DOI https://doi.org/10.15407/usim.2020.02.066
UDK 519.237.5

H. PIDNEBESNA, junior research scientist,
International Research and Training Centre of Information Technologies and Systems
of the NAS and MES of Ukraine,
Glushkov ave., 40, Kyiv, 03680, Ukraine,
pidnebesna@irtc.org.ua

BIOPRODUCTIVITY OF DNIEPER RESERVOIRS ANALYSIS BY INDUCTIVE METHODS

The task was to determine the factors that have the most significant influence on the state of water in the Dnieper reservoirs by constructing a model of dependence of the concentration of chlorophyll a in phytoplankton according to long-term observations in Kremenchug and Kakhovka reservoirs. The results of observations of the Institute of Sciences in 1976–1993. Various inductive methods were used to obtain a satisfactory result. Algorithms: linear regression of (LR), LASSO, combinatorial algorithm of (COMBI) GMDH and correlation-rating algorithm (CRA).

Keywords: *inductive modeling methods, correlation-rating algorithm CRA, combinatorial algorithm GMDH COMBI, LASSO, phytoplankton, chlorophyll "a" concentration.*

Introduction

A pressing issue in today's world is environmental security and maintaining ecological balance. Water resources are extremely important in solving this problem. Investigation of patterns of functioning of aquatic ecosystems, formation of their biological productivity and influence of natural factors on this process under conditions of anthropogenic loading are necessary for the development of scientifically methods of management and prediction of water quality [1 – 3].

The reservoir ecosystem is a complex system. Water quality in reservoirs depends on many factors. Therefore, it is necessary to identify the most influential ones by constructing models of their functioning. However, environmental modeling is a complicated process. Example of problematic issues are: the uniqueness of each body of water (rivers, lakes, reservoirs, etc.); problems of conducting system studies; small number of observations and errors of measurements of physicochemical parameters; incomplete knowledge of ecosystem functioning.

To study the functioning of ecosystems in the Dnieper reservoirs, the task was to determine the factors that have the most significant influence on the water status, by constructing a model of dependence of *chlorophyll "a"* concentration in phytoplankton according to long-term observations in Kremenchug and Kakhovka reservoirs. The results of observations for 1976 –1993 were provided by the Institute of Hydrobiology of the NAS of Ukraine. Unfortunately, the difficult economic situation in the country and the absence of a laboratory vessel makes it impossible to obtain up-to-date monitoring data. Therefore, it is even more important to identify the most important factors that affect the environmental status of reservoirs.

As mentioned above, the small amount of observational data and measurement errors make it difficult to solve the problem. Several methods

of inductive modeling were applied in the study. One of the well-known effective approaches to finding dependencies by short sample data is the Group Method of Data Handling (GMDH). In the research, one of the considered algorithms is the combinatorial algorithm of GMDH (algorithm of complete search). However, the results of such analysis are highly dependent on sampling. Given the small number of measurements, the results of the modeling at different divisions significantly differ. To overcome this problem, another GMDH sorting algorithm was used, a correlation algorithm with factor rating analysis that uses resampling, that is, a repeated sample split. In addition, one of the popular modern methods, the LASSO algorithm (Least absolute shrinkage and selection operator), was performed.

The biological basis of the problem

The primary autotrophic link of reservoirs is phytoplankton (algae of various systematic groups that inhabit the water column). By assimilating solar radiation and converting it into organic matter in the process of photosynthesis, it creates primary products. Due to this, there are other aquatic organisms living in water through the supply chain. The response of planktonic algae to the effects of natural and anthropogenic factors can be investigated by the content of chlorophyll a, the main photosynthetic pigment of phytoplankton, since it is an indicator of the level of vegetation and primary production. The increase in organic matter in the water is due to the intensive development of phytoplankton. To a certain extent, this creates the basis for the development of a forage base for fish and other aquatic organisms and helps to increase their numbers. When the "flowering" of water (mass development of blue-green algae – cyanobacteria) begins, the water quality deteriorates and the oxygen content decreases. Oxygen deficiency causes processes by which a number of substances are more actively released into the water. When the content of nitrogen, phosphorus, potassium in water exceeds a critical level, the life processes of aquatic organisms are accelerated, that is, the process of eutrophication intensifies. The negative consequences of this process are deterioration of water quality by organoleptic, hydrochemical and sanitary-microbiological parameters, accumulation of biologically active substances (vital secretions of algae and products of their decomposition), including toxic, allergenic and carcinogenic ones. Water gets a bad smell and taste, its transparency decreases, the color increases, the content of dissolved and suspended organic matter increases. It stimulates the development of saprophytic bacteria (including especially dangerous pathogens), aquatic fungi, dramatically exacerbating the epidemiological situation in water bodies [4, 5]. A high level of eutrophication causes fish and other hydrobionts to die. That is, from a certain point in time, eutrophication causes degradation of lake systems and reservoirs.

According to the world statistics, about 40–50% of cases of "flowering" in water accumulate significant concentrations of compounds that cause diseases of humans and animals [6].

The use of water bodies during "flowering" of water for recreational purposes and as sources of drinking water supply significantly increases the risk to human health. It also increases the risk of the formation of harmful substances also in the process of water treatment by existing technologies (eg, formation dioxins in the chlorination of water contaminated with phenolic compounds). This indicates the ecological and social significance of the problem of anthropogenic eutrophication of surface waters.

Formulation of task

Our *purpose* is to study the processes of ecosystem functioning in the Dnieper reservoirs to determine the factors that have the most significant impact on the water status. We do it by constructing a model of formation of chlorophyll a concentration in phytoplankton according to long-term observations in Kremenchuk and Kakhovka reservoirs.

The modeling was based on long-term observations of mean values of *chlorophyll "a"* and phytoplankton content per unit volume of water. The following physicochemical factors were measured for the study [7]:

• x_1 – total content of dissolved inorganic nitrogen, *mg N*/l;

• x_2 – content of dissolved inorganic phosphorus, *mg P*/l;

• x_3 – *N/P* – ratio of nitrogen content to phosphorus content, relative units;

• x_4 – water temperature, $t^\circ C$;

• x_5 – volume of water runoff, $m \times 10^9$/month;

• x_6 – total solar radiation, $MJ/m^2 \times$ month.

DESCRIPTION OF THE RESEARCH

Environmental processes forme complex system. The complex interaction of many factors gives them the character of a random process. Therefore, the calculations relating to the study of the laws of the functioning of water resources are probabilistic, statistical in nature. A large number of different methods have been developed, which makes the choice of the optimal method of finding patterns for a particular case one of the key problems in designing the statistical model of the process under study.

The linear regression approach remains one of the most powerful tools for analysis. Therefore, it is advisable to use regression analysis methods in the study of reservoir ecological processes. However, it can often produce erratic results. Regularization methods prevent unnecessary complexity while maintaining model adequacy. In addition, a small amount of observations requires methods that work satisfactorily on short samples. GMDH is one of these methods. Several inductive modeling methods were applied in the study: linear regression LR [8], LASSO [9], combinatorial algorithm COMBI GMDH [10], and correlation-ranking algorithm GMDH CRA [11].

Several methods for constructing a linear model have been applied to determine the most influential factors for modeling. In order to evaluate the adequacy of the obtained models, the coefficient of determination R^2 and the corresponding multiple correlation coefficient R are calculated, which reflects the degree of dependence of factor X and the dependent variable Y.

Note that the coefficient of determination R^2 is calculated as follows [12]

$$R^2 = 1 - \frac{ESS}{TSS},$$

$$ESS = \sum_{t=1}^{n} (y_{real} - y_{est})^2,$$

$$TSS = \sum_{t=1}^{n} (y_{real} - \overline{y}),$$

$$\overline{y} = \frac{1}{n} \sum_{t=1}^{n} y_i .$$

In the second stage of the study, modeling was performed using one of the modern methods of regression LASSO (Least absolute shrinkage and selection operator). It consists of the introduction of an additional component of regularization in the optimization functionality of the model, which often allows us to obtain a more stable solution [5]. This achieves a compromise between the regression error and the complexity of the model structure. During the minimization, some coefficients become equal to zero, which, in fact, determines the selection of informative features.

GMDH are well-known class of algorithms that work well on short data samples. Therefore, the next step in the study was to use the COMBI combinatorial algorithm [6]. The result of modeling using the combinatorial algorithm COMBI GMDH highly depends on the division of the data sample into training and testing subsample. Especially when the data sample is very short.

To overcome the problem of short data sampling, bootstrap procedure is often used, i.e. multiple uses of data sampling by different divisions into training and testing subsample. This technique is also used in another algorithm (CRA) that was used in the study. This method of selecting information arguments using correlation analysis belongs to the class of GMDH algorithms with an incomplete, directed search of models.

The use of correlation ranking of the sequence of arguments in enumeration algorithms was considered in [7]. We used a different method of defining of the informative factors based on pairwise correlation and ranking.

The general idea of the CRA method

In order to assess the informativeness of the factors according to the given table of observations, we propose to use a rating of factors. At the beginning, all the factors do not have rating points: ($V(x_1) = 0,..., V(x_n) = 0$). The rating is calculated using a procedure, which is repeated many times on different parions of the initial data sample (Fig. 1). Data sampling is randomly divided number of times into training and testing parts. The model is built on the training subsample in a certain way. The model parameters are calculated by OLS.

For the selected training subsample, each step determines whether the factor should be included in the model. The best model is selected using the criterion of regularity, calculated on a verified subsample of input arguments. If a factor is chosen at this stage, we say that it get a rating score.

The described procedure is repeated many times for different ways of splitting the sample. Thus we get a rating (the sum of rating points) factors. After completing the described procedure, the rating is analyzed. The division into "informative" and "non-informative" factors is made according to this analysis. Those with a higher rating are considered informative. This can be done in different ways: automatic clustering procedures, application of frequency criteria, or expert opinion. Thus, the structure of the final model is determined, which includes factors with a large number of rating points. The parameters of the model are determined using OLS by full the data sample.

The procedure for constructing the model is schematically presented in more detail in Fig. 2. Modeling is done by gradually complicating the model by adding one factor. Step by step for each of the factors (elements of the set *currentX*) it is determined whether it should be included in the model (be an element of the set *bestIndX*). The first partial model consists of only one factor. To determine this factor, the value of the pairwise correlation of each of the factors with the initial value is calculated. The factor with the highest correlation value is selected, provided that it is not random (according to the *pVal* test). The OLS

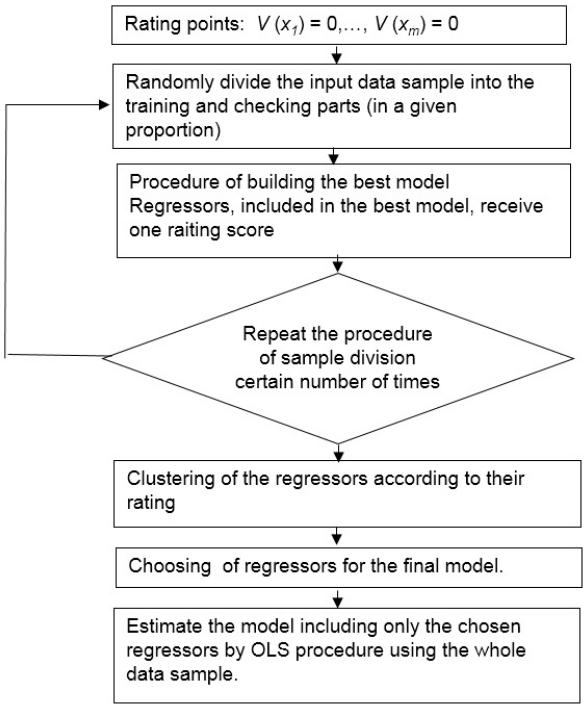

***Fig.* 1.** The scheme of the CRA method

builds a model, finds the coefficient, and calculates the value of the original variable. Next, the following factor is similarly selected. But we choose the one for which the pairwise correlation with the current residual model value y (Corr ($\Delta Y, x_i$)) has the largest value Considering all the factors, we obtain a set of partial models. The best model for the current partition is selected according to the minimum regularity criterion calculated on a verified subset of input arguments. Those factors that are included in the best model for the current sample devission receive a rating score.

The procedure is repeated a specified number of times (threshold). Thus, the rating of factors is calculated for further analysis.

Resultes of modeling

Kakhovka reservoir

The first phase of research was conducted for the Kakhovka Reservoir. Data on observations of the state of the Kakhovka reservoir for the period from 1980 to 1993 are presented in Table 1.

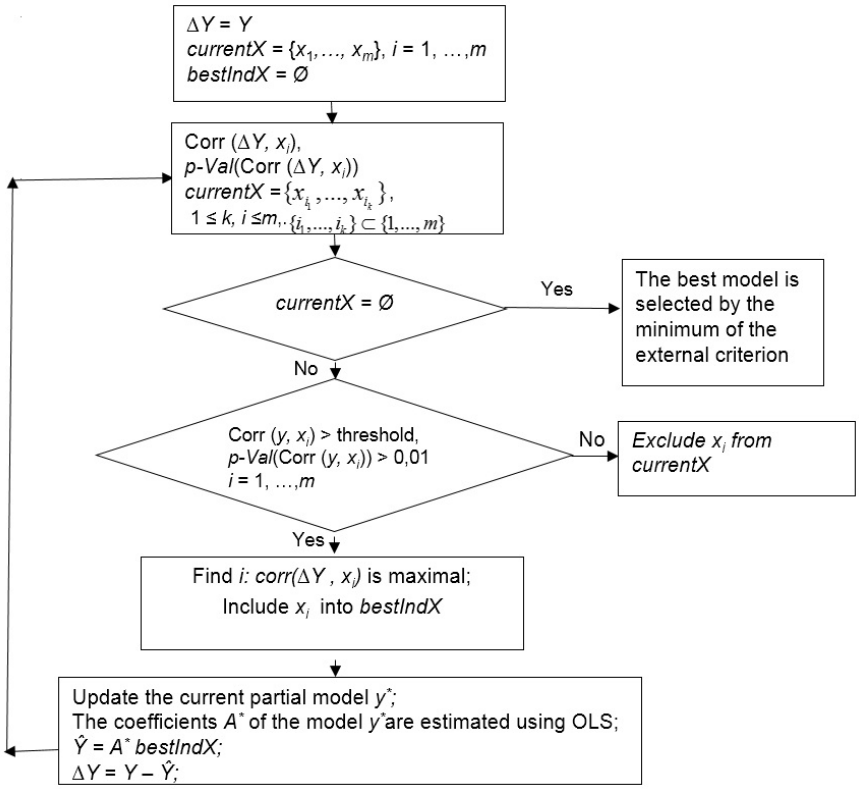

***Fig.* 2.** The scheme of the constructing model procedure by CRA

At first, linear regression models with different number of factors were constructed and correlation and determination coefficients were calculated for them [5]. The calculations showed that several models with different structures have virtually the same coefficient of determination (Table 2). This means that further complications of the model structure do not affect the value of the multiple correlation coefficient $R = 0{,}839$. Thus it is sufficient to use the model with the least number of factors. Further increase in the number of factors will mean unnecessary overhaul of the model.

Aften application the LASSO, we obtained:

$$y = f(x_1, x_3, x_6) = 5{,}16\,x_1 + 0{,}58x_3 + 0{,}02x_6.$$

The model has a value of multiple correlation coefficient $R = 0{,}803$.

To estimate the coefficients of the models by COMBI GMDH, three variants of the division of the data sample into training X_A and test X_B were considered (Table 3). The model with the hightest value R consider four factores.

The next one described CRA GMDH algorithm was used to construct a model. To identify informative factors, 50 random divisions of the sample were performed. The obtained rating of factors is presented in Fig.3.

The result model was

$$y = f(x_1, x_5) = 14{,}29x_1 + 3{,}04x_5,$$

for which the value of the correlation coefficient $R = 0{,}812$.

Thus, as a result of the analisys by various inductive methods for Kakhovka Reservoir, models were obtained (Table 4).

Table 1. ***Kakhovka reservoir input data***

#	Year, august	Mass of hlorophyll "a"	Total content of dissolved inorganic nitrogen, *mg N/l*	content of dissolved inorganic phosphorus, *mg P/l*;	*N/P*	water tempera-ture, *t °C*	volume of water runoff, $m10^9$/month;	total solar radiation, MJ/m^2 month
		Y	x_1	x_2	x_3	x_4	x_5	x_6
1	1980	58,1	3,224	0,065	49,60	19,4	5,080	552,7
2	1981	27,4	1,2312	0,076	16,20	21,1	3,200	703,8
3	1982	16,3	1,3108	0,058	22,60	24,0	4,570	620,7
4	1983	41,4	1,4348	0,068	21,10	21,0	2,540	606,8
5	1984	14,1	0,617	0,095	6,49	21,4	3,390	673,3
6	1985	47,8	1,065	0,036	29,58	23,5	4,620	645,1
7	1986	29,1	1,156	0,027	42,81	23,4	4,150	497,9
8	1987	17,7	0,081	0,026	3,12	20,8	2,830	724,3
9	1988	19,5	0,291	0,094	3,10	23,4	3,570	653,0
10	1989	11,09	0,406	0,026	15,62	22,7	4,260	753,0
11	1990	9,8	0,369	0,030	12,30	21,0	2,950	750,0
12	1991	17,9	0,578	0,123	4,70	29,0	3,360	693,0
13	1993	14,3	0,438	0,111	3,95	22,9	2,840	625,0

Table 2. ***Regression models***

Structure	Models	*R*
$y=f(x_1,x_3)$	$21{,}02 + 11{,}6*x_1 + 0{,}24*x_3$	0,831
$y=f(x_1,x_3,x_5,x_6)$	$23{,}05 + 11{,}41*x_1 + 0{,}33*x_3 - 2{,}84*x_5 - 0{,}006*x_6$	0,839
$y=f(x_1,x_3,x_4,x_5,x_6)$	$27{,}01 + 11{,}14*x_1 + 0{,}33*x3 - 0{,}14*x_4 - 2{,}66*x_5 - 0{,}008*x_6$	0,839
$y=f(x_1,x_2,x_3,x_5,x_6)$	$25{,}08 + 11{,}73*x_1 - 7{,}33*x_2 + 0{,}3*x_3 - 2{,}77*x_5 - 0{,}009*x_6$	0,839
$y=f(x_1,x_2,x_3,x_4,x_5,x_6)$	$29{,}53 + 11{,}45*x_1 - 7{,}88*x_2 + 0{,}29*x_3 - 0{,}16*x_4 - 2{,}5766*x_5 - 0{,}0111*x_6$	0,839

Table 3. ***The models by COMBI GMDH***

Structure	Model	*R*	Sample division (X_A/X_B)
$y=f(x_1)$	$y = 21{,}59x_1$	0,68	9/4
$y=f(x_1, x_2)$	$y2 = 18{,}23x_1 + 96{,}25x_2$	0,74	8/5
$y=f(x_1, x_4, x_5, x_6)$	$y3 = 16{,}1x_1 + 0{,}81x_4 - 1{,}56x_5 - 0{,}002x_6$	0,82	7/6

Table 4. ***The result of Kakhovka Reservoir modeling***

Algorithm	Model	*R*
LR	$y = f(x_1, x_3, x_5, x_6) = 11{,}41x_1 + 0{,}33x_3 - 2{,}84x_5 - 0{,}007x_6$	0,839
COMBI	$y = f(x_1, x_4, x_5, x_6) = 16{,}1x_1 + 0{,}81x_4 - 1{,}56x_5 - 0{,}002x_6$	0,820
CAR	$y = f(x_1, x_5) = 14{,}29x_1 + 3{,}04x_5$	0,812
LASSO	$y = f(x_1, x_3, x_6) = 5{,}155x_1 + 0{,}582x_3 + 0{,}015x_6$	0,803

Fig 3. Rating of arguments for Kakhovka Reservoir

Kremenchuk reservoir

In addition to Kakhovka, another of the Dnieper reservoirs was explored – Kremenchuk (from 1976 to 1992 yy) (table 5).

The first step was simulation using linear regression LR:

$$y = 2{,}12 * x_1 - 719{,}11 * x_2 - 9{,}17 * x_3 + \\ +1{,}83 * x_4 + 50{,}75 * x_5 + 0{,}02 * x_6.$$

The obtained model has a coefficient of determination $R^2 = 0{,}204$. Accordingly, the multiple correlation coefficient $R = 0{,}452$. But it contains all the factors, ie does not select the most important. And this does not meet the purpose of the study. Therefore, other modeling methods were used.

The LASSO method was also used to model the functioning of the Kremenchuk Reservoir. in our study. The obtained model has the form:

$$y = 0{,}10x_6,$$

and the coefficient of determination $R^2 = -0{,}0045$. A negative value of R^2 means that the resulting model approximates the value of the original variable worse than the mean value.

Table **5**. ***Input data of Kremenchuk Reservoir***

Year, august	Mass of *hlorophyll a", mkg*/l	Total content of dissolved inorganic nitrogen, mg *N* / l	Content of dissolved inorganic phosphorus, *mg P* / l;	*N* / *P*	Water temperature, *t* °*C*	Volume of water runoff, *m* 10^9 /month;	Total solar radiation, *MJ* / m^2 × month
	Y	x_1	x_2	x_3	x_4	x_5	x_6
1976	46,3	0,290	0,025	11,60	26,0	2,292	650,0
1977	86,8	0,408	0,040	10,20	18,6	3,210	560,5
1979	156,5	0,939	0,063	14,90	20,1	5,630	551,5
1980	54,2	0,862	0,069	12,49	19,0	3,320	507,9
1981	95,2	1,002	0,103	9,73	23,8	2,790	643,7
1982	30,4	1,238	0,048	25,79	22,0	5,470	506,5
1983	123,6	0,629	0,074	8,50	20,5	4,070	597,1
1984	17,0	0,690	0,060	11,50	20,2	2,450	530,3
1985	25,0	0,468	0,078	6,00	24,2	2,530	489,0
1986	40,1	1,067	0,084	12,70	23,6	3,970	629,1
1987	43,6	0,158	0,109	1,45	18,7	3,360	630,3
1988	39,0	0,529	0,129	4,10	22,9	3,930	649,0
1989	39,0	0,316	0,117	2,70	22,8	4,180	630,0
1990	74,3	0,459	0,062	7,40	23,1	2,964	688,0
1991	44,0	0,715	0,130	5,50	21,3	3,858	653,0
1992	55,4	0,269	0,084	3,20	24,8	2,399	649,0

The next step was modeling using COMBI GMDH. The obtained model has the form:

$y = 111{,}65x_1 - 872{,}48x_2 - 6{,}92x_3 + 124{,}93.$

The coefficient of determination $R^2 = -1{,}32$ calculated for it suggests that in this case the model also does not bring the initial variable close to the real values.

To construct the model according to the Kremenchuk Reservoir, a correlation algorithm with the calculation of the rating of CRA factors was also used. To obtain a rating 50 times, the sample was divided into educational and training (Fig. 4).

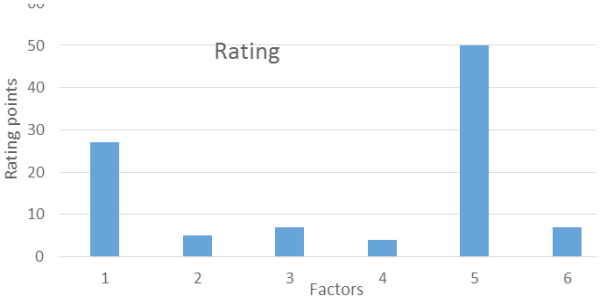

Fig. 4. Rating of arguments for Kremenchuk Reservoir

The model obtained using the CRA algorithm has the form $y = -0{,}96x_1 + 1{,}76x_5$. The multiple correlation indexis $R = 0{,}522$. According to the rating, the model includes two factors (nitrogen content and water runoff), which gives reason to believe that they play the most influential role in the formation of *chlorophyll "a"* in the phytoplankton of the Kremenchuk reservoir..

Conclusions

The paper presents a study of the functioning of two Dnieper reservoirs (Kakhovka and Kremenchuk) according to the statistics of long-term observations. The aim of the work was to build a mathematical model of the dependence of the concentration of *chlorophyll "a"* in phytoplankton on a number of different physicochemical factors. The difficulty of solving the problem is due to the small amount of observational data and measurement errors. Various inductive methods were used to obtain a satisfactory result. Modeling algorithms were performed: linear regression LR, LASSO, combinatorial algorithm GMDH COMBI and correlation algorithm for analysis of rating of CRA factors.

The analysis of the obtained results showed that in all models obtained by different methods, both reservoirs includes the variable x_1 – the mineral form of nitrogen. This is natural, when atrophying water bodies, the nitrogen content becomes a factor that limits the development of phytoplankton. This was observed in the reservoirs of the Dnieper. Also, three of the models obtained for the Kakhovka Reservoir and two models of the Kremenchuk Reservoir, which highest value of multiple correlation index, include an indicator of the volume of water runoff.

The study suggests that these factors (the mineral form of nitrogen and volume of water runoff) may be the most influential for the life of phytoplankton and determine the state of functioning of Kakhovka and Kremenchuk reservoirs. This applies to the named Dnieper reservoirs and the available sample of statistics for the summer season, which is characterized by intense “blooming” of water with blue-green algae.

REFERENCES

1. *Yang, X., Wu, X., Hao, H., Zhen-li, He.*, 2008. "Mechanisms and assessment of water eutrophication," J. Zhejiang Univ. Sci B, 9v(3), pp. 197–209
2. *Raike, A., Pietilainen, O.P., Rekolainen, S., et al.*, 2003. "Trends of phosphorus, nitrogen, and chlorophyll a concentrations in Finnish rivers and lakes in 1975–2000," The Science of the Total Environment, vol. 310, pp. 47–59.
3. *Worsfold, P.J., Monbet, P., Tappin, A.D., Fitzsimons, M.F., Stiles, D.A., McKelvie, I.D.*, 2008. "Characterisation and quantification of organic phosphorus and organic nitrogen components in aquatic systems: a review," Anal. Chim. Acta. Epub, 624 (1), pp. 37–58.
4. *Elser, J.J., Bracken, M.E., Cleland, E.E., Gruner, D.S., Harpole, W.S., Hillebrand, H., Ngai, J.T., Seabloom, E.W., Shurin, J.B., Smith, J.E.*, 2007. "Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems," Ecol. Lett., 10 (12), pp. 1135–1142.
5. *Kureishevich, A.V. Nezbrytskaya, I.N., Stanislavchuk, A.V.*, 2016. "Activity of Antioxidant Enzymes of Cyanoprokaryota and Green Microalgae Cultures under Different Temperatures Conditions", International Journal on Algae, 18 (2), pp. 169–177.
6. "Toxic Cyanobacteria in Water: A guide to their public health consequences, monitoring and management". URL: https://www.who.int/water_sanitation_health/resourcesquality/toxcyanbegin.pdf [Accessed: 26.02.2020].
7. *Kureishevich, A.V., Pidnebesna, H.A., Stepashko, V.S.*, 2015. "Analysis of the dependence of chlorophyll a content in the phytoplankton of the Kakhovka reservoir on the combined effect of factors by means of multiple regression," Inductive modeling of complex systems, Collected papers, Kyiv: IRTC ITS NASU, 7, pp. 147–153. (In Ukrainian).
8. *Meloun, M., Militk, J.,* 2011. Statistical Data Analysis. A Practical Guide. Woodhead Publishing India.
9. *Tibshirani, R.*, "Regression Shrinkage and Selection via the LASSO," Journal of the Royal Statistical Society, Series B (Methodological), 1996, vol. 58, no. 1, pp. 267–288.
10. *Stepashko, V.,* 2018. "Developments and Prospects of GMDH-Based Inductive Modeling, In: Advances in Intelligent Systems and Computing II" / N. Shakhovska, V. Stepashko, Editors, AISC book series, Cham: Springer, 2018, vol. 689, pp. 474–491.
11. *Pidnebesna, H.A.*, 2019. "Correlation-Based Sorting Algorithm of Inductive Modeling using Argument Rating," Proceedings of the IEEE 14th Internnational Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT–2019), Lviv, Ukraine: LNPU, September 17–20, vol. 1, pp. 211–214.
12. *Rawlings, J.O., Sastry, G.P., Dickey, D.A.*, 1998. Applied Regression Analysis: A Research Tool, Second Edition, Springer.

Received 20.03.2020

Г.А.Піднебесна, молодший науковий співробітник,
Міжнародний науково-учбовий центр
інформаційних технологій та систем НАН та МОН України,
просп. Академіка Глушкова, 40, Київ, 03187, Україна,
pidnebesna@ukr.net

МОДЕЛЮВАННЯ БІОПРОДУКТИВНОСТІ ДНІПРОВСЬКИХ ВОДОСХОВИЩ ІНДУКТИВНИМИ МЕТОДАМИ

Вступ. Дослідження закономірностей функціонування водних екосистем, впливу природних чинників на процес формування біологічної продуктивності водойм в умовах антропогенного навантаження відіграють надзвичайно важливе значення для оцінки екологічної безпеки та підтримки екологічної рівноваги, і, як наслідок, для розробки науково обґрунтованих методів управління і прогнозування якості води.

Мета. Була поставлена задача визначення факторів, які мають найвагоміший вплив на стан води в Дніпровських водосховищах, шляхом побудови моделі залежності концентрації хлорофілу а у фітопланктоні за даними багаторічних спостережень в Кременчуцькому та Каховському водосховищах. Результати спостережень за 1976–1993 роки надані Інститутом гідробіології НАН України.

Методи. Малий обсяг даних спостережень та похибки вимірювань значно ускладнює розв'язання задачі. Для отримання задовільного результату було застосовано різні індуктивні методи. Проведено моделювання алгоритмами: лінійна регресія *LR*, *LASSO*, комбінаторний алгоритм МГУА *COMBI* та кореляційний алгоритм з аналізу рейтингу факторів *CRA*. Для оцінки адекватності отриманих моделей застосовано коефіцієнт детермінації R^2 та відповідний коефіцієнт множинної кореляції R.

Результати. Для Кременчуцького водосховища виявилось, що моделі, побудовані за допомогою *LASSO* та *COMBI* мають негативне значення коефіцієнту детермінації R^2, тобто недостатню адекватність.

Модель, отримана за допомогою лінійної регресії *LR,* має коефіцієнт детермінації $R^2 = 0{,}204$ (відповідно, множинної кореляції $R = 0{,}452$). Це означає, що модель має задовільну адекватність. Але при цьому має в своєму складі всі чинники, тобто не відбирає найвагоміші.

Модель, отримана за допомогою кореляційного алгоритму з розрахунком рейтингу регресорів *CRA*, має коефіцієнт детермінації $R^2 = 0{,}273$ (відповідно, множинної кореляції $R = 0{,}522$).

Для Каховського водосховища моделі, отримані застосованими методами, мають високий ступінь адекватності ($R_{LR} = 0{,}838$, $R_{COMBI} = 0{,}820$, $R_{CAR} = 0{,}812$, $R_{LASSO} = 0{,}803$). Крім того, моделі мають схожі структури.

Висновки. Аналіз отриманих результатів показав, що до складу всіх моделей, отриманих різними методами, для обох водосховищ входить фактор – мінеральна форма азоту. Це закономірно, при евтрофуванні водойм вміст азоту стає тим чинником, що лімітує розвиток фітопланктону. Також до складу трьох моделей з отриманих для Каховського водосховища та двох моделей Кременчуцького, які мають коефіцієнт множинної кореляції $R > 0{,}5$, входить показник об'єму стоку води.

Проведене дослідження дає підстави вважати, що ці фактори (мінеральна форма азоту та об'єм стоку води) можуть бути найвпливовішими для життєдіяльності фітопланктону і визначати стан функціонування Каховського та Кременчуцького водосховищ. Сказане стосується саме названих дніпровських водосховищ та наявної вибірки статистичних даних літнього сезону, який характеризується інтенсивним «цвітінням» води синьозеленими водоростями.

Ключові слова: *індуктивні методи моделювання, кореляційний алгоритм з розрахунком рейтингу регресорів CRA, комбінаторний алгоритм МГУА COMBI, LASSO, фітопланктон, концентрація хлорофілу "а".*

Г.А. Поднебесная, младший научный сотрндник,
Международный научно-учебный центр
информационных технологий и систем НАН и МОН Украини,
просп. Академика Глушкова, 40, Киев, 03187, Украина,
pidnebesna@ukr.net

МОДЕЛЮВАННЯ БІОПРОДУКТИВНОСТІ ДНІПРОВСЬКИХ ВОДОСХОВИЩ ІНДУКТИВНИМИ МЕТОДАМИ

Введение. Исследование закономерностей функционирования водных экосистем, влияния природных факторов на процесс формирования биологической продуктивности водоемов в условиях антропогенной нагрузки играют чрезвычайно важное значение для оценки экологической безопасности и поддержания экологического равновесия, и, как следствие, для разработки научно обоснованных методов управления и прогнозирования качества воды.

Цель. Была поставлена задача определения факторов, которые имеют весомое влияние на состояние воды в Днепровских водохранилищах, путем построения модели зависимости концентрации хлорофилла а в фитопланктоне по данным многолетних наблюдений в Кременчугском и Каховском водохранилищах. Результаты наблюдений за 1976–1993 годы предоставлены Институтом гидробиологии НАН Украины.

Методы. Малый объем данных наблюдений и погрешности измерений значительно усложняет решение задачи. Для получения удовлетворительного результата были применены разные индуктивные методы. Проведено моделирование такими методами как: линейная регрессия *LR*, *LASSO*, комбинаторный алгоритм МГУА *COMBI* и корреляционно- рейтинговый алгоритм МГУА *CRA*. Для оценки адекватности полученных моделей применен коэффициент детерминации R^2 и соответствующий коэффициент множественной корреляции R.

Результаты. Для Кременчугского водохранилища оказалось, что модели, построенные с помощью *LASSO* и *COMBI* имеют отрицательное значение коэффициента детерминации R^2, т. е. недостаточной адекватности.

Модель, полученная с помощью линейной регрессии *LR*, имеет коэффициент детерминации $R^2 = 0{,}204$ (соответственно, множественной корреляции $R = 0{,}452$). Это означает, что модель имеет удовлетворительную адекватность. Но при этом имеет в своем составе все факторы, т. е. не отбирает наиболее значимые.

Модель, полученная с помощью корреляционного алгоритма с расчетом рейтинга регрессоров *CRA*, имеет коэффициент детерминации $R^2 = 0{,}273$ (соответственно, множественной корреляции $R = 0{,}522$).

Для Каховского водохранилища модели, полученные примененными методами, имеют высокую степень адекватности ($R_{LR} = 0{,}838$, $R_{COMBI} = 0{,}820$, $R_{CRA} = 0{,}812$, $R_{LASSO} = 0{,}803$). Кроме того, модели имеют схожие структуры.

Выводы. Анализ полученных результатов показал, что во все модели, полученные различными методами, для обоих водохранилищ входит фактор – минеральная форма азота. Это закономерно, поскольку при эвтрофировании водоемов содержание азота становится фактором, лимитирующим развитие фитопланктона. Также в состав трех моделей из полученных для Каховского водохранилища и двух моделей Кременчугского, которые имеют коэффициент множественной корреляции $R < 0{,}5$, входит показатель объема стока воды.

Проведенное исследование дает основания полагать, что эти факторы (минеральная форма азота и объем стока воды) могут быть самыми влиятельными для жизнедеятельности фитопланктона и определять состояние функционирования Каховского и Кременчугского водохранилищ. Сказанное касается именно названных днепровских водохранилищ и имеющейся выборки статистических данных летнего сезона, который характеризуется интенсивным «цветением» воды синезелеными водорослями.

Ключевые слова: *индуктивные методы моделирования, корреляционный алгоритм с расчетом рейтинга регрессоров CRA, комбинаторный алгоритм МГУА COMBI, LASSO, фитопланктон, концентрация хлорофилла "а".*