The Modified Allan Variance as Time-Domain Analysis Tool for Estimating the Hurst Parameter of Long-Range Dependent Traffic

Stefano Bregni, *Senior Member, IEEE,* Luca Primerano

Politecnico di Milano, Dept. of Electronics and Information, P.zza L. Da Vinci 32, 20133 Milano, ITALY
Tel.: +39-02-2399.3503 – Fax: +39-02-2399.3413 – E-mail: bregni@elet.polimi.it

Abstract — **Experimental measurements show that Internet traffic exhibits self-similarity and long-range dependence (LRD). A delicate issue is the estimation of traffic statistical quantities that characterize self-similarity and LRD, such as the Hurst parameter *H*. In this paper, we propose to use the Modified Allan Variance (MAVAR), a well-known time-domain tool originally studied for frequency stability characterization, for estimating the power-law spectrum and thus the *H* parameter of LRD traffic time series. This novel method is validated by comparison to one of the most widely adopted algorithms for analyzing LRD traffic: the log-scale diagram technique based on wavelet analysis. Both methods are applied to pseudo-random data series, generated with known values of *H*. MAVAR exhibits outstanding accuracy in estimating *H*, better than the classical log-scale method. Finally, both techniques are applied to a real IP traffic trace, providing a further example of the capabilities of MAVAR.**

Index Terms — **Fractals, fractional noise, Internet, long-range dependence, self-similarity, traffic control (communication).**

I. Introduction

Experimental measurements have revealed that Internet traffic exhibits high temporal complexity. Far from being dominated by well-identifiable pseudo-periodic components, Internet traffic features intriguing temporal scale-invariance properties, such as self-similarity and long memory (long-range dependence) on various time scales [1]—[4]. Such examples include, but are not limited to, time series of successive TCP connection durations, cumulative data count transmitted over time, inter-arrival time series of successive TCP connections or IP packets, etc.

In a self-similar random process, a dilated portion of a realization (sample path) has the same statistical characterization than the whole. “Dilating” is applied on both amplitude and time axes of the sample path, according to a self-similarity parameter called Hurst parameter. On the other hand, Long-Range Dependence (LRD) is a long-memory model for scaling observed in the limit of largest time scales: LRD is usually equated with an asymptotic power-law decrease of the autocovariance function.

A delicate issue is the estimation of statistical parameters that characterize self-similar and LRD random processes. Among such quantities, the Hurst parameter *H* has been devoted particular attention in literature. By definition of self-similarity, it enters for example into characterization of Fractional Brownian Motion (FBM), fractal and multifractal processes. Several algorithms have been proposed to estimate *H* under various hypotheses [1][4][5].

In time and frequency measurement theory, a well-known time-domain tool for stability characterization of precision oscillators is the Modified Allan Variance (MAVAR) [6]—[12]. This variance was designed to discriminate effectively power-law noise types, recognized very commonly in frequency sources. Moreover, international standard bodies specify stability requirements for telecommunications network synchronization in terms of a quantity directly derived from MAVAR (Time variance, TVAR) [13].

In this work, we propose to use MAVAR for estimating the power-law spectrum and thus the *H* parameter of LRD traffic time series. This novel method is validated by comparison with one of the most widely adopted algorithms for the analysis of long-range dependent traffic: the log-scale diagram technique based on wavelet analysis [4]. To this purpose, both methods are applied to pseudo-random data series, generated with known values of the *H* parameter. The MAVAR method exhibits outstanding accuracy in estimating *H*, higher than the classical log-scale method, in spite of its computational lightweight. Finally, both techniques are applied to a real traffic trace, providing a further example of the capabilities of the MAVAR method.

II. Self-Similarity and Long-Range Dependence

A process *X*(*t*) (e.g., cumulative packet arrivals in the time interval 0-*t*) is said to be *self-similar*, with self-similarity or Hurst parameter *H*>0, if

$$X(t) \overset{d}{=} a^{-H} X(at) \qquad t \in \Re, \forall a > 0 \tag{1}$$

where $\overset{d}{=}$ means equality for all finite dimensional distributions [1][2]. In other terms, the statistical description of the process *X*(*t*) does not change by scaling simultaneously its amplitude by a^{-H} and the time axis by *a*. Self-similar processes are by definition non stationary, since the moments of *X*, provided they exist, behave as power laws of time, i.e.

$$E\left[\left|X(t)\right|^q\right] = E\left[\left|X(1)\right|^q\right] \; |t|^{qH} \tag{2}.$$

Work partialy supported by Ministero dell’Istruzione, dell’Università e della Ricerca (MIUR), Italy, under FIRB project TANGO.

0-7803-8794-5/04/$20.00 (C) 2004 IEEE

In practical applications, the class of self-similar processes is usually restricted to that of *self-similar processes with stationary increments* (or H-sssi processes), which are integral of some stationary process. For example, the δ-increment process of $X(t)$ is defined as $Y_\delta(t) = X(t)\text{-}X(t\text{-}\delta)$ (e.g., packet arrivals in δ time units). For a H-sssi process $X(t)$, we have $0<H<1$.

Long-Range Dependence (LRD) of a process is defined by an asymptotic power-law decrease of its autocovariance function [1][2]. Let $Y(t)$, with $t\in\Re$, be a second-order stationary stochastic process. The process $Y(t)$ exhibits LRD if either its autocovariance function follows

$$r_Y(\delta) \sim c_1|\delta|^{\gamma-1} \qquad \delta\to+\infty, \gamma\in(0,1) \tag{3}$$

or its spectral density follows

$$S_Y(f) \sim c_2|f|^{-\gamma} \qquad f\to 0, \gamma\in(0,1) \tag{4}.$$

In most practical cases, $0.5<H<1$ and $\gamma=2H\text{-}1$. All H-sssi processes $X(t)$ with $0.5<H<1$ have long-range dependent increments $Y(t)$.

Several techniques have been proposed to detect LRD and to estimate the Hurst parameter H in a given time series (e.g., traffic trace). In the time domain, the so-called *variance-time plot* method studies the covariance function of aggregated time series, made of samples computed by averaging windows of the original data set, as a function of the window width. By analyzing the covariance decay, as in (3), it is then possible to infer the spectrum power law and thus H. In the frequency domain, a simple *periodogram* plot allows to estimate H rather straightforward, according to the expression (4). Nevertheless, one of the most interesting and considered methods is the so-called *log-scale diagram*, based on wavelet decomposition [4].

III. The Modified Allan Variance

In time and frequency measurement theory, a well-known tool in the time domain for stability characterization of precision oscillators is the Modified Allan Variance (MAVAR) $\mathrm{Mod}\,\sigma_y^2(\tau)$ [6]—[12]. This variance was proposed in 1981 by modifying the definition of the two-sample variance recommended by IEEE in 1971 for characterization of frequency stability [8], following the pioneering work of D. W. Allan in 1966 [7]. Compared to the poor discrimination capability of the original Allan variance against white and flicker phase noise, the MAVAR discriminates effectively all power-law noise types recognized very commonly in frequency sources. Since then, international standard bodies have specified several stability requirements for telecommunications clocks in terms of a quantity directly derived from MAVAR (Time Variance, TVAR) [13].

Given an infinite sequence $\{x_k\}$ of samples evenly spaced in time with sampling period τ_0, the MAVAR is defined as

$$\mathrm{Mod}\,\sigma_y^2(n\tau_0) = \frac{1}{2n^2\tau_0^2}\left\langle\left[\frac{1}{n}\sum_{j=1}^{n}\left(x_{j+2n}-2x_{j+n}+x_j\right)\right]^2\right\rangle \tag{5}$$

where the observation interval is $\tau=n\tau_0$. In time and frequency stability characterization, the data sequence $\{x_k\}$ is made of samples of random time deviation $x(t)$ of the chronosignal under test. To summarize, modified Allan variance differs from basic Allan variance in the additional average over n adjacent measurements. For n=1 ($\tau=\tau_0$), the two variances coincide.

In practical measurements, given a finite set of N samples $\{x_k\}$, spaced by sampling period τ_0,an estimate of MAVAR can be computed using the ITU-T standard estimator [6][13]

$$\mathrm{Mod}\,\sigma_y^2(\tau) = \frac{1}{2n^4\tau_0^2(N-3n+1)}\sum_{j=1}^{N-3n+1}\left[\sum_{i=j}^{n+j-1}\left(x_{i+2n}-2x_{i+n}+x_i\right)\right]^2 \tag{6}$$

with n=1, 2,..., $\lfloor N/3\rfloor$. A recursive algorithm for fast computation of this estimator exists [6], which cuts down the number of operations to $\sim N$ instead of $\sim N^2$.

IV. Using the Modified Allan Variance for Estimating the Hurst Parameter

Following its definition, the MAVAR can be seen as the mean-square value of the signal output by a hypothetical filter, with proper impulse response shaped according to (5), receiving the data sequence $\{x_k\}$. Hence, the MAVAR can be also defined in the frequency domain [6][12], as the area under the spectral density of the signal output by such filter, i.e.

$$\mathrm{Mod}\,\sigma_y^2(\tau) = \int_0^\infty S_x(f)(2\pi f)^2\frac{2\sin^6\pi\tau f}{(n\pi\tau f)^2\sin^2\pi\frac{\tau}{n}f}df \tag{7}$$

where $S_x(f)$ is the power spectral density of input signal $x(t)$.

We are interested in power-law processes, whose spectral density behaves asymptotically for $f\to 0$ as

$$S_x(f) \sim k_1 f^\alpha \tag{8}$$

for $-1\le\alpha\le 0$ (LRD). Under this hypothesis (actually for the whole range $-4\le\alpha\le 0$, $\alpha\in\Re$), the MAVAR obeys asymptotically to a power law of the observation time τ, as

$$\mathrm{Mod}\,\sigma_y^2(\tau) \sim k_2\tau^\mu \tag{9}$$

where $\mu = -3-\alpha$ [6][12].

By definition of LRD, the autocovariance of a long-range dependent process $x(t)$ follows the asymptotical behavior (3) and its spectral density follows (4), with $0<\gamma<1$. Since $\gamma = 2H\text{-}1$ $(0.5<H<1)$, considering a relatively high number of samples (ideally $n\to\infty$) we obtain

$$\mathrm{Mod}\,\sigma_y^2 \sim k_2\tau^\mu \tag{10}$$

yielding the remarkable linear relation in a log-log plot

$$\log\sigma_y(\tau) \sim k_2 + (H-2)\log\tau \tag{11}.$$

Therefore, we can estimate the Hurst parameter of a LRD sample realization $\{x_k\}$ following this procedure:

1) compute MAVAR(τ) with the estimator (6), based on the data sequence $\{x_k\}$;
2) estimate its average slope μ in a log-log plot, at least in some intervals of values of τ, by best fitting a straight line to the curve (e.g., by least square error);
3) get the estimate of the Hurst parameter as

$$H = \frac{\mu}{2} + 2 \qquad (12).$$

It is worthwhile noticing that the estimate of MAVAR computed from a finite number of samples is a random variable itself. Its variance can be computed and used to assess the uncertainty of the estimate of H [10][6]. In our tests, nevertheless, we used the values of MAVAR computed for small n (left portions of plots), which have negligible uncertainty.

V. Method Validation and Accuracy Evaluation

In this section, this method for estimating the Hurst parameter is validated by comparison with one of the most widely adopted algorithms for the analysis of long-range dependent traffic: the log-scale diagram technique based on wavelet analysis [4].

To this purpose, both methods were applied to LRD pseudo-random data series $\{x_k\}$ of length N, each generated with assigned spectrum $S_x(f) \sim 1/f^\gamma$ ($0<\gamma<1$), according to specific values of $H=(\gamma+1)/2$. The generation algorithm is by Paxon [14]. In very short, it is based on spectral shaping: a vector of random complex samples, each with mean amplitude equal to the square root of the desired value of $S_x(f_k)$ and phase uniformly distributed in $[0, 2\pi]$, is inversely Fourier-transformed to yield the time-domain sequence $\{x_k\}$.

In our first test, five sequences $\{x_k\}$ of length N=50000, with mean 0 and variance 1, were generated for each value of H={0.50, 0.55, 0.60, ..., 0.95}. On the resulting 45 traces, we applied both the MAVAR and the log-scale diagram methods, getting two sequences of estimates. We then calculated the absolute and relative inaccuracy of these estimates with respect to the generation value H.

Fig. 1 shows three sample plots of Mod $\sigma_y(\tau)$ (Modified Allan Deviation, MADEV, square root of MAVAR) evaluated on pseudo-random LRD sequences for H=0.55, 0.75, 0.95. The three curves are almost linear and with expected slope, confirming the correctness of the simulation procedure and the effectiveness of MAVAR in discriminating power-law spectra.

In Fig. 2, the H values imposed in generating the pseudo-random sequences are compared to the H values estimated by the MAVAR method. Bars of max/min values, out of the 5 estimates per each H value, are superposed to the ideal line. In Fig. 3, the same comparison is made between imposed H and H estimated by the log-scale diagram method.

Moreover, Figs. 4 and 5 show the estimation errors of H by the two methods (cf. Figs. 2 and 3). Bars of max/min errors and mean error, out of the 5 estimates per each H value, are plotted. By inspection of these graphs, it is evident that, while both methods feature good accuracy, the MAVAR technique is better performing, because it achieves smaller error bars and no error bias towards positive or negative values.

Now, it is interesting to assess the accuracy of the two methods with short sequences. With sequences made of few samples, the use of the log-scale diagram may be misleading, as it is based on the estimation of the variance of wavelet coefficient details, which are only $\log_2 N$ at most, where N is the number of samples. The MAVAR method, on the other hand, may suffer poor confidence in variance estimates.

Therefore, we repeated the same test as before, but by generating 45 sequences of length N=200000 and by truncating them to the first 1000 samples. Analogously to previous graphs, Figs. 6 and 7 compare the H values estimated with the two methods to the values imposed in noise generation, while Figs. 8 and 9 compare estimation errors. In this case, the better accuracy of the MAVAR technique is even more evident, considering both the error amplitude and its bias.

To further point out the impact of the sample sequence length N on the accuracy of the H estimate, we applied the MAVAR method to 4 different pseudo-random sequences, generated with H=0.75 and truncated to increasing lengths, up to N=50000. The graph in Fig. 10 plots the error of the resulting estimates as a function of N.

VI. Application to a Real IP Traffic Trace

Finally, we compared the behavior of the MAVAR and log-scale methods on a real IP traffic trace, obtained by counting the packets transmitted per time unit on a transoceanic system (MAWI Project [15]). The data sequence is made of $N=2^{16}$=65536 samples, acquired with sampling period τ_0=8 ms, thus spanning a measurement interval $T\cong$524 s.

First, we computed the log-scale diagram shown in Fig. 11, using scripts available at [16] (Daubechies' wavelet with two vanishing moments), where vertical bars represent 95%-confidence intervals. The slope of the straight line best-fitting the left portion of the curve yields the estimate H=0.588. First, we notice that, because of the irregular trend of the curve, changing the interval on which line best-fitting is calculated yields slightly different results. Moreover, at the right side, the uncertainty of the measurement is too high to infer meaningful results. However, despite the multi-slope trend and the wide confidence bars, an average slope increase is evident there.

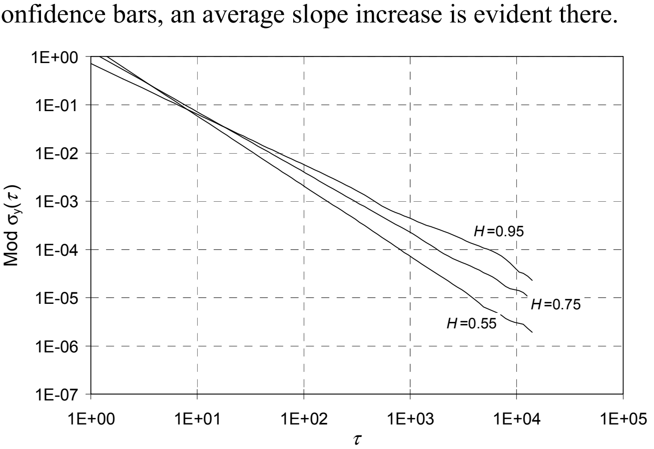

Fig. 1: Modified Allan Deviation (MADEV) evaluated on three pseudo-random LRD noise sequences for H=0.55, 0.75, 0.95.

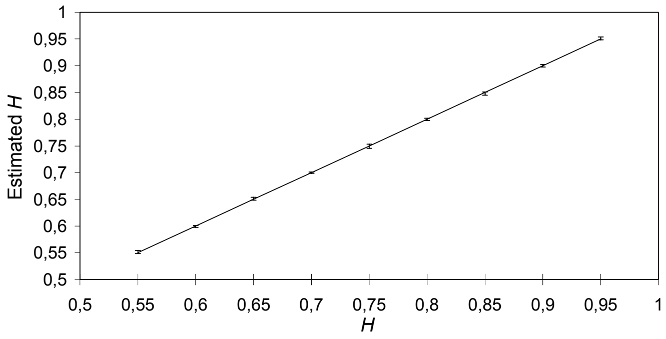

Fig. 2: Comparison of *H* values imposed in generating data sequences with *H* values estimated by the MAVAR method (*N*=50000 samples). Bars of max/min values, out of the 5 estimates per each *H* value, and ideal line.

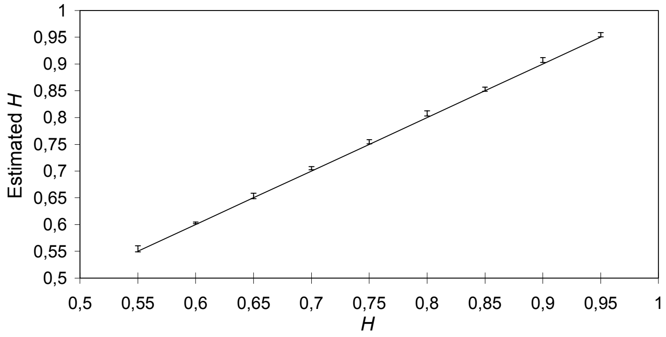

Fig. 3: Comparison of *H* values imposed in generating data sequences with *H* values estimated by the LOG-SCALE method (*N*=50000 samples). Bars of max/min values, out of the 5 estimates per each *H* value, and ideal line.

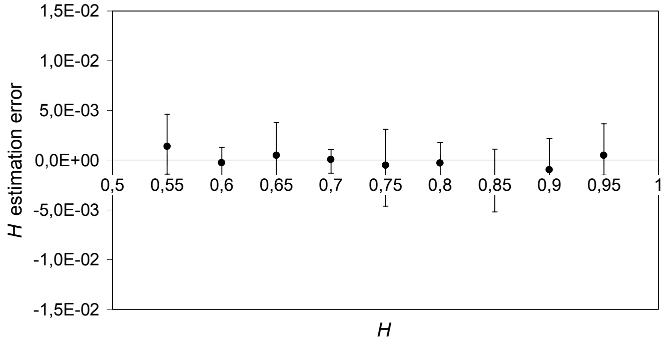

Fig. 4: Estimation error of *H* by the MAVAR method (*N*=50000 samples). Bars of max/min errors and mean error, out of the 5 estimates per each *H* value, are plotted. Cf. Fig. 2.

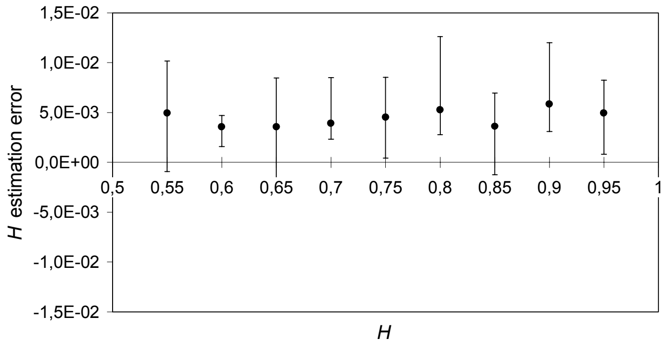

Fig. 5: Estimation error of *H* by the LOG-SCALE method (*N*=50000 samples). Bars of max/min errors and mean error, out of the 5 estimates per each *H* value, are plotted. Cf. Fig. 3.

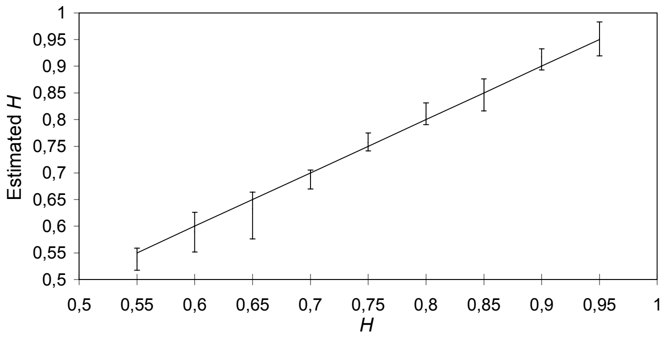

Fig. 6: Comparison of *H* values imposed in generating data sequences with *H* values estimated by the MAVAR method (*N*=1000 samples). Bars of max/min values, out of the 5 estimates per each *H* value, and ideal line.

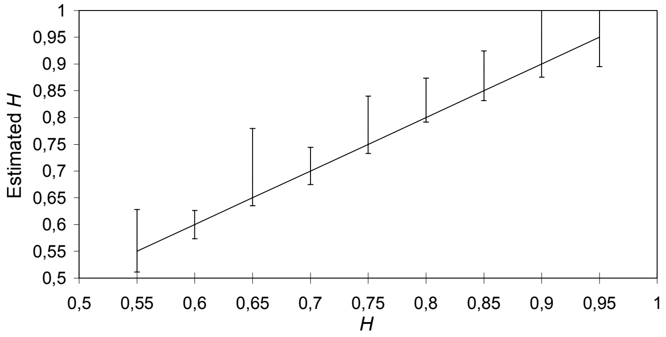

Fig. 7: Comparison of *H* values imposed in generating data sequences with *H* values estimated by the LOG-SCALE method (*N*=1000 samples). Bars of max/min values, out of the 5 estimates per each *H* value, and ideal line.

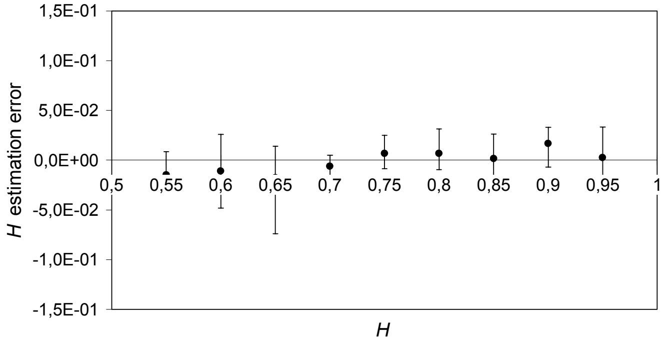

Fig. 8. Estimation error of *H* by the MAVAR method (*N*=1000 samples). Bars of max/min errors and mean error, out of the 5 estimates per each *H* value, are plotted. Cf. Fig. 6.

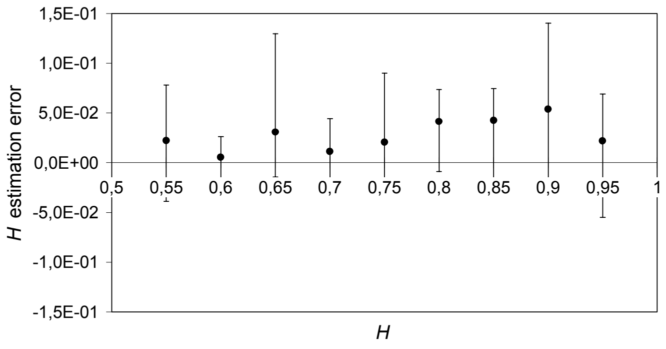

Fig. 9: Estimation error of *H* by the LOG-SCALE method (*N*=1000 samples). Bars of max/min errors and mean error, out of the 5 estimates per each *H* value, are plotted. Cf. Fig. 7.

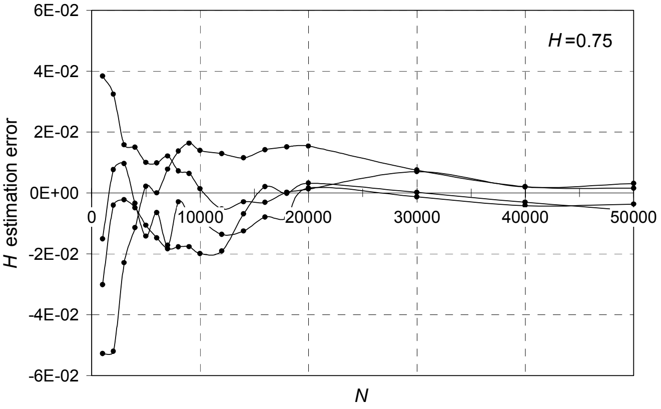

Fig. 10: Convergence of the MAVAR method in estimating H of 4 different pseudo-random sequences (H=0.75 and truncated to increasing lengths).

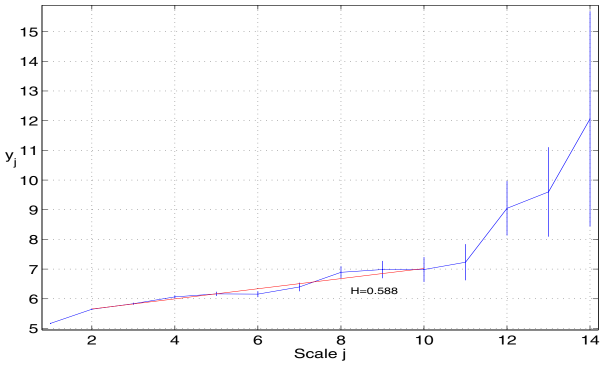

Fig. 11: Log-scale diagram of a real IP packet/time trace (MAWI [15], N=65536, τ_0=8 ms, T≅524 s).

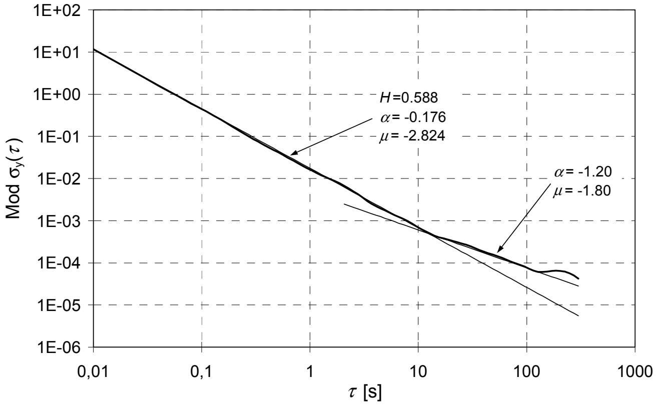

Fig. 12: Modified Allan Deviation (MADEV) of a real IP packet/time trace (MAWI [15], N=65536, τ_0=8 ms, T≅524 s).

Then, we computed MAVAR on the same traffic trace, obtaining the curve in Fig. 12. Opposite to log-scale, MAVAR gives an astonishingly clear picture of the spectral characteristics of the traffic sequence under analysis. The MAVAR curve, in log-log scale, is nearly precisely made of two linear segments, with slopes μ=-2.824 (τ<10 s) and μ=-1.80 (10 s<τ<100 s). Thus, the data sequence analyzed is revealed to be almost exactly sum of two simple components with power-law spectrum (8), with α=-0.176 dominant for τ<10 s and α=-1.20 dominant for 10 s<τ<100 s. The former noise (α=-0.176) is LRD-type, with H=0.588 (12).

VII. Conclusions

In this paper, we proposed to use the Modified Allan Variance, a time-domain tool originally studied for frequency stability characterization, for estimating the power-law spectrum and thus the H parameter of LRD traffic time series. This novel method was compared to one of the most widely adopted algorithms for analyzing LRD traffic: the logscale-diagram technique based on wavelet analysis.

Both methods were applied to pseudo-random data series, generated with known values of H. The MAVAR method proved very accurate in estimating H, better than the classical log-scale method, even despite its computational lightweight.

Finally, both techniques were applied to a real IP traffic trace (MAWI [15]), providing a further example of MAVAR capabilities. While the log-scale method produced uncertain results and H estimates, the fine accuracy of MAVAR spectral analysis allowed to recognize that the IP trace under test is precisely made of two simple fractional noise components, having power-law spectrum $k_1/f^{0.176} + k_2/f^{1.20}$. The first term is LRD, with H=0.588.

References

[1] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, D. Veitch, "The Multiscale Nature of Network Traffic: Discovery Analysis and Modelling" *IEEE Signal Processing Magazine*. April 2002.

[2] *"Self-similar Network Traffic and Performance Evaluation"*. K. Park and W. Willinger, Eds. Wiley Interscience, 2000.

[3] V. Paxson, S. Floyd, "Wide-Area Traffic: the Failure of Poisson Modeling", *IEEE/ACM Trans. on Networking,* vol. 3, no. 6, June 1995, pp. 226-244.

[4] P. Abry, D. Veitch, "Wavelet Analysis and Long Range Dependent Traffic", *IEEE Trans. Inform. Theory*, vol. 4, no.1, pp. 2-15, 1998.

[5] M. S. Taqqu, V. Teverovsky, W. Willinger, "Estimators for Long-Range Dependence: an Empirical Study", *Fractals*, vol. 3, no.4, 1995.

[6] S. Bregni, *Synchronization of Digital Telecommunications Networks*. Chichester, UK: John Wiley & Sons, 2002.

[7] D. W. Allan, "Statistics of Atomic Frequency Standards", *Proceedings of the IEEE*, vol. 54, no. 2, July 1966.

[8] J. A. Barnes, A. R. Chi, L. S. Cutler, D. J. Healey, D. B. Leeson, T. E. McGunigal, J. A. Mullen Jr., W. L. Smith, R. L. Sydnor, R. F. C. Vessot and G. M. R. Winkler, "Characterization of Frequency Stability," *IEEE Trans. on Instr. and Meas.*, vol. IM-20, no. 2, May 1971.

[9] D. W. Allan, J. A. Barnes, "A Modified Allan Variance with Increased Oscillator Characterization Ability", *Proc. of the 35th Annual Frequency Control Symposium*, 1981.

10] P. Lesage and T. Ayi, "Characterization of Frequency Stability: Analysis of the Modified Allan Variance and Properties of Its Estimate," *IEEE Trans. on Instr. and Meas.*, vol. IM-33, no. 4, Dec. 1984.

[11] L. G. Bernier, "Theoretical Analysis of the Modified Allan Variance," *Proc. of the 41st Annual Frequency Control Symposium*, 1987.

[12] J. Rutman, F. L. Walls, "Characterization of Frequency Stability in Precision Frequency Sources", *Proc. of the IEEE*, vol. 79, no. 6, 1991.

[13] ITU-T Rec. G.810 *"Definitions and Terminology for Synchronisation Networks"*, Rec. G.811 *"Timing Characteristics of Primary Reference Clocks"*, Rec. G.812 *"Timing Requirements of Slave Clocks Suitable for Use as Node Clocks in Synchronization Networks"*, Rec. G.813 *"Timing Characteristics of SDH Equipment Slave Clocks (SEC)"*, 1996-2003.

[14] Paxson, V. "Fast Approximation of Self-Similar Network Traffic", *ACM/SIGCOMM Computer Communication Review*, vol. 27, no. 7, Oct. 1997, pp. 5-18.

[15] MAWI (Measurement and Analysis on the WIDE internet) project. Available at URL: http://tracer.csl.sony.co.jp/mawi/

[16] Darryl Veitch, *"Code for The Estimation of Scaling Exponents"*. http://www.cubinlab.ee.mu.oz.au/~darryl/secondorder_code.html.