What is the Required Series Length for Correct Self-similarity Analysis?

Julio C. Ramírez Pacheco[1,2] and Deni Torres Román[1]

[1] CINVESTAV Unidad Guadalajara, Av Científica 1145, 45010
Col. el Bajío, Zapopán, Jalisco, México
cramirez@gdl.cinvestav.mx, dtorres@gdl.cinvestav.mx
[2] Universidad del Caribe, SM 78, Manzana 1, lote 1, 77528, Cancún, QRoo, México
jramirez@ucaribe.edu.mx

Abstract. It is well-known that *self-similar* and *Long-memory* signals appear in many fields of science. LAN, VBR sources, WWW traces, wireless traffic, etc. are among the ones with this behaviour in computer networking. An important question in these applications is how long a measured trace should be to obtain reliable estimates of the *Hurst*-index. This paper addresses this question by first providing a thorough study of estimators for short series based on the behaviour of bias, σ, $\sqrt{\text{MSE}}$ and convergence when using *Gaussian H-sssi* signals. Results show that *Whittle*-type estimators behave the best when estimating H for short signals. Based on the results, empirically derived minimum trace length for the estimators is proposed. Finally for testing the results, the application of estimators to real traces is accomplished. Inmediate applications from this can be found in the real-time estimation of the *Hurst*-index which is useful in agent-based control of *QoS* parameters.

1 Introduction

Self-similar stochastic processes are the ones which present scale-invariant statistical behaviour [27] [12] [5]. These processes are widely applied as models for different types of phenomena in a wide range of fields of science [23] [21] [22] [14] [24]. In the computer networking area, these processes are used for the modelling of aggregate LAN, VBR video, wireless and WWW traffic among others [17] [6] [8] [20] [19]. In these studies, traffic was measured and then analyzed in order to find whether it fits the *self-similar* model or not. The traces used in these studies consisted of hundreds of thousands of points in the LAN case and nearly 100000 for each VBR video trace. Obtaining such lengths usually implies a long measurement time. For off-line study, the above lengths are acceptable while for applications of real-time administration of *QoS* metrics based on accurate *Hurst*-index estimation, the above are unacceptable. The paper first studies the behaviour of estimators to short time series and then adresses the problem of obtaining the minimum length required for accurate real-time estimations of the *Hurst*-index. The requirement is thus obtaining high accuracy with minimum length. The accuracy shoud be comparable with that of long series, where accuracy in this case, is based on metrics such as standard deviation,

© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 297-307

Received 08/03/09
Accepted 27/04/09
Final version 06/05/09

bias and $\sqrt{MSE}$. Convergence analysis is also a useful tool for accomplishing the above. Thus, the paper addresses the following issue: given specifications of bias, variance or MSE, what the series length N should be?, i.e., suppose stochastic process Ψ posseses *Hurst*-index H, find the minimum length N_{min} such that for every realization(of length N_{min}) the estimated *Hurst*-indexes $\hat{H}$ are similar to that of H. For accomplishing the above, the paper is organized as follows. Section 2 briefly summarizes *self-similar* stochastic processes and estimation methods. Section 3 provides description of the methodology used for finding the minimum value N while section 4 presents a detailed study of the behaviour of estimators to short time series, the problem of finding the minimum length for the estimators and the application of these results to real LAN traces. Finally section 5 concludes the paper.

2 Self-*similar* Signals and Estimation of *H*

Self-similar processes are the ones whose distributional properties are invariant to dilations in time and suitable compression of amplitude. Let $Z = \{Z_t\}_{t \in I}$, where $I = \mathbb{R}$ or $\mathbb{R}_+$, be a real-valued stochastic process, it is said that Z is ***self-similar iff*** there exist an $H \in \mathbb{R}$ such that for any $a \in \mathbb{R}_+$ the following holds $\{Z_{at}\}_{t \in I} \overset{d}{=} \{a^H Z_t\}_{t \in I}$, where $\overset{d}{=}$ is in the distributional sense. Usually, the interest is in *H-ss* processes with stationary increments for which the above holds with $H > 0$. The above definition is called the strict one. A relaxed version of the above is obtained by the second-order definition one which requires invariance on second-order statistical properties under scaling. Formally let Z_t be a continuous-time stochastic process, it is said that Z_t is second-order *self-similar* if $EZ_t = a^{-H} EZ_{at}$ and $R_{zz}(t, s) = a^{-2H} R_{zz}(at, as)$. Computer networking requires discrete-time models, then discrete versions of the above are needed. Let $X = \{X_t, \in \mathbb{Z}\}$ be a discrete-time process, possibly obtained by sampling a continuous time random signal, it is said that X is strictly ***self-similar iff*** there exists an $H \in (0, 1)$ such that for any $m \geq 1$ $X \overset{d}{=} m^{1-H} \Gamma_m(\{X\})$. $\Gamma_m(.)$ is the block aggregation process which receives as input a length N time series and outputs an length N/m time series. A relaxed version of strict discrete *self-similarity* is given by second-order *self-similarity* in the exact sense. Let X be a discrete-time stochastic process, it is said that X is exact second-order *self-similar* if its correlation coefficient satisfies

$$\rho(k) = \frac{1}{2}\{(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}\}. \tag{1}$$

A process having a correlation function of the above form also satisfies $\mathrm{Var}(X) = m^{2-2H}\mathrm{Var}(\Gamma_m(\{X\}))$ and $\mathrm{Cov}(\Gamma_m(X_t), \Gamma_m(X_{t+k})) = m^{2-2H}\mathrm{Cov}(X_t, X_{t+k})$. In computer networking an even more relaxed version of (1) is used. A process X is said to be asymptotically second-order *self-similar* if the correlation function of $\Gamma_m(\{X\})$ as $m \to \infty$ is equal to that of an exact second-order *self-similar* stochastic process in discrete-time i.e., equal to (1). If in the exact or asymptotic definition of discrete *self-similarity* we let $k \to \infty$, then, $\rho(k) = ck^{2-2H}$

which implies *long-range* dependency. It means an asymptotic or exact *self-similar* process is *long-range* dependent provided $H \in (0,1)$ and $k \rightarrow \infty$. The greater the parameter H the smoother the process is and the slower the decay to zero of the autocorrelations. Several methods of estimation of the parameter H have been proposed, the methods can be classified as time-domain, frequency-domain and time-scale methods. Among the time-domain methods it is found the R/S statistic [18] [13], variance-time plot(dispersional analysis), variance of residuals(DFA), absolute moment, MAVAR, Higuchi's method, scaled window variance [7], Whittle, etc [26] [25]. GPH, Periodogram and other modified periodograms methods are found in the frequency-domain class which in turn take advantage of the power law behaviour of the *self-similar* processes near the origin. Time-scale methods include wavelet based estimators such as Abry-Veitch estimator [2] [3] [1] [4] [28]. Software tools for *self-similarity* analysis are also important since they collect a number of estimators and methodologies for improving the anaysis of *self-similarity*. In this context, studies have demostrated that `fARMA` and `SelQoS` are the most robust and accurate ones while a widely used one, `Selfis` [16] [15], is inaccurate and not robust. This paper, makes use of the R package `fARMA` for obtaining the results of estimations. The choice of `fARMA` is based on the excellent programming capabilities of the `S` language.

3 Methodology

We first provide a detailed study of the behaviour of estimators to short series and then propose minimum lengths for the estimators. The behaviour of estimators is studied by applying a given estimator to N time series and then obtaining BIAS, σ and $\sqrt{\text{MSE}}$. Comparison of the behaviour of the estimators subject to these statistics and for varying length aids in finding the minimum length. Also a convergent analysis shows the evolution of estimators in time and will be useful in this paper. Next we describe these steps in more detail. In order to apply the estimators, N time series should be obtained. Synthetic signals with known H are obtained by the simulation of Gaussian *H-sssi(fGn)* series using the Davies and Harte method [9]. The considered lengths for the traces were $N = \{2^i, i = 6, 7, \ldots, 16\}$ and for each length 100 traces with *Hurst*-index $H \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$ were generated. Thus, a total of 6600 'exact' fractal signals were generated. For each set of estimations of a particular H, the following statistics are computed $\text{BIAS} = H_0 - \bar{X}$, where H_0 is the nominal value, standard deviation σ and $\sqrt{\text{MSE}} = N^{-1} \; {}_{i=1}^{N} (x_i - H_0)^2$. Based on the results of BIAS, σ and $\sqrt{\text{MSE}}$ a minimum length is proposed. The length N_{min} is obtained from accurate estimations(BIAS ~ 0.03,$\sigma \sim 0.015$). In addition, the classification of estimations based on these values(that of BIAS and σ) is also proposed. *High accuracy* estimations are obtained when $\text{BIAS} \leq 0.03$ and $\sigma \leq 0.015$, *acceptable* estimations when $\text{BIAS} \in (0.03, 0.05)$ and $\sigma \leq 0.02$ and *biased* estimations when $\text{BIAS} > 0.1$. Once the minimum length is obtained for *fGn*-type series, the application of these results is performed to long synthetic and real traces. For such series Z of length M, $M >> N_{min}$, the following is studied: let $t_0, t_1, \ldots, t_k$ be a sequence of points in the x-axis, where $t_{i+1} > t_i$ and $(t_{i+1} - t_i) < N_{min}$, to each block of Z of length N_{min}, $\{Z_j\}_{j=t_i}^{t_i+N_{min}-1}$, apply a *Hurst* estimation method $\Theta_{t_i}^{N_{min}}(.)$ to these blocks. Repeat until $t_k + N_{min} > M$ for any k. A plot of t_i versus $\Theta_{t_i}^{N_{min}}(.)$ should result in a signal with little variation(the variation should conform σ) if N_{min} is correctly set. The proposed length N_{min} is related to convergence analysis of a series, which is also studied. Convergence of any estimator is obtained by first partitioning the original series Z into blocks of size $m << M$ to obtain $Z = \{\Psi_1^m, \Psi_2^m, \ldots, \Psi_i^m\}$, where $\Psi_i^m = \{Z_{(i-1)m}, Z_{(i-m)m+1}, \ldots, Z_{im}\}$. Next apply a Hurst methodology $\Theta(.)$ to $\cup_{j=1}^{N/m}\{\Psi_i^m\}$, $j = 1, 2, \ldots, N/m$ to obtain $\Theta_1^m, \Theta_1^{2m}, \ldots, \Theta_1^{jm}$. Plot Θ_i^{jm} versus jm for $j = 1, 2, \ldots, N/m$ to visualize the convergent behaviour of estimator $\Theta(.)$.

4 Simulation Results

4.1 Perspective, Bias and σ Plots

Figure 1 shows a perspective plot of estimations of the *Hurst*-index for traces with $H = 0.90$ and for varying length when applying *wavelet*-based and Whittle techniques. Left plot corresponds to wavelet technique while right plot to Whittle. Note from the left plot that wavelet-based techniques experience high bias and variability when estimating the *Hurst*-index for short time series. The length of the traces for these highly variable estimations is in the order of $N < 2^{12}$. When $N \in (2^{11}, 2^{12})$, the estimations show aceptable results and only when $N > 2^{13}$, the estimations are highly accurate. Similar behaviour is obtained for traces with *Hurst*-index value different that $H = 0.90$. The same kind of plot

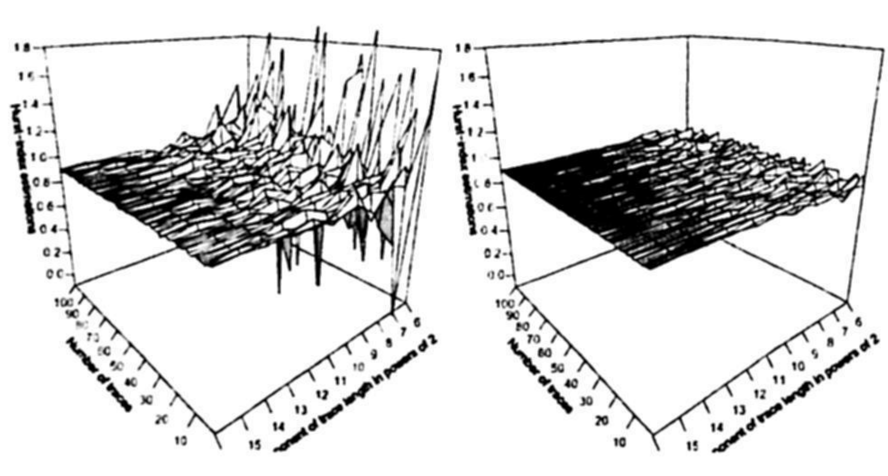

Fig. 1. Wavelet and Whittle Estimations for 100 fGn traces with $H = 0.90$, left plot corresponds to Wavelet estimations and the right to Whittle-type estimations.

when using *Whittle*-type estimations is shown in right plot of the figure for traces with $H = 0.90$. As can be seen from the figure, in general, Whittle-type estimations present high accuracy and low variability for short *long-range dependent* time series. For values of H below 0.90, the estimated *Hurst*-index present high variability when $N < 2^{10}$ and when $H \geq 0.90$, generally, the estimations are accurate for $N > 2^9$. Comparing the results of *Whittle*-type estimators with that of *wavelet*-based techniques is seen that the *Whittle* ones are more robust to short series in the context of exact simulated *fGn* traces. Figure 2 shows a perspective plot of estimations of the *Hurst*-index for traces with $H = 0.90$ and for varying length when applying periodogram and R/S Statistic method. Recall that periodogram method is based on the behaviour of *long-memory* series' PSD near origin. As seen in the left plot of the figure, the periodogram method behaves similarly to wavelet method for short series, i.e., with high bias and variability. A similar behaviour is obtained when using different values of the *Hurst*-index H. The right plot of figure also shows the same kind of plot as the Periodogram for R/S algorithm. Note from the plot that R/S method present high variablity no matter what the length of the trace is. The variablity diminishes when the lengths of the traces are longer but it is difficult to establish a minimum length for accurate estimations of H. In contrast to the other methods, the surface in the R/S perspective plot is always rough. The results then imply that additional analysis methods should be applied to R/S statistic method in order to deal with this biased behaviour. Finally, figure 3 shows the perspective plots for variance-type method. Variance-type method presents a similar behaviour to those of wavelet and periodogram methods. Unless wavelet and periodogram, the range of accurate estimations for variance-type in the *fGn* context is different. For accurate estimations, the length of the trace should be at least $N \geq 2^{15}$.

Figure 4 shows the bias experienced by every method studied for $H = 0.90$.

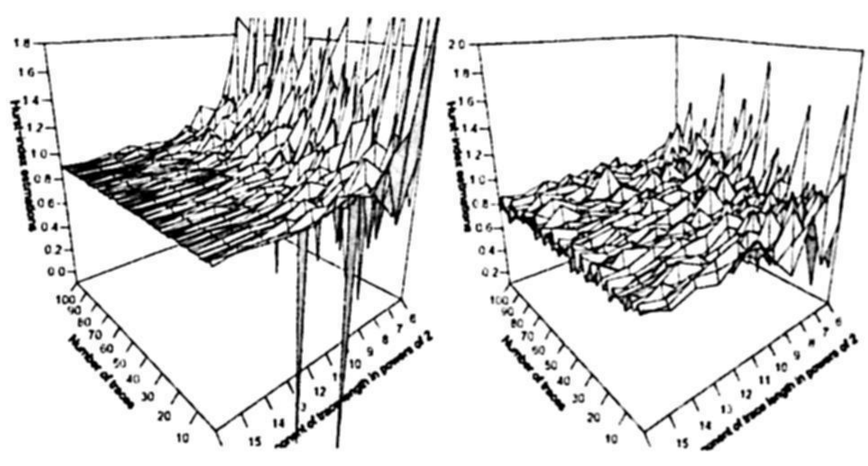

Fig. 2. Periodogram and R/S Statistic estimations for 100 fGn traces with $H = 0.90$, left plot corresponds to Periodogram estimations and the right to R/S Statistic.

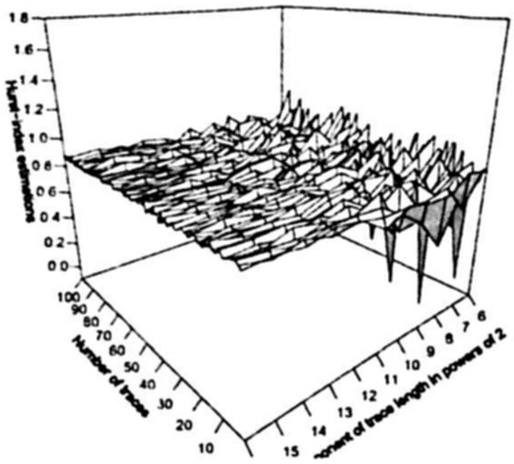

Fig. 3. Variance estimations for 100 fGn traces

Note from the figure that Whittle and Wavelet methods are the methods whose behaviour for short series is better than the others. Unless Wavelet, Whittle method behaves with less irregularity for short series and for $N \geq 2^8$, the biases are not significant. For wavelet methods, the bias is irregular for short series and stabilizes on $N \geq 2^{12}$. The other methods show irregular behaviour and high bias and unless $N \geq 2^{16}$, the estimations are not aceptable. R/S statistic method bias seems not to have stabilizing behaviour while periodogram and variance seems to stabilize for high N. Figure 5 illustrates the standard deviations for traces with $H = 0.90$ and varying length N. Note that Whittle-type method

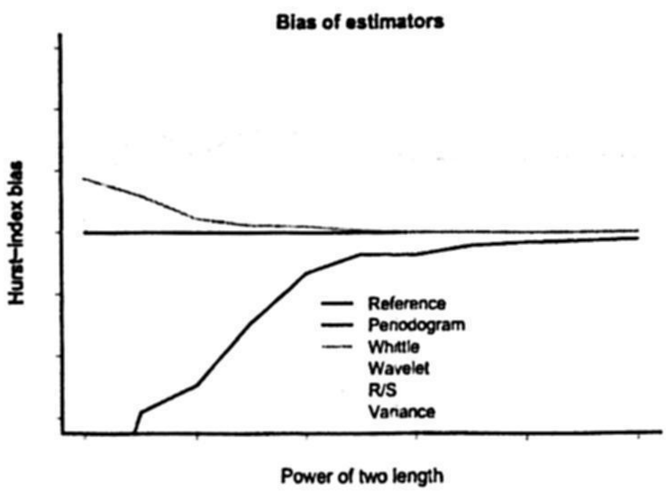

Fig. 4. Bias of all methods

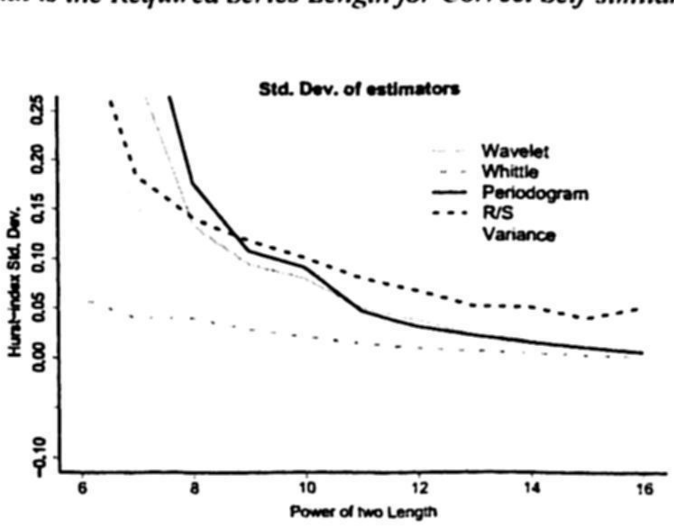

Fig. 5. Standard deviation of all methods

is the estimator which presents less variability and for $N \geq 2^8 = 256$ points, the estimations are accurate enough. Periodogram and Wavelet methods are the ones that follow in accuracy and the length required is the same when variability is taken into account. However, bias and standard deviation together indicate that the best estimator for short time series is Whittle method. Whittle method presents high accuracy when $N \geq 2^8$, Wavelet methods presents accuracy when $N \geq 2^{13}$, Periodogram acceptable estimations when $N \geq 2^{15}$ and Variance and R/S statistic method, biased estimations when $N \in (2^6, 2^{16})$.

4.2 Convergence of Estimators

For each series of length $M = 2^{16}$, a variation of the above defined convergent analysis is performed. The analysis is first applied to the first $\tau_0 = 64$ points of the series X_j, i.e., a *Hurst* estimation method $\Theta(.)$ is applied to the first τ_0 points of X_j. Then, we repeatedly apply $\Theta(.)$ to the next $\tau_0 + i\tau_u$ points of X_j, where $\tau_u = 200$ and $i = 1, 2, \ldots, k$ such that $\tau_0 + k\tau_u \leq M$. This analysis is done to 100 *fGn* series, thus obtaining the convergent behaviour of each. Once the convergent analysis is performed for each of the studied traces, the mean convergence analysis is performed. it means the mean for the 100 estimations of τ_0 and so on. The mean convergence plot $\Theta(.)$ versus $\tau_i, i = 0, 1, 2, \ldots, k$ is now an indicator of how well the estimators convergence to the theoretical *Hurst*-index value. This mean convergent behaviour was applied to the studied estimators in this paper. Figure 6 shows the mean convergent behaviour of the R/S-statistic method and variance-type method. Note from figure that R/S statistic stabilizes quickly but is biased by 0.05, thus it is difficult to propose a minimum length for this method. Unlike R/S statistic, variance-type method converges to the theoretical value($H = 0.9$ in this case) as the length increases. From the figure and the bias and σ plots it is inferred that the required length for this method should be at least 2^{16} points. Figure 7 shows the same kind of mean convergent plots for the other methods studied. Note that *Whittle*-type method stabilizes quickly

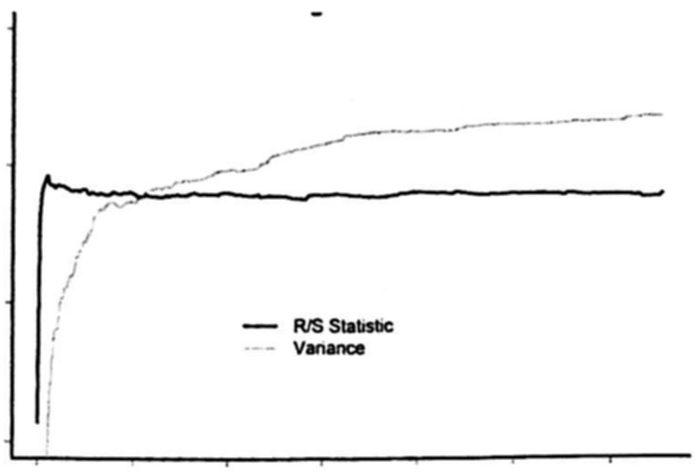

Fig. 6. Mean convergent behaviour

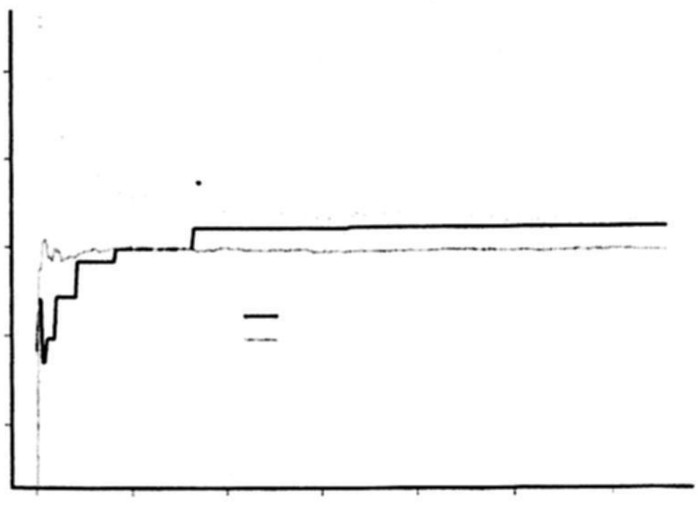

Fig. 7. Mean convergent behaviour

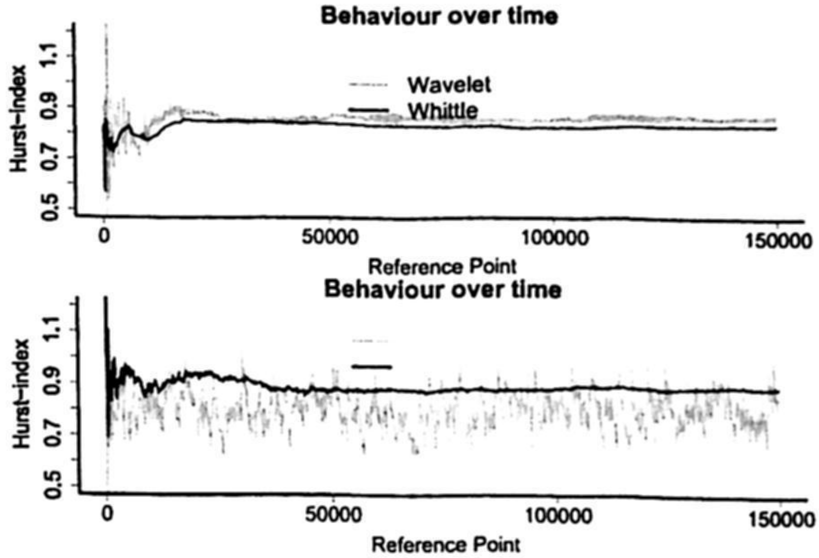

Fig. 8. LAN trace behaviour in time

with very low bias. The behaviour is similar to that of R/S statistic but unlike R/S statistic, the estimations in *Whittle*-type are not biased. The behaviour of Periodogram and wavelet method is similar. Periodogram, according to definition in section 3 is negatively biased, while wavelet is in principle positively biased, but in the long term it becomes negatively biased. From the above figure it is inferred that for accurate estimation of the *Hurst*-index at least 512 points are needed for the *Whittle*-type method. For the wavelet method based on the figures it is inferred that at least 2^{12} points are needed for accurate estimations. Periodogram method, on the other hand, needs at least 2^{13} points for accurate estimations.

4.3 Application to Real Traces

The application of the above results to a real trace is now done. The trace selected is Bellcore's Ethernet trace measured in August 1989. The trace represents one hour of traffic on a LAN. The analysis performed is the same explained in section 3, let $1 = t_0, t_1, \ldots, t_k$ be a sequence of points in the x-axis, where $t_{i+1} > t_i$ and $\left(t_{i+1} - t_i\right) = 256$, to each block of trace X_j of length 256, $\{X_j\}_{j=t_i}^{t_i+256}$, apply a *Hurst* estimation method $\Theta(.)$ and finally construct the graph of the behaviour of the estimator $\Theta(.)$. Figure 8 shows the results of this analysis. Note from figure that periodogram overestimates and R/S statistic shows respectively irregular behaviour. Whittle-type estimator follows the H value reported in [17] and wavelet based method follow the reported value with high variability. Among the possible application areas of the current results are physiological time series where short time series are obtained [10] [11], administration of QoS parameters in real time, where a short measured trace is required in order to make performance decisions and in every discipline where time series length is a considerable factor affecting the performance of a system.

5 Conclusions and Future Work

This paper presented a study of the behaviour of estimators under short series in the context of *fGn* traces. Based on the study of the behaviour of thousands of *fGn* time series under bias, σ, convergent analysis and behaviour under time for a given length N_{min}, supossed to be the minimum length we arrive at the folllowing conclusions. The *Whittle*-type method behaves the best for short and long time series presenting both minimum bias and variability. The wavelet and periodogram method behave well when the time series is medium length. Variance and R/S statistic method behave with high bias and are not suitable for short-length measurements. The minimum length for accurate estimation of the *Hurst*-index was proposed for the estimation methods. Based largely on the above mentioned analyses the minimum length for *Whittle*-type method is at least 2^8, 2^{13} for the wavelet method, 2^{15} for the periodogram one and 2^{15} for the variance-type method. No minimum length was obtained for the R/S statistic method due to the high bias and variability in lengths of 2^{16}. The testing of these results in real Ethernet traces was also done. In the future we expect to study the same behaviour for f-*ARIMA* time series and also to study the effects of nonstationarities and trends on estimating the *Hurst*-index for short time series.

Acknowledgements. We finally acknowledge the reviewers for the useful comments that enriched the final version of the paper. All their comments and suggestions were taken into account in the final version of this paper. We also acknowledge CINVESTAV, Unidad Guadalajara and CONACYT for the support received.

References

1. Abry, P.: Wavelets, Spectrum Estimation and 1/f processes. Lecture Notes in Statistics **105** (1995) 15–30
2. Abry, P., Veitch D.: Wavelet Analysis of Long-Range Dependent Traffic. IEEE Transactions on Information Theory **44** (1998) 2–15
3. Abry, P., Veitch D., Flandrin, P.: Long-Range Dependence: Revisiting Aggregation with Wavelets. Journal of Time Series Analyis **19** (1998) 253–266
4. Audit, B., Bacry, E., Muzy, J., Arneodo, A.: Wavelet-Based Estimators of Scaling Behaviour. IEEE Transactions on Information Theory **48** (2002) 2938–2954
5. Beran, J.: Statistics for Long-memory Processes. New York, Chapman & Hall (1994)
6. Beran, J., Sherman, R., Taqqu, M. S., Willinger, W.: Long-range dependence in variable-bit-rate video traffic. IEEE Transactions on communications **43** (1995) 1566–1579

7. Cannon, M., Percival, D., Caccia, D., Raymond, G., Bassingthwaighte, J.: Evaluating Scaled Window Variance Methods for Estimating the Hurst Coefficient of Time Series. Physica A. **247** (1997) 606–626
8. Crovella, M., Bestavros, A.: Self-similarity in Word-Wide-Web traffic: evidence and possible causes. IEEE/ACM Transactions on Networking. **5** (1997) 835–846
9. Davies, R. B., Harte D.: Tests for Hurst Effect. Biometrika. **74** (1987) 95–101
10. Deligneres, D., Ramdani, S., Lemoine, L., Torre, K., Fortes, M., Ninot, G.: Fractal Analyses for Short Time Series: A Reassesment of Classical Methods. Journal of Mathematical Psychology. **50** (2006) 525–544
11. Eke, A., Herman, P., Bassingwaighte, J., Raymond, G., Percival, D., Cannon, M., Balla, I., Ikrenyi, C.: Physiological Time Series: Distinguishing Fractal Noises from Motions. Pflugers Archives. **439** (2000) 403–415
12. Embrechts, P., Maejima, M.: Self-Similar Processes. Princeton, Princeton University Press. (2002)
13. Giraitis, L., Kokoszka, P., Leipus, R., Teyssiere, G.: On the Power of R/S-Type Tests Under Contiguous and Semi Long-Memory Alternatives. Acta Applicandae Mathematicae. **78** (2003) 285–299
14. Hosking, J. R. M.: Modelling Persistence in Hydrological Time Series using Fractional Differencing. Water Resources Research. **20** (1984) 1898–1908
15. Karagiannis, T., Faloutsos, M.: SELFIS: A Tool for Self-similarity and Long-range Dependence Analysis. 1st Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches.(2002)
16. Karagiannis, T., Faloutsos, M., Molle, M.: A User-Friendly Self-similarity Analysis Tool. Special section on Tools and Technologies for Networking Research and Education, ACM SIGCOMM Computer Communication Review. (2003)
17. Leland, W. E., Taqqu, M.S., Willinger, W., Wilson, D. V.: On the self-similar nature of Ethernet traffic(Extended version). IEEE/ACM Transactions on Networking **2** (1994) 1–15
18. Mandelbrot, M., Wallis, J.: Robustness of R/S in Measuring Non-Cyclic Global Statistical Dependence. Water Resources Research. **5** (1969) 967–988
19. Park, K., Willinger W.: Self-similar Network Traffic and Performance Evaluation. Wiley-Interscience (2000)
20. Paxson, V., Floyd, S.: Wide-area traffic: The failure of poisson modelling. IEEE/ACM Transactions on Networking. **3** (1995) 226–244
21. Percival, B. P.: Stochastic Models and Statistical Analysis of Clock Noise. Metrologia. **40** (2003) S289–S304
22. Podesta, J.: Self-Similar Processes in Solar Wind Data. Advances in Space Research. **41** (2008) 148–152
23. Rangarajan, G., Ding, M.: Processes with Long-Range Correlations. Berlin, Springer (2003)
24. Serletis, A., Rosengerg, A.: The Hurst Exponent in Energy Future Prices. Physica A. **380** (2007) 325–332
25. Taqqu, M. S., Teverovsky, V.: Semi-parametric graphical estimation techniques for long-memory data Athens Conference on Applied Probability and Time series analysis. **115** (1996) 420–432
26. Taqqu, M. S., Teverovsky, V., Willinger, W.: Estimators for long-range dependence: An empirical study. Fractals. **3** (1995) 785–798
27. Tsybakov, B., Georganas, N.: Self-similar Processes in Communications Networks. IEEE Transactions on Infotmation Theory. **44** (1998) 1713–1725
28. Veitch, D., Abry, P.: A Wavelet Based Joint Estimator of the Parameters of Long-Range Dependence. IEEE Transactions on Infotmation Theory. **45** (1999) 878–897