# ESTIMATION OF THE SELF-SIMILARITY PARAMETER IN LONG MEMORY PROCESSES

**M. M. A. Sarker**
Department of Mathematics, BUET, Dhaka-1000, Bangladesh.
Corresponding email: masarker@math.buet.ac.bd

***Abstract:*** *Long memory processes where positive correlations between observations far apart in time and space decay very slowly to zero with increasing time lag, occur quite frequently in fields such as hydrology and economics. Stochastic processes that are invariant in distribution under judicious scaling of time and space, called self-similar process, can parsimoniously model the long-run properties of phenomena exhibiting long-range dependence. Four of the heuristic estimation approaches have been presented in this study so that the self-similarity parameter, H that gives the correlation structure in long memory processes, can be effectively estimated. Finally, the methods presented in this paper were applied to two observed time series, namely Nile River Data set and the VBR (Variable- Bit-Rate) data set. The estimated values of H for two data sets found from different methods suggest that all methods are not equally good for estimation.*

***Keywords:*** *Long memory process, long-range dependence, Self-similar process, Hurst Parameter, Gaussian noise.*

## INTRODUCTION

The analysis of experimental data that have been observed at different time points leads to new and unique problems in mathematical/statistical modeling and inference. The obvious correlations introduced by the sampling of adjacent time points can severely restrict the applicability of the many conventional statistical methods those traditionally depend on the assumption that the adjacent observations are independent and identically distributed. The accuracy in forecasting depends on actively detecting the underlying trend in a time series data. Numerous efforts have been made in this regard, the analysis of the behavior of Nile river and several similar time series led to the discovery of Hurst effect[1,3]. Motivated by Hurst's empirical findings, Mandelbrot and co-researchers[4-6] later introduced fractional Gaussian noise as a statistical model with long-range dependence. The works of J. Beran[7] on statistics for long memory processes is considered a pioneering one in this arena. Estimating the self-similarity parameter by different methods applying in two data sets namely, Nile river data set and the VBR data set is the main goal of this paper. Autocorrelation is a useful tool for finding repeating patterns not only in statistics but also in a signal, such as determining the presence of a periodic signal which has been buried under noise. The autocorrelation function (ACF) of a random process describes the correlation between the processes at different points in time. Let $X_t$ be the value of the process at time $t$ (where $t$ may be an integer for a discrete-time process or a real number for a continuous-time process). If $X_t$ has mean μ and variance $\sigma^2$ then ACF can be defined as

$$r(k) = E\left[(X_i - \mu)(X_{i+k} - \mu)\right] / \sigma^2$$

where $E$ is the expected value operator, $k$ is the lag, $|t - s|$. Brief descriptions of each data set along with the relevant figures are included here to get a preliminary idea about the behavior of the data.

***Nile River data set:*** This data set is based on the minimal water level of the Nile river for the years (622 AD-1284 AD) measured at Roda Gause near Cairo, Egypt[7]. An ACF (Auto-Correlation Function)[8] plot of Nile data set is shown in Fig.1 where it is evident that there is a long period when the minimum water level tends to be high and another long period where water level seems to be relatively low but the whole series looks to be stationary. ACF plots show that autocorrelation decreases towards zero at a very slower rate which may be considered as a hint of long-range dependence structure.

***VBR data set:*** The VBR (Variable-Bit-Rate) data set is based on video traffic measurements over asynchronous transfer mode (ATM). Data set contains the amount of coded information per frame for a certain video scene. The scene consists of a conversation among three people sitting around a table. This data set is a part of a longer series, which contains only 1000 coded information per frame based on about 30 minutes of video film. About 25 frames per second are processed. The data set was generated by engineers at Siemens, Munich using VBR codec that was especially designed for high-speed networks[3]. Time series of VBR in Fig. 2 indicates that there are some local periods with a large number of ATM cells and in some periods with a small number of cells. The ACF plot indicates that there may be long-lasting strong dependence and the dependence structure seems to be more complicated than that for the Nile river data set.

## FORMULATION OF THE METHODS

Long-range dependence had been known long before suitable stochastic models were developed. It was observed empirically in many cases that correlation between observations which are far apart in time or space decay to zero at a slower rate than expected from independent observations.

### The Hurst Phenomenon

Hurst Phenomenon was brought to light through the calculation procedure of the capacity of a fictitious reservoir that would have been 'perfect' for the time span between j and j + k (j, k ∈ N, the set of positive integers). To avoid any complexity, it was assumed that time is discrete and that there are no storage losses owing to evaporation, leakage etc. 'Perfect capacity' means that the outflow is uniform, that is, at time t + k the reservoir is as full as at time t and that the dam never overflows. If $X_i$ denote the inflow at time i (i ∈N, i ≥ 1) and the cumulative inflow up to time j is

$Z_j = \sum_{i=1}^{j} X_i$ , then the 'perfect capacity' is defined as

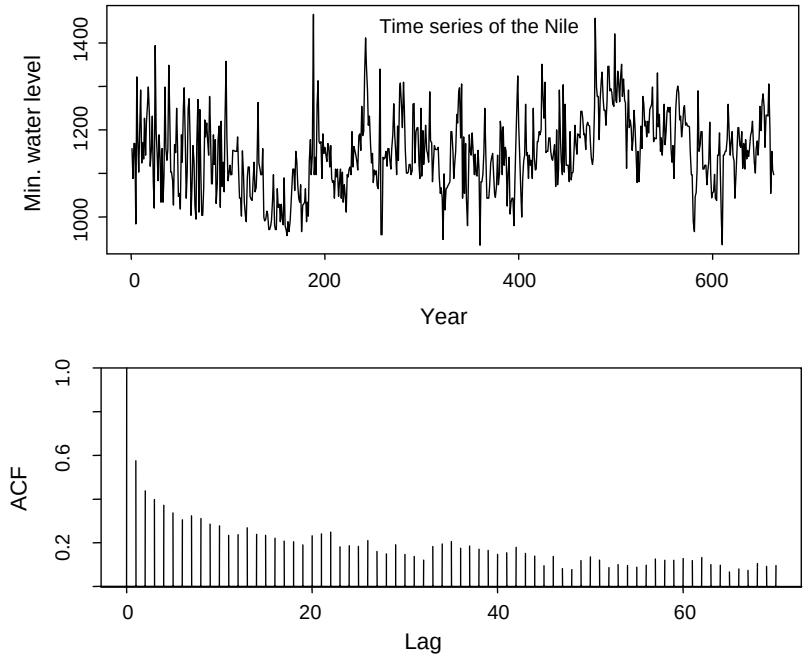


Fig. 1: Yearly minimum water level of Nile River [Top] and ACF up to Lag 70 [Bottom]

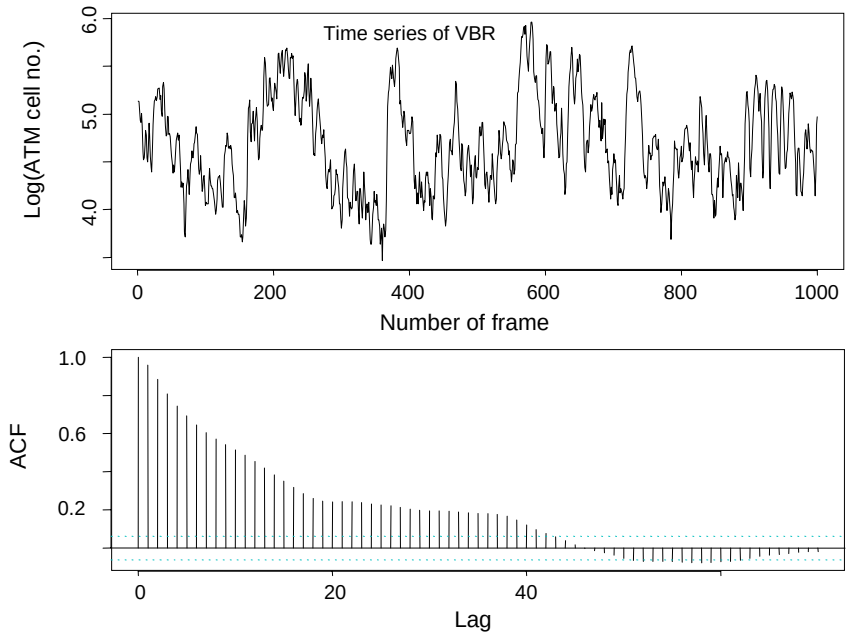


Fig. 2: Plot of VBR data using ln [no. of ATM] (top) along with its ACF (bottom).

$$R(t,k)=\max_{0\le i\le k}\left[z_{t+i}-z_t-\frac{i}{k}\left(z_{t+i}-z_t\right)\right]-\min_{0\le i\le k}\left[z_{t+i}-z_t-\frac{i}{k}\left(z_{t+i}-z_t\right)\right] \quad (1)$$

*R(t, k)* is called adjusted range. The definition also applies to cases where $X_i$ is negative. To study the scale-independent properties, it is required to divide *R(t, k)* by the empirical standard deviation of $X_i$, given by

$$S(t,k)=\sqrt{\frac{1}{k}\sum_{i=t+1}^{t+k}X_i^2-\left(\frac{1}{k}\sum_{i=t+1}^{t+k}X_i\right)^2} \quad (2)$$

The quotient $R/S=R(t,k)/S(t,k),(k\ge 2)$ is called the rescaled adjusted range or R/S–statistics. For many hydrological, geophysical and climological records, Hurst plotted the logarithm of R/S against several values of time lag k and observed that for large values of k, log R/S was scattered around a straight line with slope greater than ½, i.e., R/S behaves like a constant time $K^H$ for some H >1/2. This feature is known as the Hurst phenomenon after the name of famous Hydrologist H. E. Hurst [9-11].

**Slowly decreasing variances of sample means**

Long memory processes include spatial data, which

exhibit long-range dependence. In the result of an agricultural experiment[12], Smith and Graf showed that *V(k)*, the variance of the average yield for a plot of size k can be estimated by the sample variance of the averages for plots and suggested a relationship of the form

$$\log V(k) = a + b \log k \tag{3}$$

with b around –0.749. Relation (3) implies that *V(k)* converges to zero at a slower rate of convergence than it would have happened if the observations were weakly dependent or independent. Same behavior was noted for several dozens of other agronomical uniformity trials. Assuming that for each pair of positions (t, s), correlation ρ(t, s) depends on the Euclidian distance between t and s only, it has been proved that if $\rho(t, s) = \rho(|t, s|)$ decays asymptotically like $|t - s|^{4H-4}$ for some H in between ½<H<1, then V(k) converges to 0 (zero) like a constant times $k^{2H-2}$. If the observations for different plots are independent or the correlation decay fast, then *V(k)* decays like a constant times $K^{-1}$.

## RESULTS AND DISCUSSION

***Long memory processes:*** Intuitively, long-range dependence or long memory processes means that the positive correlation of a stationary process decays very slowly to zero than expected from independent observations. Let $X_t$ be a stationary process and let us assume that there exists a real number $\alpha \in (0,1)$ and a constant $c_\rho > 0$ such that

$$\lim_{k\to\infty} \frac{\rho(k)}{\left[c_\rho k^{-\alpha}\right]} = 1 \tag{4}$$

Then $X_t$ is called a stationary process with long memory or long-range dependence or strong dependence or stationary process with slowly decaying correlation or long-range correlation. Equation (4) simply implies that the correlations ρ(k) are asymptotically equal to a constant $c_\rho$ time $k^{-\alpha}$, $0<\alpha<1$. In terms of the Hurst parameter $H = 1 - \alpha/2$, long memory occurs for ½<H<1.

Imposing a restriction on spectral density, long-range dependence can also be equivalently defined in terms of spectral density. Let $X_t$ be a stationary process and let there exists a real number $\beta \in (0,1)$ and a positive constant $c_f$ such that

$$\lim_{\lambda\to 0} \frac{f(\lambda)}{\left[c_f |\lambda|^{-\beta}\right]} = 1 \tag{5}$$

Equation (5) indicates that the spectral density f(λ) has a pole at zero that is equal to a constant $c_f$ times $\lambda^{-\beta}$ for some $0<\beta<1$.

It may be noted that the definition given above is an asymptotic one that only tells about the ultimate behavior of the correlation as the lag tends to infinity but does not specify the correlation for any fixed finite lag. It merely determines the rate of convergence, not the absolute size.

***Self-similar process:*** A body is called geometrically self-similar if the same geometric structures are observed irrespective of the distance from which the body has been looked at. A stochastic process X(t) ($\forall t \in R^+$, the set of real numbers greater than or equal to zero) is called self-similar with self-similarity parameter H>0, if for c>0,

$$X(ct) \overset{d}{\cong} c^H X(t) \tag{6}$$

i.e., the finite dimensional distributions of X(ct) are identical to the finite dimensional distributions of $c^H$ X(t). H is also called Hurst parameter, a scaling index. Equation (6) explains the fact that for all ($t_1$ , $t_2$ , …., $t_k$ ) $\in R^n$ with $n \in N$ and c>0, $c^H$ ($X(t_1)$, $X(t_2)$,…, $X(t_k)$) has the same distribution as ($X(ct_1)$, $X(ct_2)$, …$X(ct_k)$). That means the typical sample paths of a self-similar process look qualitatively the same, independent of the distance from which one looks at them.

### *Estimation of self-similarity parameter*

The self-similarity parameter, H, which determines the strength of the correlation, needs to be estimated in many cases. Efforts have been made to approach the problem of testing for and estimating the degree of self-similarity from four different angles: (1) Time domain analysis based on the R/S statistics; (2) Analysis of the variances of the aggregated processes; (3) Periodogram based analysis in the frequency domain and (4) Correlogram based analysis on the time domain. A brief description of the corresponding statistical and graphical methods and their applications in analyzing the Nile and VBR data set are given next.

### *Rescaled adjusted range R/S*

To infer the degree of self-similarity is the objective of the R/S analysis of an empirical record. This graphical heuristic approach tries to exploit the information in a given record as much as possible. Let us consider a sample of n observations and subdivide the sample into p non over-lapping blocks and compute R(t, k)/S(t, k) for different time points t and lags k satisfying $t + k \le n$. For each lag k, we should have many samples of R/S, as many as p for small k and as few as 1 when k is close to the total sample size n. Then the plot of the R/S versus k in logarithmic scale with $k \approx 10$ is called rescaled adjusted range plot or pox plot of R/S. The points on the pox plot are expected to fluctuate around a straight line with certain slope. A straight line is then fitted to the plot. The asymptotic slope of the simple least square fitted line serves as the estimate of the self-similarity parameter, which can take any value from ½<H<1. Since any short-range dependence in the series typically results in a temporary zone near the lower end of the plot, so it is better to set a cut-off point for the purpose of estimating H. Similarly, it is not wise to use the extreme higher end of the plot, because there may be too few points on the plot to make reliable estimates.

Despite some minor difficulties, Pox plots are highly useful and give a fair and square view of the self-similar nature of the underlying time series and about the degree of self-similarity for large samples but unreliable for empirical records with small sample sizes. Pox plot has a nice property that its asymptotic behavior remains unaffected by long-tailed marginal distributions. That is, if time series has a long-tailed marginal distribution, the R/S statistics still reflect the independence in that the asymptotic slope in the pox plot remains to be ½.

### *Estimation of H for Nile river data set by pox plot*

For Nile River minima, there are 663 observations but the plot has been produced with n=660 for the convenience of computation. Pox plot is given in Fig. 3 choosing k = 10p where p = (1,2,..20), t = 60m + 1, with m = (1,2,..). Fig. 3 shows that, for increasing k, the value of R/S statistics

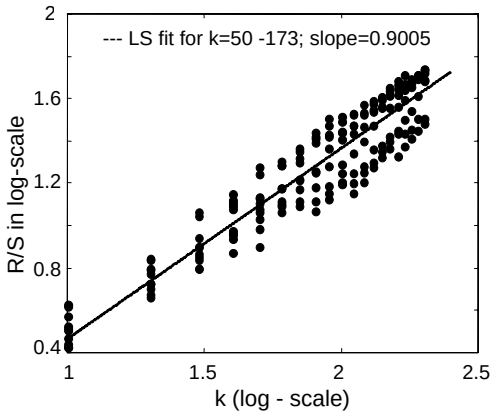


Fig. 3: Pox plot of R/S for Nile River

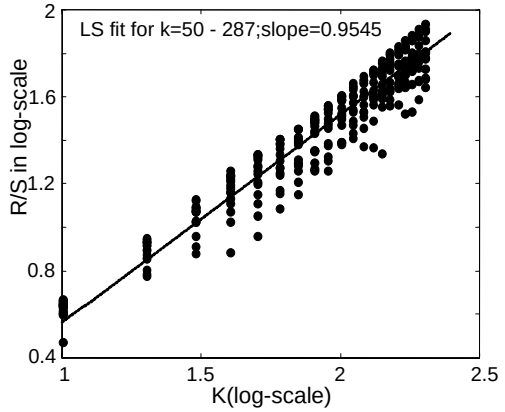


Fig. 4: Pox plot of R/S for VBR data set.

are scattered around a least square fitted line with a slope of approximately equal to 0.90. Therefore estimated H for Nile River is 0.90. This H which belongs to ½<H<1 clearly indicates that Nile River data has a very strong long-range dependence as was anticipated from Fig. 1.

***Estimation of H for VBR data set by pox plot***

VBR data set has 1000 observations, therefore, n = 1000. Taking k = 10p where p = (1,2,..20), t = 60 m + 1, with m = (1,2,..), Fig. 4 gives the Pox plot for VBR data. The estimated H for VBR data set found from this figure is 0.95. This value of H once again implies that VBR data has a very strong long-range dependence as was guessed from the Fig. 2 (ACF Plot). From pox plots of R/S, it is clear that both Nile and VBR data set has long-range dependence. To confirm it further and for the comparison purposes of H values, the dependence structure is once again checked with the help of other techniques starting with Variance-time plot followed by Periodogram and Correlogram methods.

***Variance-time plot***

One of the salient features of self-similar process is that sample mean decreases more slowly than the reciprocal of the sample size. Variance of the sample mean is given by $\mathrm{Var}\left(\overline{X_n}\right) \approx cn^{2H-2}, (c>0)$. For estimating H, calculate sample means $\overline{X}_1$, $\overline{X}_2$,.....$\overline{X}_{mk}$ for $m_k$ sub series of length k with the integer time lag k lying in $2 \le k \le n/2$. Overall mean can be given by,

$$\overline{X}(k) = \frac{1}{m_k}\sum_{i=1}^{m_k} \overline{X_i}(k) \qquad (7)$$

Now sample variance, $s^2(k)$ of the sample means $\overline{X_i}(k)$, i = 1, 2, ..$m_k$ can be calculated as[13-14],

$$s^2(k) = (m_k - 1)^{-1} \sum_{j=1}^{m_k} \left(\overline{X_j}(k) - \overline{X}(k)\right)^2 \qquad (8)$$

The variance-time plots are obtained by plotting log $s^2(k)$ against log k and by fitting a simple least square line through the resulting points in the plane ignoring the small values of k. For large values of k, the points in the plot are expected to be scattered around a straight line with a negative slope 2H-2. For short-range dependence or independence among the observations, the slope is = -1. Values of the estimated asymptotic slope between –1 and 0 (zero) suggest self-similarity, and an estimate for the degree of self-similarity is given by H = 1 + ½(slope). It is demonstrated here that with sample sizes of the magnitude of the Nile river data set or VBR data set, variance-time plots give reasonably accurate picture about the self-similar nature of the underlying time series and about the degree of self-similarity.

***Estimating H for Nile data by Variance-time plot:*** Variance time plot of Nile river minima is shown in Fig. 5, where the points in the plot are scattered around a straight line with a slope of –0.28, which clearly indicate the presence of long–term dependence among the observation of Nile River. The estimated value of H for Nile is H = 1 +(-0.28/2) = .86 approximately.

***Estimating H of VBR data by Variance-time plot:*** Fig. 6 gives the Variance time plot of VBR data set and suggests that variance of sample means converges to zero at a slower rate than $K^{-1}$ and the points in the plot are scattered around a straight line with a slope of –0.29 indicating the presence of long–term dependence among the observation. In this case, the estimated value of H is 0.855.

***Periodogram method***

In the frequency domain, analysis of time series is merely the analysis of a stationary process by means of its spectral representation. From the modifications of Herglotz' theorem first by Blomfield followed by Brockwell & Davis [14], the periodogram can be given by

$$I_N(\lambda) = \frac{1}{2\pi N}\left|\sum_{j=0}^{N-1} X_j e^{j\lambda}\right|^2 \qquad (9)$$

where λ is the Fourier frequency, N is the number of terms in the time series and $X_j$ is the data of the given series. To estimate H, first, one has to calculate this periodogram. Since $I_N(\lambda)$ is an estimator of the spectral density, a series with long-range dependence should have a periodogram, which is proportional to $|\lambda|^{1-2H}$ close to the origin. Then a regression of the logarithm of the periodogram on the logarithm of the frequency λ should give a coefficient of 1-2H. The slope of the fitted straight line is the estimate of 1-2H.

***Estimating H of Nile by Periodogram method:*** The points in Fig. 7 are randomly scattered around a LS fitted line that has a slope of –0.095 indicating the presence of the long memory in the data of Nile river. However, estimating H based on this straight line may not be wise because the points don't necessarily imply that the trend should be well represented by a straight line.

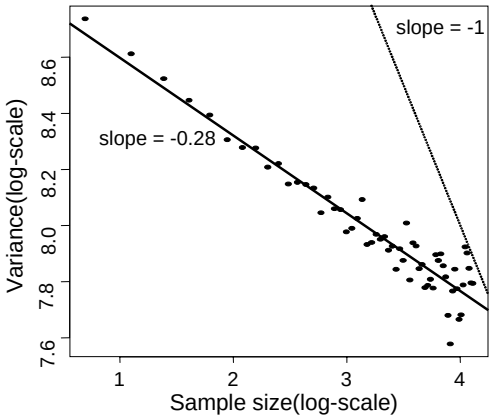


Fig. 5: Variance time plot of Nile river minima along with a LSE fitted straight line.

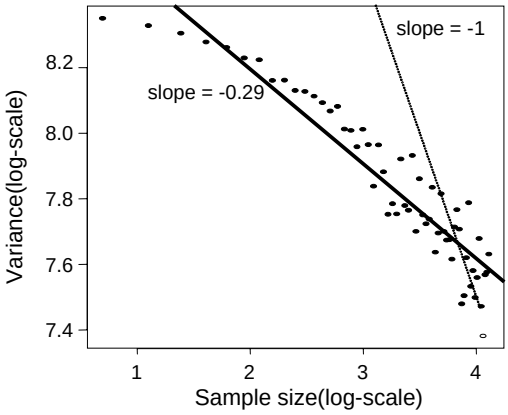


Fig. 6: Variance time plot of VBR data set along with a simple LES line

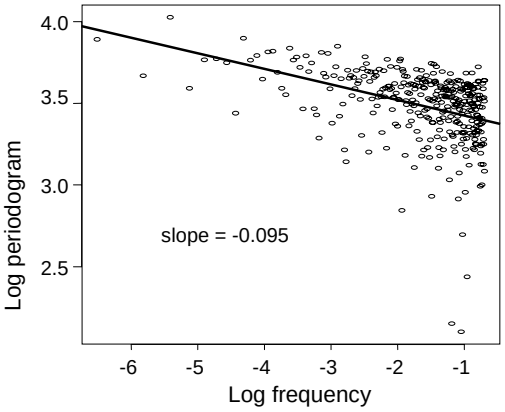


Fig. 7: Periodogram of Nile River minima

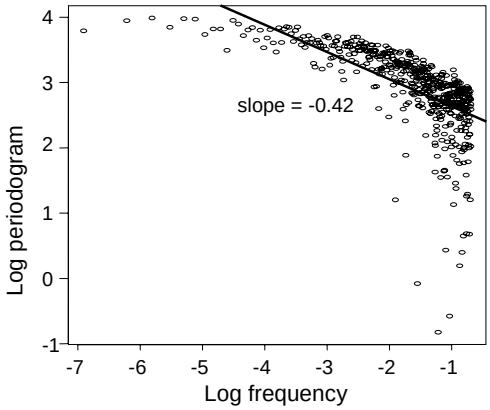


Fig. 8: Periodogram of VBR data.

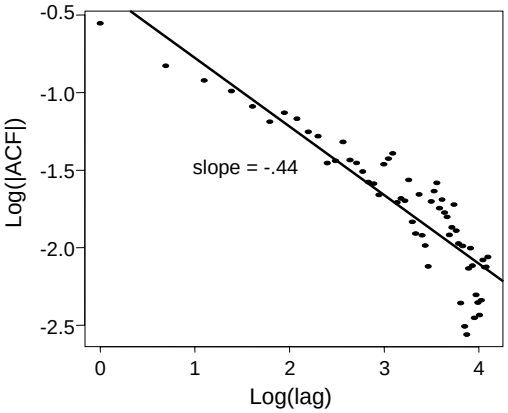


Fig. 9: The Correlogram of Nile data set.

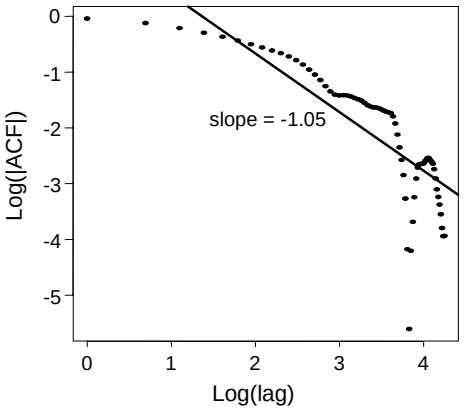


Fig. 10: Correlogram of VBR data set.

***Estimating H of VBR by Periodogram method:*** From Fig. 8 it is evident that the points are randomly scattered. It is neither fair nor wise to estimate H by fitting a line because the points don't seem to follow the line adequately. However, a negative slope of –0.42 indicates the presence of strong long-range dependence in the VBR data.

***Correlogram method***

In time series analysis, plot of ACF (autocorrelation function) is known as correlogram where the estimated correlation can be given in terms of auto-covariance function $\gamma(k)$ as[14]

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)} \tag{10}$$

It has already been observed that slow decay of correlation, which is proportional to $k^{2H-2}$ for ½<H<1 indicates the long-memory process. Therefore, the plot of the sample autocorrelation should exhibit this property. A much better plot for the handling of long-range dependence is the plot of ACF in logarithmic scale. If the asymptotic decay of the correlation is hyperbolic, then the points in the plot should be approximately scattered around a straight line with a negative slope of 2H-2 for the long memory processes but for short memory, the points should tend to diverse to minus infinity at an exponential rate. If the time series is long enough or if the series has strong long-range dependence, then this log-log correlogram is useful. Correlogram is useful as a preliminary heuristic approach to the data. Some pitfalls of sample correlation which are less known can be found in Mandelbrot[3,7]. Even though it is neither widely used nor attractive method for estimation, still H, the self-similarity parameter, can be estimated by this method deriving an equation of the form $\hat{\rho}(k) = \hat{H}(2\hat{H}-1)k^{2\hat{H}-2}$.

***Estimating H of Nile by Correlogram method:*** ACF plot in logarithmic scale in Fig. 9 suggests a slow decay of correlation. Points in the plot are scattered around a LSE (Least Squared Error) line with a slope of –0.44 once again confirming the long-range dependence structure of Nile river. Estimated H (equal to 0.78) is required to analyze carefully as there are large number of points lying a bit far from the fitted line.

***Estimating H of VBR data by Correlogram:*** In Fig. 10, the ACF plot in log-scale is not showing any clear linear trend, rather suggesting some curvilinear nature that may be an indication of the second order self-similarity. It would be unfair to estimate H by this method especially for VBR data set.

**CONCLUSIONS**

A number of heuristic methods to estimate self-similarity parameter H have been presented and applied to two different kinds of data set (Nile and VBR) and observed that pox plot of R/S and variance-time plot give better estimates. It may be mentioned that, these are heuristic graphical methods and there is no guarantee that all methods should work for every data set with the same sort of accuracy. For Nile River data set, it has been observed that the estimated value of H vary for different methods. For Nile, estimated H are 0.90 from R/S statistics, 0.86 from variance-time plot and 0.78 from the correlogram while periodogram method failed to produce any reasonable estimate. Variation in H values may be brought about due to the fact that no cut-off points were set while fitting a straight line.

For VBR data set, H = 0.95 from R/S statistics and 0.86 from variance-time plot while two other methods were not simply good enough to provide any realistic estimate. Analysis of this H values suggests that R/S statistics and Variance-time plots give reliable estimate than other two and can be used for estimating self-similarity parameter H. In some cases, periodogram and correlogram can serve the purpose of estimation as correlogram did for Nile River but failed with VBR data set. So, while estimating H using these heuristic methods, one has to be very careful and analytic about the derived value of H as well as the nature of the data set.

**RFERENCES**


1. L. Delbeke, "*Wavelet based estimators for the Hurst parameter of a self-similar process,"* Ph. D. thesis, Department of Mathematics, KULeuven, Belgium (1997).
2. H. P. Graf, "*Long-range correlations and estimation of the self-similarity parameter*," Ph. D. thesis, Swiss federal institute of technology, Zurich (1983).
3. J. Beran, M. S. Taqqu and W. Willinger, "*Long–range dependence in variable bit rate traffic,"* IEEE Trans. on Communications, Vol. 43, pp.1566-1579 (1995).
4. Mandelbrot, B. B. and Van Ness J. W., *"Fractional Brownian motions, fractional noises and applications"*, SIAM Rev. Vol. 10 pp. 422-437 (1968).
5. Mandelbrot, B. B. and Wallis, J. R., "*Noah, Joseph and operational hydrology*". Water Resource Research Vol. 4 pp.909-918 (1968).
6. Mandelbrot, B. B. and Wallis, J. R. "*Computer experiments with fractional Gaussian noises"*. Water Resource Research Vol. 5, pp.228-267 (1969).
7. J. Beran, "*Statistics for long-Memory Processes",* Chapman & Hall (1994).
8. W. Leland, M. Taqqu and W. Willinger, D. Wilson, "*On the self - similar nature of Ethernet traffic (extended version)*," IEEE/ACM Transactions on Networking, Vol. 2, pp. 1-15 (1994).
9. H. P. McKean, "*Stochastic Integrals"*, Academic Press, New York (1969).
10. M. Maejima, "*Self-similar processes and limit theorems"*, Sugaku Expositions, Vol. 2, pp. 103-123 (1989).
11. M. S. Taqqu, V. Teverosky and W. Willinger, "*Estimators for long-range dependence: an empirical study,"* Fractals, Vol. 3, No. 4, pp.785-788 (1995).
12. G. Samorodnitsky and M. S. Taqqu, "S*table non-Gaussian processes, stochastic models with infinite variance",* Chapman & Hall, NY (1994).
13. H. C. Tijms, "*Stochastic models, an algorithmic approach",* J. Wiley (1994).
14. P. J. Brockwell and R. A. Davis, "*An introduction to time series and forecasting",* Springer - Verlag, New York (1996).