# A Parameterized Long Range Dependence Trace Generator

Abdullah Balamash

Department of Electrical Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia
asbalamesh@kau.edu.sa

**Abstract:** Many real data in different disciplines follow a self-similar model that is based on the Hurst parameter. Self-similar models exhibits the long range dependence feature (LRD). Many simulation experiments rely on real LRD traces instead of assuming the traditional Poisson model. Such traces are not always available and can be difficult to obtain. Accordingly, there is a need for synthetic LRD data generators. In this work, we present a simple and fast LRD data generator based of the Haar wavelet transform. The generator is parametric where the only inputs needed are the mean, the variance, and the Hurst parameter of the target synthetic trace. We prove the goodness of our model through comprehensive simulation results.

[Abdullah Balamash. **A Parameterized Long Range Dependence Trace Generator.** *Life Sci J* 2013;10(3):2446-2449] (ISSN: 1097-8135). http://www.lifesciencesite.com. 354

**Keywords:** Data modeling; data generation; Long range dependence; Self-similarity; Wavelet

## 1. Introduction

Self-similarity has become the key model in describing numerous physical phenomenon including communication network traffic, flow rates in rivers (Hosking 1984), economical, geographical, and biological data (Mandelbrot 1967, Mandelbrot and Van Ness 1968, Wornell 1993). The usual Poisson model has failed to capture the main features of such data (Paxson and Floyd 1995). For example, the average queue size at a given network device was found to be higher than the expected one assuming Poisson arrival process (Shugong, Xinwei et al. 1998). The reason is that the Poisson model assumes that the data are uncorrelated while in most of these phenomenon the data exhibits an LRD nature (long term temporal correlation) (Karagiannis, Molle et al. 2004).

Mathematical analysis of self-Similar processes is difficult and is intractable and accordingly, researchers depend on simulation for mining information from real data sequences that are not always available and can be costly or impossible to capture. In this paper we present a simple and fast algorithm to synthesis LRD time series.

There are several methods for generating LRD synthetic traces. Some of these methods are very slow and need high processing powers to generate a moderate length trace (for example; the fractional ARIMA method (Hosking 1984) and the Mandelbrot method (Mandelbrot 1971)). Others are approximate methods (Lau, Erramilli et al. 1995). Our algorithm is based on some properties of the Haar Wavelet transform similar as in Riedi et al. method (Riedi, Crouse et al. 1999) that assumes the knowledge of the second moment of the wavelet coefficients at each scale, which our model does not require.

The strength of our method lies in its simplicity where the inputs to our generator are only the mean, the variance, and the Hurst parameter of the target synthetic trace, which makes it a very simple model. Moreover, our algorithm is very fast where we could generate traces of more than million data points in less than a second and for this reason, we do not intend to compare the processing time of our algorithm with the others, but will prove the accuracy of our algorithm through comprehensive synthetic trace generation comparing the intended Hurst parameters of the generated traces with their estimated ones.

The rest of the paper is organized as follows. Section 2 is a background about the self-similar processes and the concept of LRD. We present our data generator in section 3. In section 4, we present some experimental results showing the goodness of our model. We conclude our paper in Section 5.

## 2. Self-Similarity and LRD

Informally, self-similarity refers to the degree of randomness where a non-self-similar sequence is totally random and a self-similar sequence exhibits a degree of non-randomness. This degree of non-randomness can be utilized towards better system prediction and management. Mathematically, a stationary sequence $x = \{x(i), i \geq 1\}$ is called exactly/asymptotically self-similar sequence if

$$x \overset{d}{=} m^{1-H} x^{(m)}, 0.5 < H \leq 1. \qquad (1)$$

holds for all m/as $m \rightarrow \infty$ where

$$x^{(m)}(k) = \frac{1}{m}\sum_{i=(k-1)m+1}^{km}\left(x(i) - E[x]\right), k = 1, 2, \ldots \quad (2)$$

is the aggregated sequence of $x$ with level of aggregation $m$ and $H$ is a constant called the Hurst parameter. $x$ is also called exactly/asymptotically second-order self-similar sequence if the variance and the correlation function (ACF) of $x$ is equal to the variance and the ACF of $m^{1-H}x^{(m)}$ for all m/as $m \rightarrow \infty$ (Taqqu, Teverovsky et al. 1997). The Self-similarity manifests itself in several equivalent ways as shown in the subsequent sections.

LRD describes the long term correlation in a sequence of data where the ACF slowly decays with the lag towards zero and may not reach the zero (non-summable ACF). For Poisson processes, the ACF exponentially decays with the lag (short-Range Dependence (SRD)).

The Hurst effect describes the degree of perseverance of a second order statistic of the self-similar sequence across different time scales. This effect has different representations. For example, the variance of the aggregated self-similar sequence can be described as a function of the Hurst parameter as follows:

$$Var[x^m(k)] \sim cm^{2(H-1)}. \quad (3)$$

Another representation of the Hurst effect describes the relationship between the variance of the wavelet coefficients of the Haar Wavelet transform at a given scale $j$ and the Hurst parameter as follows:

$$Var\left[W_j\right] = c2^{j(2H-1)}. \quad (4)$$

where $W_j$ is the wavelet coefficients sequence at scale $j$, and $c$ is a constant value. The Hurst effect as described by equations (3) and (4) indicates whether the sequence exhibits SRD (H ≤ 0.5) or LRD (H > 0.5).

## 3. Synthetic Data Generator

Our sequence generator is based on the Haar Wavelet transform. The idea behind the wavelet transform is to express a discrete sequence by an approximated version (scale coefficients) and a detail (wavelet coefficients). The approximation process is repeated at various scales by expressing the approximated version of the sequence at scale $j$ by a coarser approximation and a detail at a scale $j+1$. Assuming that a scale $j = 0$ represents the original sequence and the sequence length is a multiple of 2, then the scale coefficients at a given scale are computed as a function of the scale and the Wavelet coefficients at the higher scale as follows:

$$U_{j,2k} = \frac{U_{j+1,k} + W_{j+1,k}}{\sqrt{2}}, \quad (5)$$

$$U_{j,2k+1} = \frac{U_{j+1,k} - W_{j+1,k}}{\sqrt{2}}, k = 0, 1, 2, \ldots \quad (6)$$

Where $U_{j,2k}$ and $U_{j,2k+1}$ are the two computed scale coefficients at scale $j$ from one scale coefficient ($U_{j+1,k}$) and one Wavelet coefficient ($W_{j+1,k}$) at scale $j+1$ as illustrated in figure 1.

The data generation process of our model is a function of three parameters, the target mean (μ), variance ($\sigma^2$), and the target Hurst parameter (H). To generate a sequence of length $N$, we need to have $log_2N$ scales. We start by a single value (at scale $j_h = log_2N - 1$), which is the approximation of the signal at the highest scale $j_h$.

From the definition of the Haar Wavelet transform, the expected value of the scale coefficient at a scale $j_h$ (the scale coefficient at the highest scale) is computed as follows (assuming that we deal with a sequence $x_i$, $i=0,1,\ldots$):

$$E\left[U_{jh}\right] = \frac{\sum_{i=0}^{N-1} x_i}{\frac{N}{2^{jh}}\left(\sqrt{2}\right)^{jh}}$$

$$= \frac{\frac{\sum_{i=0}^{N-1} x_i}{N}}{2^{-\frac{jh}{2}}}$$

$$= \mu 2^{jh/2} \quad (7)$$

Accordingly, this single scale coefficient at the highest scale jh is taken as the single value that we start with.

The Wavelet coefficients at a given scale can be modeled as normally distributed random numbers with mean 0 and variance equal to $Var(W_j)$ (defined in equation (4)). The constant value $c$ is computed as follows:

Since $E[W_j] = 0$, the second moment of the wavelet coefficient at scale $j$ is defined as:

$$E\left[W_j^{\,2}\right] = Var\left[W_j\right] = c2^{j(2H-1)} \quad (8)$$

And accordingly,

$$E\left[W_1^{\,2}\right] = c2^{(2H-1)} \quad (9)$$

And the $c$ value is computed as:

$$c = \frac{E[{W_1}^2]}{2^{(2H-1)}} \qquad (10)$$

From the Haar Wavelet transform definition, $E[{W_1}^2]$ is defined as:

$$E[{W_1}^2] = \frac{\sum_{i=0}^{\frac{N}{2}-1}(x_{2i} - x_{2i+1})^2}{\frac{N}{2}2}$$

$$= \frac{\sum_{i=0}^{\frac{N}{2}-1}({x_{2i}}^2 + {x_{2i+1}}^2) - 2\sum_{i=0}^{\frac{N}{2}-1}(x_{2i}\, x_{2i+1})}{N}$$

$$= \frac{\sum_{i=0}^{N-1} x_i^2}{N} - \frac{\sum_{i=0}^{\frac{N}{2}-1}(x_{2i}\, x_{2i+1})}{N/2}$$

$$\approx E[x^2] - E[x_i\, x_{i+1}]$$

$$= (\sigma^2 - \sigma^2\rho_1)$$

$$= \sigma^2(1-\rho_1) \qquad (11)$$

And accordingly,

$$c = \frac{\sigma^2(1-\rho_1)}{2^{(2H-1)}} \qquad (12)$$

Where $\rho_1$ represents the ACF of the sequence *x* at lag equal to 1. Although any model can be assumed for $\rho_1$, we use the model of exactly self-similar process for $\rho_1$, which was found to be equal to $2^{(2H-1)} - 1$ (Beran, Sherman et al. 1995). Accordingly, *Var[W<sub>j</sub>]* is computed as:

$$Var[W_j] = \frac{2\sigma^2(1-2^{(2H-2)})}{2^{(2H-1)}} 2^{j(2H-1)}$$

$$= \sigma^2(1-2^{(2H-2)})2^{j(2H-1)-2H+2} \qquad (13)$$

In summary, the parameters of the model are the target mean, the target variance, and the target Hurst parameter. Equation (7) is used to compute a single value, the scale coefficient at the highest scale ($j_h$). Equation (13) is used to compute the variance of the Wavelet coefficients at each scale starting from scale $j_h$ down to scale 1. Knowing the variance and the mean (zero mean) of the wavelet coefficients, the Wavelet coefficients can be generated as normally distributed random numbers and the synthesis process (of scale coefficients) can proceed using equations (5) and (6) (Figure 1).

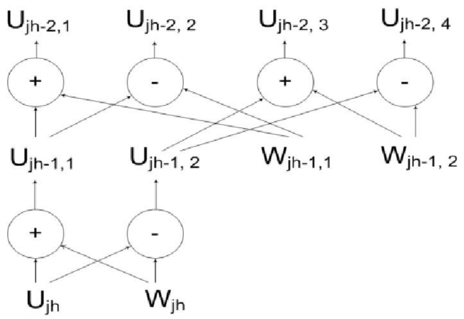

Figure 1. The generation process.

## 4. Experimental Results

To test our generator, we first use an LRD real trace that we try to mimic through its sample mean, variance, and the Hurst parameter. This is a public trace that contains of a million packet arrivals seen on an Ethernet network at the Bellcore Morristown Research and Engineering facility (Leland, Taqqu et al. 1994). The data set contains two columns, the first column is the time stamp in seconds of the packet arrival and the second column is the number of bytes in each packet. From this data set, we extracted a sequence of the number of bytes seen every 10 milliseconds. Table 1 shows some statistics computed over this sequence including the Hurst parameter. There are several methods to estimate the Hurst parameter (Karagiannis, Faloutsos et al. 2003). We use the Abry-Veitch method (Abry and Veitch 1998) .

Table 1: Basic statistics of the BELCORE trace

| Sample Mean | Sample St. Deviation | Estimated H |
|---|---|---|
| 1381.9 | 2227 | 0.81 |

Using our generator, we generated 50 data sets with *H* equals to 0.81 and with the same mean and variance as the Bellcore trace. Figure 2 shows the average ACF of the generated traces compared to the ACF of the Bellcore trace. They both exhibit similar LRD behavior.

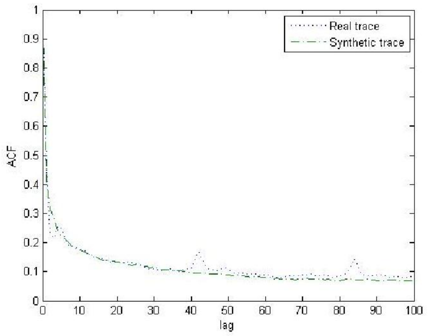

Figure 2: The ACF of the BELCORE trace compared to the ACF of the Synthetic traces

Now we test our generator further by generating several traces with Hurst parameters ranges from 0.6 to 0.9. For each value of the target Hurst parameter, we generated 50 traces and estimated the *H* value and took the average. Table 2 shows the estimations of the Hurst parameters of these traces. From Table 2, it is clear that our generator can accurately generate synthetic traces with any arbitrary Hurst parameter.

Table 2: The Hurst parameters estimation for the synthetic traces

| Target H | Sample mean | 95% Confidence Interval |
|---|---|---|
| 0.6 | 0.6002 | [0.5972, 0.6032] |
| 0.65 | 0.6455 | [0.6412, 0.6497] |
| 0.7 | 0.6995 | [0.6961, 0.7029] |
| 0.75 | 0.7493 | [0.7459, 0.7527] |
| 0.8 | 0.7994 | [0.7958, 0.8030] |
| 0.85 | 0.8495 | [0.8464, 0.8527] |
| 0.9 | 0.8955 | [0.8919, 0.8990] |

It worth mentioning that this generator produces positive and negative data and there are so many applications, especially in network traffic modeling, where we need positive data. A simple way is to replace all negative data by zeros. Table 3 shows the Application of this simple method to the above experiment. Although this simple method produce traces with slightly deviated Hurst parameters from the target ones, it still gives a very good accuracy.

Table 3: The Hurst parameters estimation for the synthetic positive traces

| Target H | Sample mean | 95% Confidence interval |
|---|---|---|
| 0.6 | 0.5943 | [0.5910, 0.5977] |
| 0.65 | 0.6412 | [0.6382, 0.6442] |
| 0.7 | 0.6916 | [0.6874, 0.6957] |
| 0.75 | 0.7389 | [0.7365, 0.7414] |
| 0.8 | 0.7899 | [0.8772, 0.7927] |
| 0.85 | 0.8336 | [0.8300, 0.8372] |
| 0.9 | 0.8864 | [0.8830, 0.8899] |

## 4. Conclusions

In this paper, we developed a simple self-similar data generator that is based on a few number of parameters, the mean and the variance of the modeled data, and the needed Hurst parameter H. We showed some numerical results that prove the goodness of our generator based on the measured Hurst parameter and correlation structure of the generated synthetic traces. The generated traces were normally distributed and for a future work, we need to find a way to generate data with different distributions.

9/12/2013

**Acknowledgements:**

This article was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah. The authors, therefore, acknowledge with thanks DSR technical and financial support.

**Corresponding Author:**
Dr. Abdullah Balamash
Department of Electrical Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia
E-mail: geetakh@gmail.com

## References

1. Abry, P. and D. Veitch (1998). "Wavelet analysis of long-range-dependent traffic." Information Theory, IEEE Transactions on **44**(1): 2-15.
2. Beran, J., R. Sherman, M. S. Taqqu and W. Willinger (1995). "Long-range dependence in variable-bit-rate video traffic." Communications, IEEE Transactions on **43**(234): 1566-1579.
3. Hosking, J. R. M. (1984). "Modeling persistence in hydrological time series using fractional differencing." Water resources research **20**(12): 1898-1908.
4. Karagiannis, T., M. Faloutsos and M. Molle (2003). "A user-friendly self-similarity analysis tool." ACM SIGCOMM Computer Communication Review **33**(3): 81-93.
5. Karagiannis, T., M. Molle and M. Faloutsos (2004). "Long-range dependence ten years of Internet traffic modeling." Internet Computing, IEEE **8**(5): 57-64.
6. Lau, W. C., A. Erramilli, J. L. Wang and W. Willinger (1995). Self-similar traffic generation: The random midpoint displacement algorithm and its properties, IEEE.
7. Leland, W. E., M. S. Taqqu, W. Willinger and D. V. Wilson (1994). "On the self-similar nature of Ethernet traffic (extended version)." Networking, IEEE/ACM Transactions on **2**(1): 1-15.
8. Mandelbrot, B. (1967). "How long is the coast of Britain? Statistical self-similarity and fractional dimension." Science **156**(3775): 636-638.
9. Mandelbrot, B. B. (1971). "A fast fractional Gaussian noise generator." Water Resources Research **7**(3): 543-553.
10. Mandelbrot, B. B. and J. W. Van Ness (1968). "Fractional Brownian motions, fractional noises and applications." SIAM review **10**(4): 422-437.
11. Paxson, V. and S. Floyd (1995). "Wide area traffic: the failure of Poisson modeling." IEEE/ACM Transactions on Networking (ToN) **3**(3): 226-244.
12. Riedi, R. H., M. S. Crouse, V. J. Ribeiro and R. G. Baraniuk (1999). "A multifractal wavelet model with application to network traffic." Information Theory, IEEE Transactions on **45**(3): 992-1018.
13. Shugong, X., H. Xinwei and H. Zailu (1998). "Experimental Queuing Analysis withLong-Range Dependent Traffic [J]." Acta Electronica Sinica **4**.
14. Taqqu, M. S., V. Teverovsky and W. Willinger (1997). "Is network traffic self-similar or multifractal?" Fractals-an Interdisciplinary Journal on the Complex Geometry **5**(1): 63-74.
15. Wornell, G. W. (1993). "Wavelet-based representations for the 1/f family of fractal processes." Proceedings of the IEEE **81**(10): 1428-1450.