ARSIMA MODEL

Vitalii Shchelkalin

***Abstract**: In presented work the further development of Box-Jenkins technique for models constructing and improvement of themselves ARIMA models is produced. A novel autoregressive – spectral integrated moving average (ARSIMA) model founded on joint use of the Box-Jenkins method (ARIMA models) and "Caterpillar"-SSA method with model trained on competitive base is developed.*

***Keywords**: modeling, filtering, forecasting, control, "Caterpillar»-SSA method, ARIMA model, "Caterpillar»-SSA – ARIMA – SIGARCH method, ARSIMA model, ARSIMA – SIGARCH model, heteroskedasticity, Levenberg-Marquardt method.*

Introduction

The research progress constantly develops economic, technical, social, medical and other systems, complicating their structure and enlarging amount complex their internal intercoupling and external factor, from which they hang and majority from which take into account impossible. Therefore, it is actual to develop and use the universal mathematical models and techniques allowing to modeling, forecast and control the wide class of the processes since this will allow to raise efficiency of control and planning state of working complex systems, spare the significant facilities and resources not by development of new energy-saving resources, but by way of any mathematical subterfuges, as well as offloads the work of the personnel of various organizations and allows to anticipate emergencies and brings many other benefits.

The main requirements to the mathematical models construction in 70 - 90 of the last century, is an economical number of parameters, the velocity of the model determination and its resource-intensive for use on available then computers with low productivity. However, modern computer technology and mathematical modeling methods provide a great possibilities for analysis, modeling and forecasting time series of the different nature. Therefore, at present these requirements are not crucial and modern computing tools and systems allow to stand on the first plan the requirement of modeling accuracy, quality of the analysis and forecasting.

One of the most widely used models that corresponding to above requirements are the models ARIMA. ARMA method works only with pre-reduced to the stationary form time series. Nonstationary series are usually characterized by the presence of high power at low frequencies. However, in many practical applications of interest information may be concentrated at high frequencies. In such cases, all that was done - it is filtered out non-stationary low-frequency components and was used the remainder of the series for further analysis. At the same time as a filter to eliminate low-frequency component in the ARIMA model used a filter of the first differences or maximum second. Watching the gain of the filter can be seen that low frequency considerably weakened and, therefore, be less visible at the filter output. So the method of seasonal ARIMA model was satisfactorily predicted only with a relatively simple structure time series.

In the 80's years of last century Granger and Djoyo [Granger 1980] proposed a new class of ARFIMA models is convenient to describe the financial and economic time series with the effects of long and short memory.

2000's years are characterized by the using for a wide range of models for time series analyzing and forecasting, as well as ensembles of models with different structures. With the advent of high-speed computer was occurred the transition from ensembles of predictive models to its combination. The difference between the combined models and its ensembles lies in the simultaneous adjustment of model parameters.

The most important characteristics of models at analysis and choice of the most appropriate mathematical models of the following main important features are:

- method of modeling the trend component of the time series;
- method of nonlinear modeling of time series;
- method of modeling the random component of the time series;
- way of accounting for the influence of external factors on the process.

Therefore, the priority type of models are combined probability and deterministic prediction models with nonlinear complexity, because In this models simultaneously used as statistical and deterministic components that allows to reach the best quality of the forecasting [Седов, 2010] . Among the deterministic models the priority variant is a deterministic model of the spectral decomposition, which implements simulation-based expansion in the deterministic orthonormal basis different from that of harmonic functions. While the most priority among the probabilistic models is the model of auto regression - integrated moving average.

In the scientific literature have long been known combined probability and deterministic model presented in [Седов, 2010] . In given paper is offered next modification of the ARIMA and the GARCH models and at first proposed "Caterpillar"-SSA – ARIMA – SIGARCH method and combined probabilistic and deterministic model of auto regression – spectrally integrated moving average with spectrally integrated generalized autoregressive heteroskedasticity (ARSPSS – SIGARCH) and the method of its construction, which allows greater flexibility in analyzing, modeling and forecasting time series in comparison with the ARIMA – GARCH models.

Summary of the main material of the study

The essence of the method was at first in a multi-dimensional decomposition of exogenous time series, and the propagated time series for basic latent components, including their extents and combinations, obtained by the principal (PCA), smooth (SCA) and independent components (ICA), in the selection of the basic components of design method of fast orthogonal search (FOS), one of the most effective and economical methods for time spent, and cutting off the destructive, thus forming the transfer function of a mathematical model of the process, to further identify the noise of a mathematical model of the process due to the seasonal autoregressive models –- moving average, and the simultaneous parameter identification of model structure obtained by the Levenberg-Marquardt algorithm [Щелкалин, Тевяшев, 2010]. As it became known later, the proposed method is similar to the decomposition method of modeling (DMM) [Седов, 2010] . The method of "Caterpillar"-SSA also uses the decomposition of time series of singular values (SVD). Known publications using the "Caterpillar"-SSA method in various branches of science and technology as a method of fairly good description of non-stationary time series with linear, parabolic or exponential trend with not always stable oscillatory component, however studies have identified a number of significant shortcomings of the method, greatly limiting its applicability. Method for modeling uses suboptimal in terms of accuracy of some time series of orthogonal basis vectors of the trajectory matrix. Therefore, the main idea of the origin of the proposed method was first concluded in joint use the "Caterpillar"-SSA method and models of autoregressive - moving average, trained on a competitive base, with account generalized criterion of the accuracy and adequacy. Using such combination was dictated by the fact that individually, these approaches have several disadvantages, but their joint use brings synergy, increasing their efficiency, robustness and adequacy. However, the trend separation by "Caterpillar"-SSA method, as well as any other method, the residual component of the series in most cases is non-stationary, and therefore hereinafter "Caterpillar"-SSA method has been used in combination with the model of autoregressive - integrated moving average (ARIMA). In this case, joint use of the above methods imply that the parametric identification computes the parameter estimates ARIMA and nonlinear generalizations of principal components of the autoregression,

minimizing the sum of the squares error modeling, taking into account the time series, obtained by the "Caterpillar"-SSA, which during modeling does not have the model and parameters, respectively, and Levenberg-Marquardt method, calculating the parameters of the remaining parts of the additive model helps to define the exterior of a time series of "Caterpillar"-SSA method, acting as assistant to the definition of a deterministic (trend) component of the process.

The proposed approach is a variant of the priority to date combined probability and deterministic approaches, because herewith are simultaneously used as statistical and deterministic components that allows to reach the best quality of the forecasting. It also implemented the so-called trend approach, where the process is modeled as the deviation of actual values from the trend (which is presented here as time series obtained by the "Caterpillar"-SSA method), and which ensures the stability of the model and obtained the required accuracy of modeling, whereas previously the probabilistic model ARIMA tried to describe the entire process. Thus, a successful attempt made after more than thirty years after the establishment of the Box-Jenkins method and the "Caterpillar"-SSA method to combine them. However, for satisfaction of such requirements to models, as: learning rate, labor content, resource use, presentation models, ease of use and interpretability, time- and resource-consuming method "Caterpillar"-SSA was later offered to be used only for preliminary structural identification and rough parametric identification of the so-called integrating a polynomial of the operator of the delay L of proposed model as well as for a rough structural and parametric identification of a polynomial of the delay, whose presence is distinguishes more general polynomial model from the Box-Jenkins model, structure and whose coefficients are equal to those of recurrence prediction model of the "Caterpillar"-SSA method.

The "Caterpillar"-SSA method is also offered to use for preliminary generalized co-integration of multiply time series modeling processes, as well as for separation for a finite and separately for deadbeat regulators in the case of use the proposed model in control theory [Щелкалин, Тевяшев, 2011]. It is also possible nonlinear complication of the model transfer function of one of the ways: FOS, GMDH, RBF, LARS, built on the principal component of their degrees and combinations. So, first of all, the author proposed a model aimed at the automatic control theory, modeling and forecasting of technical systems and technological processes, due to the fact that their transfer functions are more determined and have a complex nonlinear structure.

Description of the proposed mathematical models

A mathematical model of the processes that depend from several exogenous factors in the operator form can be presented as a model of the seasonal autoregressive - integrated moving average (SARIMA) [Евдокимов, Тевяшев, 1980] :

$$z_t^y = \sum_{i=1}^{N} \frac{b^i_{n_{b^i}}(L)}{a^i_{n_{a^i}}(L)} \cdot z^{x^i}_{t-m_i} + \frac{c^{\Pi}_{n_c^{\Sigma}}(L)}{d^{\nabla}_{n_d^{\nabla}}(L)} \cdot e_t \,, \tag{1}$$

where L – the shift operator in time by one unit back, so that $L^i x_t = x_{t-i}$, N – the number of exogenous variables; z_t^y – normalized from 0 to 1 according to the formula $z_t^y = \frac{y_t - y_t^{\min}}{y_t^{\max} - y_t^{\min}}$ or some other way a time series y_t of simulated and predicted process subtracted the average value; $z^{x^i}_{t-m_i}$ – normalized in the same way the i-th exogenous time series x_t^i subtracted from the average; m_i – the delay of the i-th exogenous time series x_t^i in time relative to the forecasted time series y_t; $a^i_{n_{a^i}}(L)$, $b^i_{n_{b^i}}(L)$ – polynomials from L of n_{a^i} and

n_{b^i} degrees respectively; $c^{\Pi}_{n_c^{\Sigma}}(L)=c^1_{n_c^1}(L^{s_1})\cdot c^2_{n_c^2}(L^{s_2})\times\ldots\times c^{n_s}_{n_c^{n_s}}(L^{s_{n_s}})=\prod_{i=1}^{n_s} c^i_{n_c^i}(L^{s_i})$ – polynomial from L^{s_i} of n_c^i degree, defining component of the moving average of the periodic component with a periods s_i, $n_c^{\Sigma}=\sum_{i=1}^{n_s} n_c^i\cdot s_i$;

$$d^{\nabla}_{n_d^{\nabla}}(L)=d^{\Pi}_{n_d^{\Sigma}}(L)\nabla^{D_1}_{s_1}\nabla^{D_2}_{s_2}\ldots\nabla^{D_{n_s}}_{s_{n_s}}=$$

$$=d^1_{n_d^1}(L^{s_1})\cdot d^2_{n_d^2}(L^{s_2})\cdot\ldots\cdot d^{n_s}_{n_d^{n_s}}(L^{s_{n_s}})\nabla^{D_1}_{s_1}\nabla^{D_2}_{s_2}\ldots\nabla^{D_{n_s}}_{s_{n_s}}=\prod_{i=1}^{n_s} d^i_{n_d^i}(L^{s_i})\nabla^{D_1}_{s_1}\nabla^{D_2}_{s_2}\ldots\nabla^{D_{n_s}}_{s_{n_s}},$$

$n_d^{\nabla}=n_d^{\Sigma}+\sum_{i=1}^{n} D_i\cdot s_i$, $d^{\Pi}_{n_d^{\Sigma}}(L)=d^1_{n_d^1}(L^{s_1})\cdot d^2_{n_d^2}(L^{s_2})\times\ldots\times d^{n_s}_{n_d^{n_s}}(L^{s_{n_s}})=\prod_{i=1}^{n_s} d^i_{n_d^i}(L^{s_i})$ – polynomial from L^{s_i} of n_d^i degree, defining component of the autoregressive seasonal component with a period s_i, $n_d^{\Sigma}=\sum_{i=1}^{n_s} n_d^i\cdot s_i$; e_t – residual errors model: D_i – the procedure of taking the differences s_i; ∇_{s_i} and L^{s_i} – simplify the operators such that $\nabla_{s_i} y_t=\left(1-L^{s_i}\right)\cdot y_t=y_t-y_{t-s_i}$.

Equation (1) in more compact form can be represented as:

$$\tilde{a}(q)\cdot z_t^y=\sum_{i=1}^{k}\tilde{b}^i(q)\cdot z^{x^i}_{t-m_i}+\tilde{c}(q)\cdot e_t$$

where $\tilde{a}(L)=d^{\nabla}_{n_d^{\nabla}}(L)\cdot\prod_{i=1}^{N} a^i_{n_{a^i}}(L)$, $\tilde{b}^i(L)=b^i_{n_{b^i}}(L)\cdot d^{\nabla}_{n_d^{\nabla}}(L)\cdot\prod_{j=1,i\neq j}^{N} a^j_{n_{a^j}}(L)$,

$\tilde{c}(q)=c^{\Pi}_{n_c^{\Sigma}}(q)\cdot\prod_{i=1}^{N} a^i_{n_{a^i}}(q)$, $i=\overline{1,N}$.

The expression for the prediction of pre-emption l using the proposed model of the joint use to models of the "Caterpillar"-SSA method and seasonal autoregressive model - integrated moving average with exogenous variables, after adduction it from rational form to the difference equation takes the type:

$$\hat{y}_t(l)=\hat{y}_t^{SSA}(l)+\sum_{j=1}^{n_d^{\Sigma}+\sum_{i=1}^{N} n_{a^i}}\tilde{a}_j\cdot h^y_{t+l-j}+$$

$$+\sum_{i=1}^{N}\left(\tilde{b}^i_0\cdot x^i_{t+l-m_i}-\sum_{j=1}^{n_d^{\Sigma}+n_{b^i}+\sum_{p=1}^{N} n_{a^p}}\tilde{b}^i_j\cdot x^i_{t+l-m_i-j}\right)-\sum_{j=1}^{n_c^{\Sigma}+\sum_{i=1}^{N} n_{a^i}}\tilde{c}_j\cdot e_{t+l-j};\qquad(2)$$

$$\hat{x}^i_t(l)=\hat{x}^k{}_t^{SSA}(l)+\sum_{j=1}^{n^{x^k}{}_d^{\Sigma}}\tilde{a}_{x^k{}_j}\cdot z^{x^k}_{t+l-j}+\sum_{j=1}^{n^{x^k}{}_c^{\Sigma}}\tilde{c}_{x^k{}_j}\cdot e_{x^k{}_{t+l-j}},k=\overline{1,N},$$

where $y_{t+j} = \begin{cases} y_{t+j}, j \le 0, \\ \hat{y}_t(j), j > 0, \end{cases}$ $x_{t+j}^i = \begin{cases} x_{t+j}^i, j \le 0, \\ \hat{x}_t^i(j), j > 0, \end{cases}$ $i = \overline{1,N}$; $e_{t+j} = \begin{cases} e_{t+j}, j \le 0, \\ 0, j > 0. \end{cases}$; $h_i^y = y_i - \widetilde{w}_i^{N+1}$;

$\hat{y}_t^{SSA}(i) = \sum_{j=1}^{L-1} f_j^y \cdot \widetilde{w}^y{}_{t+i-j}^{N+1}$, $i = \overline{1,l}$; $\widetilde{w}^y{}_i^{N+1} = \begin{cases} \hat{y}_i, i > t, \\ \widetilde{w}^y{}_i^{N+1}, i \le t; \end{cases}$

$(f_{L-1}^y, f_{L-2}^y, ..., f_1^y)^T = \frac{1}{1 - \sum_{i=1}^{R} (u_L^i)^2} \sum_{i=1}^{R} u_L^i \cdot (u_1^i \quad u_2^i \quad \dots \quad u_{L-1}^i)^T$, where $(u_1^i \quad u_2^i \quad \dots \quad u_{L-1}^i)^T$ – a vector consisting of the first $(L-1)$ elements of the singular value decomposition eigenvector U^i of the trajectory matrix of exogenous and simulated processes $X = (X_1 \quad X_2 \quad \dots \quad X_N \quad Y) = (X_{1,1} \quad X_{1,2} \quad \dots \quad X_{1,K} \quad X_{2,1} \quad X_{2,2} \quad \dots$

$\dots \quad X_{2,K} \quad \dots \quad X_{N,1} \quad X_{N,2} \quad \dots \quad X_{N,K} \quad Y_1 \quad Y_2 \quad \dots \quad Y_K)$;

$Y_j = (y_{j-1} \quad y_{j-2} \quad \dots \quad y_{j+L-2})^T$, $X_{i,j} = (x_{j-1}^i \quad x_{j-2}^i \quad \dots \quad x_{j+L-2}^i)$, $i = \overline{1,N}$, $j = \overline{1,K}$ – transformation of the $(N+1)$-dimensional time series (N – number of exogenous time series and one predictable) in the sequence L^y-dimensional vectors (L^y – width of the window), whose number is equal $(N+1) \cdot K$, $K = n - L^y + 1$, where n – the length of time series; u_L^i – the last element of the vector U^i; R – the number of elements of the singular value decomposition; $\hat{y}_t^{SSA}(i) = \sum_{j=1}^{L-1} f_j^y \cdot \widetilde{w}^y{}_{t+i-j}^{N+1}$ – model of recurrent prediction "Caterpillar"-SSA method of time y_t, $t = \overline{0, n-1}$; which in turn can often be economically recorded by seasonal ARIMA (ARI) or by Almon model of distributed lags, where $\widetilde{w}^y{}_0^{N+1}$, $\widetilde{w}^y{}_1^{N+1}$, ..., $\widetilde{w}^y{}_{n-1}^{N+1}$ – a time serie corresponding to the transformation of the predicted time serie y_t; time series $\widetilde{w}^{x^i}{}_0$, $\widetilde{w}^{x^i}{}_1$, ..., $\widetilde{w}^{x^i}{}_{n-1}$, $i = \overline{1,N}$ correspond to the transformations i-th time series of the exogenous time series using singular spectrum analysis at the stage of diagonal averaging which takes the matrix $\widetilde{Z}^i$, $i = \overline{1,N+1}$ consisting of K columns from $(i-1) \cdot K$-th to $i \cdot K - 1$-th matrix Z to series $\widetilde{w}^{x^i}{}_0$, $\widetilde{w}^{x^i}{}_1$, ..., $\widetilde{w}^{x^i}{}_{n-1}$, according to the formula

$$\widetilde{w}^y{}_i^k = \begin{cases} \frac{1}{i+1} \sum_{j=1}^{k+1} \widetilde{z}_{j,i-j+2}^i, i = \overline{0, \min(L,K)-2}; \\ \frac{1}{\min(L,K)} \sum_{j=1}^{\min(L,K)} \widetilde{z}_{j,i-j+2}^i, i = \overline{\min(L,K)-1, \max(L,K)-1}; \\ \frac{1}{n-i} \sum_{j=k-\max(L,K)+2}^{n-\max(L,K)+1} \widetilde{z}_{j,i-j+2}^i, i = \overline{\max(L,K), n-1}, \end{cases}$$

$k = \overline{1,N+1}$; where $Z = \widetilde{Z}^i + \dots + \widetilde{Z}^j$ – the amount of decomposition matrices $\widetilde{Z}^i = \left(U^i \cdot (U^i)^T \cdot X_1 \quad U^i \cdot (U^i)^T \cdot X_2 \quad \dots \quad U^i \cdot (U^i)^T \cdot X_N \quad U^i \cdot (U^i)^T \cdot Y \right)$, selected by standard analysis of eigenvalues of the trajectory matrix in the "Caterpillar"-SSA method. There is also a modification of the

recursive method of SSA-forecasting – vector SSA-prediction [Голяндина, 2004], which in some cases provides more accurate forecasts. $\hat{x}^k{}_t^{SSA}(i)=\sum_{j=1}^{L^y-1} f_j^{x^k}\cdot\widetilde{w}^{x^k}{}_{t+i-j}^{N+1}$, $k=\overline{1,N}$, same thing, only for k-th exogenous time series, as well as all the variables and parameters with an index x^i, $i=\overline{1,N}$ similarly interpreted as for the time series y_t, $t=\overline{1,n}$;

With the nonlinear complexity of the transfer function the first equation (2) becomes:

$$\hat{z}_t^y(l)=\hat{y}_t^{SSA}(l)+\sum_{i=1}^{r} g_i\cdot p_t^i+\sum_{j=1}^{n_d^{\Sigma}+\sum_{i=1}^{N} n_{a^i}}\widetilde{a}_j\cdot h_{t+l-j}^y+$$

$$+\sum_{i=1}^{N}\left(\widetilde{b}_0^{\,i}\cdot z_{t+l-m_i}^{x^i}-\sum_{j=1}^{n_d^{\Sigma}+n_{b^i}+\sum_{p=1}^{N} n_{a^p}}\widetilde{b}_j^{\,i}\cdot z_{t+l-m_i-j}^{x^i}\right)-\sum_{j=1}^{n_c^{\Sigma}+\sum_{i=1}^{N} n_{a^i}}\widetilde{c}_j\cdot e_{t+l-j},$$

$$\hat{x}_t^i(l)=\hat{x}^k{}_t^{SSA}(l)+\sum_{j=1}^{n_d^{x^k\Sigma}}\widetilde{a}_{x^k j}\cdot z_{t+l-j}^{x^k}+\sum_{j=1}^{n_c^{x^k\Sigma}}\widetilde{c}_{x^k j}\cdot e_{x^k\,t+l-j},k=\overline{1,N},$$

where $g_i\cdot p_t^i$ – the members of the Kolmogorov-Gabor polynomial:

$$\sum_{i=1}^{r} g_i\cdot p_i^t=FOS\left(\sum_{i=1}^{M}\widetilde{g}_i\cdot\widetilde{x}_t^i+\sum_{i=1}^{M}\sum_{j=1}^{M}\widetilde{g}_{ij}\cdot\widetilde{x}_t^i\cdot\widetilde{x}_t^j+\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M}\widetilde{g}_{ijk}\cdot\widetilde{x}_t^i\cdot\widetilde{x}_t^j\cdot\widetilde{x}_t^k+\ldots\right),$$

$g_n=\widetilde{g}_{ij\ldots k}, n=\overline{1,r}$ or $\sum_{i=1}^{r} g_i\cdot p_i^t=FOS\left(\sum_{i=1}^{M_2}\widetilde{g}_i\cdot\varphi_i\left(\widetilde{\widetilde{x}}_t\right)\right)$, where

$\varphi_i\left(\widetilde{\widetilde{x}}_t\right)=\frac{1}{(2\pi)^{-\frac{M}{2}}\cdot|\Sigma_i|^{-\frac{1}{2}}}\cdot e^{-\frac{1}{2}\left(\widetilde{\widetilde{x}}_t-\overline{c}_i\right)\Sigma_i^{-1}\cdot\left(\widetilde{\widetilde{x}}_t-\overline{c}_i\right)^T}$, $\overline{c}_i$ – the vector of mathematical expectation of time series, representing the principal components, Σ_i – the covariance matrixes, $i=\overline{1,M}$; or $\sum_{i=1}^{r} g_i\cdot p_i^t=GMDH\left(\widetilde{x}_t^1,\widetilde{x}_t^2,\ldots,\widetilde{x}_t^M\right)$ – structural identification of GMDH method, $p_l^t=\widetilde{x}_t^i\cdot\widetilde{x}_t^j\cdot\ldots\cdot\widetilde{x}_t^k$, consisting of the principal components identified as $\widetilde{x}_t^i, i=\overline{1,M}$, and their degrees and combinations, , selected by FOS-algorithm, determining the nonlinear part of the proposed model; FOS – the function of the structural simplification of the model, written in her argument, given the nature of the behavior of time series with fast orthogonal search algorithm; $h_i^y=z_i^y-\widetilde{w}_i^{N+1}-\sum_{i=1}^{r} g_i\cdot p_t^i$.

The process of finding such combined models (2) of the joint use of ARIMA and of the "Caterpillar"-SSA method can be prolonged due to the resource-intensive of the "Caterpillar"-SSA . Therefore, the "Caterpillar"-SSA method analysis is proposed to use only for pre-structural identification and rough parametric identification of integrating polynomial $w(L)$ (hence the name of an autoregressive model - spectrally integrated moving average) of the delay operator L of model, which can also be interpreted as an transfer operator into the state space.

$$f(L)\cdot w(L)\cdot y_t=\frac{c(L)}{d(L)}\cdot e_t \text{ – ARSIMA model,} \qquad (3)$$

and the recursion SSA-forecasting method for gross-structural and parametric identification of a polynomial $f(L)$, whose structure and coefficients are equal to those of first recurrence prediction model of the "Caterpillar"-SSA method and the presence of which distinguishes the more general polynomial model from the Box-Jenkins model and in conjunction with $w(L)(\frac{w_1(L)}{w_2(L)}$ which may have, generally, a rational view), determining the long-term memory model, describing a more wide class of processes of long-term memory than fractional integration in the ARFIMA model, which in turn was invented to overcome the lack of ARIMA models for modeling and forecasting processes in a long memory - loss (distortion) of long-term information in the when taking the incrementations. Polynomials $d(L)$ and $c(L)$, in turn, determine the short-term depending of the process.

When analyzing and forecasting time series, depending on several other essential balance of the dynamic properties of the variables on the left-and right-hand sides of the equation of model. In this case the ideas of the "Caterpillar"-SSA method stand for pre-generalized cointegration of time series and model is divided as follows:

$$\hat{\tilde{w}}_t^y = \frac{b^y_{n_{b^y}}(q)}{a^y_{n_{a^y}}(q)} \cdot z_t^{\,y} + \sum_{i=1}^{N} \frac{b^{w^{y\,i}}_{n^{w^y}_{b^i}}(q)}{a^{w^{y\,i}}_{n^{w^y}_{a^i}}(q)} \cdot z^{x^i}_{t-m_i} + \frac{c^{w^{y\,\Pi}}_{n^{w^{y\,\Sigma}}_{c}}(q)}{d^{w^{y\,\Pi}}_{n^{w^{y\,\Sigma}}_{d}}(q)} \cdot e_t^{w^y};$$

$$z_t^y = f^{\,y}(q) \cdot \hat{\tilde{w}}_t^y + \sum_{i=1}^{N} \frac{b^i_{n_{b^i}}(q)}{a^i_{n_{a^i}}(q)} \cdot z^{x^j}_{t-m_i} + \frac{c^{\Pi}_{n^{\Sigma}_{c}}(q)}{d^{\Pi}_{n^{\Sigma}_{d}}(q)} \cdot e_t; \qquad (4)$$

$$f^{x_i}(q) \cdot \omega(q) \cdot z_t^{x^j} = \frac{c^{x^{j\,\Pi}}_{n_{x^j}{}^{\Sigma}_{c}}(q)}{d^{x^{j\,\Pi}}_{n_{x^j}{}^{\Sigma}_{d}}(q)} \cdot e_t^{x^j}, j = \overline{1,N},$$

$\hat{\tilde{w}}_i^y$ – approximation by time series $\hat{y}_t$ and by exogenous time series x^j_t, time series $\tilde{w}_t^y$ with seasonal ARIMAX model (or ARIX - integrated auto regression with exogenous variables), which was originally obtained by the "Caterpillar"-SSA method and, subsequently, adjustable by the optimization method with competitive learning of model; $\omega(L)$ – integrating polynomial that takes the time series x_t^j in time series $\hat{\tilde{w}}_t^{x^j}$ – an approximation of the time serie $\tilde{w}_{x^k}{}^{N+1}_{t}$ by ARIMA model; the initial rough values of the polynomials coefficients $f^{\,y}(L)$, $f^{x_i}(L)$ and their number can be taken equal to the coefficients f_j^y и $f_j^{x_i}$, $j = \overline{1,N}$ of models SSA-recursive prediction method $\hat{y}_t^{SSA}(i) = \sum_{j=1}^{L^y-1} f_j^y \cdot \tilde{w}^{y\,N+1}_{t+i-j}$ and $\hat{x}^k{}_t^{SSA}(i) = \sum_{j=1}^{L^{x^k}-1} f_j^{x^k} \cdot \tilde{w}^{x^k\,N+1}_{t+i-j}$ respectively;

L^y and L^{x_i} – appropriate lengths of windows, and then iteratively tune with the rest of the coefficients of model (4) using of Levenberg-Marquardt method. Model (4) benefits significantly by the time training a combined model of sharing seasonal ARIMAX model and the method of "Caterpillar"-SSA (2), but slightly inferior to it by the statistical properties with regard to the manner of its construction is called as seasonal autoregressive model - spectrally integrated moving average model with exogenous variables (ARSIMAX) and can be written as follows:

$$\hat{z}_t^y(l) = \sum_{j=1}^{L^y + n_{b^y} + \sum_{i=1}^{N} n^{w^y}_{a^i} + n^{w^y\,\nabla}_{d} + \sum_{i=1}^{N} n_{a^i} + n^{\nabla}_{d}} \tilde{\tilde{a}}_j \cdot z^y_{t+l-j} +$$

$$+\sum_{i=1}^{N}\left(\tilde{\tilde{b}}_0^{\,i}\cdot z_{t+l-m_i}^{x^i}-\sum_{j=1}^{\tilde{\tilde{n}}^{\tilde{\tilde{b}}}}\tilde{\tilde{b}}_j^{\,i}\cdot z_{t+l-m_i-j}^{x^i}\right)-\sum_{j=1}^{n_c^{\Sigma}+n_{a^y}+\sum_{i=1}^{N}n_{a^i}^{w^y}+n^{w^y}{}_d^{\nabla}+\sum_{i=1}^{N}n_{a^i}+n_d^{\nabla}}\tilde{\tilde{c}}_j\cdot e_{t+l-j}-$$

$$-\sum_{j=1}^{L^y+n^{w^y}{}_c^{\Sigma}+n_{a^y}+\sum_{i=1}^{N}n_{a^i}^{w^y}++\sum_{i=1}^{N}n_{a^i}+n_d^{\nabla}}\tilde{\tilde{d}}_j\cdot e^{w^y}{}_{t+l-j};$$

$$\hat{z}_t^{x^k}(l)=\sum_{j=1}^{L^{x^k}+n_{x^k}{}_d^{\nabla}}\tilde{\tilde{a}}_j^{x^k}\cdot z_{t+l-j}^{x^j}+\sum_{j=1}^{n^{x^k}{}_c^{\Sigma}}\tilde{\tilde{c}}_j^{x^k}\cdot e_{t+l-j}^{x^k},\ k=\overline{1,N},$$

where $\tilde{\tilde{a}}_j$, $j=\overline{1,L^y+n_{b^y}+\sum_{i=1}^{N}n_{a^i}^{w^y}+n^{w^y}{}_d^{\nabla}+\sum_{i=1}^{N}n_{a^i}+n_d^{\nabla}}$ – coefficients of the polynomial $\tilde{\tilde{a}}(L)=f^y(L)\cdot b_{n_{b^y}}^{y}(L)\cdot\prod_{i=1}^{N}a^{w^y}{}_{n_{a^i}^{w^y}}^{i}(L)\cdot d_{n_d^{\nabla}}^{\nabla}(L)$; $\tilde{\tilde{b}}_j^{\,i}$, $i=\overline{1,N}$, j = 1, 2, ..., $n^{\tilde{\tilde{b}}}$; $n^{\tilde{\tilde{b}}}$ =

$= L^y +$

$$+\max\left(n_{b^i}^{w^y}+\sum_{j=1,j\neq i}^{N}n_{a^j}^{w^y}+n_{a^y}+n^{w^y}{}_d^{\nabla}+\sum_{j=1,j\neq i}^{N}n_{a^j}+n_d^{\nabla},j=\overline{1,N},n_{b^i}+n_{a^y}+\sum_{j=1,j\neq i}^{N}n_{a^j}^{w^y}+n^{w^y}{}_d^{\nabla}+\sum_{j=1,j\neq i}^{N}n_{a^j}+n_d^{\nabla}\right)$$

– coefficients of the polynomial

$$\tilde{\tilde{b}}^{\,i}(L)=f^y(L)\cdot\sum_{i=1}^{N}\left(b^{w^y}{}_{n_{b^i}^{w^y}}^{i}(L)\cdot\prod_{j=1,j\neq i}^{N}a^{w^y}{}_{n_{a^j}^{w^y}}^{j}(L)\cdot a_{n_{a^y}}^{y}(L)\cdot d^{w^y}{}_{n^{w^y}{}_d^{\nabla}}^{\nabla}(L)\cdot\prod_{j=1,j\neq i}^{N}a_{n_{a^j}}^{j}(L)\cdot d_{n_d^{\nabla}}^{\nabla}(L)\right)-$$

$$-\sum_{i=1}^{N}\left(b_{n_{b^i}}^{i}(L)\cdot a_{n_{a^y}}^{y}(L)\cdot\prod_{j=1,j\neq i}^{N}a^{w^y}{}_{n_{a^j}^{w^y}}^{j}(L)\cdot d^{w^y}{}_{n^{w^y}{}_d^{\nabla}}^{\nabla}(L)\cdot\prod_{j=1,j\neq i}^{N}a_{n_{a^j}}^{j}(L)\cdot\cdot d^{w^y}{}_{n^{w^y}{}_d^{\nabla}}^{\nabla}(L)\right)$$

$\tilde{\tilde{c}}_j$, $j=\overline{1,n_c^{\Sigma}+n_{a^y}+\sum_{i=1}^{N}n_{a^i}^{w^y}+n^{w^y}{}_d^{\nabla}+\sum_{i=1}^{N}n_{a^i}+n_d^{\nabla}}$ – coefficients of the polynomial $\tilde{\tilde{c}}(L)=c_{n_c^{\Sigma}}^{\Pi}(L)\cdot a_{n_{a^y}}^{y}(L)\cdot\prod_{i=1}^{N}a^{w^y}{}_{n_{a^i}^{w^y}}^{i}(L)\cdot d^{w^y}{}_{n^{w^y}{}_d^{\nabla}}^{\nabla}(L)\cdot\prod_{i=1}^{N}a_{n_{a^i}}^{i}(L)\cdot d_{n_d^{\nabla}}^{\nabla}(L)$, $i=\overline{1,N}$; $\tilde{\tilde{d}}_j$,

$j=\overline{1,L^y+n^{w^y}{}_c^{\Sigma}+n_{a^y}+\sum_{i=1}^{N}n_{a^i}^{w^y}++\sum_{i=1}^{N}n_{a^i}+n_d^{\nabla}}$ – coefficients of the polynomial $\tilde{\tilde{d}}(L)=f^y(L)\cdot c^{w^y}{}_{n^{w^y}{}_c^{\Sigma}}^{\Pi}(L)\cdot a_{n_{a^y}}^{y}(L)\cdot\prod_{i=1}^{N}a^{w^y}{}_{n_{a^i}^{w^y}}^{i}(L)\cdot\prod_{i=1}^{N}a_{n_{a^i}}^{i}(L)\cdot d_{n_d^{\nabla}}^{\nabla}(L)$, $i=\overline{1,N}$; $\tilde{\tilde{a}}_j^{x^i}$,

$j=\overline{1,L^{x^k}+L^{\omega}+n^{x^k}{}_d^{\nabla}}$ – coefficients of the polynomial $\tilde{\tilde{a}}^{x^i}(L)=f^{x^i}(L)\cdot\omega(L)\cdot d^{x^i}{}_{n^{x^i}{}_d^{\nabla}}^{\nabla}(L)$; $\tilde{\tilde{c}}^{x^i}{}_j$,

$j=\overline{1,n^{x^k}{}_c^{\Sigma}}$ – coefficients of the polynomial $\tilde{\tilde{c}}^{x^i}(L)=c^{x^i}{}_{n^{x^i}{}_c^{\Sigma}}^{\Pi}(L)$.

To account for heteroskedasticity of the process (changing the variance in time) applied the generalized model with autoregressive conditional heteroskedasticity GARCH(m,r), which has the form [Перцовский, 2003]:

$$\sigma_t^2=w+\theta(L)\varepsilon_t^2+\varphi(L)\sigma_t^2,$$

where σ_t^2 – a time series of process dispersion changes y_t, $\theta(L)=\theta_1 L+\theta_2 L^2+\ldots+\theta_p L^m$, $\varphi(L)=\varphi_1 L+\varphi_2 L^2+\ldots+\varphi_r L^r$, ε_t^2 – the remainders of model. The GARCH(m,r) model can be expressed through the ARMA model as follows [Bollerslev, 1986]:

$$\varepsilon_t^2=\frac{w+(1-\varphi(L))}{(1-\theta(L)-\varphi(L))}v_t,$$

where $s=\max(r,m)$, $v_t=\varepsilon_t^2-\sigma_t^2$.

Fractal integrated GARCH process can be written as follows:

$$(1-L)^{2-H}\varepsilon_t^2=\frac{w+(1-\varphi(L))}{(1-\theta(L)-\varphi(L))\cdot(1-L)^{-1}}v_t,$$

where H – the Hurst factor.

The proposed spectrally integrated generalized model with autoregressive conditional heteroskedasticity is as follows:

$$f^{\varepsilon^2}(L)\cdot\omega^{\varepsilon^2}(L)\cdot\varepsilon_t^2=\frac{w+(1-\varphi(L))}{(1-\theta(L)-\varphi(L))}\cdot v_t,$$

where $\omega^{\varepsilon^2}(L)$ – integrating polynomial of delay operator, with which approximated the time series variance of the noise ε_t^2 into transformed smoothed by the "Caterpillar"-SSA method time series $w_t^{\varepsilon_t^2}$, and preliminary structural identification and a rough parametric identification of a polynomial of delay $f^{\varepsilon^2}(L)$, With which approximated itself series ε_t^2 is made in determining the coefficients of the recursive prediction formula "Caterpillar"-SSA method; w – the average value or the level of time series ε_t^2.

Thus, the autoregressive - spectrally integrated moving average with the spectrally integrated generalized autoregressive conditional heteroskedasticity and exogenous variables model (ARSIMA – SIGARCH model) takes the form:

$$\hat{z}_t^y(l)=\sum_{j=1}^{L^y+n_{b^y}+\sum_{i=1}^{N}n_{a^i}^{w^y}+n^{w^y}{}_d^{\nabla}+\sum_{i=1}^{N}n_{a^i}+n_d^{\nabla}}\tilde{\tilde{a}}_j\cdot z_{t+l-j}^y+$$

$$+\sum_{i=1}^{N}\left(\tilde{\tilde{b}}_0^i\cdot z_{t+l-m_i}^{x^i}-\sum_{j=1}^{\tilde{\tilde{b}}^i}\tilde{\tilde{b}}_j^i\cdot z_{t+l-m_i-j}^{x^i}\right)-\sum_{j=1}^{n_c^{\Sigma}+n_{a^y}+\sum_{i=1}^{N}n_{a^i}^{w^y}+n^{w^y}{}_d^{\nabla}+\sum_{i=1}^{N}n_{a^i}+n_d^{\nabla}}\tilde{\tilde{c}}_j\cdot e_{t+l-j}-,$$

$$-\sum_{j=1}^{L^y+n^{w^y}{}_c^{\Sigma}+n_{a^y}+\sum_{i=1}^{N}n_{a^i}^{w^y}++\sum_{i=1}^{N}n_{a^i}+n_d^{\nabla}}\tilde{\tilde{d}}_j\cdot e^{w^y}{}_{t+l-j};$$

$$\hat{z}_t^{x^k}(l)=\sum_{j=1}^{L^{x^k}+n_{x^k}{}_d^{\nabla}}\tilde{\tilde{a}}_j^{x^k}\cdot z_{t+l-j}^{x^j}+\sum_{j=1}^{n^{x^k}{}_c^{\Sigma}}\tilde{\tilde{c}}_j^{x^k}\cdot e_{t+l-j}^{x^k},\ k=\overline{1,N},$$

$$\varepsilon_t^2=\sigma_t z_t,$$

$$z_t \sim N(0,1),$$

$$f^{\varepsilon^2}(L)\cdot\omega^{\varepsilon^2}(L)\cdot\varepsilon_t^2 = \frac{w+(1-\varphi(L))}{(1-\theta(L)-\varphi(L))}\cdot v_t .$$

The model can be generalized to the multidimensional case.

Results

Testing the proposed model (4) was carried out on real data of daily consumption of gas from air temperature changes over a three year period.

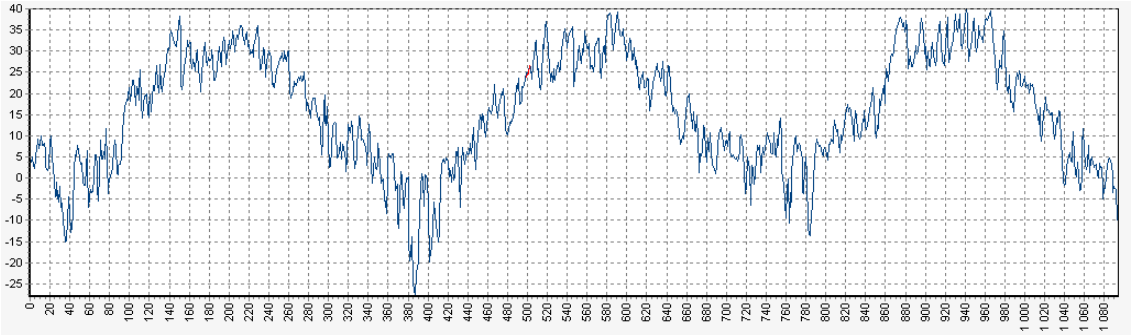

Fig. 1. The graph of daily changes in air temperature data

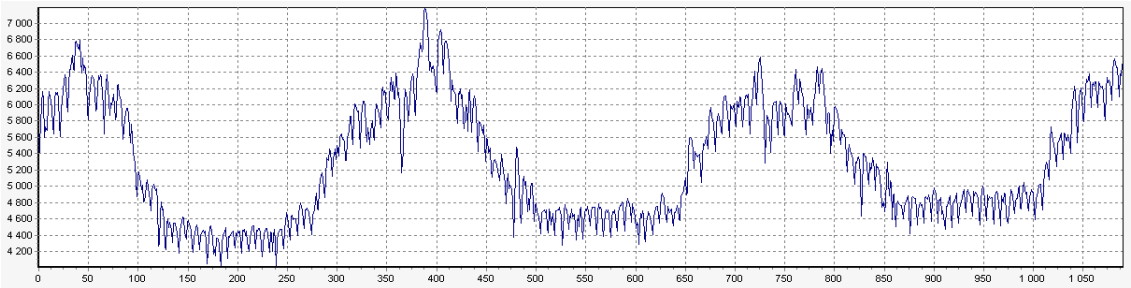

Fig..2. The graph of daily natural gas consumption data

Factoring time series corresponding to large and close-largest Eigen values of the information matrix of data can be judged weekly and yearly seasonal components of time series data.

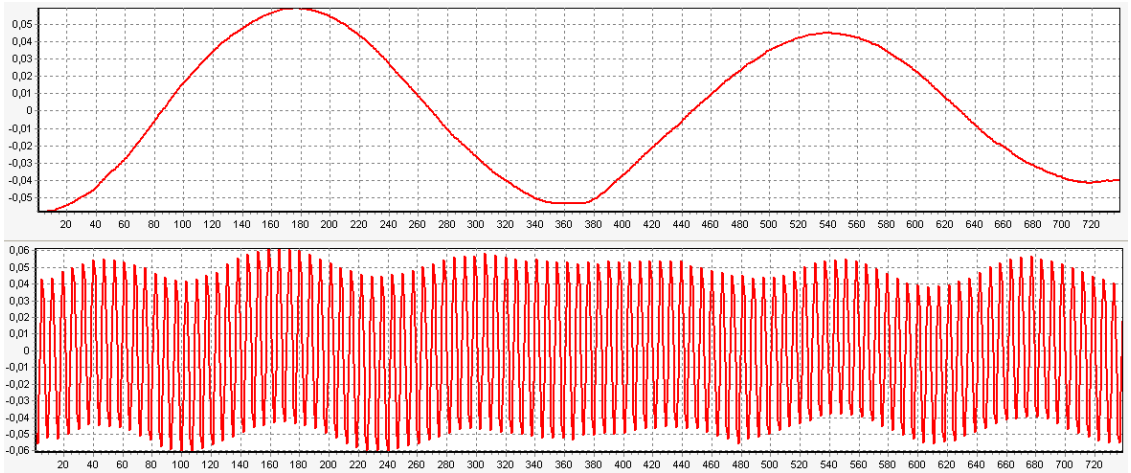

Fig. 3. Some of factor graphs of time series of given processes

Was obtained for the prediction SARIMAX model of natural gas consumption, taking into account changes in air temperature. The average percentage forecast errors was 1.87%.

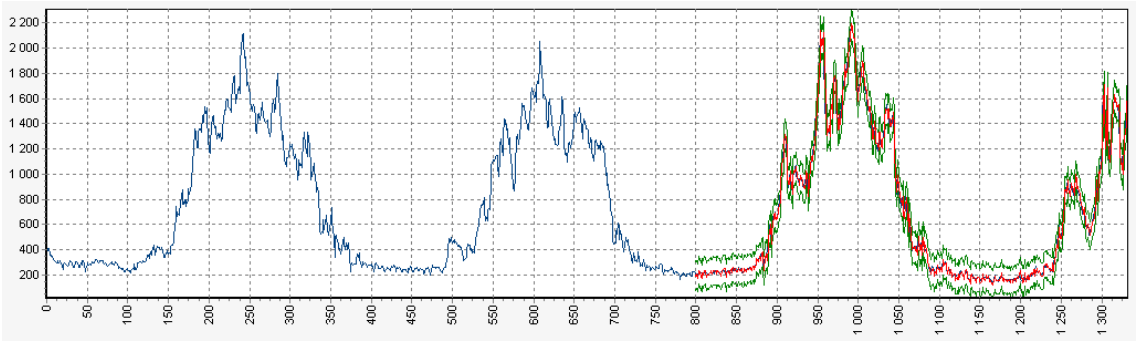

Fig. 4. Charts of forecasts of natural gas consumption by model SARIMAX and 95% confidence intervals

The average percentage error of forecasting natural gas consumption, taking into account changes in air temperature with proposed model was 1.12%. Together with a decrease in forecast errors and confidence intervals are narrowed.

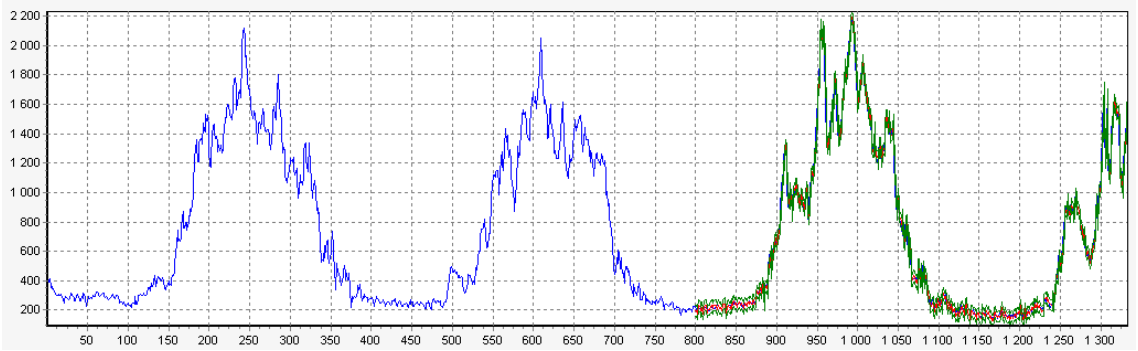

Fig. 5. Charts forecasts natural gas consumption of the proposed model and the 95% confidence intervals

In [Щелкалин, Тевяшев, 2011] presented the application of these models in various fields of science:

- operational forecasting of target products consumption processes in a housing and utilities infrastructure;
- simulation, prediction and control quasi–steady mode of gas–transport systems;
- automatized control for the construction of plants growing single crystals;
- for analysis and forecasting time series in economics;
- to describe and predict the physiological and psycho–physiological processes;
- to simulate the radio–processes and processes in the noise radar, etc;

Conclusion

Thereby, to obtain adequate models of the complex processes, high-quality forecasts it is necessary to combine the models with miscellaneous structures, including nonlinear models, which are complementary in their competitive learning. The proposed method of "Caterpillar"-SSA – ARIMA – SIGARCH is a modification of the method of "Caterpillar"-SSA with automatic separation of short-term memory and periodic components and can be interpreted as the development of models in state space and the proposed model ARSIMA – SIGARCH – as a

model ARFIMA – FIGARCH is a next modification of the model ARIMA – GARCH, a method of constructing the proposed model is an extension of the method of Box-Jenkins, but to build a broader class of models.

The proposed model ARSIMA – SIGARCH and method of its construction some intermediate approach boundary classical regression and modern neural networks, but more formalized at the choice of structure, being herewith optimum in detail with provision for existing on the date of mathematical, human and machine as the strengths and achievements and shortcomings and limitations.

Summing up the above-described advantages of the proposed method, once again it should be noted that the basic idea is the effect of synergy, which arises from the combined use of two methods: the "Caterpillar"-SSA method and the Box-Jenkins method.

The main advantage of the proposed method of constructing an adequate model of the process under study is its rigorous formalization and, consequently, the ability to fully automate all phases of construction and use of the model.

Bibliography

[Granger, 1980] Granger C.W.J., Joyeux R. An Introduction to Long-Memory Time Series Models and Fractional Differencing // Journal of Time Series Analysis. 1980. N 1(1). P. 15-29.

[Седов, 2010] Седов А.В. Моделирование объектов с дискретно-распределёнными параметрами: декомпозиионный подход / А.В. Седов; Южный научный центр РАН. – М. : Наука, 2010. – 438 с.

[Щелкалин, Тевяшев, 2010] Щелкалин В.Н., Тевяшев А.Д. «Автоматизированная система анализа и оперативного прогнозирования процессов потребления целевых продуктов в жилищно-коммунальном хозяйстве». Международный конкурс инновационных проектов “Харьковские инициативы”, 2010.

[Щелкалин, Тевяшев, 2011] Щелкалин В.Н., Тевяшев А.Д. Модель авторегрессии – спектрально проинтегрированного скользящего среднего со спектрально проинтегрированной обобщенной авторегрессионной условной гетероскедастичностью для моделирования, фильтрации, прогнозирования и управления процессами в современных системах автоматизации. // Труды Международной научно-практической конференции «Передовые информационные технологии, средства и системы автоматизации и их внедрение на российских предприятиях» AITA-2011. Москва, 4 – 8 апреля 2011 г. М.: Институт проблем управления им. В.А. Трапезникова РАН, 2011. с. 996 – 1022.

[Евдокимов, Тевяшев, 1980] Евдокимов А. Г., Тевяшев_А._Д. Оперативное управление потокораспределением в инженерных сетях. – Х. : Вища школа, 1980. – 144 с.

[Голяндина, 2004] Голяндина Н. Э. Метод «Гусеница»-SSA: прогноз временных рядов: Уч. Пособие,СПб, 2004. – 52 с.

[Перцовский, 2003] Перцовский О.Е. Моделирование валютных рынков на основе процессов с длинной памятью: Препринт WP2/2004/03 – М.: ГУ ВШЭ, 2003. – 52 с.

[Bollerslev, 1986] Bollerslev T. Generalized autoregressive conditional heteroscedasticity // Journal of econometrics. – 1986. – V. 31. – PP. 307 – 327.

Authors' Information

Vitalii Shchelkalin *– graduate student, Department of Applied Mathematics, Kharkiv national university of radioelectroniks, Lenina str., 14, Kharkiv, Ukraine; e-mail: vitalii.shchelkalin@gmail.com*

Major Fields of Scientific Research: Mathematical modeling, prediction, control theory, data analysis, neural networks, data mining