36

INLAND WATERS BRANCH

DEPARTMENT OF ENERGY, MINES AND RESOURCES

C.C.I.W.
MAR 15 1971
LIBRARY

Data Generating Methods in Hydrology

G.W. KITE and R.L. PENTLAND

TECHNICAL BULLETIN NO.36

GB
707
C338
no. 36

CANADA

TECHNICAL BULLETIN NO. 36

Data Generating Methods in Hydrology

G.W. KITE and R.L. PENTLAND

INLAND WATERS BRANCH
DEPARTMENT OF ENERGY, MINES AND RESOURCES
OTTAWA, CANADA, 1971

CONTENTS

ABSTRACT

It is now generally accepted that traditional empirical solutions to flood, drought, and storage problems are inadequate. On the other hand, analytical methods of solution are often intractable and therefore cannot be used in real-life problems. A third solution, data generation or simulation, offers a computationally easy and highly efficient alternative. This report summarizes the background to simulation and shows some of the many methods of generating data now available.

Data Generating Methods in Hydrology

G.W. KITE and R.L. PENTLAND

INTRODUCTION

"Any similarity between the events described here and real streamflow records is not a pure and sublime coincidence but rather the result of deliberate forethought.". Fiering (1967).

In most cases there are three methods by which problems involving stochastic hydrology may be solved:

(a) Empirical solutions

(b) Analytical methods

(c) Data generation

The empirical method of problem solving has traditionally been carried out with some variation of a mass curve analysis (Rippl, 1883). This approach has the following principal defects:

(i) It is unrealistic to design a project on the basis of a single hydrologic sequence which is highly unlikely to re-occur during the lifetime of the project.

(ii) The mass diagram does not facilitate the analysis of risk of water shortages during periods of low flow.

(iii) The storage capacity determined from the Rippl Method increases with the length of the record. This causes arbitrary adjustments to be made when the length of the historic record differs from the economic life of the proposed structure.

For solving problems, analytical methods rapidly become very complex as the distributions of variables involved depart from the simplest possible. Since in practical problems, distributions are invariably complex and hydrologists never have enough data, simulation or data generation is often the only solution.

Data generation can provide no new statistical information about a process, but by generating many long periods of record while preserving the properties of the original data, it can help solve the problem. To quote

Fiering (1967, p. 2) again, "in practice, simulation is a tactical, not a strategic, victory over analytic insufficiency."

Simulated data, if used properly, will prevent the "tailoring" of water resources systems to one historical sequence of data and can thus provide a realistic economic and engineering appraisal of a project. That is, if a project is tested over only a single sequence of water supplies obtained from the historical record, then only a single estimate of its performance is obtained but no information is gained about its performance during other equally likely time series.

The general idea of simulation, or data generation, is an old one. Even the more specific idea of simulation with the help of some particular stochastic process can be traced back several hundred years to near the beginning of probability theory, to Buffon's needle. However, the idea of using random sampling to estimate distribution functions is a more recent development due to 'Student' in 1908.

One of the first uses of data generation in hydrology was made by Hazen (1914). Later, Sudler (1927), employed decks of cards and other sampling devices to generate non-historic flow patterns which were then analyzed by the mass diagram method to develop probability distributions of reservoir capacities. These researchers all recognized the usefulness of synthetically increasing the length of the hydrologic record. However, due to obvious flaws in the suggested methods, and because of computational difficulties, none of their methods were generally accepted by water resources designers.

DATA GENERATING

In order to simulate any process involving a specified random process, a sequence of random variables corresponding to some fixed distribution function, usually the normal, must be constructed. In order to obtain a value of a random variable with this fixed distribution function, uniformly distributed random numbers are commonly used. The five steps involved in a simulation study then are:

A Model build-up.

B Generating uniformly distributed pseudo-random numbers.

C Randomness tests.

D Generating a set of normally distributed random numbers.

E Testing generated data.

Each of these steps are described on the following pages.

A. Model Build-up

General Discussion

Any phenomenon that undergoes continuous change, particularly with respect to time, may be called a process. As practically all hydrologic phenomena change with time, they are hydrologic processes. If the chance of occurrence of the variables involved in such a process is ignored and the model is considered to follow a definite law of certainty, but not any law of probability, the process and its model are described as deterministic.

If, on the other hand, the chance of occurrence of the variables is taken into consideration and the concept of probability is introduced in formulating the model, the process and the model are described as stochastic or probabilistic. For example, conventional flood routing through a reservoir, and also unit hydrograph theory, are deterministic models since no probability theory is involved. Using a queuing theory model for probability routing, however, is stochastic or probabilistic.

Stochastic processes are generally considered to be time-dependent while probabilistic processes are thought of as time-independent, that is, the sequence of occurrence of the events involved in the process is ignored and the chance of their occurrence is assumed to follow a definite probability distribution in which the variables are considered pure random. For a time-dependent stochastic process, the sequence of occurrence of the variates is observed and used in the process. The variables may be either pure-random or non pure-random and the probability distribution of the variables may or may not vary with time. If pure-random, the members of the time series are independent among themselves and so constitute a random sequence. If non pure-random, the members of the time series are dependent among themselves, and are composed of a deterministic component and a pure-random component, and so constitute a nonrandom sequence.

The general model of a hydrologic process can be described as:

$$X_t = R_t + P_t + \varepsilon_t \qquad \text{......(1)}$$

where R_t is a trend component,

P_t is a deterministic component, and

ε_t is a stochastic component.

Generally a trend component can be isolated and eliminated and so model (1) can be reduced to

$$X_t = P_t + \varepsilon_t \qquad \text{......(2)}$$

for further analysis.

In reality, all hydrologic processes are more or less stochastic; they have been assumed deterministic or probabilistic only to simplify their analysis. Mathematically, a stochastic process is a family of random variables X(t) which is a function of time, or other parameters, and whose variate X_t is changing in time t within the range of time T.

Quantitatively, the stochastic process may be a discrete or a continuous time series and can be sampled either continuously or at discrete or uniform intervals.

If the time series is sampled continuously, then there are two options open for analyzing the series:

1. to use an analog computer or,

2. to digitize the data and use a digital computer; this, however, involves loss of information.

Similarly, if the time series is sampled at discrete intervals, then the data can be analyzed by

1. digital computer or,

2. by interpolating, using an analog computer.

Hybrid computers are also available which accept analog input and produce digital output.

Just as the stochastic process can be divided into pure-random and non pure-random classes, so the deterministic time series component can be subdivided. Most deterministic components are periodic functions but other functions such as time trends and jumps exist. The deciding property of a deterministic component is that the value of the variable can be precisely computed at any time by fitting a mathematical equation. The subdivisions of a deterministic component can be:

1. Trend: This is a unidirectional diminishing or increasing change in the average value of a hydrologic variable, such as the trend of annual precipitation often visible. A common, though often misused, method of analyzing trend is by moving averages.

2. Periodicity: This represents a regular or oscillatory form of variation, such as the diurnal, seasonal, or secular changes that frequently exist in hydrologic data. These variations are of nearly constant length and may be analyzed by Fourier analysis. A Fourier series is used to represent the time series $X_1, X_2, \ldots X_n$ with a total period T:

$$X_t = \frac{A_o}{2} + \sum_{j=1}^{n} \left(A_j \cos \frac{2\pi jt}{T} + B_j \sin \frac{2\pi jt}{T} \right) \quad \ldots\ldots(3)$$

where A_o is a constant, t is the time, and the amplitudes A_j and B_j are expressed as

$$A_j = \frac{2}{n} \sum_{t=1}^{n} y_t \cos \frac{2\pi jt}{T} \quad \ldots\ldots(4)$$

and $$B_j = \frac{2}{n} \sum_{t=1}^{n} y_t \sin \frac{2\pi jt}{T} \quad \ldots\ldots(5)$$

where y_t is the deviation of X_t from the arithmetic straight line trend for the period selected, and with $j = 1,2,\ldots n$ being the number of harmonics used in the analysis.

The sum of the squared amplitudes is

$$\sum R_j^2 = \sum \left(A_j^2 + B_j^2 \right) \quad \ldots\ldots(6)$$

and if the series is pure-random with no periodic fluctuations, the mean squared amplitude of the series is

$$R_m^2 = \frac{4\sigma^2}{n} \quad \ldots\ldots(7)$$

where σ^2 is the variance of the deviations y, and

n is the number of harmonics used.

Three tests of periodicity are available:

(i) Schuster test: The hypothesis to be tested is that the series is not significantly different from pure-random. If $k = \frac{R_j^2}{R_m^2}$, then the probability in percent that R_j^2 is k times R_m^2 is given by Schuster (1898) as

$$P_S = e^{-k} \qquad \ldots\ldots(8)$$

Taking $P_S = 10\%$ as the level of significance, the value of R_j^2 for a given series can be tested to see if it differs from R_m^2 derived for a pure-random series.

Corresponding to $P_S = 10\%$; $k = 2.303$, thus $R_j^2 = 2.303R_m^2 = 9.212\sigma^2/n$. Substituting R_j^2 the value of j can be computed and the possible hidden periodicity found as T/j.

(ii) Walker test: According to Walker (1925), the probability that at least one squared amplitude R_j^2 will be k times R_m^2 is

$$P_w = 1 - (1-e^{-k})^{n/2} \qquad \ldots\ldots(9)$$

which may be used for a periodicity test, as in the Schuster test.

(iii) Fisher test: Let R_j^2 be the largest of the squared amplitudes R_j^2. From Fisher (1929) the probability P_f that $R_J^2/2\hat{S}^2$ (where $\hat{S}^2$ is the unbiased estimate of σ^2) is greater than a given value g is:

$$P_f = \sum_{i=o}^{m} (-1)^i \binom{j}{i} (1-ig)^{j-1} \qquad \ldots\ldots(10)$$

where m is the greatest integer less than 1/g and j=1,2,... is the number of periods. The probability may be used for a periodicity test as in the Schuster test.

3. Persistence: Persistence means that successive members of a time series are linked among themselves in some persistent manner resulting in non pure-randomness. Persistence is the tendency of variables to have a carryover effect, for the immediate antecedent conditions to influence later conditions. The persistence and carryover effect are related to the time interval between observations of such effects.

Stationarity

A simplification of the analysis of stochastic time series is obtained if, for all times t, statistical parameters of the distribution do not change. Such processes, for which the statistics do not depend on the instant t at which the sample is being taken, are called stationary. A stationary process is thus a process in which ensemble averages, such as the autocovariance, are dependent only on the time difference $\tau = t_1 - t_2$, that is, they are invariant with respect to a translation of the time origin.

Most hydrologic time series are regarded as being stationary because many mathematical methods developed for treating random time series either require that the series be stationary or are concerned with reducing the series to approximate stationarity.

Two types of stationarity are commonly distinguished; strict stationarity implies that all population parameters are independent of the time itself. This type of stationarity is difficult to prove and so the idea of a weakly stationary process has been introduced in which the mean, variance, and autocorrelation are independent of time. This is sometimes termed a second order stationary process. Similarly, a process in which the third moment or the covariance of three values is independent of time, is called a third order stationary series.

Since most hydrologic processes in practice show periodic or seasonal fluctuations, they do not satisfy even the weakly stationary conditions of having a constant mean. A modified model taking this into account is described by the equation

$$X(t) = \mu_X(t) + y(t) \qquad \ldots\ldots(11)$$

where $\mu_X(t)$ is a stochastic property of some parameter of X, and

$y(t)$ is a stationary series.

A process in which all the properties change with time is termed non-stationary or evolutive.

As an example, an annual virgin flow with no significant change in river basin characteristics or climatic conditions for the period of record is considered as a stationary time series. If it is affected by man's activities in the river basin, or natural catastrophe, or slow modifications of the rainfall and runoff conditions the recorded or historical flow is a non-stationary time series. As such, the mathematics is complicated which further explains why most hydrologic processes are treated as stationary.

As a test for stationarity, the sample may be split into four or five parts and the means and variance of each section tested to see if they are from the same population, or if they are significantly different from each other. If it is shown that all the parts are from the same population, then the series is called quasi-stationary or self-stationary.

Having discussed time series in general, the generation models so far developed can be briefly summarized. These models can be subdivided into single station analysis and multivariate analysis.

Single Station Analysis

One of the simplest models to assume is a Markov linear model of lag one; that is, it is assumed that the variable at time t is a function of the antecedent value only.

$$Z_t = \rho_1 Z_{t-1} + \varepsilon_t \qquad \ldots\ldots(12)$$

where $Z = \frac{X-\bar{X}}{S}$ is the standardized value of the time series X.

ρ_1 is the lag one autocorrelation coefficient of the time series.

ε is an independent random variable.

In terms of variance, equation (12) can be expressed as

$$\text{var } Z_t = \rho_1{}^2 \text{ var } Z_{t-1} + \text{var } \varepsilon_t \qquad \ldots\ldots(13)$$

Because var Z_t = var Z_{t-1} = 1 the variance of the random component is

$$\text{var } \varepsilon_t = 1 - \rho_1^2 \qquad \text{.....(14)}$$

Therefore because it is more convenient in data generation to always use a random variable with unit variance, equation (12) can be modified to

$$Z_t = \rho_1 Z_{t-1} + (1 - \rho_1^2)^{\frac{1}{2}} \varepsilon_t \qquad \text{.....(15)}$$

This equation is suitable for generating values of a time series belonging to one population. If, however, the time series is composed of several sub-series, as for example monthly streamflow, where it might be said that the January streamflow belongs to a different population than the July streamflow, then equation (15) can be modified as:

$$Y = \bar{Y} + \rho_1 \frac{Sy}{Sx} (X-\bar{X}) + Sy (1 - \rho_1^2)^{\frac{1}{2}} \varepsilon_t \qquad \text{.....(16)}$$

where Y, $\bar{Y}$ and Sy are the flow to be simulated and the long term mean and standard deviation of the flows in the month being simulated, and

X, $\bar{X}$ and S_x are the flow in the antecedent month and the long term mean and standard deviation of the flows in the antecedent month.

In equation (16), Sy $(1 - \rho_1^2)^{\frac{1}{2}}$ represents the standard error of estimation and ε_t is a number drawn at random from a distribution with mean = 0 and standard deviation = 1.

The main disadvantage of this model is that, while preserving the serial correlation of lag 1, it does not preserve the correlation between other non-adjacent time periods. Roesner and Yevjevich (1968) found that series of monthly runoff in the western United States were well fitted by a Markov first order logarithmic model. Yagil (1963) introduced a method in which he considered annual flows to be independent of one another, and then generated monthly flows with a multiple lag model.

Let $a_{i,j}$ be the multiple regression coefficient of month j on month i (j = 2,3 --- 12 and i = 1,2 --- (j - 1))

R_j is the multiple correlation coefficient between month j and all preceding months.

$\bar{X}_j$ is the mean historical flow in month j.

X_j is the generated flow for month j in the year under consideration.

ε is a random normal deviate with a mean = 0 and a variance = 1.

S_j is the standard deviation of the historical flows of month j.

Then, X can be computed as:

$$X_1 = \bar{X}_1 + \varepsilon_1 S_1$$

$$X_2 = \bar{X}_2 + a_{1,2} (X_1 - \bar{X}_1) + \varepsilon_2 S_2 (1 - R_2{}^2)^{\frac{1}{2}}$$

$$\vdots \qquad \qquad \qquad \qquad .....(17)$$

$$X_n = \bar{X}_n + a_{1,n} (X_1 - \bar{X}_1) + \text{---} \; a_{n-1,n} (X_{n-1} - \bar{X}_{n-1}) + \varepsilon_n S_n (1 - R_n{}^2)^{\frac{1}{2}}$$

This model could easily be adapted to generate monthly flows without considering annual flows to be independent. A linear regression analysis with any number of desired lags could be done for each month, and the generation for each month carried out with an equation similar to the last equation above. Such an investigation was in fact described by Fiering (1967). However, despite the introduction of 20 lags in his regression analysis, he still failed to introduce as much long-term persistence into his record as existed in historical flows. This weakness, that is, the lack of long-term persistence in generated streamflows, is common to all methods developed prior to 1968.

Recent efforts have been concentrated on developing models which would preserve long-term persistence. Mandelbrot and Wallis (1968) introduced a model termed "self-similar fractional Brownian motion". This is a form of moving average representation using an extremely long memory.

Multivariate Analysis

If flows at more than one site are to be generated, it is necessary not only to take into account the interrelationship between flows at different stations but also to preserve the relevant characteristics at each site. Three multivariate models suggested by Fiering (1964), Beard (1965), and Matalas (1967) are discussed in the following sub-sections:

(a) Fiering (1964): The original variables X_1, X_2 --- X_n representing standardized flows at n stations are subjected to a principal component analysis.

$$Z_1 = a_{11} X_1 + a_{12} X_2 + \text{---} \; a_{1n} X_n$$

$$Z_2 = a_{21} X_1 + \text{------------} \; a_{2n} X_n$$

$$\vdots \qquad \qquad \qquad \qquad .....(18)$$

$$Z_n = a_{n1} X_1 \; \text{--------------} \; a_{nn} X_n$$

The coefficients a_{ij} in the above equations are determined from a standard eigenvector program. It can be shown that the principal components Z_1, Z_2 --- Z_n are independent of each other. Therefore, the principal components can be generated independently with any of the single station models described previously. The generated principal components can then be converted to standardized flows with the inverse matrix of eigenvectors.

(b) Beard (1965): Streamflows are first converted to standard normal deviates by taking logarithms, standardizing, and conversion with a Pearson Type III function. The deviates are then generated with a regression equation:

$$X_{1,t} = aX_{1,t-1} + bY_i + cX_{2,t} + dX_{3,t} + \text{-----} + \varepsilon_t (1 - R^2)^{\frac{1}{2}} \qquad \text{.....(19)}$$

where $X_{1,t}$ = deviate being generated.

$X_{1,t-1}$ = antecedent deviate at station being generated.

Y_i = logarithm of total flow for all stations for the 6 months preceding the antecedent month, transformed to a normal standard deviate.

$X_{2,t}$, $X_{3,t}$, --- $X_{n,t}$ = deviates for the same period at the other stations.

For each month, flows are generated at each station in turn. Those at the first station require only the first two terms on the right side of the equation, those at the second, the first three terms, etc.

The generated normal deviates are finally reconverted into flows.

(c) Matalas (1967): This model is an extension to Fiering's model. In addition to preserving the relevant correlations in the month being generated, the model developed by Matalas also preserves the cross-correlations of lag 1.

The basic equation (in matrix notation) given below is used:

$$X_{i+1} = A\ X_i + B\ e \qquad \text{.....(20)}$$

For m stations, X_{i+1} and X_i are standardized flows in successive time periods (each an m x 1 matrix). A and B are m x m matrices to be defined and e is an m x 1 matrix.

The matrix Mo is defined as the variance-covariance matrix, and M_1 as the covariance matrix with a lag of 1 time period. The matrix A can then be calculated.

$$M_1 = A\ Mo \qquad \text{.....(21)}$$

The matrix B can be calculated by solving the following equation with a principal component analysis:

$$B\ B^T = Mo - M_1\ Mo^{-1}\ M_1{}^T \qquad \text{.....(22)}$$

Young and Pisano (1968) presented a similar model which avoided the use of principal components.

B. Generating Uniformly Distributed Random Numbers

There are, generally speaking, two methods of generating uniformly distributed random numbers. The first involves some physical form of generating device, such as radiation from radioactive substances or electron noise in valves. The results of these random physical processes could then be transformed into a sequence of binary digits within a computer. The register in which the random numbers are generated is usually assigned an address within the general system of addresses in the computer storage. Then a reference to the random number device reduces to a reading from that store in the machine. The use of the random number device increases the speed of a computation because at every stage of operation of the computer a new random number appears in a fixed standard cell. There are, however, two disadvantages in such devices: first, there is a risk of instability, which must be countered by constant preventative maintenance; second, the results of a computation on the machine can never be exactly reproduced. Other random number generators, not involving computers, are dice, roulette, and other number games, and taking numbers from a telephone directory.

The second technique of generating uniformly distributed random numbers is to write a computer program to use some recurrence relation. This means that each successive number r_{j+1} is formed from the preceding number r_j (or from a group of preceding numbers) by applying some algorithm consisting of arithmetic and logical operations. Such a sequence of numbers is not random, but nevertheless, it may satisfy various statistical criteria of randomness and so is termed pseudo-random. The advantages of programming techniques for generating uniformly distributed pseudo-random numbers are the simplicity of the algorithm and the possibility of duplicating any computation.

Any program written to compute uniformly distributed pseudo-random numbers must comply with the following three requirements:

1. The program must generate numbers with extremely weak statistical autocorrelation. Any program written to produce a set of pseudo-random numbers must satisfy the established criteria for testing randomness.

2. The distribution function of the pseudo-random numbers generated by the program must approximate, as closely as possible, a uniform distribution.

3. The program must be stable. The distribution function of the pseudo-random numbers must not change during the running of the program.

A technique for generating uniformly distributed pseudo-random numbers based on the extraction of middle digits of products was proposed by J. von Neumann (1951). An arbitrary number α_o is taken as the start of the recurrence process, where α_o consists of an even number, 2k, of binary digits. The number α_o is squared, producing a 4k digit number α_o^2. The next number in the sequence, α_1, is taken as the central 2k binary digits [from the (k+1)th to the 3kth inclusive] of α_o^2. A significant improvement in the results is obtained if a pair of numbers α_o and α_1 are chosen arbitrarily, multiplied together and the central digits used as the number α_2. The process would then continue, calculating α_3 from α_1 and α_2, etc.

Lehmer (1949) developed a more sophisticated method that produces a sequence which differs less from a uniform distribution than do the earlier methods. Each of the above methods by truncation generates a periodic sequence, with a period not exceeding 2^{2k}.

A second technique for generating pseudo-random numbers is based on the application of residues. Lehmer (1949) used the recurrence relation

$$\alpha_{n+1} = k\alpha_n (\text{mod } m) \qquad .....(23)$$

The following recurrence relation has been used to generate pseudo-random numbers uniformly distributed over the interval (0,1):

$$\alpha_{n+1} = 2^{-42}\ \beta_n \qquad .....(24)$$

$$\beta_{n+1} = 5^{17}\ \beta_n (\text{mod } 2^{42}) \qquad .....(25)$$

with $\beta_o = 1$. Such a relation has a period of around 2^{40} or approximately 10^{12}.

There are also several techniques for generating uniformly distributed pseudo-random numbers which exploit the peculiarities of particular computers. These methods are based on the imitation of random chaotic processes by shifting the digits of the mantissae of pseudo-random numbers.

Sobol (1958) published a program that, in three operations, computes one number in the sequence:

1. A number α_k is multiplied by 10^{17}.

2. The product $10^{17}\ \alpha_k$ is shifted seven places to the left.

3. The modulus of the resulting number is normalized, and the result is taken as α_{k+1}.

The I.B.M. 360 system uses a program in which α_o is any odd integer with nine or less digits. The integer α_o is then multiplied by 65,539. If the resulting integer is negative then the figure 2,147,483,647 is added to it. The figure 65,539 (α_o) (plus 2,147,483,647 if necessary) is now converted to floating point and multiplied by 0.4656613×10^{-9}. The result is a pseudo-random number belonging to a distribution function uniformly distributed between 0 and 1. This procedure has a periodicity of 2^{29} terms.

C. Randomness Tests

Once a set of supposedly uniformly distributed random numbers or pseudo-random numbers have been generated it is necessary to subject them to various tests of their randomness, before using them.

Kendall and Babington-Smith (1938, 1939a, 1939b) have developed a series of four tests for checking the randomness of a distribution. In all tests the random numbers are classified according to criteria, which vary from test to test, and the empirically generated distribution is compared with a theoretical distribution, generally using as means of comparison the χ^2 criterion, the Kolmogorov criterion, and the ω^2 criterion.

1. Frequency test: This consists of counting the numbers of pseudo-random numbers being sampled that fall within the intervals of a dissection of the domain of definition of the pseudo-random numbers. Usually the range of the distribution is divided into ten or twenty equal intervals.

2. Serial test: The number of zeros and ones in the various places of the set of supposedly random numbers being tested are counted. In the case of random numbers, which are uniformly distributed over the interval (0,1), the mathematical expectation of the digits occurring in each place of the mantissa of a random number is $\frac{1}{2}$ since the probability of the occurrence of a zero or a one is $\frac{1}{2}$.

3. Gap test: This is a sequential test in which the generated random numbers less than $\frac{1}{2}$ are assigned to class A, and numbers greater than $\frac{1}{2}$ are assigned to class B. The sampled values of the numbers of each class, which occur in sequences and the lengths of sequences of numbers of the first and second classes, are compared with their theoretical limits.

4. Poker test: The numbers of various combinations of binary digits in a large sample set are counted. For example, the distribution test might be made on 10 binary places; ten zeros, nine zeros, and one one, eight zeros and two ones, etc.

If the supposedly random numbers are really pseudo-random, i.e., have been generated by a computer program, then an additional test must be made to ensure that the pseudo-random numbers imitate truly random numbers as far as periodicity parameters are concerned. It is advisable that the number of pseudo-random numbers used should not exceed this periodicity, otherwise the statistical process would be simulated by recurring pseudo-random numbers.

Let the periodicity be L so that the first L successively generated pseudo-random numbers are distinct but the (L+1)th pseudo-random number coincides with one of the α_i earlier numbers. Thereafter the sequence of pseudo-random numbers, beginning with α_1 and ending with α_L, will recur periodically. For pseudo-random numbers distributed over the interval (0,1), the number of distinct pseudo-random numbers, N, available is $N = 2^l$ where l is the number of digits in the mantissa on the computer. If successive lengths of the interval of aperiodicity $L_1, L_2, \ldots L_m$ are arrived at experimentally using a certain computer then the variable

$$z = \frac{\sum_{i=1}^{m} L_i^2}{N} \qquad \ldots\ldots(26)$$

will be distributed as Chi-square with 2m degrees of freedom, and so a test using the Chi-square criterion will indicate the value of the length of the interval of aperiodicity.

A set of pseudo-random numbers which satisfies all these tests is termed "locally random".

D. Generating Random Numbers Following Other Distributions

Given a set U_i, i = 1,2,... of random numbers uniformly distributed over the interval (0,1), it is possible to transform this set mathematically to a set of random numbers with any specified distribution. First consider the case in which the uniformly distributed random numbers are used to generate a sequence of normally distributed random numbers.

The most obvious approach to this transformation, as termed by Müller (1957), is the inverse method. To generate a normal deviate X from a uniform deviate U this approach derives an inverse relationship X = X(U), given that

$$U = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} e^{-\frac{t^2}{2}} dt \quad .....(27)$$

The relation X = X(U) is approximated stepwise by dividing the interval of U(0,1) into sub-intervals and using Chebyshev polynomials. If this approach is to be efficient, the approximations to X = X(U) should be designed to work over sub-intervals of U such that the lengths of sub-intervals are a negative power of two for computers operating in the binary mode. At the expense of utilizing a large memory space, it is possible to develop a good degree of accuracy with an extremely fast procedure.

Box and Müller (1958) have developed a direct method of transformation that gives a higher degree of accuracy than the inverse method, at a comparable speed. If U_1 and U_2 are two independent random numbers from the same uniformly distributed density function in the interval (0,1), then the random variables

$$X_1 = (-2 \log_e U_1)^{\frac{1}{2}} \cos 2\pi U_2 \quad .....(28)$$

and $$X_2 = (-2 \log_e U_1)^{\frac{1}{2}} \sin 2\pi U_2 \quad .....(29)$$

will be a pair of independent random numbers from the same normal distribution with mean zero and variance unity. This is justified from the inverse relationships

$$U_1 = e^{\frac{-(X_1^2 + X_2^2)}{2}} \quad .....(30)$$

$$U_2 = -\frac{1}{2\pi} \tan^{-1} \frac{X_2}{X_1} \quad .....(31)$$

from which the joint probability density of X_1, X_2 is

$$f(X_1,X_2) = \frac{1}{2\pi} e^{\frac{-(X_1^2+X_2^2)}{2}} \quad .....(32)$$

$$f(X_1,X_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{X_1^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{X_2^2}{2}} \quad .....(33)$$

$$f(X_1,X_2) = f(X_1) \quad f(X_2) \ , \quad .....(34)$$

i.e. X_1, and X_2 are normally distributed with mean zero and variance unity and are independent.

A rejection approach has been originated by von Neumann (1951) and developed by Teichroew (1953); normal deviates in the truncated region $-b \leq X \leq b$ are generated from

$$Y = -2b^2(U_1 - \tfrac{1}{2})^2, \quad .....(35)$$

if $\log_e U_2 \leq Y$ then the normal deviate is

$$X = b(2U_1 - 1), \qquad \text{.....(36)}$$

if $\log_e U_2 > Y$ then the pair of uniformly distributed random numbers (U_1,U_2) is rejected and the process repeated. For the normal distribution, this is an inefficient generation technique, especially if precise tail values are required. The probability that a pair (U_1,U_2) will generate a normal deviate, that is,

$$P \left[U_2 \leq e^{-2b^2(U_1-\frac{1}{2})^2}\right] \qquad \text{.....(37)}$$

is assymptotically

$$\frac{1}{b}\sqrt{\frac{\pi}{2}} \qquad \text{.....(38)}$$

A further method, known as "approximation by curve fitting" has been developed by Teichroew (1953). A fixed number of uniform deviates is summed and an improved approximate normal deviate is obtained using an interpolating Chebyshev polynomial.

The approximate normal deviate X appears, using a truncated series for ease of computation, as

$$X = a_1r + a_3r^3 + a_5r^5 + a_7r^7 + a_9r^9 \qquad \text{.....(39)}$$

where $a_1 = 3.949846138$,

$a_3 = 0.252408784$,

$a_5 = 0.07652912$,

$a_7 = 0.008355968$,

$a_9 = 0.029899776$,

and $r = (z-6)/4$ .....(40)

where $z = \sum_{i=1}^{12} U_i$ .....(41)

and U_i being the uniformly distributed random number as before.

The disadvantage of this method is that the value of z must be restricted to the range (2,10), meaning that normal deviates cannot be generated much beyond four standard deviations from the mean.

Schreider (1960) lists a further algorithm,

$$X = Z - \frac{41}{13400n^2}(Z^5 - 10Z^3 + 15Z) \qquad \text{.....(42)}$$

where $$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \qquad \ldots..(43)$$

which for practical purposes needs n to be only 2 to give a very good approximation to the normal distribution.

A method of rational approximation to transform a uniform deviate to a normal deviate has been suggested by Hastings (1965):

$$X = X^* \ (q) = Z - \frac{(a_o+a_1z+a_2z^2)}{(1+b_1z+b_2z^2+b_3z^3)} \qquad \ldots..(44)$$

where $$Z = \sqrt{\log \frac{1}{q^2}} \qquad \ldots..(45)$$

and $$q = \frac{1}{\sqrt{2\pi}} \int_{x(q)}^{\infty} e^{-\frac{t^2}{2}}, \ 0 < q \leq 0.5 \qquad \ldots..(46)$$

and $a_o = 2.515517$ $b_1 = 1.432788$

$a_1 = 0.802853$ $b_2 = 0.189269$

$a_2 = 0.010328$ $b_3 = 0.001308$

This method is very reliable, generally producing an absolute error of less than 4×10^{-4}. However, faster procedures requiring less memory storage are available.

Probably the most well known transformation technique utilizes the Central Limit Theorem. Given a set of identically distributed variables $U_1, U_2 \ldots U_k$, with each having a mathematical expectation of a* and a variance of $(\sigma^*)^2$, then the sum

$$X = U_1 + U_2 + \ldots U_n \qquad \ldots..(47)$$

will be asymptotically normal with a mathematical expectation

$$a = a^*n \qquad \ldots..(48)$$

and standard deviation

$$\sigma = \sigma^*n^{\frac{1}{2}} . \qquad \ldots..(49)$$

An initial set of random variables uniformly distributed over the interval (0,1) will have a mean of 0.5 and a standard deviation of $\frac{1}{2\sqrt{3}}$; therefore, the sum of n of such random numbers will have an expected value of

$$a = 0.5n , \qquad \ldots..(50)$$

and a standard deviation of

$$\frac{1}{2}\sqrt{\frac{n}{3}} \quad . \qquad .....(51)$$

If, however, the uniform distribution is machine generated, and therefore pseudo-random, the standard deviation will be

$$\sigma^* = \frac{1}{2\sqrt{3}} \sqrt{\frac{2^k+1}{2^k-1}} \qquad .....(52)$$

where k is the number of digits in the mantissa of the computer; the standard deviation of the sum of n pseudo-random numbers will be

$$\sigma = \frac{1}{2}\sqrt{\frac{n}{3}}\sqrt{\frac{2^k+1}{2^k-1}} \quad . \qquad .....(53)$$

For most cases, where the pseudo-random numbers are generated with an adequate number of digits, equation (51) will be more than adequate.

An increase in the number, n, of terms in the summation will result in the distribution of X being a better fit to the normal, although this will lead to an increase in the number of arithmetic operations needed for the transformation.

To produce a normal distribution with a mathematical expectation of zero and a variance of unity, a common expression is

$$X = \sum_{i=1}^{12} U_i - 6.0 \quad . \qquad .....(54)$$

The expectation is then seen to be $\frac{12}{2} - 6 = 0$ and the standard deviation is $\frac{1}{2}\sqrt{\frac{12}{3}} = 1$.

The problem of comparing the accuracy of this approach with others is complicated because the Central Limit Theorem is concerned with an asymptotic convergence in probability. A direct measure of accuracy is available by comparing the actual distribution function of the sum of a finite number of uniform deviates to the limiting normal distribution function. For example, using equation (54), the probability of X being greater than 3.0 above the mean is 0.100700×10^{-2}. Yet this probability point, for a normal distribution with the same mean and variance, gives a value of 3.0882. The difference is -0.0882.

Despite some loss of accuracy above 3 standard deviations from the mean (which could be reduced by increasing n in equation (54)), the method is very convenient, quick, and requires little memory space.

Now consider two common cases in which the distribution to be generated is non-normal.

(a) Log-normal: For a 2 parameter log-normal distribution with the mean of the logarithms, U_n, and the standard deviation of the logarithms,

σ_n, the value of any deviate is given by

$$X_i = e^{\mu_n + \varepsilon_i \sigma_n} \qquad .....(55)$$

where ε_i is a standardized normally distributed random number.

(b) A gamma distribution with 2 or 3 parameters: for any gamma distribution a deviate can be defined as

$$X_i = \frac{1}{2} \sum_{i=1}^{2\alpha} \varepsilon_i^2 \qquad .....(56)$$

where α is a multiple of 1/2.

Using the gamma 2 parameter function the distribution is then

$$P(X) = \frac{m^\alpha X^{\alpha-1} e^{-mX}}{\Gamma\alpha} \qquad .....(57)$$

or the gamma 3 parameter function by

$$P(X) = \frac{1}{m\Gamma\alpha} \left(\frac{X-b}{m}\right)^{\alpha-1} e^{-\left(\frac{X-b}{m}\right)} \qquad .....(58)$$

where m is a scale factor, and

b is a lower boundary.

To generate a gamma function of N values it is necessary to produce 2α. N normally distributed random numbers.

Finally, consider the case in which the random number distribution to be generated is empirical, that is, it cannot be defined, or it is not worthwhile to define the distribution mathematically. If, for example, a simple linear regression equation

$$y = a + bX \qquad .....(59)$$

were being used to relate two streamflow records, then in order to use one streamflow record to extend or fill in gaps in the second record it would be necessary to generate random numbers defined only over the period of joint record as

$$\varepsilon_i = y_i - \hat{y}_i \qquad .....(60)$$

where y_i is the recorded streamflow, and

$\hat{y}_i$ is the computed streamflow.

A mathematical description of the distribution of the residuals is not easily available but the cumulative probability distribution can be easily plotted on linear graph paper, smoothed, and extrapolated to either extreme. Similarly the cumulative probability distribution of generated uniformly distributed random numbers can easily be plotted since this must be a straight line joining the points (0, min.), (1, max.) where the first coordinate in each bracket refers to cumulative probability and the second to the expected minimum and maximum values of the residuals. Now, the graph is entered at the generated value of the uniformly distributed random number and the two cumulative probability distributions are used to convert this value to a random number following the empirical distribution of the original residuals.

The greatest inaccuracy of the method occurs, of course, in the extension of the empirical distribution to its extremes. This extension of the empirical distribution can be either by extrapolation or by fitting a theoretical distribution.

In the latter case it would be most important that the theoretical distribution and the observed frequencies were well-matched at the extreme being studied. Either a theoretical distribution could be chosen, a priori, for some reason, or several distributions could be tested and the one giving the best fit would be used. The Chi-square test of goodness of fit between empirical and theoretical distributions could be used with weighted values of class observations, most weight being given to the class covering the extreme values. The sum of the weights would be unity.

E. Testing Generated Data

Once a set of data has been generated to specified requirements, it must be tested to ensure that it meets those requirements and faithfully duplicates the statistical properties of the 'parent' data. Assuming that tests for circularity and extremes as well as the Kendall tests have been carried out on the random numbers used in the generation, then the only tests that are required at this stage are comparisons between properties of the generated data and the parent data such as the following.

1. Comparison of basic statistics: The basic statistics that should theoretically be preserved in the simulated data are twofold;

   a) Parameters of the probability distribution of the original data such as mean and variance.

   b) Parameters of the time dependence such as serial correlation coefficients, up to whatever order of lag was simulated.

2. Autocorrelation test: The autocorrelation test should test calculated autocorrelation coefficients up to a lag of 25 or 30 units for both recorded and simulated data, and compare the two series.

3. Spectral analysis test: The spectrum demonstrates the proportion of the total variance contributed by each frequency (Granger, 1964). The analysis may be carried out by applying harmonics to the autocorrelation function to reveal cycles and long term trends.

4. Test for long-term persistence: Hurst (1951, 1956), proposed a measure of long-term persistence as

$$\frac{R}{S} = \left(\frac{N}{2}\right)^k \qquad \ldots\ldots(61)$$

where R is the volume of storage necessary to maintain the average flow for the period of record,

S is the standard deviation of the data,

N is the number of years of record, and

K is a measure of persistence.

The values of K computed for the observed and simulated data,should be compared.

5. Duration analysis.

6. Non-parametric tests:

a) Number of values above the mean compared to the number of values below the mean.

b) Number of quartile changes.

c) Cluster test.

CONCLUSION

The report has covered the background information necessary to the use of data generating techniques. Several of the techniques currently used have been described. Most of these methods are available as computer programs and are relatively easy to use. One of the most important uses of data generation is in estimating floods, droughts or storage requirements at a given return period. While empirical or analytical techniques can give only one estimate of the design parameter, data generation can be used to obtain a best estimate of the design parameter and confidence limits on the estimate.

As an example of the efficiency of the data generation technique, consider the problem of estimating the maximum monthly discharge in any year, which will have a likely return period of, say, 100 years from a record consisting of 30 years of mean monthly flows. An empirical or analytical solution, such as fitting a distribution to the annual maxima, would use only 30 pieces of information. On the other hand the data generation technique would use all 30 x 12 segments of information in order to simulate a sequence of data from which the required parameter could be determined.

The accuracy of the results of the data generation method depends on how accurately the recorded time series can be broken down into mathematically describable terms. The most important question is the distribution of the random component in the model.

A further point of interest is how to determine the number of deviates to simulate. In many cases this can be determined from known objectives or by comparing the cost of increasing the sample size with the expected benefits due to the increased accuracy. Chow and Ramaseshan, (1965), describe a purely statistical technique.

If P_n is the proportion of a sample size n from a population which can be estimated within an error level of α % of its true value at a β % confidence level, then the required sample size is

$$n \geqslant \left(\frac{t_\beta}{\alpha}\right)^2 \left(\frac{1-P_n}{P_n}\right)$$

where t_β is the standard normal deviate corresponding to the β % confidence level.

BIBLIOGRAPHY

Beard, L.R. 1965. Use of interrelated records to simulate streamflow. ASCE, September 1965.

Box, G.E.P., and M.E. Müller. 1958. A note on the generation of normal deviates. Ann. Math. Stat., 28: 610-611.

Chow, V.T., and S. Ramaseshan. 1965. Sequential generation of rainfall and runoff data. ASCE, vol. 91, HY4, July 1965.

Fiering, M.B. 1964. Multivariate technique for synthetic hydrology. ASCE, September 1964.

Fiering, M.B. 1967. Streamflow synthesis. Harvard University Press.

Fisher, R.A. 1929. Tests of significance in harmonic analysis. Roy. Soc. London, Proc., Ser. A, vol. 125, no. 796: 54-59.

Granger, C.W.J. 1964. Spectral analysis of economic time series. Princeton University Press.

Hastings, C. 1965. Approximations for digital computations. Princeton University Press.

Hazen, A. 1914. Storage to be provided in impounding reservoirs for municipal water supply. ASCE.

Hufshmidt, M., and M.B. Fiering. 1966. Simulation techniques for design of water-resources systems. Harvard University Press.

Hull, T.E., and A.R. Dobell. 1962. Random number generators. SIAM Review, 4, 229-254.

Hurst, H.E. 1951. Long-term storage capacities of reservoirs. Trans. ASCE, 116, no. 776.

Hurst, H.E. 1956. Methods of using long-term storage in reservoirs. Proc. Inst. Civil Engineers, Paper 6059.

Hurst, H.E., R.P. Black, and Y.M. Simaika. 1965. Long-term storage. Garden City Press.

Kendall, M.G., and B. Babington-Smith. 1938. Randomness and random sampling numbers. J. Roy. Statistical Soc., 101, no. 1: 147-166.

Kendall, M.G., and B. Babington-Smith. 1939a. Second paper on random sampling numbers. J. Roy. Statistical Soc., 6, Suppl. no. 1: 51-61.

Kendall, M.G., and B. Babington-Smith. 1939b. Random sampling numbers. Tracts for computers, 24, London.

Johnson, D.L. 1956. Generating and testing pseudo-random numbers on the IBM Type 701. Math. Tables Aids Comp. 10, 8-13.

Kite, G.W. 1969. A method of statistical inference. M.S. thesis, Colorado State University.

Lehmer, D.H. 1949. Mathematical methods in large scale computing units. Proc. Symposium on Large Scale Digital Calculating Machinery, Harvard University Press.

Mandelbrot, B.B., and J.R. Wallis. 1968. Noah, Joseph and operational hydrology. Water Resources Research, October 1968.

Matalas, N.C. 1967. Mathematical assessment of synthetic hydrology. Water Resources Research, Fourth Quarter 1967.

McCracken, D.D. 1955. The Monte Carlo method. Sci. Am., 192, 90, May 1955.

Megerian, E., and R.L. Pentland. 1968. Simulation of Great Lakes basin water supplies. Water Resources Research.

Meyer, H.A., ed. 1956. Symposium in Monte Carlo methods. Wiley and Sons, New York.

Moshman, J. 1954. The generation of pseudo-random numbers on a decimal calculator. J. Assoc. Computing Machinery, 1: 88-91.

Müller, M.E. 1958. An inverse method for the generation of random normal deviates on large-scale computers. Math. Tables Aids Comp. 12: 167-174.

Müller, M.E. 1957. Generation of normal deviates. Technical Report No. 13, Statistical Techniques Research Group, Department of Mathematics, Princeton University.

Müller, M.E. 1959. A comparison for generating normal deviates on digital computers. J. Assoc. Computing Machinery, 6: 376-383.

von Neumann, J. 1951. Various techniques used in connection with random digits. NBS Appl. Math. Series, no. 12: 36-38.

Pentland, R.L., H. Rosenberg and G. Cavadias. 1968. Digital computer applications to Great Lakes regulation. Proceedings of the Symposium on the Use of Analog and Digital Computers in Hydrology, Tucson, Arizona.

Rand Corp. 1955. One million random digits and 100,000 normal deviates. The Free Press, Glencoe, Illinois.

Rippl, W. 1883. The capacity of storage reservoirs for water supply. Proc. Institute of Civil Engineers.

Roesner, L.A., and V. Yevjevich. 1968. Mathematical models for time series of monthly precipitation and monthly runoff. Hydrology Paper No. 15, Colorado State University.

Schreider, Y.A., ed. 1960. The Monte Carlo method, the method of statistical trials. Translated from the Russian by G.J. Tee, Pergamon Press, London.

Schuster, A. 1898. On the investigation of hidden periodicities with application to a supposed 26 day period meteorological phenomena. Terrestrial Magnetism, vol. 3, no. 1: 13-41.

Sobol, I.M. 1958. Pseudo-random numbers for the machine "STRELA". In Russian: Teoriya Veroyatn. I Yehe Primen., 3, No. 2: 205-211.

Sudler, C.E. 1927. Storage required for the regulation of streamflow. Trans. ASCE, 91, 622.

Taussky, O., and J. Todd. 1956. Generation and testing of pseudo-random numbers. Symposium in Monte Carlo methods, H.A. Meyer, ed., John Wiley and Sons, New York, 15-27.

Teichroew, D. 1953. Distribution sampling with high speed computers. Ph.D. Thesis, University of North Carolina.

Walker, Sir Gilbert. 1925. On periodicity. Roy. Meteorol. Soc. J., vol. 51: 337-346.

Yagil, S. 1963. Generation of input data for simulations. IBM Systems Journal, September-December 1963.

Young, G.K., and W.C. Pisano. 1968. Operational hydrology using residuals. ASCE, July 1968.

Environment Canada Library, Burlington
3 9055 1017 3394 6

Current Technical Bulletins

No. 25 Stream Gauging Techniques for Remote Areas Using Portable Equipment. R. Kellerhals, 1970.

A review of streamflow measuring techniques applicable to rivers with peak flows up to 10,000 cfs.

No. 26 The Control of Eutrophication. Prepared by staff of the Canada Centre For Inland Waters, Burlington, Ontario; Fisheries Research Board of Canada, Winipeg; Inland Waters Branch, Ottawa, 1970.

A discussion of the respective roles of phosphorus, nitrogen and carbon as critical elements in limiting the eutrophication process.

No. 27 An Automated Method for Determining Mercury in Water. P.D. Goulden and B.K. Afghan, 1970.

A report describing a method for determining the mercury content in water containing mercury concentrations as low as 0.05 μg/l.

No. 28 An Assessment of the Wave Agitation in the Small Boat Basin at the Canada Centre For Inland Waters. T.M. Dick, 1970.

A discussion of the results obtained from a model study of wave action in the small boat basin at the Canada Centre For Inland Waters, Burlington, Ontario.

No. 29 Measurement of Discharge Under Ice Conditions. P.W. Strilaeff and J.H. Wedel, 1970.

An outline of the difficulties encountered in the measurement of discharge under ice cover and a discussion of a possible technique for estimating river discharge using a single velocity in a cross-section.

No. 30 Prediction of Saturation Precipitation of Low Solubility Inorganic Salts from Subsurface Waters under Changing Conditions of Total Concentration, Temperature and Pressure. R.O. van Everdingen, 1970.

A report containing graphs that enable the determination of the degree of (under- or over-) saturation of aqueous solutions with respect to $BaSO_4$, $CaSO_4$, $SrSO_4$, BaF_2, CaF_2 and MgF_2 under a variety of conditions of temperature, pressure and total salt concentrations. Also presented are examples of the influence of temperature changes, dilution, evaporation, addition of common salt, and mixing, on the degree of saturation of the above solutions.

No. 31 A Hydrologic Model of the Lake Ontario Local Drainage Basin. D. Witherspoon, 1970.

A discussion of a hydrologic model proposed for the Lake Ontario local drainage area. The basic principles used in the model are those of water and energy balances. Using estimates of actual evaporation, realistic values of the regional moisture are obtained which, when routed, simulate the measured outflow.

No. 32 Identification of Petroleum Products in Water. A. Demayo, 1970.

A description of an extraction method used in the Water Quality Division laboratory to analyze water samples, and activated carbon samples through which water has been passed, for the presence of crude oil or other petroleum product.

No. 33 Seasonal Variations, Sulphur Mountain Hot Springs, Banff, Alberta. R.O. van Everdingen, 1970.

A study of seasonal variations in the physical and chemical parameters of the sulfurous hot springs on Sulphur Mountain, near Banff, Alberta. In the absence of accurate discharge measurements, only a "minimum required" mixing ratio could be calculated, leading to minimum ion concentrations and a minimum temperature for the cooler water.

No. 34 Instrumentation for Study of Energy Budget of Rawson Lake. R. Chapil, 1970.

A report describing the equipment and procedures used, and some graphical results obtained mainly during 1960 in the Rawson Lake study, a hydrological study of a small research basin in northwestern Ontario.

No. 35 Precipitation of Heavy Metals from Natural and Synthetic Acidic Aqueous Solutions during Neutralization with Limestone. R.O. van Everdingen and J.A. Banner, 1970.

A report describing a method in which iron, aluminum, manganese, copper, lead and zinc in natural and synthetic acidic water with H^+ concentrations ranging from 4.0×10^{-3} to 6.3×10^{-4} are circulated through crushed limestone, resulting in the neutralization of the acidity and removal of varying amounts of the metals.

A complete list of titles in the Technical Bulletin Series and copies of any of these publications may be obtained from the Director, Inland Waters Branch, Department of Energy, Mines and Resources, Ottawa, Ontario.