PERIODICA POLYTECHNICA SER. EL. ENG. VOL. 43, NO. 4, PP. 227–250 (1999)

ON THE EFFECTS OF NON-STATIONARITY IN LONG-RANGE DEPENDENCE TESTS

Trang Dinh DANG and Sándor MOLNÁR

High Speed Networks Laboratory
Dept. of Telecommunications and Telematics
Budapest University of Technology and Economics
H–1117, Budapest, Pázmány Péter sétány 1/D, Hungary
Tel: (+361) 463 3889, Fax: (+361) 463 3107
E-mail: {trang,molnar}@ttt-atm.ttt.bme.hu

Received: Dec. 8, 1999

Abstract

Careful statistical analyses indicate that the measured traffic traces from live packet networks often contain non-stationary effects like level shifts or polynomial trends. In these cases several popular tests for long-range dependence can result in wrong conclusions and unreliable estimate of the Hurst parameter. In this paper both analytical and simulation investigations of the implications of these effects on several tests are presented. The results are also demonstrated with examples based on measured ATM traces. The use of these results can be utilized to avoid pitfalls in LRD traffic modeling.

Keywords: long-range dependence, non-stationarity, statistical tests.

1. Introduction

It has been widely recognized on the basis of a series of comprehensive analyses of real-time traffic measurements from working packet switched networks that packet traffic fluctuates over a number of time scales [26]. This behaviour is called *burstiness*. However, the unique definition and characterization of burstiness have not been established yet in the teletraffic literature [11, 18].

A very promising approach to capture this burstiness phenomenon in a parsimonious manner is to use *fractal traffic models* [16, 26]. These models have dynamics governed by *power law* distribution functions and *hyperbolically* decaying autocorrelation [26]. The important characteristics of these models are *self-similarity* and *long-range dependence* [11, 16].

Self-similar stochastic processes have been defined in a number of ways in the literature [11, 16, 26]. From a practical point of view the long-range dependent (LRD) processes constitute one of the most important classes of these processes [11, 16]. In this paper we consider this class of fractal processes which is defined in the next section.

The important practical issues are the identification of LRD phenomena and the estimation of LRD parameters, especially the estimation of the Hurst parameter.

Unfortunately, testing for LRD of measured data is not possible by simply checking the definitions. Instead, we can use some methods for testing the presence of some characteristics of the data which can or cannot support the LRD property and also can or cannot give a reliable estimate of the Hurst parameter. Moreover, if all methods support the assumption of the presence of LRD with a parameter H it is still possible that this observation is caused by non-stationarities present in the data and not due to the LRD (since LRD is only defined for the set of stationary processes). In this case it is possible to end up with wrong conclusions and build wrong models. In order to avoid such pitfalls we address this problem in this paper and give both analytical and simulation investigations of these effects with different non-stationarities in the data.

The issue is not new and was also addressed in the hydrology literature (e.g. [13]) after the application of LRD processes in the modeling of natural storage systems by HURST [10], MANDELBROT and others [15, 25]. However, after the discovery and first application of LRD processes in the teletraffic research a number of papers have been published just by blind application of some LRD tests assuming the stationarity for hours of the traffic and taking no care of this important question. We note that the problem was also addressed in the recent teletraffic literature, e.g. in [6, 9, 8, 18, 16] and also see the related references in [26], but stationarity tests and the validation techniques of fractal models have not widely been applied in today's teletraffic practice.

There are some approaches to deal with this problem. One practical solution is based on the notion of *local stationarity*. Here we assume stationarity only over some short periods of time. Therefore our model parameters are valid only for such a period and should be updated in the next period. A measurement-based approach with periodic real-time parameter estimation is a possible solution. Local stationarity with traditional models can also be used to capture the observed characteristics [24, 23].

An alternative but rather difficult solution is to use *non-stationary models*, e.g. [7].

Some authors argue that this topic is somewhat philosophical from the applied point of view [8, 13]. Indeed, if the modeling alternative can provide useful practical tools to dimension our networks, then this can be a non-questionable proof for a proposed model. However, if more alternatives can work, then we may prefer the parsimonious one which is a nice feature of fractal models. We believe that besides these factors the final choice of the proposed model and understanding about the nature of network traffic should be made not only by the analysis of the measured data but our *a priori* knowledge about the traffic generation process.

The contribution of this paper is to reveal the implications of the most important non-stationary effects which occur in practice on the most frequently used LRD tests in order to have a good understanding of these phenomena and investigate the robustness of these tests against non-stationarities. The practical use of our findings is to support teletraffic engineers with guidelines not to mistake actual non-stationarities for stationary fractal behaviour.

In Sections 2 and 3 the methods and the non-stationary effects under investigation are described, respectively. Our analytical investigations for the tests of variance-time plot and R/S plot with level shifts and linear trends are given in Section 4. Our simulation study with several examples is presented in Section 5, and Section 6 concludes our paper.

2. Long-Range Dependence Tests

In this section we give a short overview of LRD processes with the most frequently used test methods which are analyzed in the paper.

Let $X = (X_k : k \geq 0)$ be a covariance-stationarity process with autocorrelation function $r(k)$. X is said to exhibit *long-range dependence (LRD)* if $r(k) \sim k^{2H-2}L(k)$ as $k \to \infty$, $1/2 < H < 1$, where L is slowly varying at infinity, i.e., $\lim_{k\to\infty} [L(tk)/L(k)] = 1, \quad t > 0$ and $a(x) \sim b(x)$ means $a(x)/b(x) \to 1$ as $x \to \infty$. The class of LRD processes is equivalent to the class of *asymptotically second-order self-similar* processes [22] defined as follows. For all integer $m \geq 1$ let $X_k^{(m)} = \frac{1}{m}\sum_{i=(k-1)m+1}^{km} X_i$ be the aggregated process with autocorrelation function $r^{(m)}(k)$. X is called asymptotically second-order self-similar if $\lim_{m\to\infty} r^m(k) = 1/2(|k+1|^{2H} - 2\,|k|^{2H} + |k-1|^{2H})$ for all $k \geq 0$. The most important property of these processes is that the aggregated process has a non-degenerate autocorrelation structure as $m \to \infty$. In contrast, in the case of short-range dependent processes (e.g. Markov processes) this aggregate process tends to second-order pure noise.

As discussed in the previous section, the task for testing of LRD and the estimation of the Hurst parameter are not simple in practice. The main problem is that it is rather difficult to distinguish between non-stationary processes and stationary LRD processes due to the fact that LRD processes appear to have local trends, cycles, etc., many of the characteristics of non-stationary processes. These properties disappear after some time but if we have a finite and sometimes also short data set, this identification is almost impossible. Having a longer data set this identification becomes easier but we know for sure that in a larger measured data set non-stationary effects are present due to the daily cycles of traffic characteristics. The assumption about stationarity with high reliability may only be supported in the *busy periods* of the traffic. However, in some cases (e.g. IP traffic in a LAN) the notion of busy period cannot be applied [3].

There are methods developed to test stationarity (e.g. [23, 19, 17]) and to distinguish between LRD and non-stationarities (e.g. [21, 14, 2, 20]) but application of these tests is not easy in practice. Moreover, such tests can seldom support their results with high reliability. We review here four widely used LRD tests: the variance-time plot, the R/S analysis, the periodogram and the wavelet based H-estimator. More detailed description of these methods can be found e.g. in [2] and [1].

2.1. Variance-Time Plot

The variance-time plot is constructed based on the following asymptotic property of LRD processes [2],

$$\mathrm{Var}(X^{(m)}) = m^{2H-2}\,\mathrm{Var}(X) \qquad \text{as } m \to \infty, \tag{1}$$

where H is the Hurst parameter of X, $X^{(m)}$ denotes the m-aggregated process of X, m is the aggregation level, $m = 1, 2, \ldots$

In practice, for a given time series X of size n, one chooses the maximum value of m so that $\lfloor n/m \rfloor$ is still large enough and then logarithmically increases m from 1 to that value. For those successive values of m, the logarithm of the sample variance of $X^{(m)}$ is plotted versus the logarithm of m. If X is LRD, this *variance-time plot* should be a straight line with a slope of $2H - 2$. An estimation of the Hurst parameter can be calculated by fitting a least-squares line to points of the plot over the large values of m.

Since $0.5 < H < 1$, the asymptotic slope of the variance-time has a value between -1 and 0. The variance-time plot with slope -1 suggests that the series has no LRD and it has finite variance. The Poissonian and Markovian processes are typical examples of such short-range dependent processes.

2.2. R/S Analysis

Consider a time series X of size d with sample mean $\bar{X}_d$ and sample variance $S^2(d) = (1/d)\sum_{i=1}^{d}(X_i - \bar{X}_d)^2$. The *rescaled adjusted range* [2] R/S statistics of X is given by the ratio:

$$\frac{R(d)}{S(d)} = \frac{\max\{W_i : i = 1, 2, \ldots, d\} - \min\{W_i : i = 1, 2, \ldots, d\}}{S(d)}, \tag{2}$$

where $W_i = \sum_{k=1}^{i}(X_k - \bar{X}_d)$. It can be proven for any stationary process with LRD of parameter H that the R/S ratio has the following characteristics for large n:

$$\mathbf{E}\left[\frac{R(d)}{S(d)}\right] \sim \left(\frac{d}{2}\right)^{H}, \tag{3}$$

which is known as *Hurst effect* [2].

Given an empirical time series of length n ($X_j : j = 1, 2, \ldots, n$), subdivide the series into K blocks of size $\lfloor n/K \rfloor$. Then for each lag $d := \lfloor n/K \rfloor$, compute the $R(t_i, d)/S(t_i, d)$ ratios, where t_i denotes the starting point of the data block d, i.e., $t_i = \lfloor n/K \rfloor(i - 1) + 1,\ i = 1, 2, \ldots, d$. Thus one has K estimates of $R/S(d)$ for each value of d. Choosing logarithmically spaced values of d ($d < n$) and plotting $\log[R/S(t_i, d)]$ versus $\log d$ results in the R/S plot, also known as *pox diagram*.

Next, a least squares regression line should be fitted to points of the R/S plot. The slope of the regression line gives an estimate of the Hurst parameter of LRD. The smallest values of d should be disregarded because these points are dominated by short-range dependence in the series. We do not use the high end of the plot either because only a few points in this region may make the estimate unreliable. In practice, values of d in the middle region of the R/S plot are used to estimate H.

2.3. Periodogram

A typical property of LRD exhibited in the frequency domain is called $1/f$ *noise*. The $1/f$ noise expresses the behaviour of the power spectral density of an LRD process at small frequencies, i.e.,

$$\log f(\nu) \sim -\gamma \log \nu, \qquad \text{as} \quad \nu \to 0, \tag{4}$$

where $f(.)$ denotes the power spectral density and $\gamma = 2H - 1$.

An estimate of the spectral density can be obtained by the Fourier transform of the estimate of the autocorrelation function. This estimator is referred to as a periodogram [2], and is defined as

$$I(\nu) = \frac{1}{2\pi n} \left| \sum_{k=1}^{n} (X_k - \bar{X}) e^{ik\nu} \right|^2, \tag{5}$$

for a discrete time series $X = \{X_1, X_2, \ldots, X_n\}$.

The main idea of periodogram analysis is simply to plot the periodogram in a log-log grid and to estimate H from the slope of the regression line fitted to the plot at low frequencies. Practically, the periodogram plot is the graph of $\{\log \nu_j, \log I(\nu_j)\}$, $j = 1, 2, \ldots, M$, where $\nu_j = 2\pi j/n$ and M is always chosen to be $n/4$, $n/8$, $n/16$ or $n/32$ and so on depending on how large n is. According to *Eq.* (4), the plot should be a straight line with slope $-\gamma = 1 - 2H$ in the case of LRD processes .

2.4. Wavelet-Based Estimator

Wavelet analysis of LRD traffic is introduced by P. ABRY and D. VEITCH in [1]. The estimator is found to be very unbiased and highly robust against the presence of deterministic trends. The description of the wavelet estimator is briefly reviewed here.

The discrete wavelet transform (DWT) represents a discrete series $\{X_1, X_2, \ldots, X_n\}$ by a combination of the scaled and delayed versions of the mother wavelet function $\psi(.)$. At scale level j the wavelet coefficients $d_x(j,k)$ are defined as follows:

$$d_x(j,k) = 2^{j/2} \sum_{i=1}^{n} X_i \psi(2^{-j}n - k) \qquad j = 1, 2, \ldots; \quad k = 1, 2, \ldots, 2^{-j}n. \tag{6}$$

Let X be a second-order stationary process. Then its wavelet coefficients $d_x(j,k)$ satisfy:

$$E\left[d_x(j,k)^2\right] = \int f(\nu) 2^j |\Psi(2^j \nu)|^2 d\nu, \tag{7}$$

where $f(\nu)$ and $\Psi(\nu)$ are the power spectrum of X and the Fourier transform of the wavelet function $\psi(.)$, respectively. Based on *Eq.* (4) we have

$$E\left[d_x(j,k)^2\right] \sim 2^{j(2H-1)} c_f C(H, \psi), \tag{8}$$

where $C(H,\psi) = \int |\nu|^{-(2H-1)} |\Psi(\nu)|^2 d\nu$ is a constant which depends on H and ψ.

If the length of X is n, then the available number of wavelet coefficients at octave j is n_j, $n_j = 2^{-j}n$. Then,

$$\mu_j = E\left[d_x(j,k)^2\right] \approx \frac{1}{n_j} \sum_{k=1}^{n_j} |d_x(j,k)|^2. \tag{9}$$

Eq. (8) provides a possible way to estimate the Hurst parameter of the LRD processes:

$$\log_2 \mu_j \approx \log_2 \left(\frac{1}{n_j} \sum_{k=1}^{n_j} |d_x(j,k)|^2 \right) \sim (2H-1)j + c, \tag{10}$$

where $c = \log_2(c_f C(H,\psi))$ is a constant. This means that if X is LRD with Hurst parameter H, then the graph of $\log_2(\mu_j)$ versus j, called the Logscale Diagram (LD), should be linear with slope $2H-1$. (In practice, the raw data can be preprocessed to make the result of the LD more reliable and $\log \mu_j$ is replaced by a corrected version denoted by y_j in the implementation of the authors of [1].)

As discussed in detail in [1], the effects of polynomial trends with the degree P on this estimator can be avoided by increasing the vanishing moment N of the wavelet function so that $N \geq P+1$. This observation is justified by our simulations presented later.

3. Types of Non-Stationarities

The analysis of measured packet traffic can reveal various deterministic changes in the data on different time scales. These traffic variations are not stochastic by nature but rather caused by deterministic mechanisms like protocols [12]. These mechanisms can, for example, introduce quasi-periodic patterns in the traffic data which can be, if not detected and removed, the reasons for several statistical pitfalls, e.g. the conclusion of slowly decaying correlations.

On longer time scales we can observe also a regular character of the traffic due to daily or weekly variations. These traffic trends should also be identified and removed prior to any statistical analysis. These are not easy but important parts of a comprehensive statistical analysis [5, 4].

Different trend models are candidates for investigations, e.g. linear trend, parabolic trend, exponential trend, logistical trend or Gompertz trend, etc. We have chosen the non-stationary effects and trends which are frequently observed in practice. These are the *level shift*, which can be observed when during our traffic measurements suddenly a new source starts to emit traffic to the aggregation and the *linear and parabolic trends*, which can be observed in daily traffic variations. For example, when people start to work in their office between 8 and 10 am a monotonic increase of the total load of the aggregated traffic can be observed.

4. Analytical Investigations

In this section we present our analytical study which shows how some non-stationarities can change the results of some widely used LRD tests. We concern here three cases: variance-time plot of LRD data with level shift, with linear trend, and R/S analysis of LRD data with level shift.

Consider an $\{X_1, X_2, \ldots X_n\}$ series which is LRD with Hurst parameter H. To make the later calculations simple we use two assumptions: (1) n is large enough so that aggregated series of $\{X\}$ used in computation of the variance-time plot still contains a large amount of data; (2) the mean of $\{X\}$ is zero, i.e., $\bar{X}_n = 0$. The second assumption can be taken into account because the non-zero mean of LRD data does not change the result of LRD tests (see their definitions in Section 2).

4.1. Variance-Time Plot of LRD Data with Level Shift

The variance-time plot is the log-log plot of the variance of data versus the aggregation level. The corrected sample variance of $\{X\}$ series:

$$\mathrm{Var}(X) = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2, \tag{11}$$

since the mean of $\{X\}$ is zero. Similarly, the corrected sample variance of the m-aggregated series of $\{X\}$ is:

$$\mathrm{Var}(X^{(m)}) = \frac{1}{\lfloor n/m \rfloor - 1} \sum_{j=1}^{\lfloor n/m \rfloor} \left(X_j^{(m)} \right)^2, \tag{12}$$

where $\lfloor z \rfloor$ denotes the greatest integer smaller than or equal to z. *Eq.* (12) holds because we have assumed that $\lfloor n/m \rfloor$ is still large enough so $\bar{X}^{(m)}_{\lfloor n/m \rfloor} \approx \bar{X}_n = 0$[1]. The following relation holds for LRD series:

$$\mathrm{Var}\left(X^{(m)}\right) = \frac{\mathrm{Var}(X)}{m^{2-2H}}. \tag{13}$$

After adding a level shift to the series X, the i-th element of the new series, denoted by X_i^{LS} has the value:

$$X_i^{LS} = \begin{cases} X_i & \text{if } i \le \lfloor n/2 \rfloor \\ X_i + t_m & \text{if } i > \lfloor n/2 \rfloor \end{cases},$$

where t_m denotes the value of the level shift occurred in the middle of the investigated time period[2]. It is easy to observe that the mean of the X^{LS} series is $t_m/2$. Thus its variance is of the form:

$$\begin{aligned}
\mathrm{Var}\left(X^{LS}\right) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^{LS} - t_m/2\right)^2 \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{\lfloor n/2 \rfloor}(X_i - t_m/2)^2 + \sum_{i=\lfloor n/2 \rfloor+1}^{n}(X_i + t_m - t_m/2)^2\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 + \frac{n}{n-1}\frac{t_m^2}{4} + \frac{t_m}{n-1}\left[\sum_{i=\lfloor n/2 \rfloor+1}^{n}X_i - \sum_{i=1}^{\lfloor n/2 \rfloor}X_i\right] \\
&\approx \mathrm{Var}(X) + \frac{t_m^2}{4}, \qquad \text{if } n \text{ is large enough.}
\end{aligned} \tag{14}$$

The j-th element of the m-aggregated series $X^{(m)}$ is given by $X_j^{(m)} = \sum_{k=(j-1)m+1}^{jm} X_k$, therefore

$$X_j^{LS\ (m)} = \begin{cases} X_j^{(m)} & \text{if } j \le \lfloor n/2m \rfloor \\ X_j^{(m)} + t_m & \text{if } j > \lfloor n/2m \rfloor \end{cases},$$

with the only exception when the element contains the location of the shift. Since $\lfloor n/m \rfloor$ is large enough this exception does not change the result. Thus doing the same calculations as in *Eq.* (14) we have

$$\mathrm{Var}\left(X^{LS\ (m)}\right) \approx \mathrm{Var}\left(X^{(m)}\right) + \frac{t_m^2}{4}. \tag{15}$$

[1] This equality only holds when dealing with stationary data series.

[2] The location of the level shift jump has no effect on analytical and simulation results.

By inserting *Eq.* (14) and *Eq.* (15) into *Eq.* (13) we get the final result:

$$\mathrm{Var}\left(X^{LS\ (m)}\right) = \frac{\mathrm{Var}\left(X^{LS}\right) - t_m^2/4}{m^{2-2H}} + \frac{t_m^2}{4}. \tag{16}$$

By plotting $\log\left[\mathrm{Var}\left(X^{LS\ (m)}\right)\right]$ against $\log m$ we get a convex curve bounded by two lines: the line with slope $2H-2$ and ordinate $\log\left[\mathrm{Var}\left(X^{LS}\right) - t_m^2/4\right]$ as $m \to 0$ and a horizontal line with ordinate $t_m^2/4$ as $m \to \infty$. The estimation of H for LRD processes should be performed at large m (in theory as $m \to \infty$). Therefore we can conclude that the estimation is highly destroyed in the presence of level shifts. More details about this distortion demonstrated by examples are given in Section 5.

4.2. *Variance-Time Plot of LRD Data with Linear Trend*

In this case we also denote the maximum value of the linear trend by t_m. The LRD series with linear trend X^L can be given as:

$$X_i^L = X_i + \frac{(i-1)\ t_m}{n-1}\ . \tag{17}$$

Because the mean of X is zero, i.e., $\bar{X}_n = 0$, the mean of the new series is $\bar{X}^L = t_m/2 \approx \bar{X}^{L\ (m)}_{\lfloor n/m \rfloor}$, where $\bar{X}^{L\ (m)}_{\lfloor n/m \rfloor}$ denotes the mean of m-aggregated series of X^L. The variance of X^L can be calculated as follows:

$$\begin{aligned}\mathrm{Var}\left(X^L\right) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^L - \frac{t_m}{2}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[X_i + \frac{(i-1)\ t_m}{n-1} - \frac{t_m}{2}\right]^2 \\ &= \mathrm{Var}(X) + \frac{2t_m}{(n-1)^2}\sum_{i=1}^{n} i\ X_i + \frac{t_m^2}{4(n-1)^3}\sum_{i=1}^{n}(2i-n-1)^2 \\ &\approx \mathrm{Var}(X) + \frac{2t_m}{(n-1)^2}\sum_{i=1}^{n} i\ X_i + \frac{t_m^2}{12} = \mathrm{Var}(X) + C_1\ ,\end{aligned} \tag{18}$$

where C_1 is a constant independent of m for a given data.

Similarly, for m-aggregated series we have:

$$X_j^{L\ (m)} = X_j^{(m)} + \frac{t_m}{2(n-1)}(2jm - m - 1) \qquad j = 1, 2, \ldots, \lfloor n/m \rfloor \tag{19}$$

and

$$\begin{aligned}\mathrm{Var}\left(X^{L\ (m)}\right) &= \frac{1}{\lfloor n/m\rfloor - 1}\sum_{j=1}^{\lfloor n/m\rfloor}\left[X_j^{L\ (m)} - \frac{t_m}{2}\right]^2\\ &= \frac{1}{\lfloor n/m\rfloor - 1}\sum_{j=1}^{\lfloor n/m\rfloor}\left[X_j^{(m)} + \frac{t_m}{2(n-1)}(2jm - m - 1) - \frac{t_m}{2}\right]^2\\ &\approx \mathrm{Var}\left(X^{(m)}\right) + \frac{2t_m\, m}{(n-1)\,(\lfloor n/m\rfloor - 1)}\sum_{j=1}^{\lfloor n/m\rfloor} j\ X_j^{(m)} +\\ &+ \frac{t_m^2}{4(n-1)^2\,(\lfloor n/m\rfloor - 1)}\sum_{j=1}^{\lfloor n/m\rfloor}(2jm - m - n)^2. \end{aligned} \tag{20}$$

Using the condition that $\lfloor n/m\rfloor$ is large enough, i.e., $m \ll n$ or $n/m \to 0$ and the approximation $\lfloor n/m\rfloor \approx n/m$, *Eq.* (20) can be simplified:

$$\mathrm{Var}\left(X^{L\ (m)}\right) \approx \mathrm{Var}\left(X^{(m)}\right) + \frac{2t_m\, m^2}{(n-1)(n-m)}\sum_{j=1}^{\lfloor n/m\rfloor} j\ X_j^{(m)} + \frac{t_m^2}{12}\frac{(7m-6)}{m}. \tag{21}$$

Finally, by inserting *Eq.* (18) and *Eq.* (21) into *Eq.* (13) we get the following:

$$\begin{aligned}\mathrm{Var}\left(X^{L\ (m)}\right) &\approx \frac{\mathrm{Var}(X^L) - C_1}{m^{2-2H}} + \frac{2t_m\, m^2}{(n-1)(n-m)}\sum_{j=1}^{\lfloor n/m\rfloor} j\ X_j^{(m)} + \frac{t_m^2}{12}\frac{(7m-6)}{m}\\ &= \frac{\mathrm{Var}(X^L) - C_1}{m^{f_L(m)}},\end{aligned} \tag{22}$$

where

$$f_L(m) = \frac{\log\left(\dfrac{\mathrm{Var}(X_L) - C_1}{\dfrac{\mathrm{Var}(X_L) - C_1}{m^{2-2H}} + \dfrac{2t_m\, m^2}{(n-1)(n-m)}\displaystyle\sum_{j=1}^{\lfloor n/m\rfloor} j\ X_j^{(m)} + \dfrac{t_m^2}{12}\dfrac{(7m-6)}{m}}\right)}{\log m}. \tag{23}$$

Eq. (22) shows that the presence of a linear trend in LRD data turns the result of variance-time plot to be quite different from its original form. Plotting $\log\left[\mathrm{Var}(X^{L\ (m)})\right]$ versus $\log m$ instead of a straight line with slope $(2H-2)$ we should observe a curve described by $f_L(m)$, which is a complicated function of m. The estimation of the Hurst parameter of LRD from the variance-time plot should be done by fitting a regression line to the plot at large values of m, so from *Eq.* (22) and using the fact that $X_j^{(m)}$ is close to the sample mean $\bar{X}$ we get:

$$\begin{aligned}
&\frac{2t_m\, m^2}{(n-1)(n-m)} \sum_{j=1}^{\lfloor n/m \rfloor} j\, X_j^{(m)} + \frac{t_m^2}{12} \frac{(7m-6)}{m} \\
&\approx \frac{2t_m\, m^2}{(n-1)(n-m)}\, C_2 \frac{n/m(n/m+1)}{2} + \frac{t_m^2}{12} \frac{(7m-6)}{m} \\
&\approx C_2\, t_m + \frac{7\, t_m^2}{12}, \quad \text{as } m \to \infty,
\end{aligned} \tag{24}$$

where C_2 denotes a constant close to 0. Thus *Eq.* (22) can be rewritten as

$$\mathrm{Var}\left(X^{L\ (m)}\right) \approx \frac{\mathrm{Var}(X^L) - C_1}{m^{2-2H}} + C_2\, t_m + \frac{7\, t_m^2}{12}, \quad \text{as } m \to \infty. \tag{25}$$

Eq. (25) concludes that the variance-time plot of a LRD process with linear trend asymptotically approaches a horizontal line with ordinate $C_2 t_m + 7\, t_m^2/12$, where the constant C_2 is independent of m. The variance-time plots of the LRD process and a process with no LRD are no longer distinguishable in the presence of a linear trend. For more details see our examples in Section 5.

4.3. *R/S Plot of LRD Data with Level Shift*

The R/S analysis of an $\{X_1, X_2, \ldots, X_n\}$ data series is defined by the log-log plot of the *rescaled adjusted range* R/S ratio versus the actual data window size d. For a certain window size d of data the R/S value is given by:

$$\frac{R}{S} = \frac{\max\{W_i;\ i = 1, 2, \ldots, d\} - \min\{W_i;\ i = 1, 2, \ldots, d\}}{\sqrt{\mathrm{Var}(X_{\mathit{off},d})}}, \tag{26}$$

where $X_{\mathit{off},d}$ denotes the considered sub-series $\{X_{\mathit{off}+1,}, X_{\mathit{off}+2}, \ldots, X_{\mathit{off}+d}\}$ and $W_i = \sum_{k=1}^{i} (X_{\mathit{off}+k} - \bar{X}_{\mathit{off},d})$. With a given value of d we calculate several R/S ratios by sliding the window size d throughout the X series. The R/S ratios of LRD data have the following characteristics $R/S \sim C_H d^H$ as $n \to \infty$, where C_H is an infinite positive constant independent of d.

By adding a level shift to the series X, we get the new series denoted by X^{LS}, $X^{LS} = \{X_1, X_2, \ldots, X_{\lfloor n/2 \rfloor - 1}, X_{\lfloor n/2 \rfloor}, X_{\lfloor n/2 \rfloor + 1} + t_m, \ldots, X_n + t_m\}$, where t_m denotes the value of the level shift. According to the definition of the R/S ratio we can observe that this ratio does not change if the data window d does not cover the level shift point. It is simply due to the fact that the k-th element of $X_{\mathit{off},d}$ is $X_{\mathit{off}+k} + C_3$ where C_3 is a constant. More precisely, $C_3 = 0$ if the data window is placed entirely at the first level and $C_3 = t_m$ if it stays entirely at the region of the second level.

Thus we see that $L = \max\{W_i\} - \min\{W_i\} = \max\{W_i + C_3\} - \min\{W_i + C_3\}$, where $i = 1, 2, \ldots, d$, and $S = \sqrt{\mathrm{Var}(X_{\mathit{off},d})} = \sqrt{\mathrm{Var}(X_{\mathit{off},d} + C_3)}$. Therefore the R/S ratio holds its original value.

The situation is different when the data window d contains the jump of the level shift. We concern here the simple case when the location of the shift is placed at the centre of the window:

$$X^{*\,LS}_{\mathit{off}+k} = \begin{cases} X_{\mathit{off}+k} & \text{if } k \le \lfloor n/2 \rfloor \\ X_{\mathit{off}+k} + t_m & \text{if } k > \lfloor n/2 \rfloor \end{cases},$$

where $k = 1, 2, \ldots, d$ and (*) means that it only relates to those d-windows mentioned above. As proven in subsection 4.1,

$$\mathrm{Var}\left(X^{*\,LS}_{\mathit{off},d}\right) \approx \mathrm{Var}\left(X_{\mathit{off},d}\right) + \frac{t_m^2}{4}. \tag{27}$$

Moreover, for the new series

$$\begin{aligned} W_i^{*\,LS} &= \sum_{k=1}^{i} \left[X^{*\,LS}_{\mathit{off}+k} - \left(\bar{X}_{\mathit{off},d} + \frac{t_m}{2} \right) \right] \\ &= \begin{cases} W_i - i\; t_m/2 & \text{if } i = 1, 2, \ldots, \lfloor d/2 \rfloor \\ W_i - (d-i)\; t_m/2 & \text{if } i = \lfloor d/2 \rfloor + 1, \ldots, d \end{cases}. \end{aligned}$$

We compute different R/S values by increasing the window size d and moving this window along the data. There is one window d that contains the shift location at most when d has a small value. However, the change of only one value of the R/S ratio at a fixed d can be counted as a noise and it does not change the look of the plot. In contrast, when, d assumes a large enough value, the following can be justified:

$$\begin{aligned} R^{*\,LS} &= \max\{W_i^{*\,LS}\} - \min\{W_i^{*\,LS}\} = W_d^{*\,LS} - W^{*\,LS}_{\lfloor d/2 \rfloor} \\ &= W_d - W_{\lfloor d/2 \rfloor} + \lfloor d/2 \rfloor \frac{t_m}{2} \approx d\frac{t_m}{4}. \end{aligned} \tag{28}$$

Moreover, as d is large,

$$S^{*\,LS} = \sqrt{\mathrm{Var}\left(X^{*\,LS}_{\mathit{off},d}\right)} \approx \sqrt{\mathrm{Var}\left(X_{\mathit{off},d}\right) + \frac{t_m^2}{4}} \approx \sqrt{\mathrm{Var}(X_n) + \frac{t_m^2}{4}}. \tag{29}$$

Therefore,

$$\left(\frac{R}{S}\right)^{*\,LS} \approx \frac{d\dfrac{t_m}{4}}{\sqrt{\mathrm{Var}(X_n) + \dfrac{t_m^2}{4}}} = d\; C_4, \tag{30}$$

where C_4 is a constant independent of d. These points create a separate part on the log-log plot which should be placed closely around a straight line with slope 1. The other large cluster of points remains at the same place as before adding the level shift and this part of the R/S plot of LRD data with level shift looks similar to the R/S plot of the original LRD data.

This result shows that the R/S plot can also be used for detection of level shifts in the data. Moreover, the linear part with slope 1 in the plot should be disregarded in the estimation of Hurst parameter of LRD processes. In this way, in the cases when this separation is feasible, we can make a reliable estimate of H even in the presence of level shifts.

5. Simulations

5.1. Setup

Reference data sets. A sample series of Fractional Gaussian Noise (FGN) [2] was used as a reference for data exhibiting LRD. In this generated set the Hurst parameter was set to be 0.7. The other reference set is generated by Poisson process. In order to make a good comparison, these data sets were set to have the same mean and variance of the value 10. Both data sets have the same size of 32,768 data.

Measured SUNET ATM cell-traffic. A series of ATM cell arrivals obtained from a real-time traffic measurement on the Swedish University NETwork (SUNET) [16] was also analyzed. Data traces were collected in 1996 based on a custom-built measurement tool which is able to record more than 8 million consecutive cell arrivals. In our tests the traces of the number of cell arrivals in a 1 ms time window were considered. The analysis of these data traces can illustrate the non-stationary effects in LRD estimation of real traffic.

Types of non-stationarities. There are two typical classes of non-stationarities observed in real traffic data: the level shift and the polynomial trends. In our simulation study we only concerned three simple cases: *level shift with two states*, *linear trend*, and *parabolic trend*. These effects were added to both data sets (see *Fig. 4* for the detailed information of these non-stationarities). We denote by 0.7-FGN the original FGN set, by 0.7-FGN_L the FGN set with linear trend, by 0.7-FGN_P the one with parabolic trend and by 0.7-FGN_LS the one with level shift. The Poisson sets are marked with the same notations: POISS, POISS_L, POISS_P, POISS_LS. *Table 1* gives more information about these data sets.

The datagram of these data sets can be seen in *Fig. 1*.

5.2. Empirical Results

Variance-time plot. Results of variance-time analysis can be seen in *Fig. 2*. The variance-time plot estimated on the original data sets, 0.7-FGN and POISS, gives us

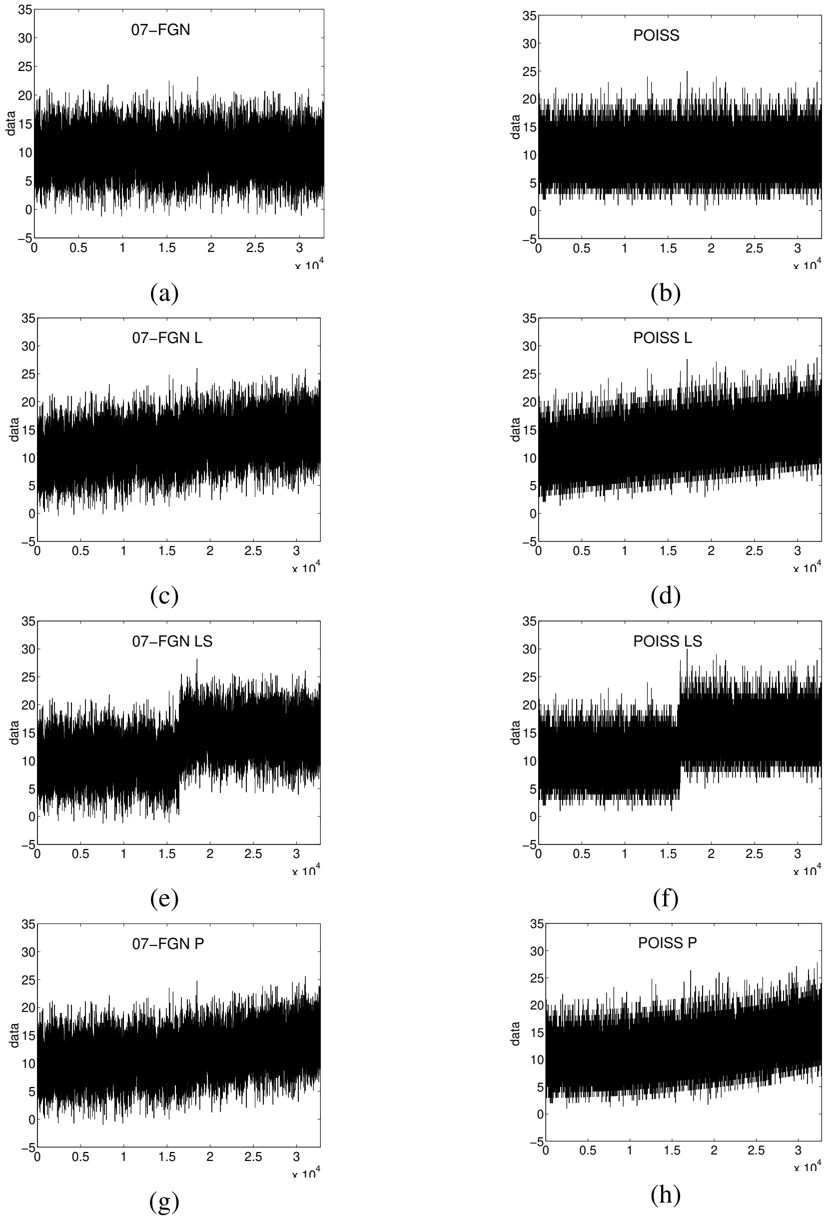

Fig. 1. The datagrams of the investigated data sets. The 0.7-FGN means m + FGN(σ^2, H), where $m = 10$, $\sigma^2 = 10$, $H = 0.7$; the POISS means Poissonian samples with $\lambda = 10$ ($\sigma^2 = 10$).

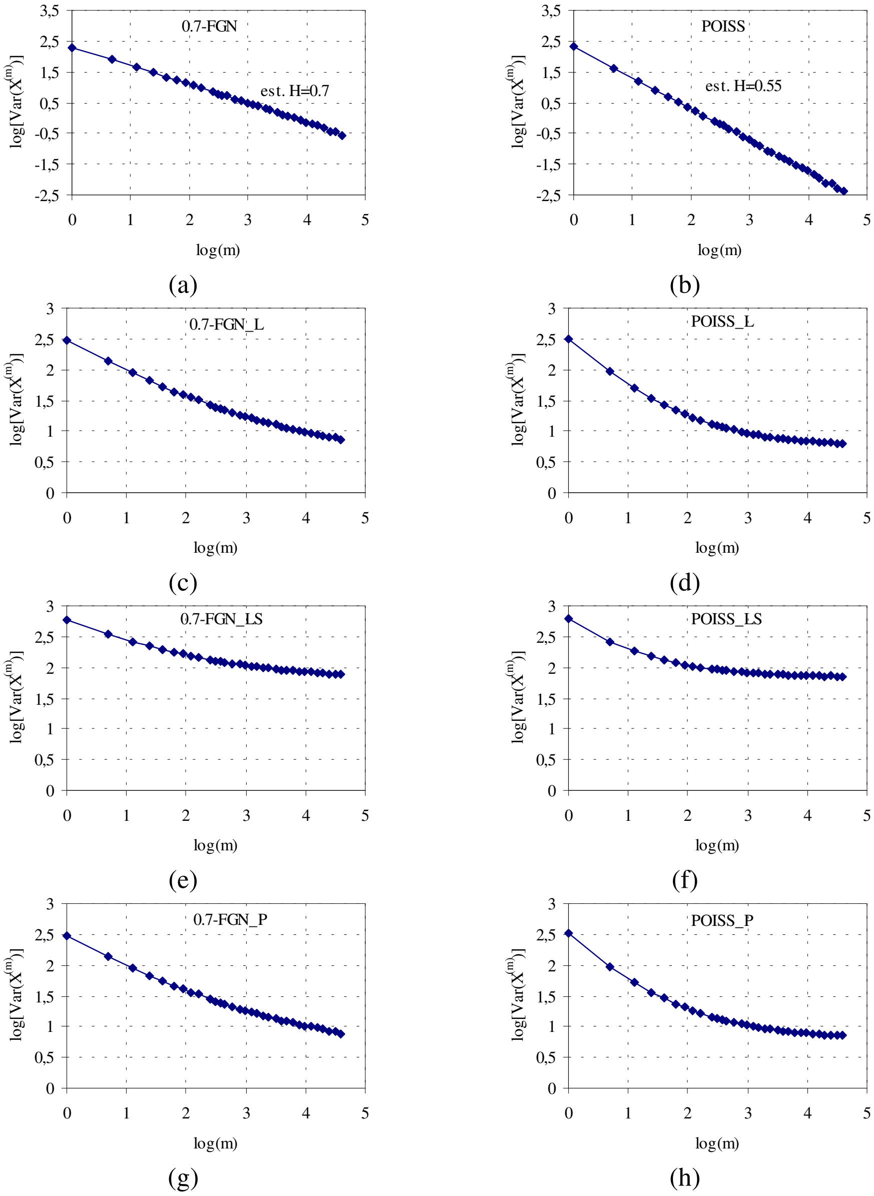

Fig. 2. The variance-time plots. Left: variance-time plots of 0.7-FGN, 0.7-FGN_L, 0.7-FGN_LS, and 0.7-FGN_P. Right: variance-time plots of POISS, POISS_L, POISS_LS, and POISS_P.

the exact values of Hurst parameter we expected: $H = 0.7$ for 0.7-FGN data set and

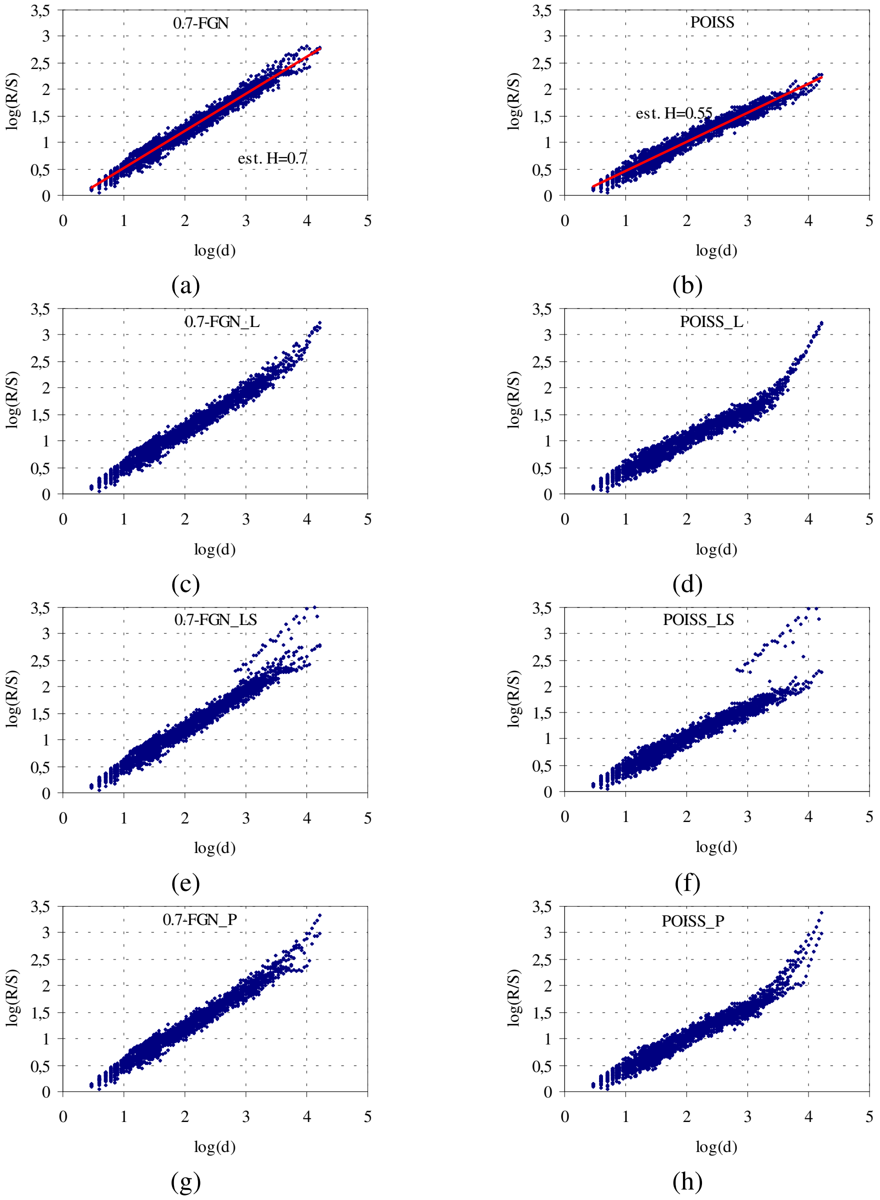

Fig. 3. The R/S plots. Left: R/S plots of 0.7-FGN, 0.7-FGN_L, 0.7-FGN_LS, and 0.7-FGN_P. Right: R/S plots of POISS, POISS_L, POISS_LS, and POISS_P.

$H = 0.5$ for POISS one (see *Figs. 2.a* and *2.b*). However, variance-time plot is very sensitive to the investigated non-stationarities. As seen in *Figs. 2.c, 2.d, 2.e, 2.f,*

Table 1. The detailed information of investigated data sets ($\hat{\mu}$ and $\hat{\sigma}^2$ denote the sample mean and the sample variance, respectively).

Data sets	t_m	$\hat{\mu}$	$\hat{\sigma}^2$	Data sets	t_m	$\hat{\mu}$	$\hat{\sigma}^2$
0.7-FGN	-	10	10	POISS	-	10	10
0.7-FGN_L	5	12.5	11.81	POISS_L	5	12.5	12.19
0.7-FGN_LS	5	12.5	16.02	POISS_LS	5	12.5	16.35
0.7-FGN_P	5	11.66	11.89	POISS_P	5	11.71	12.33

2.g, and *2.h*, the variance-time plots are all convex curves and a careful observation reveals that no linear parts are found in these plots. Moreover, comparing *2.c* with *2.d*, *2.e* with *2.f*, and *2.g* with *2.h*, there are no significant differences between the variance-time plots of the data sets of the FGN with trends and level shift and the Poisson with trends and level shift, respectively.

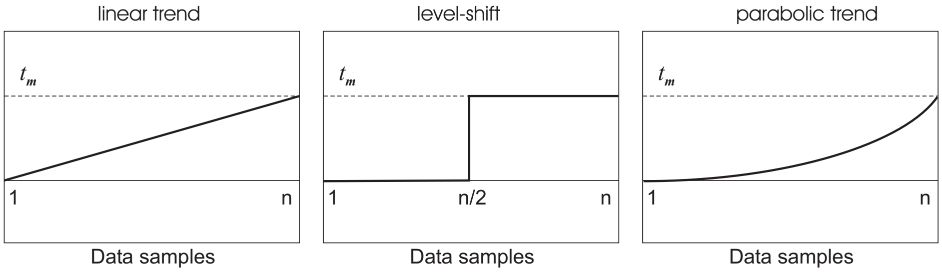

Fig. 4. Types of non-stationarities: linear trend (left), level shift (middle), and parabolic trend (right). The value of t_m is set to be 5 in each case.

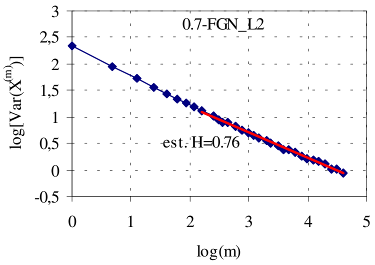

Fig. 5. The variance-time plot of the 0.7-FGN_L2 set ($t_m = 2.5, \hat{\mu} = 11.25, \hat{\sigma}^2 = 10.38$)

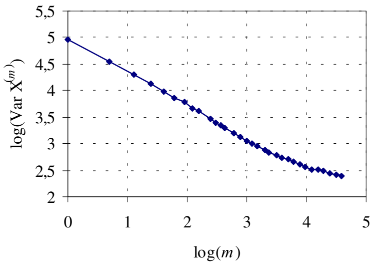

Fig. 6. The variance-time plot of the SUNET ATM data

We also made the variance-time plots for the 0.7-FGN data set with a smaller linear trend (the value of t_m of this linear trend is 2.5, see *Fig. 5*). In this case the plot of the new series 0.7-FGN_L2 seems to be linear which is tempting to make an estimate but the estimated value of H is 0.76, which is far from the real value. It means that getting an estimate from the linear part (at large time-scales) of the variance-time plot (as usually done in practical analysis) can produce misleading results.

As an example, we show the result of variance-time analysis of the SUNET ATM data presented in *Fig. 6*. The measured ATM traffic is bursty in nature and although several pre-processing procedures were done in this trace it is difficult to detect a certain trend. However, the curve is very similar to those obtained with level shift or trends in *Figs. 2.e* and *2.f*. As discussed above to avoid misleading results estimation of H cannot be applied in this case.

These simulation results confirm our analytical results and also show that short-range dependent (SRD) processes with non-stationarities can produce the same variance-time plot as LRD processes. Moreover, in the case of LRD processes trends can significantly destroy the accuracy of the estimation of the H parameter.

R/S plot. *Fig. 3* shows the results of the R/S tests. The effects of the linear and the parabolic trend are revealed in the rise of the upper tail of the R/S plots (see *Figs. 3.c, 3.d, 3.e*, and *3.f*). However, if we extract this part from the plot, the linear rest of the plot shows the exact slope which is seen in the R/S plot of the original data sets, *Figs 3.a* and *3.b*.

The interesting results are found in the plots of data sets with level shift, see *Figs. 3.e* and *3.f*. On the one hand, these plots seem to be constructed from two parts which are independent of each other. The lower parts look almost like the R/S plot of the original sets as comparing *3.e* with *3.a* and *3.f* with *3.b*. On the other hand, the upper parts are nearly the same in both plots. So we assumed that the lower parts belong to the original data sets and the upper parts are due to the level shift. We applied the R/S plot for the level shift only and our assumption seems to be justified. *Fig. 7*, which is the R/S plot of the level shift contains the upper part only. Our results are in good agreement with our analytical investigation presented in Section 4. We can see that the presence of level shifts can be revealed by R/S plot and the reliable estimation of H parameter for LRD processes with level shifts is possible if the cluster produced by the level shift is separable.

An illustrative example from practice of such an effect can be seen in the R/S plot of the SUNET ATM series (*Fig. 8*). The plot contains a breakpoint where the slope of the curve changes approximately in the middle of the figure. If one tries to estimate H from the upper part of the plot it will result in a wrong value as we demonstrate it in the following. *Fig. 9* shows the R/S plot of a subset of the SUNET ATM set. The subset is gained from the original set after erasing some suspected non-stationary parts of the data selected by our stationarity analysis [19]. In *Fig. 9* the part of the curve with the higher slope disappeared and the lower part continues growing nearly as a straight line. An explanation of this phenomenon is the possible presence of several local level shifts in the original SUNET ATM data. The result also demonstrates that the important part for LRD parameter estimation is distorted by level shifts.

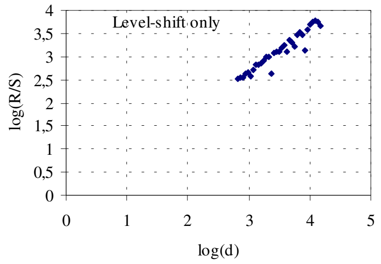

Fig. 7. The R/S plot of a pure level shift.

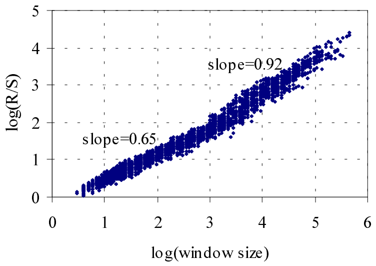

Fig. 8. The R/S plot of the SUNET ATM data series

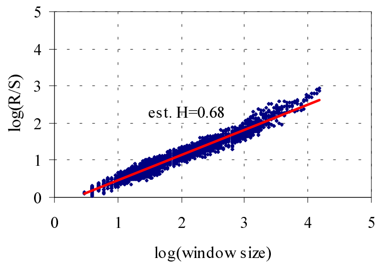

Fig. 9. The R/S plot of the 'stationary' subset of the SUNET ATM data set

Periodogram plot. In the frequency domain, adding trend to data produces the increase of low frequency components. Thus we were not surprised when observing the rise of the lower tail of the periodograms under the influence of different trends (see *Figs. 10.c, 10.d, 10.e, 10.f, 10.g*, and *10.h*). This affects the periodogram in both cases (the 0.7-FGN sets and the POISS sets) and the estimation of the Hurst parameter. Besides, the periodogram plot of the FGN with trend can be easily confused with the Poisson case.

Since the periodogram at low frequencies should be counted for estimation of the Hurst parameter, the presence of trends in LRD data destroys the testing and estimating capability of the periodogram plot.

Wavelet based estimator. We investigated the LRD test based on the wavelet transformation (which is also called the Logscale diagram). As shown in [1], the influence of polynomial trends on this kind of LRD test can be avoided by an adequate choice of the vanishing moment N of the wavelet function. Our empirical work has justified this observation. In *Fig. 11* we see that the Logscale Diagrams give the robust estimate of the Hurst parameter around 0.72 for the 0.7-FGN sets and 0.5 for POISS sets independently of the presence and of the type of trends. Moreover, our simulation also shows that the LD is still robust in the presence of level shifts. As seen in *Figs. 11.e* and *11.f* the level shift added to the 0.7-FGN and the POISS sets slightly changes the result: the estimation of H is 0.72 with confidence interval (0.71, 0.73) for the 0.7-FGN_LS set and 0.5 with confidence interval (0.49, 0.51) for POISS_LS set.

6. Conclusions

Based on both analytical and simulation studies and examples from measured traffic we have shown that the presence of different non-stationarities (level shifts, linear and polynomial trends) in the data can deceive several detecting and estimating methods of LRD.

These effects result in poor estimates of the Hurst parameter in the case of the variance-time plot and periodogram. Moreover, the estimated results can be confused with processes having short-range dependence with non-stationary effects. We suggest that the variance-time plot and the periodogram methods should not be used without a stationarity analysis.

We have found that the wavelet based method (the Logscale diagram) provides a very robust estimation of H. Its estimation results are almost independent of the presence of the investigated trends and level shifts. The R/S analysis was also found to be a robust estimator of the Hurst parameter of LRD processes. In addition, we have also demonstrated that the level shift can be detected by the R/S analysis therefore this method can also be well utilized in stationarity analysis.

We recommend the use of the R/S plot and the Logscale diagram for the estimation of Hurst parameter of LRD processes in the possible presence of the investigated non-stationarities.

Acknowledgement

Authors would like to thank István Maricza for his comments on the paper and Darryl Veitch for the code of the wavelet-based method.

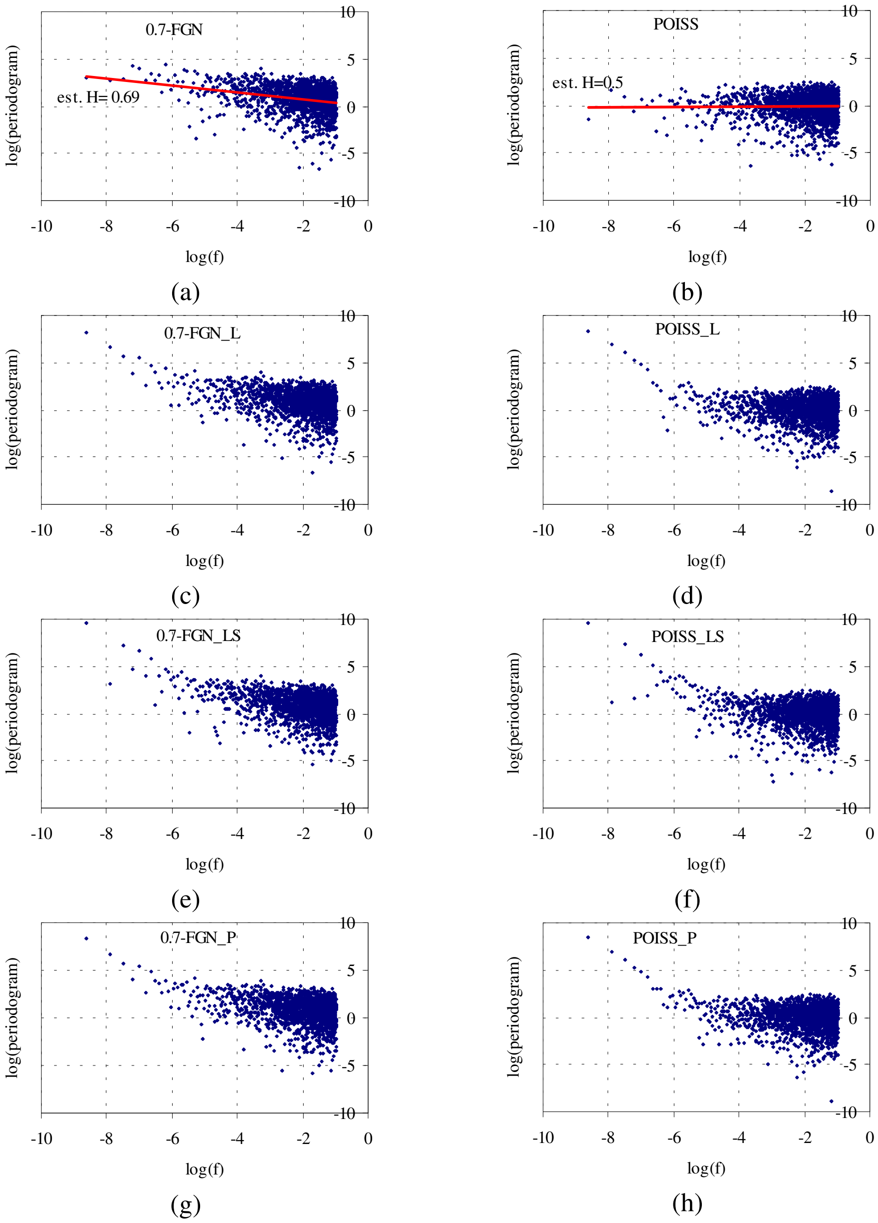

Fig. 10. The periodogram plots. Left: periodogram plots of 0.7-FGN, 0.7-FGN_L, 0.7-FGN_LS, and 0.7-FGN_P. Right: periodogram plots of POISS, POISS_L, POISS_LS, and POISS_P.

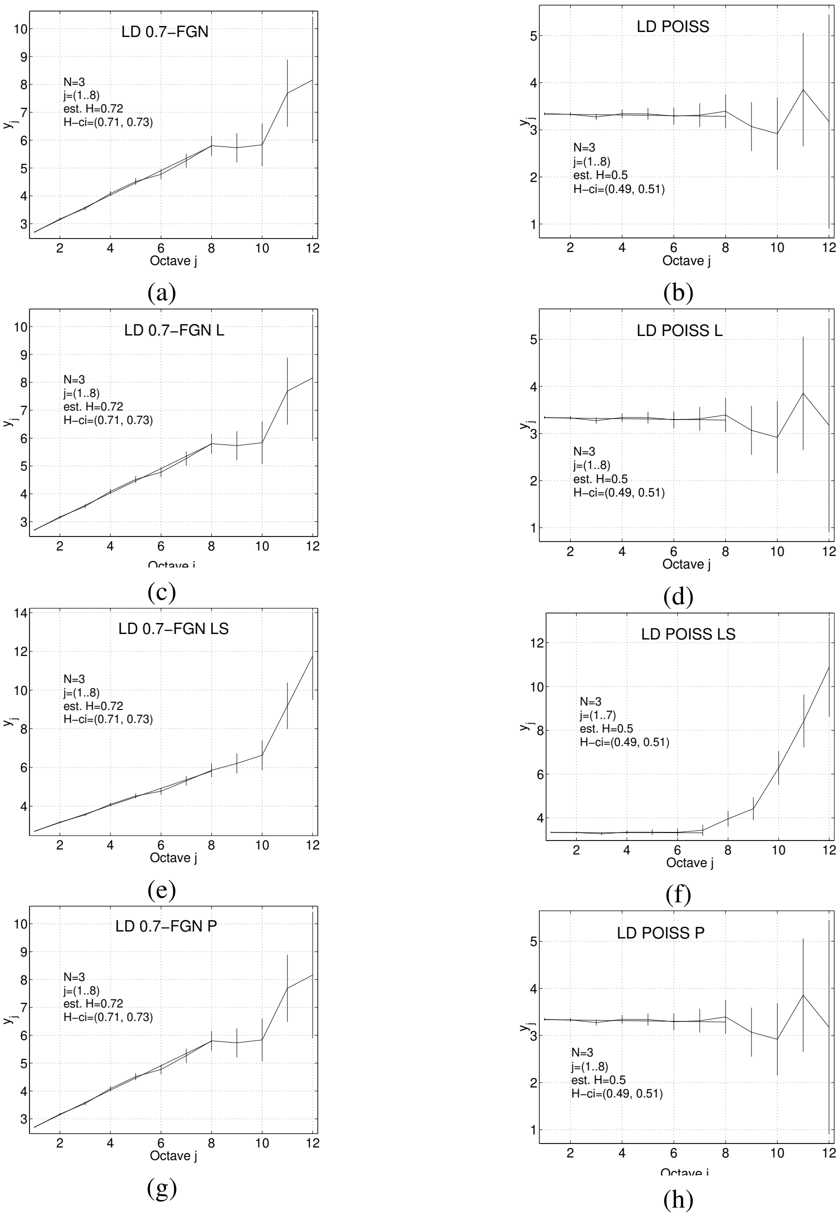

Fig. 11. The Logscale diagrams. Left: Logscale diagrams of 0.7-FGN, 0.7-FGN_L, 0.7-FGN_LS, and 0.7-FGN_P. Right: Logscale diagrams of POISS, POISS_L, POISS_LS, and POISS_P.

References

[1] ABRY, P. – VEITCH, D: Wavelet Analysis of Long Range Dependent Traffic. *IEEE Trans. Inform. Theory*, **44** (1) (1998), pp. 2–15.

[2] BERAN, J.: *Statistics for Long-Memory Processes*. Chapman & Hall, One Penn Plaza, New York, NY 10119, 1995.

[3] BOLOTIN, V. – COOMBS-REYES, J. – HEYMAN, D. – LEVY, Y. – LIU, D.: IP Traffic Characterization for Planning and Control. In P. Key and D. Smith, editors, *Proc. ITC 16*, pp 425–436. Elsevier Science B. V., 1999.

[4] BROCKWELL, P. J. – DAVIS, R. A.: *Time Series: Theory and Methods*. Springer, 1996.

[5] COX, D. R. – LEWIS, P. A. W.: *The Statistical Analysis of Series of Events*. Methuen, 1966.

[6] DUFFIELD, N. G. – LEWIS, J. T. — O'CONNELL, N. – RUSSEL, R. – TOOMEY, F.: Statistical Issues Raised by the Bellcore Data. In *11th Teletraffic Symposium*, Cambridge, 23-25 March 1994.

[7] DUFFIELD, N. G. – MASSEY, W. A. – WHITT, W.: A Nonstationary Offered-Load Model for Packet Networks. In *Sel. Proc. of the 4th INFORMS Telecomm. Conf.*, 1999.

[8] ERRAMILLI, A. – PRUTHI, P. – WILLINGER, W.: Self-Similarity in High Speed Network Traffic Measurements: Fact or Artifact? In *Proc. of the 12th Nordic Teletraffic Seminar, NTS12*, Espoo, Finland, 22-24 August 1995.

[9] GRASSE, M. FRATER, M. R. – ARNOLD, J. F.: Implications of Non-Stationarity of MPEG2 Video Traffic. Technical report, COST 257 TD(97)01, January 1997.

[10] HURST, H. E.: Long-Term Storage Capacity of Reservoirs. *Proc. Amer. Soc. Civil Eng.*, **76** (1950), (11).

[11] JAGERMAN, D. L. – MELAMED, B. – WILLINGER, W.: Stochastic Modeling of Traffic Processes. In *Frontiers in Queueing*, pp. 271–370. CRC Press, 1997.

[12] JORMAKKA Jorma: On Self-Similar Models for ATM Traffic. In *ITC Specialist Seminar on Control in Communications*, pp. 277–288, 1996.

[13] KLEMES, W.: The Hurst Phenomenon: A Puzzle? *Water Resources Research* **10**, pp. 675-688.

[14] KUNSCH, H.: Discrimination between Monotonic Trends and Long-Range Dependence. *J. Appl. Prob.*, **23** (1986).

[15] MANDELBROT, B. B. – VAN NESS, J.W.: Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Rev.*, **10** (1968), pp. 422–437.

[16] MOLNÁR, S. – MARICZA, I. eds.: Source Characterization in Broadband Networks. Interim report, COST 257, Vilamoura, Portugal, January 1999.

[17] MOLNÁR, S. – GEFFERTH, A.: On the Scaling and Burst Structure of Data Traffic. In *8th Int. Conference on Telecommunication Systems, Modelling and Analysis*, Nashville, Tennessee, USA, 9-12 March 2000.

[18] MOLNÁR, S. – VIDÁCS, A. – NILSSON, A. A.: Bottlenecks on the Way towards Fractal Characterization of Network Traffic: Estimation and Interpretation of the Hurst Parameter. In *Proc., PMCCN'97*, pp. 125–144, Tsukuba, Japan, 1997.

[19] ÉLTETŐ, T. – MOLNÁR, S.: On the Distribution of Round-trip Delays in TCP/IP Networks. In *The 24th Annual Conference on Local Computer Networks (LCN'99)*, Lowell/Boston, MA, October 1999.

[20] ROUGHAN, M. – VEITCH, D.: Measuring Long-Range Dependence under Changing Traffic Conditions. Extended version of paper in: Proc. Infocom'99, Manhattan, April 1999.

[21] TEVEROVSKI, V. – TAQQU, M.: Testing for Long-Range Dependence in the Presence of Shifting Means or Slowly Declining Trends, using Variance Type Estimator. *J. of Time Series analysis*, **18** (3) (1997), pp. 279–304.

[22] TSYBAKOV, B. – GEORGANAS, N. D.: On Self-Similar Traffic in ATM Queue: Definitions, Overflow Probability Bound, and Cell Delay Distribution. *IEEE/ACM Trans. on Networking*, **5** (3) (1997), pp. 397–409.

[23] VATON, S. – MOULINES, E.: A Locally Stationary Semi-Markovian Representation for Ethernet Lan Traffic Data. In *IFIP TC6/WG6.2 4th Int. Conference on Broadband Communications*, Stuttgart, Germany, 1-3 April 1998.

[24] VATON, S. – MOULINES, E. – KOREZLIOGLU, H. – KOFMAN, D.: Statistical Identification of Lan Traffic Data. In *ATM97, 5th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks*, Ilkley, UK, 21-23 July 1997.

[25] WALLIS, J. R. – O'CONNELL, P. E.: Firm Reservoir Yield: How Reliable are Historic Hydrologic Records? *Hydrol. Sci. Bull.*, (1973), pp. 347–365.

[26] WILLINGER, W. – TAQQU, M. – ERRAMILLI, A.: A Bibliographical Guide to Self-Similar Traffic and Performance Modeling for Modern High-Speed Networks Stochastic Networks: Theory and Applications. In *Royal Statistical Society Lecture Notes Series*, Vol. 4. Oxford University Press, 1996.