Identification of Structure-Predictability Relations in Time Series with Pattern Recognition Techniques

E. Bautista-Thompson and J. Figueroa-Nazuno

Centro de Investigación en Computación
Instituto Politécnico Nacional
Unidad Profesional "Adolfo López Mateos"
Ciudad de México, D. F., C. P. 07738, México
ebautista@correo.cic.ipn.mx, jfn@cic.ipn.mx

Abstract. The predictability deals with the difficulty that can be assigned to a time series in order to be forecasted by a model. In this work, the identification of relations between predictability and time series structure is done by means of two pattern recognition techniques: Multidimensional Scaling and Recurrence Plots. The first technique allows the clustering of the time series by their predictability degree that is associated with a set of time series parameters, the second technique allows the visualization of the spatial and temporal dynamics hidden in the structure of the time series. The results shows that the predictability is related with the structural features of the time series through a set of basic structural patterns, these patterns show different kinds of associations with groups of time series that possess a similar predictability.

1 Introduction

A time series represents the behavior of an observable for a system from natural or artificial origin [1, 2], the dynamics represented by a time series have a richness of structural patterns that can be studied with different techniques such as Fourier Analysis, Principal Component Analysis, etc. [3]; the identification and analysis of these patterns are basic steps towards a better understanding of the time series dynamics. The time series dynamics is related with the predictability, this is a quantitative estimation of the difficulty that implies to model and forecast such dynamics [4, 5, 6, 7]. The identification of relations between predictability and structure of time series is the main goal of the present work, this is done with the help of two pattern recognition techniques: Multidimensional Scaling and Recurrence Plots. The Section 2, explains the concept of predictability used in this work and the way it is estimated. The Section 3, describes the two pattern recognition techniques and their advantages. The Section 4, shows the experimental results and their interpretation. Finally, the Section 5 presents the conclusions of this work.

2 Predictability of Time Series

The predictability deals with the difficulty that can be assigned to a time series in order to be forecasted by a model. Traditionally, this feature has been associated with the forecast error, for example the root mean square error (RMSE) or others definitions reported in the literature [8, 9], these kind of measurements assume that the predictability is due to the forecasting model and not to the time series characteristics, others works consider that the predictability can be estimated with only one feature of time series such as Lyapunov exponent or some definitions of information entropy (an illustrative example is showed in Section 4), based on the hypothesis that with only one parameter there is enough information in order to characterize the predictability [3, 6, 10].

In this work, the predictability was studied with a set of parameters that characterize the time series from different points of view: Non Linear Dynamics, Statistics, Fourier Analysis, Information Theory and Computation Theory. In this way, a holistic estimation of the predictability is achieved. The estimation of the predictability is not explicit in nature, in the sense of a scalar value such as a forecast error can provide, instead the Multidimensional Scaling was used in order to identify groups of time series with similar predictability in an implicit way. The Table 1 shows a brief description of the fifteen parameters selected to estimate the predictability. The parameters are classified by: its theoretical basis, the generic type of the feature that is computed, the reach of the parameter, and the type of information that it provides about the time series [11, 12, 13, 14, 15, 16, 17, 18].

3 Pattern Recognition Techniques

In this section the pattern recognition techniques used in this work are described and its advantages high lined. The first technique is Multidimensional Scaling, it allows the grouping of objects with similar features, in this case similar predictability represented by the set of time series parameters, the second technique is Recurrence Plots, it allows the visualization of patterns that represent the temporal and spatial correlation between the points that form a time series, enabling a better visualization of the time series dynamics. Associating the information provided by these two techniques, knowledge about the relations between the patterns of time series dynamics and its predictability was extracted.

3.1 Multidimensional Scaling (MDS)

This is a method that represents similarity metrics between pairs of objects as distances between points in a multidimensional space into one space of lower dimension (2-D or 3-D). The graphical representation allows the observation and

Table 1. Description of parameters related with time series predictability.

Parameter	Theoretical Basis	Feature	Reach	Type of Information
Pearson Correlation	Statistics	Statistical	Global	Correlation degree between time series points
Hurst Exponent	Statistics	Statistical	Global	Trend
Dominant Frequency	Fourier Analysis	Temporal	Global	Signal frequency patterns
Lyapunov Exponent	Non Linear Dynamics	Topological	Global	Forecast horizon
Correlation Dimension	Non Linear Dynamics	Topological	Local	Local spatial correlation
Capacity Dimension	Non Linear Dynamics	Spatial	Local	Self similarity degree
Fractal Dimension	Non Linear Dynamics	Spatial	Local	Local average dimension in a sphere of radius epsilon
Embedded Dimension	Non Linear Dynamics	Topological	Global	Degrees of freedom
Spatial Temporal Entropy	Non Linear Dynamics	Spatial Temporal	Global Local	Degree of non spatial and temporal correlation
Recurrence	Non Linear Dynamics	Spatial Temporal	Global Local	Periodicity and structure
Determinism	Non Linear Dynamics	Spatial Temporal	Global Local	Degree of determinism
Shannon Entropy	Information Theory	Probabilistic	Global Local	Information extracted from a measurement into a system
Average Mutual Information	Information Theory	Probabilistic	Global	Information contained in one variable for two instants of time
Lempel-Ziv Complexity	Computation Theory	Computational	Global Local	Structure an hierarchy in data strings
Production Rules	Computation Theory	Computational	Global Local	Computational complexity

exploration by the expert of the data structure in search of hidden patterns and also the discovery of dimensions that represent parameters of similarity [19].

The MDS takes as input information a proximity matrix of the form,

$$\Delta = \begin{bmatrix} \delta_{11} & \delta_{12} & .... & \delta_{1n} \\ \delta_{21} & \delta_{22} & .... & \delta_{2n} \\ . & . & .... & . \\ \delta_{n1} & \delta_{n2} & .... & \delta_{nn} \end{bmatrix}. \quad (1)$$

Where n is the number of objects to be compared. Each element δ_{ij} is a distance measurement (usually an Euclidean metric but not restricted to it) that represents the proximity between the features of the objects i and j. The goal is to adjust an initial random distance matrix $X \in M_{nxm}$ where n is the number of objects (in this particular case the number of time series) and m is the number of dimensions,

$$X = \begin{bmatrix} x_{11} & x_{12} & .... & x_{1m} \\ x_{21} & x_{22} & .... & x_{2m} \\ . & . & .... & . \\ x_{n1} & x_{n2} & & x_{nm} \end{bmatrix}. \quad (2)$$

Each value x_{ij} represents the coordinate of the *i-th* time series. Now, the distance between two time series i and j can be calculated, and a distance matrix $D \in M_{nxx}$ is obtained,

$$D = \begin{bmatrix} d_{11} & d_{12} & .... & d_{1n} \\ d_{21} & d_{22} & .... & d_{2n} \\ . & . & .... & . \\ d_{n1} & d_{n2} & .... & d_{nn} \end{bmatrix}. \quad (3)$$

The solution must satisfy that there is a maximum correspondence between the proximity matrix Δ and the distance matrix D. This is achieved by adjusting iteratively the matrix X in the next way,

$$X = \frac{BX}{2n}. \quad (4)$$

Where B has as elements,

$$b_{ij} = \frac{-2\delta_{ij}}{d_{ij}} \text{ if } i = j. \quad (5)$$

$$b_{ii} = \sum\sum \frac{2\delta_{ik}}{d_{ik}} \text{ if } i = j. \tag{6}$$

$$b_{ij} = 0 \text{ if } d_{ij} = 0 \tag{7}$$

The optimal adjustment is achieved when a precision function called S-Stress reach a precision value determined a priori by the expert [19],

$$S - Stress = \sqrt{\frac{\sum_{i,j}\left(f(\delta_{ij}) - d_{ij}^2\right)^2}{\sum_{i,j}\left(d_{ij}^2\right)^2}} \tag{8}$$

In particular, MDS was applied to a set of time series (the objects) represented by fifteen parameters (dimensions) that characterize each time series, these parameters as was mentioned earlier, are related with the predictability of time series.

3.2 Recurrence Plots

The Recurrence Plots were first described by J. P. Eckman, S. O. Kamphorst and D. Ruelle in 1987 [20]. This qualitative (by means of the visualization of the plots) and quantitative analysis (by means of some parameters derived from the plots), allows the detection of hidden patterns and structural changes inside the time series data [16, 20, 21]. The basic idea that supports a Recurrence Plot, is that a time series is the product of a dynamical process where the relevant variables interact, and that it is possible to recover the information of such multivariate process from only one time series [1, 2]. A Recurrence Plot represents an expansion of a time series in a multidimensional space, in this space the dynamics of the time series is visualized through the multiplication of the available information. In order to build this plot, a reconstruction of the phase space is necessary, this is done with the embedding of the time series, that consists of the identification of a dimension m that resembles closely the original phase space dimension of the process that the time series represents, and the building of a set of vectors that corresponds with the set of original points that formed the trajectory of states for the original process in the phase space. The vectors have the form,

$$y(1) = \{x(i), x(i-d), x(i-2d), ..., x(i-(m-1)d\}. \tag{9}$$

Where i corresponds with the time index, m is the embedding dimension and d is a time delay. As a result, a time series formed by a set of vectors is generated,

$$Y = \{y(1), y(2), y(3), ..., y(N-(m-1)d)\}. \tag{10}$$

Where N is the size of the original time series.

Once the dynamics is reconstructed, a Recurrence Plot allows to show what vectors in the phase space are closer or more separated between them. This is established with an Euclidean distance between all the pairs of vectors, and it is codified with a color scale (e.g. a gray scale). Essentially, the Recurrence Plot is a color coding matrix, where each input (i, j) corresponds with a distance between the vectors $y(i)$ and $y(j)$. This distance is associated with a predefined color code that is displayed in the position of temporal character (i, j), for example, a light gray color corresponds with a small distance between vectors, and a black color corresponds with a big distance between vectors. A Recurrence Plot can be interpreted as a graphical representation of a correlation integral. The advantage compared with such correlation is that a Recurrence Plot preserves the temporal and spatial dependency between the points. The interpretation of the information displayed by a plot, has a qualitative nature: structured patterns are related with recurrent dynamics inside the time series and more determinism, non structured patterns corresponds with a changing dynamics (non necessarily random in nature) and less determinism, also combinations of these patterns can exist.

4 Identification of the Structural Patterns and their Relations

The advantage of exploit a set of parameters and not just only one in the study of the predictability is illustrated with the Figure 1, here the behavior of the Lyapunov exponent for the thirty time series used in this work is showed. This exponent estimates the rate of propagation of the forecast error and the related forecast horizon (one of the predictability definitions) [3, 6]. The time series are representative of different dynamics: periodic, quasi periodic, chaotic, complex and stochastic [3, 12, 15, 22]. There is not a clear relation between the Lyapunov exponent and the time series dynamics, then using only one parameter does not provide enough information in order to extract knowledge about the predictability and its behavior.

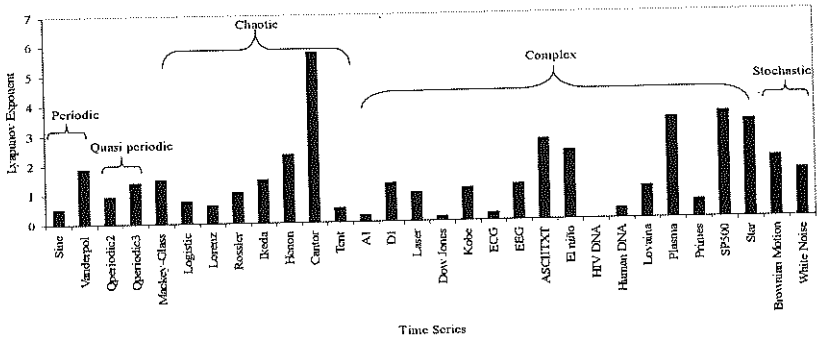

Fig. 1. Behavior of the Lyapunov exponent for the experimental set of time series

The Figure 2 shows the five clusters identified with the technique of Multidimensional Scaling, the clusters were formed by the similarity between the

time series due to the set of parameters that characterize their dynamics. The Recurrence Plots of the time series were associated to the Multidimensional Scaling Plot, the Recurrence Plots were generated with a codification of gray colors, more analysis follows below, but in Figure 2 a general view of the distribution of patterns is visualized. Some time series are not associated with a particular cluster, these time series have basic patterns too, but they were not study in the present work.

In order to identify the basic structural patterns that are present in the different clusters of time series, a comparison between the Recurrence Plots for the same cluster was first done, once a basic pattern was identified, a second comparison was done with the Recurrence Plots of the rest of time series that do not belong to the cluster of origin. In order to easily visualize the patterns a preprocessing was made in the Recurrence Plots images, each one was converted as an 8-bit image and then a binary threshold operation was applied, in this way the main structural features of each pattern were high lined [23].

The Figure 3 shows an example of the comparison between the Recurrence Plots of two time series that belong to the cluster 3, the similarity between the patterns of their Recurrence Plots is represented by the basic pattern in form of a dotted "L" inside the circles.

The Figures 4 to 8, show the basic structural patterns identified by the combination of the MDS analysis and the Recurrence Plots.

The time series in cluster 1, have a mix of patterns P1a and P1b at different levels in their corresponding Recurrence Plots. The patterns that are show in Figure 4 corresponds to the ASCIITXT time series for the pattern P1a and the Cantor time series for the pattern P1b.

In the cluster 3, the time series: A1, Lovaina and Laser have a mix of the P3a and P3b patterns in their Recurrence Plots, and the time series Lorenz and D1 have only the P3b pattern. The patterns in the Figure 6 corresponds to the Lovaina time series for the pattern P3a and Lorenz time series for the pattern P3b.

In the cluster 4, the time series: Qperiodic 2 and Rossler, have the pattern P4a; and the time series: ECG and Human DNA, have the pattern P4b in their Recurrence Plots. The patterns in the Figure 7 corresponds to the Rossler time series for the pattern P4a and Human DNA time series for the pattern P4b.

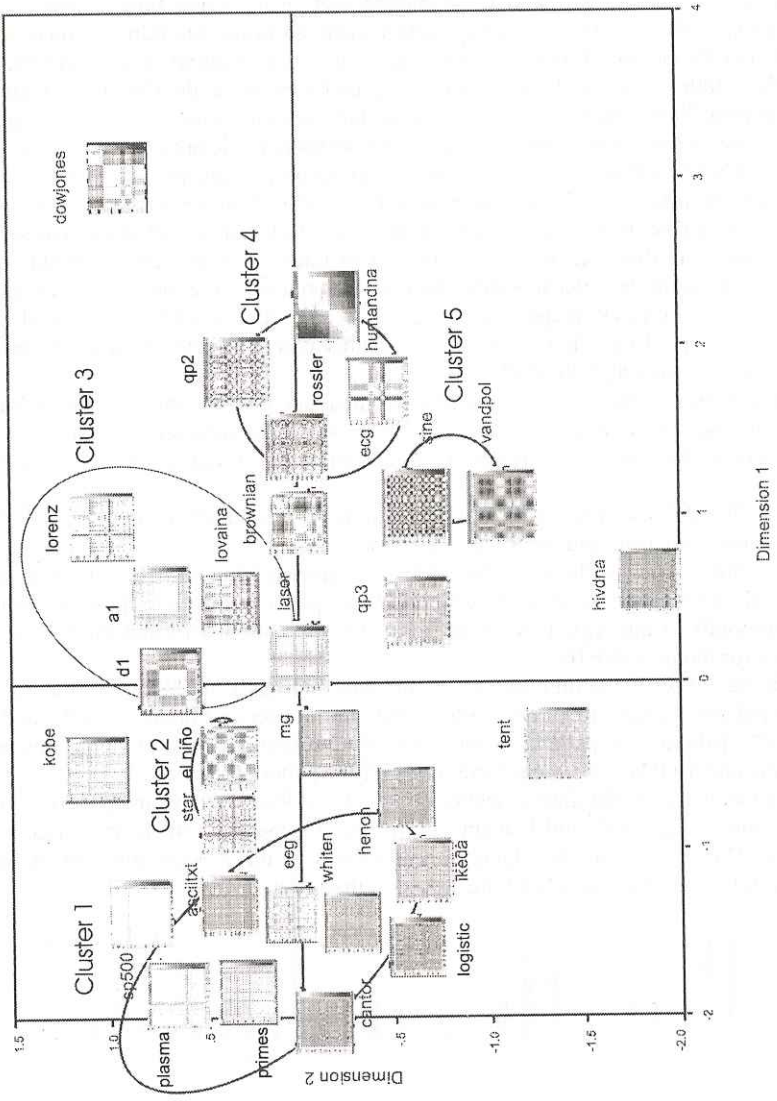

Fig. 2. Multidimensional Scaling combined with the Recurrence Plots of the time series. Five similarity clusters were identified

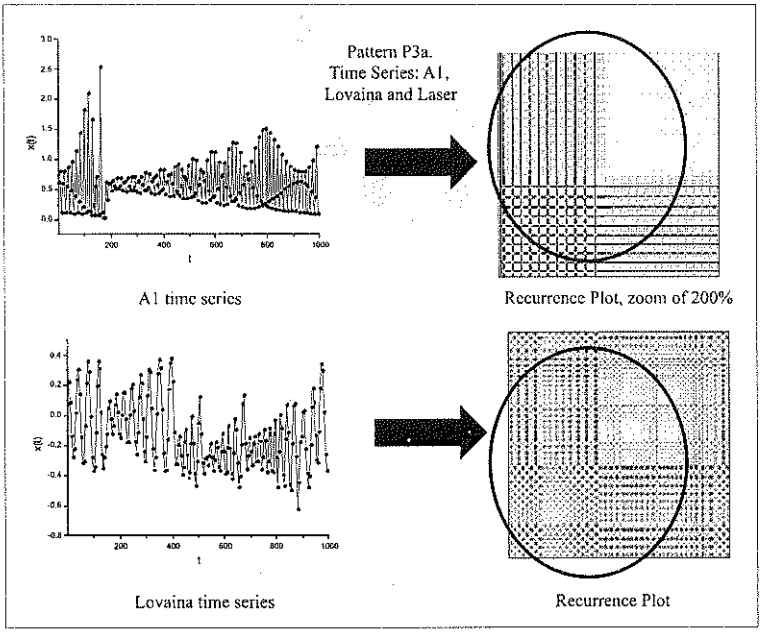

Fig. 3. Example of the comparison between two time series that belong to cluster 3 with their Recurrence Plots

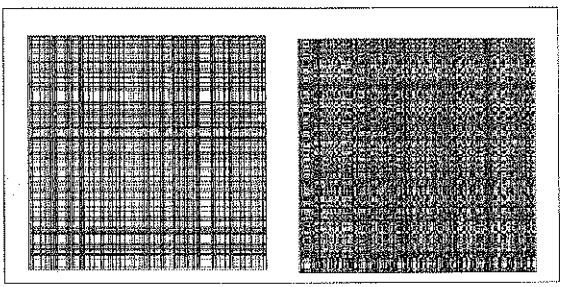

Fig. 4. Pattern P1a to the left side and pattern P1b to the right side, the cluster 1 formed by the time series: Plasma, SP500, ASCIITXT, Primes, EEG, White Noise, Cantor, Logistic, Henon and Ikeda have these patterns

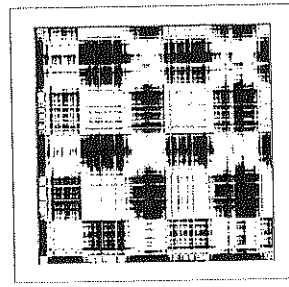

Fig. 5. Pattern P2, the cluster 2 formed by the time series: Star and El niño, have this pattern in their Recurrence Plots. The pattern corresponds to the El niño time series

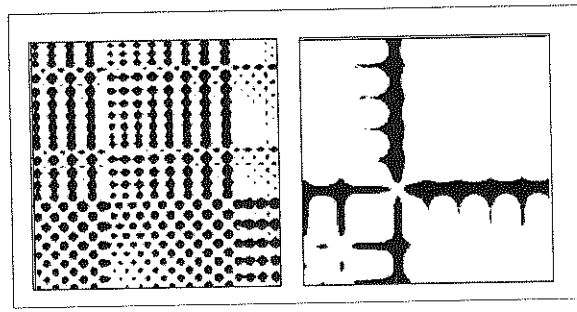

Fig. 6. Pattern P3a to the left side and pattern P3b to the right side

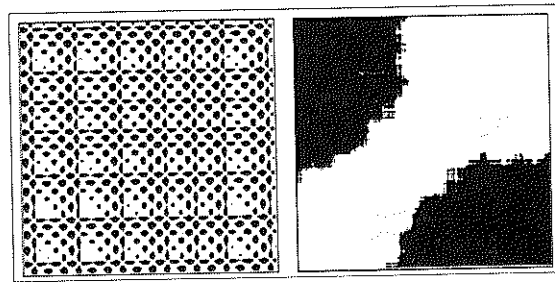

Fig. 7. Pattern P4a to the left side and pattern P4b to the right side

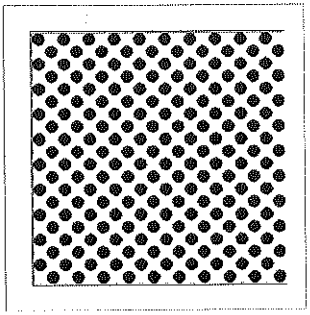

Fig. 8. Pattern P5, in the cluster 5 the time series: Sine and Vanderpol have this pattern in their Recurrence Plots. The pattern corresponds to the Sine time series

5 Conclusions

The application of Multidimensional Scaling in search of hidden patterns inside a set of time series, and Recurrence Plots in search of hidden patterns inside each time series allowed the identification of relations between the clustering of time series by their similarity with respect to the predictability, expressed through a set of parameters that characterize the time series, and the time series structure visualized with the Recurrence Plots. These relations of similar predictability are represented by basic structural patterns identified as: P1a, P1b, P2, P3a, P3b, P4a, P4b and P5; the relations between predictability and structure are resumed in the Table 2.

Table 2. Structure-Predictability Relations found in the similarity clusters.

Cluster of Time Series	Structure-Predictability Relation
1	Time series have a mix of two basic patterns
2 and 5	Time series have only one basic pattern
3	Time series have one or more basic patterns
4	Time series have different basic patterns

The relations show the existence of basic dynamics for the different phenomena represented by the time series, these dynamics are expressed as structural patterns. Also, the diversity of relations show that a richness of dynamic behaviors are related with the predictability of time series. In the future with an increment of the number and variety of time series, a better definition of the clusters will be achieved and new relations could be discovered.

References

1. Whitney, H.: Differentiable Manifolds. Annals of Mathematics 37 (1934) 645-680.
2. Takens, F.: Detecting Strange Attractors in Turbulence. Lecture Notes in Mathematics 898 (1981) 366-381.
3. Kantz, H., Schreiber, T.: Nonlinear Time Series Analysis. Cambridge University Press, Cambridge (2000).
4. Diebold, F. X., Kilian, L.: Measuring Predictability: Theory and Macroeconomic Applications. Working Papers Series Number 97-23 Federal Reserve Bank of Philadelphia, (1997).
5. Brooks, C.: Linear and Non-linear (Non-) Forecastability of High-frequency Exchange Rates. Journal of Forecasting 16 (1997) 125-145.
6. Smith, L. A.: Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems. In: Mees, A. (ed.): Nonlinear Dynamics and Statistics. Birkhäuser, Boston (2000) 31-64.
7. Avramov, D. : Stock-Return Predictability and Model Uncertainty. Rodney L. White Center for Financial Research Working Papers Number 12-00 The Wharton School University of Pennsylvania, (2000).
8. Granger, C. W. J., Newbold, P.: Forecasting Economic Time Series. Academic Press, Orlando (1986).
9. Kaboudan, M. A.: A Measure of Time Series' Predictability Using Genetic Programming Applied to Stock Returns. Journal of Forecasting 18 (1999) 345-357.
10. Palus, M.: Coarse-grained Entropy Rates for Characterization of Complex Time Series. Physica D 93 (1996) 64-77.
11. Kachigan, S. K.: Multivariate Statistical Analysis A Conceptual Introduction. Radius Press, New York (1991).
12. Pollock, D. S. G.: A Handbook of Time Series Analysis, Signal Processing, and Dynamics. Academic Press, San Diego (1999).
13. Peters, E. E.: Fractal Market Analysis: Applying Chaos Theory to Investment and Economics. John Wiley & Sons, New York (1994).
14. Urbach, R. M. A.: Footprints of Chaos in the Markets. Financial Times-Prentice Hall, London (2000).
15. Hilborn, R. C.: Chaos and Nonlinear Dynamics. Oxford University Press, Oxford (2000).
16. Trulla, L. L., Giuliani, A., Zbilut, J. P., Webber, C. L.: Recurrence Quantification Analysis of the Logistic Equation with Transients. Physics Letters A 223 (1996) 255-260.
17. Lempel, A., Ziv, J.: On the Complexity of Finite Sequences. IEEE Transactions on Information Theory 22 (1976) 75-81.
18. Badii, R., Politi, A.: Complexity: Hierarchical Structures and Scaling in Physics. Cambridge University Press, Cambridge (1999).
19. Borg, I., Groenen, P.: Modern Multidimensional Scaling. Springer-Verlag, New York (1997).
20. Eckmann, J. P., Kamphorst, S. O., Ruelle, D.: Recurrence Plots of Dynamical Systems. Europhys. Lett. 4 (1987) 973-977.
21. Gao, J., Cai, H.: On the Structures and Quantification of Recurrence Plots. Physics Letters A 270 (2000) 75-87.
22. Ott, E.: Chaos in Dynamical Systems. Cambridge University Press, Cambridge (2000).
23. Gonzalez, R. C., Woods, R. E.: Digital Image Processing. Prentice Hall, Upper Saddle River (2002).