Brief Reports

Small Sample Estimation of ρ_1

JAMES R. WALLIS AND P. ENDA O'CONNELL*

IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598

Abstract. Hydrologists frequently need to estimate lag 1 correlation ρ_1 from rather short observed records of length n. The most commonly used algorithms for estimating ρ_1 can give seriously biased estimates if n is small and $|\rho_1|$ is large. Small sample bias corrections based on the assumption of a lag 1 Markov generating process are available for $\hat{\rho}_1$, but the corrected estimates tend to have larger variance than the uncorrected estimates do.

The optimal design of water resource systems would be comparatively easy if we had complete, accurate knowledge of all future inputs into the system. Lacking prescience hydrologists have in recent years employed synthetic sequences, which according to probabilistic theory are believed to yield likely projections of future flows. The generation of synthetic sequences involves the choice of a generating mechanism and frequently the postulation that this choice represents the underlying mechanism for the historic flows. However, statistically similar short traces can come from very different generating mechanisms, and similarity does not constitute proof that the chosen generating mechanism is that of the real world.

Over the last several years the lag 1 Markov process has found occasional application. Pertinent to the use of a Markov generating process is the estimation of the lag 1 autocorrelation ρ_1. Unfortunately the usual algorithms for estimating ρ_1 from short historic records yield estimates that are biased toward zero. This bias tends to give a distorted view of hydrologic persistence that is further accentuated in the synthetic sequences used in design simulations [*Wallis and Matalas*, 1971*a*].

In hydrology given a record of length n the most commonly recommended procedure for estimating the autocorrelation function for lag k, ρ_k, is given by

$$\hat{\rho}_k = \frac{\sum_{i=1}^{n-k} x_i x_{i+k} - [1/(n-k)] \sum_{i=1}^{n-k} x_i \sum_{i=1}^{n-k} x_{i+k}}{\left\{\sum_{i=1}^{n-k} x_i^2 - [1/(n-k)]\left(\sum_{i=1}^{n-k} x_i\right)^2\right\}^{0.5} \left\{\sum_{i=1}^{n-k} x_{i+k}^2 - [1/(n-k)]\left(\sum_{i=1}^{n-k} x_{i+k}\right)^2\right\}^{0.5}} \quad (1)$$

For the lag 1 Markov process defined as

$$x_i = \rho x_{i-1} + (1 - \rho^2)^{1/2} \epsilon_i \quad (2)$$

where ϵ_i is NIP (0, 1) and $\rho = \rho_1$, synthetic sequences were generated for preselected values of ρ_1 and n, and ρ_1 was estimated by using (1). Values of the mean bias defined as $[\rho_1 - \bar{E}(\rho_1)]$, where $\bar{E}(\rho_1)$ denotes the mean ρ_1 observed for 20,000 sequences as functions of ρ_1 and n, are presented in Table 1, and the corresponding variances of the estimates of ρ_1 are presented in Table 2. For n small, the biases are relatively large, and if the design is sensitive to measured values of ρ_1 in the synthetic sequences, the preservation of uncorrected estimates of ρ_1 could have an appreciable effect on the design. However, if a correction is applied for the expected bias in ρ_1 for both the historic and synthetic sequences, more realistic design results should be obtained. For some monthly multisite generating processes, corrected estimates of ρ_1

* Now at Imperial College, London, England.

may violate the constraints to be satisfied between the lag and cross correlations of the generating mechanism [*Matalas and Wallis*, 1971]; such a violation is prima facie evidence that an inappropriate mechanism has been chosen.

Jenkins and Watts [1968] and *Box and Jenkins* [1970] suggest that an estimator of ρ_k more efficient than (1) can be obtained by using

$$\hat{\rho}_k = \frac{\sum_{i=1}^{n-k}\left[x_i - (1/n)\sum_{i=1}^{n} x_i\right]\left[x_{i+k} - (1/n)\sum_{i=1}^{n} x_i\right]}{\sum_{i=1}^{n}\left[x_i - (1/n)\sum_{i=1}^{n} x_i\right]^2} \tag{3}$$

The sampling experiments carried out for (1) were repeated for (3), and the results are presented in Tables 3 and 4. Comparison of Table 1 with Table 3 and of Table 2 with Table 4 shows that (3) yields estimates of ρ_1 that have greater bias and smaller variance than those yielded by (1). The increased efficiency of (3) may be insufficient evidence to convince some hydrologists to stop using (1), but we hope it does not discourage them from reading further.

Other procedures for estimating ρ_1 are occasionally encountered, e.g., the algorithm,

$$\hat{\rho}_1 = \frac{\sum_{i=1}^{n-1} x_i x_{i+1} - (n-1)(\sum x_i/n)^2}{[(n-2)/(n-1)]\left[\sum_{i=1}^{n} x_i^2 - (\sum x_i/n)^2\right]} \tag{4}$$

which has been criticized by *Fiering and Jackson* [1971] and which should not be used, since it can be shown to yield estimates of ρ_1 that are both more biased and more variable than those obtained by using either (1) or (3).

Corrections for bias. A number of bias corrections for ρ_k estimated from small samples have appeared in time series analysis literature, and bias corrections for $\hat{\rho}_1$ have received particular attention. *Quenouille* [1956] has suggested the following rough method of removing bias in $\hat{\rho}_1$; it requires no assumption about the generating process and may be applied to any of the standard procedures for estimating ρ_1.

Let $\hat{\rho}_{1,1}$ denote the estimated lag 1 correlation for the first half of the series and $\hat{\rho}_{1,2}$ the corresponding quantity for the second half. If $\hat{\rho}_1$ is denoted as the estimated lag 1 correlation coefficient for the whole series, the statistic

$$\hat{\rho}_1^{\,q} = 2\hat{\rho}_1 - \tfrac{1}{2}(\hat{\rho}_{1,1} + \hat{\rho}_{1,2}) \tag{5}$$

is calculated and may be shown to reduce the bias in $\hat{\rho}_1^{\,q}$ to order n^{-2}. However, the use of this adjustment inevitably implies some loss of efficiency [*Marriott and Pope*, 1954]. The efficiency of the procedure was investigated for small samples by using (2) as the generating process and a number of algorithms for estimating ρ_1, including (1) and (3). The general conclusion was that on the average the variance of $\hat{\rho}_1^{\,q}$ was greater than the variance of unbiased ρ_1 estimated by other procedures, although the mean bias correction proved remarkably good. Consequently the procedure cannot be universally recommended, but if many short realizations of the same process are available, mean $\hat{\rho}_1^{\,q}$ may be very good.

Marriott and Pope [1954] derived a bias correction for ρ_k when a lag 1 Markov generating process is assumed. However, their result was based on a rather unusual estimator of ρ_k. In a discussion of the Marriott and Pope paper *Kendall* [1954] derived a bias correction to order n^{-1} based on the definition of $\hat{\rho}_k$ given in (1). For the lag 1 Markov process this correction was expressed as

$$E\{\hat{\rho}_k\} = \rho^k - \frac{1}{(n-k)} \cdot \left\{\frac{1+\rho}{1-\rho}(1-\rho^k) + 2k\rho^k\right\} \tag{6}$$

TABLE 1. Values of $(\rho_1 - \hat{\rho}_1)$ Obtained from Equations 1 and 2

ρ	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 60$	$n = 80$	$n = 100$
0.1	0.064	0.042	0.033	0.024	0.024	0.016	0.013
0.2	0.081	0.054	0.041	0.034	0.027	0.020	0.017
0.3	0.095	0.065	0.048	0.040	0.032	0.025	0.019
0.4	0.113	0.074	0.056	0.045	0.037	0.029	0.022
0.5	0.129	0.086	0.064	0.050	0.043	0.032	0.026
0.6	0.147	0.098	0.072	0.058	0.047	0.036	0.028
0.7	0.167	0.109	0.081	0.065	0.054	0.040	0.031
0.8	0.189	0.126	0.092	0.073	0.060	0.045	0.036
0.9	0.212	0.143	0.105	0.083	0.069	0.050	0.040

Each entry is based on 20,000 runs.

which for ρ_1 reduces to

$$E\{\hat{\rho}_1\} = \rho - \frac{1}{(n-1)}\{1 + 3\rho\} \quad (7)$$

Kendall also derived a correction for $\hat{\rho}_k$ where $\hat{\rho}_k$ was based on a circular definition of a time series. The circular estimator of ρ_k may be written as

$$\hat{\rho}_k = \frac{\sum_{i=1}^{n}\left[x_i - (1/n)\sum_{i=1}^{n} x_i\right]\left[x_{i+k} - (1/n)\sum_{i=1}^{n} x_i\right]}{\sum_{i=1}^{n}\left[x_i - (1/n)\sum_{i=1}^{n} x_i\right]^2} \quad (8)$$

The bias correction to order n^{-1} may be expressed as

$$E\{\hat{\rho}_k\} = \rho^k - \frac{1}{n}\left\{\frac{1+\rho}{1-\rho}(1-\rho^k) + 3k\rho^k - k\rho^{n-k}\right\} \quad (9)$$

and in particular for $k = 1$

$$E\{\hat{\rho}_1\} = \rho - (1/n)(1 + 4\rho) \quad (10)$$

For the lag 1 case, (8) and (3) are extremely close in that (8) includes only one extra product in the numerator. Consequently it was quite possible that (10) would provide an adequate bias correction for ρ_1 estimated from (3). However, estimates of ρ_1 derived from (1) and corrected by using (7) have their variance increased by a factor $[(n-1)/(n-4)]^2$, whereas estimates derived from (3) and corrected by using (10) have the variance amplification factor $[n/(n-4)]^2$. Application of these correction factors to the variances in Tables 2 and 4 shows that in general the corrected estimates yielded by (3) and (10) have only

TABLE 2. Variance of $\hat{\rho}_1$ Obtained from Equations 1 and 2

ρ	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 60$	$n = 80$	$n = 100$
0.1	0.048	0.032	0.024	0.020	0.016	0.012	0.010
0.2	0.048	0.032	0.024	0.019	0.016	0.012	0.010
0.3	0.047	0.031	0.024	0.019	0.015	0.012	0.009
0.4	0.046	0.030	0.022	0.018	0.014	0.011	0.009
0.5	0.045	0.029	0.021	0.016	0.014	0.010	0.008
0.6	0.042	0.026	0.019	0.015	0.012	0.009	0.007
0.7	0.040	0.024	0.017	0.013	0.011	0.007	0.006
0.8	0.037	0.022	0.015	0.011	0.008	0.006	0.005
0.9	0.035	0.019	0.012	0.008	0.006	0.004	0.003

Each entry is based on 20,000 runs.

TABLE 3. Values of $(\rho_1 - \hat{\rho}_1)$ Obtained from Equations 3 and 2

ρ	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 60$	$n = 80$	$n = 100$
0.1	0.067	0.044	0.035	0.026	0.025	0.017	0.014
0.2	0.088	0.060	0.044	0.035	0.030	0.022	0.018
0.3	0.106	0.074	0.054	0.045	0.036	0.028	0.022
0.4	0.130	0.086	0.065	0.052	0.043	0.034	0.026
0.5	0.151	0.102	0.075	0.060	0.051	0.038	0.031
0.6	0.175	0.117	0.087	0.070	0.057	0.043	0.034
0.7	0.202	0.133	0.099	0.079	0.066	0.049	0.038
0.8	0.233	0.156	0.114	0.090	0.075	0.056	0.044
0.9	0.269	0.182	0.134	0.106	0.088	0.064	0.050

Each entry is based on 20,000 runs.

marginally smaller variance than the estimates yielded by (1) and (7).

Before we proceeded to derive the appropriate bias correction for estimates of ρ_1 derived from (3), we decided to examine the ability of (10) to correct for bias in such estimates. Rearrangement of (10) yields

$$\rho = \frac{E\{\hat{\rho}_1\} + (1/n)}{1 - (4/n)} \qquad (11)$$

The replacement of $E\{\hat{\rho}_1\}$ in (11) by its sample estimate from (3) results in an estimate of ρ, $\hat{\rho} = \hat{\rho}_1$; if (11) represents an adequate correction, this corrected estimate of ρ_1 should be unbiased. Table 5 gives the residual bias in $\hat{\rho}_1$ based on 20,000 samples of size n for selected values of ρ_1. It can reasonably be concluded that the circular bias correction applied to (3) yields unbiased estimates of ρ_1. The derivation of an exact bias correction for $\hat{\rho}_1$ in this case would probably yield a quite similar analytic result; thus the rather tedious derivation was not undertaken. The few mean residual biases in Table 5 may be attributed to the violation of the underlying assumption of normality to the distribution of $\hat{\rho}_1$ for large ρ_1 or to the order of approximation of the bias correction. In the application of either (7) or (10) as bias corrections, sampling error in the uncorrected estimates may lead to corrected values of $|\rho_1| > 1.0$. To avoid such a result, upper and lower constraints for uncorrected $\hat{\rho}_1$ values may be derived by using (7) and (10); or alternatively higher order terms could be included in the bias correction algorithms. It is to be expected that the constraints on the uncorrected $\hat{\rho}_1$ will probably not be exceeded with hydrologic records unless unrealistically short sequences are used.

Those who postulate that the lag 1 Markov model is adequate for generating synthetic hydrologic traces should use values of $\hat{\rho}_1$ estimated from either (3) and (10) or (1) and (7); both combinations yield estimates that are essentially unbiased and that have similar variance. Furthermore to generate unbiased synthetic sequences, values of $\hat{\rho}_1^*$ measured from synthetic sequences of length n, where n is the length of the historic sample, should be such that

TABLE 4. Variance of $\hat{\rho}_1$ Obtained from Equations 3 and 2

ρ	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 60$	$n = 80$	$n = 100$
0.1	0.043	0.030	0.023	0.019	0.016	0.012	0.010
0.2	0.043	0.030	0.023	0.019	0.016	0.012	0.010
0.3	0.042	0.029	0.022	0.018	0.015	0.011	0.009
0.4	0.042	0.028	0.021	0.017	0.014	0.011	0.009
0.5	0.040	0.027	0.021	0.016	0.013	0.010	0.008
0.6	0.038	0.025	0.019	0.014	0.012	0.009	0.007
0.7	0.037	0.023	0.016	0.013	0.010	0.007	0.006
0.8	0.033	0.021	0.015	0.011	0.009	0.006	0.005
0.9	0.031	0.018	0.012	0.009	0.007	0.004	0.003

Each entry is based on 20,000 runs.

TABLE 5. Value of the Residual Bias ($\rho_1 - \hat{\rho}_1$) When $\hat{\rho}_1$ Is Estimated by the Box and Jenkins Algorithm and the Kendall Circular Bias Correction

	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 60$	$n = 80$	$n = 100$
Value of ρ							
−0.9	−0.02	−0.01	−0.00	−0.00	−0.00	−0.00	−0.00
−0.8	−0.01	−0.01	−0.00	−0.00	−0.00	−0.00	−0.00
−0.7	−0.01	−0.01	−0.00	−0.00	−0.00	−0.00	−0.00
−0.6	−0.01	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
−0.5	−0.01	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
−0.4	−0.01	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
−0.3	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
−0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
−0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
+0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
+0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
+0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00
+0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00
+0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00
+0.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00
+0.7	0.01	0.01	0.00	0.00	0.00	0.00	0.00
+0.8	0.02	0.02	0.01	0.01	0.01	0.00	0.00
+0.9	0.04	0.03	0.02	0.01	0.01	0.01	0.00
Upper constraint (10)	+0.75	+0.83	+0.87	+0.90	+0.92	+0.94	+0.95
Lower constraint (10)	−0.85	−0.90	−0.92	−0.94	−0.95	−0.96	−0.97

Each entry is the mean of 20,000 sequences generated by a lag 1 Markov process defined by equation 2.

$$E[\hat{\rho}_1^*] = \hat{\rho}_1 \qquad (12)$$

where both $\hat{\rho}_1^*$ and $\hat{\rho}_1$ are corrected for bias.

Unbiased $\hat{\rho}_1$ as a test for independence. Numerous tests for establishing statistical independence in a hydrologic time series are available in the literature [*Matalas*, 1967]. All the commonly advocated tests lack power when they are confronted with small samples from processes that are noncyclic and that have small expected values for ρ_1. Type 2 errors for several of the more common tests have been documented [*Wallis and Matalas*, 1971*b*] for both the lag 1 Markov generating process and for filtered fractional noises. A small sample test more powerful than the conventional ones can be obtained by combining (3) and (10) or (1) and (7) and using the confidence limits for $\hat{\rho}_1$ given by

$$\text{CL}\,[\hat{\rho}_{(1)}] = [-1 \pm Z_\alpha(n-2)^{1/2}]/(n-1) \qquad (13)$$

where Z_α is the standard normal deviate corresponding to a probability level α. Simulation results for 2000 runs from a lag 1 Markov process for $n = 25$, 50, and 100 and $\alpha = 0.05$ are presented in Figure 1; it is evident that misclassification errors for $\rho_1 \neq 0.0$ are large when n is small and ρ_1 large, though much smaller than those from similar tests reported elsewhere.

Small sample bias correction of $\hat{\rho}_1$ when Hurst's $h \neq 0.5$. For those who postulate that hydrology can best be modeled by using stochastic processes for which $h \neq 0.5$, the unbiased small sample estimation of $\hat{\rho}_1$ is a more complex task. For instance, it is known that for given ρ_1 the

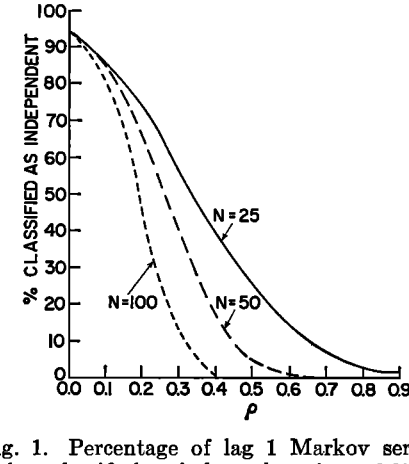

Fig. 1. Percentage of lag 1 Markov series of length n classified as independent ($\alpha = 0.05$) and based on unbiased ρ_1.

TABLE 6. Mean Bias ($\rho_1 - \hat{\rho}_1$) Obtained by Different Estimating Procedures for a Filtered Fractional Noise for $M = 1000$, $p = 1$, and $h = 0.9$ and Various Small n

$\hat{\rho}_1$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 60$	$n = 80$	$n = 100$
Equation 1	+0.38	+0.31	+0.26	+0.23	+0.21	+0.18	+0.16
Equations 3 and 10	+0.24	+0.21	+0.19	+0.17	+0.16	+0.14	+0.13
Equations 1 and 5	+0.19	+0.16	+0.14	+0.12	+0.11	+0.10	+0.09

bias in ρ_1 for filtered fractional noises is a positive function of the parameter h [*Wallis and Matalas*, 1971*a*], but small sample estimates of h are also biased [*Wallis and Matalis*, 1970]. Good theoretical work on small sample biases when $h \neq 0.5$ is definitely needed. The expected value of ρ_1 for a filtered fractional noise when $M = 1000$, $p = 1$, and $h = 0.9$ is 0.867, but the mean estimated value when (1) is used is much lower (Table 6). The Quenouille correction applied to (1) leaves only a small residual bias but cannot be recommended unless many separate realizations are available, since the averages shown used many individual estimates of $\rho_1^q > 1.0$. The corrected estimates when (3) and (10) are used have a larger residual bias as would be expected. It should be realized that a filtered fractional noise when $h = 0.9$ and $\rho_1 = 0.867$ is an extreme process that is quite unlikely to be used for hydrologic simulations. The residual bias in $\hat{\rho}_1$ for noises near annual hydrology ($h = 0.7$, $\rho_1 = 0.3$) is not expected to be excessive.

Acknowledgment. Partially supported by Office of Water Resources Research grant 14-31-001-3691, 1972.

REFERENCES

Box, J. P., and G. M. Jenkins, *Time Series Analysis Forecasting and Control*, 553 pp., Holden-Day, San Francisco, Calif., 1970.

Fiering, M. B, and B. B. Jackson, *Synthetic Streamflows, Water Resour. Monogr. 1*, p. 46, AGU, Washington, D. C., 1971.

Jenkins, G. M., and D. G. Watts, *Spectral Analysis and Its Applications*, 525 pp., Holden-Day, San Francisco, Calif., 1968.

Kendall, M. G., Note on the bias in the estimation of autocorrelation, *Biometrika, 42*, 403–404, 1954.

Marriott, F. H. C., and J. A. Pope, Bias in the estimation of autocorrelations, *Biometrika, 42*, 390–402, 1954.

Matalas, N. C., Some aspects of time series analysis in hydrologic studies, *Proc. Hydrol. Symp., 5*, 271–309, 1967.

Matalas, N. C., and J. R. Wallis, Correlation constraints for generating processes, *Proc. Warsaw Symp., 2*, 697–707, 1971.

Quenouille, M. H., Notes in bias in estimation, *Biometrika, 43*, 353–360, 1956.

Wallis, J. R., and N. C. Matalas, Small sample properties of H and K—Estimators of the Hurst coefficient h, *Water Resour. Res., 6*(6), 1583–1594, 1970.

Wallis, J. R., and N. C. Matalas, In hydrology h is a household word, *Proc. Warsaw Symp., 1*, 375–393, 1971*a*.

Wallis, J. R., and N. C. Matalas, Correlogram analysis revisited, *Water Resour. Res., 7*(6), 1448–1459, 1971*b*.

(Manuscript received January 31, 1972; revised February 9, 1972.)