Classification of Time Series Based on their Inner Structures

H. Solís-Estrella, E. Bautista-Thompson and J. Figueroa-Nazuno

Centro de Investigación en Computación, Instituto Politécnico Nacional
07738, Mexico D.F
{habacuc, ebautista}@correo.cic.ipn.mx, jfn@cic.ipn.mx

Abstract. There are techniques such as the Singular-Spectrum Analysis (SSA) which jointly with Principal Component Analysis (PCA), help us to analyze the underlying structures of a time series and condense them for their study. Using such information with the Multidimensional Scaling (MDS), we can find a representation that allows locating hidden regularities and thus classify the analyzed data. In this paper we present the study of time series of diverse nature using the three abovementioned techniques and show how they group in a bidimensional plane according to similar patterns within their components disregarding the dynamics of the series.

1 Introduction

Finding similarity among signals in a time series database has drawn plenty of attention in the data mining area in view of the fact that it is a very useful tool for analysis and knowledge extraction of phenomena [1]. A main aspect in similarity search is to define a set of relevant characteristics and find a metric or scheme to classify the data in accordance with defined criteria, i.e. the time series classification.

The time series classification problem is still open because no specific scheme suits all possible required criteria nor takes in account all the essential parameters, thus for particular needs new classification proposals must be developed. Traditionally, for the classification of time series, two approaches are taken, the first one is to classify according to scalar parameters of the data [2], the second is to classify according to the time series' morphology using a transformation [3] or lower dimensionality representation of the data [4,5]. Here we propose a scheme that attempts to uncover hidden relations among time series classifying them according to the underlying structures rather than the external waveform. The paper layout is as follows: in section 2 we review the analysis techniques, SSA and PCA, in section 3 the visualization and classification technique, MDS. Finally in section 4 we discuss the experimental results.

2 Analysis of the Time Series

2.1 Singular-Spectrum Analysis

Singular Spectrum Analysis (SSA) is a technique for time series analysis that incorporates elements of several disciplines such as multivariate statistics, multivariate geometry, dynamical systems and signal processing. It aims at a decomposition of the data into a sum of a small number of interpretable components such as trends, oscillatory patterns and random constituents [6]. The basic algorithm of the SSA technique consists of four steps. Given a time series $F=(f_0, f_1, \ldots, f_{N-1})$ of length N and L an integer called "window length"

1. Construct the trajectory matrix X of the time series as follows:
   set $K = L+N+1$ and define the L-lagged vectors $X_j = (f_{j-1}, \ldots, f_{j+L-2})^T, j = 1,2,\ldots,K$

$$\mathbf{X} = (x_{i+j-2})_{i,j=1}^{L,K} = [X_1,\ldots,X_K] \tag{1}$$

2. Obtain the Singular Value Decomposition (SVD) of the matrix X via eigenvalues and eigenvectors of the matrix $S = XX^T$. We thus obtain a representation of X as a sum of rank-one biorthogonal matrices X_i ($i = 1, \ldots, d$), where d is the number of nonzero singular values of X
3. Split the set of indices $I = \{1, \ldots, d\}$ into several groups $I_1,\ldots,I_m$ and sum the matrices Xi whitin each group. The result of the step is the representation

$$\mathbf{X} = \sum_{k=1}^{m} \mathbf{X}_{I_k} \quad \text{where} \quad \mathbf{X}_{I_k} = \sum_{i \in I_k} \mathbf{X}_i \tag{2}$$

4. Average over the diagonals $i+j$=const of the matrices $\mathbf{X}_{I_k}$. This gives us a decomposition of the original series F into a sum of series

$$f_n = \sum_{k=1}^{m} f_n^{(k)}, \; n = 0,\ldots, N-1 \tag{3}$$

where for each k the series $f_n^{(k)}$ is the result of diagonal averaging of the matrix $\mathbf{X}_{I_k}$.

2.2 Principal Component Analysis

Related to the time series decomposition with the SSA technique, is the Principal Component Analysis (PCA). Once the series has been separated in different component series, PCA is employed as a complementary tool. The PCA consists in the search of lineally independent components which provide the maximum variance of the data set and are not correlated [7]. PCA is generally used to condense the amount of data and to extract important features from it. Colebrook applied a form of SSA to biological oceanography and noted the duality with the principal component analysis [8].

The procedure to compute the principal components is as follows: let X be a nxp matrix whose rows represent cases and its columns variables, in addition, X must be

mean-centered. Let a be the yet to be determined projection weights column vector, which will result in the greatest variance when X is projected into a. The variance along a is defined as:

$$\sigma_a^2 = (Xa)^T(Xa) = a^T X^T Xa = a^T Va \tag{4}$$

where $V = X^T X$ is the covariance matrix. The projected data variance, σ_a^2 is expressed as a function of V and a.

To maximize σ_a^2 we need to apply a normalization restriction over a vectors which is $a^T a = 1$.

Consequently, the optimization problem can be rewritten as the maximization of the quantity:

$$u = a^T Va - \lambda(a^T a - 1) \tag{5}$$

where λ is a Lagrange multiplier. Differentiating with respect to a, we have:

$$\frac{\partial u}{\partial a} = 2Va - 2\lambda a = 0 \tag{6}$$

which is reduced to its eigenvalues form as:

$$(V - \lambda I)a = 0 \tag{7}$$

Therefore, the first principal component is the eigenvector associated with the largest eigenvalue in the covariance matrix V. The second principal component is the eigenvector associated with the second largest eigenvalue of V, and so on.

3 Multidimensional Scaling

Multidimensional Scaling (MDS) is a method that represents similarity metrics between pair of objects as distances between points in a lower dimensional space. This representation allows the expert observe and explore the data structure and find hidden regularities not easily distinguished in the raw numerical data [9].

MDS takes as input a proximity matrix $\Delta \in M_{n,n}$. Each element δ_{ij} of Δ represents the proximities between the parameters where $\delta_{ij} = (c_i - c_j)^2$ being c_k the aforementioned parameters.

The algorithm begins with a random matrix $X \in M_{n,m}$, where m is the desired number of dimensions and n is the number of variables. Each value x_{ij} represents the coordinates of the variable F_i in the j-th dimension.

Parting from such matrix, the distance among any two variables, i and j, can be calculated using the Minkowsky general distance equation:

$$d_{ij} = \left[\sum_{l=1}^{m} (x_{il} - x_{jl})^p \right]^{1/p} \tag{8}$$

Thus, a distance matrix $D \in M_{n,n}$ can be obtained from X.

The MDS solution must be such that there is a maximum correspondence between the initial proximities matrix Δ and the distance matrix D, which is accomplished iteratively modifying X according to:

$$X = \frac{BX}{2n} \tag{9}$$

where the elements bij of B are calculated using the following criteria:

$$b_{ij} = \frac{-2\delta_{ij}}{\delta_{ij}} \quad \text{if } i = j \tag{10}$$

$$b_{ii} = \sum\sum \frac{-2\delta_{ik}}{d_{ik}} \quad \text{if } i = j \tag{11}$$

$$b_{ij} = 0 \text{ if } d_{ij} = 0 \tag{12}$$

Then a monotonous increasing relation $\delta_{ij} < \delta_{kl} \Rightarrow d_{ij} < d_{kl}$ is assumed. To determine the precision, a function that relates distances with similarity $f(\delta_{ij})$ is constructed:

$$\begin{aligned} f &: \delta_{ij} \rightarrow d_{ij} \\ f(\delta_{ij}) &= a + b\delta_{ij} \end{aligned} \tag{13}$$

where *a* and *b* are constants to establish.

Therefore, we can apply the S-Stress precision measure, defined as:

$$S-Streess = \sqrt{\frac{\sum_{i,j}\left(f(\delta_{ij})^2 - d_{ij}^2\right)}{\sum_{i,j}\left(d_{ij}^2\right)}} \tag{14}$$

To complete MDS process, if a desired precision is not met, X is modified according to (15) and start over until the required S-Stress value is achieved.

4 Experimental Results

For the results we present, the experimental data set consists of 30 time series of diverse dynamics as shown in Table 1, the reference nominal classification is according to Sprott [10] and Figueroa [11]. All the time series have a length of 1000 data samples.

The methodology is as follows: the series were decomposed using SSA and then de data was used to obtain the principal components using PCA. The matrix of the first five principal components was fed to the MDS algorithm to obtain the time series

representation in a bidimensional plane according to their structure. Only five components were used since nearly 90% of the variance is contained in those components as we can observe in Table 2.

In Figure 1 we show the final clustering of the series and the three main groups that were identified and the dynamic of the time series that are part of the cluster. In the figure we can appreciate that despite having different dynamics, several signals can belong to a group since they share similar inner structures, the local interactions of those basic structures are what make the time series behave in a particular way.

Table 1. Experimental data set

Time Series	*Dynamics*	*Time Series*	*Dynamics*
Sine	periodic	Dow Jones	complex
Vanderpol	periodic	Kobe	complex
Qperiodic2	quasiperiodic	ECG	complex
Qperiodic3	quasiperiodic	EEG	complex
Mackey-Glass	chaotic	ASCII	complex
Logistic	chaotic	El niño	complex
Lorenz	chaotic	HIV DNA	complex
Rossler	chaotic	Human DNA	complex
Ikeda	chaotic	Lovaina	complex
Henon	chaotic	Plasma	complex
Cantor	chaotic	Primes	complex
Tent	chaotic	SP500	complex
A1	complex	Star	complex
D1	complex	Brownian motion	random
Laser	complex	White Noise	random

For illustrative purposes, in Figure 2 we show the same plane, but now showing the waveforms of the time series. Here we can observe how the elements of the principal clusters are related by the intricacy of the time series. In the leftmost cluster there are very intricated, noise-like time series, at the right extreme there are somewhat smooth waveforms and in the middle top are time series that are not as intricate or as regular, i.e. a blend of the first two clusters. From this results we can also speculate that a discrete and qualitative classifications such as "chaotic", "complex" and "random" are not always the best suited as the dynamics of the phenomena can form a continuous space if we use a quantitative approach.

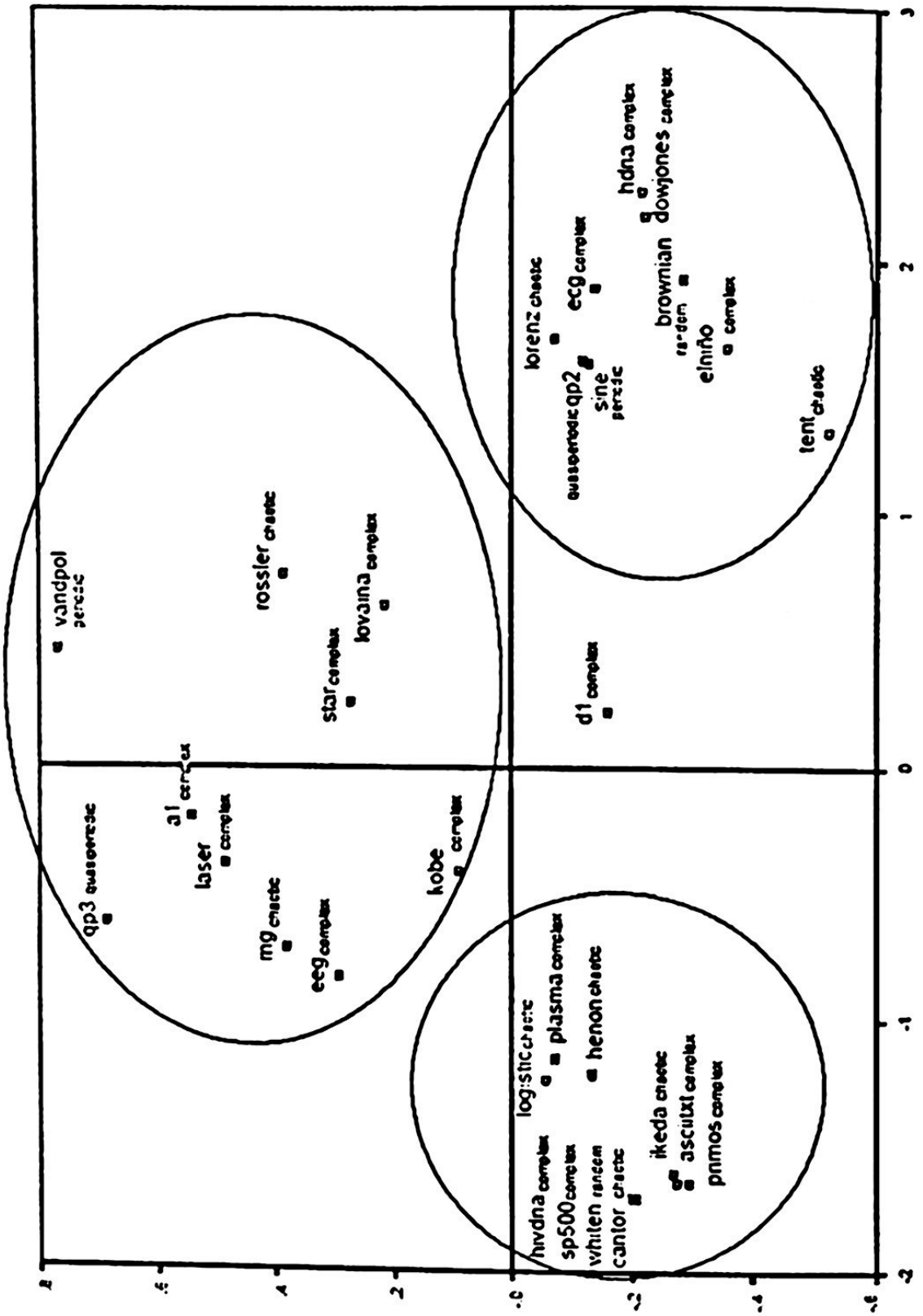

Fig. 1. Clustering of the series in the MDS plane

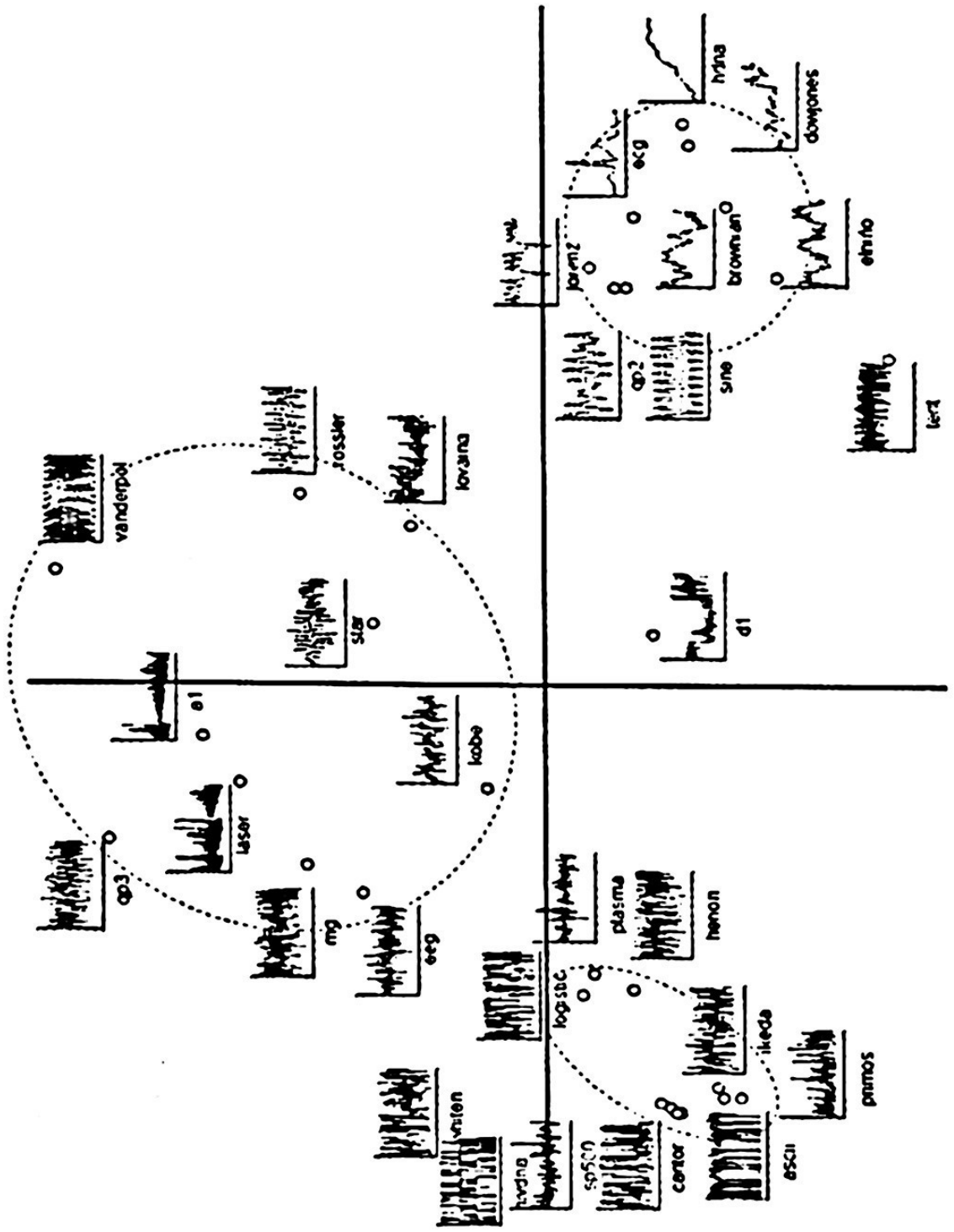

Fig. 2. Waveforms of the experimental data set in the MDS plane

Table 2. Variance percentage of the first five principal components

Time Series	*%Var PC1*	*%Var PC2*	*%Var PC3*	*%Var PC4*	*%Var PC5*
Sine	88.158	11.842	0	0	0
Vanderpol	59.752	38.884	0.698	0.609	0.028
Qperiodic2	88.116	11.608	0.247	0.021	0.005
Qperiodic3	33.545	31.809	19.528	13.291	1.156
MackeyGlass	29.777	27.577	16.732	12.074	3.14
Logistic	16.341	15.725	10.029	9.394	6.667
Lorenz	89.332	10.031	0.604	0.031	0.002
Rossler	67.399	31.077	1.113	0.316	0.068
Ikeda	8.037	8.03	7.72	7.589	6.976
Henon	16.398	14.592	10.519	6.5	6.448
Cantor	6.404	6.375	5.757	5.619	5.343
Tent	80.689	2.263	2.23	1.758	1.745
A1	43.903	35.677	6.724	6.425	2.145
D1	65.134	12.674	11.936	4.46	2.148
Laser	38.393	34.912	5.853	5.391	3.606
Dowjones	96.723	2.114	0.431	0.227	0.149
Kobe	23.285	22.457	11.749	10.45	6.366
Ecg	93.729	6.047	0.21	0.012	0.002
Eeg	26.944	24.338	18.28	5.541	3.208
Ascii	7.435	7.297	6.284	6.069	6.047
El niño	87.321	5.593	2.999	1.8	0.908
HIV DNA	6.452	6.332	6.077	5.491	5.297
HumanDNA	99.848	0.094	0.023	0.01	0.006
Lovaina	52.332	38.383	8.265	0.835	0.125
Plasma	18.074	17.677	7.626	6.517	5.984
Primos	7.343	6.482	6.359	6.107	5.526
Sp500	7.407	7.353	6.588	6.528	5.906
Star	56.218	31.681	1.665	1.143	1.07
Brownian m.	94.099	3.936	0.723	0.348	0.247
Whitenoise	6.449	6.178	6.044	5.309	5.296

5 Conclusions

We have analyzed a set of several time series with diverse dynamics by means of two complementary analysis techniques, SSA and PCA and employed the multidimensional scaling for the process of clustering and visualization of the obtained data.

As we can see in the results, the time series did not form clusters based on their dynamic behavior. Instead, the grouping was based in the similarity of basic structures that are common to the series. This means that phenomena with different dynamics can share similar internal patterns although the local interactions of patterns lead to diverse behavior

It is also worth of notice that the three main clusters that were identified are conformed by time series of comparable intricacy, one group was formed from very intricate data another from somewhat smooth waveforms and a third one of a merge of the previous.

Thus, the proposed scheme can be useful to classify data of diverse nature with hidden patterns, disregarding the outward appearance of the time series.

References

1. J. Rodríguez-Elizalde, J. Figueroa-Nazuno, "Quirón: Búsqueda de Similitud en Series de Tiempo por Métodos Espaciales de Acceso",IEEE ROC&C'2003, (2003)
2. J. Friedman, "Regularized Discriminant Analysis" Journal of the American Statistical Association (1989) 165-175
3. Agrawal, R., Lin, K. I., Sawhney, H. S. & Shim, K "Fast similarity search in the presence of noise, scaling, and translation in time-series databases" In proceedings of the 21st Int'l Conference on Very Large Databases (1995). pp 490-50.
4. E. Keogh, and M. Pazzani, "An enhanced representation of time series which allows fast and accurate classification clustering and relevance feedback". In 4th International Conference on Knowledge Discovery and Data Mining, (1998) 239-243.
5. J. Rodríguez-Elizalde, J. Figueroa-Nazuno "Agrupación de Series de Tiempo con Semejanza por Representación Simbólica", IEEE ROC&C'2003, (2003)
6. N Golyandina, V. Nekrutkin, A. Zhiglajavsky, "Analysis of Time Series Streuture", Chapman & Hall, (2001)
7. D. Hand, H. Mannila, P. Smyth, "Principles of data mining", MIT Press, (2001)
8. J.M. Colebrook,: "Continuous plankton records - zooplankton and environment, northeast Atlantic and North-Sea", Oceanol. Acta, 1, (1978) 1948-1975
9. I. Borg, P. Groenen, "Modern Multidimensional Scaling" Springer Verlag, New York, (2001)
10. J. C. Sprott "Chaos and time series analysis" Oxford University Press, (2004)
11. A. Espinosa-Contreras and J. Figueroa-Nazuno, "Análisis del comportamiento de la pérdida de paquetes en la red Internet con técnicas de la dinámica no-lineal", Memorias del Congreso Internacional de Computación CIC2000, (2000), 529-535