Time Series Pattern Recognition Based on MAP Transform and Local Trend Associations

Ildar Batyrshin and Leonid Sheremetov

Mexican Petroleum Institute, Av. Lázaro Cárdenas, 152,
Col. San Bartolo Atepehuacan, Mexico D.F., CP 07730, Mexico
{batyr, sher}@imp.mx

Abstract. The methods of pattern recognition in time series based on moving approximation (MAP) transform and MAP image of patterns are proposed. We discuss main properties of MAP transform, introduce a concept of a MAP image of time series and distance between time series patterns based on this concept which were used for recognition of small patterns in noisy time series. To illustrate the application of this technique to recognition of perception based patterns given by sequence of slopes, an example of recognition of water production patterns in petroleum wells used in expert system for diagnosis of water production problems is considered.

Keywords: Moving approximation transform, local trend association, time series, pattern recognition.

1 Introduction

Time series contain important information about measured parameters of systems changing in time. Such systems can be found in meteorology, economics, finance, geophysics, industry, and telecommunications. Time-series analysis is an important research area in these domains and more accurate analysis tools are consistently sought. Conventional techniques of time series analysis are based on statistical modeling, Fourier and wavelet transforms [4-6], [8], [12-14], [16], [17]. Techniques developed for noise suppression, data filtering and pattern recognition often suppose that time series describe some signal propagation or oscillating processes, contain large patterns, small noise, or noise with known statistical parameters. Many of these conditions are not fulfilled for time series describing economic or industrial systems [7], [9], [10]. In these application domains (i) time series are already obtained as a result of averaging of some parameter during given time intervals; (ii) they do not represent some oscillating processes; (iii) recognition of small patterns is important for decision making; and (iv) small patterns in time series can exist in presence of a noise. Moreover, in decision making procedures related with time series analysis in these domains, a trend or a tendency in a change of a parameter during some time interval usually becomes an important characteristic and the powerful signal processing techniques can not be applied.

J.F. Martínez-Trinidad et al. (Eds.): CIARP 2006, LNCS 4225, pp. 910 – 919, 2006.

The formal technique for analysis of such "local trends" was introduced in [1], [2]. The basis of this technique is an analysis of slopes of linear approximations of time series in a sliding window. This technique consists of two parts: i) moving approximation (MAP) transform calculates slopes in sliding window of a given size; and ii) a "local trend" association measure calculates a similarity between time series or time series patterns. This technique was applied in [1], [2] for analysis of associations between time series and representation of these associations as an association network of time series. The paper [3] considered the method of application of this technique to time series forecasting. In this paper we discuss the methods of application of MAP transform and local trend association measures to the recognition of time series patterns.

The rest of the paper is organized as follows. In the next section, we consider the basic notions of MAP transform and local trend associations. In section 3, we propose novel methods of coding of time series patterns by local trends and a pattern recognition procedure for extraction of given patterns from time series. Several measures of local trend association between time series patterns (*E*-distance, *J*-distance and *EJ*-distance) are introduced and compared on examples of time series pattern recognition in the presence of noise. Finally we discuss an application of this approach to the recognition of water production patterns in petroleum wells which are important for diagnosis of water production problems in petroleum industry. In Conclusions we discuss the main results presented in the paper and possible application and extension of the proposed technique.

2 Basic Notions of MAP Transform and Local Trend Associations

A time series (y,t) is a sequence $\{(y_i,t_i)\}$, $i \in I = (1,\ldots, n)$, such that $t_i < t_{i+1}$ for all $i = 1,.., n\text{-}1$, where y_i and t_i are real numbers called time series values and time points, respectively. A time series (y,t) will be denoted also as y. A window W_i of a length $k > 1$ is a sequence of indexes $W_i = (i, i+1,\ldots, i+k\text{-}1)$, $i \in \{1,\ldots, n\text{-}k+1\}$. The sequence $y_{W_i} = (y_i, y_{i+1},..., y_{i+k-1})$ of the corresponding values of time series y is called a partial time series induced by window W_i. A sequence $J = (W_1, W_2,\ldots, W_{n-k+1})$ of all windows of size k, $(1 < k \le n)$, is called a moving (or sliding) window. Such mowing window is used, for example, in statistics in moving average procedure for smoothing time series when the value in the middle of the window is replaced by the mean of values from this window.

Suppose J is a moving window of size k and $y_{W_i} = (y_i, y_{i+1},..., y_{i+k-1})$, $i \in (1,2,\ldots, n\text{-}k+1)$, are corresponding partial time series in time points $(t_i, t_{i+1},\ldots, t_{i+k-1})$. A linear function $f_i = a_i t + b_i$ with parameters $\{a_i,b_i\}$ minimizing the criterion

$$Q(f_i, y_{W_i}) = \sum_{j=i}^{i+k-1} \left(f_i(t_j) - y_j\right)^2 = \sum_{j=i}^{i+k-1} \left(a_i t_j + b_i - y_j\right)^2 , \qquad (1)$$

is called a moving (least squares) approximation of y_{Wi}. The solution of (1) is well known and optimal values of parameters a_i, b_i can de calculated as follows:

$$a_i = \frac{\sum_{j=i}^{i+k-1}(t_j - \overline{t_i})(y_j - \overline{y_i})}{\sum_{j=i}^{i+k-1}(t_j - \overline{t_i})^2},\ b_i = \overline{y_i} - a_i\overline{t_i}, \tag{2}$$

where $\overline{t_i} = \frac{1}{k}\sum_{j=i}^{i+k-1} t_j$, $\overline{y_i} = \frac{1}{k}\sum_{j=i}^{i+k-1} y_j$.

Definition 1. A transformation $MAP_k(y,t) = a$, where $a = (a_1, \ldots, a_{n-k+1})$ is a sequence of slope values obtained as a result of moving approximations of time series (y,t) in moving window of size k is called a moving approximation (MAP) transform of time series y. The slope values $a_1, \ldots, a_{n-k+1}$ are called local trends.

Elements a_i, $(i = 1,\ldots, n-k+1)$ from $MAP_k(y,t)$ will be denoted as $MAP_{ki}(y,t)$.

In many applications time points $t_1, \ldots, t_n$ are increasing with a constant step h such that $t_{i+1} - t_i = h$ for all $i = 1,\ldots, n-1$. In such cases, in MAP transform the set of time points $t = (t_1, \ldots, t_n)$ can be replaced by the set of indexes $I = (1,\ldots, n)$ as follows: $MAP_k(y,t) = (1/h)MAP_k(y,I)$ and the formula (2) for local trends can be simplified as follows [2].

Theorem 1. Suppose time points $t_1, \ldots, t_n$ are increasing with a constant step h then the values of MAP transform $MAP_k(y,t)$ can be calculated as follows:

$$a_i = \frac{6\sum_{j=0}^{k-1}(2j - k + 1)y_{i+j}}{hk(k^2 - 1)},\ i \in (1,2,\ldots, n-k+1). \tag{3}$$

Formula (3) gives a simple method to calculate MAP transform for time series with a fixed step. Further, for such time series we will replace time points by indexes $I = (1,\ldots, n)$ and in (3) the value $h = 1$ will be used. We will denote time series also as $y = (y_1,\ldots,y_n)$ and will use a notation $MAP_k(y)$ for $k \in \{2,\ldots,n-1\}$.

As a measure of similarity between time series one can use measures of similarity between their MAP transforms.

Definition 2. Suppose $y = (y_1,\ldots,y_n)$, $x = (x_1,\ldots, x_n)$ are two time series and $MAP_k(y) = (a_{y1}, \ldots, a_{ym})$, $MAP_k(x) = (a_{x1}, \ldots, a_{xm})$, $(k \in \{2,\ldots,n-1\}, m = n - k + 1)$, are their MAP transforms. The following function is called a measure of local trend associations:

$$coss_k(y,x) = \frac{\sum_{i=1}^{m} a_{yi} \cdot a_{xi}}{\sqrt{\sum_{i=1}^{m} a_{yi}^2 \cdot \sum_{j=1}^{m} a_{xj}^2}} \tag{4}$$

Suppose p,q,r,s, $(p,r \neq 0)$ are real values and (y,t) is a time series. Denote $py+q = (py_1+q, \ldots, py_n+q)$ and $rt+s = (rt_1+s, \ldots, rt_n+s)$. A transformation $L(y,t) = (py+q, rt+s)$ is called a linear transformation of time series (y,t).

Theorem 2. Suppose L_1 and L_2 are two linear transformations of time series *(y,t)* and *(x,t)* given by the sets of parameters (p_1,q_1,r_1,s_1) and (p_2,q_2,r_2,s_2), respectively, where $p_1, p_2, r_1, r_2 \neq 0$, then

$$coss_k(L_1(y,t), L_2(x,t)) = sign(p_1)\cdot sign(r_1)\cdot sign(p_2)\cdot sign(r_2)\cdot coss_k((y,t), (x,t)). \quad (5)$$

From this Theorem it follows a very nice invariance property of local trend association measure under various types of normalization of time series. For the lack of space, we leave the proving of the Theorems 1 and 2 out of the scope of this paper.

Analysis of associations between time series is based on the analysis of associations between them for different window size. The sequence of association values $AV(y,x)= (coss_2(y,x),\ldots, coss_n(y,x))$ for all sizes of window is called an association function [2]. A specific measure of association between time series is defined by the subset of window sizes $J \subset \{2,\ldots,n\}$ as a maximum or average of all associations $coss_k(y,x)$, $k \in J$. Examples of application of this association measure to the classification of time series are considered in [2]. In Section 3 it will be introduced a *J*-MAP image of time series pattern defined by a subset of window sizes *J*.

3 MAP Image and *J*-MAP Distance in Time Series Pattern Recognition

Suppose $y = (y_1, \ldots, y_n)$ is a time series with a constant time step. Our goal is to propose a method to find a pattern $p = (p_1,\ldots, p_m)$ in time series *y*. This problem is trivial if the pattern *p* coincides with some subsequence $x = (y_i, y_{i+1},\ldots,y_{i+m-1})$ of *y* such that $y_i= p_1$, $y_{i+1} = p_2,\ldots, y_{i+m-1} = p_m$. In real applications, such trivial cases usually do not take place and the problem is to find a subsequence in *y* which is most similar to the goal pattern. A simple algorithm to find such pattern is following: to move a window W_i of size *m* along time series and calculate a distance between goal pattern and partial time series $y_{W_i} = (y_i, y_{i+1},..., y_{i+m-1})$ induced by this window. The partial time series minimizing this distance can be chosen as a found pattern. As a commonly used distance one can use Euclidean distance between time series patterns:

$$d_E(x,p) = \sqrt{\sum_{k=1}^{m}(y_{i+k-1} - p_k)^2}\ . \quad (6)$$

Before searching patterns in time series one can apply some methods of data smoothing, data filtering or noise suppression, but in the presence of large errors such procedures developed mainly for signal processing can essentially deform time series which do not describe some wave propagation or oscillating process but, instead, describe a change of some time dependent economical, financial or industrial parameter. Fig. 1 gives an example of such time series distorted by different types of errors.

Here we propose a novel method of pattern recognition in time series based on MAP transformation of patterns. MAP transform smoothes data in a sliding window. For this reason we can suppose that it is less sensitive to errors in data than original data.

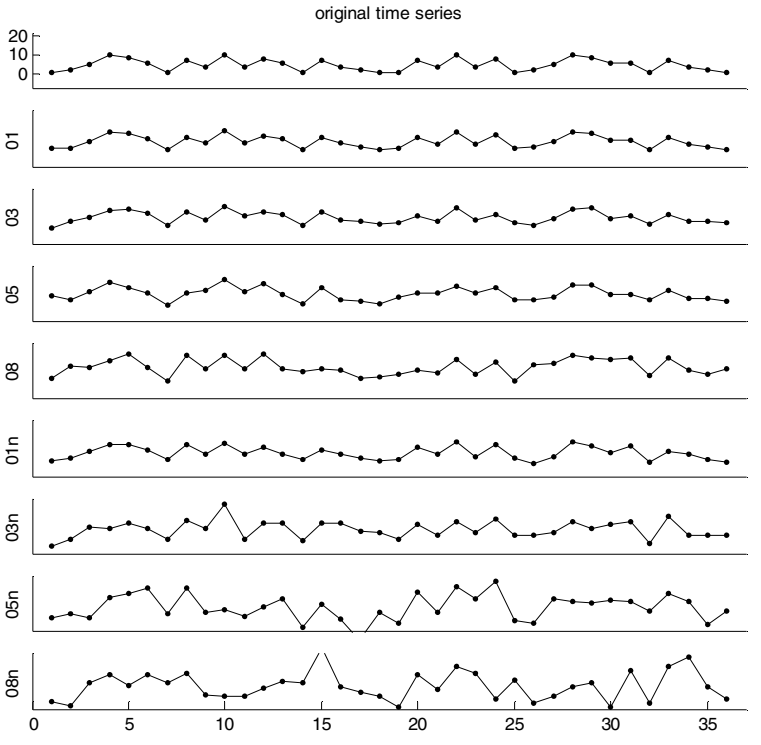

Fig. 1. Example of a synthetic time series (on the top of figure) taking values between 1 and 10 and noisy time series obtained from it by adding the following errors: 01, 03, 05, 08 denote time series with errors uniformly distributed in the intervals [0,1], [0,3], [0,5], [0,8], respectively; 01n, 03n, 05n, 08n denote time series with standard normal distribution errors multiplied by 1, 3, 5, 8, respectively.

Definition 3. A MAP image of pattern $p = (p_1,\ldots, p_m)$ is a sequence $MAPI(p)= (MAP_2(p),\ldots, MAP_m(p))$ of MAP-transforms of p for all possible window sizes. Suppose J is a subset of indexes $\{2,\ldots, m\}$. A J-MAP image of p is a sequence of MAP-transforms of p for all possible window sizes from J.

Fig. 2 shows an example of moving approximations of a pattern for different window sizes. The MAP transforms corresponding to these moving approximations contain the following sets of slope values of approximating lines: $MAP_2(p)$ = (1, 3, 5, -1, -3), $MAP_3(p)$ = (2, 4, 2, -2), $MAP_4(p)$ = (3, 2.6, 0.2), $MAP_5(p)$ = (2.4, 1.2), $MAP_6(p)$ = 1.4571. The corresponding MAP image of p will be represented by a sequence: ((1, 3, 5, -1, -3), (2, 4, 2, -2), (3, 2.6, 0.2), (2.4, 1.2), (1.4571)). For J = {3, 4} a J-MAP image of p will be equal to: ((2, 4, 2, -2), (3, 2.6, 0.2)).

Instead of a distance (6) between goal pattern p and partial time series $x = y_{Wi}$ in sliding window we propose to calculate a distance between J-MAP images of goal pattern p and partial time series $x = y_{Wi}$. This distance, called J-MAP distance, will be defined as follows:

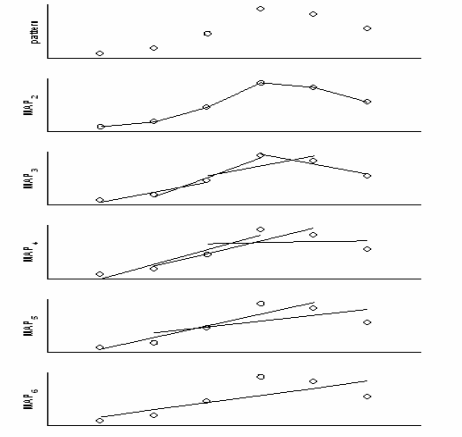

Fig. 2. Moving approximations of a pattern (given on the top of the figure) for all possible window sizes

$$d_J(x,p) = \sum_{k\in J} \sqrt{\frac{1}{m-k+1} \sum_{i=1}^{m-k+1} (MAP_{ki}(x) - MAP_{ki}(p))^2} \,. \tag{7}$$

As a distance measure combining both Euclidean distance and *J*-MAP distance we propose the following distance, which will be called *EJ*-MAP distance:

$$d_{EJ}(x,p) = d_J(x,p) + \frac{1}{\sqrt{m}} d_E(x,p) \,. \tag{8}$$

It is clear that instead of Euclidean distance one can use a normalized distance:

$$d_{NE}(x,p) = \sqrt{\frac{1}{m} \sum_{k=1}^{m} (y_{i+k-1} - p_k)^2} \,. \tag{9}$$

In this case (8) will be presented as follows:

$$d_{EJ}(x,p) = d_J(x,p) + d_{NE}(x,p) \,. \tag{10}$$

Experiments show that results of pattern recognition based on these distances can be different. Fig. 3 depicts an example of recognition of time series patterns in time series with noise. Fig. 3a shows search patterns in original time series and Fig. 3b depicts results of recognition of these patterns in time series distorted by noise.

Table 1 contains the results of comparison of considered three distance measures in their ability to recognize time series patterns in time series distorted by random errors. From time series shown on top of Fig. 1 were generated 300 random time series for each type of error described in the caption of Fig. 1. The distances for two sets of indexes J = {3,4,5,6} and J= {4,5,6} were applied. The results obtained for J = {3,4,5,6} are slightly better than for J = {4,5,6}. As one can see from Table 1, the *J*-MAP distance is more suitable for recognition of patterns in the presence of large

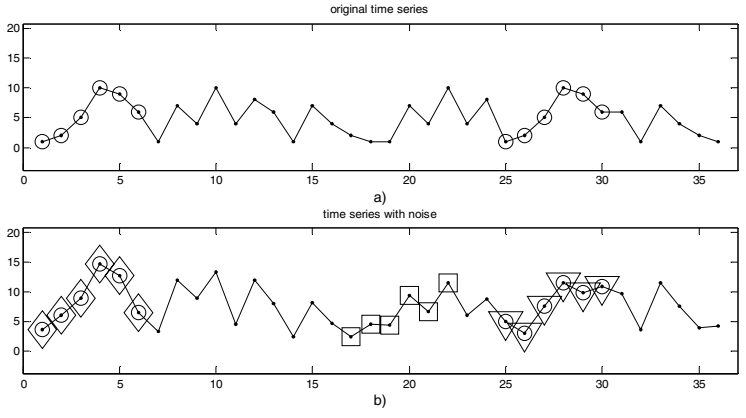

Fig. 3. a) Goal patterns (shown by circles "o") in original time series depicted also on top of Fig. 1; b) Two goal patterns in time series distorted by random errors are shown by circles (o). Patterns found by *J*-distance, *E*-distance and *EJ*-distance are shown by diamonds (◊), squares (□) and triangles (∇) respectively

uniformly distributed errors and Euclidean distance has some advantage in recognition of patterns with normally distributed errors. *EJ*-MAP distance can be used as a compromise between these two distances when the type of error is unknown.

4 Recognition of Water Production Patterns in Petroleum Wells

The proposed methods of time series pattern representation and recognition were realized as modules of the Percept-Miner toolbox [15]. They were also used in SMART-Agua, an expert system for diagnosis of water production problems in petroleum wells.

In petroleum industry it is convenient to describe a water production patterns in petroleum wells by slopes. MAP transform and local trend associations based on MAP give a natural method for analysis of such patterns. As an example of the application of the proposed technique we consider the case of the recognition of the four water entrance patterns important for diagnosis of water production problems. The real data are analyzed against four typical patterns describing "quick increase" and "slow increase" of water production related to the certain problems of excessive water production (Fig. 4). A methodology for interpreting the behavior of waterflooding is applied in order to analyze the behavior of the water oil ratio vs. time curve in various time domains (for example, following the breakthrough).

Since expert patterns were given by sequences of slopes, *J-MAP* distance was applied for search patterns in time series closest to the goal patterns. A screenshot (Fig. 4) shows found patterns most similar to given ones. The first column shows goal patterns defined by expert. The second column shows found patterns in time series most similar to the goal patterns. The columns 3 and 4 give numerical values of slopes in goal patterns and in found patterns. Depending on the importance of the

particular time interval for the expert, he can select the most appropriate pattern based on the similarity measure. This tool was developed as a support for knowledge engineer in describing his perceptions about water production patterns important both for diagnosis and for testing the expert rules considering such patterns.

Table 1. Percentage of correct recognition of patterns for different types of errors

Type of error	*E*-distance	*EJ*-distance	*J*-distance	*J* =
08	36	58	67	4,5,6
08	32	61	70	3,4,5,6
05	83	92	90	4,5,6
05	75	91	95	3,4,5,6
03	100	100	99	4,5,6
03	100	100	100	3,4,5,6
08n	17	16	16	4,5,6
08n	19	14	13	3,4,5,6
05n	28	24	21	4,5,6
05n	27	23	22	3,4,5,6
03n	64	62	55	4,5,6
03n	74	71	62	3,4,5,6
01n	100	100	96	4,5,6
01n	100	100	100	3,4,5,6

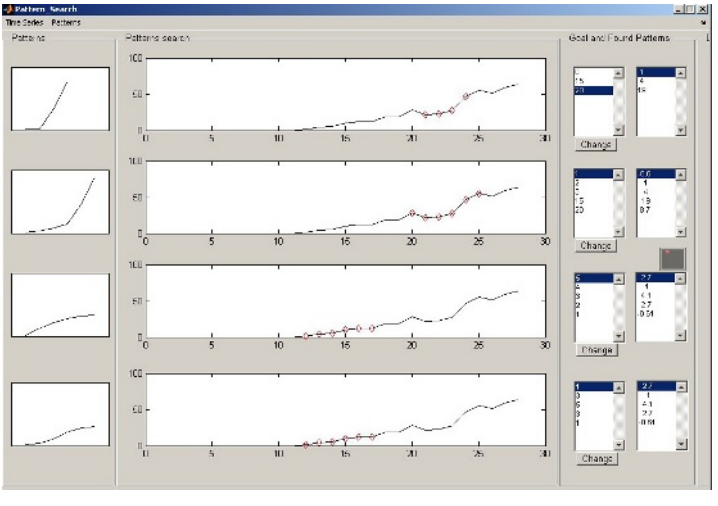

Fig. 4. Moving approximations of four patterns for all possible window sizes

5 Discussion and Conclusions

Human decision making in different application domains like economics, finance, or industry is often based on analysis of time series data bases. The main reasons for exploring a pattern recognition tools for time-series analysis are [17]: i) pattern

recognition methods are more flexible than other available tools in signal processing, statistics and neural networks and offer the user the ability to optimize their design for the best results; ii) results using such tools are easy to explain to the users as opposed to neural networks whose behavior is often difficult to rationalize; and iii) pattern recognition methodologies already offer a range of existing techniques that are well suited for time-series analysis. Pattern recognition methods offer different services for time-series analysis including its recognition, classification and prediction, e.g. speech classification, source separation or forecasting.

Such analysis usually contains description of time series patterns important for decision making. Two tasks arise from searching for such patterns in time series. First, time series data can be noisy and can contain large errors. Many of conditions supposed by techniques developed in time series data mining, signal processing and time series analysis for noise suppression, data filtering and pattern recognition are not satisfied in time series describing economic or industrial time series. Moreover, the goal patterns can be small and known methods of smoothing of time series can delete them. New methods of recognition of such patterns in noisy time series are proposed in this paper. Section 3 introduces the new concepts of MAP-image and *J*-MAP distance which give possibility to recognize small patterns in time series in the presence of large evenly distributed errors. For noisy time series *J*-MAP image acts as a filter. This novel technique is based on analysis of slopes of moving approximations of time series in a sliding window. It can be used also for recognition of patterns in time series given by sequences of slopes. It is a natural way to describe time series patterns important for decision making in several application areas. As an example, in Section 4 the use of this technique for recognition of water production patterns in petroleum wells given by sequences of slopes is discussed.

Several directions of future research can considered: 1) combination of proposed methods with traditional time series analysis technique like smoothing and filtering; 2) development of special methods based on the proposed technique for recognition of specific complex patterns in real time series; 3) extension of the proposed technique to solution of perception based time series data mining tasks where linguistically described patterns are used in decision making procedures.

Acknowledgements

The research work was supported by projects D.00006 and D.00322. Special thanks to R. Herrera for his invaluable support in the development of the software. Authors also highly appreciate the valuable comments of the reviewers of the paper.

References

1. Batyrshin I., Herrera-Avelar R., Sheremetov L., Suarez R.: Moving Approximations in Time Series Data Mining. In: Proc. of the Int. Conf. Fuzzy Sets and Soft Computing in Economics and Finance, June 17-20, St. Petersburg, Russia, Vol. I (2004) 62-72
2. Batyrshin I., Herrera-Avelar R., Sheremetov L., Panova A.: Association Networks in Time Series Data Mining. In: Soft Computing for Real World Applications, Proc. of the Int. Conf. NAFIPS, June 22-25, Ann Arbor, Michigan, USA. IEEE Comp. Soc. (2005) 754-759

3. Batyrshin I., Sheremetov L.: Perception Based Time Series Data Mining with MAP Transform. In: Advances in Artificial Intelligence, Lecture Notes in Computer Science, Vol. 3789, Springer-Verlag, Berlin Heidelberg New York (2005) 514 – 523
4. Bowerman, B.L., O'Connell, R.T.: Time Series and Forecasting; An Applied Approach. Duxbury Press, Massachusetts (1979)
5. Brockwell, P.J., Davis, R.A., Fienberg, S.E.: Time Series: Theory and Methods, Springer Series in Statistics, Springer-Verlag, Berlin Heidelberg New York (1991)
6. Chatfield, C.: The Analysis of Time Series, an Introduction, Sixth Edition. London: Chapman and Hall/CRC (2004)
7. Cheung, J.T.: Representation and Extraction of Trends from Process Data. D.Sci. Th., Massachusetts Institute of Technology, Cambridge/MA, USA (1992)
8. Hand, D.J.: Intelligent Data Analysis: Issues and Opportunities. In: Proc. of the Int. Conf. IDA97. Lecture Notes in Computer Science, Vol. 1280. Springer-Verlag, Berlin Heidelberg New York (1997) 1-14
9. Kivikunnas, S.: Overview of Process Trend Analysis Methods and Applications. In: Proc. of Workshop on Applications in Chemical and Biochemical Industry. Aachen, Germany. (1999)
10. Konstantinov, K.B., Yoshida, T.: Real-time Qualitative Analysis of the Temporal Shapes of (Bio) Process Variables. J. of Am. Inst. of Chem. Eng., Vol. 38 No. 11 (1992) 1703-1715
11. Least Squares Fitting. Wolfram Research. Mathworld. URL: http://mathworld.wolfram.com/LeastSquaresFitting.html
12. Mörchen, F.: Time Series Feature Extraction for Data Mining Using DWT and DFT. Data Bionics, Philipps-University Marburg, Germany, October 9 (2003)
13. Ponomaryov, V., Gallegos-Funes, F., Sansores-Pech, R., Sadovnychiy, S.: Real-time noise suppression in 3D ultrasound imaging based on order statistics. Electronics Letters, Vol. 42, No. 2 (2006) 80-82
14. Ponomaryov, V., Pogrebnyak, O.: Novel Robust RM Filters for Radar Image Filtering. J. of Electronic Imaging, Vol. 5 No 3, (1996) 410-421
15. Sheremetov, L.B., Batyrshin, I.Z, Filatov, D.M.: Perception Based Hybrid Intelligent Systems in Petroleum Applications. In: Proc. of the NAFIPS'06 Int. Conf., June 3-6, Montréal, Canada. IEEE Comp. Soc. (2006)
16. Shumway, R.H., Stoffer, D.S.: Time Series Analysis and its Applications, Springer-Verlag, Berlin Heidelberg New York (2000)
17. Singh, S.: Noise Impact on Time-Series Forecasting Using an Intelligent Pattern Matching Technique, Pattern Recognition, Vol. 32, Issue 8 (1999) 1389-1398