# Attractor-based models for individual and groups' forecasting

**N N Astakhova[1], L A Demidova[1, 2], A V Kuzovnikov[3] and R V Tishkin[1]**

[1] Ryazan State Radio Engineering University, Gagarin Str., 59/1, Ryazan, Russian Federation, 390005
[2] Moscow Technological Institute, Leninskiy pr., 38A, Moscow, Russian Federation, 119334
[3] JSC Academician M.F. Reshetnev Information Satellite Systems, Lenin Street, 52, Zheleznogorsk, Krasnoyarsk region, Russian Federation, 662972

E-mail: asnadya@yandex.ru

**Abstract**. In this paper the questions of the attractors' application in case of the development of the forecasting models on the base of the strictly binary trees have been considered. Usually, these models use the short time series as the training data sequence. The application of the principles of the attractors' forming on the base of the long time series will allow creating the training data sequence more reasonably. The offered approach to creation of the training data sequence for the forecasting models on the base of the strictly binary trees was applied for the individual and groups' forecasting of time series. At the same time the problems of one-objective and multiobjective optimization on the base of the modified clonal selection algorithm have been considered. The reviewed examples confirm the efficiency of the attractors' application in sense of minimization of the used quality indicators of the forecasting models, and also the forecasting errors on 1 – 5 steps forward. Besides, the minimization of time expenditures for the development of the forecasting models is provided.

## 1. Introduction

Usually the forecasting models on the base of the strictly binary trees use the short time series (about 15 – 30 elements) as the training data sequence [1 – 4]. Therefore, it is necessary to allocate reasonably some part of the analyzed time series (TS), if its length is significantly bigger. Obviously, that application of the long training data sequences will be followed by increase in time expenditures for the development of the forecasting models. Also, it can lead to receive of the not quite adequate forecasting model which will yield the bad forecasting results. This fact can be explained, firstly, by the influence of already insignificant data on the forecasting results, and, secondly, by the difficulties of matching of the forecasting models on the base of the SBT.

In this paper it is offered to use the attractor-based approach to find the adequate length of the training data sequence for the development of the forecasting models on the base of the strictly binary trees. The offered approach to the choice of the training data sequence can be applied for the individual and groups' forecasting of TSs [1 – 6], that will allow minimizing the time expenditures for the development of the forecasting models and the obtaining of the forecasting values.

The rest of this paper is structured as follows. Section 2 describes the main ideas of the development of the forecasting models on the base of the strictly binary trees (SBT) and the modified clonal selection algorithm (MCSA). Herewith, the main principles of one-objective and multiobjective

optimization are considereted. Section 3 details the attractor-based approach to the choice of the training data sequence for the forecasting model on the base of the SBT. Besides, the main recommendations on the attractors' application for the individual and groups' forecasting of TSs are provided. The experimental results which clearly highlight the efficiency of the offered approach to the choice of the training data sequence for the forecasting model on the base of the SBT follow in Section 4. Finally, conclusions are drawn in Section 5.

## 2. Forecasting models on the base of the strictly binary trees and the modified clonal selection algorithm

The principles of the development of the forecasting models on the base of the one-objective MCSA were investigated in [2]. The main ideas of the multi-objective MCSA are described in [6]. The MCSA allows forming an analytical dependence on the base of the SBT at an acceptable time expenses. This analytical dependence in the one-objective MCSA describes the certain TS values and provides a minimum value of the average forecasting error rate (*AFER*):

$$AFER = \frac{\sum_{j=k+1}^{n} |(f^j - d^j)/d^j|}{n-k} 100\%, \quad (1)$$

where $d^j$ and $f^j$ are respectively the actual (fact) and forecasted values for the $j$-th element of the TS; $n$ is the number of TS elements; $k$ is the model order.

Also the *AFER* (1) can be named as the affinity indicator $Aff$ [16].

The multiobjective MCSA (MMCSA) uses two quality indicator: the mentioned above affinity indicator $Aff$ and the tendencies discrepancy indicator *Tendency* :

$$Tendency = \frac{h}{n-k-1}, \quad (2)$$

where $h$ is the number of negative multiplications $(f^{j-1} - f^j) \cdot (d^{j-1} - d^j)$; $j = \overline{k+2, n}$; $d^j$ and $f^j$ are respectively the actual (fact) and forecasted values for the $j$-th element of TS; $n$ is the number of TS elements; $k$ is the model order; $n-k-1$ is the total number of multiplications $(f^{j-1} - f^j) \cdot (d^{j-1} - d^j)$.

This indicator describes the rate of discrepancy between the tendencies of two time series (real TS and model TS). Both indicators must be minimized.

Possible options for analytical dependences are presented in the form of antibodies $Ab$ which recognize antigens $Ag$ (certain TS values). An antibody $Ab$ is selected as "the best one". It provides the minimum value of the affinity indicator $Aff$ [1, 2]. Coding of an antibody $Ab$ is carried out by recording signs in a line. The signs are selected from three alphabets [1, 2]: the alphabet of arithmetic operations (addition, subtraction, multiplication and division) – $Operation = \{'+','–','\cdot','/'\}$; the functional alphabet, where letters $'S','C','Q','L','E'$ define mathematical functions "sine", "cosine", "square root", "natural logarithm", "exhibitor", and the sign $'\_'$ means the absence of any mathematical function, – $Functional = \{'S','C','Q','L','E','\_'\}$; the alphabet of terminals, where letters $'a','b',...,'z'$ define the arguments required analytical dependence and the sign $'?'$ defines a constant, $Terminal = \{'a','b',...,'z','?'\}$. The use of these three signs alphabets provides a correct conversion of randomly generated antibodies into the analytical dependence. The structure of such antibodies can be described with the help of the SBT [1, 2].

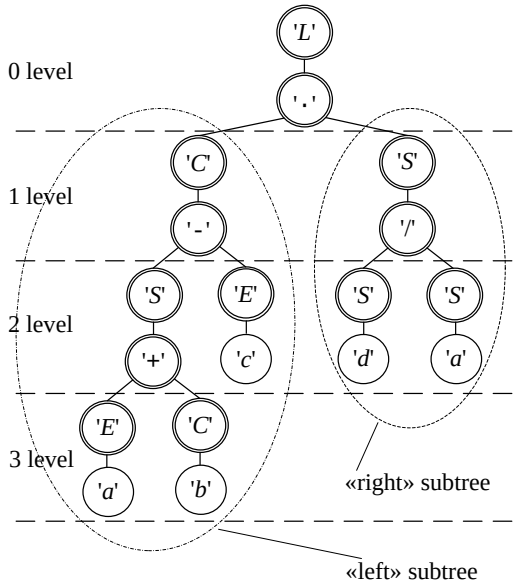



**Figure 1.** An example of a strict binary tree, which is used to form antibodies.

The number of signs in the alphabet of terminals *Terminal* in the antibody *Ab* determines maximal possible order $K$ of the forecasting models with $K \geq k$, where $k$ is the real model order), i.e. having the value of the element $d^j$ in the forecasting TS at the $j$-th moment of time, $K$ values of TS elements can be used as: $d^{j-K}$,…, $d^{j-2}$, $d^{j-1}$ [1 – 6].

The use of the SBT, illustrated in figure 1, allows building the complex analytical dependence and provides high accuracy of the forecasting TS [2]. Such SBT can be generated as a composition result of one "left" subtree of the maximum possible order $K = 3$ and some "right" subtrees of the maximum possible order $K = 2$. Thus the term "left" subtree ("right" subtree) is used for the branch (left or right) of SBT level in which it is necessary to include a new subtree. In this case it is rational to form antibodies by subdividing SBT into subtrees, then execute the subtree-walk of each vertex forming the ordered symbol lists on its vertices and then combining these lists consecutively [1 – 6].

Forming the symbol ordered list on the base of a subtree the consecutive double subtree-walk is carried out: at first moving the subtree bottom-up left to right we walk the vertices containing the alphabetic terminal signs *Terminal* in pairs and correspondingly above placed vertices containing the alphabetic functional symbols *Functional* and then moving in the same direction it is necessary to go around in pairs the vertices containing the alphabetic arithmetic operation signs *Operation* and correspondingly above placed vertices containing the alphabetic functional signs *Functional*. The first two signs in such an antibody contain the pair of zero level SBT from the functional alphabet *Functional* and arithmetic operation alphabet *Operation*.

Then there are the lists of the signs corresponding to the "right" maximum possible ordered subtrees $K = 2$ (moving the SBT bottom-up) and finally the symbol list of the "left" maximum possible ordered subtree $K = 3$. Using such a way of antibody formation we ensure the visualization of the SBT structure representation in the form of the subtrees union, and the antibody is easily interpreted in the analytical dependence.

For example, the antibody formed on the base of the SBT as shown in figure 1 is coded by the line of signs: $L \cdot S/SeSdC - S + EcCbEa$, which can be transformed into an analytical dependence:

$$f(a,b,c,d) = ln(cos(sin(exp(a) + cos(b)) - exp(c)) \cdot sin(sin(d) / sin(a))).$$

For the $k$-order forecasting model with $k = 4$ this analytical dependence can be written as:

$$f(d^{j-1}, d^{j-2}, d^{j-3}, d^{j-4}) = ln(cos(sin(exp(d^{j-1}) + cos(d^{j-2})) - exp(d^{j-3})) \cdot sin(sin(d^{j-4}) / sin(d^{j-1}))).$$

Interpreting the antibodies into the analytical dependences it is rational to use the recursive procedure of interpretation [1, 2]. The MCSA applied to the searching for "the best" antibody defining "the best" analytic dependence includes the preparatory part (realizes the formation of the initial antibody population) and iterative part (presupposes the ascending antibodies ordering of affinity *Aff* the selection and cloning the part of "the best" antibodies, that are characterized by the least affine value *Aff* the hypermutation of the antibodies clones; self-destruction of the antibodies clones "similar" to the other clones and antibodies of the current population; calculating the affinity of the antibodies clones and forming the new antibodies population; suppression of the population received; generation of the new antibodies and adding them to the current population until the ingoing size; the conditional test of the MCSA completion.

The forecasting model on the base of the SBT can be applied for the individual and groups' forecasting of TSs [3 – 6]. In the second case this model is used as the general forecasting model for describing the clusters' centroids. Herewith, the general forecasting model can be specified for some individual TS during the additional iterations of the MCSA.

## 3. Attractors

In the mathematical field of dynamical systems, an attractor is a set of numerical values toward which a system tends to evolve, for a wide variety of starting conditions of the system. System values that get close enough to the attractor values remain close even if slightly disturbed [7]. Attractor can be used for the solution of many complex applied problems, including the forecasting problems [8 – 11].

In applied mathematics, the phase space method is a technique for constructing and analyzing solutions of dynamical systems, that is, solving time-dependent differential equations. The method consists of first rewriting the equations as a system of differential equations that are first-order in time, by introducing additional variables. The original and the new variables form a vector in the phase space. The solution then becomes a curve in the phase space, parameterized by time. The differential equation is reformulated as a geometrical description of the curve, that is, as a differential equation in terms of the phase space variables only, without the original time parameterization. Finally, a solution in the phase space is transformed back into the original setting [1].

One of approaches to creation of the phase curve suggests to pass from the differential equations to the difference ones and assumes that value $d^j$ ( $j = \overline{2,n}$ ) of element of the analyzed by TS is postponed on the abscissa axis in each concrete timepoint, and the corresponding chain pure gain:

$$\Delta d^j = d^j - d^{j-1} \ ( \ j = \overline{2,n} \ ) \tag{3}$$

is postponed on the ordinate axis. It is not a derivative in the classical understanding, it is its discrete analog.

The attractors found thus can be used to form reasonably the training data sequence in the context of the forecasting problem. In this case, it is necessary to find all attractors and analyze the last attractor, if there are several attractors.

If attractor is single, it is necessary to consider it. Further, it is necessary to estimate the length of such attractor, i.e. the number of elements forming it.

If the attractor length corresponds to the condition imposed at the length of the training data sequence for the forecasting model on the base of the SBT, then such attractor can be recommended for the use.

If the attractor length is significantly less, than required, then it is necessary to refuse the use of the forecasting model on the base of the SBT. Also, it is possible to unite two or more attractors in one general training data sequence with the acceptable length. If the attractor length is significantly more, than required, then it is necessary to be ready to the possible side effects in case of its use as the training data sequence for the forecasting model on the base of the SBT. Besides, the attempt accomplishment of the reasonable reduction of length of the training data sequence created on the base of the attractor is possible. Also, the reasonable step is search of the alternative approaches to creation of the forecasting model.

In case of the individual forecasting it is necessary to use the attractor-based approach to the choice of the training data sequence for the forecasting model on the base of the SBT.

In case of the group's forecasting it is necessary, firstly, to clusterize the TS group by means of some clustering alrorithm (for example, by means of fuzzy *c*-means algorithm (FCM-algorithm) [3, 4] or *k*-means algorithm [5]) and, secondly, to use the attractor-based approach to the choice of the training data sequence for the forecasting model on the base of the SBT for the general forecasting model corresponding to the the cluster centroid of each formed cluster.

Also, it is possible to analyze all TS in the cluster to find attractors in them, to find lengths of all attractors in the cluster, to reveal the most often found attractors' lengths in the cluster and to use their maximum value (or their average value rounded to the next bigger integer) for determination of the length of the training data sequence. It is obviously, that the training data sequences for the general forecasting models on the base of the SBT in each cluster can have the different length.

## 4. Experimental studies

The experimental studies executed with the use of TSs, describing various socio-economic indexes, confirm the expediency and prospects of the attractor-based approach to creation of the training data sequence for the forecasting models on the base of the SBT.

The offered approach to creation of the training data sequence was applied for individual and groups' forecasting of TSs, describing the number of references of the E-Commerce systems in the requirements to vacancies posted on the websites of 2 famous recruiting network services – HeadHunter.ru (Russia) and Indeed.com (USA) [12].

In particular, the TS "Magenta (USA)" was considered (figure 2, a). This TS contains information on the number of vacancies which include a specific keyword "Magenta". This keyword defines the name of E-Commerce system for development of online stores. The obtained forecasting results can be used for the analysis of tendencies of the labour market.

The TS "Magenta (USA)" contains 64 elements (the supervision period: 03.02.2016 – 06.04.2016, a unit of measure – the number of references). Herewith, the values of 5 last elements must be predicted, and the others can be used for development of the forecasting model. It is possible to allocate two attractors on the phase curve (figure 2, b). The values by dates are designated by the sequence numbers of the TS elements.

The first attractor is determined during the period from 04.02.2016 to 14.03.2016. As the transition period it is possible to consider the period from 15.03.2016 to 16.03.2016, during which the transition to the new attractor has been executed. From 17.03.2016 to 06.04.2016 the second attractor is determined (from the 43-th value to the 64-th value in figure 2, b).

Let the TS values in the last 5 counting of time (from 02.04.2016 to 06.04.2016, that is from the 60-th value to the 64-th value in figure 2, b) are unknown and it is necessary to predict them.

In this case the length of the training data sequence received on the base of the second attractor with the length of 22 elements will be equal to 17 elements. Such length of the training data sequence can be considered as acceptable for development of the forecasting model on the base of the SBT.

In this example it is visible practically directly what data should be used as the training data sequence (though usually it isn't so obviously). Nevertheless, for confirmation of correctness of the decision on the choice of the training data sequence the attractor-based approach has been used.

Firstly, two forecasting models on the base of the SBT and the one-objective MCSA have been developed: the first model on the base of the elements' values of the initial TS with the length of 59 elements and the second model on the base of the elements' values of the second attractor with the length of 17 elements. Then, the same forecasting models on the base of the SBT and the multiobjective MCSA have been developed.

In case of the one-objective MCSA the average time of development of the required forecasting model in the first and second cases have constituted 1342.924 seconds and 1094.638 seconds respectively. Hence, the average time expenditures have been reduced more, than by 1.23 times. The average value of the affinity indicator in the first and second cases have constituted 2.619% and

1.893% respectively. The average value of the forecasting error on 5 steps forward in the first and second cases have constituted 2.136% and 1.043% respectively.

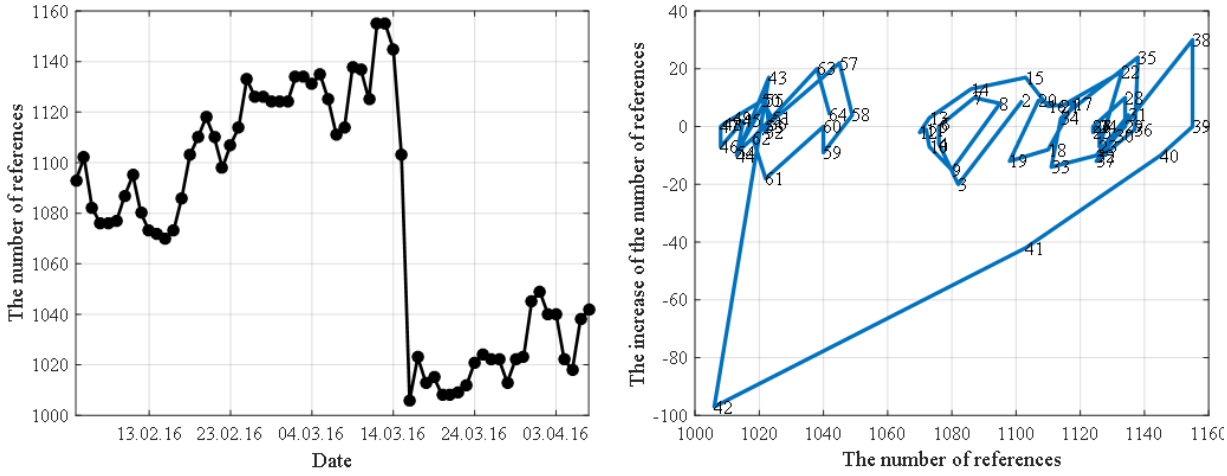
In case of the multiobjective MCSA the average time of development of the required forecasting model in the first and second cases have constituted 1377.805 seconds and 1107,07 seconds respectively. Hence, the average time expenditures have been reduced more, than by 1.24 times.

The average value of the affinity indicator the tendencies discrepancy indicator *Aff* in the first and second cases have constituted 2.411% and 1.676% respectively. The average value of the forecasting error on 5 steps forward in the first and second cases have constituted 2.007% and 0.988% respectively.

It is visible, that the one-objective MCSA works a little faster (as expected) than the multiobjective MCSA, but the last algorithm provides smaller values of errors of all types thanks to the corresponding correction (by means of the accounting of the second quality indicator – the tendencies discrepancy indicator *Tendency*) of the search direction of the forecasting model. Herewith, the use of attractor allows mimimizing the values of errors of all types, because the elements of TS, which lost the relevance were excluded from the consideration during the process of creation of the forecasting model. Also, the use of attractor allows mimimizing the time expenditures for the development of the forecasting model. For both MCSAs the execution of one run was considered as successful if the value of the affinity indicator *Aff* did not exceed 5%. It has as a result turned out that the share of the successful runs when using the elements' values of the initial TS as the training data sequence is equal to 32%, and the share of the successful runs when using the elements' values of the second attractor as the training data sequence is equal to 54%.

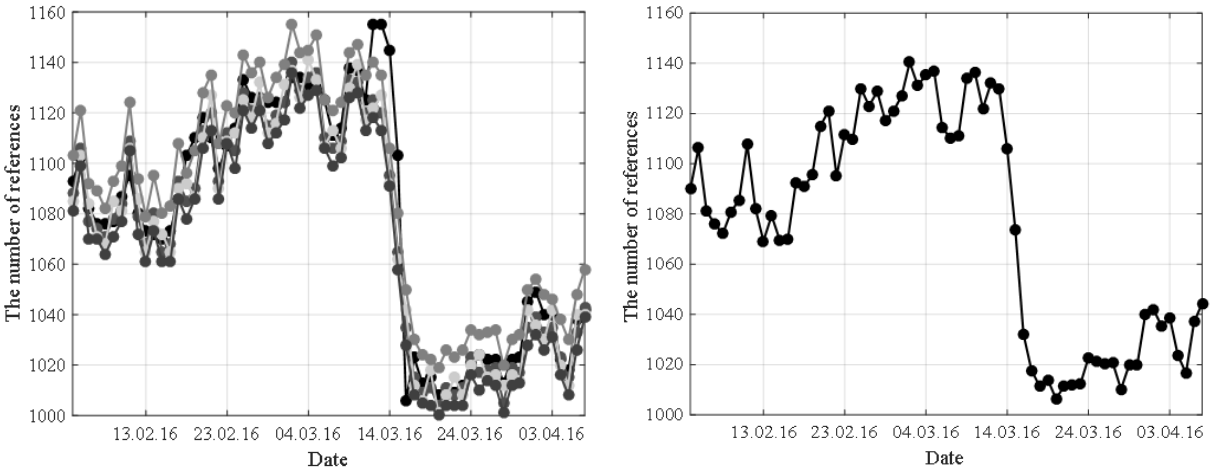
In the reviewed example 1000 iterations of the MCSA for population of 20 antibodies were executed. Coefficient of antibodies' cloning was equal to 0.3. Coefficient of clones' reproduction was equal to 0.8. Computer working under the 64-bit Windows 7 version with RAM of 2 Gb and the two-nuclear Pentium 4 processor with a clock frequency of 3.4 GHz was used for experiment.

Also, the attractor-based approach was applied for groups' forecasting on the base of the model data, which were generated with the use of the data about the references of the E-Commerce systems in the requirements to vacancies. Figure 3, a shows one of clusters, received during the clusterization of the group of TSs. Figure 3, b presents the TS "Cluster's centriod" for this cluster. Figure 3, c shows the phase curve of the TS "Cluster's centriod". It is possible to allocate two attractors on the phase curve (figure 3, c). As in case of the individual forecasting, it is expediently to use the second attractor, which is determined from the 43-th value to the 64-th value in figure 3, c.



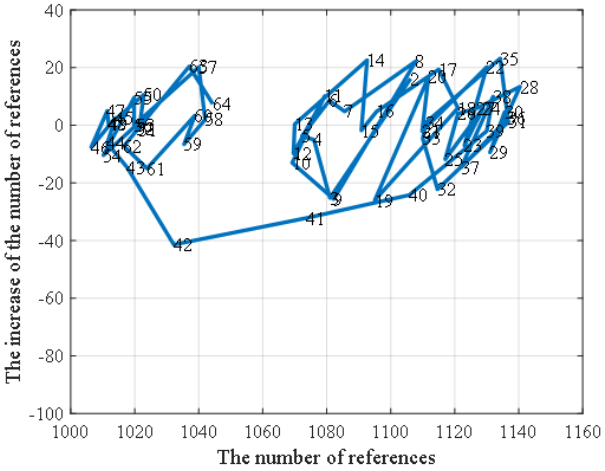

a – The TS "Magenta (USA)" b – The phase curve of the TS "Magenta (USA)"

**Figure 2.** Identification of attractors.




a – The cluster of TSs

b – The TS "Cluster's centriod"




c – The phase curve of the TS "Cluster's centriod"

**Figure 3.** Identification of attractors for the cluster of TSs.

The rest data should be excluded from the analysis. It is visible (figure 3), that data from the first attractor have the absolutely other law of change. Therefore, their accounting would lead to forming the analytical dependences, not absolutely urgent for the present moment of time. Besides, it will increase the time expenditures for creation of the forecasting model on the base of the SBT. It is necessary to say, that the general forecasting model, which describes the cluster's centroid can be specified for the individual TS, if it is necessary.

The obtained results can be explained by the fact that the length of the training data sequence has been reduced, and the elements with the little actuality for the present moment of time have been excluded from the consideration by the use of the mathematically reasonable approach.

## 5. Conclusions

The experimental results confirm the expediency and prospects of the attractor-based approach to the choice of the training data sequence for the forecasting model on the base of the SBT, which can be applied for individual and groups' forecasting.

In case of the individual forecasting it is neccessary to find the attractor in the TS and use it to form the training data sequence. It allows reducing the time expenditures on the creation of the forecasting models and the obtaining of the forecasting results. In case of the group's forecasting it is necessary to apply the attractor-based approach to clusters which are formed on the base of the TS group. Herewith, the attractors for the different clusters' centroids can have the different length. As a result, it can lead

to the additional reduction of the time expenditures on the creation of the forecasting models and the obtaining of the forecasting results. Herewith, it is possible to minimize at the same time the value of the affinity indicator of the forecasting model on the base of the SBT, the values of forecasting errors on 1 – 5 steps forward, and also the time expenditures on the creation of the forecasting models.

Further, it is planned to use the attractor-based approach for forecasting as the values of TS elements as the values' increments of TS elements with aim to create a technique for individual and groups' forecasting using the forecasting models on the base of the SBT and the MMCSA.

**Acknowledgments**
The reported study was funded by RFBR according to the research project № 16-08-00771.

**References**

[1] Demidova L A 2014 Time series forecasting models on the base of modified clonal selection algorithm *2014 International conference on computer technologies in physical and engineering applications* 33–34.
[2] Demidova L A 2014 Assessment of the quality prediction models based of the strict on binary trees and the modified clonal selection algorithm *Cloud of Science* pp 202–222
[3] Astakhova N N, Demidova L A and Nikulchev E V 2015 Forecasting Of Time Series' Groups With Application Of Fuzzy C-Mean Algorithm *Contemporary Engineering Sciences* **8** (**35)** pp 1659–1677.
[4] Astakhova N N, Demidova L A and Nikulchev E V 2015 Forecasting Method For Grouped Time Series With The Use Of K-Means Algorithm *Applied Mathematical Sciences* **9 (97)** pp 4813–4830.
[5] Astakhova N, Demidova L and Konev V 2015 The Description Problem Of The Clusters' Centroids 2015 *International Conference "Stability and Control Processes" in Memory of V.I. Zubov (SCP)* pp 448–451.
[6] Astakhova N and Demidova L 2016 Using of the notion "Pareto set" for development of the forecasting models based on the modified clonal selection algorithm *6th Seminar on Industrial Control Systems: analysis, modeling and computation* art 02001.
[7] Kolmogorov A, Petrovskii I and Piscounov N 1937 A study of the diffusion equation with increase in the amount of substance, and its application to a biological problem *In V.M. Tikhomirov editor, Selected Works of A.N. Kolmogorov I* pp 248-270.
[8] Palmer T N 2003 Predicting uncertainty in forecasts of weather and climate *Meteorological Training Course Lecture Series, ECMWF* pp 1-48.
[9] Marcelo Espinoza, Suykens Johan A K and Bart De Moor 2005 Short Term Chaotic Time Series Prediction using Symmetric LS-SVM Regression *2005 International Symposium on Nonlinear Theory and its Applications (NOLTA2005) (Belgium: Bruges)* pp 606-609.
[10] Nikulchev E 2014 Robust Chaos Generation on the Basis of Symmetry Violations in Attractors *2nd International Conference on Emission Electronics (ICEE)* pp 59-61.
[11] Nikulchev E V, Kondratov A P 2015 Method of Generation of Chaos Map in the Centre Manifold *Advanced Studies in Theoretical Physics* **9** (**16)** pp 787-792.
[12] Nikulchev E, Ilin D, Biryukov D and Bubnov G 2016 Monitoring of Information Space for Professional Skills *Demand Contemporary Engineering Sciences* **9** (**14)** pp 671-678.
[13] Deb K, Pratap A, Agarwal S, and Meyarivan T 2002 A Fast and Elitist Multiobjective Genetic Algorithm: NSGA II *IEEE Transactions on Evolutionary Computation* **6** (**2)** pp 182–197
[14] Deb K 2001 Multi-objective Optimization using Evolutionary Algorithms *Chichester* pp 221-232.
[15] Seada H, and Deb K 2015 U-NSGA-III: A Unified Evolutionary Optimization Procedure for Single, Multiple, and Many Objectives: Proof-of-Principle Results Evolutionary Multi-Criterion Optimization *Lecture Notes in Computer Science* **9019** pp 34-49.