Anniversary Symposium

Proceedings SASM 2005 - Celebrating 25 years from the birth of the Seminar Gr.C. Moisil and 15 years from the establishing of the Romanian Society for Fuzzy Systems & A.I.

Iasi, Romania, May 5-7, 2005

The DBSOW Method in Language Analysis

Lucian-Iulian Fira[*], Horia-Nicolai Teodorescu[*,**],

(*) *Institute for Theoretical Informatics of the Romanian Academy*
(**) *Technical University of Iasi*
lfira@etc.tuiasi.ro, hteodor@etc.tuiasi.ro

Abstract. *In several previous papers, the second author has suggested that the dynamics of the words might be analyzed by mean of Distance Between Successive Occurrence of the Words (DBSOW). In this paper, the natural language is analyzed using one-step-ahead predictors for the distance between words time series.*

1. Introduction

During the past decades, a large amount of research has been performed on the statistical properties of the natural language. A statistical analysis of the m-grams and the words in the Romanian language, including analysis of texts form different domains (literature and science) or from the same domain (for different authors or between texts for the same author) can be found in [1-3].

In previous papers, the second author [4-6] suggested a dynamic approach for the natural language analysis. Statistical and nonlinear analysis of the distance between successive occurrence of the words was made.

In this paper, the natural language is analyzed using a neuro-fuzzy approach to predict the distance between words time series. One of the goals of this paper is to develop a preprocessing method and a bi- or multi-block predictor to perform high quality predictions for the time series obtained by DBSOW method.

Like in [6], the focus is on the words that connect phrases, ideas. Such connecting words are assumed to play a cognitive role in the discourse generation. Such words, i.e. SI (AND), are quite frequent in natural language texts. We hypothesize that, the analysis of the DBSOW time series for connecting words one may help determining the domain of the text, the author of the text, and the papers of the same author.

A mechanism to use a modeling technique to classify a text by the domain or by the author was suggested in [6]. A predicting system, trained to model a given DBSOW time series, learns the statistic and the series. In turn, the dynamics may be a fingerprint for the text and, consequently, for the author or for the domain. If a text assumed to belong to a given author does not match the model and yields a high modeling error, the authorship hypothesis is rejected [6].

A similar method was already developed and tested for the genomic sequences recognition [7-13]. Notice that for the DNA sequences a set of four one-step-ahead predictors was trained, one for each nitric basis type (Adenine, Cytosine, Guanine, and Thymine). On an upper hierarchical level, a decision system mixes the information from the individual predictors [7]. The above described methodology, applied to the genomic sequences recognition, can be extent for the texts classification by using a list of models / predictors corresponding to a set of characteristic words for the texts.

The paper is structured as follows: the next section is devoted to the description of the methodology. The third section contains several simulation results. In the fourth section, conclusions are outlined.

2. Methodology

An already learned sequence will give a small prediction error at a subsequently testing. A foreign sequence might be rejected due of high prediction error. To verify the methodology, we tested linear predictors (linear combiner), neuronal predictors (RBF or MLP type), and neuro-fuzzy predictors (Yamakawa model based).

2.1. The predicting systems

The class of a predictor is given by the input-output function of the predicting system. We tried several predictors as: linear predictors based on linear combiners, RBF predictors, and neuro-fuzzy predictors. In the case of linear combiner predictor, the characteristic function is a linear weight sum of the delayed inputs:

$$f_1(x_n, x_{n-1}, x_{n-2}, \ldots, x_{n-k}) = w_0 + \sum_{j=1}^{k} w_j x_{n-j+1} \tag{1}$$

where k represents the predictor order, w_0 is the bias, and w_j are the weights of the linear combiner.

An RBF network with Gaussian neurons in the hidden layer has the characteristic function as a linear combination of Gauss functions:

$$f_2(x_n, x_{n-1}, x_{n-2}, \ldots, x_{n-k}) = w_0 + \sum_{i=1}^{H} w_i \cdot \exp\left(-\frac{\sqrt{\sum_{j=0}^{k} (x_{n-j} - c_{ij})^2}}{\sigma} \right)^2 \quad (2)$$

where k represents the predictor order, w_0 is the bias, and w_j are the weights of the output neuron. H is the hidden Gauss type neurons number. σ are the spreading of the Gauss type functions and c_{ij} are the centers.

In case of the neuro-fuzzy predictor, the architecture is a multi-fuzzy system network with inputs represented by the delayed samples. The fuzzy cells acting as multipliers of inputs are Sugeno type 0, with Gauss input membership functions. The input-output function is a ratio with sums of exponentials at the nominator and the denominator.

$$f_3(x_n, x_{n-1}, x_{n-2}, \ldots, x_{n-k}) = \sum_{j=0}^{k} w_j \frac{\sum_{i=1}^{N} \beta_{ij} \cdot \exp\left(-\left(x_{n-j} - c_{ij}\right)/\sigma\right)^2}{\sum_{i=1}^{N} \exp\left(-\left(x_{n-j} - c_{ij}\right)/\sigma\right)^2} \quad (3)$$

where k is the predictor order, N is the input membership function number for each Sugeno fuzzy system, c_{ij}, β_{ij} are the centers of the Gauss type input membership function, respectively the output singleton $\#i$ of the fuzzy system $\#j$. σ are the spreadings of the Gauss type functions. w_j represent the weights associated to the output of the system $\#j$.

2.2. The preprocessing of the distances between words time series

In order to make prediction in similar conditions for all predictors, the distances between words time series was normalized to the [-1,1] interval. We notice that for the neuro-fuzzy predictors, the centers of the seven input membership functions uniformly cover the [-1,1] interval; thus, the input values must be in this interval.

Two methods are adopted for the distance series prediction: the direct prediction of the original time series and the prediction by components followed by the prediction results cumulating. The decomposition is made using a causal MA filter.

Another processing stage consists in a separation of the original time series into a slow varying (also named trend) and a fast varying component. This decomposition is made using a low pas moving average filter. The aim of the separation is to improve the prediction quality by developing individual predictors for each component. Then, the individual prediction results are cumulated. Several filters are tested for the decomposition task.

The main required condition was to use causal filters. If that non-mandatory condition is not satisfied, the results consist in a false prediction. We notice that, in several works, e.g. [14], the methodology used for the time series preprocessing for prediction implies the use of non-causal filters. Below, in the results section, a contra-example will be presented, to show that that choice is not acceptable.

Another condition for used filter is to delay not the filtered signal. A convenient filter satisfying both requirements is a 3-order MA filter given by the equation (4).

$$y[n] = (10x[n] + 5x[n-1] + 3x[n-2] + 2x[n-3])/20 \quad (4)$$

where x is the original signal and y is the slow varying component.

The fast varying component is obtained by the subtraction of the slow component from the original signal.

3. Results

We used a time series consisting in the distances between successive occurrence of 'SI' word in Bible – Genesis. The obtained time series and the components, resulted after separation are illustrated in Fig. 1.

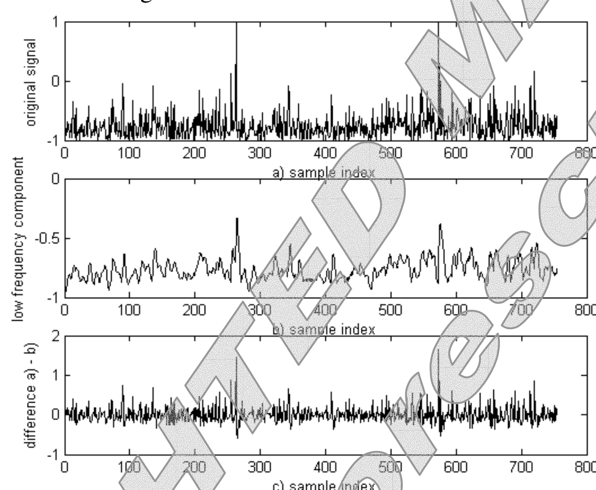

Fig. 1. Original signal and component separation

In the upper panel of the Fig. 1, the normalized original component is shown. In the middle panel, the trend component obtained using the filter in Equation (4) is illustrated. The fast varying component is shown in the lower panel.

In Table 1, the results obtained by searching of a predictor for the fast varying component are shown.

TABLE 1. The searching of a predictor for the fast varying component

RBF neurons	Spread	Train MSE	Test MSE	Theil Coeff.
45	0.5	0.009193	0.020882	0.644684
36	0.5	0.009529	0.019884	0.629274
27	0.5	0.009949	0.019356	0.620724
18	0.5	0.010911	0.016085	0.56587
17	**0.5**	**0.011022**	**0.015949**	**0.563127**
16	0.5	0.011082	0.016001	0.564139
15	0.5	0.011169	0.01601	0.564242
14	0.5	0.01126	0.016146	0.566654
9	0.5	0.011894	0.016349	0.569956

The optimum configuration was obtained for 17 RBF neurons with spreads of 0.5. The evaluation of the performances was made by means of the Mean Square Error (MSE), both, for train and test period. Also, the Theil coefficient was computed.

The Theil coefficient compares the RMSE error for the obtained prediction and for the naive prediction. If the current value of a time series is y_t, then the naive prediction will be $y'_{t+1} = y_t$.

If we have a desired series $\{y_{t,\, t=1,N}\}$ and a predicted series $\{y'_{t,\, t=1,N}\}$, then the Theil coefficient is defined as ([15] quoting [16]):

$$T = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - y'_t)^2} \Big/ \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - y_{t-1})^2} . \qquad (5)$$

A value of 1 for the Theil coefficient means that the current prediction is similar to naive prediction and the quality of prediction is improved to zero.

In Fig. 2, the prediction result obtained for the optimum predictor of the fast varying component is shown. The full line represents the desired signal and the dotted line represents the predicted signal.

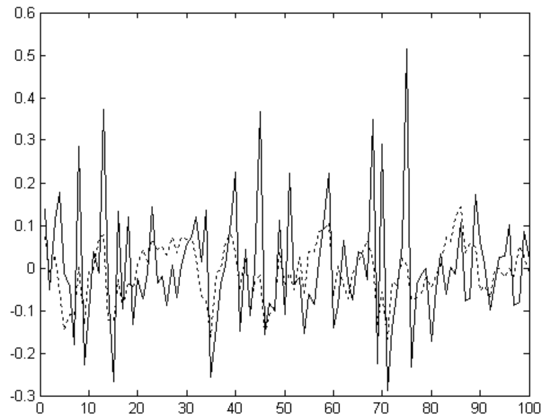

Fig. 2. Prediction results obtained for the optimum predictor of the fast varying component

TABLE 2. The searching of a predictor for the slow varying component

RBF neurons	Spread	Train MSE	Test MSE	Theil Coeff.
18	0.25	0.012493	0.014568	0.851778
18	*0.5*	*0.011559*	*0.012615*	*0.792969*
18	0.75	0.010836	0.01394	0.83354
18	1	0.010468	0.015182	0.869831
18	1.25	0.010561	0.013955	0.833641
27	0.25	0.012002	0.014708	0.85564
27	0.5	0.011059	0.014155	0.839982
27	0.75	0.010223	0.016999	0.920517
27	1	0.01015	0.017915	0.944953
27	1.25	0.010211	0.016537	0.907887
27	1.5	0.010267	0.016513	0.907277
9	0.5	0.013303	0.014641	0.853392
12	0.5	0.012033	0.013889	0.832042
15	0.5	0.011677	0.012883	0.801383
16	0.5	0.011618	0.012691	0.795359
17	**0.5**	**0.011559**	**0.012618**	**0.793073**

The optimum configuration was obtained for 17 RBF neurons with spreads of 0.5 from both MSE and Theil coefficient on the test period. Since the RBF with 18 neurons and spreads 0.5 has the MSE for test period with 3×10^{-6} less than the RBF with 17 neurons, the performance advantage is insignificant compared to the model complexity increasing.

In Fig. 3, the prediction result is illustrated for the optimum predictor obtained for the trend component.

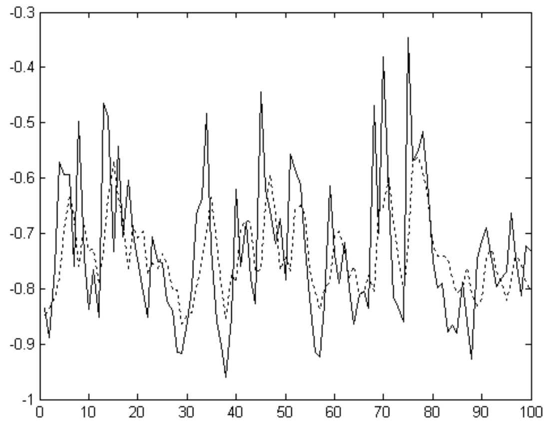

Fig. 3. Prediction results obtained for the optimum predictor for the slow varying component

In Fig. 4, the prediction results obtained for the slow and the fast varying component are cumulated.

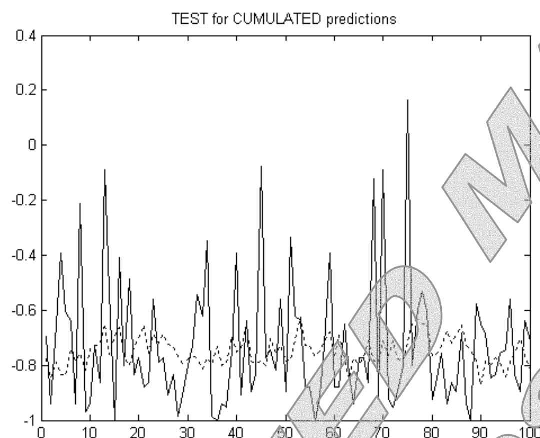

Fig. 4. Prediction results obtained by cumulating the individual predictions

In Fig. 5, the prediction results obtained by directly using the original (non-decomposed) distances series are shown.

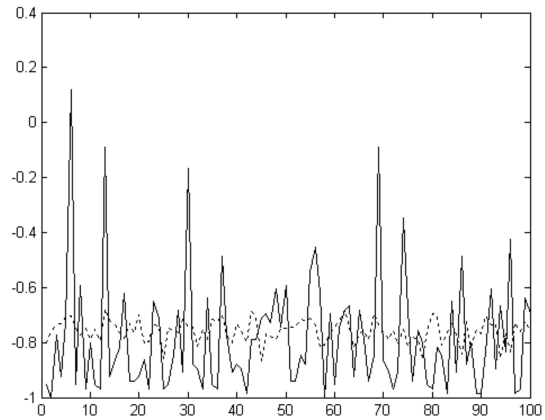

Fig. 5. Prediction results obtained by using of the original series for prediction

In Table 3, a comparison between the two methods for prediction using RBF predictors is made. The directly use of the original time series for the prediction task shows better results.

TABLE 3. 9th order RBF predictors comparison

Time series	Train MSE	Test MSE	Theil Coeff.
Fast varying component	0.011022	0.015949	0.563127
Slow varying component	0.011559	0.012618	0.793073
Cumulated predictions	0.042888	0.05456	0.673375
Original series prediction	0.044334	0.044614	0.669992

In Table 4, a comparison between performances obtained using the direct prediction and the prediction by components is made. In this case, the method of direct prediction is still better.

TABLE 4. 9th order adaptive linear combiner predictors comparison

Time series	Train MSE	Test MSE	Theil Coeff.
Fast varying component	0.011831	0.015097	0.547715
Slow varying component	0.011081	0.01391	0.832082
Cumulated predictions	0.045074	0.057239	0.689287
Original series prediction	0.044333	0.055711	0.680215

In Table 5, performances obtained using the prediction by components are shown.

TABLE 5. 9th order neuro-fuzzy predictors performances

Time series	Train MSE	Test MSE	Theil Coeff.
Fast varying component	0.015675	0.020086	0.629931
Slow varying component	0.012275	0.015921	0.890671
Cumulated predictions	0.048784	0.065672	0.737696

For all predictor models, a 9-order predictor is tested. The performances for the direct prediction were better in the cases of the RBF and the ALC predictors. Overall performance was obtained for RBF predictor, using direct prediction.

The method of prediction by decomposition of time series is not useful in that case.

A contra-example

The prediction results can be falsified by using of a non-causal filter. In our example, a MA filter computes the average between the current sample and their neighbors.

In the Fig. 6, the prediction result obtained for the predictor trained for the fast varying component is shown.

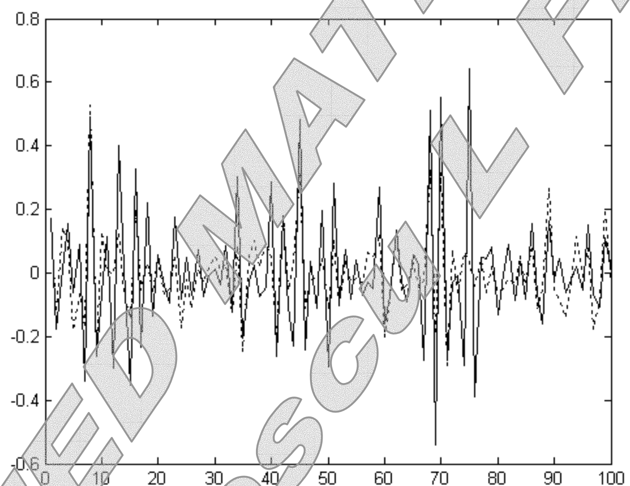

Fig. 6. Prediction results obtained for the fast varying component

In Fig. 7, a prediction result is illustrated for the predictor trained for the trend component. Notice that the prediction error is quite small and no significant average delay between the predicted and the actual values occurs.

In Fig. 8, the prediction results obtained for the slow and the fast varying component are cumulated.

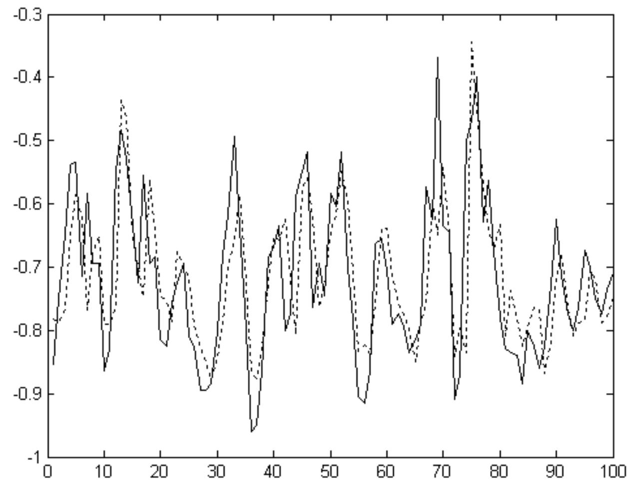

Fig. 7. Prediction results obtained for the trend component

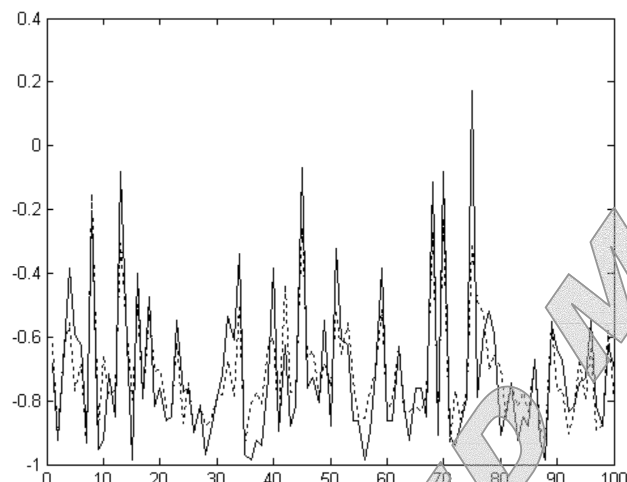

Fig. 8. Prediction results obtained by cumulating the individual predictions

TABLE 6. 9 order RBF predictors comparison

Time series	Train MSE	Test MSE	Theil Coeff.
Fast varying component	0.009412	0.01953	0.370735
Slow varying component	0.007098	0.008272	0.7956
Cumulated predictions	0.00856	0.013588	0.338273
Original series prediction	0.044334	0.044614	0.669992

The performances obtained using the prediction by components are shown in Table 6. Since the results are quite good for the method of predicting by components, these results are false. As we have noticed, the method of prediction by components, for the used distances time series, is applicable neither in case of causal filters, nor in case of non-causal filters.

4. Discussion and conclusions

In this paper, an approach to predict the distance between words time series was addressed. The time series are representing the distances between the successive occurrence of 'SI' and have been obtained from the Bible, Genesis.

Two methods were adopted for the distance series prediction: the direct prediction of the original time series and the prediction by components followed by prediction results cumulating. The decomposition was made using a causal MA filter.

For prediction performances were tested several predictor models as adaptive linear combiner, RBF, and neuro-fuzzy predictors.

An example of how prediction results can be falsifies by the use of non-causal filters was also shown, contradicting what appears to be a quite popular belief in the literature.

We notice that our previous results, reported in [10] might be affected by the use of the classical decomposition method [7]. In the present paper we showed that the use of the non-causal filtering falsifies the predictions.

The second author speculates that it should and it might be some level of predictability of the cognitive processes related to the natural language, moreover that the predictability may be a measure for the cognitive process basing the language communication process.

The high prediction accuracy obtained for the trend component might suggest that it exists some predictability level in the series. The high prediction performances on the trend series might be viewed as indication that the natural language is predictable on the long term.

Acknowledgments.

The CNCSIS Grant 149/2005 "System for the analysis and prediction of genomic sequences based on neuro-fuzzy data-mining methods" has supported part of the research for this paper, namely the experimenting by the second author. This research is party performed for the Romanian Academy prioritary grant "Cognitive systems and applications" ("Sisteme cognitive şi aplicații"); however, there was no financial support from this Grant for this research. The second author has received no grant or other financial support for the research reported here; consequently he reserves all the rights on this research.

References

[1]. A. Vlad, A. Mitrea, M. Mitrea, Limba română scrisă ca sursă de informație, Ed. Paideia, Bucuresti, România, 2003.

[2]. A. Vlad, A. Mitrea, M. Mitrea, Printed Romanian Modelling: the m-Grams and the Word Information Sources, In Speech Technology and Human-Computer Dialogue

Proceedings of the 2-nd Conference, Bucharest, April 10-11, 2003, Romanian Academy Publishing House, Bucharest, 2003, pp. 79-98.

[3]. H.-N. Teodorescu, "Grant: A2105, Imbunatatirea aspectelor prozodice in sinteza text-to-speech pentru limba romana", Revista Politica Stiintei si Scientometrie, Numar Special 2005- ISSN –1582-1218. In colaborare (la grant) cu D. Cristea, C. Zamfir, V. Apopei, L. Fira, Al. Ceausu, M. Stavila, C. Branzila.

[4]. H. N. Teodorescu, The Dynamics of the Words. The 11th Conference on Applied and Industrial Mathematics (CAIM 2003): 29 - 31 May, 2003. University of Oradea

[5]. H.N. Teodorescu, What the (DDWO) distribution of the distances of the word occurrences can tell us? Simpozionul National de Sisteme Inteligente si Aplicatii, Iasi, 19-20 sept. 2003

[6]. H.N. Teodorescu, C.B. Brânzilă, Analysis of the Dynamics of the Words in the Romanian and English Languages, CD-Proc. ATM2003 6-7 Nov. 2003, pp. 239-246

[7]. H.N. Teodorescu: Genetics, Gene Prediction, and Neuro-Fuzzy Systems-The Context and a Program Proposal. F.S.A.I., Vol. 9, Nos. 1-3, 15-22, 2003.

[8]. H.N. Teodorescu, L.I. Fira, Predicting the Genome Bases Sequences by means of distance sequences and a Neuro-Fuzzy Predictor, Fuzzy Systems & A.I. - Reports and Letters, vol. 7 (2003).

[9]. L.I. Fira, H.N. Teodorescu, Genome Bases Sequences Characterization by a Neuro-Fuzzy Predictor, Proceedings IEEE-EMBS 2003 Conference, 17-21 September, Cancun, Mexico.

[10]. H.N. Teodorescu, L.I. Fira, A Hybrid Data-Mining Approach in Genomics and Text Structures, The Third IEEE International Conference on Data Mining ICDM '03, Melbourne, Florida, USA, November 19 - 22, 2003.

[11]. L.I. Fira, H.N. Teodorescu, Analiza unor secvente de baze din genom cu un predictor neuro-fuzzy, Proceedings SIA'2003, Simpozionul National de Sisteme Inteligente si Aplicatii, 19-20 Septembrie 2003, Iasi.

[12]. H.N. Teodorescu, L.I. Fira, DNA Sequence Pattern Identification using A Combination of Neuro-Fuzzy Predictors, 11th International Conference on Neural Information Processing, ICONIP2004, November 22-25, 2004 Science City, Calcutta.

[13]. L.I. Fira, H.N. Teodorescu, Neural Network Implementation of a Decision Block for the Evaluation of Genomic Sequences Recognition Scores, CSCS15, The 15th International Conference on Control Systems and Computer Science, May 25-27, 2005, Bucharest, Romania

[14]. T. Popescu, Serii de timp – aplicatii in analiza sistemelor, Ed. Tehnica, Bucuresti, 2000.

[15]. A. Foka, Time Series Prediction Using Evolving Polynomial Neural Networks, A dissertation submitted to the University of Manchester Institute of Science and Technology for the degree of MSc, 1999.

[16]. Farnum N. R., Stanton L. W., Quantitative forecasting methods, PWS-KENT, 1989.