Use of Correlation to Improve Estimates of the Mean and Variance

GEOLOGICAL SURVEY PROFESSIONAL PAPER 434-C

Use of Correlation to Improve Estimates of the Mean and Variance

By MYRON B. FIERING

STATISTICAL STUDIES IN HYDROLOGY

GEOLOGICAL SURVEY PROFESSIONAL PAPER 434-C

An examination of the criteria for extending streamflow records by correlation

UNITED STATES GOVERNMENT PRINTING OFFICE, WASHINGTON : 1963

UNITED STATES DEPARTMENT OF THE INTERIOR

STEWART L. UDALL, *Secretary*

GEOLOGICAL SURVEY

Thomas B. Nolan, *Director*

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington 25, D.C.

CONTENTS

ILLUSTRATIONS

TABLES

658684—63

STATISTICAL STUDIES IN HYDROLOGY

USE OF CORRELATION TO IMPROVE ESTIMATES OF THE MEAN AND VARIANCE

By Myron B. Fiering

ABSTRACT

Statistical techniques are utilized to examine the validity of using correlation analysis to augment streamflow data when concurrent and additional data are available on a nearby stream. The investigation demonstrates that this use of correlation can yield significant improvement in estimates of streamflow population parameters, and confirms previous indications that the indiscriminate use of poor correlations may produce poorer estimates of parameters than could be obtained from the original data alone. The mathematical analysis leads to complicated equations for judging the relative gain in information, these, equations being functions of the length of the original record, the length of extension, and the coefficients of correlation. To aid in the application of these equations to actual problems, tabulated solutions are available.

INTRODUCTION

Frequency curves of various streamflow characteristics are commonly used in planning water-resource developments. Parameters of the frequency curve for an individual site are usually computed from streamflow data obtained at that site. If the streamflow record is short the parameters may be poor estimates of the population parameters. Under certain conditions the correlation with data obtained at other sites may be used to improve the estimates of the two parameters, mean and variance, of the population. The use of a least-squares regression line is typical of the several methods available for applying this technique to hydrologic problems. This method assumes that there exists a linear relationship of the form

$$\eta=\gamma+\beta\xi \tag{1}$$

where η and ξ are concurrent flows on each of two streams, Y and X, respectively, and where γ and β are constants. If additional records of flow from stream X are available, it is possible to estimate the corresponding values of flow for stream Y by means of equation 1 and to use these values to aid in determining estimates of the mean and variance for stream Y.

This paper reviews the previous work that has been done on this problem and extends the analysis to the case of the relationship between streamflow at one stream site and the concurrent flows at sites on two other streams (the three-stream model).

The analysis indicates that the extension of a short record for the purpose of improving the estimate of a population parameter is not always desirable. Under certain conditions the parameter based on the extended record will be less reliable than that based on the unextended record. Criteria are derived which define the range of useful application of least-squares estimates of the type embodied in equation 1 and in the three-stream model. These criteria depend on the length of original record, the length of the proposed extension, and the coefficients of correlation.

The research reported in this paper was done at Harvard University under the sponsorship of the U.S. Geological Survey. The author acknowledges the assistance of Dr. Nicholas C. Matalas and Mr. Walter B. Langbein, of the Geological Survey, and Dr. Joan R. Rosenblatt, of the National Bureau of Standards, for advice, guidance, and encouragement. Mr. Robert Gemmell, of Harvard University, checked derivations and analyses and Mr. John Burton, of the University of New South Wales in Australia, reviewed the manuscript. Finally, the author expresses his sincere appreciation to Professor Harold A. Thomas, Jr., of Harvard University, under whose guidance the entire project, of which this paper summarizes only a part, was conducted.

NOTATION

The following list includes the majority of the symbols which appear in this paper. In addition to this summary, graphic representation of sample sizes—that is, record lengths—is provided by arrays where appropriate, and terms which appear rarely are defined in the text as necessary. Where no confusion can exist, subscripts are dropped in the text for convenience.

X_i Independent variate value

Y_i Dependent variate value (bivariate case) or independent variate value (trivariate case)

Z_i	Dependent variate value (trivariate case)
$(\ )$	Estimated value of $(\ )$
n_1	Length of record during which measurements are available for all variables
n_2	Length of record during which measurements are available for one variable (bivariate case), or for two variables (trivariate case)
n_3	Length of record during which measurements are available for one variable (trivariate case)
μ	Population mean of the dependent variable
$\mu_{n_1+n_2}$	Estimate of μ utilizing n_2 regression estimates and n_1 measured values
σ_i^2	Population variance of the parameter i
$(\bar{\ })_j$	Measured sample mean of $(\ )$ during period j
ρ_{xy}	Population correlation coefficient between the X_i and Y_i (bivariate case)
ρ_{ij}	Population correlation coefficient between the measured values of variate i and j (trivariate case)
R	Multiple correlation coefficient (trivariate case)
β	Population regression coefficient of Y on X (bivariate case)
b	Least-squares estimate of β
r_{ij}	Least-squares estimate of ρ_{ij}
$E(\ )$	Expected value of $(\ )$
I	Relative-information ratio
η, a	Population values of dependent variate and mean, respectively, in regression equations
Var$(\ )$	Variance of $(\ )$
Cov$(\ )$	Covariance of $(\ )$
MSE$(\ )$	Mean-square-error of $(\ )$

STATISTICAL PRINCIPLES

Because the theory of mathematical statistics enters prominently into this research, a brief survey of some of the more commonly used results is given here.

1. The mean of a sample of N items, each of which has a numerical value X_i, is given by

$$\text{mean}=\bar{X}=\frac{1}{N}(X_1+X_2+\ldots+X_N)=\frac{1}{N}\sum_1^N X_i \quad (2)$$

The mean can be described as a measure of central tendency, or average, of the data. The mean of a population is generally denoted by μ.

2. The variance of the sample in (1) above is a measure of the degree of dispersion of the individual items about their mean value, and is defined by

$$\text{variance}=\text{Var}(X)=\frac{1}{N}\sum_1^N (X_i-\bar{X})^2 \quad (3)$$

The variance of a population is generally denoted by σ^2. Data which possess a small variance are closely grouped about their mean, whereas data with a large variance are widely spread about their mean. Thus, if it is desired to estimate a population parameter—for example, the mean—from sample data and if several sets of data are available, each possessing a mean, these several means form another sample from which it is possible to compute the variance of the mean. The smaller the value of this variance, the more reliable is the estimated value of the mean, since a small variance implies a small dispersion and greater precision. Extensions of this logical process will be used throughout this paper.

Correlation techniques involving the relation between two variables, say X and Y, require the following defined terms.

1. The regression coefficient, b, equals

$$\frac{\sum_1^N X_iY_i-N\bar{X}\bar{Y}}{\sum_1^N X_i^2-N\bar{X}^2} \quad (4)$$

2. The correlation coefficient r equals

$$\frac{\sum_1^N X_iY_i-N\bar{X}\bar{Y}}{\left[\sum_1^N X_i^2-N\bar{X}^2\right]^{1/2}\left[\sum_1^N Y_i^2-N\bar{Y}^2\right]^{1/2}} \quad (5)$$

The theory of least squares defines b as the best estimator of β in equation (1) and r as a measure of the goodness of fit between the variables X and Y. A value of r equal to ± 1.0 indicates that the variables are perfectly correlated; a value of 0.0 indicates that there is no relation between the sample values of the two variables. Intermediate values indicate the closeness of fit between X and Y. The correlation coefficient of the joint population of X and Y is denoted by ρ, and R is an estimator of ρ.

To clarify the problem, consider a simplified example involving the mean annual flows, X_i and Y_i, of two nearby streams, X and Y respectively. The following assumptions are made:

1. A stable hydrologic regime is known to exist, so that significant correlation between the X_i and Y_i can reasonably be expected,
2. The mean annual flows, or some suitable transform thereof, are normally distributed,
3. Serial correlation of the mean annual flows is zero,
4. n_1 mean annual flows are available on stream Y,
5. n_1 mean annual flows measured concurrently with those on stream Y, are available on stream X; n_2 additional mean annual flows are also available on stream X, and
6. The X_i and Y_i are linearly related by a regression equation of the form

$$\eta=a+\beta(X-\mu_x) \quad (6)$$

and a fitted trend line is

$$Y=\bar{Y}_{n_1}+b(X-\bar{X}_{n_1}) \quad (7)$$

The data may be arrayed with regard to time in the

following form:

$$X_1, X_2, \ldots X_{n_1}, \ldots X_{n_1+n_2}$$
$$Y_1, Y_2, \ldots Y_{n_1} \qquad \text{(A–I)}$$

The major issue is whether or not the supplementary data from stream X can be used to provide a better estimate of the population parameters on stream Y, whose record is shorter. Since the data or their transforms are assumed normally distributed, the first two statistical moments serve to specify completely the distribution of the data. Thus we deal here only with these two moments, since small-sample correction terms involving the higher moments and cumulants are neglected.

Since the data of array (A–I) or their transforms are assumed to be normally distributed, it follows that the estimates of population moments for this model will be normally distributed about their respective population values. Thus the magnitude of the variance of each such estimate is a measure of the precision associated with it, this being a logical extension of the method of the simple example given above, wherein the variance of the mean was defined as a measure of the precision of the estimate of the mean. Well-known results from the theory of sampling give the variance of the first two statistical moments for a record of length n_1, taken to include only the measured data. If correlation estimates are used to provide n_2 additional values, the result is a combined record consisting of n_1 measured values and n_2 estimated values. If the variance of a parameter computed from this combined, or blended, record exceeds that computed from the record of size n_1 alone, it is clear that the combined record provides a less precise estimate of the parameter. However, if the variance is less than that computed from the original record alone, the corrrelation technique has provided a more precise estimate and should be utilized. For any particular parameter, with known values of n_1, n_2 and ρ (the population correlation coefficient), it is a simple matter to define the relative-information ratio, I, as the ratio of the variance of that parameter estimated from the original record to that estimated from the combined record. When I exceeds unity, it implies that the variance of the estimate of a moment made from the original record alone is larger than that of the estimate made from the combined record, and therefore the more precise estimate is computed from the combined data. On the other hand, if I is less than unity, the implication is that the original estimate is more precise than that computed from the blended data and that correlation analysis actually introduces additional variance, or loss of precision, into the estimate. Correlation should, therefore, not be used to augment the original data when $I < 1.0$.

HISTORICAL REVIEW—THE TWO-STREAM MODEL

Statistical literature contains several references to the use of regression estimates and, more generally, the use of correlation to improve estimates of parameters.

Wilkes (1952) assumes the following array of data to be a large sample from a bivariate normal population, and derives maximum likelihood estimators of the population means and variances:

$$X_1, X_2, \ldots X_{n_1}, \ldots X_{n_1+n_2}$$
$$Y_1, Y_2, \ldots Y_{n_1} \qquad Y_{n_1+n_2+1}, \ldots Y_{n_1+n_2+n_3} \qquad \text{(A-II)}$$

If n_2/n_1 and n_3/n_1 remain constant as n_1 approaches infinity,

$$\hat{\mu}_x = \frac{1}{\Delta\sigma_y}\left[\frac{(1+\epsilon)\overline{X}}{\sigma_x(1-\rho^2)} + \frac{\delta\overline{\overline{X}}}{\sigma_x}\left(\frac{1}{1-\rho^2}+\epsilon\right) + \frac{\epsilon\rho(\overline{\overline{Y}}-\overline{Y})}{\sigma_x(1-\rho^2)}\right] \qquad \text{(8a)}$$

and similarly

$$\hat{\mu}_y = \frac{1}{\Delta\sigma_x}\left[\frac{(1+\delta)\overline{Y}}{\sigma_y(1-\rho^2)} + \frac{\epsilon\overline{\overline{Y}}}{\sigma_y}\left(\frac{1}{1-\rho^2}+\delta\right) + \frac{\delta\rho(\overline{\overline{X}}-\overline{X})}{\sigma_y(1-\rho^2)}\right] \qquad \text{(8b)}$$

wherein

$$\Delta = \frac{1}{\sigma_x\sigma_y}\left[\frac{1}{(1-\rho^2)}\left[1+\delta+\epsilon+\delta\epsilon(1-\rho^2)\right]\right]$$
$$\delta = n_2/n_1$$
$$\epsilon = n_3/n_1$$

and

$\overline{X}, \overline{Y}$ are sample means during the period n_1,

$\overline{\overline{X}}, \overline{\overline{Y}}$ are sample means during the periods n_2 and n_3, respectively.

The variance of the estimator of the mean of X is given by

$$\sigma^2_{\hat{\mu}_x} = \frac{[n_1+n_3(1-\rho^2)]\sigma_x^2}{(n_1+n_2)(n_1+n_3)-\rho^2 n_2 n_3} \qquad (9)$$

and similarly for $\sigma^2_{\hat{\mu}_y}$. Additional results are given for the estimators of the population variances.

For independence between X and Y, that is $\rho=0$, equation 9 reduces to $\sigma_X^2/(n_1+n_2)$, which is equal to the variance of the mean for an original record of length n_1+n_2. In other words, Wilks' results indicate that the absence of correlation does not increase the variance of the estimated mean, and therefore that the use of correlation is always justifiable since it cannot decrease the precision of the estimate but can, at worst, effect no change. It will be shown that this conclusion is erroneous for small samples, in which the variance of the regression coefficient must be considered.

Matthai (1951), Edgett (1955), Nicholson (1957), and Lord (1956), have given solutions for different arrays of data, and their results are based on the assumption of the asymptotic behavior, as n_1 becomes large, of the ratio of n_i/n_1, where n_i is the length of

additional record on a nearby stream. Lord's results are typical of the rest. The data may be arrayed as follows:

$$\begin{array}{l} X_1, X_2, \ldots\ldots\ldots\ldots\ldots\ldots X_{n_1+n_2} \\ Y_1, Y_2, \ldots\ldots\ldots Y_{n_1} \\ \qquad\qquad\qquad\qquad Z_{n_1+1}, \ldots\ldots\ldots Z_{n_1+n_2} \end{array} \qquad \text{(A-III)}$$

Maximum likelihood estimators are given for the eight parameters μ_x, μ_y, μ_z, σ_x^2, σ_y^2, σ_z^2, ρ_{xy}, ρ_{xz}. Since there is no overlap, ρ_{yz} cannot be estimated from the sample. Of particular interest is a matrix of the sampling variances and covariances of the maximum likelihood estimators of the population means, from which it is deduced that the reciprocal of the relative-information ratio is:

$$\left.\begin{array}{l} 1-\dfrac{n_2}{n_1+n_2}\,\rho_{xz}^2 \text{ (when estimating } \mu_z) \\ 1-\dfrac{n_1}{n_1+n_2}\,\rho_{xy}^2 \text{ (when estimating } \mu_y) \end{array}\right\} \qquad (10)$$

It can further be shown that I^{-1} for the variance is:

$$\left.\begin{array}{l} 1-\dfrac{n_2}{n_1+n_2}\,\rho_{xz}^4 \text{ (when estimating } \sigma_z^2) \\ 1-\dfrac{n_1}{n_1+n_2}\,\rho_{xy}^4 \text{ (when estimating } \sigma_y^2) \end{array}\right\} \qquad (11)$$

Note that again these values of I^{-1} cannot exceed unity, and therefore I cannot be less than 1.0. The result of this lower boundary on I is to indicate that the correlation technique cannot decrease the precision of the estimates of the mean and variance, no matter how poor the correlation may be. This conclusion, drawn from maximum-likelihood (large-sample) theory, is in agreement with the result of Wilks above, and is incorrect for small samples. The appropriate values for computing the relative information ratio will be given below.

Langbein and Hardison (1955) propose a semi-graphical method for providing estimates of additional data. Their analysis indicates that when correlation is poor, the error in these estimates may increase the sampling error of the mean.

Defining

e=relative reduction in the variance in the mean$=(I-1)/I$,

they derive the following approximate formula for useful extension when estimating the mean:

$$\frac{n_1+n_2}{n_1}=\frac{r^2}{r^2-e} \qquad (12)$$

With known values of the record lengths (n_1 and n_2) it is possible to compute the required correlation as a function of the relative reduction in variance. Langbein (written communication) indicates that equation 12 applies only when $n_1 \gg 2$, say of the order of 10 or more, and may be derived from the more general form

$$e=1-\frac{n_1}{(n_1+n_2)}-\frac{(1-\rho^2)n_1n_2}{(n_1-2)(n_1+n_2)} \qquad (13)$$

From equation 13 it may be deduced that $e \geqq 0$ (or, equivalently, $I \geqq 1.0$) when

$$\rho^2 \geqq \frac{2}{n_1}$$

The Langbein-Hardison equations are approximate relations, but are significant in that they represent the first solution which suggests that correlation may not always be useful for improving estimates of the mean. One minor inconsistency arises from equation 12 when $e=0$, or when I is unity. For this case, $(n_1+n_2)/n_1$ is always equal to 1.0, which implies that n_2 is identically zero. This singularity is avoided by use of equation 13.

Professor H. A. Thomas, Jr. (written communication), arrived at a formulation which reduces to results found earlier and independently by Cochran (1953). For the array of data given in array A–1 above, Professor Thomas derived the relative-information ratio, I, for an estimate of the mean based on a combined sample of n_1 original measurements and n_2 regression estimates. For those combinations of n_1, n_2 and ρ_{xy} which yield a value of I less than unity, correlation introduces a retrogression of information and should not be utilized. The pertinent results are:

$$\mu_y=\overline{Y}_{n_1}+\frac{bn_2}{n_1+n_2}(\overline{X}_{n_2}-\overline{X}_{n_1}) \qquad (14a)$$

$$\sigma_{\mu_y}^2=\frac{\sigma_y^2}{n_1}\left[1-\frac{n_2}{n_1+n_2}\left(\rho^2-\frac{(1-\rho^2)}{n_1-3}\right)\right] \qquad (14b)$$

and

$$I=\left[1-\frac{n_2}{n_1+n_2}\left(\frac{\rho^2(n_1-2)-1}{n_1-3}\right)\right]^{-1} \qquad (14c)$$

Equations 14 are exact solutions, and do not depend upon any approximations assuming that the data conform to assumptions (1) through (6) given above. Solving equation 14c for the case of $I=1.0$, it is found that the critical, or cut-off value of ρ^2, is

$$\rho^2=\frac{1}{(n_1-2)}$$

For the range of values which n_1 generally assumes in hydrologic problems, this expression gives values of ρ in fair agreement with those obtained from the earlier Langbein and Hardison (1955) approximate solution for which $\rho^2=2/n_1$.

Correlation is also used for improving estimates of the variance of a population. Dr. J. R. Rosenblatt

(written communication, 1959) gives the following results based on the Thomas model.

$$E(\text{Var}\,(Y))=\sigma_y^2\left[1-\frac{n_2(n_1-4)(1-\rho^2)}{(n_1+n_2-1)(n_1-3)}\right] \quad (15a)$$

$$\text{MSE}\,(\text{Var}\,(Y))=\frac{2\sigma_y^4}{n_1-1}+\frac{n_2\sigma_y^4}{(n_1+n_2-1)^2}\left[2A+(n_2+2)B+(n_1+n_2-1)C-\frac{(n_1+1)(2n_1+n_2-2)}{(n_1-1)}\right] \quad (15b)$$

in which

$$A=(n_1-1)\rho^4+(n_1+4)\rho^2(1-\rho^2)+\frac{n_1+1}{n_1-3}(1-\rho^2)^2$$

$$B=\rho^4+\frac{6\rho^2(1-\rho^2)}{(n_1-3)}+\frac{3(1-\rho^2)}{(n_1-3)(n_1-5)}$$

$$C=\frac{2(n_1-4)(1-\rho^2)}{n_1-3}$$

and

$$I=\frac{2\sigma_y^4/(n_1-1)}{\text{MSE}\,(\text{Var}\,(Y))} \quad (15c)$$

The Rosenblatt equations comprise an exact solution, and are the first clear statement of the fact that correlation, which can decrease the precision of estimates of the variance, should be used only when I exceeds unity. Previous solutions—for example, equations 11—yield information ratios which exceed unity for all nonzero values of the correlation coefficient. For typical values of n_1 and n_2, I exceeds 1.0 when ρ is of the order of 0.8, according to the Rosenblatt solution. Thus, conditions suitable for improving estimates of the variance are much more restrictive than those for improving estimates of the mean.

THE THREE-STREAM MODEL

Previous work has considered only the correlation between flow characteristics at two sites (the two-stream model). The following development defines the conditions under which estimates of the mean and variance at one site can be improved by correlation with flow characteristics at two other sites. The three-stream model is used.

Improving estimates of the mean.—The following data, or some suitable transform of data, constitute an arrayed sample from a trivariate normal population in which serial correlation is zero.

$$X_1, X_2, \ldots X_{n_1}, \ldots X_{n_1+n_2}$$

$$Y_1, Y_2, \ldots Y_{n_1}, \ldots Y_{n_1+n_2}$$

$$Z_1, Z_2, \ldots Z_{n_1} \quad \text{(A-IV)}$$

It is required to examine the effect on the estimate of the population mean of Z of using the correlation between Z and Y and X. A linear trivariate regression model is predicated to represent the data:

$$\hat{Z}_i=\overline{Z}_{n_1}+b_x(X_i-\overline{X}_{n_1})+b_y(Y_i-\overline{Y}_{n_1}) \quad (16)$$

wherein b_x and b_y are the partial regression coefficients computed from the n_1 data on three streams by means of well-known formulae from the theory of least squares. The theory gives:

$$b_x=[\text{Var}\,(Y)\cdot\text{Cov}\,(X,Z)-\text{Cov}\,(X,Y)\cdot\text{Cov}\,(Y,Z)]\div\Delta$$

$$b_y=[\text{Var}\,(X)\cdot\text{Cov}\,(Y,Z)-\text{Cov}\,(X,Y)\cdot\text{Cov}\,(X,Z)]\div\Delta$$

and

$$\Delta=\text{Var}\,(X)\cdot\text{Var}\,(Y)-[\text{Cov}\,(X,Y)]^2.$$

An unbiased estimate of the combined mean, $\mu_{n_1+n_2}$, is given by

$$\hat{\mu}_{n_1+n_2}=\overline{Z}_{n_1}+\frac{n_2}{n_1+n_2}[b_x(\overline{X}_{n_2}-\overline{X}_{n_1})+b_y(\overline{Y}_{n_2}-\overline{Y}_{n_1})] \quad (17)$$

and it is necessary to calculate the variance of the mean, σ_μ^2. The result may be written, after considerable simplification, as

$$\text{Var}\,(\hat{\mu}_{n_1+n_2})=\frac{\sigma_z^2}{n_1}\left[1+\frac{n_2}{N}\frac{(R^2(2-n_1)+2)}{(n_1-4)}\right] \quad (18)$$

where $N=n_1+n_2$, the total length of record, and R is the total correlation coefficient of Z on X and Y, defined by

$$R^2=\frac{\rho_{zx}^2+\rho_{zy}^2-2\rho_{xy}\rho_{xz}\rho_{yz}}{1-\rho_{xy}^2}.$$

From equation 18 it may be deduced that I, the relative information with respect to the mean, is given by

$$I=\frac{\sigma_z^2}{n_1}\div\text{Var}\,(\hat{\mu})$$

or

$$I=\left[1-\frac{n_2}{N}\left(\frac{R^2(n_1-2)-2}{n_1-4}\right)\right]^{-1} \quad (19)$$

This result may be generalized to a model represented by a multiple regression on p-independent variables, each of equal length $N=n_1+n_2$. For this case, the variance of the mean of the combined sample may be expressed as

$$\text{Var}\,(\hat{\mu}_{n_1+n_2})=\text{Var}\,(\overline{Z}_{n_1})\left[1+\frac{n_2(pR^2-n_1R^2+p)}{N(n_1-p-2)}\right] \quad (20)$$

The families of curves in figures 1 and 2 represent the relative information of the mean, I, for typical values of n_2 and R^2, with n_1 set equal to 6 and 10, respectively.

Equation 19 can be solved for R^2 and evaluated when I is set equal to unity. This gives

$$R^2=\frac{2}{n_1-2}$$

as the critical values of R^2, at which it is equally advantageous to correlate or not for estimating the mean.

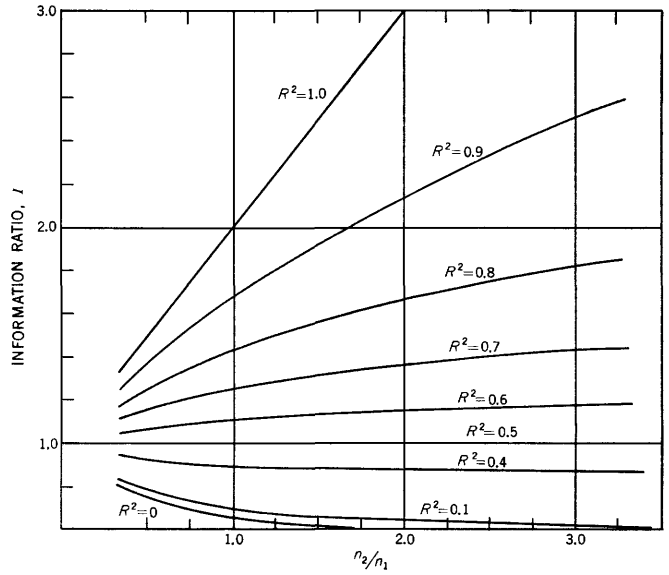

FIGURE 1.—Relative information with respect to the mean obtained by using the three-stream correlation model, $n_1=6$.

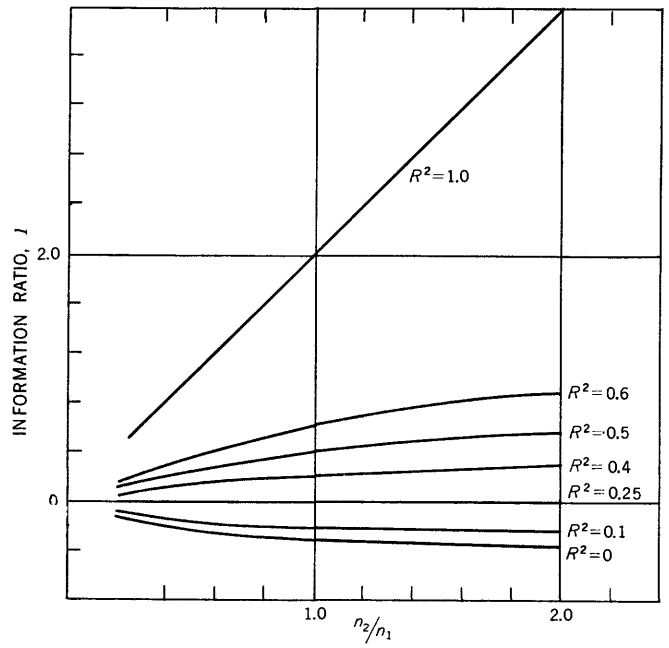

FIGURE 2.—Relative information with respect to the mean obtained by using the three-stream correlation model, $n_1=10$.

This is an exact solution, and the similarity to the solutions for the two-stream case is apparent.

The above analysis may be extended to the case in which each of the three streams has a record of different length. The following array of data obtains from a trivariate normal population, in which serial correlation is assumed to be zero:

$$X_1, \ldots X_{n_1}, \ldots X_{n_1+n_2}, \ldots X_{n_1+n_2+n_3}$$

$$Y_1, \ldots Y_{n_1}, \ldots Y_{n_1+n_2}$$

$$Z_1, \ldots Z_{n_1} \qquad \text{(A-V)}$$

In hydrologic practice, the effect of the use of correlation would probably be investigated in two stages, with the equations for the two- and three-stream models being applied separately to the applicable lengths of record. Analytical solutions are available for testing the use of a combined regression model in which Z is estimated by means of a trivariate equation for the range n_1+1 to n_1+n_2, and estimated by means of a bivariate equation for the range n_1+n_2+1 to $n_1+n_2+n_3$, but owing to the formidable statistical problems encountered they are approximate solutions only. However, for the sake of completeness, two solutions are presented here. It must be emphasized that these solutions are not exact, and that the first of the two leads to overestimation of the relative information by a factor of $(1/3)\cdot(\rho_{xx}^2)$. The solutions are:

$$I=\left\{1+\frac{n_2}{n_1+n_2}\left(\frac{2R^2-n_1R^2+2}{n_1-4}\right)+\frac{n_3}{N}\left[\frac{n_1+n_3}{N}\left(\frac{(n_1-4)\rho_{xx}^2+1}{n_1-3}\right)-2\rho_{xx}^2 n_1\sqrt{(n_1+n_2)}\right]\right\}^{-1} \qquad (21)$$

$$I=N\left\{n_1\left[1+\frac{n_2}{n_1-2}(1-R^2)\right]+(n_2+n_3)(1-\rho_{xx}^2)\right\}^{-1} \qquad (22)$$

Equation 21, derived by the author, is based on the single simplifying assumption that the partial regression coefficients, b_x and b_y, are each uncorrelated with the bivariate regression coefficient, b, which defines the dependence of Z on X (as in Equation 7). Equation 22, suggested by Langbein (written correspondence), has somewhat less theoretical justification but provides exact solutions for the case of $R^2=\rho_{xx}^2=1.0$, whereas equation 21 does not. Neither formulation is recommended.

To facilitate solution for the relative information, the equations representing I for the two exact solutions for the mean are evaluated by means of programs written by the author for the UNIVAC I computer at the Harvard University Computation Laboratory. Equation 14c for the Thomas bivariate model, and equation 19 for the trivariate model are tabulated for many combinations of the several values of n_1 and correlation coefficients. To use the tables one enters with known values of the n_i and the correlation coefficient, and reads the corresponding value of I directly. Complete sets of tables, on deposit at the Widener Library and Gordon McKay Library, at Harvard University and the U.S. Geological Survey, require some 600 pages and are not reproduced here. The range and mesh, or spacing between successive entries, of each argument in the complete tables is as follows:

TABLE 1.—*Range and mesh of arguments for tables of the relative information ratio of the mean*

Model	Argument	Range	Mesh
2-stream	n_1	4–30	2
2-stream	n_2	0–20	2
2-stream	ρ_{xy}	0. 05–1. 00	0. 05
3-stream	n_1	6–20	2
3-stream	n_2	2–20	2
3-stream	R^2	0. 0–1. 0	0. 05

Condensed results for the two-stream model are given in table 2, from which the general behavior of I may be deduced. In this table, I is given as a function of n_1, n_2, and ρ_{xy}. A condensed version of the complete table for the three-stream model with $n_3=0$ is given in table 3.

Estimating the Variance.—A solution is presented for the three-stream model in which the two independent records have equal lengths. As before, the data or their transforms are assumed normally and independently distributed, and may be arrayed as in array (A-IV) above:

$$\begin{array}{l} X_1,\ldots X_{n_1},\ldots X_{n_1+n_2} \\ Y_1,\ldots Y_{n_1},\ldots Y_{n_1+n_2} \\ Z_1,\ldots Z_{n_1} \end{array} \qquad \text{(A-VI)}$$

It is desired to estimate the variance of stream Z, σ_z^2 using correlation techniques. As above, a linear trivariate regression model is postulated, and may be expressed as

$$\hat{Z}_i=\overline{Z}_{n_1}+b_x(X_i-\overline{X}_{n_1})+b_y(Y_i-\overline{Y}_{n_1}) \qquad (23)$$

TABLE 2.—*Relative information of the mean, two-stream model*

n_2	ρ	n_1							
		4	6	8	10	12	14	16	18
2	0.2	1.000	0.934	0.970	0.984	0.990	0.994	0.996	0.997
	.4	1.000	.970	.998	1.006	1.009	1.010	1.010	1.010
	.6	1.000	1.038	1.048	1.046	1.043	1.039	1.035	1.032
	.8	1.000	1.149	1.128	1.108	1.093	1.082	1.073	1.065
	1.0	1.000	1.333	1.250	1.200	1.166	1.142	1.125	1.111
6	.2	.644	.877	.938	.964	.978	.986	.990	.994
	.4	.710	.943	.996	1.015	1.022	1.025	1.026	1.026
	.6	.856	1.079	1.110	1.112	1.106	1.099	1.092	1.086
	.8	1.201	1.351	1.321	1.283	1.250	1.222	1.200	1.182
	1.0	2.500	2.000	1.750	1.600	1.500	1.428	1.375	1.333
10	.2	.603	.851	.922	.953	.970	.980	.987	.991
	.4	.673	.930	.995	1.020	1.031	1.036	1.038	1.038
	.6	.833	1.100	1.148	1.155	1.151	1.143	1.135	1.127
	.8	1.250	1.481	1.461	1.417	1.375	1.338	1.308	1.282
	1.0	3.500	2.666	2.250	2.000	1.833	1.714	1.625	1.555
14	.2	.582	.836	.911	.946	.965	.976	.984	.989
	.4	.654	.922	.994	1.023	1.037	1.043	1.046	1.047
	.6	.821	1.114	1.173	1.185	1.184	1.177	1.169	1.161
	.8	1.278	1.572	1.566	1.522	1.477	1.436	1.400	1.368
	1.0	4.500	3.333	2.750	2.400	2.166	2.000	1.875	1.777
18	.2	.570	.826	.904	.941	.961	.974	.982	.988
	.4	.642	.917	.994	1.026	1.041	1.049	1.053	1.054
	.6	.813	1.123	1.191	1.208	1.209	1.204	1.196	1.188
	.8	1.297	1.639	1.648	1.608	1.562	1.518	1.479	1.445
	1.0	5.500	4.000	3.250	2.800	2.500	2.285	2.125	2.000

If the values of Z_{n_1+1} to $Z_{n_1+n_2}$ are computed using this model, and the variance of Z estimated without regard to the fact that these values are not measured

TABLE 3.—*Relative information of the mean, three-stream model, $n_3=0$*

n_2	R^2	n_1							
		6	8	10	12	14	16	18	20
2	0.0	0.800	0.909	0.947	0.966	0.976	0.982	0.986	0.989
	.2	.870	.962	.989	1.000	1.005	1.007	1.009	1.009
	.4	.952	1.020	1.034	1.037	1.036	1.034	1.032	1.030
	.6	1.053	1.087	1.084	1.077	1.070	1.063	1.057	1.053
	.8	1.176	1.163	1.139	1.120	1.105	1.093	1.084	1.076
	1.0	1.333	1.250	1.200	1.667	1.143	1.125	1.111	1.110
6	.0	.667	.824	.889	.923	.943	.957	.966	.972
	.2	.769	.921	.976	1.000	1.012	1.019	1.022	1.024
	.4	.909	1.045	1.081	1.091	1.092	1.089	1.085	1.081
	.6	1.111	1.207	1.212	1.200	1.185	1.170	1.157	1.145
	.8	1.429	1.429	1.379	1.333	1.295	1.264	1.239	1.218
	1.0	2.000	1.750	1.600	1.500	1.429	1.375	1.333	1.300
10	.0	.615	.783	.857	.898	.923	.940	.951	.960
	.2	.727	.900	.968	1.000	1.017	1.026	1.032	1.034
	.4	.889	1.059	1.111	1.128	1.132	1.130	1.126	1.121
	.6	1.143	1.286	1.304	1.294	1.277	1.258	1.241	1.224
	.8	1.600	1.636	1.579	1.517	1.463	1.418	1.380	1.348
	1.0	2.667	2.250	2.000	1.833	1.714	1.625	1.556	1.500
14	.0	.588	.759	.837	.881	.909	.928	.941	.951
	.2	.704	.887	.963	1.000	1.020	1.032	1.039	1.043
	.4	.877	1.068	1.132	1.156	1.163	1.163	1.159	1.154
	.6	1.163	1.341	1.374	1.368	1.351	1.331	1.311	1.293
	.8	1.724	1.803	1.748	1.677	1.613	1.557	1.509	1.469
	1.0	3.333	2.750	2.400	2.167	2.000	1.875	1.778	1.700
18	.0	.571	.743	.824	.870	.899	.919	.933	.944
	.2	.690	.878	.959	1.000	1.023	1.037	1.045	1.050
	.4	.870	1.074	1.148	1.176	1.187	1.189	1.186	1.182
	.6	1.176	1.383	1.429	1.429	1.413	1.393	1.373	1.352
	.8	1.818	1.940	1.892	1.818	1.747	1.683	1.628	1.580
	1.0	4.000	3.250	2.800	2.500	2.286	2.125	2.000	1.900

but estimated variates, the following expressions are obtained:

$$(N-1)S^2=\sum_{i=1}^{n_1}(Z_i-\hat{\mu}_{n_1+n_2})^2+\sum_{i=1}^{n_2}(Z_{n_1+i}-\hat{\mu}_{n_1+n_2})^2 \qquad (24)$$

$$E(S^2)=\sigma_z^2\left[1-\frac{n_2(n_1-6)(1-R^2)}{(N-1)(n_1-4)}\right] \qquad (25)$$

whereupon

$$\text{MSE}\ (S^2)=\frac{\sigma_z^4}{(N-1)^2}\left\{2(N-1)+4n_2(1-R^2)+n_2(n_2-6)(1-R^2)^2 + \frac{4n_2(n_2+3)}{n_1-4}(1-R^2)-\frac{4n_2(2n_2+1)}{n_1-4}(1-R^2)^2 + \frac{8n_2(n_2+2)}{(n_1-4)(n_1-6)}(1-R^2)^2\right\} \qquad (26)$$

and the relative information may be written

$$I=\frac{2\sigma_z^4}{n_1-1}\div\text{MSE}\ (S^2) \qquad (27)$$

As in the case of the mean, the relative information of the variance is tabulated by means of UNIVAC I evaluation of equation 15b for the Rosenblatt two-stream solution and of equation 27 for the three-stream solution. Table 4 gives a concise summary of the relative information of the variance for the two-stream model. Table 5 is a sample of the complete tables which are available in the Harvard University libraries. Entering the tables with values of n_1, n_2, and correlation coefficients, the relative information is read directly.

Table 6 is used in the same manner, except that n_1, n_2, and R^2 are required to read the relative information in the 3-stream case.

The range and mesh of each argument is tabulated in table 7.

TABLE 4.—*Relative information of the variance, two-stream model*

n_2	ρ	n_1							
		6	8	10	12	14	16	18	20
2	0.2	0.887	1.155	1.163	1.146	1.129	1.115	1.103	1.093
	.4	.884	1.098	1.112	1.104	1.094	1.084	1.076	1.069
	.5	.908	1.053	1.068	1.066	1.061	1.056	1.051	1.047
	.8	1.013	1.075	1.076	1.070	1.063	1.057	1.052	1.048
	1.0	1.400	1.285	1.222	1.190	1.154	1.133	1.118	1.105
6	.2	.628	.955	.969	.954	.940	.931	.925	.921
	.4	.659	.940	.963	.957	.950	.944	.940	.938
	.6	.742	.961	.989	.990	.988	.985	.983	.981
	.8	.977	1.110	1.121	1.114	1.104	1.095	1.086	1.079
	1.0	2.200	1.858	1.668	1.545	1.461	1.400	1.354	1.316
10	.2	.520	.782	.769	.741	.720	.809	.700	.697
	.4	.561	.801	.800	.782	.767	.845	.753	.750
	.6	.661	.870	.885	.877	.869	.923	.859	.857
	.8	.946	1.105	1.118	1.108	1.096	1.093	1.074	1.065
	1.0	3.000	2.428	2.111	1.909	1.769	1.667	1.589	1.526
14	.2	.466	.680	.648	.610	.583	.628	.554	.547
	.4	.509	.714	.695	.666	.643	.686	.618	.612
	.6	.616	.808	.808	.790	.774	.810	.755	.750
	.8	.925	1.095	1.104	1.091	1.074	1.073	1.047	1.036
	1.0	3.800	3.000	2.556	2.273	2.079	1.934	1.825	1.737
18	.2	.434	.617	.573	.528	.497	.476	.462	.453
	.4	.478	.658	.627	.589	.561	.542	.528	.519
	.6	.587	.765	.753	.727	.704	.688	.675	.667
	.8	.911	1.084	1.089	1.071	1.050	1.032	1.016	1.003
	1.0	4.600	3.570	3.000	2.638	2.386	2.200	2.060	1.948

TABLE 5.—*Sample of table of relative information of the variance, two-stream model*

n^1	n^2	ρ	I	n^1	n^2	ρ	I
16	14	0.70	0.880	16	18	.40	.542
16	14	.75	.939	16	18	.45	.568
16	14	.80	1.059	16	18	.50	.601
16	14	.85	1.188	16	18	.55	.640
16	14	.90	1.358	16	18	.60	.688
16	14	.95	1.593	16	18	.65	.747
16	16	.10	.502	16	18	.70	.820
16	16	.15	.508	16	18	.75	.912
16	16	.20	.516	16	18	.80	1.032
16	16	.25	.527	16	18	.85	1.190
16	16	.30	.541	16	18	.90	1.407
16	16	.35	.559	16	18	.95	1.718
16	16	.40	.581	16	20	.10	.430
16	16	.45	.607	16	20	.15	.436
16	16	.50	.638	16	20	.20	.444
16	16	.55	.676	16	20	.25	.455
16	16	.60	.722	16	20	.30	.470
16	16	.65	.779	16	20	.35	.487
16	16	.70	.848	16	20	.40	.509
16	16	.75	.935	16	20	.45	.536
16	16	.80	1.045	16	20	.50	.569
16	16	.85	1.190	16	20	.55	.609
16	16	.90	1.384	16	20	.60	.658
16	16	.95	1.657	16	20	.65	.718
16	18	.10	.462	16	20	.70	.795
16	18	.15	.468	16	20	.75	.892
16	18	.20	.476	16	20	.80	1.018
16	18	.25	.488	16	20	.85	1.189
16	18	.30	.502	16	20	.90	1.426
16	18	.35	.520	16	20	.95	1.776

The most significant result of the derivation and tabulation of I for the variance is the fact that the maximum information gain in the three-stream model for estimating the variance rarely exceeds, and generally is less than, the gains associated with the two-stream model. Owing to the formidable analytic and computational difficulties which would be encountered, the case of three different lengths is not solved.

Numerical Checking of Results.—A check on the derivations is provided by examining a multiple regression model predicated on a table of normal random sample deviates with mean=5.0 and variance=1.0. A 540-item sequence of these deviates was divided into 18 sets of 30 items. Ten of the 30 items in each set were discarded in favor of estimated values, computed by means of a multiple regression—that is, 3-stream—model using assumed population values of the regression coefficients. The variance of the mean and the variance of the variance are computed from a sample of 18 sets, using the combined 20–10 item record, where each such record consists of 20 original and 10 estimated deviates. These measured values are then compared to theoretical values of the variance of both the mean and the variance as computed from assumed population parameters. A trial was made with $n_1=20$, $n_2=10$, $\rho_{zx}=0.9$, $\rho_{zy}=0.9$, and $\rho_{xy}=0.8$. Theoretical values taken from the tables are $I=1.42$ for the mean and 1.34 for the variance, which compare favorably with measured values of 1.33 and 1.14.

All the derivations are made using ρ instead of its estimator, r. To evaluate the effect of bias in estimating ρ, a simulation technique was devised. A 540-item sequence of normal random sample numbers of mean=5.0 and variance=1.0 is used to generate two additional records of equal length by means of a linear regression model with a random additive component using assumed population values of the ρ's. This random component is numerically equal to the product of the standard error of estimate and a new normal random sampling

TABLE 6.—*Relative information of the variance, three variables, for R^2=0.2, 0.4, 0.6, 0.8, 1.0*

n_2	R^2	n_1		
		8	12	16
2	0.2	0.891	1.045	1.046
	.4	.921	1.045	1.044
	.6	.976	1.051	1.044
	.8	1.090	1.092	1.077
	1.0	1.279	1.180	1.138
6	.2	.698	.954	.970
	.4	.778	.985	.998
	.6	.918	1.058	1.068
	.8	1.192	1.220	1.193
	1.0	1.840	1.545	1.410
10	.2	.598	.809	.811
	.4	.700	.905	.904
	.6	.879	1.045	1.042
	.8	1.250	1.309	1.261
	1.0	2.430	1.895	1.660

TABLE 7.—*Range and mesh of arguments for tables of the relative information ratio of the variance*

Model	Argument	Range	Mesh
2-stream	n_1	6–20	2
2-stream	n_2	2–20	2
2-stream	ρ	0. 10–0. 9	0. 05
3-stream	n_1	8–16	4
3-stream	n_2	2–10	4
3-stream	R^2	0. 2–1. 0	0. 2

deviate. These three concurrent sequences are then divided into 18 sets of 30 items, from each of which 10 values of the dependent variate are discarded in favor of estimates computed using the least-squares partial regression coefficients measured from the 20 values remaining in each set. It does not follow that the partial regression coefficients computed in this manner will produce the same value of dependent variable given in the original sequence, owing to the nature of the random-number table. As before, the variance of the mean and of the variance are computed from the sample of 18 sets by consideration of the population values which obtain from the original sample of 20 versus its combined 20-10-item counterpart. The results of several runs are summarized in table 8.

TABLE 8.—*Comparison of theoretical and simulated values of the information ratio*

ρ_{xx} and ρ_{xy}	I (mean)		I (var)	
	Theory	Simulated	Theory	Simulated
0. 9	1. 42	1. 42	1. 22	1. 21
0. 8	1. 34	1. 47	1. 09	1. 17
0. 7	1. 26	1. 46	. 97	1. 09
0. 6	1. 20	1. 45	. 89	1. 07

These results are not conclusive. For example, although agreement seems to be quite close as to order of magnitude, the simulated data for I (mean) exhibits a trend opposite to that which is predicted. However, it must be noted that all the entries in table 8 are based on the same sequence of random sampling numbers and any bias in the mean or variance of the random deviates would tend to distort all the results in the same fashion. It should also be noted that the theoretical solution for I (var) is approximate so that perfect agreement is not expected. Based on these considerations, these test results are accepted as an aid in verifying the analyses. An analytical approach to the sampling variance of I due to the variance of r is hopelessly complicated.

CONCLUSIONS

1. Much of the statistical literature, of which the work of Lord (1956) is typical, does not admit the possibility of loss of information on statistical parameters by correlation, since, as can be seen in equations 10 and 11, efficiencies are always less than unity. This implies that at worst the application of correlation techniques to augment data will not lead to a retrogression of information about parameters, and always implies a gain of information when ρ is not zero. As indicated approximately by Langbein and Hardison (1955), and verified subsequently by the exact analyses of Professor Thomas, Cochran (1953), and Dr. Rosenblatt, the inclusion of the variance of the regression coefficient introduces the possibility of dilution of good information by estimates based on poorly correlated data. This can lead to a loss of hydrologic information.

2. For the case of a three-stream model, in which either the data or a suitable transform thereof are normally distributed, expressions for the relative information of the mean and of the variance are derived. Both solutions are exact.

3. The several cases which have been considered, and the equations to be used for analysis in each case, are summarized below in table 9.

TABLE 9.—*Summary of equations to be used for the several cases considered*

Case	Equation No.
2-stream, mean, evaluation of I	14c
2-stream, variance, evaluation of I	15b, 15c
3-stream, mean, evaluation of I	19
3-stream, variance, evaluation of I	26, 27

4. Owing to the great complexity of the several functions described above, a table of the relative information versus appropriate arguments has been prepared in each case by a program written for the UNIVAC computer. Entering the tables with known values of the n_1 and ρ or R^2, the corresponding I is read directly. When I exceeds unity, the variance of the moment under consideration is reduced. Conversely, it is increased by the correlation technique when the tabulated value of I lies below 1.0. In the latter case, the technique should not be used to augment records.

5. Results given in this paper apply to hydrologic data which may be, or are transformed to be, reasonably normally distributed and without serial correlation.

REFERENCES

Cochran, W. G., 1953, Sampling techniques: New York, John Wiley & Sons, Inc.

Edgett, G. L., 1955, Multiple regression with missing observations among the independent variables: Statist. Assoc. Jour., v. 50, p. 122–131.

Langbein, W. B., and Hardison, C. H., 1955, Extending streamflow data; Am. Soc. Civil Engineers Proc., v. 81, Paper No. 826, 13 p.

Lord, F. M., 1956, Estimation of parameters from incomplete data, Statist. Assoc: Jour., v. 51, p. 870–876.

Matthai, A., 1951, Estimation of parameters from incomplete data with application to design of sample surveys: Sankhya, v. 11, p. 145–152.

Nicholson, G. E., Jr., 1957, Estimation of parameters from incomplete multivariate samples: Statist. Assoc. Jour., v. 52, p. 523–526.

Wilks, S. S., 1952, Moments and distributions of estimates of population parameters from fragmentary samples: Annals of Math. Statistics, v. 23, p. 163–190.