Evaluation of Hurst exponent for precipitation time series

ALINA BĂRBULESCU
Faculty of Mathematics and Computers Science
Ovidius University of Constantza
124, Mamaia Boulevard, Constantza
ROMANIA
alinadumitriu@yahoo.com http://www.alina.ilinc.ro

CRISTINA SERBAN (GHERGHINA)
Faculty of Civil Engineereing
Ovidius University of Constantza
124, Mamaia Boulevard, Constantza
ROMANIA
cgherghina@gmail.com

CARMEN MAFTEI
Faculty of Civil Engineereing
Ovidius University of Constantza
124, Mamaia Boulevard, Constantza
ROMANIA
cmaftei@univ-ovidius.com

Abstract: - A major issue in time series analysis and particularly in the study of meteorological time series behaviour is the long range dependence (LRD). Various estimators of LRD have been proposed. Their accuracy have been generally tested by using simulated time series since sometimes only their asymptotic property are known, or worse, no asymptotic property have been proved. It is well – known that the Hurst exponent (*H*) is a statistical measure used to classify time series. In this article we determine the Hurst exponent for precipitation time series collected in Dobrudja region, for 41 years and we compare the results.

Key-Words: - Long – range dependence, Hurst coefficient, precipitation, detrended fluctuation

1 Introduction

The study of time series with long-range dependence have been extensively developed for applications in nature sciences, as well as in DNA sequences, cardiac dynamics, internet traffic [1] and finance [2].

The Hurst [3] exponent provides a measure for long term memory and fractality [4] of a time series. For details on the topics, see the volumes of Beran [5], Embrechts and Maejima [6], and Palma [7] and the collections of Doukhan et al. [8] and Robinson [9].

The values of the Hurst exponent range between 0 and 1. Based on the Hurst exponent value *H*, the following classifications of time series can be realized:

- $H = 0.5$ indicates a random series;
- $0 < H < 0.5$ indicates an anti - persistent series, which means an up value is more likely followed by a down value, and vice versa;
- $0.5 < H < 1$ indicates a persistent series, which means the direction of the next value is more likely the same as current value.

2 Measuring Hurst exponent

Let $(X_t)_{t\in\mathrm{N}}$ be a time series, shortly denoted by (X_t).

We say that (X_t) it is weakly stationary if it has a finite mean and the covariance depends only on the lag between two points in the series.

Let $\rho(k)$ be the autocorrelation function (ACF) of (X_t). If the time series is weakly stationary, ACF is given by:

$$\rho(k) = \frac{E[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma^2},$$

where $E(X_t)$ is the expectation of X_t, μ is the mean and σ^2 the variance.

It is said that the time series (X_t) has the long range dependence (LRD) property if $\sum_{k=-\infty}^{\infty}\rho(k)$ diverges.

LRD can be thought of in two ways: [10]

- In the time domain it manifests as a high degree of correlation between distantly separated data points.
- In the frequency domain it manifests as a significant level of power at frequencies near zero.

The following methods are used for LRD analysis:

- R/S and Lo's modified R/S statistic;
- Aggregated Variance;
- Absolute Moments;
- Detrended fluctuation analysis (variance of residuals);
- Ratio of variance of residuals;
- Periodogram;
- Whittle's approximate MLE and Local Whittle estimators;
- Wavelets.

We shall discuss here the results obtained using some of the methods described in the following.

2.1 R/S analysis

The Hurst exponent can be calculated by rescaled range analysis (R/S analysis) [3].

To study the LRD in a time series, the following algorithm is used.

A time series $(X_k)_{k\in\overline{1,N}}$ is divided into d sub-series of length m. For each sub-series $n = 1, \dots, d$:

- Find the mean, E_n and the standard deviation, S_n ;

- Normalize the data (X_{in}) by subtracting the sub-series mean:

$$Z_{in} = X_{in} - E_n,\ i = 1,\dots,m;$$

- Create a cumulative time series:

$$Y_{in} = \sum_{j=1}^{i} Z_{jn},\ i = 1,\dots,m;$$

- Find the range

$$R_n = \max_{j=1,m} Y_{jn} - \min_{j=1,m} Y_{jn};$$

- Rescale the range R_n / S_n;

- Calculate the mean value of the rescaled range for all sub-series of length m:

$$(R/S)_m = \frac{1}{d}\sum_{n=1}^{d} R_n / S_n .$$

Hurst found that (R/S) scales by power - law as time increases, which indicates

$$(R/S)_t = c \cdot t^H .$$

In practice, in classical *R/S* analysis, H can be estimated as the slope of log/log plot of $(R/S)_t$ versus t.

Although Mandelbrot [11] gave a formal justification for the use of this test, Lo [12] showed that this statistic was not robust to short memory dependence and modified this statistic.

Lo defined *modified R/S* statistic by:

- instead of considering multiple lags, only focus on lag N, the length of the series:

After finding the overall mean

$$\overline{X} = \sum_{j=1}^{N} X_j ,$$

create the cumulative time series:

$$Y_k = \sum_{j=1}^{N}(X_k - \overline{X}),\ k = 1,\dots,N;$$

and find the range

$$R(N) = \max_{k=1,N} Y_k - \min_{j=1,N} Y_k ;$$

- instead of using the standard deviation to normalize $R(N)$, he uses the following sum:

$$S_q(N) = \sqrt{\sum_{j=1}^{N}(X_j - \overline{X})^2 + 2\sum_{j=1}^{q}\omega_j(q)\left[\sum_{i=j+1}^{N}(X_j - \overline{X})(X_{i-j} - \overline{X})\right]}$$

where

$$\omega_i(q) = 1 - \frac{j}{q+1}, q < N.$$

The Lo's modified *R/S* statistic is defined by:

$$V_q(N) = \frac{1}{\sqrt{N}} R(N) / S_q(N) .$$

Lo uses the interval [0.809, 1.862] as the 95% asymptotic acceptance region for testing the null hypothesis:

H_0 : the absence of LRD,

against the alternative:

H_1 : the presence of LRD.

As discussed in [13], the right choice of q in Lo's method is essential. For study, the following values are used:

$$q = \left[\left(\frac{3N}{2}\right)^{\frac{1}{3}}\left(\frac{2\hat{\rho}}{1-\hat{\rho}^2}\right)^{\frac{2}{3}}\right], [14] \qquad (*)$$

and

$$q = \left[\left(\frac{N}{10}\right)^{\frac{1}{4}}\left(\frac{2\hat{\rho}}{1-\hat{\rho}^2}\right)^{\frac{2}{3}}\right], [15] \qquad (**)$$

where:
- N is the length of the series,
- $\hat{\rho}$ is the estimated first order correlation coefficient,
- [ ] is the greatest integer function.

2.2. Aggregated variance method [13]

A series of length N is divided into d sub-series of length m. For each subseries, the aggregated series, formed by the means

$$X^m(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i, k = 1,2,...,d$$

is calculated, as well as, its sample variance:

$$VarX^{(m)} = \frac{1}{d} \sum_{k=1}^{d} (X^{(m)}(k) - \overline{X})^2 .$$

For successive values of m, the sample variance is plotted against m on a log-log plot. Fitting a least squares line to the points of the plot, the Hurst coefficient is calculated, knowing that the straight line slope is $2H$-2.

In order to distinguish between the nonstationarity and LRD, Teverovski and Taqqu [10] proposed to use this method, together with the study of successive differences of variances or fitting a function $C_1 + C_2 m^{2H-2}$ to the $VarX^{(m)}$.

2.3. Absolute Moments Method [13]

This method is analogous to 2.3, but instead of $VarX^{(m)}$, the n^{th} absolute moments are calculated for the aggregated series,

$$AM_n^{(m)} = \frac{1}{d} \sum_{k=1}^{d} \left| X_k^{(m)} - \overline{X} \right|^n .$$

For successive values of m, the sample absolute moment is plotted versus m on a log-log plot. Fitting a least squares line to the points of the plot, the Hurst coefficient is calculated, knowing that the straight line slope is $n(H$-1).

2.4. Detrended fluctuation analysis (DFA)

Detrended fluctuation analysis was originally proposed as a technique for quantifying the nature of long-range correlations by Peng *et al.* [16]. It was introduced in order to permit the detection and quantification of long-range correlations in DNA sequences.

In recent years the DFA method has been applied in analysis of different time series that appear in different fields as DNA sequences [17], heart rate study [18], human gait, meteorology, economics, and physics.

DFA method, involves the following steps [19]:

- Starting with a time series, (X_t) with the length N, it is integrated, obtaining:

$$Y_k = \sum_{i=1}^{k} (X_k - \overline{X}) .$$

- The integrated series in divided into d sub-series of equal length m.
- In each sub-series, fit X_t, using a polynomial function of order l which represents the *trend* of that sub-series. The ordinate of the fit line in each box is denoted by $y_m(k)$.
- The integrated series is detrended by subtracting the local trend $y_m(k)$ in each sub-series of length m.
- For a given sub-series length, m, the root mean-square fluctuation for the integrated and detrended series is calculated:

$$F(m) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} [Y_k - y_m(k)]^2} .$$

- The above computation is repeated for a broad range of scales (m) to provide a relationship between $F(m)$ and the box size m.

A power-law relation between the average root-mean-square fluctuation function $F(m)$ and the box size m indicates the presence of scaling: $F(m) \sim m^{2H}$.

Ratio of Variance of Residuals

This method estimates the parameter alpha characterizing the intensity of heavy tails, instead of estimating the long-range dependence parameter H. The method is based on the Variance of Residuals method. It calculates the Variance of Residuals in two ways, and takes their ratio to obtain a statistic, and then fits a least-squares line to the logarithm of that statistic. This enables one to estimate alpha.[13]

2.5. Periodogram

Geweke and Porter-Hudak [20] proposed a semi-parametric approach to test for long-memory memory of a fractionally integrated process. The fractional difference parameter d can be estimated by regression equations.

Let $(X_k)_{k \in \overline{1,N}}$ be a time series, and

$$I(\lambda) = \frac{1}{2\pi N} \left| \sum_{j=1}^{N} X_j e^{ij\lambda} \right|^2 ,$$

where λ is a frequency.

It was shown that a series with LRD should have a periodogram proportional to $|\lambda|^{1-2H}$ in the origin neighbourhood, so a log-log plot of a periodogram

against the frequency should give the coefficient $1-2H=d$.

Modified periodogram, as well *cumulative periodogram* have also been used [13].

3. Results and discussions

In the following we present the results of the calculus of Hurst coefficient, for ten annual precipitation time series (Fig.1), collected between 1965 and 2005, in Dobrudja, Romania and we discuss the results.

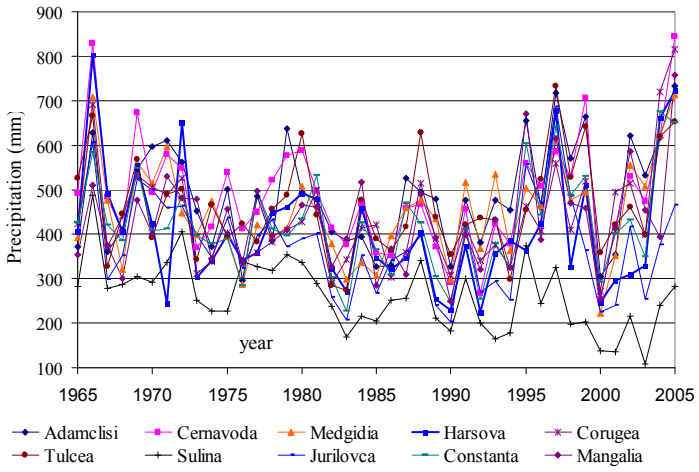

Fig.1. The series of annual mean precipitation in Dobrudja region (1965 - 2010)

For R/S analysis, a part of results is given in [21]. and those obtained by using KaotiXL [22], in Table 1.

The R/S charts and the evolution of V - statistics associated are exemplified in Figs. 2 – 7, where R^2 is the determination coefficient.

Table 1. Hurst coefficients for annual series, calculated by KaotiXL (rescaled method)

Station	Adamclisi	Cernavoda	Constanta	Corugea	Harsova
H	0.7207	0.9775	0.9738	0.9224	0.8028
Station	Jurilovca	Medgidia	Mangalia	Sulina	Tulcea
H	0.9301	0.8727	0.7049	0.9567	0.7057

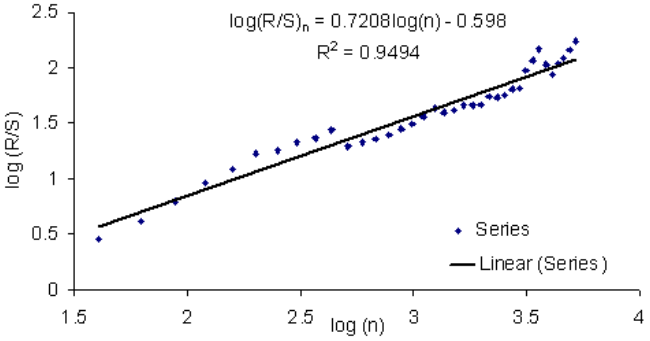

Fig. 2. The R/S chart of Adamclisi series

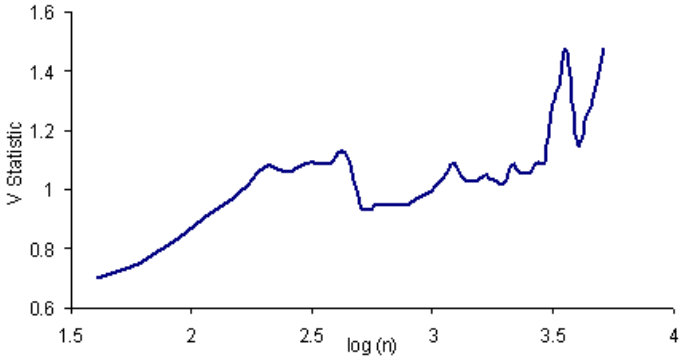

Fig. 3. The V - statistic associated to Adamclisi series

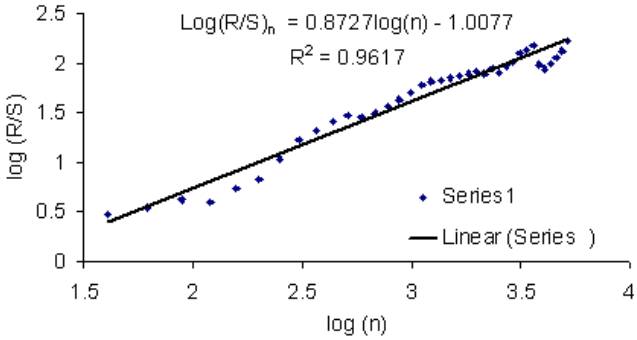

Fig. 4. The R/S chart of Medgidia series

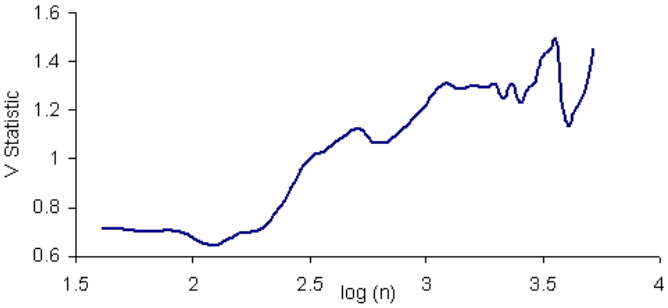

Fig. 5. The V - statistic associated to Medgidia series

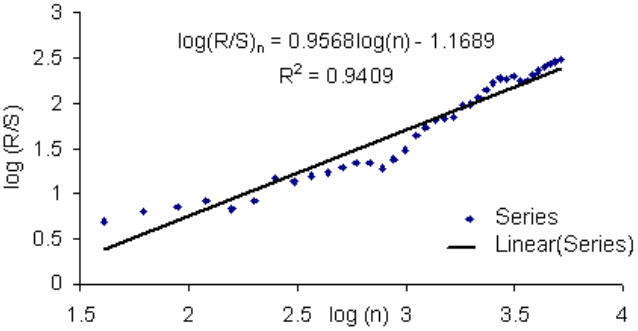

Fig. 6. The R/S chart of Sulina series

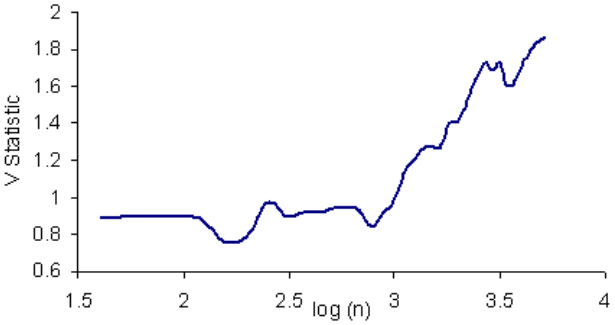

Fig. 7. The V - statistic associated to Sulina series

For Lo's modified statistic, the values used for q were calculated by (*) and (**). As result of formula (*), q was respectively: 3 – for Sulina and Jurilovca series and 2 – for the rest. Using formula (**), the coefficients were not very different.

The results are presented in Table 2 and those obtained by periodogram method, in Table 3.

Table 2. Hurst coefficients for annual series, calculated by Lo's method

Station	Adamclisi	Cernavoda	Constanta	Corugea	Harsova
H	0.0988	0.1177	0.11	0.1262	0.1167
Station	Jurilovca	Medgidia	Mangalia	Sulina	Tulcea
H	0.1104	0.1304	0.1165	0.119	0.1121

Table 3. Hurst coefficients for annual series, calculated by periodogram method

Station	Adamclisi	Cernavoda	Constanta	Corugea	Harsova
H	0.44505	0.4491	0.36745	0.2826	0.55745
Station	Jurilovca	Medgidia	Mangalia	Sulina	Tulcea
H	0.46905	0.3967	0.4559	0.32415	0.58925

It can be remarked that the results for each series are very different and a correct conclusion on the long range dependence property of these time series could not be extract applying only one method.

Discussing, for example, about Adamclisi, Cernavodă or Medgidia series, after the application of normality, homoscedasticity and correlation tests, it was concluded that they are Gaussian white noises [21]. But not all the values from the previous tables are in concordance with these conclusions. In [21] a FARIMA model was found for Sulina series. But the values of H, in Tables 2 and 3 are discordance to the LRD property of Sulina series.

The reasons for which these results are so different could be: q is too small in the case of Lo's statistic (and it does not account for the autocorrelation of the process) and the small number of data.

The results of DFA are closer to those of Lo's method (for example, for Adamclisi series), other being closer to those of periodogram method.

4. Conclusion

In this article we presented the results of LRD analysis for mean annual precipitation time series. They were compared to the results of statistic tests. We remark that they are not concordant. As consequence, it is indicated to use with circumspection different LRD analysis methods, since sometimes the statistics attached to them have properties that have been proved only for some well established time series or their properties are not known.

References:

[1] W. Willinger et all, Self-similarity in high-speed packet traffic: analysis and modeling of Ethernet traffic measurements, *Statistical Science,* 10, 1995, pp.67-85

[2] A. Lo, Fat tails, long memory, and the stock market since the 1960s, Economic Notes, Banca Monte dei Paschi di Siena, 2001.

[3] H. E. Hurst, Long-term storage of reservoirs: an experimental study, *Transactions of the American society of civil engineers,* 116, 1951, pp. 770 – 799

[4] B.B. Mandelbrot, J. R. Wallis, Robustness of the rescaled range R/S in the measurement of noncyclic long-run statistical dependence, *Water Resources*, 5, 1969, pp. 967 – 988.

[5] J. Beran, *Statistics for Long - Memory Processes*, Chapman and Hall, New York, 1994

[6] M. Embrechts, P. Maejima, *Self - similar processes*. Princeton, University Press, Princeton and Oxford, 2002

[7] W. Palma, *Long-Memory Time Series Theory and Methods*, Wiley - Interscience, 2007

[8] P. Doukhan, G. Oppenheim, and M. Taqqu, *Theory and Applications of Long-Range Dependence*, Birkhauser, 2003

[9] P. M. Robinson, *Time Series with Long Memory*, Oxford University Press, 2003.

[10] V. Teverovsky, M. S. Taqqu, W. Willinger, On Lo's modified R/S statistic, *Preprint*, 1997

[11] B. Mandelbrot, Statistical methodology for non - periodic cycles: from the covariance to R/S analysis, *Annals of Economic and Social Measurement*, vol.1, 1972

[12] A. W. Lo, Long term memory in stock market prices, *Econometrica*, vol. 59, 1991, pp.1279 - 1313

[13] 13 M. S. Taqqu, V. Teverovsky, W. Willinger, Estimators for long range dependence: an empirical study, *Fractals,* vol.3, no.4, 1995, pp.785 – 788

[14] D. W. K. Andrews, Heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica*, 59, 1991, pp. 817–858

[15] W. Wang et all, Detecting long-memory: Monte Carlo simulations and application to daily streamflow processes, www.hydrol-earth-syst-sci-discuss.net/3/1603/2006/

[16] C.- K. Peng et all, Mosaic organisation of DNA nucleotides, *Physical Review E*, vol. 49, no.2, 1994, pp. 1685 – 1689

[17] S. V. Buldyrev et all. Long-Range Correlation Properties of Coding and Noncoding DNA Sequences: GenBank Analysis, *Physical Reviews E*, vol. 51, 1995, pp. 5084 – 5091

[18] Y. Ashkenazy et all., Magnitude and Sign Correlations in Heartbeat Fluctuations, *Physical Review Letters*, vol. 86, 2001, pp. 1900-1903

[19] http://www.physionet.org/physiobank/database/synthetic/tns/paper2/node2.html
[20] J. Geweke, S. Porter-Hudak, The estimation and application of long memory time series models, Journal of Time Series Analysis, 4, 1983, pp. 221–238
[21] A. Bărbulescu, E. Pelican, ARIMA models for the analysis of the precipitation evolution, *Recent Advances in Computers*, *Proceedings of the 13th WEAS International Conference on Computers*, 2009, pp. 221 – 226
[22] http://www.soft32.com/Download/free-trial/KaotiXL/ 4-191370-1.html

Acknowledgements: This article was supported by CNCSIS – UEFISCSU, project number PNII – IDEI 262/2007.