I.J. Modern Education and Computer Science, **2020, 6, 1-15**
Published Online December 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2020.06.01

Comparative Studies of Self-organizing Algorithms for Forecasting Economic Parameters

Volodymyr Lytvynenko
Department of Informatics and Computer Science, Kherson National Technical University, Kherson, Ukraine
Email: immun56@gmail.com

Olena Kryvoruchko
Department of Economic Theories, National University of Water and Environmental Engineering, Rivne, Ukraine
Email: o.p.kryvjruchko@nuwm.edu.ua

Irina Lurie
Department of Informatics and Computer Science, Kherson National Technical University, Kherson, Ukraine
Email: lurieira@gmail.com

Nataliia Savina
Department of Economic Theories, National University of Water and Environmental Engineering, Rivne, Ukraine
Email: n.b.savina@nuwm.edu.ua

Oleksandr Naumov
University of State Fiscal Service of Ukraine, Irpin, Ukraine
Email: abnaumov75@gmail.com

Mariia Voronenko
Department of Informatics and Computer Science, Kherson National Technical University, Kherson, Ukraine
Email: mary_voronenko@i.ua

Received: 25 September 2020; Accepted: 28 October 2020; Published: 08 December 2020

Abstract: This manuscript presents the economic research results based on their input-output characteristics and functional description with inductive modeling methods and tools. There are a wide plethora of methods to be used for solving this type of problem, including various neural network models, linear and nonlinear regressions, reference vectors' methods, fuzzy models, etc. The main disadvantage of these methods is that the obtained models cannot always interpret and obtain a model of optimal complexity. Unlike the mentioned methods and tools, the group method of data handling (GMDH) allows building models directly from a data sample without the attraction of additional a priori information. This algorithm admits finding internal dependencies in the data and determining optimal model complexity. There is a broad range of iterative GMDH algorithms that have been developed and studied. Oversampling algorithms are applicable for solving the structural identification problems for a limited number of arguments. Iteration algorithms are suitable for solving tasks with many arguments, but they do not guarantee proper structure development. Multi-row GMDH iteration algorithms are the most popular ones. However, they have several sufficient defects, such as informative argument loss or non-informative argument inclusion, as well as a polynomial degree of exponential growth. In this context, the applicability of the GMDH-based iterative and combined architectures for solving the model's interrelation problems between a volume of capital investments and GDP by activity types in the transport branch is considered. The determination coefficient is utilized for the estimation of the obtained models based on a complicated evaluation procedure. The Kolmogorov-Smirnov criterion estimates the model's adequacy. The F-criterion Fisher assesses the significance of polynomial models. The demonstrated results proved that the combined iterative and combinatorial algorithms turned out to be the most effective solution for all evaluation criteria.

Index Terms: Group Method of Data Handling; Iterative Algorithm; Gross Domestic Product; Investment; Model Adequacy.

1. Introduction

Successful implementation of the investment policy will contribute to implementing one of the leading countries' economic tasks to increase the number of primary domestic investment resources sources. This will create the necessary prerequisites for the production growth and expanded reproduction of GDP to increase the population's well-being.

Therefore, it would be advisable to determine the informative investment indicators that have the most significant impact on the dynamics of Ukraine's GDP. It is necessary to develop a model of the relationship between capital investment and GDP.

In [1], the author defined investment as "the current increase in capital property values as a result of production activities during a given period," or as a share of income for a given period that was not used for consumption. The investment is said to be the material basis for the economy's structuring. Solving the problem of investment will mean the beginning of not only the economy's restructuring but also its stabilization and subsequent growth [2].

The term "investment" has several meanings. First, it means the purchase of shares, bonds with the expectation of specific financial results. Secondly, real assets like machinery, equipment necessary for the production and sale of any goods seem to be a great option. In the broadest sense, investments provide the mechanism necessary to finance the growth and development of the country's economy, region, industry, or enterprise.

Successful implementation of the investment policy will contribute to implementing one of the main tasks of the country's economy to increase the number of primary sources of domestic investment resources. This will create the necessary prerequisites for the production growth and expanded reproduction of GDP to increase the population's well-being.

Complex economic processes are characterized by input and output variables and parameters that determine the object's internal state and depend on numerous random factors. It is known that economic processes are characterized by instability and instability. At the same time, well-known models in practice are often unsuitable for forecasting. In such situations, for the analysis, modeling, and forecasting of these processes, it is advisable to apply direct construction methods from observational data (statistics). The purpose of such methods is the identification of implicit causal relationships and patterns hidden in the data and the construction of mathematical models in explicit form. Therefore, obtaining an adequate mathematical description of such processes is a complex research project.

Static methods are used to describe stochastic processes, which allow us to extract the necessary results with incomplete information about the economic process's mechanisms. One of the most acceptable methods for solving complex tasks related to modeling investment processes is the group method of data handling (GMDH) [3, 4].

The GMDH algorithms are widely used for solving identification problems of high order nonlinear dependencies between input and output variables. Even though these algorithms can cope with high-dimensional data and successfully cope with forecasting and modeling complex nonlinear processes, the questions of selecting a subset of variables necessary for a significant dimensionality reduction are still unclear. At present, there are a plethora of methods for the implementation of the self-organization idea [5, 6]. However, there is a problem in choosing the most effective models [7]. Therefore, a comparative study is carried out by algorithms of the GMDH iterative combinatorial type to identify the hidden dependencies between different types of capital investments in the transport industry and GDP.

In this study, we did not limit the study to evaluating the accuracy of the obtained models but assessed their adequacy and significance.

The novelty of this research has to do with:

- A new extended method based on the comparative analysis of the GMDH iteration algorithms has been developed;
- A methodology for estimating GDP from the volume of investments in the transport sector has been proposed.

The work aims to develop and research models of the relationship between the volume of capital investments and GDP by type of activity in the transport industry.

This article is arranged as follows. Section II gives a formal statement of the problem. Section III presents a detailed analysis of the work based on models of complex systems using inductive simulation tools to describe the system's input and output characteristics. Section IV describes methods of applying linear and nonlinear inductive GMDH models, as well as methods of assessing their quality. Section V presents the obtained experimental results. Finally, concluding remarks are given in Section VI.

2. Problem Statement

Such a statement is to find the extremum of some criterion *CR* on the set of different models $\Im$:

$$f^* = \arg\min_{f \in \Im} CR(f). \quad (1)$$

Since (1) is not completed by the formulation of the problem, it needs to be further identified, in particular: to ask a priori expert or expert information about the kind, character, and volume of the initial information to be known from the analysis of the experiment; to specify a class of essential functions from which the set must be formed $\Im$; to determine how to generate models *f*; to set the method of estimating parameters; to set the model comparison criterion $CR(f)$ and specify a method for minimizing it. Specify this statement, assuming that the given sample $W = \left[X_y \right]$ contains observation *n* points that form the matrix $X = \left\{ x_{ij}, i = 1,..., n;\ j = 1,..., m \right\}$ and the vector $y = (y_1 ... y_m)^T$, for which $n \geq m$.

In general, the process of constructing models according to experimental data (1) includes, in particular, the following main stages:

Specifying a sample of experimental data, a priori, and expert information, and dividing the data table into at least two non-disjoint subsets;

Definition of a class of essential functions;

Generation different patterns of models in the selected class;

Evaluation of the parameters of the generated structures and the formation of the set $\Im$;

Minimization of the given criterion $CR(f)$ and choice of the optimal model f^* ;

Checking the adequacy of the optimum model obtained;

The decision to complete the simulation process.

3. Related Works

In the problems of constructing models of complex systems in the conditions of incomplete information, methods and means of inductive modeling are actively used, primarily intended for the functional description of the input-output characteristics of the systems.

The first part of the preliminary literature review was described in the Introduction. Furthermore, some more specific issues are dwelled on in this section.

The scientific direction is called "inductive modeling of complex processes and systems," formed by Alexei Ivakhnenko [8].

The author called this new direction differently: first, "heuristic self-organization" [9], later "self-organization of models based on experimental data" [10], then "inductive self-organization of models of complex systems" [11].

Finally, the various methods of modeling that can be classified as inductive are highlighted by the group method of data handling (GMDH), allowing to build models directly on data sampling without attracting additional a priori information [12, 13].

GMDH is successfully used in data analysis and identification of patterns, modeling, forecasting [14], structural identification, clustering, classification, and pattern recognition [15, 16].

It allows you to automatically find interdependencies and patterns implicitly reflected in the data, and present them in an explicit form of mathematical models of optimal complexity.

As of today, many varieties of algorithms GMDH of robotic [17] and iterative [3] types have been developed and investigated. Reordering algorithms are practical as a means of structural identification, but only for a limited number of arguments since they are based on a full or directed enumeration of all possible variants of model structures [18].

Iterative algorithms work with a sufficiently large number of arguments. However, the specificity of their architecture does not guarantee the construction of a proper structure model, since they are based on incomplete inductive procedures for hierarchically complicating models. For a long time, these two classes of algorithms developed independently, without a detailed study of the possibility of combining their strengths while eliminating shortcomings [19].

4. Applied Methods

The following macroeconomic indicators for the period from 2012 (1st quarter) to 2017 (4th quarter) - 24 points were taken as experimental data for calculating the dependence between the Ukraine's GDP growth and the volume of investment in the transport industry (http://www.ukrstat.gov.ua/).

In Table I, there are numerical values used to build models of statistic data on capital includity and calculation of an including product on types of activities "transport, warning, leading and courier activities".

x_1 is the volume of investments in land and pipeline;

x_2 is the volume of investment in water transport;

x_3 is the volume of investment in air transport;

x_4 is the volume of investments in warehousing and auxiliary activities in the field of transport;

x_5 is the volume of investments in postal and courier activities.

In general, the data set is divided into two parts: 16 measurements are training sample A, eight measurements are a test sample B.

It is necessary to establish a model of the relationship between the volume of capital investments and GDP by type of activity "Transport, warehousing, postal and courier activities."

In this paper, a model simulation class is considered for a statistical sample containing information about *n* observations for *m* input variables *X* [*n*×*m*] and one output variable *y* [*n*×1].

It is necessary to find a model of the input-output dependence in the conditions of incompleteness and uncertainty of information on the structure of this dependence.

From the available positions, the task of identification is to form a sample of experimental data for a particular set $\Im$ of models of the different structure of the species [20]:

$$\hat{y}_f = f\left(X, \hat{\theta}_f\right) \tag{2}$$

According to [20], in the formation $W = (X \vdots y)$ of a sample $\Im$ of a plurality of different structures models of the species $\hat{y}_f = f(X, \hat{\theta}_f)$ and the search for an optimal model for a minimum of a given criterion according to the data of a sample $CR(\cdot)$:

$$f^* = \arg \min_{f \in \Im} CR(y, f(X, \overset{J}{\theta}_f)), \tag{3}$$

where estimating the parameters $\overset{J}{\theta}_f$ for each model $f \in \Im$ is the solution to another problem of the form:

$$\hat{\theta}_f = \arg \min_{f \in R^{s_f}} Q(y, X, \theta_f) \tag{4}$$

where $Q(\cdot) \neq CR(\cdot)$ – the quality criterion for solving the problem of parametric identification of a partial model that has the complexity S_f generated in the task of structural identification (3).

For example, the external criteria $CR(\cdot)$ for regularity or unbounding, based on a sample split, can be used as a criterion in the GMDH.

The criterion of regularity (expresses the model error in different parts of the sample). Since the second principle of inductive modeling of complex systems on the principles of GMDH is the principle of external complement, which requires breakdown of the table of output data into the training (A) and control (test - B) part, then as a criterion Regularity, and all the others that will be mentioned here, still have the application of "external," which emphasizes this principle.

The regularity criterion is used to solve pattern recognition tasks, identification, and prognosis, and in one of the most straightforward variants, it has the form:

$$\Delta^2(B) = \frac{\sum_1^{N_B} (q_F - \hat{q}_M)^2}{\sum_1^{N_B} q_F^2} \to \min \tag{5}$$

where q_F — actual (tabular) data; $\hat{q}_M$ — output of the model; $N=N_A+N_B$ — the set of points of the output data is divided into two parts: N_A and N_B.

The regularity criterion (5) was also used in the form of the correlation coefficient between the variables q_F and $\hat{q}_M$ inside the interval of data *B* or the correlation index (for nonlinear models).

To simulate the dependence of GDP growth in Ukraine (UAH Million) on the volume of investments, iterative algorithms were used, including three linear algorithms and six with a quadratic function, including combined algorithms for constructing models:

Linear:

1) multi-row iterative (lin m);

2) relaxation (lin r);

3) combined iterative (lin c);
With quadratic function:
4) multi-row iterative with quadratic function (sq m);
5) a relaxation quadratic function (sq r);
6) a relaxation-combinatorial quadratic function (sq c);
7) multi-row iterative-combined (sq combi m);
8) relaxation iterative-combined (sq combi r);
9) combined iterative-combinatorial (sq combi m).

A. Multi order Iterative Algorithm of GMDH

In the classical multi-row algorithm of **GMDH,** the problem of constructing an optimal model is solved inductively: the model of a gradually complicated structure is constructed, and the process of complication has the character of iterations when the best preliminary results are used in the next series (iterations).

The complication is the only rule that allows you to build an arbitrarily complex model from a large number of variables (arguments) that characterize the object of modeling [18].

Consider the more complex operations performed on the 1st and arbitrary (r + 1)-th ranks.

Table 1. Statistical Data of Capital Investments and Gross Domestic Product by Type of Activity "Transport, Warehousing, Postal and Courier Activity"

Years	Capital investment, mln. UAH					GDP, mln. UAH
	Ground and pipeline transport	Water transport	Air transport	Warehousing and auxiliary transport activities	Postal and courier activities	
1	2	3	4	5	6	7
	X1	X2	X3	X4	X5	Y
1st quarter 2012	1992,1	35,7	275,3	2820,1	20,4	292894
2nd quarter 2012	4378,5	26,8	176,0	3642,6	35,4	347842
3rd quarter 2012	2601,8	41,7	133,5	3581,9	168	389213
4th quarter 2012	3691,8	28,4	195,2	4140,5	172,4	381289
1st quarter 2013	587,8	6,3	98,6	1779,2	5,0	303753
2nd quarter 2013	1189,4	37,3	137,6	2042,0	5,1	354814
3rd quarter 2013	1440,2	31,8	131,6	3039,0	8,2	398000
4th quarter 2013	1650,3	21,4	155,3	3929,4	202,1	408631
1st quarter 2014	590,3	29,3	73,3	1876,8	3,8	316905
2nd quarter 2014	1140,6	41,0	79,2	2230,4	4,1	382391
3rd quarter 2014	652,3	89,0	71,4	1820,4	9,1	440476
4th quarter 2014	1072,6	48,8	96,2	3921,9	105,5	447143
1st quarter 2015	1805,6	23,1	116,3	904,4	4,9	375991
2nd quarter 2015	874,6	96,2	193,7	1935,6	2,8	456715
3rd quarter 2015	2386,9	111,6	125,6	2083,2	5,3	566997
4th quarter 2015	2150,9	102,9	223,1	3058,7	72,6	588841
1st quarter 2016	1952,2	36,6	99,2	1439,6	13,3	455298
2nd quarter 2016	2256,5	38,0	177,7	2114,6	22,5	535701
3rd quarter 2016	3365,5	53,3	218,4	2646,5	18,7	671456
4th quarter 2016	7383,0	106,2	202,3	2527,1	66,5	722912
1st quarter 2017	3693,0	50,8	210,2	1454,0	6,6	591008
2nd quarter 2017	4181,9	55,4	260,4	1979,0	51,8	664760
3rd quarter 2017	4627,7	56,9	372,2	2852,1	53,7	833130
4th quarter 2017	9455,6	74,6	340,4	5640,6	285,3	894022

The first iteration (first row). From the set of inputs $X = \{x_1, x_2, ..., x_m\}$, pairs of arguments are chosen, and partial views of the form are formed:

$$z = f(u,v) = a_0 + a_1 u + a_2 v;$$
$$z = f(u,v) = a_0 + a_1 u + a_2 v + a_3 uv; \quad (6)$$
$$z = a_0 + a_1 u + a_2 v + a_3 uv + a_4 u^2 + a_5 v^2.$$

That is $y_l^1 = f(x_i, x_j)$, $l = 1, 2, \text{K}, C_m^2$, and according to the MNA, for each description, according to the data of the training sample, estimates of unknown parameters (coefficients) $\hat{a}_0, \hat{a}_1, \hat{a}_2, ..$ are presented. By the chosen criterion, on the test sample, F chooses the best models $\hat{y}_k^1$, $k = \overline{1, F}$; that is, they implement the selection procedure, where F is

called the freedom of choice. The outputs of these models are arguments-inputs for constructing models of the next series. The following is the minimum CR_{min}^{1} among all $\boldsymbol{F}$ values of the criterion in the first row.

An arbitrary (r + 1) iteration (series r + 1): From the argument vectors $\hat{y}_k^1, k=\overline{1,F}$, the previous r-th row, all possible partial description of the form (6) is formed, that is:

$$y_l^{r+1} = f(y_i^r, y_j^r),\ l=1,2,\cdots,C_F^2,\quad i,j=\overline{1,F} \tag{7}$$

Furthermore, according to the MNA, $\boldsymbol{A}$ parameters are evaluated. According to the selection criterion, $\boldsymbol{F}$ is selected as the best model $\hat{y}_k^1, k=\overline{1,F}$, among which is $CR_{\min}^{r+1}$ the condition $CR_{\min}^{r+1} \geq CR_{\min}^{r}$ that the iteration process stops; otherwise the transition to the next series is checked. In the case of stopping the process, the model corresponding to the value $CR_{\min}^{r}$ in the previous r-series is accepted as the optimal one.

C. The Relaxation Iterative Algorithm (RIA)

The name of the relaxation iterative algorithm (RIA) corresponds to the analogy with optimization algorithms. In this algorithm, in each row, the intermediate arguments are combined in pairs with the original [21], which prevents loss of informative arguments, with such possible variants of linear, bilinear, or partial quadratic descriptions:

$$\begin{gathered} y_l^r = a_0 + a_1 y_i^{r-1} + a_2 x_j \\ y_l^r = a_0 + a_1 y_i^{r-1} + a_2 x_j + a_3 y_i^{r-1} x_j \\ y_l^r = a_0 + a_1 y_i^{r-1} + a_2 x_j + a_3 (y_i^{r-1})^2 + a_4 y_i^{r-1} x_j + a_5 (x_j)^2 \end{gathered} \tag{8}$$

Description of the algorithm. The first row is executed as in the classic multi-row algorithm.

An arbitrary $(r+1)$ - this iteration (row $r+1$). From the argument vectors $\hat{y}_k^r, k=\overline{1,F}$, the previous r-series, all possible partial description of the form (8) is formed, i.e.:

$$y_l^{r+1} = f(y_i^r, x_j),\ l=1,2,\ldots,C_F^2,\quad i,j=\overline{1,F} \tag{9}$$

Moreover, for the MNA on A, there are parameter estimates. Then, according to the selection criterion, $\boldsymbol{F}$ selected the best models $\hat{y}_k^{r+1}, k=\overline{1,F}$, and is located $CR_{\min}^{r+1}$ and verifies the condition $CR_{\boldsymbol{min}}^{r+1} \geq CR_{\boldsymbol{min}}^{r}$ in which the iterative process stops, otherwise the transition to the next row. In the case of a stop, an optimal model is adopted, which corresponds to the value CR_{min}^{r} in the previous r -series.

In this algorithm, it is also possible to use different partial descriptions on different rows to reduce the complexity of models, for example, using the second-order polynomial in the first row and only the linear form for the following series.

D. Combined Iterative Algorithm

This algorithm from the previous is different in that each pair of pairs are formed only from intermediate arguments and from intermediate and initial [22].

Description of the algorithm. The first row is executed as in the classic multi-row algorithm.

An arbitrary $(r+1)$ -this iteration (row $r+1$). From vector arguments $\hat{y}_k^r, k=\overline{1,F}$, the previous r-th row generates all possible partial descriptions of the form (1) or (8), i.e.:

$$\begin{gathered} y_l^{(r+1)} = f(y_i^r, y_j^r) \vee f(y_i^r, x_j), \\ l=1,2,\cdots,C_F^2,\quad i,j=\overline{1,F} \end{gathered} \tag{10}$$

Moreover, for the MNA on A, there are parameter estimates. Then, according to the selection criterion, $\boldsymbol{F}$ selected the best models $\hat{y}_k^1, k=\overline{1,F}$, and is located CR_{min}^{r+1} and verifies the condition $CR_{\boldsymbol{min}}^{r+1} \geq CR_{\boldsymbol{min}}^{r}$ in which the

iterative process stops, otherwise the transition to the next row. In the case of a stop, an optimal model is adopted, which corresponds to the previous r-series's value CR_{min}^{r}.

E. Multi-Row Iterative-Combinatorial Algorithm

The combinatorial optimization of partial descriptions, that is, the application of the idea of active neurons - linear, bilinear, or nonlinear - can be applied in each of the varieties of iterative algorithms [23].

Consider the more complex operations performed on an arbitrary $(r+1)$ -th series (including on the first one).

An arbitrary $(r+1)$ -this iteration (row $r+1$). The previous r -th row (or input arguments, if it is the first row), all possible partial descriptions (1) are formed from vector arguments. Combinatorial optimization consists in the fact that on each row models are considered, for example, such a kind (for a partial linear description):

$$y_l^{(r+1)} = f_{opt}(y_i^r, y_j^r) = a_0 d_1 + a_1 d_2 y_i^{r-1} + a_2 d_3 y_j^{r-1} \tag{11}$$

where d_k, $k = 1,2,3$ are elements of the binary structural vector d that accept values 1 or 0 (inclusion or not the inclusion of the corresponding argument):

$$d_k = \{0,1\}, \quad d_{opt} = \arg \min_{l=1,q} CR_l, \quad q = 2^p - 1, \quad f_{opt}(u,v) = f(u,v,d_{opt}).$$

The layout of the survey then has the form:

$$\begin{array}{ll} 100 \rightarrow f_1 = a_0 & CR_1 \\ 010 \rightarrow f_2 = a_1 y_i^{r-1} & CR_2 \\ 001 \rightarrow f_3 = a_1 y_j^{r-1} & CR_3 \\ 110 \rightarrow f_4 = a_0 + a_1 y_i^{r-1} & CR_4 \\ 101 \rightarrow f_5 = a_0 + a_1 y_j^{r-1} & CR_5 \\ 011 \rightarrow f_6 = a_1 y_i^{r-1} + a_2 y_j^{r-1} & CR_6 \\ 111 \rightarrow f_7 = a_0 + a_1 y_i^{r-1} + a_2 y_j^{r-1} & CR_7 \end{array} \Rightarrow \min f_{opt}$$

At the same time, the best option for the CR criterion minimum is selected, i.e., the complexity of the partial model is optimized:

$$d_{opt} = \arg \min_{l=1,q} CR_l, \quad q = 2^p - 1, \quad f_{opt}(u,v) = f(u,v,d_{opt}) \tag{12}$$

Then, according to the selection criterion, $\boldsymbol{F}$ selected the best models $\hat{y}_k^{r+1}$, $k = \overline{1,F}$ and is located CR_{min}^{r+1} and verifies the condition$CR_{min}^{r+1} \; r_{min}$ in which the iterative process stops, otherwise the transition to the next row. In the case of a stop, an optimal model is adopted, which corresponds to the value CR_{min}^{r} in the previous *r*-series.

F. Relaxation Iterative-Combinatorial Algorithm [24]

By adding in the multi-row iterative-combinatorial algorithm, the possibility to combine intermediate arguments in each row in pairs with the initial ones receive a relaxation iterative-combinatorial algorithm.

An arbitrary $(r+1)$ -this iteration (row $r+1$). From vector arguments $\hat{y}_k^1$, $k = \overline{1,F}$, The previous r-th row (or input arguments, if it is the first row), all possible partial descriptions (1) are formed. Combinatorial optimization consists in the fact that on each row models are considered, for example, such a kind (for a partial linear description):

$$y_l^{(r+1)} = f_{opt}(y_i^r, x_j^r) = a_0 d_1 + a_1 d_2 y_i^{r-1} + a_2 d_3 x_j^{r-1} \tag{13}$$

where d_k, $k = 1,2,3$ are elements of binary structural vector *d*, which accept values 1 or 0 (inclusion or not the inclusion of the corresponding argument):

$$d_k = \{0, 1\}, d_{opt} = \arg \min_{l=1,q} CR_l, \quad q = 2^p - 1, \quad f_{opt}(u,v) = f(u,v,d_{opt}).$$

Then, according to the selection criterion, F is selected as the best model of the series.

G. Combined iterative-combinatorial algorithm (CICA)

CICA or generalized iterative algorithm (GIA). The combinatorial optimization of partial model structures in the combined algorithm gives the algorithm the full name of the "combined iterative-combinatorial algorithm" of CICA. His further generalization, which takes into account both CICA and all its partial cases, as well as various variants of operating modes implemented through the user interface, are called "Generalized iteration algorithm" or GIA GMDH [13, 19, 25].

Formally, in the general case, for a series r, it is possible to define the GIA of the GMDH as follows [26]:

1) the input matrix is $X_{r+1} = (y_1^r, ..., y_F^r, x_1, \cdots, x_m)$;
2) the transition type operators are used:

$$\begin{aligned} y_l^{r+1} &= f(y_i^r, y_j^r), \ l = 1,2,\text{K}, C_F^2, \quad i, j = \overline{1,F}, \\ y_l^{r+1} &= f(y_i^r, x_j), l = 1,2,\text{K}, Fm, i = \overline{1,F}, \ j = \overline{1,m} \end{aligned} \tag{14}$$

with a linear, bilinear, or partial quadratic description (1);

for each description is the optimal structure, n-d, for the linear form:

$$f(u,v) = a_0 d_1 + a_1 d_2 u + a_2 d_3 v \tag{15}$$

where $d_k, k = 1,2,3$ are elements of the binary structural vector d that accept values 1 or 0 (inclusion or non-inclusion of the corresponding argument):

$$d_k = \{0,1\}, d_{opt} = \boldsymbol{arg\ min}_{l=1,q} CR_l, \quad q = 2^p - 1, \quad f_{opt}(u,v) = f(u,v,d_{opt}).$$

4) the algorithm stops when the condition $CR_{\boldsymbol{min}}^{r+1} \geq CR_{\boldsymbol{min}}^r$ is fulfilled, and the optimal model corresponds to the value $CR_{\min}^r$ on the r-th row.

H. The Accuracy of the Models

The accuracy of the models obtained was based on the formula of the determination coefficient (R^2 - statistics):

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2} 100\%, \tag{16}$$

where $\bar{y}$ is an average value and $\hat{y}_i$ is an output of the model.

The determination coefficient characterizes the fraction of the variance of the resultant variable Y, an explanation of the polynomial regression, in the overall variance of the resultant variable *Y*. Accordingly, the magnitude 1-R^2 characterizes the fraction of the variance of the variable Y caused by the influence of other factors not taken into account in the model.

I. Adequacy of the Obtained Models

The criterion by Kolmogorov-Smirnov is intended to determine that the sample has a corresponding distribution; in our case, is expected. For example, the hypothesis is checked that the sample has a normal distribution. The small value of probability means the rejection of the hypothesis of normality. For a sample with a normal distribution, the probability value tends to be unity.

Let X_n be a sample of independent equally distributed random variables, $F_n(x)$ is an empirical distribution function, $F(x)$ stands for a specific "true" distribution function with known parameters. The statistics of the criterion are determined by the following expression:

$$D_n = \sup_x \left| F_n(x) - F(x) \right| \tag{17}$$

We denote by the H_0 hypothesis that the sample is subject to distribution $F(x) \in C^l(X)$. Then Kolmogorov's theorem is validated by the hypothesis:

$$\forall t > 0: \lim_{n \to \infty} P\left(\sqrt{n}\, D_n \le t\right) = K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2 t^2} \tag{18}$$

Hypothesis H_0 rejected if statistics $\sqrt{n}\, D_n$ exceeds quantile distribution K_α the given level of significance $\alpha$and is accepted otherwise. In Kolmogorov's criterion, it is expedient to use statistics with the correction of Bolshev [27]: $\sqrt{n}\, D_n + 1 / \left(6\sqrt{n}\right)$. The distribution of these statistics with fairness is verified by the hypothesis quickly converges to the distribution of Kolmogorov, and depending on the size of the sample can be neglected. In the Kolmogorov-Smirnov test for one sample, hypotheses are tested:

- Hypothesis 0 (zero): Deviation from a normal distribution is significant
- Hypothesis 1 (alternative): The values of the variable are well suited to the normal distribution.

The probability of an error in which it is permissible to reject the null hypothesis and accept an alternative hypothesis is defined as a value that is in the range from 0 to 1.

Usually, the probability is denoted by the letter p: $0 < p < 0$. There is a commonly used terminology that relates to confidence intervals of probability. Expressions with a probability of error $p \le 0.05$ are called meaningful; statements with a probability of error $p \le 0.01$ - very significant, and statements with the probability of errors $p < 0.001$ - the most significant. A deviation from the normal distribution is considered significant at the value of $p < 0.05$; In this case, nonparametric tests should be used for the respective variables. In the considered example (the value of $p = 0.467$), that is, the probability of error is not significant, so the value of the variable is good enough to fit into the normal distribution.

A deviation from the normal distribution is considered significant at the value of $p < 0.05$; In this case, nonparametric tests should be used for the respective variables. In the example under consideration (the value of $p = 0.616$), the probability of error is not significant, so the variable's values are well suited to the normal distribution, and parametric tests can be used [28-30].

J. The evaluation of the Obtained Models

The evaluation of the significance of the entire regression equation as a whole is carried out using Fisher's F-criterion. Fisher's F-criterion consists of checking the null hypothesis about the statistical insignificance of the regression equation**.**

This is done by comparing the actual F_{fact} , and the critical (tabular) values F_{tabl} of the F-criterion F_{fact} of Fisher. is determined from the ratio between the values of factor and residual variances, calculated for one degree of freedom:

$$F_{fact} = \frac{\sum \frac{\left(\hat{y} - \bar{y}\right)^2}{m}}{\sum \frac{\left(y - \hat{y}\right)^2}{n - m - 1}} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}, \tag{19}$$

where *n* is a number of units of the population; m describes a number of parameters with variables R^2 (a coefficient of determination).

F_{tabl} is the maximum possible value of the criterion under the influence of random factors with degrees of freedom $k_1 = m$, $k_2 = n - m - 1$ (for linear regression $m = 1$), and level of significance α. The level of significance α is the probability of rejecting the correct hypothesis provided that it is true. Usually a value α equals 0,05 or 0,01. If $F_{tabl} < F_{fact}$, then, H_0 the random nature of the evaluated characteristics is rejected, and their statistical significance and reliability are recognized. If $F_{tabl} > F_{fact}$, then, the hypothesis does not deviate and recognizes the statistical insignificance, the unreliability of the regression equation.

The Kolmogorov-Smirnov criterion is acceptable $n \geq 20$ to test the hypothesis, whether the random variable of some theoretical distribution law is obeyed if its parameters are assumed to be known (a simple hypothesis). The test can be carried out for any type of distribution.

The criterion is based on determining the maximum deviation of the accumulated frequency (empirical distribution function) from the predicted theoretical distribution function. The results are in the variation series. Find the upper and lower bounds of the corresponding deviation:

$$D_n^+ = \max\left[\frac{i}{n} - F(x_i)\right], \tag{20}$$

at $1 \leq i \leq n$

$$D_n^- = \max\left[F(x_i) - \frac{x-1}{n}\right], \tag{21}$$

at $1 \leq i \leq n$

where $F(x_i)$ is the value of the theoretical distribution function. Choose the maximum outside the range of deviations:

$$D_n = \max\left[D_n^+; D_n^-\right]. \tag{22}$$

The criterion statistics can be calculated according to the formula:

$$\lambda = \frac{6nD_n + 1}{6\sqrt{n}}. \tag{23}$$

The estimated value is compared with the table λ_{tabl}. Table values are previously known.

If $\lambda_{fact} \leq \lambda_{tabl}$ then the distribution is considered to be theoretical with the distribution function $F(x)$ with known parameters at the chosen level of significance α.

5. Experiments and Results

The quality of the constructed model was calculated on the sub-sample B as the value of the regularity criterion $AR(\square)$. The accuracy of the model was also tested on the sub-sample B as the value of the determination coefficient R^2 (Tables II and III).

For each object, the balance can be calculated $e_i = Y_i - Y_i^F$, $i = 1, 2, \text{K}, n,$ i, where $\hat{Y}_i$ is an estimated value. The balance is useful for studying the adequacy of the data model.

This means that the requirements for the residue independence for individual observations must be met; the variance should not depend on X. The results of assessing the adequacy of the models obtained using the Kolmogorov-Smirnov criteria are given in Tables IV.

The results of the evaluation of the significance of the whole equation of the obtained model using Fisher's F-criterion are given in Table VI.

Table 2. Simulation Results for Linear Dependence

№	Algorithm's building models	AR Norm	R^2	Model
1	lin m	0,098	0,558	$y = 264387,04 + 2301,94 * x_2 + 249,56 * x_5 - 62,49 * x_3 + +13,58 * x_1$
2	lin r	0,109	0,561	$y = 376393,94 + 64,00 * x_4$
3	lin c	0,096	0,529	$y = 284895,56 + 2287,49 * x_2 + 346,53 * x_5 - -83,39 * x_3 - 10,18 * x_4 + 16,85 x_1$

Table 3. Simulation Results for Nonlinear Dependence

№	Algorithm's building models	AR Norm	R^2	Model
1	sq m	0,096	0,523	$y = 466075,21 - 1728,12 * x_2 - 116,49 * x_4 + 19,28 * x_2^2 + +0,02 * x_4^2 + +0,8 * x_2 * x_4$
2	sq r	0,055	0,815	$y = 132928,72 + 12,11 * x_1 + 2243,14 * x_3 + 3543,94 * x_2 + +2,95 * x_3^2 - 39,86 * x_2 * x_3 + 4,64 * x_2^2 - 0,1 * x_3^3 + +0,17 * x_2 * x_3^2$
3	sq c	0,034	0,809	$y = 352822,33 - 10,27 * x_1 - 0,003 * x_1^2 + 4,93 * x_2 * x_3 + +0,004 * x_1 * x_2 * x_3$
4	sq combi m	0,026	0,845	$y = 354266,26 - 6,01 * x_1 - 351,13 * x_3 - 0,04 * x_3 * x_5 + +0,03 * x_5^2 - 0,29 * x_3^2 - 0,002 * x_1^2 + 0,98 * x_1 * x_2 + +2,8 * x_2 ** x_3$
5	sq combi r	0,026	0,871	$y = 358151,48 + 1,38 * x_1 - 0,01 * x_1^2 + ?,01 * x_1 * x_2 * x_3$
6	sq combi c	0,009	0.981	$y = 292416 + 0.256097 * x_1 * x_3 - 0.198123 * x_1 * x_5 + 0.704721 * x_2 * x_4$

Table 4. Statistical Indicators of the Kolmogorov-Smirnov Criterion for Residues of Projections Obtained by the Polynomial Models

Model	Lin m	Lin r	Lin c	Sq m	Sq r	sq c	Sq combi m	Sq combi r	Sq combi m
Parameter of normal distribution									
Average value	6.89	-5,31	7,61	6,96	3,69	2,07	3,48	2,13	-3,31
Standard deviation	1.17	1,62	1,28	1,95	8,9	1,28	6,66	6,01	7,23
Extreme differences									
Absolute	0.217	0,155	0,247	0,159	0,199	0,117	0,119	0,095	0,162
Positive	0.217	0,155	0,247	0,159	0,199	0,072	0,119	0,095	0,082
Negative	-0.135	-0,109	-0,175	-0,116	-0,151	-0,117	-0,072	-0,054	-0,162
Z Kolmogorov-Smirnov	1.061	0,759	1,209	0,780	0,975	0,575	0,583	0,467	0,796
Static value (2-sided)	0.210	0,612	0,108	0,578	0,297	0,896	0,886	0,981	0,551
Form of distribution	Normal	Normal	Normal	Normal	Normal	Normal	Normal	Normal	Normal

Table 5. Statistical Parameters of the Fisher's Criterion for Residues of Predictions Obtained Using the Polynomial Models

Model	Fisher's F- criterion	Significance F	F critical
lim m	51,36	0,0000048	2,40
lin r	55,5	3,1000642	2,40
lin c	17,39	0,000004811	2,40
sq m	51,36	7,87787080	2,40
sq r	102,79	0,025940648	2,40
sq c	6,20	0,000011887	2,40
sq combi m	43,56	0,000011887	2,40
sq combi r	45,44	9,46225677	2,40
sq combi c	51,36	0,00000481143	2,40

A deviation from the normal distribution is considered significant at the value of $p < 0.05$. In this case, nonparametric tests should be used for the respective variables. Figures 1 and 2 show the results of the accuracy of the constructed model.

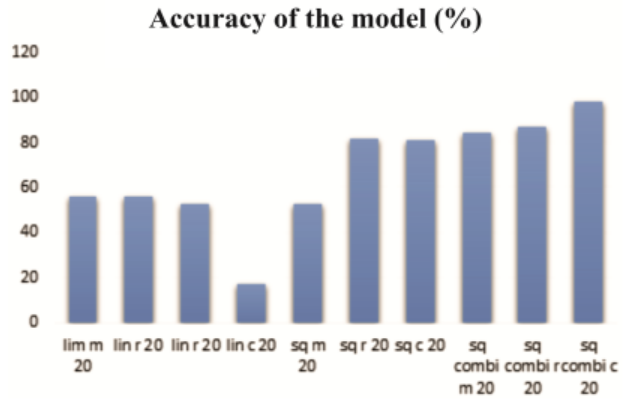

Fig. 1. An example of the graph.

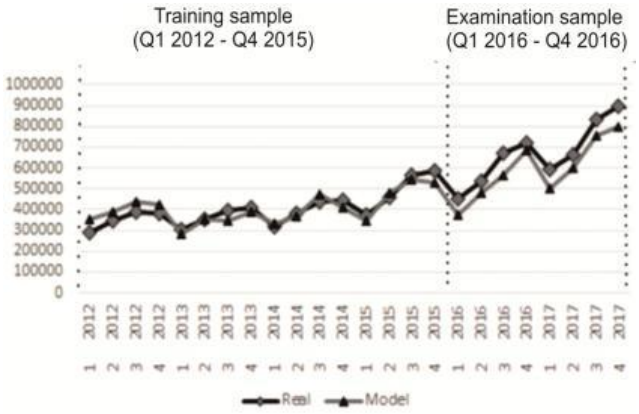

Fig. 2. Simulation results of the combined iterative-combinatorial algorithm

6. Conclusion

Based on the results of a comparative analysis, the advantages and disadvantages of the existing inductive modeling algorithms have been presented; the efficiency of applying the iterative GMDH algorithms for solving model-building problems has been established.

A complex assessment of the obtained models was carried out so that the determination coefficient was used to assess the accuracy of the obtained models; at the same time, the Kolmogorov-Smirnov criterion was used to assess their adequacy. The significance of the polynomial models was estimated according to Fisher's F-test. The results have shown that the combined iterative-combinatorial algorithms are the most efficient in all the evaluation criteria.

A comprehensive technique for comparative analysis of the iterative GMDH algorithms using computational experiments has been developed for studying the influence of the algorithms' pivot parameters on the performance indicators of the model's development process.

Based on a series of the provided numerical experiments, the combined iterative-combinatorial algorithms have demonstrated a higher level of accuracy (98.1%) compared to the multi-row iterative-combined algorithms (84.5%) and the relational iterative-combined (87.1%). The performed results' evaluations through the Kolmogorov-Smirnov criteria and the F-Fisher test emphasize that the obtained polynomial models are quite adequate and have an optimal complexity level. It was proved that the indicators used for forecasting GDP have a compatible multiplier effect on GDP, which indicates a nonlinear relationship between the utilized investment indicators and GDP. The resulting models have optimal complexity and can be interpreted easily by experts.

The complicated method of the GMDH iterative algorithms' efficiency numerical analysis allows a comprehensive study of the influence of the critical parameters of the compared algorithms on the leading indicators of the quality of the simulation results. The efficiency of using the methods of group accounting of arguments in forecasting GDP is shown.

It is proved that the indicators used in forecasting GDP have a compatible (synergistic) multiplicative effect on GDP, which indicates a nonlinear relationship between the used investment indicators and GDP.

In our further studies, we plan to use the developed methodology for solving problems in bioinformatics and immunoinformatics to determine the affinity of proteins and peptides, where the data are characterized by high dimensionality, nonlinearity, and noise.

References

[1] J. M. Keynes, "The General Theory of Employment," The Quarterly Journal of Economics, 1937.

[2] M. Zvi, J. Alan, and A. Kane, Bodie Essentials of Investments Fourth Edition, 2001.

[3] Ivakhnenko A.G. and Yurachkovsky A.A., Modelirovaniye slozhnykh sistem po eksperimental'nym dannym(rus) M.: Radio and communication, 1981.

[4] A.G. Ivakhnenko, Yu.P. Zaichenko and V.D. Dmitrov, Prinyatiye resheniya na osnove samoorganizatsii (rus), M.: Owls. radio, 1976.

[5] Zh. Hu, Ye. V. Bodyanskiy, O. K. Tyshchenko, V. O. Samitova, "Fuzzy Clustering Data Given in the Ordinal Scale", International Journal of Intelligent Systems and Applications (IJISA), Vol.9, No.1, pp.67-74, 2017.

[6] Zh. Hu, Ye. V. Bodyanskiy, O. K. Tyshchenko, V. O. Samitova, "Possibilistic Fuzzy Clustering for Categorical Data Arrays Based on Frequency Prototypes and Dissimilarity Measures", International Journal of Intelligent Systems and Applications (IJISA), Vol.9, No.5, pp.55-61, 2017.

[7] Zh. Hu, Ye. V. Bodyanskiy, O. K. Tyshchenko, O. O. Boiko, "An Evolving Cascade System Based on a Set of Neo - Fuzzy Nodes", International Journal of Intelligent Systems and Applications (IJISA), Vol.8, No.9, pp.1-7, 2016.

[8] A.G. Ivakhnenko, "Group method of data handling as competitor for the method of stochastic approximation", Soviet Automatic Control, No. 3, 1968.

[9] A.G. Ivakhnenko, Sistemy evristicheskoy samo-organizatsii v tekhnicheskoy kibernetike(rus), Kiev: Technika, 1971.

[10] A.G. Ivakhnenko, Dolgosrochnoye prognozirovaniye i upravleniye slozhnymi sistemami (rus), Kiev: Technika, 1975.

[11] A.G. Ivakhnenko, Induktivnyy metod samo-organizatsii modeley slozhnykh system(rus), K.: Naukova Dumka, 1982.

[12] O. Moroz O., V. Stepashko, "Hybrid sorting-out algorithm COMBI-GA with evolutionary growth of model complexity", Advances in Intelligent Systems and Computing II, Vol. 689. pp. 346–360, 2018.

[13] V. Stepashko, O. Bulgakova, V. Zosimov, "Construction and Research of the Generalized Iterative GMDH Algorithm with Active Neurons", Advances in Intelligent Systems and Computing II, Vol. 689, pp. 492–510, 2018.

[14] V. Lytvynenko, W. Wojcik, A. Fefelov, I. Lurie, N. Savina, M. Voronenko, O. Boskin, S. Smailova, "Hybrid Methods of GMDH-Neural Networks Synthesis and Training for Solving Problems of Time Series Forecasting", Vol. 1020, pp. 513-531, 2020.

[15] S. Yefimenko, "Building Vector Autoregressive Models Using COMBI GMDH with Recurrent-and-Parallel Computations", Advances in Intelligent Systems and Computing II, Vol. 689, pp. 601–613, 2018.

[16] V. Stepashko, O. Samoilenko, R. Voloschuk, "Informational Support of Managerial Decisions as a New Kind of Business Intelligence Systems", Computational Models for Business and Engineering Domains, pp. 269–279, 2014.

[17] A.G. Ivakhnenko and V.S. Stepashko, Pomekho-ustoychivost' modelirovaniya (rus), K.: Naukova Dumka, 1985.

[18] Moroz O., Stepashko V. Data reconstruction of seasonal changes of amylolytic microorganisms amount in copper polluted soils. Proc. of the 13th IEEE Intern. Conf. CSIT-2018 & International Workshop on Inductive Modeling. (Lviv, 11–14th of Sept., 2018), Lviv, 2018. P. 479–482.

[19] V. Stepashko, "Developments and Prospects of GMDH-Based Inductive Modeling", Advances in Intelligent Systems and Computing II, Vol. 689, pp. 474–491, 2018.

[20] V.S. Stepashko, O. Bulgakova, V. Zosimov, "Hibrydni alhorytmy samoorhanizatsiyi modeley dlya prohnozuvannya skladnykh protsesiv" (ukr), Inductive modeling of complex systems. Collection of papers, K: ISTC ITS, 2010.

[21] V.S. Stepashko, O. Bulgakova, V. Zosimov, "Construction and research of the generalized iterative GMDH algorithm with active neurons," Advances in Intelligent Systems and Computing, Vol. 689, pp. 492-510, 2018.

[22] V. Zosimov, V.S. Stepashko, O. Bulgakova, "In-ductive building of search results ranking models to enhance the relevance of the text information retrieval", Proceedings of the 26th Intern. Workshop "Database and Expert Systems Applications", Valencia, Spain, September 2015.

[23] V.S. Stepashko, "Samoorganizatsiya progno-ziruyushchikh modeley slozhnykh protsessov i system (rus)," Proceedings of the 15th All-Russian Nauch.-tech. Conf.: lectures on neuroinformatics, M.: NIIUU MEPhI, 2013.
[24] B.K. Svetalsky, P.I. Kovalchuk, "Mnogoryadnyy algoritm MGUA s selektsiyey pervichnykh argumentov" (rus), Automation, No. 4, 1979.
[25] Yu. Yurachkovsky, A.N. Gorshkov, Optimal'noye razbiyeniye iskhodnoy vyborki dannykh na obuchayushchuyu i proverochnuyu posledovatel'nosti na osnove analiza funktsii raspredeleniya kriteriya (rus), Automatics, No. 2, 1980.
[26] Reference on specific modeling programs [ed. A.G. Ivakhnenko], K.: Technique, 1980.
[27] V.M. Vysotsky, Pro naykrashchyy podil vkhidnykh danykh v alhorytmakh MHUA (ukr), Automation, No. 3, 1976.
[28] A.I. Kobzar, Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnykh rabotnikov (rus), M.: FIZMATLIT, 2006.
[29] N. Arora, J. R. Saini, "Estimation and Approximation Using Neuro-Fuzzy Systems", International Journal of Intelligent Systems and Applications (IJISA), Vol.8, No.6, pp.9-18, 2016.
[30] A. E. Khedr, S.E. Salama, N. Yaseen, "Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis", International Journal of Intelligent Systems and Applications (IJISA), Vol.9, No.7, pp.22-30, 2017.

Authors' Profiles

Volodymyr Lytvynenko Dr. Sc., Professor, Head of the Department of Informatics and Computer Science, Kherson National Technical University, Kherson, Ukraine. Area of scientific interests: Inductive modeling of complex systems, computer intelligence systems, development of hybrid algorithms, Bayesian networks

Olena Kryvoruchko Post-graduate student of the Department of Economic Theories, National University of Water and Environmental Engineering, Rivne, Ukraine. Area of scientific interests: Mathematical modeling of economic systems, estimation and forecasting of investment efficiency, Group method of data handling

Irina Lurie, Ph. D., Docent of the Department of Informatics and Computer Science, Kherson National Technical University, Kherson, Ukraine. Area of scientific interests: Inductive modeling of complex systems, computer intelligence systems, development of hybrid algorithms, Bayesian networks

Nataliia Savina, Dr.Sc., Professor, Vice-Rector for Research and International Affairs, National University of Water and Environmental Engineering, Rivne, Ukraine. Area of scientific interests: Mathematical modeling of economic systems, estimation and forecasting of investment efficiency, Group method of data handling

Oleksandr Naumov, Doctor of economics, professor, University of State Fiscal Service of Ukraine (Irpin), Research Institute of Fiscal Policy, Head of Department of Research of Problems of Tax and Customs Audit. Area of scientific interests: investment, economy of enterprises, project management

Mariia Voronenko, Ph. D., Docent of the Department of Informatics and Computer Science, Kherson National Technical University, Kherson, Ukraine. Area of scientific interests: combined iterative algorithms, GMDH

How to cite this paper: Volodymyr Lytvynenko, Olena Kryvoruchko, Irina Lurie, Nataliia Savina, Oleksandr Naumov, Mariia Voronenko, " Comparative Studies of Self-organizing Algorithms for Forecasting Economic Parameters", International Journal of Modern Education and Computer Science(IJMECS), Vol.12, No.6, pp. 1-15, 2020.DOI: 10.5815/ijmecs.2020.06.01

