Hurst Parameter Estimation Techniques: A Critical Review

Hae-Duck J. Jeong, Don McNickle[†] and Krzysztof Pawlikowski
Department of Computer Science and [†]Department of Management
University of Canterbury, New Zealand
E-mail: {{joshua, krys}@cosc, [†]dcm@mang}.canterbury.ac.nz

Abstract

Many recent studies of real teletraffic data in computer networks have shown evidence of self-similarity. The Hurst parameter H plays an important role in characterising self-similar processes. Thus, the estimation of the Hurst parameter of a sequence with a finite number of values is crucial in determining whether a process is self-similar. Parameter estimation techniques of the Hurst parameter have received great attention in recent years. In this paper a comparative analysis of the most frequently used H estimation techniques, the wavelet-based H estimator and Whittle's MLE estimator, periodogram, R/S-statistic, variance-time and IDC estimation techniques, has been reported. Our results reveal that the wavelet-based H estimator is the least biased of the H estimation techniques considered.

Keywords: Hurst parameter estimation techniques, Self-similar processes, Self-similar generators

1 Introduction

Self-similar processes have been found to be relevant in a range of areas of scientific activity such as climatology, economics, environmental sciences, geology, geophysics, hydrology, telecommunications and computer science. Historically, the importance of self-similar processes lies in the fact that they provide an elegant explanation and interpretation of an empirical law. Namely, for a given sequence of random variables $\{X_1, X_2, \cdots, \}$, one can consider the *rescaled adjusted range* $\frac{R(t,m)}{S(t,m)}$ (or *R/S-statistic*), with

$$\begin{aligned} R(t,m) &= \max_i [N_{t+i} - N_t - \frac{i}{m}(N_{t+m} - N_t), 0 \le i \le m] - \\ &\quad \min_i [N_{t+i} - N_t - \frac{i}{m}(N_{t+m} - N_t), 0 \le i \le m], \end{aligned}$$

where t is the time, m is the batch size and $N_t = \sum_{i=1}^{t} X_i$; and

$$S(t,m) = \sqrt{m^{-1} \sum_{i=t+1}^{t+m} (X_i - \bar{X}_{t,m})^2},$$

where $\bar{X}_{t,m} = m^{-1} \sum_{i=t+1}^{t+m} X_i$.

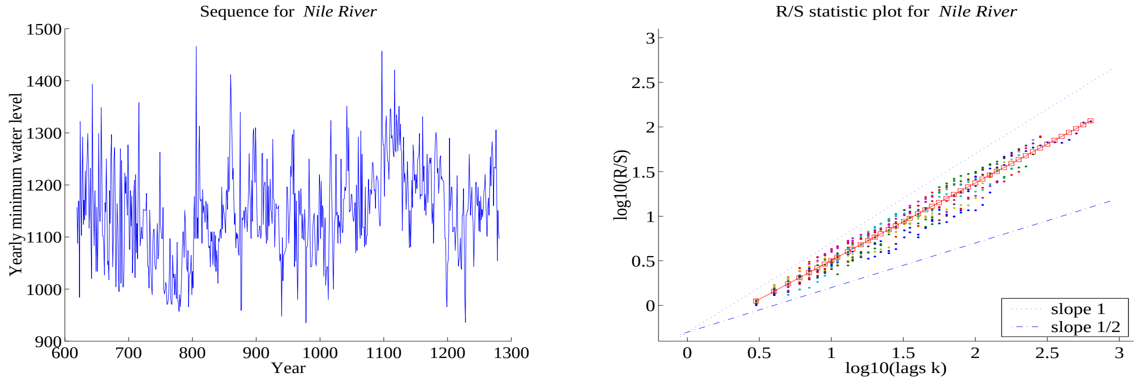
Hurst found empirically that for many time series observed in nature, the expected value of $\frac{R(t,m)}{S(t,m)}$ asymptotically satisfies the power-law relation:

$$E[\frac{R(t,m)}{S(t,m)}] \to cm^H, \text{ as } m \to \infty, \text{ with } 0.5 < H < 1,$$

where c is a finite positive constant [2]. This empirical finding was in contradiction to results for Markovian and related processes. For a stationary process with SRD, $E[\frac{R(t,m)}{S(t,m)}]$ behaves asymptotically like a constant times $m^{0.5}$. Hurst's finding that for the Nile River data, and for many other hydrological, geophysical, and climatological data, $E[\frac{R(t,m)}{S(t,m)}]$ behaves like a constant times m^H for $0.5 < H$, is known as the *Hurst effect*. Mandelbrot and Wallis [9] showed that the Hurst effect can be modelled by FGN with the self-similarity parameter $0.5 < H < 1$. For example, Figure 1 shows the yearly minimal water levels of the Nile River for the years 622-1281, measured at the Roda Gauge near Cairo [2]. The presence of the self-similar behaviour is evident. Since in this case the Hurst parameter from the R/S-statistic analysis (see Figure 1 (b)) is above 0.8690.

Many recent studies of real teletraffic data in modern computer networks have shown that teletraffic exhibits *self-similar* (or *fractal*) properties over a wide range of time scales. The properties of self-similar teletraffic are very different from the traditional models based on Poisson, Markov-modulated Poisson, and related processes, and the use of traditional models in networks characterised by self-similar processes can lead to incorrect conclusions about the performance of analysed networks. These include serious underestimations of the performance of computer networks, insufficient allocation of communication and data processing resources, and difficulties in ensuring the quality of service expected by network users.

The Hurst parameter H plays an important role in characterising self-similar processes. The estimation of the Hurst parameter of a sequence with a finite number of values is crucial in determining whether a process is self-similar. Most Hurst parameter estimation techniques are based on the idea of estimating the slope of a linear fit in a log-log plot; for detailed discussion, see [6]. For example, the R/S-statistic estimation technique is a well-known example of this approach, but has poor



(a) Sequence.

(b) R/S-statistic analysis.

Figure 1: Yearly minimum water levels of the Nile River at the Roda Gauge for the years 622-1281 (a) and their R/S-statistic analysis (b).

statistical performance; notably a high bias when the value of the Hurst parameter ($0.5 < H < 1$) is small or large. Another example is the periodogram plot based on a linear fit in a $log_{10}(P(\lambda))$ against $log_{10}(\lambda)$ plot, where $P(\cdot)$ is the periodogram and λ is frequency.

A comparative analysis of the most frequently used H estimation techniques, the wavelet-based H estimator and Whittle's MLE estimator, periodogram, R/S-statistic, variance-time and IDC estimation techniques, has been done. Wavelet-based and Whittle's estimators are asymptotically unbiased and efficient in theory, at least in the FGN case [1]. Assuming that this is the case for a Gaussian process in general, the second-order statistics (i.e., variance and ACF) of the wavelet-based and Whittle's estimators would be asymptotically equivalent under Gaussian assumptions.

To review the most commonly used estimation techniques of the Hurst parameter considered, we use many pseudo-random self-similar sequences generated by (i) a method based on the *fractional-autoregressive moving average* (F-ARIMA) process [5]; and (ii) a method based on *fractional Gaussian noise and Daubechies wavelets* (FGN-DW) [7]; for more detailed discussion, see [6].

2 Analytical Tools for Traffic Estimation

The accuracy of the most commonly used parameter estimation techniques of H are analysed empirically by applying them to output sequences generated by our generators. These techniques are as follows:

- *Wavelet-based H Estimator*: A wavelet-based H estimator is used to perform a thorough analysis of LRD in teletraffic sequences. The wavelet-based plot is obtained by plotting $log_2(2^i)$ against $log_2(1/n_i \sum_j |d_x(i,j)|^2)$ to detect LRD, the determination of the range of scales over which the power-law behaviour holds [1].

- *Whittle's approximate maximum likelihood estimate (MLE)*: Whittle's MLE is used for a more refined data analysis to obtain confidence levels for the Hurst parameter H [2]. It examines the properties in the frequency domain, while the R/S-statistic plot and variance-time plot focus on the time domain.

- *Periodogram plot*: The periodogram plot is used to show whether a generated sequence represents an LRD process or not. If the autocorrelations are summable, then, near the origin in the frequency domain, the periodogram should be scattered randomly around a constant level. If the autocorrelations are non-summable, i.e., LRD-type, the points of a sequence are scattered around a straight line with negative slope. The periodogram plot is obtained by plotting $log_{10}(periodogram)$ against $log_{10}(frequency)$. An estimate of the Hurst parameter is given by $\hat{H} = (1 - \hat{\beta}_1)/2$, where $\hat{\beta}_1$ is the slope [2].

- *R/S-statistic plot*: The Hurst parameter H can be estimated from empirical data using an R/S-statistic plot. An estimate of H is given by the asymptotic slope $\hat{\beta}_2$ [2], i.e., $\hat{H} = \hat{\beta}_2$.

- *Variance-time plot*: The variance-time plot is obtained by plotting $log_{10}(m)$ against $log_{10}(Var(X^{(m)}))$ and by fitting a simple least square line through the resulting points in the plane. An estimate of the Hurst parameter is given by $\hat{H} = 1 - \hat{\beta}_3/2$, where $\hat{\beta}_3$ is the slope of the line [2].

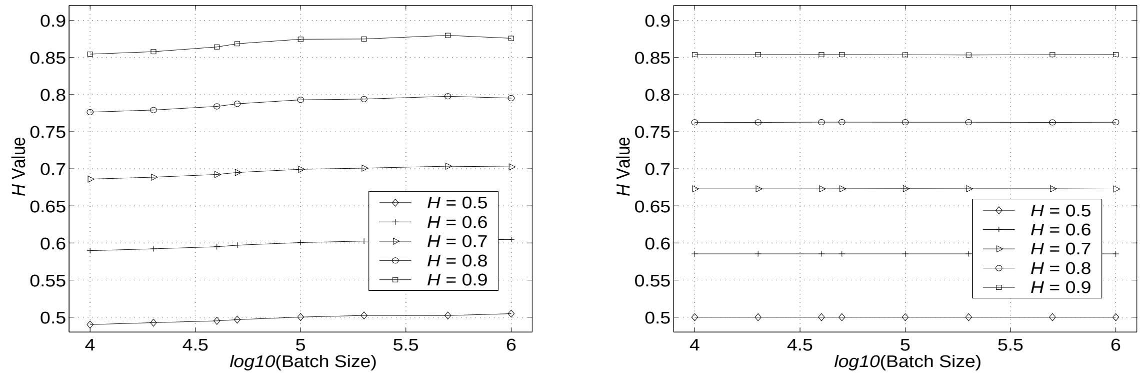
- *Index of dispersion for counts (IDC)*: The IDC captures the variability of traffic over different time scales for a count process. For a given time interval of length t, the IDC is given by the variance of the number of arrivals $\{X_t\}$ during the interval of length t divided by the expected value of the same quantity. Plotting $log_{10}(IDC(t))$ against $log_{10}(t)$ results in an asymptotic straight line with slope $2H$ - 1.

3 Finding Appropriate Sample Sizes

The estimate of the Hurst parameter H of a sequence should ideally be calculated from an infinite number of values. However, this is impossible.

Different studies of self-similar sequences have been based on different sample sizes. Mandelbrot and Wallis [9] used a sample of 9,000 numbers to measure the value of the Hurst parameter using the R/S-statistic. Leland et al. [8], [13] analysed the value of the Hurst parameter with sequences of 360,000 observations, where each observation represents the number of bytes sent over the Ethernet per 10 milliseconds. Taqqu et al. [12] used a sample size with 10,000 numbers, generated 50 different sequences for each of several values of H and compared the estimated values of H with the required ones. Garrett and Willinger [4] presented a statistical analysis of a two-hour long empirical sample of VBR video with 171,000 frames. Paxson [10] used ten samples of 32,768 numbers, to obtain estimates of the Hurst parameter. He confirmed that the stochastic dependence present in the generator was consistent with having the required value of H using Whittle's MLE. Rose [11] studied traffic modelling of VBR MPEG video and its impacts on ATM networks with different sequences of 40,000 frames of MPEG traffic data (it takes 30 minutes to obtain each sequence). An R/S-statistic estimation technique and Whittle's MLE estimator were used to estimate the Hurst parameter for MPEG traffic in [11]. Abry and Veitch [1] compared Whittle's MLE and the wavelet-based estimators using small synthetic sample sequences of 4,096 numbers and the real Ethernet LAN data set. They showed that the wavelet-based estimator is less biased than Whittle's MLE under Gaussian assumptions. A minimisation procedure is involved in the Whittle estimator which requires many repetitive calculations, leading to a significantly higher overall cost, while the wavelet-based estimator requires the simple calculation of a discrete wavelet transform, which can be done in $O(n)$ operations.

We evaluated the most commonly used methods for estimating the self-similarity parameter H to find an unbiased or the least biased estimation technique to employ in a comparative analysis of self-similar sequences produced by two generators. First, we investigated new long of sequences need to be, and then analysed parameter estimation techniques using appropriately long synthetic sequences generated by pseudo-random self-similar generators. For each H = 0.5, 0.6, 0.7, 0.8 and 0.9, each sequence with one million numbers is divided into sub-sequences (i.e., batch size) m of 10,000, 20,000, ..., 100,000, 200,000, ..., 1,000,000 numbers. For example, let a sequence be $(x_1, \ldots, x_n)$ and then divide the sequence into i sub-sequences $(x_1, \ldots, x_l)$, $(x_{l+1}, \ldots, x_{2l})$, ..., $(x_{(i-1)l+1}, \ldots, x_{il})$, $l > 0$, $i = [n/l]$. Each estimate $\hat{H}_j, j = 1, \ldots, i$ is obtained using parameter estimation techniques of H, and each of the mean estimates $\hat{H}$ and the variance estimates are given by $\hat{H} = 1/i \sum \hat{H}_j, j =$



(a) Wavelet-based H estimator.

(b) Whittle's MLE estimator.

Figure 2: For $H = 0.5$, 0.6, 0.7, 0.8 and 0.9, estimated $\hat{H}$ of H obtained from the FGN-DW method using the wavelet-based H estimator and the Whittle's MLE estimator as the batch size increases from 10,000 to 1,000,000.

$1, \ldots, i$, and $1/(i(i-1)) \sum(\hat{H}_j - \hat{H})^2$, respectively [3].

In Figure 2 for $H = 0.5$, 0.6, 0.7, 0.8 and 0.9, *log10* (Batch Size) is plotted against the estimated $\hat{H}$ of H obtained from the FGN-DW method using the wavelet-based H and Whittle's MLE estimators. We chose these estimators as they give more refined measurements than other estimation techniques [1], [8], [12]. Figure 2 (a) shows that all curves of the estimated H using the wavelet-based H estimator slowly converge toward the true values. The bottom-most three curves match the true values at $m = 100,000$. For $H = 0.8$, the curve of the estimated H matches the true value at $m = 500,000$, and for $H = 0.9$, the curve of the estimated H is the closest to the true value at $m = 500,000$. For all H values, the highest estimates of H show at $m = 500,000$, but all curves of estimates of H except for $H = 0.9$ are higher than the true values. Then these curves slowly decrease until $m = 1,000,000$. Figure 2 (b) also shows that all curves of the estimated H using the Whittle's MLE estimator converge to true values. Thus, appropriate numbers of a long sequence for analysing parameter estimation techniques for the Hurst parameter are recommended to be between 30,000 and 500,000. We have chosen as the sequence length 32,768 (2^{15}) for our review of these techniques.

4 Empirical Results

For $H = 0.5$, 0.6, 0.7, 0.8 and 0.9, each H estimate was applied in 30 sample sequences of 32,768 numbers, generated by F-ARIMA and FGN-DW self-similar generators to obtain its mean bias.

- F-ARIMA Method: In the F-ARIMA method, the wavelet-based H estimator is the most accurate (Figure 3). For $0.5 \leq H \leq 0.8$, the estimates of the wavelet-based H estimator match the true values, but when $H = 0.9$, it does not. For $H < 0.73$, the R/S-statistic H estimate is positively biased; for $0.73 < H$, it is negatively biased. The other H estimation techniques are negatively biased.

- FGN-DW Method: The wavelet-based H estimator gives the best result for the FGN-DW method (Figure 4). For $H \leq 0.8$, the estimates match the true values well, although when $H = 0.9$, it does not. The R/S-statistic estimate produces the same results as in the above method. The other estimation techniques are negatively biased.

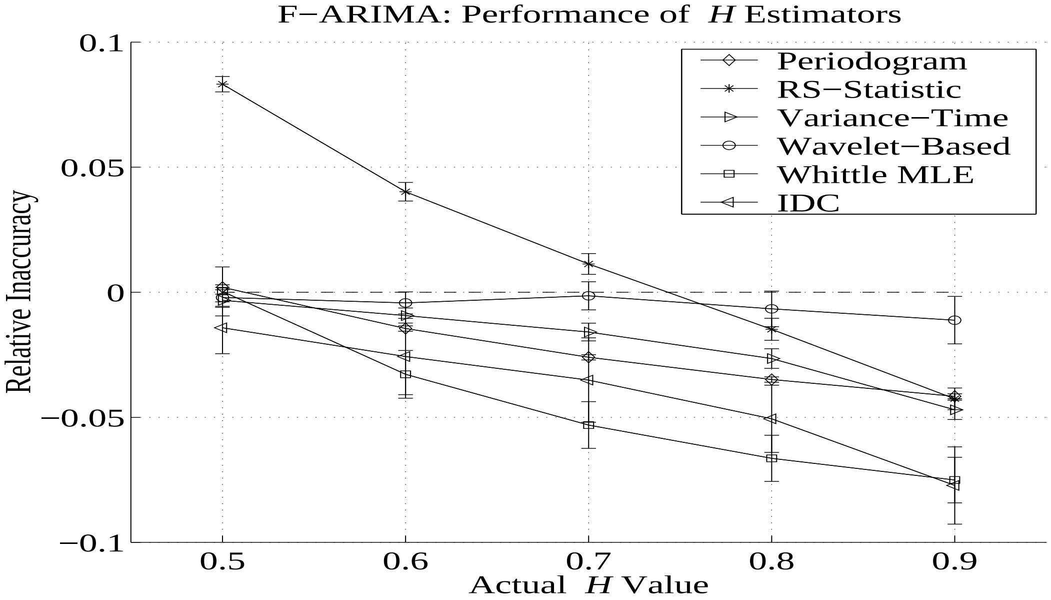


Figure 3: Bias performance of H estimation techniques for the F-ARIMA method.

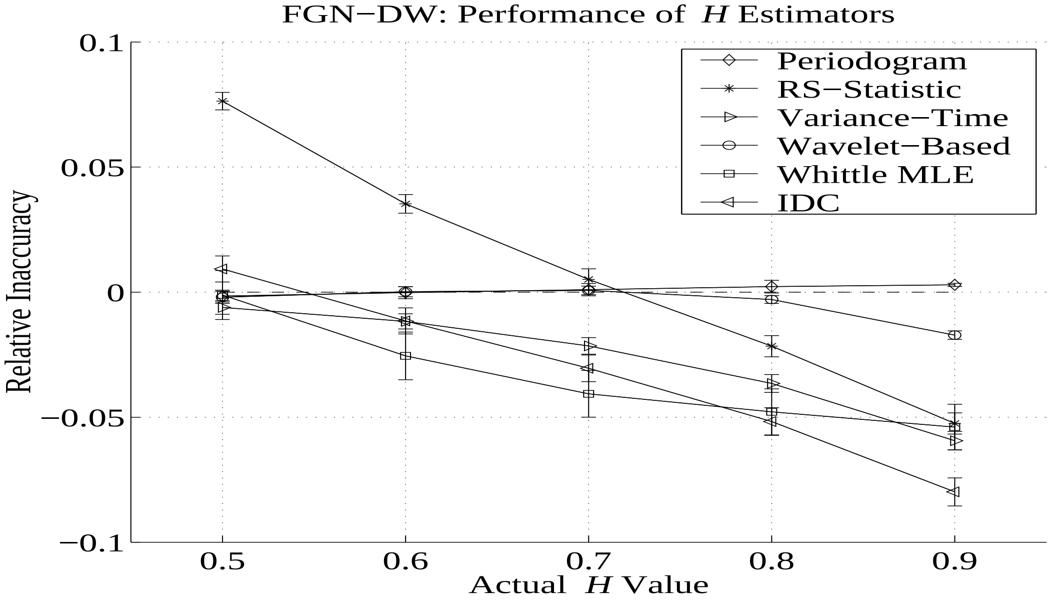


Figure 4: Bias performance of H estimation techniques for the FGN-DW method.

5 Conclusions

Our results have confirmed that the wavelet-based H estimator is the least biased of the H estimation techniques considered, supporting Abry and Veitch's results [1]. While the bias of the R/S-statistic and the periodogram estimation techniques change from positive to negative as the H value increases, the variance-time estimation technique, IDC estimation technique, wavelet-based H and Whittle's MLE estimators are steadily negatively biased. Even though they are biased, the best estimator of the self-similarity parameter H is the wavelet-based H estimator. Furthermore, the wavelet-based H estimator and the periodogram estimation technique are much faster than Whittle's MLE estimator, the R/S-statistic, variance-time and IDC estimation techniques.

References

[1] P. Abry and D. Veitch. Wavelet Analysis of Long-Range-Dependent Traffic. *IEEE Transactions on Information Theory*, 44(1):2–15, 1998.

[2] J. Beran. *Statistics for Long-Memory Processes.* Chapman and Hall, New York, 1994.

[3] J. Beran and N. Terrin. Estimation of the Long-Memory Parameter, Based on a Multivariate Central Limit Theorem. *Time Series Analysis*, 15(3):269–278, 1994.

[4] M.W. Garrett and W. Willinger. Analysis, Modeling and Generation of Self-Similar VBR Video Traffic. In *Proceedings of the ACM SIGCOMM'94*, volume 24 (4), pages 269–280, London, UK, 1994.

[5] J.R.M. Hosking. Modeling Persistence in Hydrological Time Series Using Fractional Differencing. *Water Resources Research*, 20(12):1898–1908, 1984.

[6] H.-D.J. Jeong. *Modelling of Self-Similar Teletraffic for Simulation.* PhD thesis, Department of Computer Science, University of Canterbury, 2001. Submitted.

[7] H.-D.J. Jeong, D. McNickle, and K. Pawlikowski. Fast Self-Similar Teletraffic Generation Based on FGN and Wavelets. In *Proceedings of the IEEE International Conference on Networks, ICON'99*, pages 75–82, Brisbane, Australia, 1999.

[8] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE ACM Transactions on Networking*, 2(1):1–15, 1994.

[9] B.B. Mandelbrot and J.R. Wallis. Computer Experiments with Fractional Gaussian Noises. *Water Resources Research*, 5(1):228–267, 1969.

[10] V. Paxson. Fast, Approximate Synthesis of Fractional Gaussian Noise for Generating Self-Similar Network Traffic. *Computer Communication Review, ACM SIGCOMM*, 27(5):5–18, 1997.

[11] O. Rose. *Traffic Modeling of Variable Bit Rate MPEG Video and Its Impacts on ATM Networks.* PhD thesis, Bayerische Julius-Maximilians-Universität Würzburg, 1997.

[12] M.S. Taqqu, V. Teverovsky, and W. Willinger. Estimators for Long-Range Dependence: an Empirical Study. *Fractals*, 3(4):785–788, 1995.

[13] W. Willinger, M.S. Taqqu, W.E. Leland, and D.V. Wilson. Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements. *Statistical Science*, 10(1):67–85, 1995.