Inductive Modeling of Amylolytic Microorganisms Quantity in Copper Polluted Soils

Olha Moroz, Volodymyr Stepashko

Department for Information Technologies of Inductive Modelling
International Research and Training Centre for Information Technologies and Systems of the NASU and MESU
Glushkov Avenue 40, Kyiv, 03680, Ukraine
olhahryhmoroz@gmail.com, stepashko@irtc.org.ua

Abstract: **The article presents the application results of the combinatorial-genetic algorithm COMBI-GA for modeling from experimental data of amylolytic microorganism quantity in a soil plot contaminated by copper. The constructed nonlinear mathematical models describe dependence of microorganisms concentration in soil from basic vital environmental factors and the copper concentration in the soil.**

Keywords: **GMDH, combinatorial algorithm COMBI, genetic algorithm GA, hybrid algorithm COMBI-GA, amylolytic microorganisms, soil, heavy metals, inductive modelling.**

I. Introduction

One of the main components of most ecosystems is soil, in which microorganisms play an important role in the evolution and formation of fertility. Anthropogenic pollution of the biosphere affects all living components of biogeocenoses, including soil microorganisms [1].

Modern agroecosystems are subject to the considerable technogenic influence resulting frequently in pollution of arable soils. Pollutants influence negatively on soil mikrobiotics, which causes the necessity to carry out long-term observations after its current status. In parallel with monitoring, there is the task of determination of critical deviations and prediction of the mikrobiotum state dependently on pollutants concentration in soil. Solving this task is possible only by formalization of monitoring data in the form of mathematical models [2].

Investigation of the effect of specific anthropogenic factors, such as heavy metals, on microbial community functioning is very important. Negative influence of heavy metals on microbial kenosis, soil and biological activity is well-known. In this connection, the organization and carrying out of regular control of soil state for the purpose of critical situation detecting and forecasting are actual.

Thus the dynamics of change of microorganism quantity in soil is studied under influence of different ecological factors. Modeling from observation data is the necessary condition of the ecological monitoring as it allows operative estimating current ecological situations and forecasting their development.

In this research we find models of dependence of microorganisms functioning from supporting weather conditions, this means kind of models, where inputs are hydrometeorological variables and outputs ecological indexes. Such models can be used both for restoring the omitted data and for forecasting the development of the controlled ecological processes on the basis of weather terms prediction. For construction of such models we use the Group Method of Data Handling (GMDH) as the most effective method for the analysis, modeling and forecasting of complex processes from experimental data under conditions of incompleteness of a priori information and short samples given.

To analyse of soil microbiotics of dark gray podzol soil (Kyiv region) polluted with heavy metals, the automated system of simulation ASTRID [3] with different algorithms of GMDH was initially used.

In [4] the application results of the hybrid combinatorial-genetic algorithm COMBI-GA [5, 6] for finding optimal linear models on the basis of observations of the change in the number of amylolytic microorganisms in the soil contaminated by copper are presented. But to solve prediction tasks we need more accurate approximation. So, in this paper new results of the application of the COMBI-GA algorithm for building optimal nonlinear models are presented.

Section II of this paper describes briefly task of modelling. Section III considers hybrid combinatorial-genetic GMDH algorithm COMBI-GA and their features. Section IV presents modelling results.

II. Task of Modeling

Experiments regarding functioning amylolytic microorganisms under copper contamination were carried out on small plots in deep-gray podzolic soil (Kyiv region). The The traditional chart [7] of experiments was used: several plots of the same soil type were selected for experiments and a plot remained as control non-contaminated one.

Model contamination of soil was carried out by the annual one time bringing in soil solutions of Cu_2^+ salts at the beginning of a vegetation season. The amount of the applied metal (computed as content of their ions) corresponds to contamination doses of 2 maximum permissible concentrations (MPC). The soil pieces for the analyses was taken during vegetation periods from 1993 to 1996 from the arable layer depth (0-20 cm) approximately in 2nd, 30th and 90th day after bringing of the metal salt. It was hence

received three measuring points during four years or 12 points together.

The amount of amylolytic microorganisms in the control and polluted soils was determined by the method of sowing of soil suspension on a nourishing medium consisting of a starch-ammonia agar.

Based on observations data of below listed variables, the linear mathematical models were built in [8] for description of the amylolytic bacteria quantity changes in control and polluted by the heavy metal soils.

As input (independent) variables (factors) for construction of models were used: concentration of mobile forms of Cu_2^+, decade average values of temperature, humidity of soil and air, and number of microorganisms in soil of control unpolluted plot. Quantity data of amylolytic microorganisms were output (dependent) variables in plots with model pollution of soil by copper salt. Based on the obtained data, the models of microorganisms number in soil were built.

For construction of model of changing quantity of amylolytic microorganisms such list of input variables was formed: x_1 – quantity of microorganisms in the control plot (millions in 1 g of dry soil); x_2 – concentration of copper (mg/cg soil); x_3 – number of days from the date of pollution; x_4 – current decade average temperature of air (oC); x_5 – previous decade average air temperature (oC); x_6 – current decade average humidity of air (%); x_7 – previous decade average humidity of air (%)

A definition of the inductive modelling problem in this task may be done as follows. Let us given: a data set of n observations after 7 inputs x_1, x_2, …, x_7 and one output y variables. The GMDH task is to find a model $y=f(x_1, x_2,\ldots, x_7, \theta)$ with minimum value of a given model quality criterion $C(f)$, where θ is unknown vector of model parameters. The optimal model is defined as $f^*=argmin_{\Phi}C(f)$, where Φ is a set of models of various complexity, $f\in\Phi$.

III. Hybrids GMDH-GA Algorithm

The genetic algorithm [9] is one of the meta-heuristic procedures of global optimization constructed as a result of generalization and simulation in artificial systems of such properties of living nature as natural selection, adaptability to changing environmental conditions, inheritance by offspring of vital properties from parents.

Since GA is based on the principles of biological evolution and genetics, biological terms are used actively (and sometimes incorrectly) to describe them. Here are some of these terms. *Individual* is the potential solution to the problem; *population* is a set of individuals; *offspring* is usually improved copy potential solution (father); *fitness* is usually a quality characteristic of the solution. *Chromosome* is encoded data structure of an individual in the form of an array of fixed lengths. In the simplest case it's a binary string of fixed length. The *gene* is an element of this array.

Formally, GA can be represented in such a way:

$$\mathrm{GA} = \{P_0, M, L, F, G, s\},$$

where $P_0=(a_1^0,\ldots,a_M^0)$ is an initial population; a_i^0 is an individual of this population treated as a candidate for the solution of the optimization problem presented in the form of a chromosome; M is the population size (integer number); L is the length of each chromosome of the population (integer number); F is a fitness function of an individual; G is a set of genetic operators; s is the algorithm stopping rule.

As input data for any GA initial population P_0, a finite set of chromosomes is used each of which represents a potential solution of the problem. Then the first population of offspring P_1 is formed from the parent chromosomes P_0 using some genetic operators, similarly the next population P_2 is formed from the population P_1 and so on. The process continues until the specified stopping rule of the algorithm will be satisfied.

An important feature of the GA work is that with each step the mean FF value of the current population improved and converges to the solution of the optimization problem.

The effectiveness of GA's work depends on the method of encoding genes, the composition of the initial population used by genetic operators, GA parameters, such as population size, number of chromosomes selected during selection and for crossover, probability of using genetic operators. The most important in GA are genetic operators especially the selection of which stores a certain amount of chromosomes with the best values of FF for each iteration of GA, and the operators of the creation of new offspring-chromosomes such as crossover and mutation. The crossover operator creates offspring by exchanging genetic material between the parent chromosomes, and the mutation operators by changing one chromosome in accordance with certain rules.

Formally, the hybrid of COMBI [10] and GA algorithm can be defined as follows:

$$\text{COMBI-GA} = \langle \mathbf{Z}, \mathbf{y}, \mathbf{f}, \mathbf{X}, D, CR, \mathbf{P}_0, H, M, \mathbf{G}, k, F \rangle,$$

where Z[$n\times r$] is the measurement matrix of input variables of an object, r and n are numbers of inputs and measurements respectively; $\mathbf{y}[n\times 1]$ is vector of measurements of an output variable; $\mathbf{f}$ [$m\times 1$] is vector of a given m base functions of input variables; $\mathbf{X}[n\times m]$ is the measurement matrix of base set of arguments; D is a given rule of dividing matrix $\boldsymbol{X}[n\times m]$ and vector $\boldsymbol{y}[n\times 1]$ to testing A and checking B parts; CR is an external selection criterion (as fitness function) based on dividing the sample ($\boldsymbol{X}$, $\boldsymbol{y}$); $\mathbf{P_0}$ is a set of model structures of GA initial population consisting of binary chromosomes (encoded structure of partial models); H is size of initial population of models, $H<m$; M is size of any next population, $M>H$; $\mathbf{G}$ is set of genetic operators;k is stopping rule of GA; F is number of best partial models (freedom of choice) monitored during all iterations of the algorithm, $1< F \le H$.

This algorithm consists of the following steps:

Step 1. Calculating the matrix of the base set of arguments $\boldsymbol{X}[n\times m]$ using the input matrix $\boldsymbol{Z}$ and the vector of base functions $\boldsymbol{f}$ and dividing it and the output vector of measurements $\boldsymbol{y}[n\times 1]$ according to the rule D in testing $\boldsymbol{X}_A[n_A\times m]$ and checking $\boldsymbol{X}_B[n_B\times m]$ submatrices ($n_A + n_B = n$). Obviously, in the case of linear polynomial, matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are identical ($m = r$).

Step 2. Random generating the initial population $\boldsymbol{P_0}$ of the genetic algorithm.

Step 3. Calculating the coefficients of each partial model by LSM or another method using the training matrix of base

arguments $\boldsymbol{X_A}$ and output vector $\boldsymbol{y_A}$.

Step 4. Calculating the value of an external criterion *CR* (as the GA fitness function) for each partial model using the checking matrix $\boldsymbol{X_B}$ and output $\boldsymbol{y_B}$.

Step 5. Forming the current population of partial models (chromosomes) of the size *H* with better criterion values to form the next offspring. In addition, selection the best *F* partial models that are potential solutions of the task of model building.

Step 6. Forming new population of *M* individuals applying genetic operators of crossover and mutation to individuals of the current population.

Step 7. Checking a given GA stopping rule. If it is satisfied, then go to step 8, otherwise go to step 3.

Step 8. Choosing *F* best models from the current population of the size *H*.

Step 9. The end.

IV. Modeling Results

Based on experimental data, models of quantity of amylolytic microorganisms were built in the control as well as in the copper polluted soils. In all cases we use the COMBI-GA algorithm with the following division of all data sample (12 points of observation during 4 years for the vegetation period 1993-1996): 6 points (2 years) as training set *A*, 3 points (1 year) as checking set *B*, and 3 points (1 year) as validation set *C*.

Results in the class of linear models. In case of linear modelling, the quantity of amylolytic in a control soil $Y_{contr}= x_1$ is described by the model obtained in [4]:

$$Y_{contr} = 0{,}2136x_4 - 0{,}7149x_5 + 0{,}5412x_7 .$$

Models for the quantity of microorganisms in polluted soil were built taking into account the quantity of amylolytic microorganisms at the observation of the control soil x_1.

The linear model for quantity of amylolytic microorganisms in the copper polluted soil [4]:

$$Y = 0{,}743x_1 + 1{,}6516x_2 - 0{,}9182x_5 - 0{,}2845x_7$$

Proper graphs of experimental and model data are given on Fig. 2. The characteristics of accuracy of these models are presented in the table below. This accuracy level is insufficient for quality monitoring needs. That is why we decide to build more complex nonlinear (polynomial) models. The results of this stage of modeling are presented below.

Results in the class of nonlinear models. In the case of nonlinear modelling, the quantity of amylolytics in a control soil is described by the model:

$$x_1 = -0.19297x_5 - 2.0976x_7 + 0.130x_4x_5 + 0.111x_5x_7$$

Fig. 3 shows graphs of measured and modeled data .

The model of dynamics of changing the amylolytic microorganisms in soil polluted by copper:

$$Y = 1{,}1759x_5 + 0{,}2159x_7 + 0{,}1638x_1x_5 - 0{,}0261x_2x_7 - 0{,}1252x_4x_5$$

The characteristics of all obtained models quality are presented in the table calculated according to next formulas:

$$\mathrm{MSE} = \sqrt{\frac{1}{12}\sum_{i=1}^{12}(x_i - \bar{x})^2}\ ,\ AR_B = \| y_B - X_B\hat{\theta}_A \|^2 .$$

The designation "Valid. err." means error on the independent validation set *C* calculated like AR_B.

	Linear case		Nonlinear case	
	Measured	Predicted	Measured	Predicted
MSE	0, 738	0,814	0,124	0,138
AR	0,138	0,214	0,058	0,061
Valid. err.	0, 120	0, 157	0, 094	0, 110

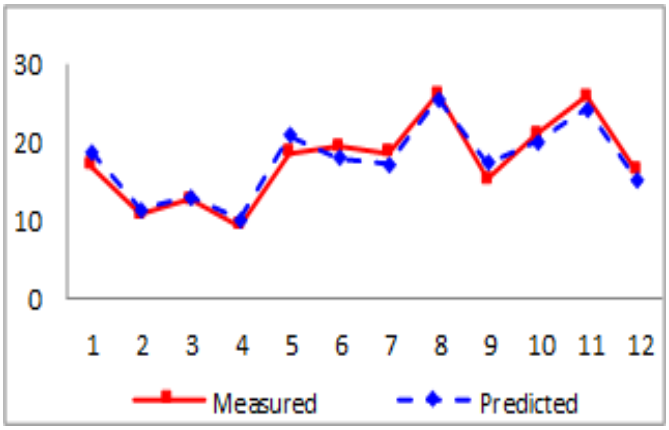

Fig. 1 Graphs of quantity change of amylolytic microorganisms on the control plot (linear model).

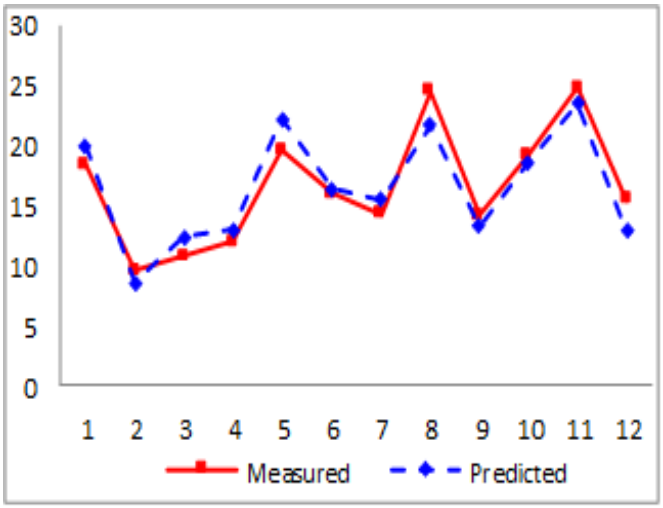

Fig. 2 Graphs of quantity change of amylolytic microorganisms on the plot polluted by copper (linear model).

Proper graphs of experimental and model data are given on Fig. 4. These graphs shows that in most points the data measured and predicted by the model coincide, that is the models adequately represents the change of microorganisms quantity. Three last three validation points on the graphs testify good results of models verification in the forecasting mode. Some distinctions can be accounted for by spatial heterogeneity of soil and other terms what could cause irregular variability of quantity.

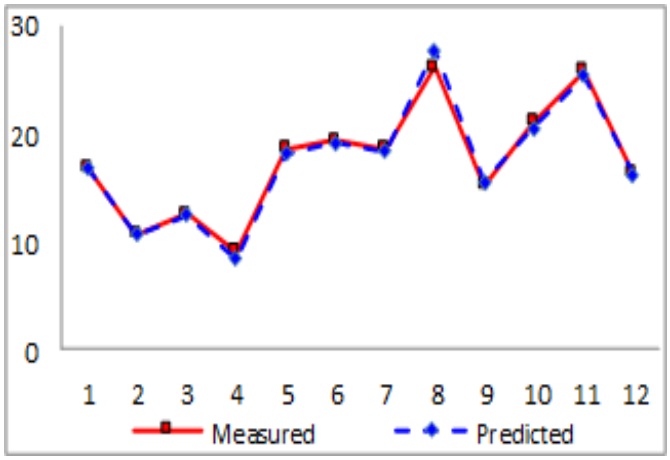

Fig. 3 Graphs of change of quantity of amylolytic microorganisms on the control plot (nonlinear model)

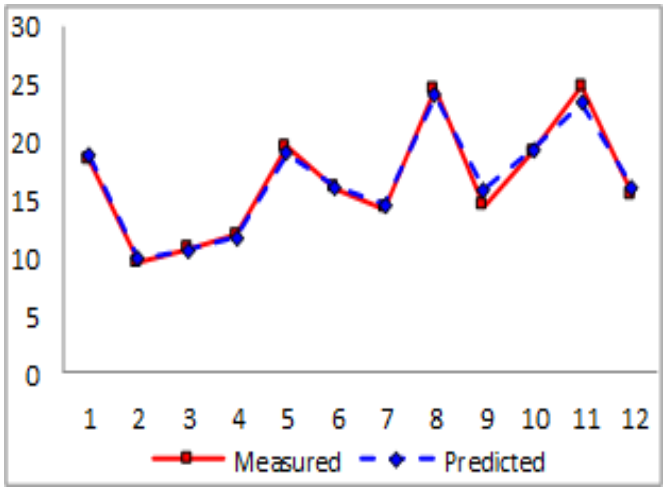

Fig. 4 Graphs of change of quantity of amylolytic microorganisms on the plot polluted by copper (nonlinear model)

As it is evident from the results, the functioning of amylolytic bacteria in soil is substantially influenced by the temperature and humidity of air.

In nonlinear case we obtain much more accurate results that can further help to effectively solve different ecological tasks based on microbial monitoring.

V. Conclusion

The carried out research manifests the possibility of formalization of the given ecological observations by construction of mathematical models. For the modeling of quantity of microorganisms in soil, application of inductive approach on the basis of GMDH is effective.

The obtained nonlinear models in a high degree coincide with experimental data that enables to use them in the system of the experiments for estimation of degree of soil contamination, renewal of intermediate or omitted data and operative forecasting the dynamics of microorganisms under various ecological conditions. Equally, these models will be helpful also for the data restoration with the purpose of obtaining the uniform series of ecological observations.

References

[1] K.I. Andreyuk, H.O. Iutynska, A.F. Antypchuk et al, "The functioning of soil microbial communities under conditions of anthropogenic load," K .: *Oberehy*, 2001, 240 p.

[2] H. G. Schlegel, "General microbiology," *7th edition, Cambridge University Press*, 1993, 655 p.

[3] V.S. Stepashko, Yu.V. Koppa, "Experience of the ASTRID system application for the modeling of economic processes from statistical data," *Cybernetics and computing technique*, vol. 117, pp.24-31, 1998. (in Russian)

[4] G. Iutynska, O.Moroz, "Inductive modeling of changes of amilolitic microorganisms on polished surface tuber," Inductive modeling of complex systems, *IRTC ITS NASU*, Kyiv, 2017, vol. 9, pp. 85-91.

[5] O. Moroz, V. Stepashko, "Hybrid Sorting-Out Algorithm COMBI-GA with Evolutionary Growth of Model Complexity," *Advances in Intelligent Systems and Computing II* / N. Shakhovska, V. Stepashko, Editors, AISC book series, Berlin: Springer Verlag, vol. 689, pp. 346-360, 2017.

[6] O.H. Moroz, "Sorting-Out GMDH algorithm with genetic search of optimal mode," *Control Systems and Machines, no. 6*, pp. 73-79, 2016. (In Russian)

[7] E.I. Andreyuk, G.A. Iutynska, Z.V. Petrusha, "Homeostasis of microbial of soils polluted by heavy metals," *Mikrobiol. Journ, vol 61, no. 6*, pp.15-21, 1991. (in Russian)

[8] A.G. Ivakhnenko, V.S. Stepashko, "Noise-Immunity of Modeling," Kiev: *Naukova Dumka*, 1985, 216 p. (In Russian)

[9] J. Holland, "Adaptation in natural and artificial systems, An introductory analysis with application to biology, control, and artificial intelligence," *University of Michigan, Computers, 1975*, 183 p.

[10] V.S. Stepashko, "Combinatorial Algorithm of the Group Method of Data Handling with Optimal Model Scanning Scheme," *Soviet Automatic Control, vol 14, no 3*, pp. 24-28, 1981.