An algorithm is given to generate additional input data for simulations when some, but insufficient, historical data are available. The additional data generated are statistically "similar" to the historical.

Motivation and application of the algorithm are demonstrated by means of a problem related to the monthly water inflow to Lake Tiberias which had to be resolved in connection with the "Israeli Integrated Water Supply" project now under construction.

The algorithm is a variant of a method previously used by Thomas and Fiering in hydrological studies.

Generation of input data for simulations

by S. Yagil

the problem in general

Simulation of large projects and of complex processes on digital computers has become a common practice. Choice of proper input data for such a simulation nearly always constitutes a serious problem because of the difficulty in predicting what the actual data will be when the project or process is put into operation. Examples are costs and prices (in market simulations and business games), enemy and own strategy and strength (in war games), atmospheric and space conditions (in flight simulations), raw material and power supply (in production simulations), climatic conditions and water flows (in hydrological and hydroelectric simulations), etc.

Often the available historical data which could be used are insufficient for conducting the desired simulation because they cover too short a period. We are usually tempted to use some average values obtained from these historical data in our simulation, possibly making some adjustments for known trends. These average values would pertain to a certain time unit: cost of a commodity in January, February, etc. (over a set of years); average power supply between 8 and 9, 9 and 10 o'clock etc. (over a set of days); etc. Note the cyclical character of such historical data resulting from some natural cycle (24 hours per day, 52 weeks or 12 months per year, etc.).

However, the use of the average values for each time unit in the cycle is in most cases unacceptable because of the presence of random fluctuations about those averages. At the other extreme,

the use of a sequence of random numbers is also unacceptable because, in general, the data should "resemble" the historical ones. Thus, we have the problem of creating a "pseudo-random" sequence of numbers having statistical properties which will be similar to those of the historical data. The statistical properties considered will usually be means (averages) and variances (fluctuations) for each time unit and correlations between time units.

specific example

A specific example, namely the "Israeli Integrated Water Supply Scheme," for which the required sequence was generated on an IBM 1620 will be used to illustrate the problem and its solution. This project, now being constructed, is a large water supply system consisting of a main conduit which delivers water from Lake Tiberias to the arid southern parts of the country. Along its route it meets a number of existing smaller projects, each of which it supplies with water, or draws water from, or both, according to seasonal water supply, demand and allocation.

The main input into the system is the monthly net inflow into Lake Tiberias (Table 1) which has been recorded for the past 35

Table 1 Monthly inflows into Lake Tiberias for 35 years in millions of cubic meters

Year	*Oct.*	*Nov.*	*Dec.*	*Jan.*	*Feb.*	*Mar.*	*Apr.*	*May*	*June*	*July*	*Aug.*	*Sept.*
1	3.0	2.0	3.0	46.0	193.0	193.0	79.0	38.0	4.0	5.0	−5.0	−3.0
2	8.0	10.0	123.0	77.0	202.0	119.0	102.0	43.0	15.0	−8.0	−13.0	−5.0
3	8.0	10.0	19.0	52.0	133.0	166.0	50.0	20.0	11.0	3.0	−3.0	−16.0
4	−24.0	45.0	69.0	135.0	400.0	102.0	116.0	39.0	44.0	16.0	13.0	−3.0
5	−28.0	65.0	72.0	65.0	155.0	87.0	48.0	43.0	13.0	3.0	−4.0	4.0
6	−11.0	−18.0	70.0	71.0	284.0	134.0	38.0	19.0	38.0	26.0	12.0	−32.0
7	6.0	36.0	35.0	51.0	156.0	31.0	27.0	20.0	5.0	−2.0	−11.0	−7.0
8	2.0	−7.0	7.0	51.0	119.0	23.0	26.0	11.0	6.0	−12.0	−11.0	−8.0
9	−47.0	34.0	28.0	33.0	121.0	72.0	27.0	17.0	−3.0	−9.0	−33.0	−6.0
10	0.0	−2.0	47.0	11.0	216.0	110.0	98.0	31.0	23.0	−2.0	1.0	0.0
11	11.0	30.0	35.0	41.0	58.0	46.0	24.0	7.0	−1.0	−7.0	−13.0	−23.0
12	−1.0	33.0	50.0	180.0	111.0	43.0	21.0	41.0	−10.0	−3.0	−4.0	−7.0
13	−9.0	21.0	−11.0	150.0	211.0	135.0	76.0	40.0	25.0	7.0	14.0	−42.0
14	17.0	42.0	27.0	73.0	109.0	130.0	81.0	40.0	10.0	4.0	−9.0	4.0
15	−5.0	21.0	36.0	119.0	149.0	102.0	67.0	34.0	14.0	8.0	−14.0	3.0
16	6.0	11.0	43.0	113.0	131.0	128.0	55.0	19.0	−30.0	8.0	−7.0	−11.0
17	−3.0	10.0	44.0	122.0	99.0	108.0	74.0	30.0	21.0	−7.0	−3.0	−1.0
18	22.0	35.0	40.0	143.0	187.0	155.0	120.0	71.0	−29.0	16.0	−10.0	−6.0
19	−9.0	14.0	30.0	144.0	130.0	88.0	59.0	37.0	10.0	0.0	−9.0	−11.0
20	−3.0	56.0	81.0	179.0	182.0	144.0	66.0	44.0	22.0	16.0	−1.0	2.0
21	13.0	11.0	40.0	58.0	121.0	124.0	61.0	55.0	23.0	2.0	1.0	5.0
22	−1.0	20.0	26.0	153.0	109.0	74.0	18.0	21.0	5.0	−2.0	−5.0	−8.0
23	−8.0	12.0	6.0	32.0	175.0	135.0	83.0	−41.0	−37.0	−26.0	−6.0	−6.0
24	−5.0	−17.0	2.0	93.0	183.0	37.0	144.0	82.0	49.0	27.0	2.0	8.0
25	27.0	20.0	25.0	173.0	107.0	91.0	60.0	44.0	25.0	8.0	0.0	1.0
26	0.0	29.0	20.0	33.0	45.0	43.0	36.0	9.0	1.0	−6.0	−14.0	−15.0
27	5.0	17.0	144.0	83.0	168.0	161.0	59.0	29.0	24.0	11.0	7.0	−5.0
28	−27.0	−7.0	35.0	94.0	115.0	174.0	113.0	41.0	34.0	10.0	7.0	4.0
29	12.0	73.0	100.0	200.0	263.0	111.0	114.0	58.0	34.0	17.0	22.0	14.0
30	18.0	43.0	70.0	41.0	49.0	50.0	35.0	19.0	3.0	−4.0	−12.0	−3.0
31	−10.0	30.0	100.0	146.0	93.0	94.0	48.0	38.0	18.0	18.0	9.0	−5.0
32	9.0	13.0	47.0	57.0	98.0	123.0	54.0	37.0	14.0	4.0	−1.0	−3.0
33	7.0	17.0	88.0	139.0	92.0	53.0	32.0	16.0	−1.0	−5.0	4.0	−2.0
34	4.0	4.0	38.0	51.0	75.0	85.0	42.0	23.0	5.0	−1.0	−4.0	0.0
35	0.0	14.0	8.0	73.0	28.0	41.0	30.0	9.0	−5.0	−24.0	−21.0	−17.0

years. The net inflow is the amount brought in by the Jordan river all year around and by flood waters in winter only, minus the amounts lost by evaporation and local pumping. In the peak summer months the net inflow may be negative.

A 1620 program which simulates the whole system has been written in cooperation with Water Planning for Israel, Ltd., which planned the entire project. The recorded data, covering 35 years, were used for the first runs. However, it was soon felt that for the purpose of long range planning, additional and larger sequences would be necessary for more simulation runs.

Monthly means and variances were computed as shown in Table 2. The correlation coefficients between all pairs of months were computed as well and are given in Table 3. Upon examination of Table 3 we note on the diagonal the highly significant correlations between adjacent months (with the exception of October–November and August–September). Correlations between months that are two apart (with the exception of October–December, January–March and July–September) are significant as well. Correlations are also high between each of the winter months December through March and each of the summer months May through August (with two exceptions).

The reasons for these significant correlations are as follows:

- The hydrological year in Israel begins at the start of winter, i.e. in October. Rainfall in the months September through November is rather erratic. Each may be quite dry or rainy independently of the other. This is the reason for the low correlations in this period. Thereafter, the year becomes either consistently rainy or consistently dry, which accounts for the high correlations between successive winter months.
- A rainy winter in the Jordan valley implies a snowy winter on Mt. Hermon, providing ample water flow in the summer thaw. This explains the high correlations between summer and preceding winter months.
- Finally, the correlations between successive summer months are probably spurious and result from the common correlation with the previous winter months.

The standard deviation of the total yearly inflows and the serial correlation between successive years were found to be 201.2 and 0.019. Assuming a normal distribution, this correlation coefficient is not satistically significant at the 5% level. This means that the hydrological years are independent of one another.

objectives

On the basis of this analysis, it was held desirable to generate a synthetic sequence of monthly inflows into Lake Tiberias, covering 400 years and having the following statistical properties:

- The total yearly inflows would be uncorrelated.
- The variance of the total yearly inflows would be close to the historical one.
- The monthly means and variance would be close to the historical ones.

Table 2 Monthly mean inflows and their standard deviations in M.C.M.

Month	Mean		Standard Deviation	
	Historical	*Generated*	*Historical*	*Generated*
Oct.	−0.4	−1.1	14.8	14.4
Nov.	20.8	21.4	20.9	21.9
Dec.	45.6	44.4	35.2	35.1
Jan.	93.8	94.0	51.6	52.0
Feb.	144.8	147.2	73.4	77.7
Mar.	100.3	100.6	44.9	43.5
Apr.	62.4	63.6	32.7	30.1
May	31.0	30.8	21.1	20.4
June	10.9	10.6	19.2	18.7
July	2.6	1.9	11.9	11.5
Aug.	−3.5	−4.5	10.7	10.8
Sept.	−5.7	−5.9	10.8	10.7

Note: "Historical" refers to the 35 recorded years appearing in Table 1. "Generated" refers to the sequence covering 400 years generated by the IBM 1620, part of which appears in Table 4.

Table 3 Correlation coefficients between inflows into Lake Tiberias

	Oct. 1	*Nov.* 2	*Dec.* 3	*Jan.* 4	*Feb.* 5	*Mar.* 6	*Apr.* 7	*May* 8	*June* 9	*July* 10	*Aug.* 11	*Sept.* 12
Oct.		.008	.050	.015	−.312	−.015	−.107	.042	−.201	−.112	−.017	.071
		.009	.053	.092	−.246	.008	−.004	.127	−.187	−.007	.054	.085
Nov.			.343	.369	.101	−.126	−.069	.158	−.093	.071	.028	.253
			.337	.353	.078	−.105	−.063	.131	−.098	.050	.006	.244
Dec.				.205	.264	.118	.019	.172	.238	.294	.347	.278
				.242	.234	.142	.054	.180	.236	.291	.337	.271
Jan.					.204	−.015	.172	.510	.209	.466	.452	.147
					.225	.066	.188	.462	.187	.440	.454	.117
Feb.						.352	.592	.230	.430	.525	.597	−.013
						.390	.583	.289	.434	.515	.548	.011
Mar.							.413	−.057	−.046	.164	.333	.072
							.470	.184	.096	.359	.368	.027
Apr.								.495	.328	.403	.415	.337
								.560	.379	.461	.413	.392
May									.535	.681	.279	.468
									.526	.710	.345	.388
June										.606	.539	.177
										.593	.591	.173
July											.580	.170
											.624	.142
Aug.												.048
												.053

Note: The lower figure in each entry pertains to the 35 years of recorded historical data, while the upper one refers to the generated sequence.

- The correlations between all pairs of months would be close to the historical ones.
- The monthly inflows could be assumed to be normally distributed about their means.

earlier work

Various methods of generating synthetic sequences of monthly flows have been suggested and used in the past. However, none of them satisfies all the requirements mentioned above. Sudler[1] suggests reshuffling the years in the historical record several times. This, however, leaves the years intact, i.e., creates no new combinations within years. Furthermore, the range for each month remains unchanged although it should grow when the record size grows.

Barnes[2] improved the method by using random numbers to generate monthly flows which were normally distributed about the monthly mean. This method still did not take into account the existing correlations between months.

Finally, a method which also takes into account correlations between successive months was developed by Thomas and Fiering[3]. This method is applicable when a Markov model can be used to represent the data. This, however, is not the case with inflows into Lake Tiberias, where the winter months are highly correlated with the following summer months as pointed out above. For example, the correlation between February and August is 0.548 (see Table 3). Also, Thomas' and Fiering's method creates a slight correlation between years, which is undesirable in our case.

algorithm

A new algorithm, which generates a sequence meeting all five objectives was developed. The algorithm is as follows:

- First, the multiple regression and multiple correlation coefficients of each month on all preceding months in the same hydrological year (i.e., going back to October) are computed from the historical data.
- Second, we let:

  $a_{i,j}$ denote the multiple regression coefficient of month j on month i, where $j = 2, \cdots, 12$ and $i = 1, \cdots, j-1$ where $j = 1$ refers to October, $j = 2$ refers to November, $\cdots$ and $j = 12$, refers to September,

  R_j denote the multiple correlation coefficient between month j and all preceding months,

  x_j denote the mean historical inflow in month j,

  y_j denote the generated synthetic inflow for month j in the year under consideration,

  t_j denote a random normal variate with mean 0 and variance 1, used in generating the inflow of month j,

  σ_j denote the standard deviation of the historical inflows of month j,

  and we compute y_i as follows:

Table 4 Part of the generated sequence of monthly inflows into Lake Tiberias in millions of cubic meters

Year	*Oct.*	*Nov.*	*Dec.*	*Jan.*	*Feb.*	*Mar.*	*Apr.*	*May*	*June*	*July*	*Aug.*	*Sept.*
1	1.6	72.1	64.2	133.6	170.0	79.1	90.8	55.3	.6	8.8	−21.4	3.9
2	−3.2	32.0	.1	101.3	47.6	66.3	49.1	29.5	−7.1	−10.7	−12.3	3.6
3	19.9	37.4	25.6	67.6	−30.0	114.7	17.2	31.5	35.8	4.3	−12.1	−2.2
4	17.2	10.5	32.9	137.2	93.7	97.3	67.6	84.6	56.5	21.9	8.6	3.1
5	−.8	49.9	45.1	90.5	176.7	132.9	49.7	43.0	13.0	−3.0	−4.9	1.8
6	−15.9	20.7	69.1	140.1	239.5	166.0	78.0	30.7	9.3	18.4	4.3	−11.4
7	−2.3	42.3	64.3	134.5	213.7	94.2	64.0	14.4	15.7	−9.2	.3	−10.1
8	−30.7	10.8	29.1	115.0	157.1	67.5	60.4	37.4	35.1	8.1	−5.1	.7
9	−23.0	15.0	43.6	147.8	306.2	132.9	74.1	24.1	35.3	10.1	.8	−38.3
10	−3.8	10.6	34.5	147.4	126.4	86.7	54.9	39.3	26.3	14.1	−.7	−19.9
11	8.4	1.6	88.7	93.5	184.9	118.7	82.9	25.0	2.4	−3.8	−6.4	−2.7
12	−11.1	44.6	53.8	162.5	211.3	37.5	73.3	49.2	9.1	1.7	1.6	−3.1
13	−4.0	46.6	−25.3	71.6	107.5	127.7	54.7	33.1	9.2	−3.1	−20.1	−8.2
14	−6.1	7.5	56.0	−.8	106.9	39.2	43.6	.8	3.4	−24.3	−33.5	−12.5
15	−35.2	18.1	100.4	46.9	232.2	157.9	48.6	−4.5	18.0	−11.1	10.2	−22.4
16	7.5	36.1	100.8	165.4	88.6	155.7	33.9	40.5	3.5	16.6	−8.7	.8
17	−13.2	−2.3	−15.1	110.4	151.8	75.7	88.4	42.3	−8.4	−9.9	−12.6	−4.0
18	−2.8	21.4	48.3	116.5	87.3	12.5	61.1	61.1	39.2	17.1	−2.8	3.3
398	13.9	19.8	81.0	93.7	−14.3	9.8	21.5	55.8	34.4	5.6	−14.3	−2.3
399	−16.7	48.2	82.0	107.4	170.1	34.1	82.7	61.9	34.5	13.4	−5.3	20.0
400	14.4	59.8	35.8	9.5	45.8	38.2	45.6	35.1	3.9	−10.2	−19.4	10.5

Table 5

	Historical	*Generated*
Mean yearly total (in MCM)	502.5	508.4
Standard deviation of yearly totals (in MCM)	201.2	195.1
Serial correlation of yearly totals (insignificant in both cases)	.019	−.011

$$y_1 = x_1 + t_1\sigma_1$$

$$y_2 = x_2 + a_{1,2}(y_1 - x_1) + t_2\sigma_2\sqrt{1 - R_2^2}$$

$$\vdots$$

$$y_n = x_n + a_{1,n}(y_1 - x_1) + a_{2,n}(y_2 - x_2) + \cdots + a_{n-1,n}(y_{n-1} - x_{n-1}) + t_n\sigma_n\sqrt{1 - R_n^2},$$

where n goes up to 12 in our case. The computation is repeated the desired number of times (400 in our example) to obtain the same number of independently generated sets of y_i.

The above procedure was programmed in FORTRAN II and, with the use of a 40K card IBM 1620 computer, a sequence corresponding to 400 years was generated. Part of this sequence is shown in Table 4. The statistical properties of the yearly totals are summarized in Table 5. The historical monthly means and variances are compared with the generated ones in Table 2 and correlations between months in Table 3. Discrepancies are in all cases very small and of no significance for practical purposes.

proof

We now establish that sequences generated by means of the algorithm satisfy the statistical objectives stated above.

It is obvious that the years are independent in this method because they are generated independently. Each y_i is normally distributed about its x_i because all the terms which are added to each x_i are normally distributed with means zero.

The proof that the generated sequence preserves the historical monthly means and variances, as well as the correlations between months, is given in an appendix.

The variance of yearly totals is the sum of all monthly variances plus twice the sum of monthly covariances. Since these are the same for the historical and generated sequences, it follows that the variance of yearly totals will also be the same for both.

Appendix: Proof of preservation of historical means, variances and correlations

We shall prove that the y_i generated by the algorithm preserve the historical monthly means, variances, and correlations.

First consider the means:

$$E(y_1) = x_1 + \sigma_1 E(t_1) = x_1$$

because t_1 is normal with mean 0. Similarly:

$$E(y_2) = x_2 + a_{1,2}E(y_1 - x_1) + \sigma_2\sqrt{1 - R_2^2}E(t_2) = x_2$$

because the last two terms are equal zero. Generally:

$$\begin{aligned} E(y_n) &= x_n + a_{1,n}E(y_1 - x_1) + \cdots \\ &\qquad + a_{n-1,n}E(y_{n-1} - x_{n-1}) + \sigma_n\sqrt{1 - R_n^2}E(t_n) \\ &= x_n. \end{aligned}$$

This shows that the expected values of generated monthly means are equal to the historical ones.

Now consider the variances. We have

$$\text{var } (y_1) = E(y_1 - x_1)^2 = E(t_1\sigma_1)^2 = \sigma_1^2$$

since t_1 is normal with mean 0 and variance 1. Generally,

$$\begin{aligned} \text{var } (y_n) &= E(y_n - x_n)^2 \\ &= E\{(y_n - x_n)[a_{1,n}(y_1 - x_1) + a_{2,n}(y_2 - x_2) + \cdots \\ &\quad + a_{n-1,n}(y_{n-1} - x_{n-1}) + t_n\sigma_n\sqrt{1 - R_n^2}]\} \\ &= a_{1,n}\sigma_{1,n} + a_{2,n}\sigma_{2,n} + \cdots + a_{n-1,n}\sigma_{n-1,n} \\ &\quad + E[(y_n - x_n)t_n\sigma_n\sqrt{1 - R_n^2}] \\ &= \left(\sum_{i=1}^{n-1} a_{i,n}\sigma_{i,n}\right) + \sigma_n^2(1 - R_n^2). \end{aligned}$$

Relative to proof of the last step, consider the general multiple regression equation,

$$z_n = a_1 z_1 + a_2 z_2 + \cdots + a_{n-1} z_{n-1} + p,$$

where p is the residual term. Assume all $E(z_i) = 0$. Let σ_n^2 denote the variance of z_n, then

$$\begin{aligned}\sigma_n^2 &= E(z_n^2) \\ &= E[z_n(a_1 z_1 + a_2 z_2 + \cdots + a_{n-1} z_{n-1} + p)] \\ &= a_1 \sigma_{1,n} + a_2 \sigma_{2,n} + \cdots + a_{n-1} \sigma_{n-1,n} + E(z_n p).\end{aligned}$$

The last term equals the variance of the residuals, $E(p^2)$, as shown in Cramer[4] page 305, which in turn equals $\sigma_n^2(1 - R_n^2)$ as shown on page 308. Thus,

$$\sigma_n^2 = \left(\sum_{i=1}^{n-1} a_{i,n} \sigma_{i,n} \right) + \sigma_n^2 (1 - R_n^2)$$

and hence,

$$\text{var } (y_n) = \sigma_n^2$$

as was to be shown.

Consider now the covariances. Let $m < n$, then

$$\begin{aligned}\text{cov } (y_m, y_n) &= E[(y_m - x_m)(y_n - x_n)] \\ &= E\{(y_m - x_m)[a_{1,n}(y_1 - x_1) + \cdots \\ &\quad + a_{n-1,n}(y_{n-1} - x_{n-1}) + t_n \sigma_n \sqrt{1 - R_n^2}]\} \\ &= a_{1,n} \sigma_{1,m} + a_{2,n} \sigma_{2,m} + \cdots + a_{n-1,m} \sigma_{n-1,m} \\ &= \sum_{i=1}^{n-1} a_{i,n} \sigma_{i,m} \text{ for all } n \text{ and for all } m < n.\end{aligned}$$

The term $t_n \sigma_n \sqrt{1 - R_n^2}$ drops out because $m \neq n$ and $E(t_n) = 0$. In the above sum $\sigma_{m,m}$, whenever it appears, means σ_m^2 for all m. Examining the last sum above for a particular n we note that we have $n - 1$ terms. Since the $a_{i,n}$ are the multiple regression coefficients of y_n on $y_1, \cdots, y_{n-1}$ they obey the so called "normal" equations,

$$\sum_{i=1}^{n-1} a_{i,n} \sigma_{i,m} = \sigma_{m,n} \qquad \text{where} \quad m = 1, \cdots, n - 1.$$

These equations are derived by means of the least squares theory as shown in Cramer page 303, in a slightly different notation. It follows that

$$\text{cov } (y_m, y_n) = \sigma_{m,n}$$

as was to be proved.

Thus the historical monthly means, variances, and covariances are preserved in the generated sequence. Consequently, correlations between months are preserved as well.

ACKNOWLEDGMENT

The author is grateful to Mr. R. Amir of Water Planning for

Israel, Ltd. for his useful comments and suggestions concerning this problem.

CITED REFERENCES

1. C. E. Sudler, "Storage Required for the Regulations of Stream Flow," *Trans. Am. Soc. Civil Engineering* **91**, 612, 1937.
2. F. B. Barnes, "Storage Required for a City Water Supply," *J. Inst. Engrs.* Australia **26,** 198, 1954.
3. H. A. Thomas, Jr. and M. B. Fiering, "Mathematical Synthesis of Streamflow Sequences for the Analysis of River Basins by Simulation," Chapter 12 in: A. Maass et al, *Design of Water Resource Systems*, Harvard Univ. Press, 1962.
4. H. Cramer, *Mathematical Methods of Statistics*, Princeton Univ. Press, 1957.

These papers introduce concepts involved in adapting the principal programming components within a single system.

After an examination of the over-all structure, the system's assembler, loader, and compilers are discussed. In this discussion (Parts I through V) attention is focused on the general design notions with minimal reference to the detail of mechanization and particular machines. Such reference, where necessary, is made to implementation of the system on the 7090.

Part VI compares implementation of the system on different machines and, to a certain extent, isolates the concepts that are independent of hardware.

Part VII is devoted to a general analysis of the system design.

Although some familiarity with the individual system components is assumed, an effort is made to address the systems engineer irrespective of his particular programming experience.

Design of an integrated programming and operating system

Part I: System considerations and the monitor

Part II: The assembly program and its language

Part III: The expanded function of the loader

Part IV: The system's FORTRAN compiler

Part V: The system's COBOL compiler

Part VI: Implementation on different machines

Part VII: Analysis of the system design

Parts III, IV and V are included here. Parts I and II appeared in June and the others are scheduled for publication in March.