NATIONAL BUREAU OF STANDARDS REPORT

NBS PROJECT		NBS REPORT
1103-40-11625	14 July 1959	6514

SOME PROPERTIES OF THE "ARRAY CORRELATION" PROCEDURE

by

Joan Raup Rosenblatt

Statistical Engineering Laboratory

Technical Report No. 2
to
Water Resources Division
U. S. Geological Survey
Department of Interior

NBS

U. S. DEPARTMENT OF COMMERCE

NATIONAL BUREAU OF STANDARDS

~~Project 11432~~ (USGS)
Technical Note No. 2
14 July 1959

Some Properties of the "Array Correlation" Procedure

Joan Raup Rosenblatt
National Bureau of Standards

1. Introduction

The expression "array correlation" refers to a proposed procedure for the construction of "synthetic records" of the type discussed in Technical Note No. 1. The essential feature of the array correlation procedure is that "synthetic" values of discharge are obtained by a transformation relating the marginal distribution of discharge for one stream to that for the other stream. The procedure described in Technical Note No. 1 uses a regression relation which explicitly takes advantage of the correlation in the bivariate distribution.

2. Assumptions and Description of Procedure

The assumptions concerning distribution of discharge are stated on page 2 of Technical Note No. 1. The notation of that note will be used.

In practice, the "array correlation" procedure is understood to be as follows.

From the paired observations

$$(X_1, Y_1), \ldots , (X_{n_1}, Y_{n_1})$$

are obtained the paired order statistics

$$(X_{(1)}, Y_{(1)}), \ldots , (X_{(n_1)}, Y_{(n_1)}) ,$$

where

$$X_{(1)} \leq \cdots \leq X_{(n_1)} ,$$

$$Y_{(1)} \leq \cdots \leq Y_{(n_1)} .$$

These points are plotted and are found to determine a straight line which is used for calculating values of Y from values of X.

For purposes of analysis, it is assumed that the slope of this line is

$$S_{y1} / S_{x1} ,$$

where

$$S^2_{y1} = \sum_{i=1}^{n_1} (Y_i - \overline{Y}_1)^2 ,$$

$$S^2_{x1} = \sum_{i=1}^{n_1} (X_i - \overline{X}_1)^2 .$$

Thus, the mean μ_y may be estimated by the average of observed and calculated values,

$$\text{(2.1)} \qquad V = \overline{Y}_1 + \frac{n_2}{n_1+n_2} \frac{S_{y1}}{S_{x1}} (\overline{X}_2 - \overline{X}_1) .$$

This is an unbiased estimator. The remaining sections of this note give the variance of V, an estimator for σ^2_y , and comparisons with the estimators treated in Technical Note No. 1.

3. Variance of V

For the ratio S_{y1} / S_{x1} , we have

$$\text{(3.1)} \qquad E \frac{S^2_{y1}}{S^2_{x1}} = \frac{\sigma^2_y}{\sigma^2_x} \left[1 + \frac{2(1-\rho^2)}{n_1-3} \right] ,$$

$$E \frac{S_{y1}}{S_{x1}} = \frac{\sigma_y}{\sigma_x} \left[1 + \frac{1-\rho^2}{2\, n_1} - \frac{3(1-\rho^2)}{2\, n_1^{\,2}} \right] + O(1/n_1^{\,3}). \tag{3.2}$$

Equation (3.2) is the only one in which an approximation is required. Accordingly, we have

$$\begin{aligned} \text{Var }(V) = {} & \frac{\sigma_y^2}{n_1} \left[1 + \frac{n_2}{N} (1-2\rho) \right] \\ & + \frac{\sigma_y^2}{n_1^2} \frac{n_2}{N} (2-\rho)(1-\rho^2) \\ & + O(1/n_1^{\,3}) \ . \end{aligned} \tag{3.3}$$

Similarly expressed, the regression estimator U has the variance

$$\begin{aligned} \text{Var }(U) = {} & \frac{\sigma_y^2}{n_1} \left[1 - \frac{n_2}{N} \rho^2 \right] \\ & + \frac{\sigma_y^2}{n_1^2} \frac{n_2}{N} (1-\rho^2) + O(1/n_1^{\,3}) \ . \end{aligned}$$

Evidently,

$$\text{Var }(V) \geq \text{Var }(U) \ . \tag{3.4}$$

Moreover, if $\rho \leq 1/2$,

$$\mathrm{Var}\ (V) \geq \sigma_y^2 / n_1 \quad . \tag{3.5}$$

4. Estimation of σ_y^2

Consider the function

$$S^2 = \sum_{i=1}^{n_1} (Y_i - V)^2 + \sum_{j=1}^{n_2} (Y^*_{n_1+j} - V)^2 , \tag{4.1}$$

where for $j = 1, 2, \ldots , n_2$,

$$Y^*_{n_1+j} = \overline{Y}_1 + \frac{S_{y1}}{S_{x1}} (X_{n_1+j} - \overline{X}_1) . \tag{4.2}$$

This is analogous to the function S_3^2 , equation (3.3) of Tech. Note No. 1. The "natural" estimator for σ_y^2 would be

$$T = S^2/(N-1) . \tag{4.3}$$

The estimator T tends to over-estimate σ_y^2 ; its expected value is

$$ET = \sigma_y^2 + \frac{2}{n_1-3} \frac{n_2}{N-1} (1-\rho^2) \sigma_y^2 . \tag{4.4}$$

The mean-squared-error of T is

$$\text{(4.5)} \qquad MSE(T) = Var\,(T_1) + \frac{n_2}{(N-1)^2}\,\sigma_y^4 \Bigg[2\,D + (n_2 + 2)F - 4\,\frac{N-1}{n_1-3}\,(1-\rho^2) - \frac{n_1+1}{n_1-1}\,(2n_1 + n_2 - 2) \Bigg],$$

where

$$\text{(4.6)} \qquad (n_1 - 3)\,D = (n_1^2 - 1) - 4(n_1 + 1)\,\rho^2 + 8\,\rho^4,$$

$$\text{(4.7)} \qquad (n_1 - 3)(n_1 - 5)\,F = (n_1^2 - 1) - 8(n_1 + 1)\,\rho^2 + 24\,\rho^4,$$

and (cf. Tech. Note No. 1),

$$(n_1 - 1)\,T_1 = S_{y1}^2 .$$

Now, let the "information ratio" be

$$\text{(4.8)} \qquad I^*(\rho, n_1, n_2) = Var\,(T_1)/MSE(T).$$

The following properties hold.

(4.9) $$I^{*}(1, n_1, n_2) = 1 + n_2/(n_1 - 1),$$

exactly as in (4.2) of Tech. Note No. 1.

(4.10) $$I^{*}(0, n_1, n_2) < 1 \quad \text{for all} \quad n_1, n_2 .$$

(4.11) $$I^{*}(0, n_1, n_2) \sim N/(N + n_2) \text{ as } n_1 \longrightarrow \infty ,$$

with the ratio n_1/n_2 held fixed.

If $n_2/n_1 = \alpha > 0$, then

(4.12) $$\rho_o^{*} \longrightarrow \frac{8 + \alpha(1 + \alpha)}{8(1+\alpha)(2+3\alpha)} \quad \text{as} \quad n_1 \longrightarrow \infty ,$$

where ρ_o^{*} is the solution of $I^{*}(\rho, n_1, n_2) = 1$. For example, with $\alpha = 1$ $(n_1 = n_2)$, $\rho_o \longrightarrow 1/8$.
For $n_1 = n_2 = 15$ or 20, one finds $\rho_o = 0.8$ (from the exact equation).