WAVELET-BASED ESTIMATION OF LONG-RANGE DEPENDENCE IN VIDEO AND NETWORK TRAFFIC TRACES

by

Nikola Cackov

Dipl. Ing., SS. Cyril and Methodius University, Skopje, Macedonia, 2001

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF APPLIED SCIENCE
in the School
of
Engineering Science

SIMON FRASER UNIVERSITY
Spring 2005

APPROVAL

Name: Nikola Cackov

Degree: Master of Applied Science

Title of thesis: Wavelet-based estimation of long-range dependence in video and network traffic traces

Examining Committee:

Dr. Jie Liang, Assistant Professor,
Simon Fraser University
Chair

Dr. Ljiljana Trajković, Professor,
Simon Fraser University
Senior Supervisor

Dr. William Gruver, Professor,
Simon Fraser University
Supervisor

Dr. Stephen Hardy, Professor,
Simon Fraser University
Examiner

Date Approved: 15.12.2004

SIMON FRASER UNIVERSITY

PARTIAL COPYRIGHT LICENCE

Abstract

Correct and efficient estimation of the Hurst parameter H of long-range dependent (LRD) traffic is important in traffic analysis. The low computational cost and the wavelets' scale invariance make wavelet transform suitable for analysis of LRD processes. In this thesis, we apply wavelet-based estimation of H to MPEG-1 and MPEG-4 encoded video sequences. Frequency-domain estimators (periodogram and wavelet-based) produce different Hurst parameters compared to time-domain estimators (R/S and variance-time plot). Wavelet-based estimators often produce Hurst parameters that are close to or greater than one. Our analysis indicates that a possible cause for the unreliable performance of the wavelet-based estimators is the non-stationarity of the scaling exponent. We also apply the monofractal wavelet-based estimator to traces of call holding and call inter-arrival times collected from a circuit-switched cellular wireless network. We test the time constancy of the scaling exponent α and compare the estimates of H from various time periods.

На мама Еми, тато Данчо и дада Ана

To my dear parents Emilija and Jordan and my sister Ana

"There are three kinds of lies: lies, damned lies and statistics"

— Benjamin Disraeli, statesman.

Acknowledgments

No work is ever individual and finished solely by only one person. All along the way, there are people who give their help and support, be it academic, financial, or emotional. Many such persons come to my mind right now and I apologize to everyone whose name will not be mentioned explicitly. One great thanks to you all!

I express my deepest gratitude to my senior supervisor, Professor Ljiljana Trajković, for accepting me in the Communication Networks Laboratory and enabling me to work in a really productive environment. Thank you for your guidance, insight, trust and above all, your patience and warmheartedness.

I would like to thank Dr. William Gruver and Dr. Stephen Hardy for serving on my examining committee and Dr. Jie Liang for chairing the thesis defense.

I would not have accomplished my goals had there not been all the love and support from my family. Although miles away from me, they were always by my side to comfort me or share my joys. Thank you very much for standing behind all my ideas. And Ana, keep up the positive attitude, you have always been my role model for optimism.

My special thanks go to my dear professors and friends Dr. Sofija and Dr. Momčilo Bogdanovi, for all their kindness and confidence in me. Without your advice and

recommendations, I would have never had this wonderful experience.

Finally, my sincere thanks go to my colleagues in the Communication Networks Laboratory: Božidar and Svetlana Vujičić, Vladimir Vukadinović, Nenad Lasković, Hui Grace Zhang, Renju Narayanan, André Dufour, Jiaqing James Song, Hao Johnson Chen, Qing Kenny Shao, Savio Lau, Hao Leo Chen, and Dongliang Tony Feng. Thank you folks for being such wonderful friends.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Analysis of traffic in communication networks is important for determining their operational status. Furthermore, it is a step toward traffic modelling, which is necessary for predicting network resources utilization and provisioning, and for planning future network developments. As networks evolved from the first manually operated telephone networks to today's high-speed Internet, so did the models of the traffic carried by those networks. The well-known Erlang models, derived and appropriate for telephone traffic, were popular due to their mathematical tractability [1]. They are based on a Poisson arrival process, with independent and exponentially distributed event-arrival times. However, in the first half of the '90s, studies of traffic in packet data networks showed that the network traffic exhibits self-similarity and long-range dependence. In their seminal work, Leland et al., [2], detected self-similarity in the aggregate backbone network traffic and linked it to the aggregation of ON/OFF traffic sources with heavy-tailed ON and OFF periods. Subsequent studies of traffic from various protocols showed that not only aggregate traffic, but also traffic from FTP or TELNET

protocols separately exhibits long-range dependence [3]. Therefore, Poisson processes proved unsuitable for modelling traffic in packet networks. Furthermore, Poisson-based Erlang models may not be applicable even in certain circuit-switched networks due to the presence of long-range dependence in the call inter-arrival times [4].

1.1 Long-range dependence in MPEG video traces

Multimedia network applications, such as video streaming and video conferencing, have gained popularity in the past years. Video traffic possesses two major characteristics: high bandwidth requirements and high variability [5]. In order to model and analyze the impact of the network performance on the quality of the received video, it is important to develop adequate models for the video sources. Analyses of the statistical properties of video sequences compressed by employing various encoding algorithms (MPEG-1, MPEG-4, H.263, and several proprietary algorithms) [6]–[10], have shown that long-range dependence is an inherent property of the video traffic.

Long-range dependent processes are characterized by the Hurst parameter H. Its estimation is a necessary first step in modelling traffic. There are several estimators for H, such as R/S plot, variance-time plot, periodogram, Whittle, and wavelet-based [11]–[13]. Wavelet-based estimator is considered to be unbiased and robust with respect to presence of deterministic trends in the analyzed process [13], [14]. However, several studies [6], [15], [16], have indicated that the wavelet-based estimator leads to $H > 1$ when applied to video traces, which contradicts the findings that the traces were long-range dependent. In this thesis, we attempt to determine the source of such behaviour of the estimator.

1.2 Long-range dependence in circuit-switched radio networks

Network traffic in circuit-switched radio networks is often modelled by the Erlang-C model [17], [18]. The model implies independent call holding and call inter-arrival times. A recent study [4] of the traffic collected from a public safety radio network operated by E-Comm [19] indicated that, while call holding times are indeed independent, call inter-arrival times exhibit certain degree of long-range dependence. The analysis was performed on the traffic data from three busy hours. In this thesis, we extend the study [4] by examining longer traces from various time periods and we investigate whether or not the assumption of independence of the call holding and call inter-arrival times holds.

1.3 Organization of the thesis

In Chapter 2, we first examine the implications of non-degenerate correlations of a stochastic process to its statistical properties, in particular, to the variance of its sample mean. Long-range dependence is introduced as a special case of correlation structure and its main characteristics are presented. Very often the term *long-range dependence* is identified with *self-similarity*. We try to avoid this misconception by following the discussion in [20] and by separately introducing self-similarity. Finally, we discuss the relationship between long-range dependent processes and self-similar processes, characterized by the Hurst parameter.

Chapter 3 provides an overview of wavelets, the wavelet transform, and the application of the discrete wavelet transform in analyzing long-range dependent processes. We begin by defining the continuous wavelet transform and its reduction to discrete wavelet transform and its practical implementation. We then introduce the monofractal and multifractal wavelet-based estimators of the Hurst parameter. Finally, we present the basic concepts of the test for time constancy of the scaling exponent α.

Chapter 4 presents the results of applying the wavelet-based estimator of H to MPEG-1 and MPEG-4 encoded video sequences. We compare the wavelet-based estimates of H with those obtained from other estimators (periodogram and R/S plot). Furthermore, we examine the source of unreliability of the estimates of H by testing the Gaussianity of the wavelet coefficients and the time constancy of α.

In Chapter 5, we present the wavelet-based estimates of the Hurst parameter for weekly, daily, and hourly traces of call holding and call inter-arrival times collected from the E-Comm network. In addition, we perform the test for time constancy of α in order to determine whether or not the estimates of H are reliable. We also compare the estimates of H for traces collected over three years. The chapter is organized chronologically by the year of the analyzed traces.

Finally, Chapter 6 concludes this thesis by outlining the most important findings and by giving directions for possible future research.

Chapter 2

Long-range dependence and Hurst parameter

2.1 Long-range dependence

Sample mean and variance of the sample mean are important quantities that characterize a discrete stochastic process X with a mean value μ. A well-known result often used in practice states that the variance of the sample mean, $\mathrm{var}\,(\bar{X})$, is inversely proportional to the sample size n [11]:

$$\mathrm{var}\,(\bar{X}) = \frac{\sigma^2}{n}, \tag{2.1}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean and $\sigma^2 = \mathrm{E}\{(X_i - \mu)^2\}$ is the variance of the observations X_i. One important assumption made in order to derive Eq. (2.1) is that the samples $X_1, X_2, \ldots, X_n$ are uncorrelated.

Treating X's as uncorrelated may not be always justified. Very often, the underlying processes have correlation structures that cannot be neglected. Let $\rho(i,j)$ be

the autocorrelation between X_i and X_j, given by

$$\rho(i,j) = \frac{\mathrm{E}\{(X_i - \mu)(X_j - \mu)\}}{\sigma^2}. \tag{2.2}$$

Then, for the general case of non-degenerate correlations between the samples, the expression for the sample variance is

$$\begin{aligned} \mathrm{var}\,(\bar{X}) &= n^{-2}\sigma^2 \sum_{i,j=1}^{n} \rho(i,j) \\ &= \sigma^2 \left[1 + n^{-1} \sum_{i \neq j} \rho(i,j)\right] n^{-1} \\ &= \sigma^2 \left[1 + \delta_n(\rho)\right] n^{-1}. \end{aligned} \tag{2.3}$$

Equation (2.3) is similar to Eq. (2.1), the only difference being the introduced correction term $[1 + \delta_n(\rho)] = c_n(\rho)$. This correction term is a result of the correlation structure that exists among the samples. Stochastic processes may exhibit three types of behaviour:

1. If the samples are uncorrelated, then $\delta_n(\rho) = 0$ and Eq. (2.3) is identical to Eq. (2.1).

2. If $\delta_n(\rho) \neq 0$, but the limit $\delta(\rho) = \lim_{n\to\infty} \delta_n(\rho)$ exists and is finite and greater than -1, then the variance of the sample mean $\mathrm{var}\,(\bar{X})$ is still proportinal to n^{-1}. For large sample sizes ($n \to \infty$), $\mathrm{var}\,(\bar{X}) \approx \sigma^2 \left[1 + \delta(\rho)\right] n^{-1}$. In this case, Eq. (2.3) differs from Eq. (2.1) only by the multiplicative constant factor $[1 + \delta(\rho)]$.

3. For certain processes, the variance of the sample mean differs from Eq. (2.1) not only by a constant term, but also by the speed of convergence to zero. For finite time series, this behavoiur can still be modelled by introducing a constant

multiplicative factor. However, the factor will increase with the increasing of the sample size. In general, a more elegant and simple way to account for the slower convergence of var $(\bar{X})$ to zero is to model it as a power-law decay:

$$\text{var}\,(\bar{X}) \approx \sigma^2 c(\rho) n^{-(1-\alpha)} \tag{2.4}$$

where α is a constant such that $0 < \alpha < 1$ and $c(\rho)$ is defined as

$$c(\rho) = \lim_{n\to\infty} n^{-(1+\alpha)} \sum_{i\neq j} \rho(i,j). \tag{2.5}$$

We now examine the third case more closely, making one additional assumption. Let the process X be a wide-sense stationary stochastic process. This implies that its mean and variance are constant and that its autocorrelation function $\rho(i,j)$ depends only on the lag $k = |i-j|$, i.e., $\rho(i,j) = \rho(k)$. It has been shown [11] that for a sample size n

$$\sum_{k=-(n-1)}^{n-1} \rho(k) \approx K n^{\alpha}, \tag{2.6}$$

where K is a positive constant. Since $\alpha > 0$,

$$\sum_{k=-\infty}^{\infty} \rho(k) \to \infty. \tag{2.7}$$

In other words, the autocorrelation function decays so slowly that the sum of all autocorrelation coefficients is infinite. Wide-sense stochastic processes for which (2.7) holds are called *long-range dependent processes* [11], [20]. For large lags k, $\rho(k)$ is modelled as a hyperbolically (power-law) decaying function

$$\rho(k) = c_\rho\, |k|^{-(1-\alpha)}\,, \quad k \to \infty, \tag{2.8}$$

where c_ρ is a positive constant.

Long-range dependent processes are wide-sense stationary. Their power spectral density function can be calculated as a Fourier transform of the autocorrelation function. Therefore, the condition for long-range dependence of a signal X, given by Eq. (2.8), can be expressed in terms of the power spectrum. The power spectral density (PSD) $f(\nu)$ of X satisfies

$$f(\nu) = c_f \, |\nu|^{-\alpha}, \quad |\nu| \to 0, \tag{2.9}$$

where c_f is a positive constant. We call α a *scaling exponent* [21], [22]. Hence, a power-law decay to zero of the autocorrelation function for large lags k implies a power-law behaviour of the PSD for low frequencies and a pole at zero.

2.2 Self-similarity and Hurst parameter

Let Y_t be a stochastic process. Y_t is called *self-similar with self-similarity parameter* H if, for any positive value c, the process $c^{-H}Y_{ct}$ is identical in distribution to the original process Y_t. This implies that for any sequence $t_1, t_2, \ldots, t_k$, the sequence $c^{-H}(Y_{ct_1}, Y_{ct_2}, \ldots, Y_{ct_k})$ has the same distribution as $(Y_{t_1}, Y_{t_2}, \ldots, Y_{t_k})$. The self-similarity parameter H is also called the Hurst parameter. Processes that are self-similar appear similar regardless of the timescale on which they are observed and analyzed. There are known examples of such processes, both in nature (yearly minimal water levels of the Nile river) and in computer communications (Bellcore Etherenet traces). Furthermore, if, for any k points $t_1, t_2, \ldots, t_k$, the distribution of $(Y_{t_1+c} - Y_{t_1+c-1}, Y_{t_2+c} - Y_{t_2+c-1}, \ldots, Y_{t_k+c} - Y_{t_k+c} - 1)$ does not depend on c, then the process Y_t is called *self-similar with stationary increments*. This type of process is important for modelling data that seem stationary.

Let Y_t be a self-similar process with stationary increments $X_i = Y_i - Y_{i-1}$, $i = 1, 2, 3, \ldots$. The autocorrelation function of the process X_i has the form

$$\rho(k) = \frac{1}{2}\left[(|k|+1)^{2H} - 2|k|^{2H} + (|k|-1)^{2H}\right] \tag{2.10}$$

where k is the lag and H is the Hurst parameter. The process X_i is also called *exactly second-order self-similar* [20]. For large lags $k \to \infty$, the behaviour of $\rho(k)$ can be described as

$$\rho(k) \approx H(2H-1)k^{2H-2}. \tag{2.11}$$

For $1/2 < H < 1$, Eq. (2.7) holds and the process X_i is long-range dependent. If $H = 1/2$ then the samples of the process X_i are uncorrelated. For values of H between 0 and $1/2$, the sum of the autocorrelations of the process X_i is finite, which implies a short-range dependent process.

2.3 Hurst parameter for LRD processes

Strictly speaking, Hurst parameter characterizes the behaviour of self-similar processes. It does not appear in the definition (Eq. (2.7)) nor in the basic property of the autocorrelation function of a long-range dependent process (Eq. (2.8)). However, it is not uncommon to attribute the Hurst parameter to a long-range dependent process. This stems from the fact that if a process is self-similar with stationary increments and $1/2 < H < 1$, then its increments are long-range dependent. Moreover, long-range dependence implies second-order self-similarity and vice versa, with the restriction [20]. Comparing Eq. (2.7) and (2.11), there is a linear relationship between the scaling exponent α and the Hurst parameter H:

$$H = 0.5(1+\alpha), \tag{2.12}$$

or, equivalently,

$$\alpha = 2H - 1 \tag{2.13}$$

with the restrictions $0 < \alpha < 1$ $(1/2 < H < 1)$.

The Hurst parameter measures the degree of long-range dependence of a process. For short-range dependent processes (including those without correlation), $0 < H \leq 0.5$. Values of H close to 1 indicate a process with a strong long-range dependence. For example, bursty network traffic has a large H [2].

As discussed in Section 2.1, a stochastic process may exhibit three types of behaviour with respect to its autocorrelation. It is important to determine the type of behaviour because even simple statistics, such as the variance of the sample mean, are highly dependent on whether the process is uncorrelated, short-range or long-range dependent. The Hurst parameter identifies the type of the process. Therefore, its correct and efficient estimation is important in statistical analysis of time series.

Chapter 3

Wavelet-based estimator of the Hurst parameter

3.1 Wavelet Transform

Let $X(t)$ be a continuous-time signal with a finite energy. Its continuous wavelet transform is given by the inner product

$$w(a,\tau) = \int_{-\infty}^{\infty} X(t)\psi_{a,\tau}(t)dt, \tag{3.1}$$

where

$$\psi_{a,\tau}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-\tau}{a}\right), \quad a \in \mathbb{R}^+, \tau \in \mathbb{R} \tag{3.2}$$

is the basis function of the transformation, called a *wavelet*. The wavelet $\psi_{a,\tau}(t)$ is obtained by dilating (by a factor of a) and time shifting (by τ time units) of a reference function $\psi(t)$ called a *mother wavelet* [21], [23]. In (3.2), a is called a scale factor and τ is a translation factor. The first notable difference between the well-known Fourier transform and the wavelet transform is that there is no unique wavelet function to

serve as a basis of the transformation. Rather, there is a freedom of choice of the mother wavelet, within certain constraints that define the behaviour of the wavelets.

A function can be a wavelet if it possesses certain properties. In order for the transformation to be invertible, the mother wavelet must satisfy the *admissibility condition*, i.e., the mean value of the mother wavelet must be zero:

$$\int_{-\infty}^{\infty} \psi(t)dt = 0. \tag{3.3}$$

This implies that the wavelets must be oscillating functions. Their frequency spectrum is bandpass and has a zero at the origin. Another property of the wavelets is their localization both in time and frequency. A function cannot be bandlimited and have a finite time support. Wavelets have most of their energy within a limited frequency band and within a limited period of time. The name *wavelet* itself summarizes the previous two properties ("a small wave").

The mother wavelet has a number of vanishing moments N, defined as the largest N for which

$$\int_{-\infty}^{\infty} t^k \psi(t)dt = 0, \quad k = 0, 1, \ldots, N-1 \tag{3.4}$$

holds. Each wavelet has at least one vanishing moment because for $k = 0$, Eq. (3.4) becomes identical to the admissibility condition (3.3). The frequency spectrum $\Psi(\nu)$ of the mother wavelet $\psi(t)$ is proportional to $|\nu|^N$ close to the origin.

Continuous wavelet transform is highly redundant. For example, it transforms a one-dimensional signal $X(t)$ into a two-dimensional continuous function $w(a, \tau)$. In some cases, it is possible to sample $w(a, \tau)$ without loss of information about $X(t)$. The conditions imposed on the mother wavelet in order to achieve the lossless sampling are stated in [23]. The sampling of the time-scale plane is performed on a dyadic grid:

$a = 2^j, \tau = 2^j k, j \in \mathcal{Z}^+, k \in \mathcal{Z}$ [21] and the resulting transformation is called *discrete wavelet transform* (DWT). The value j is called *octave* and k is *translation*. The resulting wavelet coefficients are

$$\begin{aligned} d(j,k) &= w(2^j, 2^j k) \\ &= \int_{-\infty}^{\infty} X(t) 2^{-j/2} \psi(2^{-j}t - k) dt \\ &= \int_{-\infty}^{\infty} X(t) \psi_{j,k}(t) dt. \end{aligned} \tag{3.5}$$

The DWT represents the signal $X(t)$ as a weighted sum of wavelets [21]. The reconstruction formula for the DWT is

$$X(t) = \sum_{j=0}^{\infty} \sum_{k=-\infty}^{\infty} d(j,k) \psi_{j,k}(t). \tag{3.6}$$

If the sum over j's is split in two regions, $j > j_2$ and $0 \leq j \leq j_2$, Eq. (3.6) takes the form

$$\begin{aligned} X(t) &= \sum_{j=j_2+1}^{\infty} \sum_{k=-\infty}^{\infty} d(j,k) \psi_{j,k}(t) + \sum_{j=0}^{j_2} \sum_{k=-\infty}^{\infty} d(j,k) \psi_{j,k}(t) \\ &= \sum_{k=-\infty}^{\infty} c(j_2,k) \phi_{j_2,k}(t) + \sum_{j=0}^{j_2} \sum_{k=-\infty}^{\infty} d(j,k) \psi_{j,k}(t). \end{aligned} \tag{3.7}$$

The first term in Eq. (3.7) represents an approximation of the signal at the octave j_2. The second term is a sum of details. When added to the approximation, it produces the original signal $X(t)$. The function $\phi_{j_2,k}(t)$ is called a *scaling function* at octave j_2. The corresponding coefficients $c(j_2,k)$ are called *approximation coefficients* at octave j_2. The octave j_2 measures the level of detail in the approximation. When j_2 increases, the approximations become coarser and vice versa.

An important property of the approximation and wavelet coefficients at octave j is that they can be obtained by linear, discrete-time filtering of the approximation

coefficients at octave $j-1$ (the next finer octave) [21]. This allows calculation of the DWT of a signal $X(t)$ by employing discrete-time filter-bank based pyramidal algorithms that have very low computational cost. The input to the algorithm is the sequence of approximation coefficients at the finest octave, $j = 0$. In the case of discrete-time signals $X(n)$, $n \in \mathcal{Z}$, the signal itself can be treated as the finest approximation. However, this approach introduces errors, particularly at the finest octaves. For that reason, the original signal $X(n)$ should be pre-filtered in order to obtain the initial approximation sequence [13], [21].

The DWT captures a signal at various time scales or levels of aggregation. Due to the scale invariance of the basis functions, it is suitable for analyzing properties that are present across a range of time scales, such as LRD. The low computational cost makes the DWT a popular tool for signal analysis.

3.2 Wavelet-based Hurst parameter estimator

The wavelet-based Hurst parameter estimator is based on the shape of the power spectral density (PSD) function (2.9) of the LRD signal $X(t)$. It has been shown [13], [21], that when the PSD has a power-law behavior, the relationship between the variance of the wavelet coefficients on a given octave and the octave j is

$$\mathrm{E}\left\{d(j,k)^2\right\} = 2^{j\alpha} c_f C(\alpha, \psi), \tag{3.8}$$

where the average is calculated for various k, α is the scaling exponent, and

$$C(\alpha, \psi) = \int |\nu|^{-\alpha} |\Psi(\nu)|^2 \, d\nu \tag{3.9}$$

does not depend on the octave j. In Eq. (3.9), $\Psi(\nu)$ is the Fourier transform of the mother wavelet ψ. The integral given by Eq. (3.9) converges if the number of

vanishing moments of the mother wavelet satisfies $N > 0.5(\alpha - 1)$ [12]. For LRD processes $0 < \alpha < 1$, and, therefore, the integral converges for $N \geq 1$, which is always satisfied.

The number of vanishing moments N controls the correlation between any two wavelet coefficients $d(j_1, k_1)$ and $d(j_2, k_2)$. When $N \geq \alpha/2$, the wavelet coefficients are not long-range dependent and the calculation of $\mathrm{E}\{d(j,k)^2\}$ becomes a simple time average or sample mean for all k's:

$$\mathrm{E}\Big\{d(j,k)^2\Big\} = \frac{1}{n_j}\sum_{k=1}^{n_j} d(j,k)^2, \tag{3.10}$$

where n_j is the number of wavelet coefficients available at octave j. Linear relationship with a slope α $(0 < \alpha < 1)$ between $\log_2 \mathrm{E}\{d(j,k)^2\}$ and j for a range of octaves, including the coarsest, indicates presence of LRD. Therefore, α is obtained by performing linear regression of $\log_2 \mathrm{E}\{d(j,k)^2\}$ on j over a range of octaves.

The estimator of the Hurst parameter is based on the following idealizations [13]:

1. The process $X(t)$ and its wavelet coefficients are Gaussian.

2. For fixed j, $d(j,k)$ are independent, identically distributed variables.

3. The processes $d(j_1,k)$ and $d(j_2,k)$, for $j_1 \neq j_2$ are independent.

The first assumption is important for deriving analytical expressions for the variance of the estimates of $\log_2 \mathrm{E}\{d(j,k)^2\}$. The second and the third are a basis for (3.10).

The Hurst parameter H is calculated by using Eq. (2.12). We employed publicly available MATLAB code [24] to estimate α and H. The estimator first performs DWT on the input signal, employing wavelets from the Daubechies family. We used the wavelet Daubechies3, which has three vanishing moments. After computing the

DWT, the estimator calculates the estimates of $\log_2 \mathrm{E}\{d(j,k)^2\}$ and variances of these estimates and performs a weighted linear regression. The weights are inversely proportional to the variances of the estimates of $\log_2 \mathrm{E}\{d(j,k)^2\}$. The estimator employs a weighted rather than simple linear regression because when j increases, n_j decreases, and, therefore, the variance of $\log_2 \mathrm{E}\{d(j,k)^2\}$ increases. This implies less accurate estimation of $\log_2 \mathrm{E}\{d(j,k)^2\}$ for large j's. The weighted linear regression gives more significance to the estimates of $\log_2 \mathrm{E}\{d(j,k)^2\}$ on finer octaves. This estimator is called *monofractal wavelet estimator.*

An extension to the basic monofractal wavelet estimator is the *multifractal estimator* [25]. In addition to the second moments (variances) of the wavelet coefficients, it also takes into account moments of higher order:

$$S_q(j) = \frac{1}{n_j} \sum_{k=1}^{n_j} d(j,k)^q. \tag{3.11}$$

The estimator estimates the slope α_q by performing linear regression of $\log_2 S_q(j)$ for a range of j's. H is calculated using an expression analogous to (2.12), by taking into account the order of the moment:

$$H = 0.5 + \alpha_q/q. \tag{3.12}$$

Both monofractal and multifractal estimators are used to produce logscale diagrams. They plot $\log_2 \mathrm{E}\{d(j,k)^2\}$ or $\log_2 S_q(j)$, with the corresponding confidence intervals, versus j. An example of a logscale diagram is shown in Figure 3.1. The solid line connects the estimates of $\log_2 \mathrm{E}\{d(j,k)^2\}$. The vertical lines represent the confidence intervals of the estimates and the dashed line is the slope of the linear regression performed over the range of octaves [4–13].

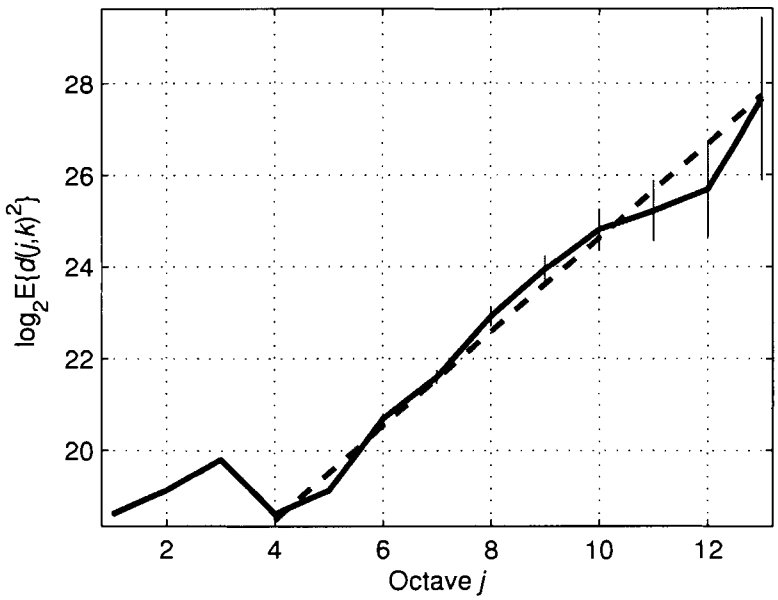

Figure 3.1: Logscale diagram for the MPEG-4 encoded "Star Wars IV" video sequence.

3.3 Test for time constancy of the scaling exponent α

Long-range dependent processes are, by definition, wide-sense stationary. However, they possess certain characteristics that, at first glance, make them *look* non-stationary. For example, LRD processes exhibit high variability [22] and there are relatively long periods where the observations stay at high and low levels [11]. The question is: how to distinguish between wide-sense stationary processes with LRD and inherently non-stationary processes?

The scaling exponenet α characterizes the behaviour the autocorrelation function of a stochastic process. Therefore, if α changes over time, then the autocorrelation function of the process also chages over time. This implies a non-stationary process. An approach to determining, with a certain probability, whether a process with

$H > 0.5$ is LRD or non-stationary is to test whether the scaling exponent α is constant over the examined trace [22]. This is performed by splitting the original trace into m sub-traces and estimating α for each sub-trace. If α is constant and $0 < \alpha < 1$, then the trace is LRD. Otherwise, the trace is non-stationary.

The test relies on the wavelet-based estimator of α. Besides the idealizations made for the estimator, the test has two additional properties that play a key role in its definition and application:

1. Estimates of α fit a Gaussian distribution, with a variance that depends only on the range of octaves where α is estimated and on the number of available wavelet coefficients at a given octave.

2. Estimates of α taken over non-overlapping blocks (sub-traces) are uncorrelated.

Experimental verifications of the above assumptions are reported in [22]. When the assumptions hold, examining whether or not α is constant becomes equivalent to testing whether or not a sequence of uncorrelated Gaussian random variables with known variances have the same mean [22].

MATLAB implementation of the test is available online [24]. Testing the time constancy of α is performed in two steps:

1. An initial estimation of α for the entire trace is performed. The range of octaves (if any) where there is a linear relationship between $\log_2 \mathrm{E}\{d(j,k)^2\}$ and j with a slope α is identified from the obtained logscale diagram.

2. The test for time constancy of α, for various values of m, is applied in the range of octaves determined in Step 1. There is no optimal way of selecting m, and, hence, the test should be repeated several times.

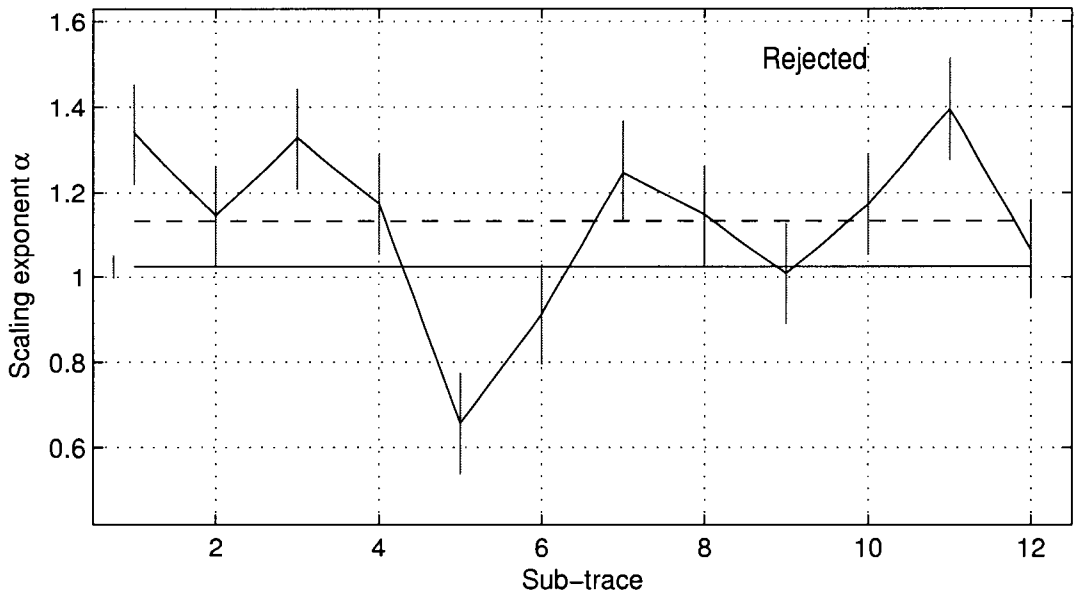

Figure 3.2: Test for time constancy of α for $m = 12$ for the MPEG-4 encoded "Star Wars IV" video sequence.

The test checks whether or not the hypothesis that α is constant can be accepted. The significance level, which determines the threshold for acceptance or rejection of the hypothesis, can be varied. We used the default value of 5%. Figure 3.2 shows a sample graphical output of the test for $m = 12$. The estimates of α are connected by a line. The vertical lines indicate the confidence intervals of the estimates. The solid horizontal line indicates the overall value of α. The dashed horizontal line shows the average of the estimates. The outcome of the test is shown in the top right-hand side of the graph.

Chapter 4

Analysis of MPEG video traces

In this chapter, we describe the application of the wavelet-based estimator of the Hurst parameter to MPEG-1 and MPEG-4 encoded video sequences. We also address the unreliability of the estimates and test the Gaussianity of the wavelet coefficients and the time constancy of the scaling exponent α.

4.1 Characteristics of MPEG traces

MPEG (Motion Picture Expert Group) is a set of standards for compression of video, or sequences of images. There are several versions of the standards. MPEG-1 is older, while MPEG-4 is more advanced and achieves better compression performances than MPEG-1. The basic principles of operation of both standards are rather similar.

Compression is achieved by reducing the spatial and temporal redundancy in the sequence of images (frames) [9]. Spatial redundancy (redundancy within an image) is reduced by applying algorithms for compression of still images (JPEG, for example). The major difference between MPEG-1 and MPEG-4 is in the algorithms and level of

reduction of the spatial redundancy. MPEG-1 coders employ discrete cosine transform on the complete original (uncompressed) image (frame-based compression). MPEG-4 coders can utilize both discrete cosine and wavelet transforms not only on the entire frame, but also on parts of it (object-based compression) [26]. Temporal redundancy (redundant information between successive images in the video sequence) is reduced by prediction of the next image based on the previous one(s). Both MPEG-1 and MPEG-4 coders create three types of frames: I, P, and B. I frames are compressed versions of the original input frames. P frames are obtained by forward prediction with motion compensation with respect to the previous I or P frame. B frames can be obtained by both forward and backward prediction with respect to the previous and next I or P frames. At the output of the coder, frames are organized in a deterministic, periodic sequence, called *Group of Pictures* (GoP) [9]. Traces that are used in this thesis are obtained from compressed video sequences whose frames form the following GoP: IBBPBBPBBPBB.

Objects of our interest are sequences of frame sizes. Typically, I frames are larger than P frames, which, in turn, are larger than B frames. This can be observed in Figure 4.1. It shows the sizes of several successive frames from the MPEG-1 encoded cartoon "Simpsons" [27].

MPEG-1 [27] and MPEG-4 [28] traces used in this thesis have a frame rate of 25 frames per second. This implies that the time interval between successive frames is 40 ms. Each sample (entry) in the traces represents the size of the corresponding frame in bits. MPEG-1 traces have 40,000 samples, corresponding to 26 minutes and 40 seconds of video. MPEG-4 traces are of variable lengths, ranging from 22,498 samples (15 minutes of video) to 89,998 samples (1 hour of video). Trace lengths are

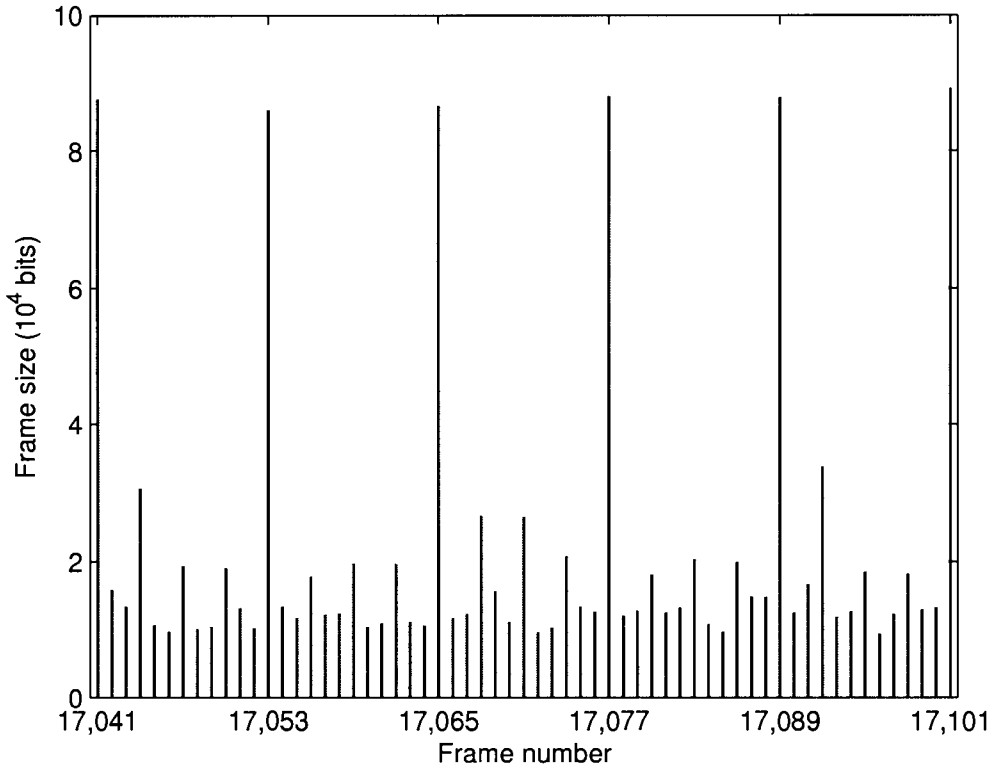

Figure 4.1: Excerpt from the series of frame sizes from the MPEG-1 encoded cartoon "Simpsons".

summarized in Table 4.1.

4.2 Hurst parameter estimation

We used wavelet-based monofractal and multifractal estimators to estimate the Hurst parameter of various traces listed in Table 4.1. The employed wavelet was Daubechies' wavelet of genus 3 (three vanishing moments). Previous studies [12], [15], have shown that wavelets with three vanishing moments are suitable for analyzing LRD processes. For the multifractal estimator, we estimated H for several values of q (order of the

Table 4.1: MPEG-1 and MPEG-4 trace lengths.

	Trace	Encoding	Length (frames)	Duration (min)
1	MTV	MPEG-1	40,000	26.67
2	Jurassic park	MPEG-1	40,000	26.67
3	Simpsons	MPEG-1	40,000	26.67
4	Mr. Bean	MPEG-1	40,000	26.67
5	Silence of the lambs	MPEG-1	40,000	26.67
6	Talk show	MPEG-1	40,000	26.67
7	ARD news	MPEG-4	22,498	15.00
8	Die hard III	MPEG-4	89,998	60.00
9	Formula 1	MPEG-4	44,998	30.00
10	Futurama	MPEG-4	30,334	20.22
11	From dusk till dawn	MPEG-4	89,998	60.00
12	First contact	MPEG-4	89,998	60.00
13	Mr. Bean	MPEG-4	89,057	59.37
14	Jurassic park	MPEG-4	89,998	60.00
15	VIVA video clips	MPEG-4	89,998	60.00
16	N3 talk	MPEG-4	89,998	60.00
17	Silence of the lambs	MPEG-4	89,998	60.00
18	Simpsons	MPEG-4	30,334	20.22
19	Star wars IV	MPEG-4	89,998	60.00

moments), with similar results. In this thesis, we report the estimates of H obtained by considering the third-order moments ($q = 3$).

Logscale diagrams from both monofractal and multifractal estimators for the MPEG-1 encoded "Simpsons" and MPEG-4 encoded "Jurassic Park" videos are shown in Figures 4.2–4.5. The remaining traces shown in Table 4.1 have similar shapes of

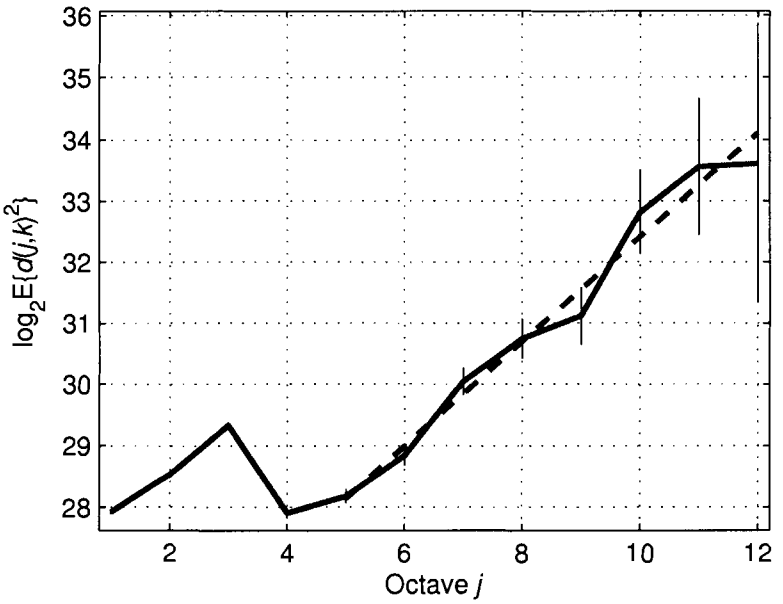

Figure 4.2: Monofractal estimator: logscale diagram for the MPEG-1 encoded "Simpsons" sequence.

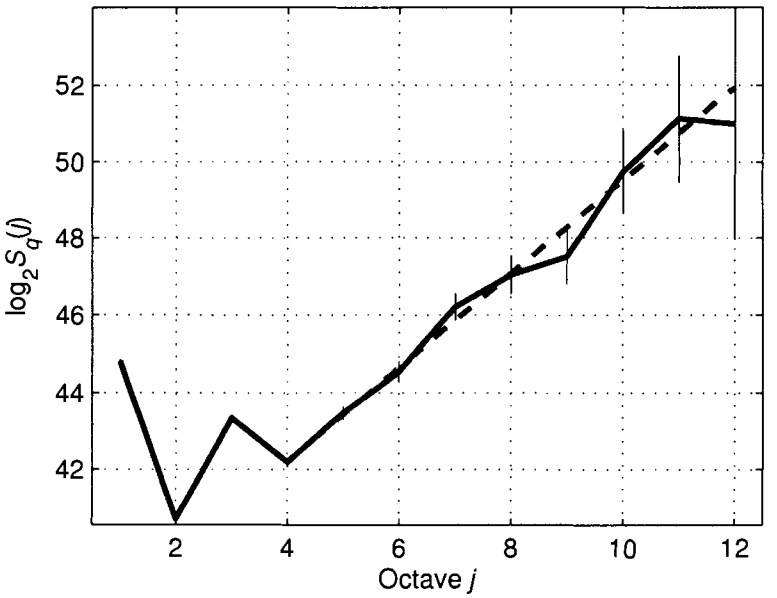

Figure 4.3: Multifractal estimator: logscale diagram for the MPEG-1 encoded "Simpsons" sequence.

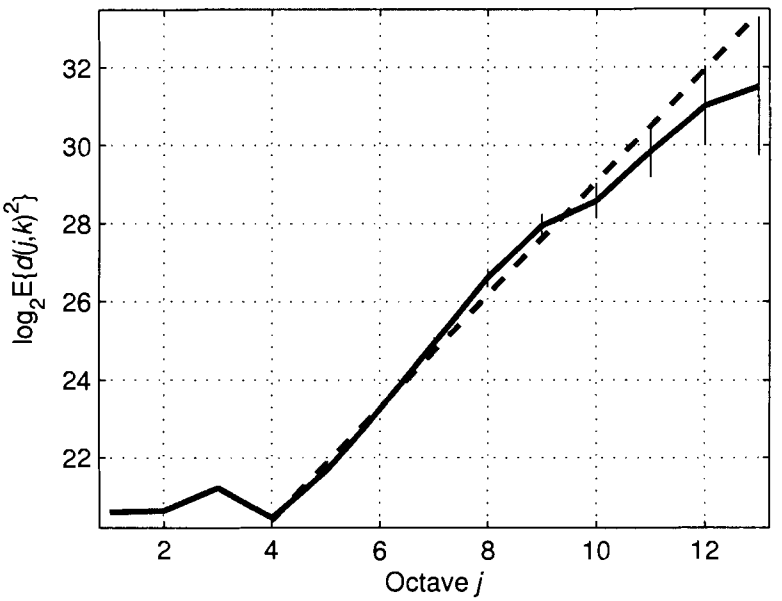

Figure 4.4: Monofractal estimator: logscale diagram for the MPEG-4 encoded "Jurassic park" sequence.

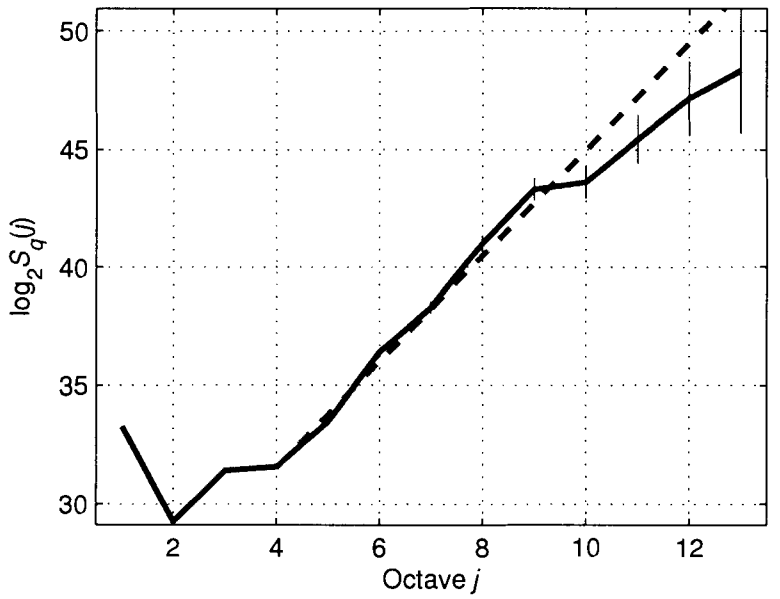

Figure 4.5: Multifractal estimator: logscale diagram for the MPEG-4 encoded "Jurassic park" sequence.

the logscale diagrams. They exhibit a linear relationship between $\log_2 \mathrm{E}\{d(j,k)^2\}$ (Eq. (3.10)) or $\log_2 S_q(j)$ (Eq. (3.11)) and j for the largest values of j (coarsest octaves or time scales). The linear region typically begins at $j = 4$ or 5. The lack of linearity over the finer octaves may be attributed to artifacts of MPEG compression algorithms or to a transition between short-term and long-term scaling behavior [6].

Numerical values of the estimates are summarized in Table 4.2. For each estimate of H ("value"), we report the range of octaves where the linear regression is performed ("range"). These ranges were chosen by visual inspection of the logscale diagrams and identification of the linear region. The "periodogram" column contains estimates of H obtained from the periodogram-based estimator [29]. Column "R/S" shows R/S estimates of H reported in [9] and [28].

Both monofractal and multifractal estimators produce similar results, as indicated by Table 4.2. They are in agreement with periodogram-based estimates. The linearity of logscale diagrams for the coarsest octaves and the match between wavelet and periodogram-based estimates indicate that PSD's of the traces exhibit power-law behavior close to the origin, with exponents α often greater than one. This implies values of H greater than one, which contradicts the LRD assumption. For LRD processes, α, and, consequently, H should be strictly smaller than one.

We compared the wavelet-based estimates of H with estimates obtained from R/S plots. R/S plots yield values of $H < 1$, except for the trace "Silence of the lambs" ($H = 1.007$). Other studies [15], [6] reported estimates of H for MPEG video traces obtained from variance-time plots. These estimates were, in general, smaller than one. This indicates that estimators of H that operate in the time domain (R/S and variance-time plots) produce similar results. Also, estimators that operate in the

Table 4.2: Hurst parameter estimates of the video traces.

Trace	Encoding	Estimates of H					
		Monofractal		Multifractal		Perio-dogram	R/S
		range	value	range	value		
MTV	MPEG-1	4–12	0.959	3–12	0.937	0.992	0.89
Jurassic park	MPEG-1	5–12	1.096	4–12	1.012	1.191	0.88
Simpsons	MPEG-1	5–12	0.926	4–12	0.906	0.988	0.89
Mr. Bean	MPEG-1	5–12	1.214	5–12	1.258	1.295	0.85
Silence of the lambs	MPEG-1	5–12	1.130	5–12	1.152	1.171	0.89
Talk show	MPEG-1	5–12	1.084	5–12	1.132	1.174	0.89
ARD news	MPEG-4	5–11	1.382	4–11	1.225	1.310	0.967
Diehard III	MPEG-4	4–13	1.190	4–13	1.208	1.233	0.969
Formula 1	MPEG-4	4–12	1.189	4–12	1.169	1.216	0.867
Futurama	MPEG-4	4–12	0.943	4–12	0.909	1.064	0.877
From dusk till dawn	MPEG-4	4–13	1.139	4–13	1.138	1.186	0.909
First contact	MPEG-4	4–13	1.194	4–13	1.213	1.268	0.931
Mr. Bean	MPEG-4	4–13	1.083	4–13	1.109	1.151	0.933
Jurassic park	MPEG-4	4–13	1.222	4–13	1.247	1.293	0.973
VIVA video clips	MPEG-4	2–13	1.000	2–13	1.120	1.119	0.961
N3 talk	MPEG-4	4–13	1.079	4–13	1.131	1.188	0.882
Silence of the lambs	MPEG-4	4–13	1.277	4–13	1.260	1.337	1.007
Simpsons	MPEG-4	4–12	0.964	4–12	0.941	1.061	0.889
Star wars IV	MPEG-4	4–13	1.013	4–13	1.051	1.138	0.903

frequency domain (periodogram and wavelet-based) yield similar estimates. However, estimates obtained from time-domain estimators (usually $H < 1$) differ from those obtained from frequency-domain estimators (often $H > 1$).

4.3 Investigating the sources of unreliability of the estimates

For LRD processes, the Hurst parameter should satisfy $H < 1$. As reported in Section 4.2, wavelet-based estimators produce values of H greater than one. In order to investigate the possible sources of the unreliability of the wavelet estimator, we test the Gaussianity of the traces and the time constancy of the scaling exponent α. We also address the results regarding the performance of the wavelet-based estimator reported in [30].

4.3.1 Testing the Gaussianity of the wavelet coefficients

One of the idealizations assumed by the wavelet estimator is that the analyzed process and its wavelet coefficients on various octaves are Gaussian [13], [21]. Therefore, we examined how close the traces and their wavelet coefficients are to a Gaussian distribution using q-q plots [31]. Q-q plots of the MPEG-4 encoded "Star Wars IV" video sequence and its wavelet coefficients on octaves 1–9 are shown in Figures 4.6 and 4.7, respectively. The vertical axis represents the quantiles of the trace (Figure 4.6), or its wavelet coefficients (Figure 4.7). The horizontal axis represents the quantiles of a Gaussian distribution with the same mean and variance as the original trace or the corresponding set of wavelet coefficients. The dashed line is the reference line with a slope of one. The vertical lines mark the 10% and 90% quantiles.

Figure 4.6 shows that the trace is highly non-Gaussian. As shown in Figure 4.7, wavelet coefficients on finer octaves (1 and 2) deviate from Gaussianity. However, in the range of octaves where H was estimated (from 4 or 5 and up), the wavelet

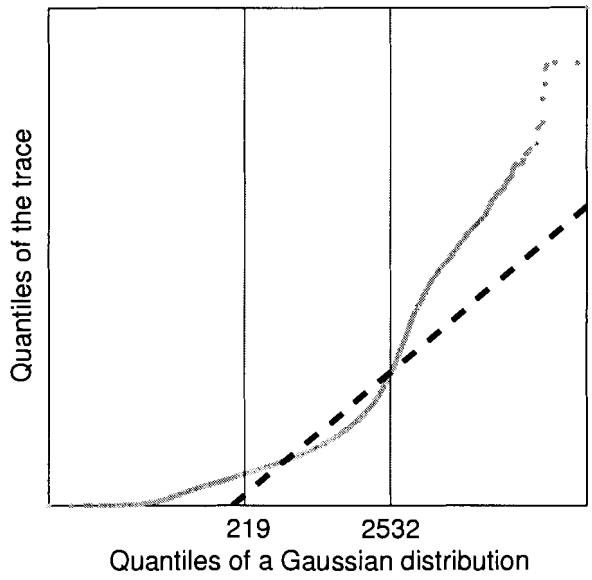

Figure 4.6: Q-q plot of the trace "Star Wars IV".

coefficients have approximately Gaussian distribution, with several outliers at each tail. This indicates that the unreliable performance of the wavelet-based estimator cannot be attributed to the lack of Gaussianity of the trace.

4.3.2 Testing the time constancy of α

We examine the time constancy of α and perform a set of tests for each video trace. We chose the number of sub-traces $m \in \{3, 4, 6, 8, 10, 12, 15\}$. The lower bound of the range where α is estimated is set to the value given in Table 4.2. It varies between 2 and 5. The upper bound depends on m. For larger m, the sub-traces are shorter and there are fewer available octaves. In our experiments, the upper octave varies between 8 and 12.

Figures 4.8–4.10 show graphical outputs of the test for time constancy of α for three traces. The top graph is the time series of the corresponding trace. The remaining

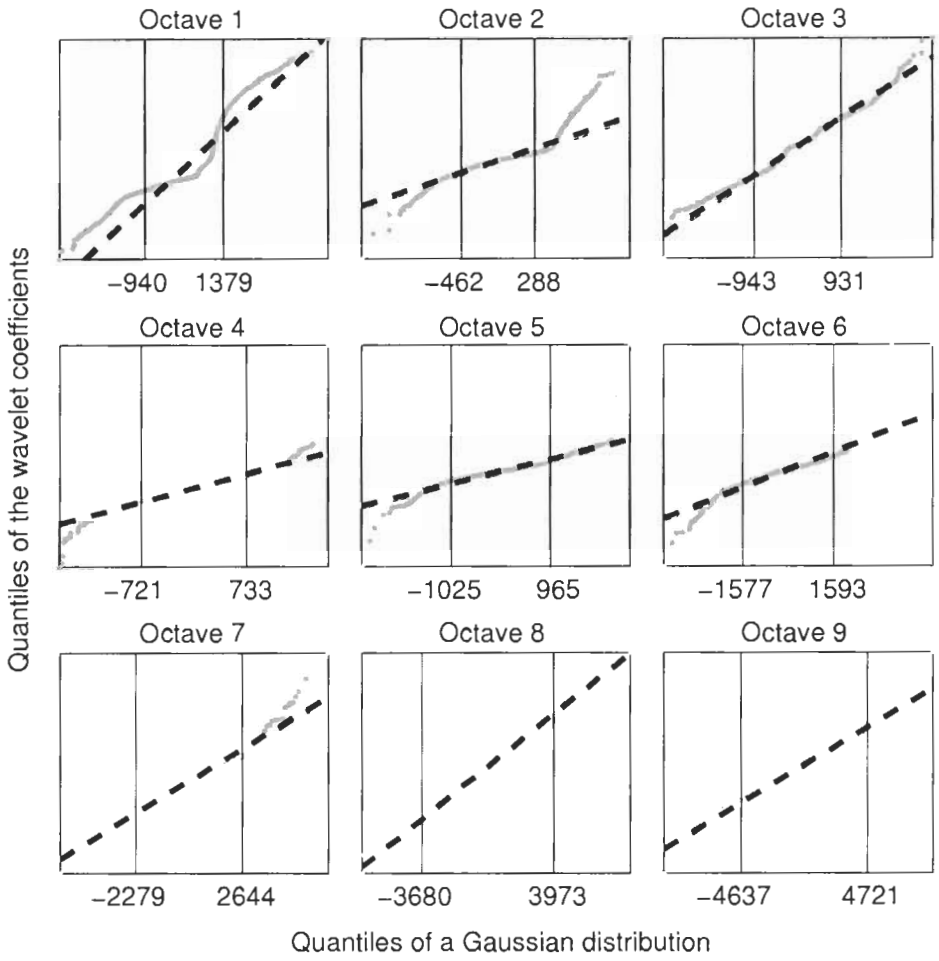

Figure 4.7: Q-q plots of the wavelet coefficients for octaves 1-9 for the trace "Star Wars IV".

three graphs show the values of α for every sub-trace, for $m = 12$, 8, and 4. The estimates of α are connected by a broken line. The vertical lines show the confidence intervals of the estimates of α. The solid horizontal line represents the overall value of α. The dashed horizontal line shows the average of the estimates of α. The outcome of the test is shown in the top left-hand side of the graph, where "rejected" means that the trace did not pass the test for time constancy of α.

Table 4.3 summarizes the results of the test for time constancy of α. For each

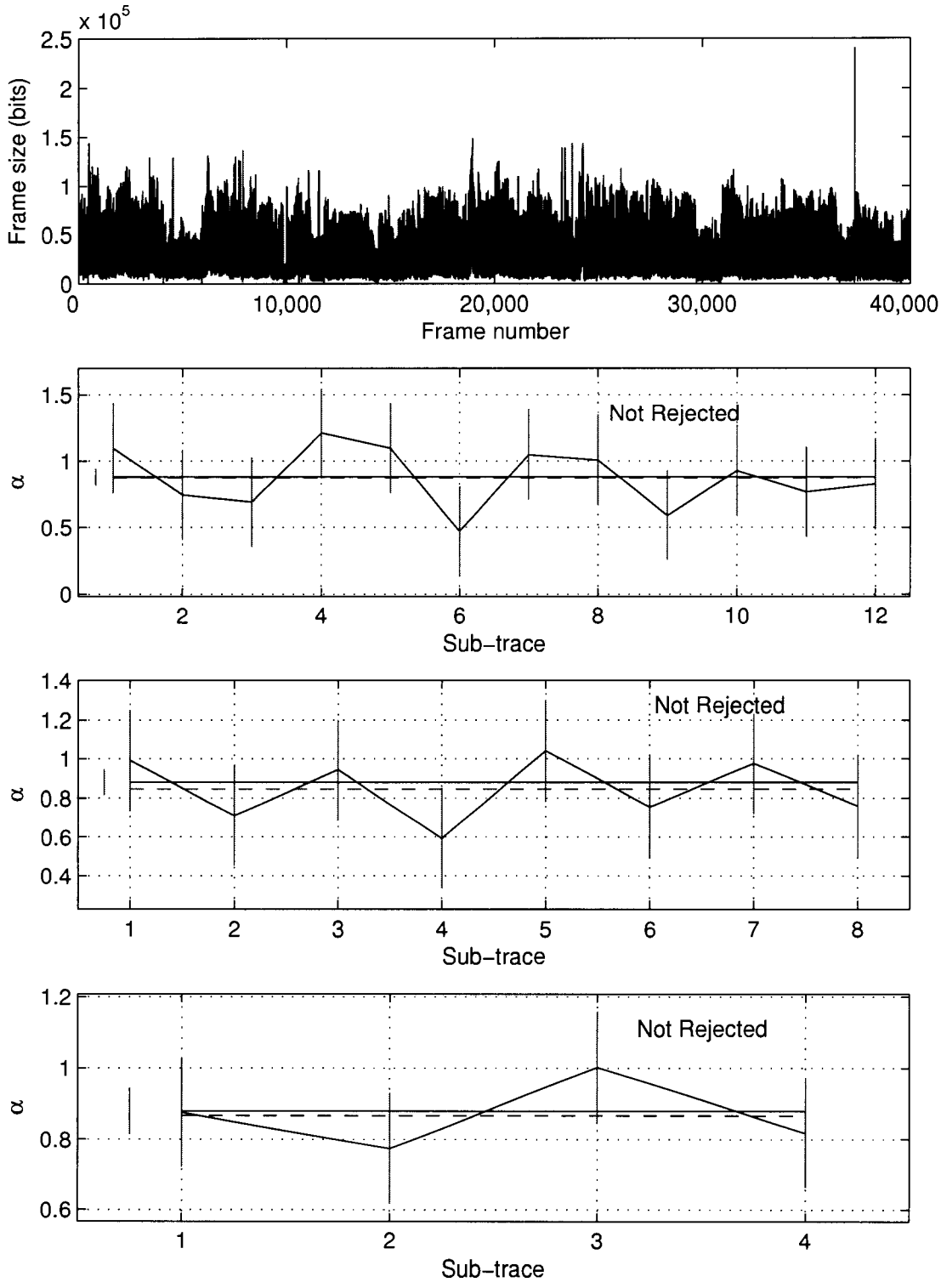

Figure 4.8: Test for time constancy of α for the MPEG-1 encoded "Simpsons" video sequence.

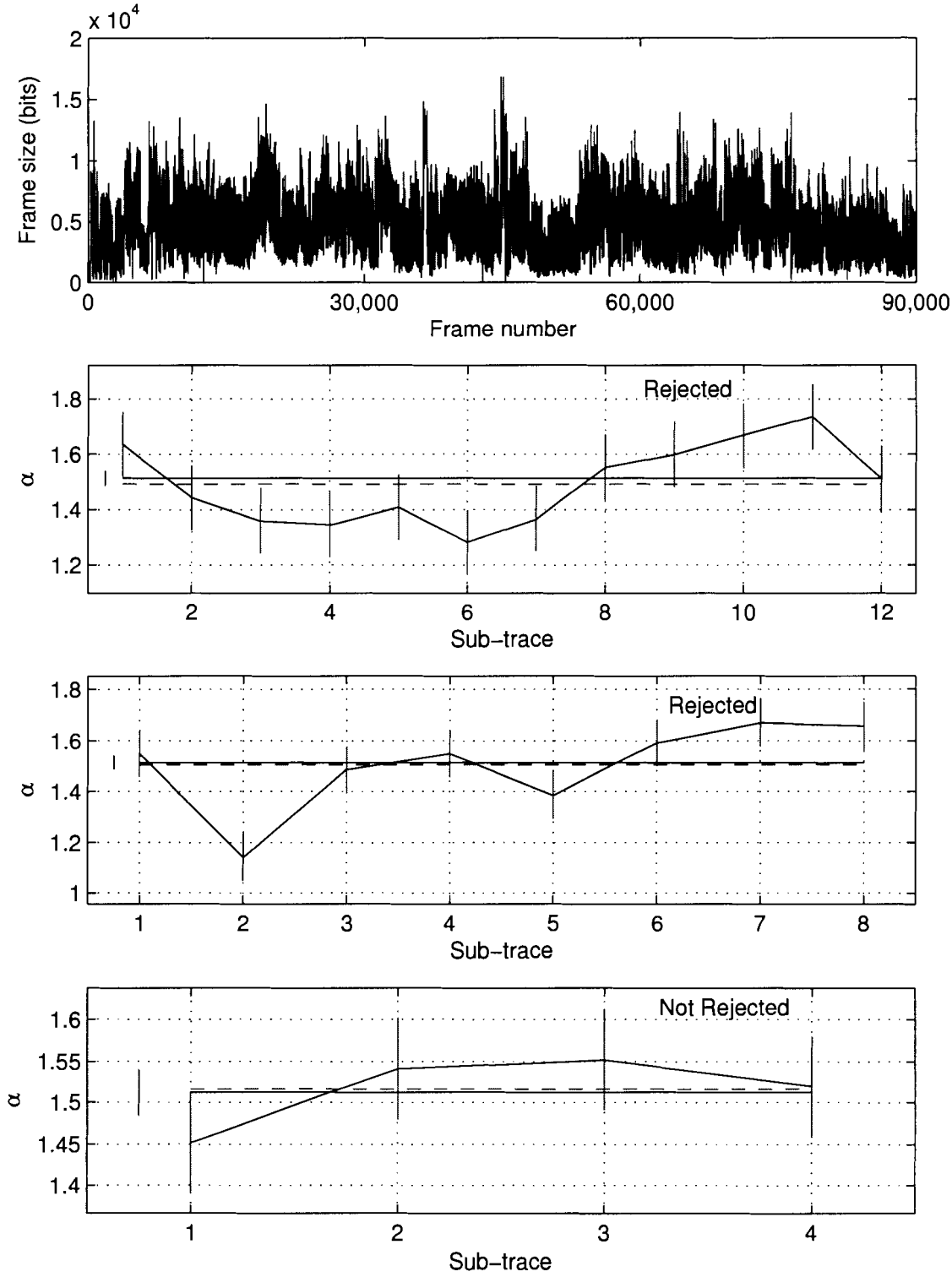

Figure 4.9: Test for time constancy of α for the MPEG-4 encoded "Jurassic park" video sequence.

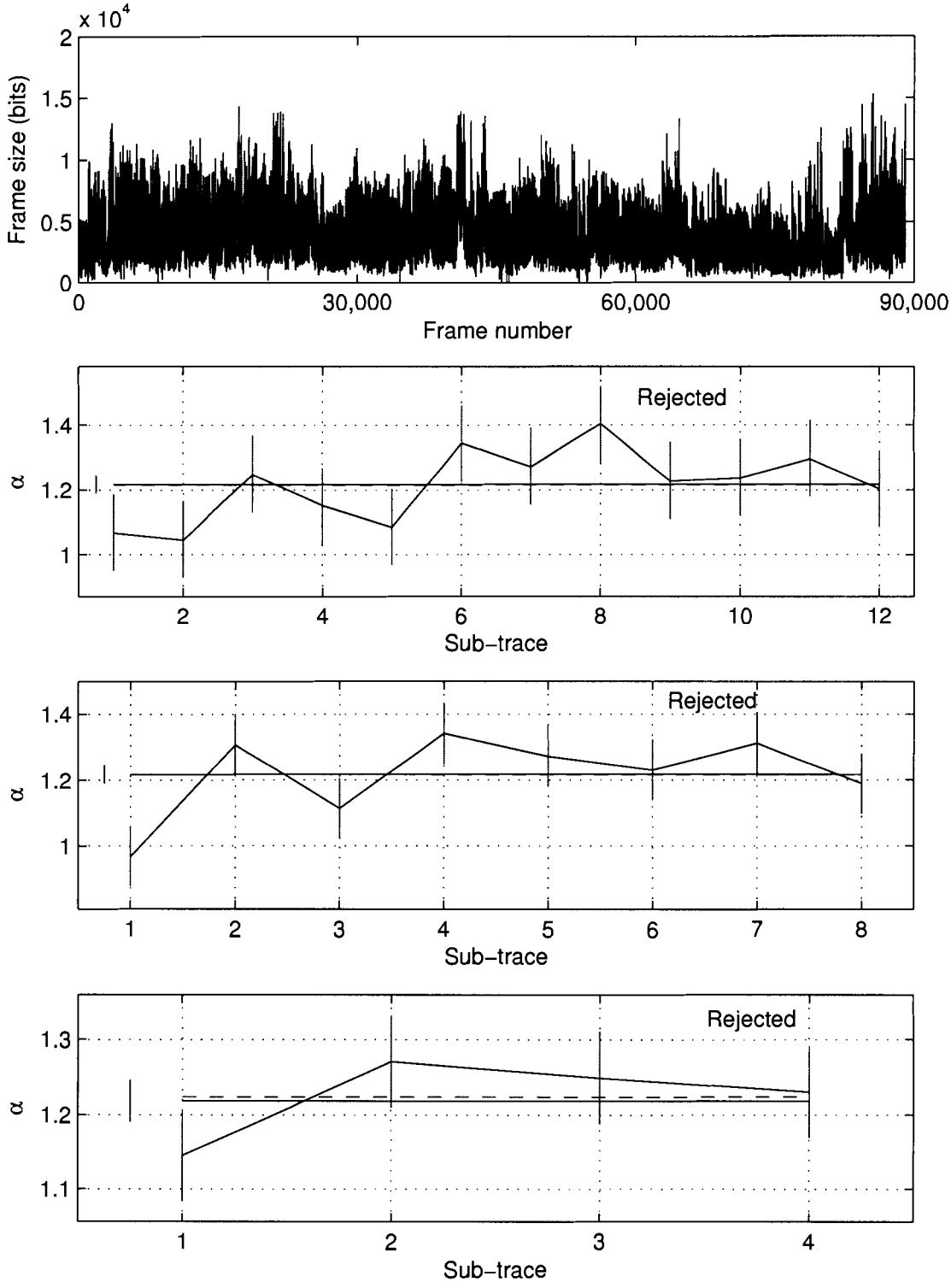

Figure 4.10: Test for time constancy of α for the MPEG-4 encoded "Mr. Bean" video sequence.

trace, we report the outcomes of the test for various m, where R stands for "rejected" and N for "not rejected". The last two columns indicate the number of times the trace failed or passed the test. Our findings indicate that 12 traces fail the test for all values of m, while others pass the test for certain values of m. MPEG-1 encoded "Simpsons" and MPEG-4 encoded "ARD news" video sequences pass the test for over 50% of m's. The remaining seven video traces pass the test for less than 50% of m's This indicates that α is not constant and varies with time. Therefore, estimating α and H for the entire trace is not meaningful and the estimates are not reliabale.

4.4 Possible causes of the unreliable estimates

The results presented in Section 4.2 show that the wavelet-based estimator produces values of $H > 1$. Estimates of H obtained from R/S plots are often greater than 0.9, which indicates strong LRD component. We conjecture that video traces also posses a strong short-range dependent (SRD) component. Video sequences consist of various scenes. Video frames representing a single scene are similar due to the identical or similar background and objects in the scene. This implies similar sizes of the adjacent frames, which indicates a strong positive correlation for small lags. It has been shown that the wavelet-based estimator produces unreliable results when applied to processes that possess both a strong SRD component and a strong LRD component [30].

Furthermore, the traces often failed the test for the time constancy of α. The linearity of the logscale diagrams may be attributed to the averaging of the nonstationarities that manifest as variability of the scaling exponent α across the traces

Table 4.3: Results of the test for time constancy of α.

Trace	Encoding	m							Rejected	Not rejected
		15	12	10	8	6	4	3		
MTV	MPEG-1	R	R	R	R	R	R	R	7	0
Jurassic park	MPEG-1	R	R	R	R	R	R	R	7	0
Simpsons	MPEG-1	R	N	N	N	R	N	N	2	5
Mr. Bean	MPEG-1	R	R	R	R	R	R	R	7	0
Silence of the lambs	MPEG-1	R	R	R	R	R	R	R	7	0
Talk show	MPEG-1	R	R	R	R	R	R	R	7	0
ARD news	MPEG-4	R	N	N	R	N	N	R	3	4
Diehard III	MPEG-4	R	R	R	R	R	R	R	7	0
Formula 1	MPEG-4	N	R	R	R	R	N	N	4	3
Futurama	MPEG-4	R	R	R	R	R	R	R	7	0
From dusk till dawn	MPEG-4	R	R	R	R	N	R	R	6	1
First contact	MPEG-4	R	R	R	R	R	R	R	7	0
Mr. Bean	MPEG-4	R	R	R	R	R	R	R	7	0
Jurassic park	MPEG-4	R	R	R	R	R	N	R	6	1
VIVA video clips	MPEG-4	R	R	R	R	R	R	R	7	0
N3 talk	MPEG-4	R	R	R	R	R	R	R	7	0
Silence of the lambs	MPEG-4	R	R	R	R	R	R	R	7	0
Simpsons	MPEG-4	R	R	R	R	N	R	R	6	1
Star wars IV	MPEG-4	R	R	R	R	N	R	N	5	2

[22]. However, as indicated by Figures 4.8–4.10, estimates of α even in the sub-traces are often greater than one. Shorter timeseries may be regarded as stationary, or at least wide-sense stationary. Outputs of the estimator indicate that close to the origin, the behaviour of the PSD function of the sub-traces can be approximated by a

power-law with an exponent greater then one. This contradicts the LRD assumption that α should be strictly smaller than one. In our opinion, there are two possible explanations for the "unreliable" estimates of H for the sub-traces:

1. Sub-traces, similarly to the whole traces, possess both strong LRD and SRD components, in which case estimates of α and H are indeed unreliable and should be abandoned.

2. Traces and sub-traces have a correlation structure that is in contrast with the property (2.8). In this case, the cause of the unreliable estimates of α lies in the choice of inappropriate model for long-term correlations in the traces. We note that this can only be conjectured, because, to the best of our knowledge, there is no other model for LRD in literature.

Chapter 5

Analysis of E-Comm traffic traces

We apply the monofractal wavelet-based estimator of the Hurst parameter to traffic traces from an operational, circuit-switched radio network operated by E-Comm [19]. We also test the time constancy of the scaling exponent α.

5.1 E-Comm system and traffic traces

E-Comm is an emergency communications centre that provides radio communications to several public safety agencies, such as police, fire department and ambulance, in the Greater Vancouver Regional District (GVRD) area. E-Comm employs the Enhanced Digital Access Communication System (EDACS). It consists of various interconnected network elements: data and PBX gateways, radio transceivers and repeaters, dispatch and management consoles, and network switches. The network is based on circuit switching and carries predominantly voice traffic. The radio interface has a cellular architecture. There are 11 cells covering disjoint areas within the GVRD. Each cell has a number of available frequencies (radio channels) that determine its capacity.

The cell covering Vancouver has the largest capacity and it handles the majority of calls [4], [32]. Therefore, we chose to examine the traffic from that cell.

We had access to traffic data from E-Comm from 2001, 2002, and 2003. The data consisted of records of network events, such as established, queued, and dropped calls. Our analysis focused on established calls in the Vancouver cell. We analyzed traffic data from one week in 2003, one week in 2002, and two days in 2001. From the traffic data, we created and analyzed traces that consisted of the call holding times and call inter-arrival times. Analyses were performed separately on weekly, daily, and hourly traces.

Figure 5.1 shows the time series of the hourly trace from 22:00:00 to 23:00:00 on March 26, 2003. The horizontal axis shows the timestamps of the calls. The vertical axis shows the call holding times in seconds. The inset graph is the one-minute interval between 22:18 and 22:19. Call inter-arrival times can be observed in the inset graph as time intervals between successive calls.

5.2 Estimating the Hurst parameter and testing the time constancy of α

Voice traffic in circuit-switched networks is often modelled as a Poisson stochastic process. This implies independent call holding and call inter-arrival times. A study on the call holding and call inter-arrival times from three busy hours in 2001 from the E-Comm network [4] indicated that the call inter-arrival times show evidence of long-range dependence. We extend this study by analyzing longer traces from various

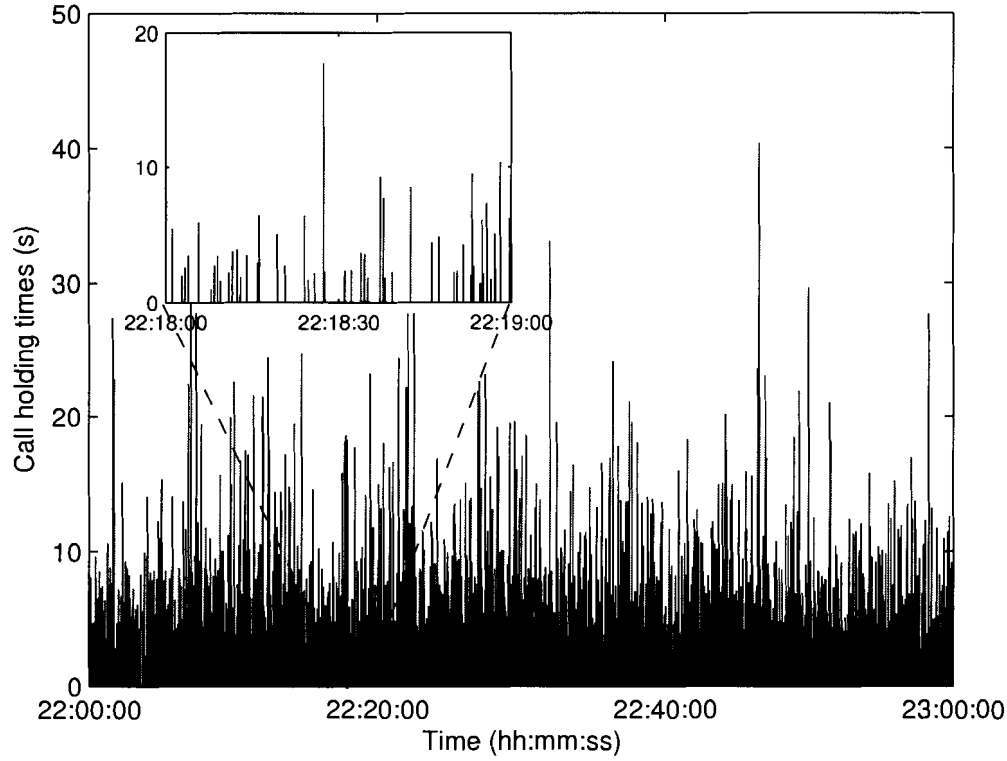

Figure 5.1: Time series of one busy hour of network traffic on March 26, 2003.

periods using the wavelet-based estimator of the Hurst parameter. The wavelet-based estimator is suitable because weekly traces contain a large number of entries (more than 370,000) and the employed discrete wavelet transform has a very low computational cost. Furthermore, we test the time constancy of the scaling exponent α in order to determine the reliability of the estimates of H.

5.2.1 Traffic traces from 2001

Traffic traces from 2001 contain information about the established calls in the Vancouver cell on November 1 and 2, 2001. The two daily traces are analyzed separately.

We create 48 hourly traces and determine the number of calls during each hour in order to identify the busiest hours. We then analyze five busiest hours in the two days.

5.2.1.1 2001: Analysis of daily traces

Estimates of the Hurst parameter for the timeseries of call holding and call inter-arrival times from the two daily traces in 2001 are summarized in Table 5.1. The table also shows the number of calls in each day. The number of calls is equal to the number of call holding times. It is one larger than the number of call inter-arrival times. The range of octaves where the linear regression is performed is determined by visual inspection of the logscale diagrams. In all four traces (two of call holding and two of call inter-arrival times), this linear region begins with octave 5 and includes the coarsest octaves.

For the traces of call inter-arrival times $H > 0.7$, which indicates presence of long-range dependence. For the trace of call holding times, $0.5 < H < 0.6$. This is an indicator of a weak long-range dependence. However, the values of H are close to 0.5 and, therefore, call holding times can be considered only short-range dependent.

We test the time constancy of the scaling exponent α by dividing the traces into m sub-traces, where $m \in \{14, 12, 10, 8, 6, 4, 3\}$. Results of the test are shown in Table 5.2. For each value of m, we report the outcome of the test. "N" stands for " not rejected", meaning that the trace passed the test for time constancy of α. Similarly, "R" stands for "rejected", or that the trace failed the test. The numbers of times the trace failed or passed the test are shown in the last two columns.

Both traces of call holding times pass the test for all 7 values of m, which implies

Table 5.1: 2001 daily traces: Hurst parameter estimates for the call holding and call inter-arrival times.

Day	Number of calls	Type of data	Range	H
01.11.2001	57,148	call holding times	5–13	0.583
		call inter-arrival times	5–13	0.732
02.11.2001	53,200	call holding times	5–13	0.561
		call inter-arrival times	5–13	0.737

Table 5.2: 2001 daily traces: results of the test for time constancy of α for the call holding and call inter-arrival times.

Day	Type of data	m							Rejected	Not rejected
		14	12	10	8	6	4	3		
01.11.2001	call holding	N	N	N	N	N	N	N	0	7
	call inter-arrival	R	R	R	R	R	R	R	7	0
02.11.2001	call holding	N	N	N	N	N	N	N	0	7
	call inter-arrival	R	R	R	R	N	N	N	4	3

that α can be considered constant throughout the traces and the estimated values of H are reliable. To the contrary, traces of call inter-arrival times fail the test for more than 50% of m's, which indicates unreliable estimates of α and, consequently, of H.

5.2.1.2 2001: Analysis of hourly traces

We create 48 hourly traces of call holding and call inter-arrival times from the two days of traffic data from 2001. We then determine the number of calls in each hour and identify the busiest hours. Estimation of the Hurst parameter and the test for time constancy of α are performed on the five busiest hours.

The logscale diagram for the trace of call holding times from the busy hour between

Table 5.3: 2001 hourly traces: Hurst parameter estimates for the call holding and call inter-arrival times.

Day/hour	Number of calls	Type of data	Range	H
02.11.2001 15:00–16:00	3,718	call holding times	2–9	0.493
		call inter-arrival times	4–9	0.907
01.11.2001 00:00–01:00	3,707	call holding times	2–9	0.471
		call inter-arrival times	4–9	0.802
02.11.2001 16:00–17:00	3,492	call holding times	2–9	0.462
		call inter-arrival times	4–9	0.770
01.11.2001 19:00–20:00	3,312	call holding times	2–9	0.467
		call inter-arrival times	4–9	0.774
02.11.2001 20:00–21:00	3,227	call holding times	2–9	0.479
		call inter-arrival times	4–9	0.663

16:00 and 17:00 on November 2, 2001 is shown in Figure 5.2. The slope of the linear regression is approximately zero, which yields Hurst parameter value of ~ 0.5. This indicates lack of long-range dependence in the call holding times.

Figure 5.3 shows the logscale diagram for the trace of call inter-arrival times from the busy hour between 16:00 and 17:00 on November 2, 2001. A linear region beginning from octave 4 and including the coarsest octaves can be clearly identified. The slope of the linear regression is greater than zero, which indicates presence of long-range dependence in the trace. The estimated value of the Hurst parameter is 0.770, as indicated in Table 5.3.

Estimates of the Hurst parameter for the five hourly traces of call holding and call inter-arrival times are shown in Table 5.3. The table also contains the number of calls in each of the busy hours. Hurst parameter estimates of the call holding times

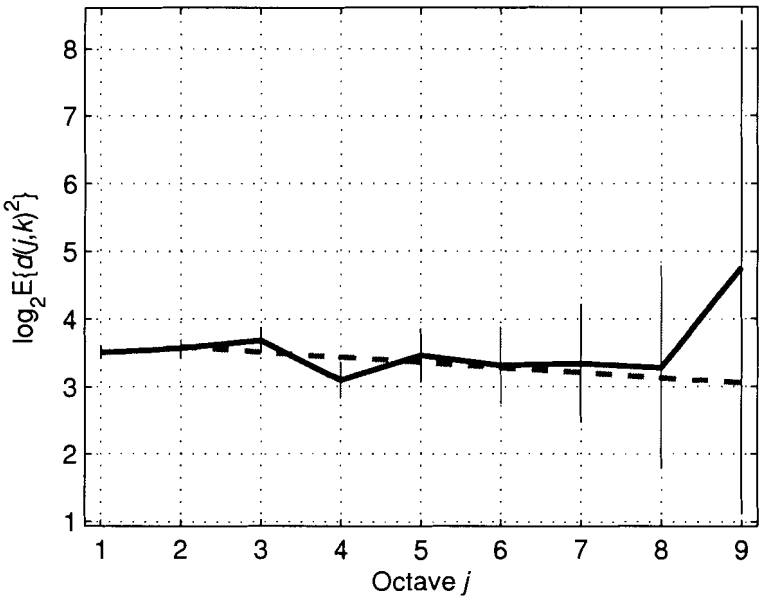

Figure 5.2: Novemer 2, 2001, busy hour 16:00–17:00: logscale diagram for the call holding times.

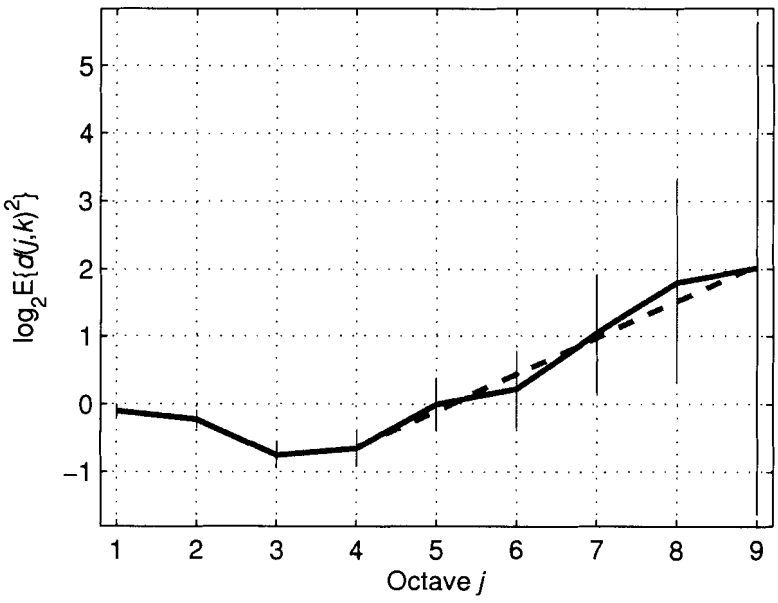

Figure 5.3: Novemer 2, 2001, busy hour 16:00–17:00: logscale diagram for the call inter-arrival times.

are close to 0.5. Hence, the call holding times are not long-range dependent. The estimates of H for the call inter-arrival times vary between 0.663 and 0.907, which indicates long-range dependence.

We perform the test for time constancy of α on the traces of call holding and call inter-arrival times. Results are shown in Tables 5.4 and 5.5, respectively. As indicated by the tables, all traces pass the test for time constancy of α for more than 50% of m's. Hence, values of α do not vary significantly across the traces and the estimates of the Hurst parameter can be considered reliable.

5.2.2 Traffic traces from 2002

The analyzed traffic data from 2002 span the period March 1–7, 2002. There were 370,510 established calls during this period. We removed from the trace of call holding times two outliers with values significantly larger than the other samples. In the trace of call inter-arrival times there was an outlier with a value greater than 3,000 s, which was due to missing traffic data between 8:59:06 and 9:49:52 on March 4, 2002. The outlier was removed prior to estimating the Hurst parameter. We first analyze the two weekly traces (one of call holding and one of call inter-arrival times). We then split each trace into seven daily traces and analyze the daily traces separately. Finally, we create and analyze five hourly traces from the busiest hours.

5.2.2.1 2002: Analysis of the weekly traces

We estimate the Hurst parameter of the weekly traces of call holding and call inter-arrival times and the logscale diagrams are shown in Figures 5.4 and 5.5, respectively. The linear region in the logscale diagram of the call holding times begins with octave 6

Table 5.4: 2001 hourly traces: results of the test for time constancy of α for the call holding times.

Day/hour	m						Rejected	Not rejected
	8	7	6	5	4	3		
02.11.2001/15:00–16:00	N	N	N	R	N	N	1	5
01.11.2001/00:00–01:00	N	R	N	N	N	N	1	5
02.11.2001/16:00–17:00	N	N	N	N	N	N	0	6
01.11.2001/19:00–20:00	N	N	N	N	N	N	0	6
02.11.2001/20:00–21:00	N	N	N	R	R	N	2	4

Table 5.5: 2001 hourly traces: results of the test for time constancy of α for the call inter-arrival times.

Day/hour	m						Rejected	Not rejected
	8	7	6	5	4	3		
02.11.2001/15:00–16:00	N	N	R	N	N	R	2	4
01.11.2001/00:00–01:00	N	N	N	N	N	N	0	6
02.11.2001/16:00–17:00	N	N	N	N	N	N	0	6
01.11.2001/19:00–20:00	N	N	N	N	N	N	0	6
02.11.2001/20:00–21:00	N	N	N	N	R	N	1	5

and extends across the coarsest octaves. The slope of the linear regression performed in the region is greater than zero. The estimated value of the Hurst parameter is 0.614. This value suggests presence of long-range dependence in the weekly trace of call holding times. The logscale diagram of the trace of call inter-arrival times, shown in Figure 5.5, shows evidence of *bi-scaling* (presence of two distinct linear regions with different slopes). The Hurst parameter estimates over the ranges of octaves [3–11] and [10–15] are 0.692 and 1.204, respectively.

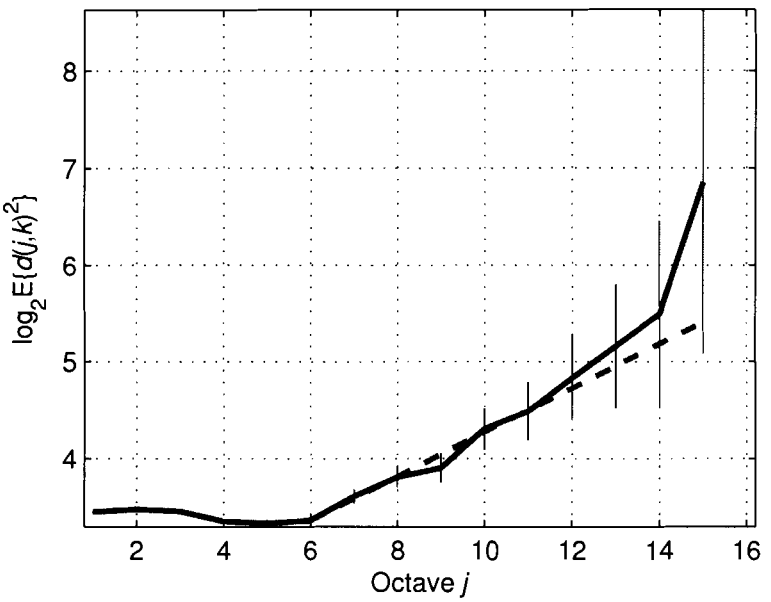

Figure 5.4: 2002 weekly trace: logscale diagram for the call holding times.

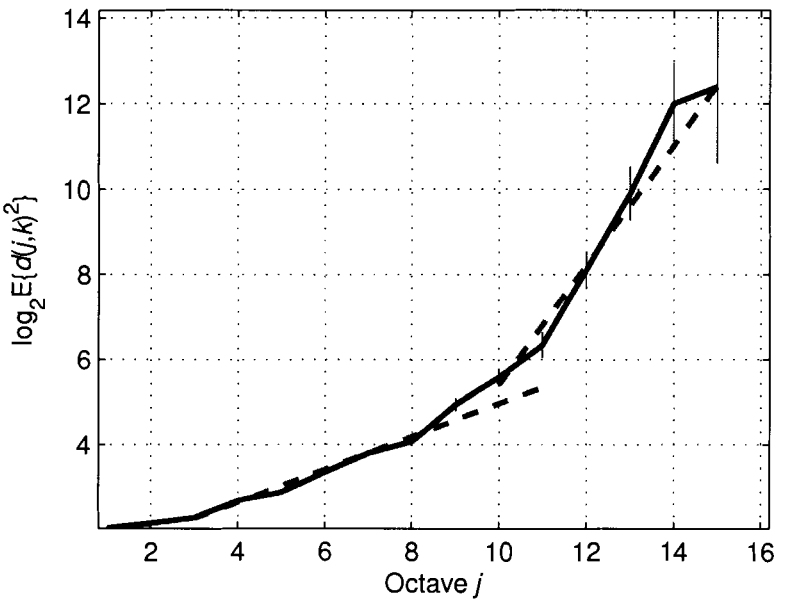

Figure 5.5: 2002 weekly trace: logscale diagram for the call inter-arrival times.

We test the time constancy of α for the 2002 weekly traces. In the case of call inter-arrival times, the test for time constancy of α is performed in the range of octaves [3–11] because there are not enough available octaves greater than 10 in the sub-traces. The values of m employed by the test are 100, 50, 40, 30, 20, and 10. The trace of call holding times passed the test for all m's. To the contrary, the trace of call inter-arrival times failed the test for all values of m. This indicates that the estimate of the Hurst parameter is reliable for the call holding times and is not reliable for the call inter-arrival times.

The graphical outputs from the test are shown in Figures 5.6 and 5.7. The top graph in each figure shows the timeseries of the corresponding trace. The remaining three graphs show the values of α in each sub-trace for $m = 100$ (second graph), $m = 50$ (third graph), and $m = 20$ (bottom graph). The solid horizontal line represents the overall value of α. The dashed horizontal line shows the average of the estimates of α. As observed in the top graph of Figure 5.7, call inter-arrival times exhibit daily cycles that are not present in call holding times (Figure 5.6). We assume that the cycles render the trace non-stationary, and, therefore, considering it long-range dependent and estimating the Hurst parameter may not produce meaningful results.

5.2.2.2 2002: Analysis of daily traces

For each day of the week between March 1 and March 7, 2002, Table 5.6 shows the number of calls and the Hurst parameter estimates for the traces of call holding and call inter-arrival times. It also shows the range of octaves where the linear regression is performed. Similarly to the results presented in Section 5.2.1.1, the linear region

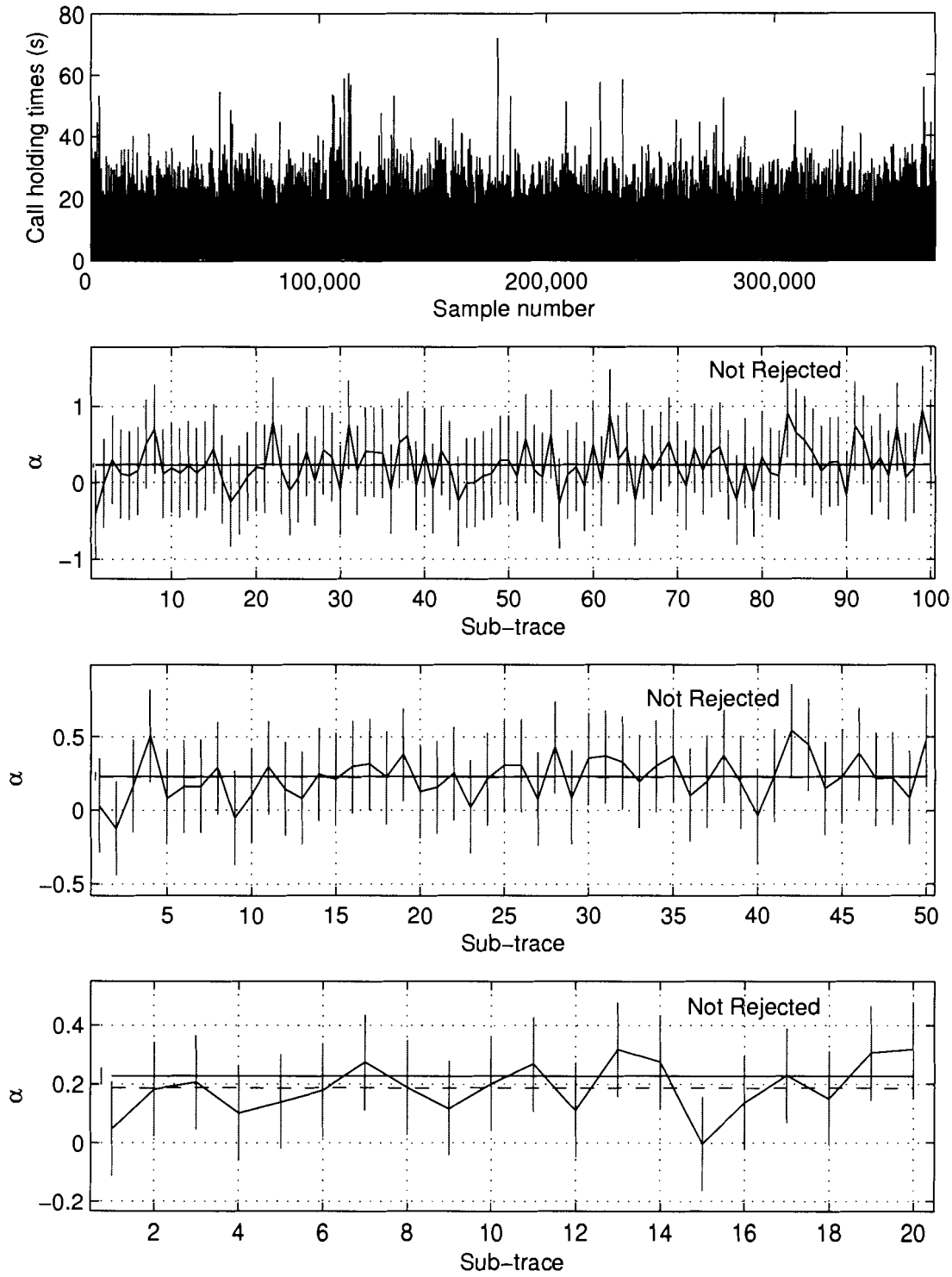

Figure 5.6: 2002 weekly trace: test for time constancy of α for the call holding times.

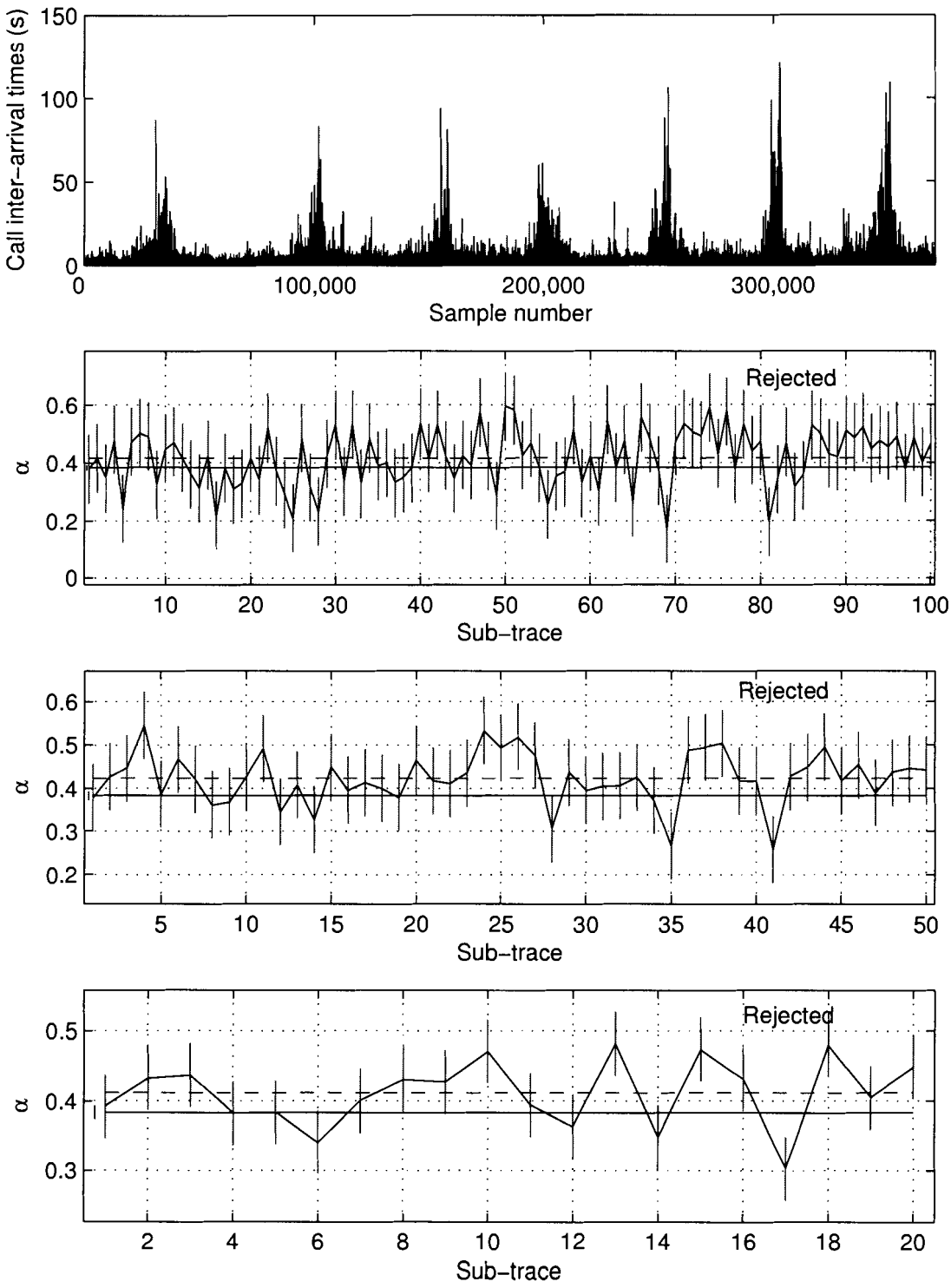

Figure 5.7: 2002 weekly trace: test for time constancy of α for the call inter-arrival times.

begins with octave 5 and extends across the coarsest octaves.

Call holding times show indication of a weak long-range dependence, with Hurst parameters between 0.5 and 0.6. The only exception is the trace from March 7, where the estimated Hurst parameter is 0.623. Hurst parameters of the call inter-arrival times are between 0.7 and 0.8, which indicates presence of long-range dependence in the traces.

We test the time constancy of α for $m \in \{14, 12, 10, 8, 6, 4, 3\}$. Results of the test are shown in Table 5.7. Traces of call holding times pass the test for all or almost all values of m. Of all traces of call inter-arrival times, one passes the test for all values of m, three pass for more than 50% of m's, and the remaining three pass the test for less than 50% of m's. This indicates that the Hurst parameter estimates

- for the traces of call holding times can be considered reliable;
- for the traces of call inter-arrival times from March 3, 4, and 6 may not be considered reliable;
- for the remaining four traces may be considered reliable.

5.2.2.3 2002: Analysis of hourly traces

We estimate the Hurst parameter and test the time constancy of α for traces of call holding and call inter-arrival times from the five busiest hours during the week March 1–7, 2002. Table 5.8 shows the busiest hours and the number of established calls during each hour. It also shows the Hurst parameter estimates together with the corresponding range of octaves where the linear regression is performed. As the table indicates, values of the Hurst parameter for the call holding times are close to 0.5, and,

Table 5.6: 2002 daily traces: Hurst parameter estimates for the call holding and call inter-arrival times.

Day	Number of calls	Type of data	Range	H
01.03.2002	63,464	call holding times	5–13	0.560
		call inter-arrival times	5–13	0.776
02.03.2002	57,339	call holding times	5–13	0.589
		call inter-arrival times	5–13	0.768
03.03.2002	53,685	call holding times	5–13	0.592
		call inter-arrival times	5–13	0.726
04.03.2002	48,803	call holding times	5–12	0.577
		call inter-arrival times	5–12	0.711
05.03.2002	52,949	call holding times	5–13	0.584
		call inter-arrival times	5–13	0.784
06.03.2002	49,752	call holding times	5–13	0.579
		call inter-arrival times	5–13	0.722
07.03.2002	44,518	call holding times	5–12	0.623
		call inter-arrival times	5–12	0.706

hence, the traces are not long-range dependent. Traces of call inter-arrival times are long-range dependent because the estimates of the Hurst parameter are significantly greater than 0.5.

Tables 5.9 and 5.10 summarize the results from the test for time constancy of α for the call holding times and call inter-arrival times, respectively. All ten traces pass the test for the majority of values of m, which indicates that the estimates of the Hurst parameter shown in Table 5.8 can be considered reliable.

Table 5.7: 2002 daily traces: results of the test for time constancy of α for the call holding and call inter-arrival times.

Day	Type of data	m							Rejected	Not rejected
		14	12	10	8	6	4	3		
01.03.2002	call holding	N	N	N	N	N	N	N	0	7
	call inter-arrival	N	N	R	R	N	R	N	3	4
02.03.2002	call holding	N	N	N	N	N	N	N	0	7
	call inter-arrival	R	N	N	N	N	N	R	2	5
03.03.2002	call holding	N	R	N	N	N	N	N	1	6
	call inter-arrival	R	R	R	R	N	R	N	5	2
04.03.2002	call holding	R	N	N	N	N	N	N	1	6
	call inter-arrival	N	N	R	N	R	R	R	4	3
05.03.2002	call holding	R	N	N	R	N	N	N	2	5
	call inter-arrival	N	N	N	N	N	N	N	0	7
06.03.2002	call holding	N	N	R	N	N	R	N	2	5
	call inter-arrival	N	R	N	R	R	R	R	5	2
07.03.2002	call holding	N	N	N	N	N	N	N	0	7
	call inter-arrival	N	N	N	N	N	N	R	1	6

5.2.3 Traffic traces from 2003

The analyzed traffic data from 2002 span the period March 24–30, 2003 and there were 387, 340 established calls during this period. From the trace of call holding times, we removed twelve outliers with values significantly larger than the other samples. Similarly to the analysis of data from 2002, we first analyze the two weekly traces, then we divide each into seven daily traces and analyze the daily traces separately. Finally, five hourly traces from the busiest hours are created and analyzed.

Table 5.8: 2002 hourly traces: Hurst parameter estimates for the call holding and call inter-arrival times.

Day/hour	Number of calls	Type of data	Range	H
01.03.2002 04:00–05:00	4,436	call holding times	2–9	0.490
		call inter-arrival times	4–9	0.679
01.03.2002 22:00–23:00	4,314	call holding times	2–9	0.460
		call inter-arrival times	4–9	0.757
01.03.2002 23:00–24:00	4,179	call holding times	2–9	0.489
		call inter-arrival times	4–9	0.780
01.03.2002 00:00–01:00	3,971	call holding times	2–9	0.508
		call inter-arrival times	4–9	0.741
02.03.2002 00:00–01:00	3,939	call holding times	2–9	0.503
		call inter-arrival times	4–9	0.747

5.2.3.1 2003: Analysis of the weekly traces

Hurst parameters of the weekly traces of call holding times and call inter-arrival times are estimated based on the logscale diagrams shown in Figures 5.8 and 5.9, respectively. In both logscale diagrams, *bi-scaling* can be observed. The distinct linear regions in the logcale diagram of the call holding times are in the ranges of octaves [5–12] and [10–15]. The corresponding values of the Hurst parameter are 0.592 and 0.751. Similarly, in the logscale diagram of the call inter-arrival times, the linear regions are within the ranges of octaves [3–11] and [10–15], with estimates of the Hurst parameter of 0.706 and 1.353, respectively.

We test the time constancy of α in order to determine whether or not the estimates of the Hurst parameter are reliable. The test for time constancy of α for the trace of call holding times is performed in the range of octaves [5–12] because of the lack of

Table 5.9: 2002 hourly traces: results of the test for time constancy of α for the call holding times.

Day/hour	m							Rejected	Not rejected
	10	8	7	6	5	4	3		
01.03.2002/04:00–05:00	N	N	N	N	N	N	N	0	7
01.03.2002/22:00–23:00	N	N	R	R	N	N	R	3	4
01.03.2002/23:00–24:00	R	N	N	N	R	N	N	2	5
01.03.2002/00:00–01:00	N	R	N	N	N	N	N	1	6
02.03.2002/00:00–01:00	R	N	N	N	N	N	N	1	6

Table 5.10: 2002 hourly traces: results of the test for time constancy of α for the call inter-arrival times.

Day/hour	m							Rejected	Not rejected
	10	8	7	6	5	4	3		
01.03.2002/04:00–05:00	R	N	N	N	N	N	N	1	6
01.03.2002/22:00–23:00	N	N	N	N	N	N	N	0	7
01.03.2002/23:00–24:00	N	R	N	N	N	N	N	1	6
01.03.2002/00:00–01:00	N	N	R	N	N	N	N	1	6
02.03.2002/00:00–01:00	N	N	N	N	N	N	N	0	7

available octaves greater than 10 in the sub-traces. For the same reason, we test the time constancy of α for the trace of call inter-arrival times in the range [3–11]. The values of m employed by the test are 100, 50, 40, 30, 20, and 10. The trace of call holding times passed the test for all m's, except for $m = 30$. Again, the trace of call inter-arrival times failed the test for all values of m. This indicates that the estimate of the Hurst parameter is reliable for the call holding times and is not reliable for the call inter-arrival times.

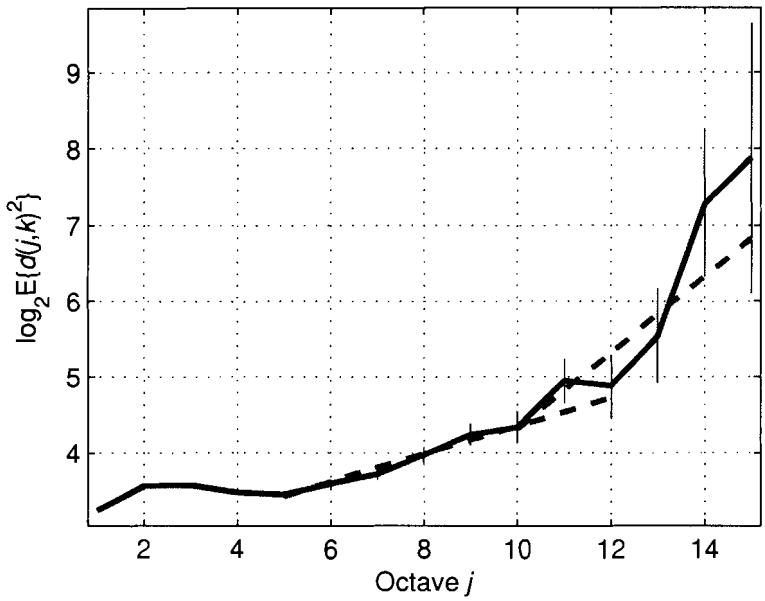

Figure 5.8: 2003 weekly trace: logscale diagram for the call holding times.

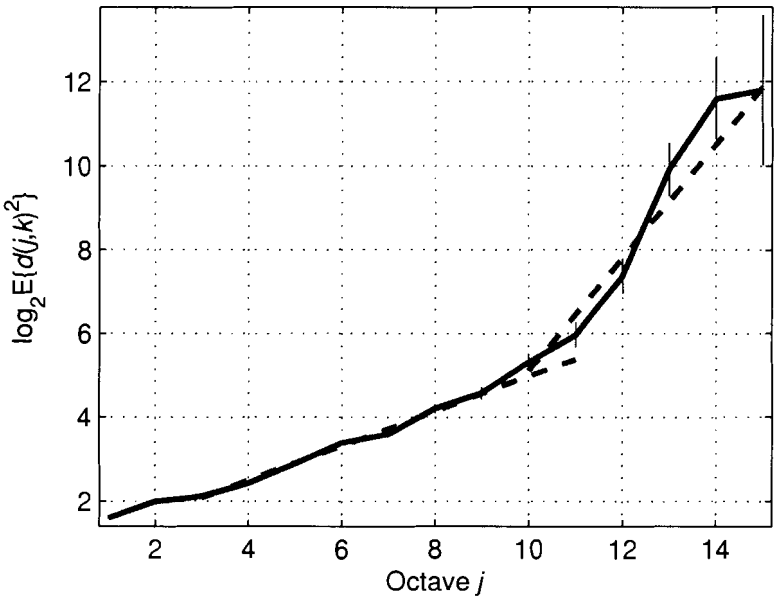

Figure 5.9: 2003 weekly trace: logscale diagram for the call inter-arrival times.

The graphical outputs from the test are shown in Figures 5.10 and 5.11. The top graph in each figure shows the timeseries of the corresponding trace. The remaining three graphs plot the values of α in each sub-trace for $m = 100$ (second graph), $m = 50$ (third graph), and $m = 20$ (bottom graph). The solid horizontal line represents the overall value of α and the dashed horizontal line shows the average of the estimates of α. As observed in the top graph of Figure 5.11, there are visible daily cycles in the trace of call inter-arrival times.

5.2.3.2 2003: Analysis of daily traces

We estimated the Hurst parameter of the daily traces from the week between March 24 and March 30, 2003. Examples of logscale diagrams for the traces of call holding and call inter-arrival times are shown in Figures 5.12 and 5.13, respectively. In both logscale diagrams, linear regions with slopes greater than zero that begin with octave 5 and include the coarsest octaves can be identified. This indicates presence of long-range dependence in the traces. Estimates of the Hurst parameter are presented in Table 5.11. The Hurst parameter of the traces of call holding times is approximately 0.6, implying weak long-range dependence. The traces of call inter-arrival times have higher values of the Hurst parameter, ranging between 0.7 and 0.8.

Table 5.12 summarizes the results of the test for time constancy of α. With the exception of the trace of call inter-arrival times from March 29, 2003, all traces pass the test for more than 50% of m's. Therefore, the majority of the estimates of H reported in Table 5.11 can be considered reliable.

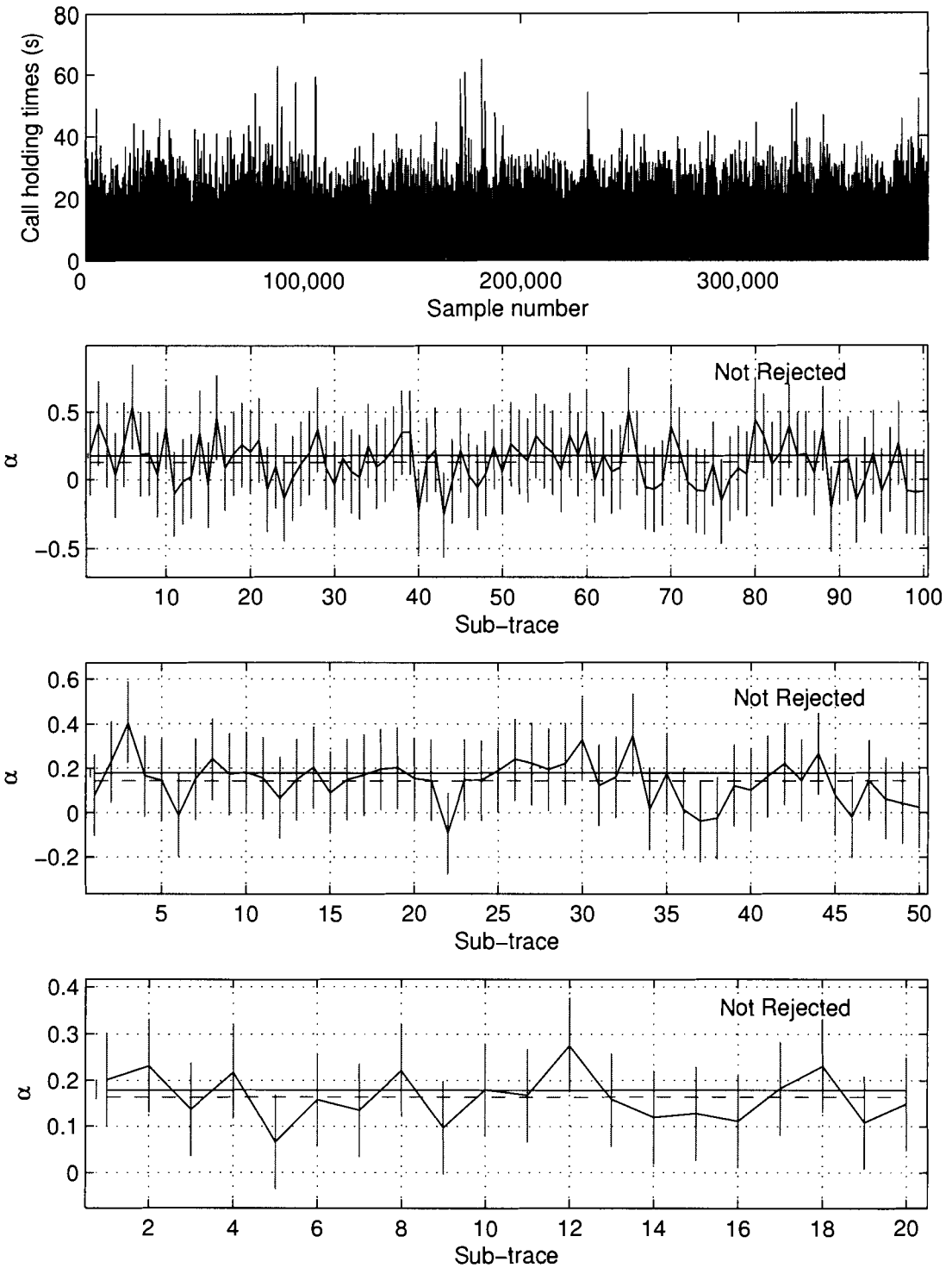

Figure 5.10: 2003 weekly trace: test for time constancy of α for the call holding times.

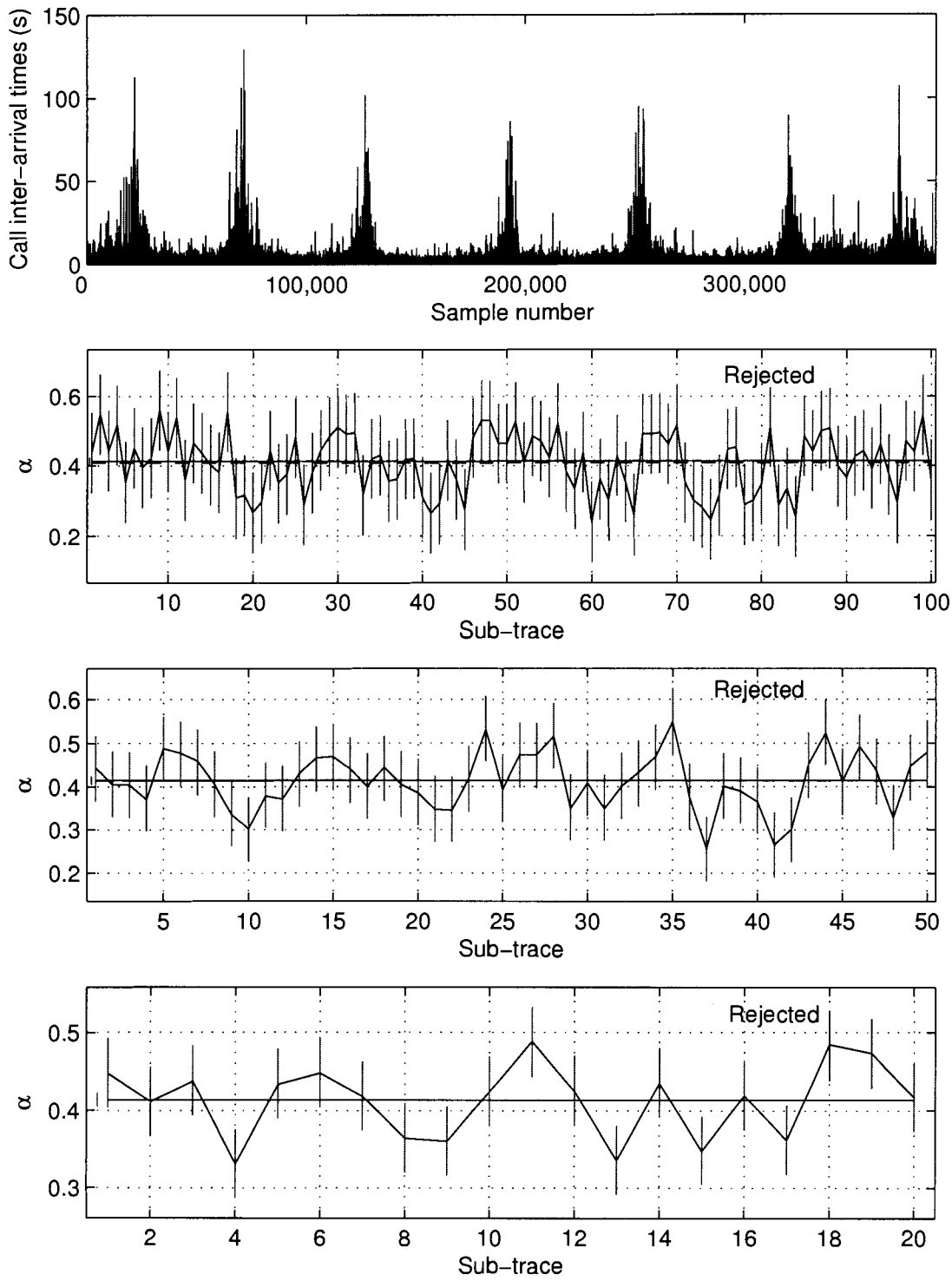

Figure 5.11: 2003 weekly trace: test for time constancy of α for the call inter-arrival times.

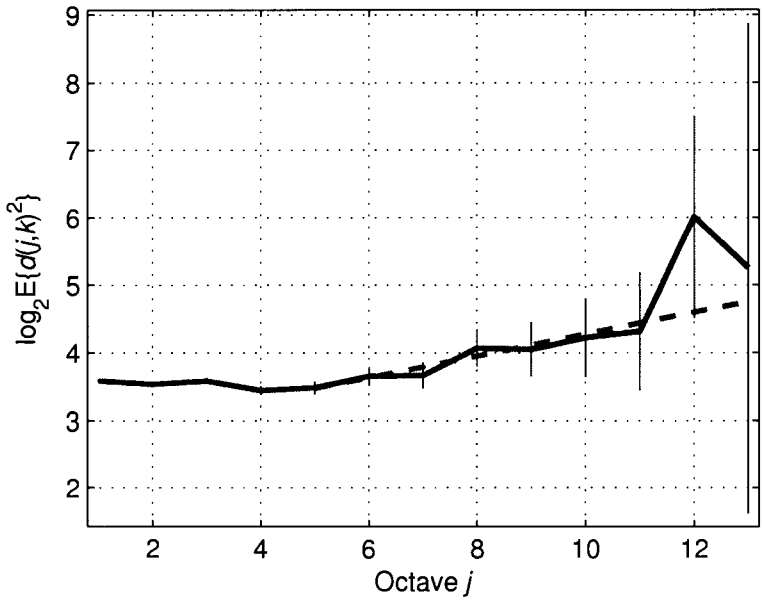

Figure 5.12: March 28, 2003: logscale diagram for the call holding times.

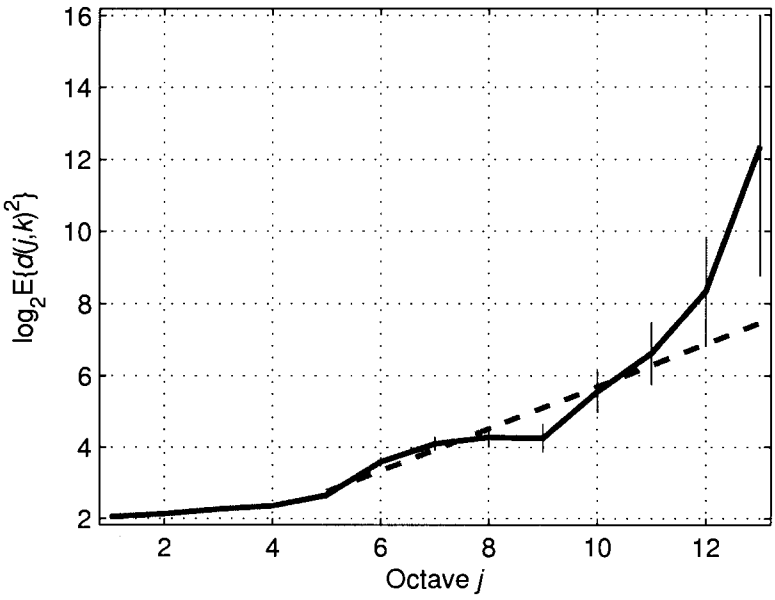

Figure 5.13: March 28, 2003: logscale diagram for the call inter-arrival times.

Table 5.11: 2003 daily traces: Hurst parameter estimates for the call holding and call inter-arrival times.

Day	Number of calls	Type of data	Range	H
24.03.2003	45,509	call holding times	5–13	0.617
		call inter-arrival times	3–12	0.715
25.03.2003	51,076	call holding times	5–13	0.587
		call inter-arrival times	5–13	0.775
26.03.2003	62,720	call holding times	5–13	0.598
		call inter-arrival times	5–13	0.764
27.03.2003	64,092	call holding times	5–13	0.605
		call inter-arrival times	5–13	0.708
28.03.2003	55,277	call holding times	5–13	0.581
		call inter-arrival times	5–13	0.794
29.03.2003	60,272	call holding times	5–13	0.571
		call inter-arrival times	5–13	0.685
30.03.2003	48,394	call holding times	5–12	0.566
		call inter-arrival times	5–12	0.739

5.2.3.3 2003: Analysis of hourly traces

We estimate the Hurst parameter and test the time constancy of α for traces of call holding and call inter-arrival times from the five busiest hours during the week March 24–30, 2003. Table 5.13 shows the busiest hours and the number of established calls in each. It also shows the Hurst parameter estimates together with the corresponding range of octaves where the linear regression is performed. Values of the Hurst parameter for the call holding times are close to 0.5, and, hence, the traces are not long-range dependent. Traces of call inter-arrival times are long-range dependent because the estimates of the Hurst parameter are significantly greater than 0.5.

Table 5.12: 2003 daily traces: Results of the test for time constancy of α for the call holding and call inter-arrival times.

Day	Type of data	m							Rejected	Not rejected
		14	12	10	8	6	4	3		
24.03.2003	call holding	N	N	N	N	N	N	N	0	7
	call inter-arrival	R	N	N	N	N	N	N	1	6
25.03.2003	call holding	N	N	N	N	N	N	N	0	7
	call inter-arrival	N	N	N	N	N	N	N	0	7
26.03.2003	call holding	N	N	N	N	N	N	N	0	7
	call inter-arrival	N	N	N	R	N	R	R	3	4
27.03.2003	call holding	N	N	N	N	N	N	N	0	7
	call inter-arrival	N	N	N	N	N	N	N	0	7
28.03.2003	call holding	N	R	N	N	N	N	N	1	6
	call inter-arrival	R	N	N	N	N	R	N	2	5
29.03.2003	call holding	N	N	N	N	N	N	N	0	7
	call inter-arrival	N	R	R	N	N	R	R	4	3
30.03.2003	call holding	N	N	N	N	N	N	N	0	7
	call inter-arrival	N	R	N	N	R	N	R	3	4

Tables 5.14 and 5.15 summarize the results from the test for time constancy of α for the call holding times and call inter-arrival times, respectively. Nine traces pass the test for the majority of values of m. The trace of call holding times from the busy hour between 23:00 and 24:00 on March 26, 2003 is the only trace that fails the test for more than 50% of m's. This indicates that most estimates of the Hurst parameter shown in Table 5.13 can be considered reliable.

Table 5.13: 2003 hourly traces: Hurst parameter estimates for the call holding and call inter-arrival times.

Day/hour	Number of calls	Type of data	Range	H
26.03.2003 22:00–23:00	4,919	call holding times	2–9	0.483
		call inter-arrival times	4–9	0.788
25.03.2003 23:00–24:00	4,249	call holding times	2–9	0.483
		call inter-arrival times	4–9	0.832
26.03.2003 23:00–24:00	4,222	call holding times	2–9	0.463
		call inter-arrival times	4–9	0.699
29.03.2003 02:00–03:00	4,150	call holding times	2–9	0.526
		call inter-arrival times	4–9	0.696
29.03.2003 01:00–02:00	4,097	call holding times	2–9	0.466
		call inter-arrival times	4–9	0.705

5.3 Summary and discussion

We created and analyzed traffic traces containing call holding times and call inter-arrival times from E-Comm's traffic data. We employed the wavelet-based estimator of the Hurst parameter and tested the time constancy of the scaling exponent α in order to determine whether or not the estimates of H are reliable. The analysis was performed on the weekly, daily, and hourly traces.

Both weekly traces of call holding times pass the test for time constancy of α. Hence, the estimates of H in the corresponding ranges of octaves, can be considered reliable. The traces of call inter-arrival times, however, fail the test and, hence, the estimates of H may not be reliable. Furthermore, as shown in Figures 5.5 and 5.9, logscale diagrams of the call inter-arrival times show evidence of *bi-scaling* and the Hurst parameter estimates in the range of octaves including the coarsest are

Table 5.14: 2003 hourly traces: Results of the test for time constancy of α for the call holding times.

Day/hour	m							Rejected	Not rejected
	10	8	7	6	5	4	3		
26.03.2003/22:00–23:00	N	R	N	N	R	N	N	2	5
25.03.2003/23:00–24:00	N	N	N	N	N	N	N	0	7
26.03.2003/23:00–24:00	N	R	R	R	R	N	N	4	3
29.03.2003/02:00–03:00	N	N	N	N	N	N	N	0	7
29.03.2003/01:00–02:00	N	N	N	N	N	N	N	0	7

Table 5.15: 2003 hourly traces: Results of the test for time constancy of α for the call inter-arrival times.

Day/hour	m							Rejected	Not rejected
	10	8	7	6	5	4	3		
26.03.2003 / 22:00–23:00	N	N	N	N	N	N	N	0	7
25.03.2003 / 23:00–24:00	N	R	N	N	R	N	N	2	5
26.03.2003 / 23:00–24:00	N	N	N	N	N	R	N	1	6
29.03.2003 / 02:00–03:00	N	N	N	N	N	N	N	0	7
29.03.2003 / 01:00–02:00	N	N	N	N	N	N	N	0	7

greater than 1.0. We assume that Hurst parameter estimates greater than 1.0 may be attributed to the non-stationarities in the traces that result from the presence of daily cycles.

Daily traces of call holding times pass the test for time constancy of α. Hurst parameter estimates of the traces are usually between 0.5 and 0.6 and may be considered reliable. This indicates that the traces are weakly long-range dependent. In the case of the daily traces of call inter-arrival times, the outcome of the test varies.

Both traces from 2001 fail the test for more than 50% of m's, which implies possibly unreliable estimates. Four out of seven traces from 2002, and six out of seven traces from 2003 pass the test for more than 50% of m's. Although not all estimates of H can be considered reliable, their values are consistent. All but one estimates are between 0.7 and 0.8. The remaining one has a value of 0.685, which is still close to 0.7.

Analyzed hourly traces of both call holding and call inter-arrival times pass the test for time constancy of α for the majority of m's. The only exception is the trace of call holding times from the busy hour between 23:00 and 24:00 on March 26, 2003, which passed the test for three and failed for four values of m. Hurst parameter estimates of the call holding times are close to 0.5, which implies absence of long-range dependence. Call inter-arrival times show evidence of long-range dependence because all estimates are greater than 0.5 and a majority of them are greater than 0.75. These results are in good agreement with already reported findings [4].

We compared the Hurst parameter estimates of the daily and hourly traces from the various datasets in order to determine their fluctuations. Figure 5.14 shows the estimates of the Hurst parameter of the daily traces from 2002 and 2003. In order to better observe if the estimates follow certain trend, we sorted the series of estimates from the corresponding year before plotting. The horizontal axis represents the rank (position in the sorted series of estimates) and the vertical axis shows the values of the estimates. For example, points in the graph with a rank of two show the second smallest Hurst parameter estimate in each series. The differences between the largest and the smallest estimates are approximately 0.05 and 0.1 for the call holding and call inter-arrival times, respectively. As shown in the graph, Hurst parameter estimates

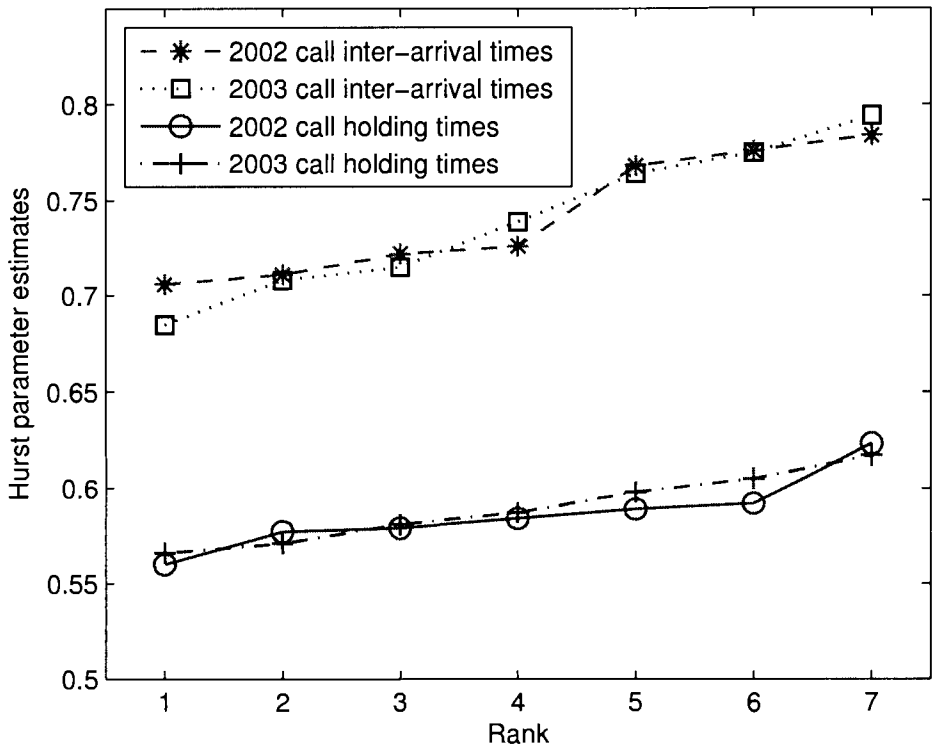

Figure 5.14: Daily traces from 2002 and 2003: Hurst parameter estimates.

for the daily traces of call holding and call inter-arrival times from 2002 and 2003 are very similar. This implies that the Hurst parameters of the daily traces remained unchanged between 2002 and 2003.

Figure 5.15 shows the sorted series of Hurst parameter estimates for the hourly traces of call holding and call inter-arrival times from 2001, 2002, and 2003. Estimates of H of the traces of call holding times from the three years are very close to each other. Similarly to the case of daily traces, the difference between the largest and the smallest estimate is approximately 0.05. Estimates of H of the hourly traces of call inter-arrival times exhibit greater variability than for the daily traces. The difference between the largest and the smallest estimate is ~ 0.2 and, except for the

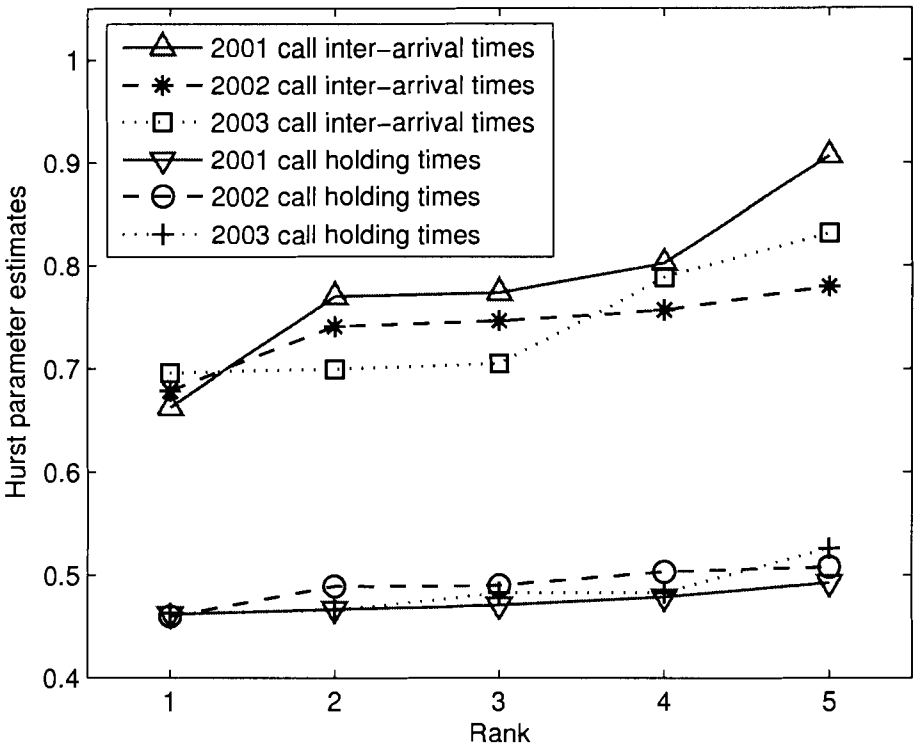

Figure 5.15: Hourly traces from 2001, 2002, and 2003: Hurst parameter estimates.

smallest estimate, estimates of H for the busiest hours from 2001 are greater than the corresponding values from 2002 and 2003.

We could not observe large differences or any trend (increase/decrease of H across the years) when comparing the Hurst parameter estimates of the daily and hourly traces from the three years. This implies that the Hurst parameter can be regarded as an invariant characteristic of the traffic traces from E-Comm's network. Based on the estimates, we conclude that:

1. daily and hourly traces of call inter-arrival times are long-range dependent with Hurst parameters of approximately 0.75

2. daily traces of call holding times show evidence of weak long-range dependence,

with Hurst parameters of 0.55–0.6

3. hourly traces of call holding times are independent.

Chapter 6

Conclusions

The Hurst parameter of a stochastic process characterizes the type of correlation structure of the process: independence, short-range, or long-range dependence. Its reliable and efficient estimation is important for the statistical analysis of the processes. When the value of the Hurst parameter is between 0.5 and 1, the process is long-range dependent. Long-range dependent processes exhibit similar behaviour when observed on various time-scales. For that reason, wavelets, with their natural scale invariance, are suitable for analyzing such processes.

In this thesis, we applied the wavelet-based Hurst parameter estimator to two datasets: MPEG-1 and MPEG-4 encoded video sequences and traffic traces from a deployed circuit-switched wireless network. We also investigated the reliability of the estimates.

The wavelet-based estimator produced values of H that were often greater than one when applied to the MPEG video sequences. We examined the possible sources of unreliability and concluded that the unreliable performance may be attributed to

presence of both strong short-range and long-range components in the traces and to the time variability of the scaling exponent α. Furthermore, comparing various estimates of H for the MPEG video traces, we conjectured that their power spectral density function has a power-law behaviour with an exponent greater than one, which does not comply with the LRD assumption .

We also applied wavelet-based estimation of the Hurst parameter to traces of call holding and call inter-arrival times from a circuit-switched network. We analyzed weekly, daily, and hourly traces from 2001, 2002, and 2003. We tested the time constancy of α and concluded that, except for the weekly traces of call inter-arrival times, α can be considered constant and the estimates of H reliable. Both daily and hourly traces of call inter-arrival times are long-range dependent with Hurst parameters of approximately 0.75. Daily traces of call holding times are weakly long-range dependent, with H ranging from 0.55 to 0.6. Hourly traces of call holding times are independent. We observed no trend when comparing the Hurst parameter estimates from 2001, 2002, and 2003. We concluded that the estimated values of the Hurst parameter can be considered invariant of the traffic traces.

There are two key areas of future research regarding the unreliable performance of the wavelet-based estimator of H. The estimator is unbiased due to an explicitly inserted bias correcting factor that is calculated based on certain assumptions, such as Gaussianity. If these assumptions do not hold, the bias may become non-negligible and the estimates of H unreliable. A possible improvement would be to introduce a more robust correction of the bias or adopt an approach that does not introduce bias. Furthermore, it is important to investigate the correlation structure of the traces under consideration and the behaviour of their power spectral density. The hyperbolically

decaying autocorrelation functions are non-summable and, therefore, fit the concept of long-range dependence. However, they are not the only non-summable functions. A possible future work may be finding functions that better model the autocorrelation function of the traces within the long-range dependence framework. Since the wavelet-based estimator of H is based on the power-law behaviour of the power spectral density (equivalent to a hyperbolically decaying autocorrelation function), the estimator may fail if the traces possess a different type of autocorrelation. In this case, a fundamental modification of the existing estimator may be necessary.

Appendix A

Other estimators of the Hurst parameter

There are several estimators for the Hurst parameter of a stochastic process, such as R/S plot, periodogram, variance-time plot, Whittle and wavelet-based. In Chapter 3, we described the wavelet-based estimator. In this Appendix, we describe two additional popular estimators that have been used to estimate H of the MPEG video traces: R/S plot and periodogram [11], [29], [33].

A.1 R/S plot

Let X_i, $i = 1, 2, \ldots, N$, be a discrete stochastic process and $Y_j = \sum_{i=1}^{j} X_i$ be its cumulative process. For every k, $0 \leq k \leq N$, called *lag*, and every starting point t such that $t + k \leq N$,

$$R(t,k) = \max_{0 \leq i \leq k} \left[Y_{t+i} - Y_t - \frac{i}{k} \left(Y_{t+k} - Y_t \right) \right] - \min_{0 \leq i \leq k} \left[Y_{t+i} - Y_t - \frac{i}{k} \left(Y_{t+k} - Y_t \right) \right] \quad \text{(A.1)}$$

is called the *adjusted range*. When $R(t,k)$ is normalized by the square root of the sample variance of the sub-series $X_{t+1},\ldots,X_{t+k}$

$$S(t,k)=\sqrt{\frac{1}{k}\sum_{i=t+1}^{t+k}\left(X_i-\bar{X}_{t,k}\right)^2}, \tag{A.2}$$

where $\bar{X}_{t,k}=k^{-1}\sum_{i=t+1}^{t+k}X_i$, the statistic thus obtained is called *rescaled adjusted range*, or R/S [11], [33], [34]:

$$R/S(t,k)=\frac{R(t,k)}{S(t,k)}. \tag{A.3}$$

If X_i is long-range dependent with Hurst parameter H, then for large k,

$$\log E\left[R/S\right]\approx a+H\log k. \tag{A.4}$$

To estimate H, the series X_i is first divided into K blocks of size N/K. For each lag k and starting points $t_i=iN/K+1$, $i=1,2,\ldots,$ and $t_i+k\leq N$, the values of $R/S(t_i,k)$ are calculated. When k is smaller than N/K, there are K values of the R/S statistic for every k. For the largest lags k, there is only one value of R/S. In practice, k's are logarithmically spaced and the estimates of R/S for the smallest and the largest k's are ignored. The plot of $\log R/S$ versus $\log k$ is called *pox plot*. H is estimated as the slope of the line fitted to the points in the pox plot.

Figure A.1 shows the pox plot of the MPEG-4 encoded "Star Wars" video sequence. The black diamonds mark the points considered in the estimation of H. The two dotted lines show the reference slopes of 1 and 0.5. Hurst parameter is estimated to be 0.955.

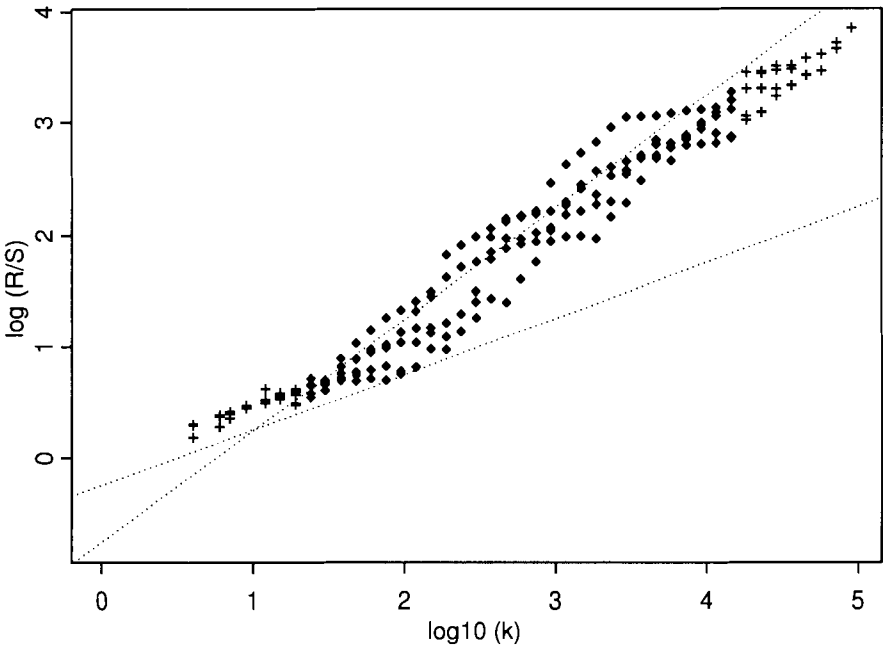

Figure A.1: Graphical output of the R/S plot for the MPEG-4 encoded "Star Wars IV" video sequence.

A.2 Periodogram

The periodogram of a discrete stochastic process X_i, $i = 1, 2, \ldots, N$, is defined as

$$I(\nu) = \frac{1}{2\pi N} \left| \sum_{j=1}^{N} X(j) e^{ij\nu} \right|^2 , \tag{A.5}$$

where ν is the frequency [29]. When X_i is wide-sense stationary, then $I(\nu)$ is an estimator of its power spectral density function (PSD). The PSD of long-range dependent processes follows a power-law for low frequencies, as shown in Eq. (2.9). This implies that there is a linear relationship between $\log I(\nu)$ and $\log \nu$ when $\nu \to 0$ with a slope equal to $-\alpha$. Therefore, α is estimated by performing linear regression of $\log I(\nu)$ on $\log \nu$. In practice, only the lowest 10% of the frequencies are considered. The Hurst

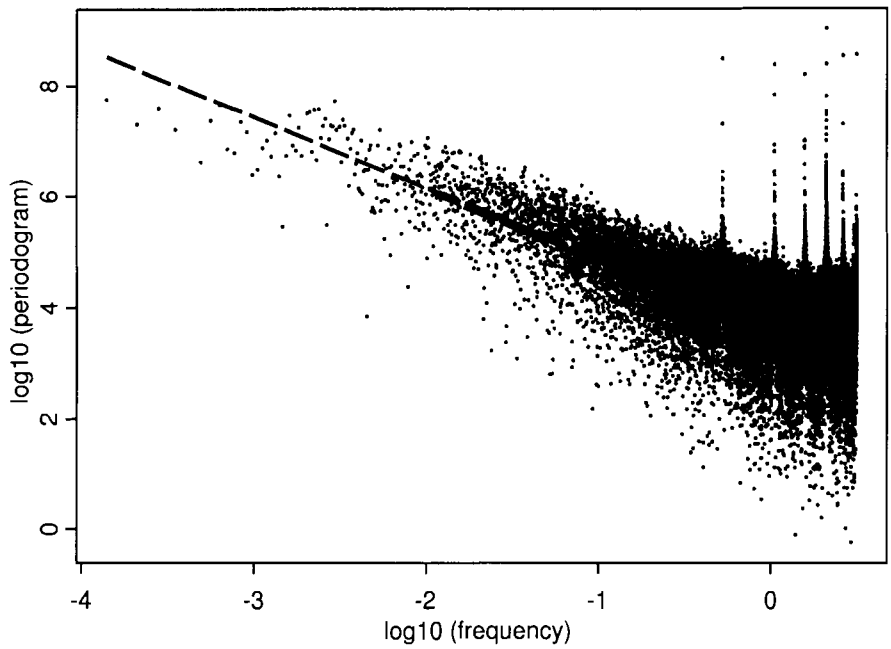

Figure A.2: Graphical output of the periodogram for the MPEG-4 encoded "Star Wars IV" video sequence.

parameter is calculated by employing Eq. (2.12).

Figure A.2 shows the plot of the periodogram of the MPEG-4 encoded "Star Wars" video sequence. The dashed line represents the result of the linear regression. Estimated value of the Hurst parameter is 1.138.

Bibliography

[1] V. Frost and B. Melamed, "Traffic modeling for communication networks," *IEEE Communications Magazine*, vol. 33, pp. 70–80, Mar. 1994.

[2] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, pp. 1–15, Feb. 1994.

[3] V. Paxson and S. Floyd, "Wide area traffic: the failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, June 1995.

[4] D. Sharp, N. Cackov, N. Lasković, Q. Shao, and Lj. Trajković, "Analysis of public safety traffic on trunked land mobile radio systems," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 7, pp. 1197–1205, Sept. 2004.

[5] M. Reisslein, J. Lassetter, S. Ratnam, O. Lotfallah, F. H. P. Fitzek, and S. Panchanathan, "Traffic and quality characterization of scalable encoded video: a large-scale trace-based study, Part 1: Overview and definitions," Arizona State University, Telecommunications Research Center, Technical Report, Dec. 2002.

[6] M. Reisslein, J. Lassetter, S. Ratnam, O. Lotfallah, F. H. P. Fitzek, and S. Panchanathan, "Traffic and quality characterization of scalable encoded video: a large-scale trace-based study, Part 2: Statistical analysis of single-layer encoded video," Arizona State University, Telecommunications Research Center, Technical Report, Jan. 2003.

[7] F. H. P. Fitzek and M. Reisslein, "MPEG-4 and H.263 video traces for network performance evaluation," Technical University Berlin, Telecommunication Networks Group, Technical Report TKN–00–06, Oct. 2000.

[8] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Transactions on Communications*, vol. 43, no. 2/3/4, pp. 1566–1579, Feb./Mar./Apr. 1995.

[9] O. Rose, "Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems," University of Würzburg, Institute of Computing Science, Report No. 101, Feb. 1995.

[10] S. H. Hong, R.-H. Park, and C. B. Lee, "Hurst parameter estimation of long-range dependent VBR MPEG video traffic in ATM networks," *Journal of Visual Communication and Image Representation*, vol. 12, no. 1, pp. 44–65, Mar. 2001.

[11] J. Beran, *Statistics for long memory processes.* New York, NY: Chapman and Hall, 1994, pp. ix–x, 1–11, 20–59, 81–87.

[12] P. Abry and D. Veitch, "Wavelet analysis of long-range dependent traffic," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 2–15, Jan. 1998.

[13] D. Veitch and P. Abry, "A wavelet-based joint estimator of the parameters of long-range dependence," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 878–897, Apr. 1999.

[14] M. Roughan and D. Veitch, "Measuring long-range dependence under changing traffic conditions," in *Proceedings IEEE INFOCOM '99*, New York, NY, Mar. 1999, vol. 3, pp. 1513–1521.

[15] Ž. Lučić, "Wavelet based estimators of long-range dependencies in traffic traces," Simon Fraser University, Vancouver, School of Engineering Science, Master of Engineering Project, Apr. 2002.

[16] N. Cackov, Ž. Lučić, M. Bogdanov, and Lj. Trajković, "Wavelet-based estimation of long-range dependence in MPEG video traces," to be presented at *IEEE International Symposium on Circuits and Systems*, Kobe, Japan, May 2005.

[17] G. Hess, *Land-mobile radio system engineering.* Norwood, MA: Artech House, 1993, pp. 249–286.

[18] G. Stone, "Public safety wireless communications user traffic profiles and grade of service recommendations," U.S. Dept. Justice, SRSC Final Report, Appendix D, Mar. 1996.

[19] E-Comm, Emergency communications for southwest British Columbia, Incorporated. (2004, September). [Online]. Available: http://www.ecomm.bc.ca.

[20] K. Park and W. Willinger, "Self-similar network traffic: an overview," *Self-similar Network Traffic and Performance Evaluation*, edited by K. Park and W. Willinger. New York, NY: Wiley, 2000, pp. 1–38.

[21] P. Abry, P. Flandrin, M. S. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation, and synthesis of scaling data," *Self-similar Network Traffic and Performance Evaluation*, edited by K. Park and W. Willinger. New York, NY: Wiley, 2000, pp. 39–88.

[22] D. Veitch and P. Abry, "A statistical test for the time constancy of scaling exponents," *IEEE Transactions on Signal Processing*, vol. 49, no. 10, pp. 2325–2334, Oct. 2001.

[23] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Magazine*, vol. 8, issue 4, pp. 14–38, Oct. 1991.

[24] D. Veitch, MATLAB code for estimation of scaling exponents. (2004, September). [Online]. Available: http://www.cubinlab.ee.mu.oz.au/~darryl/secondorder_code.html.

[25] D. Veitch, MATLAB code for the estimation of MultiScaling Exponents. (2004, September). [Online]. Available:
http://www.cubinlab.ee.mu.oz.au/~darryl/MS_code.html.

[26] J. Watkinson, *The MPEG handbook: MPEG-1, MPEG-2, MPEG-4*. Boston, MA: Focal Press, 2001, Chapter 5.

[27] University of Würzburg, Index of MPEG traces. (2004, November). [Online]. Available:
http://www-info3.informatik.uni-wuerzburg.de/MPEG/traces/.

[28] University of Berlin, MPEG-4 and H.263 video traces for network performance evaluation. (2004, June). [Online]. Available:
http://www-tkn.ee.tu-berlin.de/research/trace/trace.html.

[29] M. S. Taqqu, Statistical methods for long-range dependence: Periodogram. (2004, November). [Online]. Available:
http://math.bu.edu/people/murad/methods/per/.

[30] F. Xue and Lj. Trajković, "Performance analysis of a wavelet-based Hurst parameter estimator for self-similar network traffic," in *Proceedings SPECTS '2K*, Vancouver, BC, May 2000, pp. 294–298.

[31] Engineering statistics. (2004, November). [Online]. Available:
http://www.itl.nist.gov/div898/handbook/.

[32] N. Cackov, B. Vujičić, S. Vujičić, and Lj. Trajković, "Using network activity data to model the utilization of a trunked radio system," in *Proceedings SPECTS '04*, San Jose, CA, July 2004, pp. 517–524.

[33] M. S. Taqqu, Statistical methods for long-range dependence: R/S plot. (2004, November). [Online]. Available:
http://math.bu.edu/people/murad/methods/rs/.

[34] B. B. Mandelbrot and M. S. Taqqu, "Robust R/S analysis of long-run serial correlations," *Proceedings of the 42nd Session of the International Statistical*

Institute, Manila, Dec. 1979, Bulletin of the International Statistical Institute, vol. 48, book 2, pp. 69–104.