Hidrolojik Süreçlerin Standard Sapmalarının Taraflılığının Düzeltilmesi

Bias Correction Of The Standard Deviation Of Hydrological Processes

Zekâi ŞEN [1]

Tabiatta sürekli bir şekilde oluşan yağış, yüzeysel akış, sızma, yeraltı suyu, buharlaşma gibi hidrolojik olayların ölçümleri neticesinde elde edilen zaman serilerinin süresi oldukça kısadır. Kısa olan bu serilerden elde edilen parametre tahminleri taraflı olurlar. Bu makalede standard sapmadaki taraflılığın giderilmesi için gerekli analitik ifadeler çıkartılmıştır.

The measurements of hydrological phenomena such as the precipitation, surface flow, infiltration, groundwater, evaporation etc. which evolve continuously in the nature, constitute a time series of short length. The parameter estimation of a mathematical model suitable to predict the future values is biased due to the small samples. In this paper, necessary analytical expressions have been derived for the bias correction of standard deviation of various hydrological processes.

1 — INTRODUCTION

The generation of synthetic data by the use of various stochastic processes has assumed a very important place in the design and operation of water resources systems. In this context, a series of generating models has appeared in the hydrological literature. Currently emp-

[1] Department of Hydraulics and Water Power Technical University of Istanbul

loyed ones are the Markov process, the ARIMA (1, 0, 1) process, the Broken Line process and the white Markov process [4]. On the other hand, for the simulation of long-term persistence present in hydrological time series, the discrete fractional Gaussian process has been presented into hydrology [1]. Each one of these models has its own drawbacks, for example, Markov process fails to preserve both the long-term persistence measure, h. (Hurst coefficient) and short-term persistence measure, ρ, (first order autocorrelation coefficient), simultaneously. The ARIMA (1, 0, 1) process and the white Markov processes give a range of h values for a fixed value of ρ.

So, it can be said that, these two models are more flexible than the Markov model which is very rigid as far as the choice of h for a given ρ is concerned. The model which appears to be the first to achieve the preservation of h and ρ simultaneously, is discrete fractional Gaussian process (*dfGn*) but it has its own drawbacks in that a very large computer time and memory are required even for a short syntehtic sequence.

Another very important topic which engaged recently various hydrologists is the effect of bias on parameters of a given model. Due to the undesirable bias effect the design obtained can be either underdesigned or overdesigned; the occurrence of any one of them is associated with a loss relative to the design which would emerge with unbiased parameter estimates.

Most of the bias correction formula that appeared in the hydrology literature [2], [3], and [4] are proposed for the autoregressive models and parameters ρ and σ^2, the serial correlation coefficient and variance only. The objective of this paper is to derive analytical expressions for bias corrections of standart deviation which is one of the driving parameters in autoregressive models.

2 — BIAS EFFECT

The assumption, which has become a prerequisite in the generation of synthetic sequences, is that the historic data measured in the field is but a single sample out of the underlying population. Therefore, it has the bias in the information it provides for any kind of parameter that can be extracted from this available finite sample.

Let $\theta_1, \theta_2, \theta_3, \dots, \theta_n$ be a set of population parameters which the hydrologists are interested in. In the case of a streamflow sequence $\theta_1, \theta_2, \theta_3$ can be regarded as being the mean μ, the standard deviation σ, and the first order serial correlation coefficient, respectively. Because of this single and finite sequence of observations taken at one site, aforementioned set of parameters will have their estimated counterparts abstracted from the information provided by the historic data. Let this estimated set be $\widehat{\theta_1}, \widehat{\theta_2}, \theta_3, \dots \widehat{\theta_n}$. If the population were known to hydrologist, then, the quantitative measure of the amount of bias attached to each of the parameters would be $\hat{\theta}_1 - \theta_1$, $\widehat{\theta_3} - \widehat{\theta_3}, \dots,$ $\widehat{\theta_2} - \widehat{\theta_2}$, $\widetilde{\theta_n} - \widehat{\theta_n}$.

The bias amounts found in this way can never be eliminated unless the estimates are exactly equal to their population counterparts. This can occur only when the length of sample is infinite, that is to say, all of the information about the population is known. In practice, it is not possible to know all of the information due to the paucity of historic data. In hydrology, this type of bias is referred to as the operational bias which cannot be eliminated by any mathematical methods. The only way to deal with such a bias is to assume that estimates are all equal to the corresponding population parameters. Even at this stage there exist two alternatives one of which is to assume the small sample estimates to be equal to the population parameters without any bias correction whereas the second alternative is to correct the small sample estimates for bias and then to assume that these bias corrected estimates are the same as the population parameters. Of course, to perform such a bias correction first of all the generating mechanism of the historic data must be identified. All of the proposed bias correction procedures are dependent on the underlying generating process, the sample length and finally on the nature of estimator whether it is a maximum likelihood estimator or moment estimator or Bayesian estimator.

However, another kind of bias, which is known as the statistical bias can be mathematically eliminated. The statistical bias occurs only in the generating scheme itself, whereas the operational bias concerns the transition from single data to the generating mechanism. The statistical bias can be described in a concrete form as follows. After the hydrologists decide on the underlying generating mechanism of the available data, the next step is to work out the estimates of these parameters which appear in the structure of the model adopted. For instance, in the case of a Markov model, the number of parameters is three; when the ARIMA (1, 0, 1) process is adopted then the driving parameters are four namely, μ, σ, θ and ϕ. These parameters with or without bias corrections applied, are assumed to equal the population correspondants.

After this assumption synthetic sequences of any desired length are generated on the basis of the mathemtical model adopted. The synthetic samples of length n constitute an ensamble where each one of the member sequence is equally likely to represent the future of the phenomenon considered. Consequently, each one of the member sequence yields estimates of particularly interested parameters which turn out to be different from the assumed population parameters. The ensemble averages of the synthetic estimates will stabilize at a constant value which is denoted by $E\,(\widehat{\theta_n})$. This overall ensemble average will be different from the corresponding population parameter where the difference shows the amount of bias (statistical bias), $\theta - E\,(\widehat{\theta_n})$.

The statistics literature concerning the mathematical bias correction procedures has been reviewed by Wallis and O'Connell [3], who have presented various estimetes for the first order serial correlation coefficient and the amount of bias associated with the Markov process on the basis of computer simulations through the Monte Carlo techniques. The original form of bias correction procedure and its application to various types of models have been proposed by Kendall [2]. The application of the same procedures to the ARIMA (1, 0, 1) process and the white Markov process have been performed by Şen [4]. Bias correction for the variance of a sequence generated by the lag - one Markov process was given by Fiering [5]. In a similar way, bias correction for the variance of the ARIMA (1, 0, 1) process has first been derived by O'Connell [6].

Although the standard deviation is the positive square root of the variance, the same procedure is not valid for the bias corrections. In other words, the bias correction of the standard deviation is not the square root of the bias correction of variance. Thus, the necessary procedure for the bias correction of the standard deviation must be developed independently from the variance.

3 — BIAS CORRECTION OF STANDARD DEVIATION

For any given sequence of observations $x_1, x_2, x_3, \ldots, x_n$, the variance which is the measure of spread of observations about the mean value is expressed as,

$$S^2_n = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (1)$$

The right hand side (rhs) of this expression can be expanded which leads to,

$$S^2_n = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right] \qquad (2)$$

Although S_n^2 is uniquely obtained from a given sequence of events, when an ensemble of the sequences of the same length n is concerned, for every member of this ensemble, the value of S_n^2 is different and consequently an ensemble of S_n^2 is obtained. As a result, S_n^2 can be considered as a random variable (r.v.) and in short $A_n - S_n^2$ where A_n is a new random variable. The probability distribution function (pdf) of A_n can be shifted to its mean by,

$$a_n = A_n - E\,(A_n) \qquad (3)$$

here a_n is a newly defined r.v. which has exactly the same pdf as A_n in the shape and the moments of the two r.v. is related by Eq. 3. Thus, the expected value of a_n is $E\,(a_n) = E\,(A_n) - E\,(A_n) = 0$; and the expected value of A_n for various processes are as follows ; for the white-noise process

$$E\,(A_n) = E\,(S_n^2) = \sigma^2 \qquad (4)$$

For the Markov process the value was first provided by Fiering as,

$$E(A_n) = E(S^2) = \sigma^2 \left\{ 1 - \frac{2\rho}{n(n-1)} \left[\frac{n(1-\rho) - (1-\rho^n)}{(1-\rho)^2} \right] \right\} \qquad (5)$$

In the case of the ARIMA (1, 0, 1) process, similar expression was given by O'Connell [6], as

$$E(A_n) = E(S_n^2) = \sigma^2 \left\{ 1 - \frac{2\rho}{n(n-1)} \left[\frac{n(1-\phi) - (1-\phi^n)}{(1-\phi)^2} \right] \right\} \qquad (6)$$

From Eq. 2 the standard deviation can be written as,

$$S_n = \sqrt{E(A_n)}\left[1- \frac{a_n}{E(A_n)}\right]^{1/2} \tag{7}$$

This last expression can be expanded into Binomial series which after expectation operation leads to,

$$E(S_n)=\sqrt{E(A_n)}\left[1+\frac{E(a_n)}{2E(A_n)}-\frac{E(a^2{}_n)}{8E^2(A_n)}+\frac{E(a^3{}_n)}{16E^3(A_n)}-\frac{5.E(a^4{}_n)}{128E^4(A_n)}+\cdots\right]$$

In this expression $E\ (a_n) = 0$ and if it was easy to calculate the second and higher order central moments of r.v. a_n then the avobe expression should yield an exact value of $E\ (S_n)$. In order to reduce the burden of complicated calculations, higher order moments than two, will all be ignored and consequently the following approximate formulae is obtained.

$$E(S_n)= \sqrt{E(A_n)}\left|1- \frac{E(a^2{}_n)}{8E^2(A_n)}\right| \tag{8}$$

Or replacing A_n by $S_n{}^2$ the expression becomes,

$$E(S_n)= \sqrt{E(S^2{}_n)}\left|1- \frac{E(a^2{}_n)}{9E^2(S^2{}_n)}\right| \tag{9}$$

The only thing remains to be found is that of $E\ (a_n{}^2)$ which is expressible in terms of moments of r.v. A_n as follows,

$$a^2{}_n = A^2{}_n - 2\,E(A_n).A_n + E^2(A_n)$$

By taking expectations of both sides

$$E(a^2{}_n)=E(A^2{}_n)-E^2(A_n)=V(A_n)$$

That is the variance of A_n in turn becomes the variance of $S_n{}^2$. Therefore, the following sequence of relationship is valid,

$$E(a^2{}_n)=V(A_n)=V(S^2{}_n) \tag{10}$$

After the incorporation of this new finding the general expression in Eq. 9 becomes,

$$E(S_n)=\sqrt{E(S_n{}^2)}\left[1- \frac{V(S_n)}{8E^2(S^2{}_n)}\right] \tag{11}$$

The only unknown term is $V\ (S_n{}^2)$. This term has been given by Bailey and Hammersley [7] for the normal independent process as,

$$V(S^2{}_n)=\frac{2.\sigma^4}{(n-1)} \tag{12}$$

The substitution of Eq. 4 and Eq. 12 into Eq. 11 gives.

$$E(S_n)=\sigma\left[1-\frac{1}{4(n-1)}\right] \tag{13}$$

There exist two facts that can be proven by looking at this last expression. Firstly, as the sample size increases the second term in the brackets tends to zero and in this way the bias effect diminishes.

$$\lim_{n\to\infty} E\ (S_n)\ =0 \tag{14}$$

Secondly, the amount of bias is not the square root of the bias amount of variance of corresponding standard deviation. Final word is that, although the estimate of $S_n{}^2$ given in Eq. 1 yields an unbiased estimate of variance, the same is not valid for the standard deviation. The amount of bias in the case of normal independent process can be found from Eq. 13 as,

$$\sigma-E(S_n)=\frac{1}{4(n-1)} \tag{15}$$

If an unbiased estimate of standard deviation is required the following expression must be employed instead of Eq. 18.

$$S^2{}_n=\frac{1}{(n-1)\left[1-\dfrac{1}{4(n-1)}\right]^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{16}$$

Of course, such an estimator yields biased variance, which is expressible as

$$E(S^2{}_n)=\frac{\sigma^2}{\left[1-\dfrac{1}{4(n-1)}\right]^2} \tag{17}$$

One very important conclusion that can be drawn after all of the above calculations is that, it is not possible to have simultaneously unbiased variance and standard deviation. This is valid only for small sample.

LAG — ONE MARKOV PROCESS

The same general formula given in Eq. 11 is valid with new forms of terms. The general form of the variance of variance $V\ (S_n^2)$ is given by Bailey and Hammersley [7] with an analogy to the normal independent process case in Eq. 12 as,

$$V(S^2{}_n) = \frac{2.\sigma^4}{(n_v^* - 1)} \tag{18}$$

where $n_v{}^*$ is given by the same authors in its general form for an autoregressive process as,

$$n_v^* = n \cdot \frac{1 + \dfrac{1}{(n-1)} \displaystyle\sum_{j=1}^{n-1} \rho^2{}_j}{1 + \dfrac{n}{(n-1)} \displaystyle\sum_{j=1}^{n-1} \rho^2{}_j} \tag{19}$$

It has also been shown by the same authors that a reasonable approximation to this huge expression will be,

$$n_v^* = \frac{n}{\displaystyle\sum_{j=0}^{n-1} \rho^2{}_j} \tag{20}$$

Hence the application of the general $E\ (S_n)$ expression can be carried out with the above introduction. First, an attempt will be made to reach at a general form of $E\ (S_n)$ for an autoregressive process. To do this the following abbreviation is employed.

$$E\ (S_n^2) = \sigma^2 \cdot F \tag{21}$$

where F denotes the second product term which is in the brackets of Eq. 5 and Eq. 6. Hence the general $E\ (S_n)$ becomes

$$E(S_n) = \sigma \sqrt{F} \left[1 - \frac{1}{4(n_v^* - 1)F^2} \right] \tag{22}$$

This formulae reduces to Eq. 11 when F is set equal to one and $n = n_v{}^*$ which is the normal independent process case. For the lag-one Markov process $n_v{}^*$ becomes,

$$n_v^* = n\,\frac{(n-1)(1-\rho^2)+\rho^2[1-\rho^{2(n-1)}]}{(n-1)+\rho^2[1-n\ \rho^{(n-1)}]} \qquad (23)$$

If the approximate form of $n_v{}^*$ was considered than $n_v{}^*$ should came out as,

$$n_v^* = \frac{n(1-\rho^2)}{(1-\rho^{2n})} \qquad (24)$$

The value of F for the Markov process can be seen from Eq. 5

$$F = 1 - \frac{2\rho}{n(n-1)}\left[\frac{n(1-\rho)-(1-\rho^n)}{(1-\rho)^2}\right] \qquad (25)$$

Considering the smallest sample sizes used in hydrology, that is $n = 10$ or onwards, the bias correction factors found through the use of general expression give satisfactory results even for large values of ρ, which is not commonly used in hydrologic studies.

THE ARIMA (1, 0, 1) process

The autocorrelation structure of such a process is dependent on two parameters namely ϕ and θ.

$$\rho_k = \rho\,.\,\phi^{k-1} \quad \text{for} \quad k \geq 2$$

and where ρ is a function of both ϕ and θ in the following way

$$\rho = \frac{(1-\phi\,.\,\theta)(\phi-\theta)}{1+\theta^2-2\,.\,\phi\,.\,\theta}$$

By substituting the above autocorrelation function in Eq. 19 and Eq. 20, the exact and approximate values of $n_v{}^*$ can be obtained. Avoiding the calculations only the final results of $n_v{}^*$ will be given :

$$n_v^* = n\,\frac{(n-1)(1-\phi^2)+\rho^2[1-\phi^{2(n-1)}]}{(n-1)(1-\phi)-n\rho^2[1-\phi^{2(n-1)}]} \qquad (26)$$

and the approximate form is given as,

$$n_v^* = \frac{n(1-\phi^2)}{1-\phi-\rho^2[1-\phi^{2(n-1)}]} \qquad (27)$$

The expression of F is taken from Eq. 6 which is,

$$F = 1 - \frac{2 . \rho}{n(n-1)} \left[\frac{n(1-\phi) - (1-\phi^n)}{(1-\phi)^2} \right] \tag{28}$$

For the ARIMA (1, 0, 1) process bias correction factors provided by Eq. 22 are not valid for entire range of parameters of ϕ and θ. It is a furtunate that the analytical expressions are valid for ϕ and θ values which are employed in hydrology.

CONCLUSIONS

Although there is a quadratic relationship between the variance and the standard deviation, the same relationship is not valid for bias corrections. Hence, a different bias correction prodecure must be developed for the standard deviation.

An important fact is that, it is not possible to have unbiased values of standard deviation and variance simultaneously. A preference must be made between the two parameters. In general, it is the standard deviation that appears in the model structure hence, one gets the impression that not the variance but the standard deviation should be corrected for bias. This way has been adopted in various hydrological studies, in assessing the effect of bias on various design situations.

Another important conclusion is that the difference between bias corrections of σ and σ^2 agains importance only for small samples whereas for large samples they are the same.

REFERENCES

(1) Mandelbrot, B. B. and J. R. Wallis, Computer Experiments with Fractional Gaussian Noises, Water Resources Research, Vol. 5, 1969.

(2) Kendall, M. G., Note on the Bias in the Estimation of Autocorrelation Biometrica, Vol. 42, 1954.

(3) Wallis, J. R. and P. E. O'Connel, Small Sample Estimation of, ρ_1 Water Resources Research, Vol. 8, 1972.

(4) Şen, Z., Small Sample Properties of Stationary Stochastio Processes and the Hurst Phenomenon in Hydrology, Ph. D. Thesis, London, Imperial College, 1974.

(5) O'Connell, P. E. and J. R. Wallis, Choice of Generating Mechanism in Synthetic Hydrology with Inadequate Data, Int. Assoc. Hydrol. Sci., Madrid Symposium, June 1973.

(6) Fiering, M. B., Streamflow Synthesis, MacMillan and Company Ltd., London, 1967.

(7) Bayley, G. V. and Hammersley, J. M., The Effective Number of Independent Observations in an Autocorrelated Series, Journal of the Royal Statistica! Society, Vol. 8 (1 - B), London, 1946.