Available online at www.sciencedirect.com
ScienceDirect
Physica A 370 (2006) 808–816
PHYSICA A
www.elsevier.com/locate/physa

Language time series analysis

Kosmas Kosmidis*, Alkiviadis Kalampokis, Panos Argyrakis

Department of Physics, University of Thessaloniki, 54124 Thessaloniki, Greece

Received 17 January 2006
Available online 24 March 2006

Abstract

We use the detrended fluctuation analysis (DFA) and the Grassberger–Proccacia analysis (GP) methods in order to study language characteristics. Despite that we construct our signals using only word lengths or word frequencies, excluding in this way huge amount of information from language, the application of GP analysis indicates that linguistic signals may be considered as the manifestation of a complex system of high dimensionality, different from random signals or systems of low dimensionality such as the Earth climate. The DFA method is additionally able to distinguish a natural language signal from a computer code signal. This last result may be useful in the field of cryptography.

Keywords: Language simulations; Detrended fluctuation analysis; Grassberger–Proccacia analysis

1. Introduction

Human language has recently attracted the attention of the physical scientists. Following the advances in the theory and understanding of complex systems, it was recently realized that human language is a new emerging field for the application of methods from the physical sciences in order to achieve a deeper understanding of linguistic complexity. Important work in the field of the mathematical modelling of language and in the field of language simulations has recently been done by several groups [1–13]. There is also renewed interest in the task of discovering and explaining structural properties of the languages, such as the Zipf law [14–16] which actually deals with the probability distribution of words in spoken languages. Of course, the understanding of the complexity associated with language is not an easy task. We have to use all kinds of mathematical tools in order to gain understanding of the system we study. One of these tools is time series analysis [17]. Time series analysis plays a key role in physical sciences. Our goal is to extract information from signals that are related to real-world phenomena. Analyzing such signal allows us to achieve better understanding of the underlying physical phenomena. The methods of analyzing signals are wide spread and range from classical Fourier analysis to various types of linear time–frequency transforms, model-based and non-linear approaches.

A particularly interesting characteristic of time series associated with several physical processes is the presence of long-range correlations. Some interesting examples include DNA sequences [18–22], weather

*Corresponding author.
E-mail address: panos@physics.auth.gr (K. Kosmidis).

0378-4371/$ - see front matter
doi:10.1016/j.physa.2006.02.042

records [23] and heart rate sequences [24–30]. The common feature of all these diverse systems is that the long-range correlations decay by a power law, where a characteristic scale is absent. These findings are useful, e.g., in DNA for distinguishing between coding and noncoding sequences [22], in atmospheric science for testing state-of-the-art climate models, etc. Long-range correlations may be detected using a method called detrended fluctuations analysis (DFA) [31,32], which we will presented in Section 3.

Moreover, when we study time series and without relying on any particular model we are interested in getting an insight of the dynamics of the system solely from the knowledge of the time series. In such cases a method derived by Grassberger and Proccacia [33–35] has been proven particularly useful. This method has been applied to analyze the dynamics of climatic evolution [35 and references therein], neural network activity [36], or electric activity of semiconducting circuits [37,38]. Details on the method are presented in Section 4.

2. Mapping documents to time series

The main problem one has to deal with before applying the analysis methods is the following: Given a document written in natural language, how can one transform it in a time series and then analyze it? Although, at first, time series and natural language documents seem to be irrelevant, we will present two ways to construct time series from documents:

i. Take a document of N words. Count the length l (number of letters) of each word. The role of time is played by the position of the word in the document i.e. the first word is considered to be emitted at time $t = 1$, the second at time $t = 2$ etc. We map the word length to this time and thus a time series $l(t)$ is constructed. Henceforth, we will refer to such time series as "length time series" (LTS).
ii. Take a document of N words. Count the frequency f of appearance of each word in the document. Again the role of time is played by the position of the word in the document. We map the word frequency to this time and thus a time series $f(t)$ is constructed. Henceforth, we will refer to such time series as frequency time series (FTS).

Obviously there is a large number of ways to map a document to a time series, but in the present study we deal with the above two as there is also a physical meaning in the mapping. The length of the word is associated with speaker effort, meaning that the longer the word the higher the effort required to pronounce it. The frequency of the word is also associated with the hearer effort as frequently used words require less effort to be understood by the hearer.

Linguistic series have been studied in the past [39–41] but dealt human writing at letter level and not at word level. Human writings at word level were studied by Montemurro and Pury [42]. They use a "frequency mapping" similar—but not identical—to the one described above but they use a different method for their analysis. So their results have to be considered as complementary to ours.

3. Detrended fluctuations analysis

The DFA estimates a scaling exponent from the behaviour of the average fluctuation of a random variable around its local trend. The method can be summarized as follows. For a time series u_t, $t = 1,2,\ldots, N$, first the integrated time series Y is obtained:

$$Y(i) = \sum_{t=1}^{i} [u_t - <u>], \tag{1}$$

where $<u>$ is the sample mean.

In the second step, we divide $Y(i)$ into $N_s \equiv [N/s]$ non-overlapping segments of equal length s. Since the record length N need not be a multiple of the considered time scale s, a short part at the end of the profile will remain in most cases. In order not to disregard this part of the record, the same procedure is repeated starting from the other end of the record. Thus, $2N_s$ segments are altogether obtained.

In the third step, we calculate the local trend for each segment v by a least-square fit of the data. We denote the detrended time series for segment duration s by $Y_s(i)$. It is the difference between the original time series and the fits:

$$Y_s(i) = Y(i) - p_v(i), \tag{2}$$

where $p_v(i)$ is the fitting polynomial in the vth segment. If quadratic polynomials are used in the fitting procedure, the method is called quadratic DFA (DFA2). Linear, cubic, or higher-order polynomials can also be used (DFA1, DFA3, and higher-order DFA).

In the fourth step, we calculate the variance for each of the $2N_s$ segments

$$F_s^2(v) = < Y_s^2(i) > = \frac{1}{s}\sum_{i=1}^{s} Y_s^2[(v-1)s + i] \tag{3}$$

for the detrended time series $Y_s(i)$ by averaging over all data points i in the vth segment. Finally, we average over all segments and take the square root to obtain the DFA Fluctuation Function:

$$F(s) = \left[\frac{1}{2N_s}\sum_{v=1}^{2N_s} F_s^2(v)\right]^{1/2}. \tag{4}$$

If only short-range correlations (or no correlations) exist in the time series, then the Fluctuation function will have the statistical properties of a random walk. Thus, we expect $F(s) \sim s^{\alpha}$ with $\alpha = 1/2$, while in the presence of long-range correlations $\alpha \neq 1/2$.

4. Grassberger–Proccacia analysis

At first sight, a time series of a single variable appears to provide a limited amount of information. We usually think that such a series is restricted to a one-dimensional view of a system, which, in reality, contains a large number of independent variables. It has been shown [33–35], however, that a time series bears the marks of all other variables participating in the dynamics of the system and thus we are able to "reconstruct" the systems phase space from such a series of one-dimensional observations. When applying the Grassberger–Proccacia (GP) method to a time series we want to find the answer to the following questions:

i. Can the salient features of the system be viewed as the manifestation of a deterministic dynamics, or do they contain an irreducible stochastic element? Is it possible to identify an attractor in the system phase space from a given time series?
ii. If the attractor exists, what is its dimensionality d?
iii. What is the minimal dimensionality, n, of the phase space within which the above attractor is embedded? This defines the minimum number of variables that must be considered in the description of the underlying system.

This is done as follows: We consider a time series. Let us call this signal $x_0(t)$. We would like to reconstruct the dynamics of the system solely on our knowledge of $x_0(t)$. We consider the phase space spanned by the variables $k = 0, 1, 2, \ldots, n-1$, where k are several variables that take part in the dynamics of the system. For our problem these are the parameters related with language and we do not know which or how many they are. At a given time a state of the system is a point in phase space, while a sequence of states in time gives a trajectory. If the dynamics of the system obey some dissipative deterministic laws, then the trajectories converge to an attractor. We thus form this attractor from the $x_0(t)$ series, by successively shifting the original time series by a constant amount in time $\Delta t = \tau$, and forming n such series as

$$\begin{aligned}
&x_0 : x_0(t_1), x_0(t_2), \ldots, x_0(t_N),\\
&x_1 : x_0(t_1 + \tau), x_0(t_2 + \tau), \ldots, x_0(t_N + \tau),\\
&x_2 : x_0(t_1 + 2\tau), x_0(t_2 + 2\tau), \ldots, x_0(t_N + 2\tau),\\
&\ldots\\
&x_{n-1} : x_0(t_1 + (n-1)\tau), x_0(t_2 + (n-1)\tau), \ldots, x_0(t_N + (n-1)\tau).
\end{aligned} \tag{5}$$

These variables are expected to be linearly independent if the τ shift is properly chosen. We chose several different τ values, but in the subsequent calculations we use $\tau = 500$ time units. Notice that $x_0(t)$ is a vector made of the set of points, as given in Eq. (5). A general notation for it is x_i. We now choose a reference point in x_i and compute all the distances $|x_i - x_j|$ from the $(N-1)$ remaining points. This way we get the total of all points x_i in phase space. Doing this for all i we get

$$C(l) = \frac{1}{N^2}\Theta \sum_{i,j=1,j\neq i}^{N} (l - |x_i - x_j|), \tag{6}$$

where Θ is the Heavyside step function, $\Theta(x) = 0$, if $x<0$ and $\Theta(x) = 1$ if $x>0$. $C(l)$ is the correlation integral of the attractor, since it shows how a point in the vector x_i affects the positions of other points. Thus if the attractor is a d-dimensional manifold, then we expect $C \sim l^d$, with its dimensionality given by the exponent d.

5. Results and discussion

Both English and Greek texts were used for this research. Specifically the English texts were "the War of the Worlds" by H. G. Wells, "The Mysterious Affair at Styles" by Agatha Christie and "A Christmas Carol" by Charles Dickens. For the Greek language the translation of "Sangharakshita, Vision and Transformation" was used. All of the above texts were found using Project Gutenberg (www.gutenberg.net). The Greek corpus was complimented with extracts from publications of the Greek newspaper "Ta Nea" for the years 1997–2003.

In Fig. 1 we present an LTS signal (a) and an FTS signal (b). This series were constructed from a document written in Greek, taken from the Greek newspaper "Ta Nea". A total of 36 221 words were used for the analysis, but only a small part (100 words) of the series is shown for clarity.

In Fig. 2a we present the results of third-order DFA. In this method we use a 3rd degree polynomial in order to perform the detrending. The resulting function F_d follows a power law but exhibits a cross-over at about

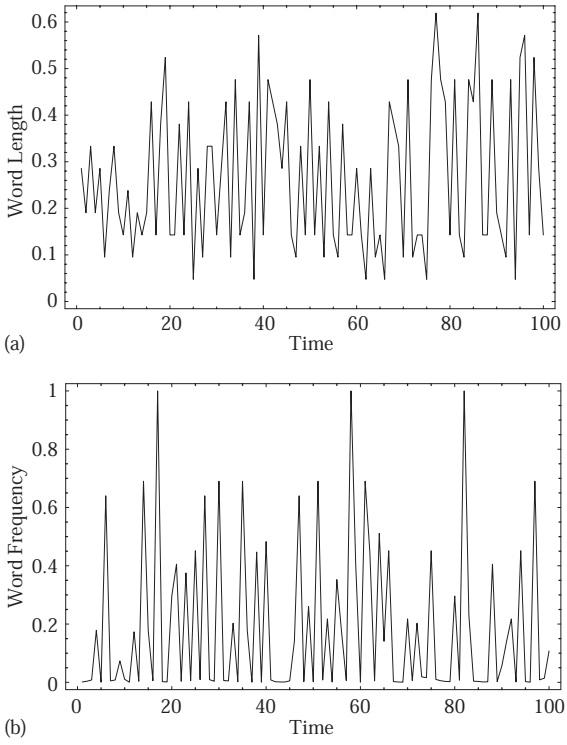

Fig. 1. (a) Word length versus time, and (b) word frequency versus time. This series was constructed from a document written in Greek. Only a part of the series is shown (100 words). The total number of words used for the analysis is 36 221 words.

$s\sim200$. The initial part has an exponent $n = 0.46$ while the remaining part has an exponent of 0.61. In Fig. 2b we present the same analysis for the shuffled series. Shuffling should destroy long-range correlations leading to an exponent $n = 0.50$. In our data the exponent is found to be 0.52 for the shuffled data, as expected. In our example the linguistic signal appears to be slightly anticorrelated in short time scales. The slope however is rather close to 0.5 which indicates absence of correlations. Thus, we have to be rather careful with the above conclusion.

Anticorrelated behaviour practically means that a word of short length has a higher probability to be followed by a long word and vice versa. In the Greek language short words (usually articles) are often followed by longer words (nouns, adjectives, etc.) so this anticorrelated behaviour is somehow anticipated. This is an indication that the above method may detect the presence of syntactic structure in a linguistic document without knowing any more details of the language.

Moreover, the existence of a crossover from an early almost uncorrelated behavior to a long-range correlated one, is characteristic of a signal composed of different patches [25]. Moreover, the location of the crossover is approximately the same as the size of the patches. Here, the crossover is found at $s\sim200$. This is roughly the size of a paragraph. This indicates that the DFA method is capable of detecting the paragraph structure of the document. See the marked difference in the results derived from computer code.

In Fig. 3a, we present the same type of analysis for an English language document namely Christmas Carols by Dickens. English grammar is somewhat different from that of the Greek. Syntax is not so restricting and articles are not used as often as in Greek documents. Here the resulting slope of the straight line in Fig. 3a is 0.48 much closer to that of an uncorrelated time series. In Fig. 3b we perform the same analysis for the frequency signal. The calculated slope is equal to 0.49, almost identical to that of an uncorrelated signal.

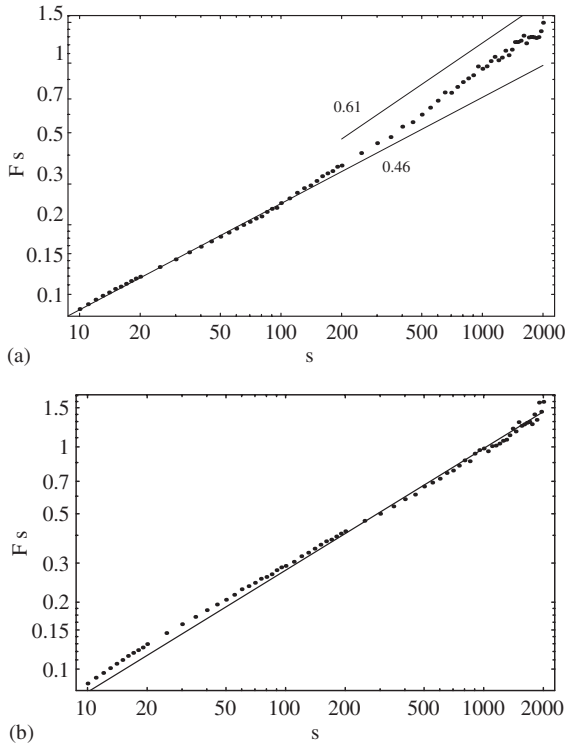

Fig. 2. (a) Plot of the DFA3 Fluctuation function $F(s)$ vs. s of the Length Time Series presented in Fig. 1. A power-law behavior is observed and the slope of the straight line is equal to 0.45. The slope change for values of s greater than 200 is probably due to (periodic) trends not eliminated from the DFA3 procedure. (b) Same plot for the shuffled Length Time Series presented in Fig. 1. A power-law behavior is observed and the slope of the straight line is equal to 0.52.

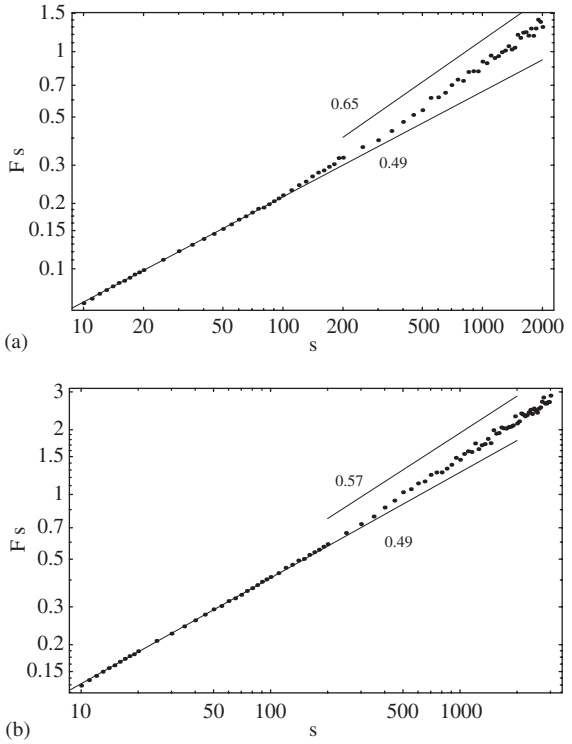

Fig. 3. (a) Plot of the DFA3 Fluctuation function $F(s)$ vs. s of the Length Time Series for an English document (A Christmas Carol by Dickens). The length of the series is 28 713 words. Again, a power-law behavior is observed and again the slope is equal to 0.48 very close to that of a signal with no or short range correlations. (b) Same analysis for the Frequency Time Series for the same English document. The initial slope here is 0.49, almost identical to that of an uncorrelated signal.

The situation is however different if instead of a spoken language we study a computer language. A spoken language is a means of communication between human beings while a computer language is a means of communication between humans and computer machines. We have constructed an LTS signal using the Linux Kernel which is written in C. It must be noted that in computer programs (as the Linux Kernel) both programming language instructions and variable names are used. The instructions are few in number and are language specific, whereas the variable names are chosen by the programmer. In the present work no distinction was made between the two, as the whole program was treated as a means of communication between the programmer and the computer.

In Fig. 4a we present the signal and in Fig. 4b we present the results of DFA3 analysis. The slope of the straight line is 0.64 indicating the presence of long-range correlations. Performing the same analysis on the FTS signal of the Linux Kernel we obtain a straight line with slope 0.55 (data not shown).

Therefore, we conclude that there is a marked difference between human and computer languages. Human language LTS and FTS signals are probably non-correlated or slightly anticorraleted whilst computer language LTS and FTS signals exhibit long-range correlations.

In Fig. 5, we present the results of our GP analysis for LTS signals. Fig. 5a shows the Correlation integral $C(r)$ as a function of r for several embedding dimensions n of the phase space for the LTS signal of a Greek language document. Notice the straight line segments of the double logarithmic plot. We use these parts in order to estimate the exponent d and determine the scaling behaviour of the correlation integral. Then we plot in Fig. 5b the obtained d values as a function of the embedding dimension n. For a white noise signal we expect to see a straight line with slope equal to one, while for a low-dimensional chaotic system we expect a saturation to some value of n which will allow us to determine the minimal dimensionality of the phase space needed to embed the attractor and consequently to determine how many independent variables we have to consider in

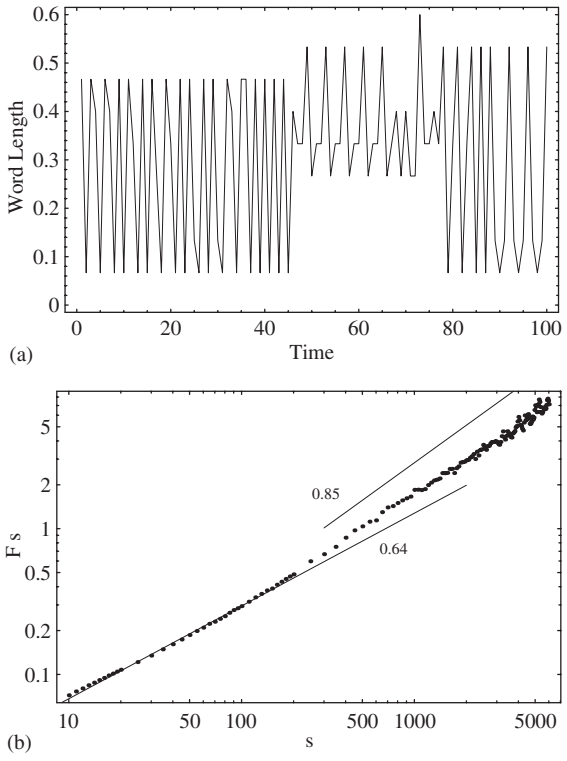

Fig. 4. (a) A Length Time Series for the Linux Kernel, and (b) DFA3 for the Linux Kernel. The signal seems to exhibit long-range correlations as the slope of the straight line is equal to 0.64.

describing a linguistic system. What we see is that there is no saturation and also that the slope of the line is 0.69, i.e., rather less than one. Is this an indication that we are dealing with a complex system of high dimensionality? If so, the dimension of the phase space of this system is high, possibly infinite. Thus, there is a marked difference between for example climatic records and LTF signals. DFA analysis of climatic records has revealed, that they are strongly long-range correlated [43] while a GP analysis has shown that they may be seen as the manifestation of a low-dimensional chaotic system (the embedding dimension was found to be no more than 5)[35]. For Language LTS signals we see that signals are not so strongly long-range correlated, but the dynamics of the system involves a large amount of parameters! Note, however, that when a complex system has a high-dimensional phase space it does not mean that it is mathematically intractable. For example, a delay differential equation is in fact an infinite dimensional system [34]. We can, nevertheless, study it numerically and obtain useful results.

In order to investigate the relation between the DFA analysis and the GP method, we have performed a GP analysis on the language LTS after shuffling the word order. Shuffling will destroy the correlation between words but will maintain the probability distribution of the word lengths. The resulting slope for the shuffled series (Fig. 5b, triangles) is practically identical to that of the original LTS. This is probably due to the fact that GP method is giving information about the geometric structure of the phase space attractor, but it is not sensitive to the particular sequence with which the points on this attractor are visited!

As a finishing remark, we would like to point to a possible application of the DFA method. It seems that the method may be used to distinguish between natural language and computer code. Suppose that we have a encrypted sequence of words. We may construct an LTS signal from this sequence and then apply the DFA method in order to decide whether this sequence is random, natural language or computer code. From our analysis seems that the DFA method may be useful in some aspects at the field of cryptography.

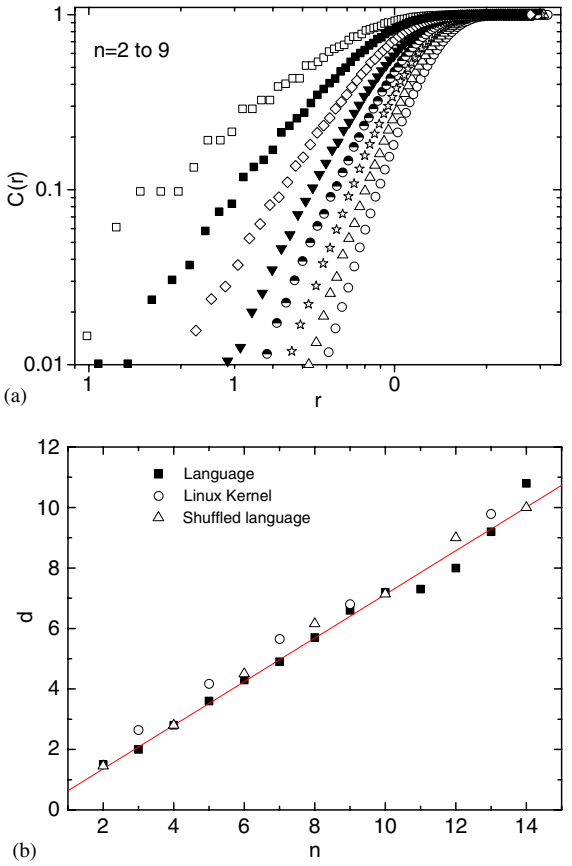

Fig. 5. (a) Double logarithmic plot of the correlation integral $C(r)$ for several embedding dimensionalities n. The exponents d are calculated from the slope of the straight line segments. (b) Plot of exponent d versus the dimensionalities n. *Rectangles*: Greek language document. *Circles*: Linux Kernel. The straight line has a slope equal to 0.69 and shows no saturation. *Triangles*: Shuffled Greek language document.

Acknowledgements

The authors want to thank Prof. Armin Bunde and Prof. Shlomo Havlin for useful discussions. This work was supported by the Greek Ministry of Education through the PYTHAGORAS project.

References

[1] M. Nowak, D. Krakauer, Proc. Natl. Acad. Sci. USA 96 (1999) 8028.
[2] D. Abrams, S. Strogatz, Nature 424 (2003) 900.
[3] C. Schulze, D. Stauffer, Int. J. Mod. Phys. C 16 (2005) 718 and AIP Conference proceedings 119, 49 (2005) (8th Granada Seminar).
[4] C. Schulze, D. Stauffer, Phys. Life Rev. 2 (2005) 89.
[5] K. Kosmidis, J.M. Halley, P. Argyrakis, Physica A 353 (2005) 595.
[6] K. Kosmidis, A. Kalampokis, P. Argyrakis, Physica A, in press (physics/051019).
[7] J. Mira, A. Paredes, Europhys. Lett. 69 (2005) 1031.
[8] V. Schwammle, Int. J. Mod. Phys. C 16 (10) (2005) 1519 physics/ 0503238.
[9] V. Schwammle, Int. J. Mod. Phys. C 17 (3) 2006, physics/ 0509018.
[10] T. Tesileanu, H. Meyer-Ortmanns, Int. J. Mod. Phys. C 17 (3) (2006), physics/0508229.
[11] M. Patriarca, T. Leppapen, Physica A 338 (2004) 296.
[12] V.M. de Oliveira, et al., Physica A 361 (2006) 361.
[13] V.M. de Oliveira et al., Physica A, in press (physics/0510249).

[14] G.K. Zipf, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley, Cambridge, MA, 1949.
[15] S. Havlin, Physica A 216 (1995) 148.
[16] R. Cancho, R. Sole, Proc. Natl. Acad. Sci. USA 100 (2003) 788.
[17] D.S.G. Pollock, Time series analysis, Signal Processes and Applications, Academic Press, London, 1999.
[18] C.-K. Peng, et al., Nature (London) 356 (1992) 168.
[19] R.F. Voss, Phys. Rev. Lett. 68 (1992) 3805.
[20] S.V. Buldyrev, et al., Phys. Rev. Lett. 71 (1993) 1776.
[21] A. Arneodo, E. Bacry, P.V. Graves, J.F. Muzy, Phys. Rev. Lett. 74 (1995) 3293.
[22] R.N. Mantegna, et al., Phys. Rev. Lett 73 (1994) 3169.
[23] S.V. Buldyrev, et al., in: A. Bunde, S. Havlin (Eds.), Fractals in Science, Springer, Berlin, 1994.
[24] E. Koscielny-Bunde, et al., Phys. Rev. Lett. 81 (1998) 729.
[25] C.-K. Peng, et al., Phys. Rev. Lett. 70 (1993) 1343.
[26] C.-K. Peng, et al., Chaos 5 (1995) 82.
[27] C.-K. Peng, et al., Physica (Amsterdam) 249A (1998) 491.
[28] S. Thurner, M.C. Feurstein, M.C. Teich, Phys. Rev.Lett. 80 (1998) 1544.
[29] L.A. Amaral, A.L. Goldberger, P. Ch. Ivanov, H.E. Stanley, Phys. Rev. Lett. 81 (1998) 2388.
[30] P.Ch. Ivanov, et al., Nature (London) 399 (1999) 461.
[31] C.K. Peng, et al., Phys. Rev. E 49 (2) (1994) 1685.
[32] J. Kantelhardt, E. Koscielny-Bunde, H. Rego, S. Havlin, A. Bunde, Physica A 295 (2001) 441.
[33] P. Grassberger, I. Procaccia, Phys. Rev. Lett. 50 (1983) 346.
[34] P. Grassberger, I. Procaccia, Physica D 9 (1983) 189.
[35] G. Nicolis, I. Prigogine, Exploring Complexity, Freeman, New York, 1989.
[36] C. Kotsavasiloglou, A. Kalampokis, P. Argyrakis, S. Baloyannis, Phys. Rev. E 56 (4) (1997) 4489.
[37] Ch. Karakotsou, A.N. Anagnostopoulos, Physica D 93 (1996) 157.
[38] Ch. L. Koliopanos, I.M. Kyprianidis, I.N. Stouboulos, A.N. Anagnostopoulos, L. Magafas, Chaos Soliton Fract 16 (2003) 173.
[39] A. Schenkel, J. Zhang, Y.-C. Zhang, Fractals 1 (1993) 47.
[40] R.F. Voss, Fractals 2 (1994) 1.
[41] M. Amit, Y. Shemerler, E. Eisenberg, M. Abraham, N. Shnerb, Fractals 2 (1) (1994) 7.
[42] M. Montemurro, P. Pury, Fractals 10 (4) (2002) 451.
[43] A. Bunde, J.F. Eichner, J.W. Kantelhardt, S. Havlin, Phys. Rev. Lett. 94 (2005) 048701.