periodicity component of a time series
Q(i,j)	Time series observation at year i and season j
r_k,	Lag k autocorrelation coefficient equal the Sample correlation coefficient
T(t)	Trend component of a time series
t	The time,
W	Number of seasons (=12 for monthly time series)
y_t	Time dependent series,
Z(i,j)	Standardized time series with mean zero and standard deviation one
$z_{v,t}$	Dependent variable series
ε_t	Time independent component
φ_1, φ_2 , φ_3 …φ_p	Autoregressive coefficients
ρ_k	Lag k population autocorrelation coefficient
λ	Transformation factor (in Box and Cox power transformation)
μ	Mean
θ_1, θ_2 , θ_3 …θq	Moving average coefficients,
σ	Standard deviation
ξ	Independent standardized normal variable