*Original Research*

# Application of Neural Networks for the Prediction of Total Phosphorus Concentrations in Surface Waters

**J. Możejko*, R. Gniot**

Institute of Chemistry and Environmental Protection, Szczecin University of Technology,
Al. Piastów 42, 71-065 Szczecin, Poland,

*Received: September 12, 2007*
*Accepted December 4, 2007*

**Abstract**

This paper describes the application of artificial neural networks (ANNs) for the time series modeling of total phosphorous concentrations in the Odra River. Data from the monitoring site Police in the lower part of the Odra were used for training, validating and testing the models. Two models are proposed to prove the satisfactory forecast of phosphorus concentrations: a simpler one with a single input variable and a more complex one with 14 input variables. Both ANN models show a high ability to predict from the new data set. On the basis of sensitivity analysis the relationships between phosphorus concentrations and other water quality variables were established.

**Keywords:** neural network, Odra River, total phosphorus

## Introduction

It is known that phosphorus plays a major role in biological metabolism ecosystems [1-5]. The enrichment of surface waters with phosphorus can have a large and undesirable impact on their tropic state, usage and appearance. Predictive models of phosphorus concentrations in waters can provide decisions that support preventive and operational control of these events. However, predicting the behavior of nutrient-enriched water bodies is difficult because of the complex physical, chemical and biological processes involved. Therefore, in engineering study, time series models are often used to make rapid preliminary estimates of water quality changes. Due to high variance and the inherent non-linear relationship of the water quality time series, it is difficult to produce a reliable model with conventional modeling approaches. The continuous development of computing facilities has made the analysis of multivariate data much more customary. Recently, alternative methods of data analysis, like artificial neural networks (ANNs), have emerged as interesting tools for time series analysis. Compared to the conventional modeling approaches in water treatment, ANN modeling has a number of distinct advantages. ANNs require no a priori assumptions about the model in terms of mathematical relationships or distribution of data. The network simply learns from the sample data and generates a black-box--type relationship. Thus ANNs have the potential to discover useful models where domain knowledge of ecosystem processes is limited. The ANN modeling approach is fast and flexible. Even a complicated neural mode can be completed relatively quickly once the data are collected. In addition, if some changes in the treatment process are necessary, the network can be quickly adjusted to the new process through model retraining, in which the new data describing the new process are added to the network's learning procedure.

*e-mail: Janina.Mozejko@ps.pl

A typical neural network consists of a large number of elements called neurons or nodes. Each neuron is connected to other neurons by means of direct communication links, each with an associated weight. The weights represent information being used by the net to solve a problem. Neurons are arranged in a layered structure. The first layer is called the input layer because the external inputs are applied here. The last layer called the output layer because it is where the outputs are processed as well as extracted. The layers between the input and output layers are called the hidden layers (not directly accessible). There can be one or more hidden layers and the number of neurons in each layer is an important parameter of the network.

Currently, many types of neural networks are known, mutually different in architecture and in the way the weights are adjusted. The most popular type of neural network is the multiple-layer perceptron (MLP). This network has a simple interpretation as a form of input-output model, with the weights and thresholds (biases) the free parameters of the model. Such networks can model functions of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining function complexity. Important issues in MLP design include specification of the number of hidden layers and the number of units in these layers.

A radial basis function network (RBF) has a hidden layer of radial units, each modeling a Gaussian response surface. Since these functions are non-linear, it is not actually necessary to have more than one hidden layer to model any shape of function: sufficient radial units will always be enough to model any function.

Generalized regression (GRNN) and probabilistic (PNN) networks are variants of the radial basis function (RBF) network. Unlike the standard RBF, the weights of theses networks can be calculated analytically. PNNs are designed for classification tasks and GRNNs for regression.

In the PNN, there are at least three layers: input, radial, and output layers. The radial units are copied directly from the training data, one per case. Each models a Gaussian function centered at the training case. There is one output unit per class. Each is connected to all the radial units belonging to its class, with zero connections from all other radial units.

Generalized regression neural networks (GRNNs) work in a similar fashion to PNNs. As with the PNN, Gaussian kernel functions are located at each training case. Each case can be regarded as evidence that the response surface is a given height at that point in input space, with progressively decaying evidence in the immediate vicinity. The GRNN copies the training cases into the network to be used to estimate the response on new points. The output is estimated using a weighted average of the outputs of the training cases, where the weighting is related to the distance of the point from the point being estimated. The first hidden layer in the GRNN contains the radial units. A second hidden layer contains units which help to estimate the weighted average. Each output has a special unit assigned in this layer which forms the weighted sum for the corresponding output. To get the weighted average from the weighted sum, the weighted sum must be divided through by the sum of the weighting factors. A single special unit in the second layer calculates the latter value and then the output layer performs the actual divisions (using special "division" units). Hence, the second hidden layer always has exactly one more unit than the output layer. In regression problems, typically only a single output is estimated, and so the second hidden layer usually has two units.

The most widely used kind of neural network is the linear neural network. In neural network terms, a linear model is represented by a network having no hidden layers, but an output layer with fully linear units (that is, linear units with linear activation function). The weights correspond to the matrix, and the thresholds to the bias vector. When the network is executed, it effectively multiplies the input by the weights matrix and then adds the bias vector [16-17].

In time series problems, the objective is to predict ahead the value of a variable that varies in time, using previous values of that and/or other variables. Any type of network can be used for time series prediction. The network can also have any number of input and output variables. However, most commonly there is a single variable that is both the input and the output. One of the major issues in neural network forecasting is how much data are necessary for neural networks to capture the dynamic nature of the process in a time series [18]. There are two facets to this issue:

- how many lagged observations should be used as inputs to the neural network
- how much past observation to use in training the neural network.

Determining an appropriate sample size for model building is not necessarily an easy task. Although a larger sample size in the form of a longer time series is usually recommended in model development, empirical results suggest that longer time series do not always yield models that provide the best forecasting performance.

In recent years there have been some successful ANN applications in water resource engineering. For example, Huang and Foo [6] applied ANN for salinity forecasting. Scardi et al.[7-8] and Jeong et al. [9] used ANN for phytoplankton primary production modeling. Wilson and Recknagel [10] applied ANN to predict algal blooms. Several authors applied ANNs for eutrophication modeling (Karul et. al., [11], Walter et. al [12]. Kuo et. al. [13]. Successful applications of ANN for forecasting water colour and pH have also been reported (Zhang et. al. [14] and Moatar et. al. [15]).

In this paper, the artificial neural network (ANN) modeling technique is used to establish a model for forecasting total phosphorus concentrations (TP) in the lower part of the Odra River. The Odra is a transboundary river and one of the longest watercourses in the Baltic Sea catchment area. Its total length amounts to 854.3 km, of which 741.9 km is in Poland. The river has its source in the Czech Republic. Its middle reach constitutes the boundary between Poland and Germany before reaching the Baltic Sea via a lagoon north of the Polish city of Szczecin. The Odra transports considerable quantities of organic matter as well as a variety of organic and inorganic contaminants in both dissolved and particulate forms. These are not expected to directly enter the Baltic, but to be deposited first in the Szczecin Lagoon. High concentrations of nutrients and organic matter result in blooms of blue–green algae, especially in summer, and gives rise to eutrophication in this area. Phosphorus is a factor limiting growth of phytoplankton in the Odra ecosystem.

## Material and Methods

### Data

Phosphorus concentrations and other water quality parameters of the lower Odra River used in the study were measured at the monitoring site in Police between 1991 and 2005. Police is a small city located about 10 km north of Szczecin. The data were collected monthly by the National Inspection Board for Environmental Protection in Szczecin. All variables are characterized by strong seasonal fluctuations.

## Results

### Statistical Analysis of Data

Statistical analysis consisted of the determination of parametric (mean, standard deviation and coefficient of variation) and non-parametric (minimum, maximum, median and quartiles) statistical parameters for the annual sets of data [19]. All analyses were performed using the computer software Statistica 6.1. The results of analysis are shown in Table 1.

Large variations in total phosphorus concentrations were observed between samples (Table 1), with a coefficient of variation from 19 to 58%. Extremely high values were observed in the years 1991-1992 and 1996. From 1991 to 2004, the annual mean P contents in the Odra decreased remarkably. Comparing the mean and median concentrations provides an indication of distributions of sample concentrations. In most cases the mean concentrations of total phosphorus are higher than median concentrations, indicating non-normal distributions with values skewed toward lower values, with a few high-concentration occurrences. This kind of distribution is frequent in ecological data.

Box-and-whisker plots (Fig. 1) for the data obtained in particular months during the period 1991-2004 illustrate the strong seasonality for TP concentrations. Box upper and lower bounds represent the 25th and 75th percentiles (first and second quartiles). In the middle of the box is the 50th percentile (the median). The upper and lower values represent the minimum and maximum total phosphorus concentrations. The highest total phosphorus concentrations were recorded in August. The median of the data reaches its minimum in April.

### Artificial Neural Networks for Modeling Total Phosphorus Concentrations in the Odra

The Statistica Neural Networks computer software was used to create neural networks in our calculations.

Table 1. Statistical parameters of the annual set of total phosphorus concentrations.

| Year | Mean mgP/dm$^3$ | Standard deviation mgP/dm$^3$ | Coeff.of variation % | Minimum mgP/dm$^3$ | Maximum mgP/dm$^3$ | First quartile mgP/dm$^3$ | Median mgP/dm$^3$ | Third quartile mgP/dm$^3$ |
|---|---|---|---|---|---|---|---|---|
| 1991 | 0.46 | 0.16 | 35 | 0.24 | 0.95 | 0.34 | 0.47 | 0.53 |
| 1992 | 0.40 | 0.22 | 55 | 0.17 | 0.91 | 0.27 | 0.35 | 0.47 |
| 1993 | 0.39 | 0.15 | 38 | 0.22 | 0.78 | 0.29 | 0.35 | 0.45 |
| 1994 | 0.36 | 0.21 | 58 | 0.13 | 0.87 | 0.21 | 0.30 | 0.44 |
| 1995 | 0.26 | 0.05 | 19 | 0.19 | 0.37 | 0.23 | 0.26 | 0.29 |
| 1996 | 0.34 | 0.19 | 56 | 0.22 | 0.91 | 0.26 | 0.29 | 0.33 |
| 1997 | 0.26 | 0.05 | 19 | 0.18 | 0.36 | 0.22 | 0.25 | 0.30 |
| 1998 | 0.31 | 0.13 | 42 | 0.18 | 0.51 | 0.20 | 0.24 | 0.46 |
| 1999 | 0.26 | 0.06 | 23 | 0.18 | 0.35 | 0.21 | 0.27 | 0.30 |
| 2000 | 0.26 | 0.08 | 31 | 0.10 | 0.39 | 0.21 | 0.23 | 0.34 |
| 2001 | 0.23 | 0.06 | 26 | 0.14 | 0.36 | 0.19 | 0.22 | 0.26 |
| 2002 | 0.23 | 0.10 | 43 | 0.12 | 0.50 | 0.16 | 0.21 | 0.27 |
| 2003 | 0.25 | 0.12 | 48 | 0.13 | 0.56 | 0.17 | 0.20 | 0.31 |
| 2004 | 0.25 | 0.10 | 40 | 0.15 | 0.50 | 0.17 | 0.21 | 0.29 |

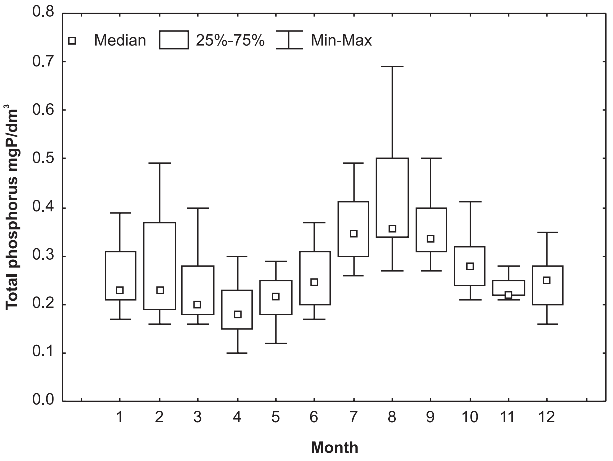

Fig.1. Box-and-whisker plots of total phosphorus concentrations in particular months.

For predicting total phosphorus concentrations in the lower Odra we built two models: a simple model with total phosphorus concentrations as input and output variable (Model A) and a model with total phosphorus and other available water quality parameters as inputs and total phosphorus as output variable (Model B). Table 2 shows the variables used for developing the models.

Five different types of ANN models were employed in this study: linear, Generalized Regression Neural Network – GRNN, Radial Basis Function – RBF, Multilayer Perceptron with one hidden layer – MLP(1) and Multilayer Perceptron with two hidden layers – MLP(2).

In Statistica Neural Networks, the network for time series prediction is configured by setting its Steps and Lookahead parameter [20]. The Steps parameter indicates how many cases should be fed in as inputs and the Lookahead parameter how far ahead the prediction should be made. Because this study focused on modeling of short-term forecasts (one month ahead), the Lookahead parameter was set to 1. The Step parameter was changed from 1 to 12.

Table 2. Neural network input and output variables (model B).

| Division | Variable | Units |
|---|---|---|
| Input | Water temperature ($T_W$) | $^oC$ |
| | Air temperature ($T_A$) | $^oC$ |
| | pH | |
| | Total Kjeldahl nitrogen ($TN_K$) | mgN/dm$^3$ |
| | Nitrate-N (N-$NO_3$) | mgN/dm$^3$ |
| | Nitrite–N (N-$NO_2$) | mgN/dm$^3$ |
| | Total phosphorus (TP) | mgP/dm$^3$ |
| | Orthophosphate (P-$PO_4$) | mgP/dm$^3$ |
| | Dissolved oxygen (DO) | mg$O_2$/dm$^3$ |
| | Biochemical oxygen demand ($BOD_5$) | mg$O_2$/dm$^3$ |
| | Chemical oxygen demand (COD) | mg$O_2$/dm$^3$ |
| | Sulphate concentration ($SO_4$) | mg$SO_4$/dm$^3$ |
| | Chloride concentration (Cl) | mgCl/dm$^3$ |
| | Total suspension concentration (TS) | mg/dm$^3$ |
| Output | Total phosphorus (TP) | mgP/dm$^3$ |

To develop the ANN models, input and output data of the years 1991–2004 (n=168) were used for training neural networks. The first few cases were only used as inputs for patterns. The remaining data set was randomly divided into three subsets: training (54% of cases), verification (23%) and test (23%). Observations in the verification set were used to perform an "independent check" of the network performance during training, to avoid over-fitting the data (i.e., to determine when to terminate training the network). The test set was not used in training at all, and was designed to give an independent assessment of the network's performance when an entire network design procedure is completed.

Network quality was estimated on the basis of following standard statistical parameters:

- Mean Error (ME) – Average error (residual between target and actual output values) of the output variable.
- Mean Absolute Error (MAE) – Average absolute error (difference between target and actual output values) of the output variable.
- Error S.D – Standard deviation of errors for the output variable.
- S.D. Ratio – The error: data standard deviation ratio. If this is 1.0 or higher, then the network is no better than a simple average model. A lower ratio indicates a better estimate.
- Correlation coefficient R – The standard Pearson-R correlation coefficient between the target and actual output values.

The optimal network was selected from the one which resulted in minimum error and the best correlation between model predictions and observations. After testing a few hundred different network topologies, the GRNN networks with two hidden layers were found to have the best performance with the best correlation. Fig. 2 shows the structures of the models. Both models use data from the previous season (Step=12 observations) as inputs. The GRNN network with single input (model A) has 106 neurons in the first hidden layer and 2 neurons in the second. Model B has 100 neurons in the first hidden layer and 2 neurons in the second.

As summarized in Table 3, the correlation coefficients between model predictions and observations is 0.803 for model A (with total phosphorus concentrations as input

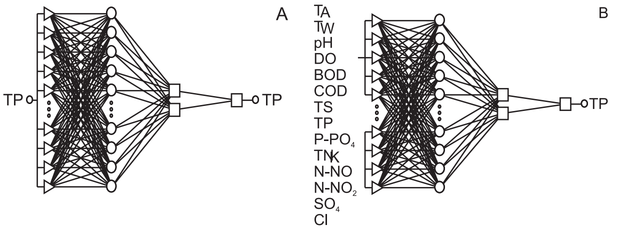

Fig.2. GRNN network structures for time-series modeling of total phosphorus in the Odra River.

Table 3. Performances of the selected GRNN networks.

| Parameter | Model A | | Model B | |
|---|---|---|---|---|
| | For data used for calibration | For unseen data from 2005 | For data used for calibration | For unseen data from 2005 |
| Mean Error (ME) | -0.0069 | -0.024 | 0.0005 | -0.020 |
| Error S.D | 0.083 | 0.044 | 0.051 | 0.038 |
| Mean Absolute Error (MAE) | 0.042 | 0.038 | 0.032 | 0.024 |
| S.D. Ratio | 0.602 | 0.605 | 0.368 | 0.516 |
| Correlation coefficient R | 0.803 | 0.799 | 0.931 | 0.865 |

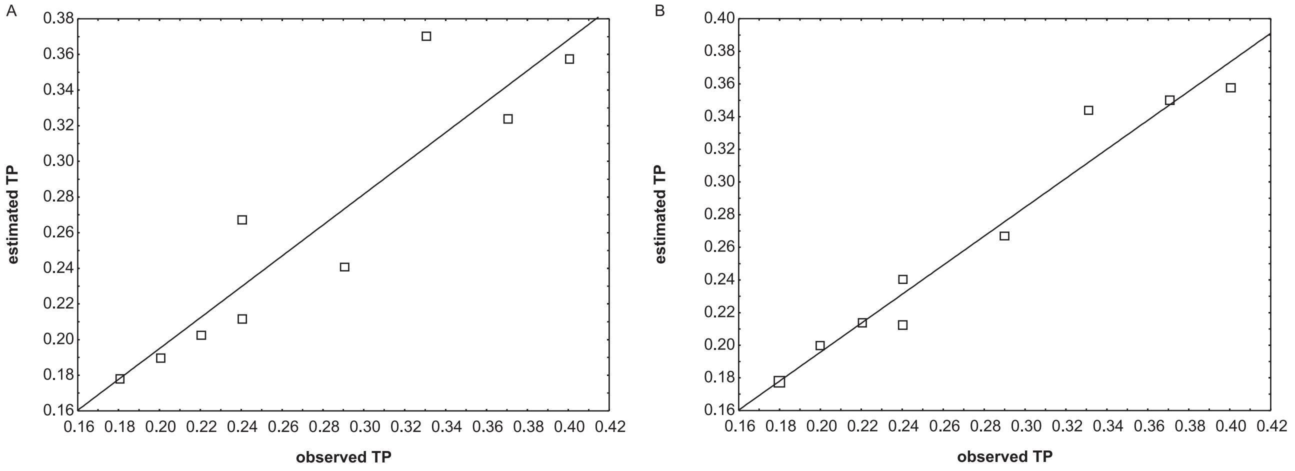

Fig.3 Scatterplots of actual versus predicted values of the TP concentrations in 2005 obtained using Models A and B.

and output variable) and 0.931 for model B (with 14 water quality parameters as inputs), and the MAE errors are 0.042 and 0.032 mgP/dm$^3$, respectively.

For model B, sensitivity analysis of input variables was performed to evaluate their relative significance in determining the forecast values. The analysis indicated that for this input data, beside previous TP concentrations, $BOD_5$ and $N\text{-}NO_3$ were the predominant variables for estimating the actual phosphorus concentrations.

In order to simulate a real-time forecasting situation, the inputs for the unseen data from January to September 2005 were presented to the selected best networks. The scatterplots of actual versus predicted values of the TP concentrations in 2005 obtained using Model A and B are shown in Fig. 3. The model's predictions matched reasonably with the observations. Corresponding values of R and MAE are given in Table 3. It can be seen that the model B predictions of TP concentrations in 2005 are quite good, with the R = 0.865 and the MAE = 0.024 mgP/dm$^3$. The predictions of model A are not quite as good, with the R = 0.799 and the MAE of 0.038 mgP/dm$^3$, but are still acceptable.

Comparisons of the time series of phosphorus from the model B predictions and observations are presented in Fig.4. The timing and magnitudes of the estimated output values compares well with the observed data.

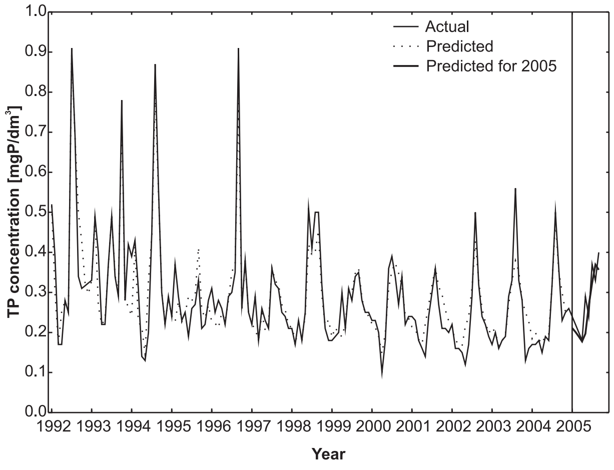

Fig.4. Comparisons of the time series of phosphorus from the model B predictions and observations.

## Conclusion

The Odra River system represents a complicated ecosystem with a distinct seasonal pattern of total phosphorus dynamics. The results from this study indicate that ANN models can be trained to provide satisfactory estimations of time series of total phosphorus concentrations. The specific ANN architecture suited well the high complexity and non-linearity of the river ecosystem. Two GRNN networks with two hidden layers were selected to have the best performance. These ANN models predicted phosphorus concentrations with good accuracy with time-delayed inputs. The best network found to forecast total phosphorus concentrations one month ahead was one with total phosphorus and other available water quality parameters as inputs and total phosphorus as output variable with the average absolute error of the output variable (MAE) of 0.032 mgP/dm$^3$. This complex model also performed satisfactorily over the range of the data used for calibration with the MAE of 0.024 mgP/dm$^3$. However, the model with total phosphorus concentrations as input and output variable can be a useful tool as well, because of its simplicity. The predictions of this model are also acceptable

Time series modeling of the Odra River by ANN proved to be suitable and useful for both prediction and elucidation of total phosphorus dynamics. The sensitivity analysis demonstrated the potential of ANN time series models to test hypothesis and elucidate causal relationships between environment-driving variables. The analysis indicated that for this input data, beside previous TP concentrations, $BOD_5$ and $N\text{-}NO_3$ were the predominant variables to estimate the actual phosphorus concentrations.

The results of this study have encouraged us to continue our modeling efforts by means of machine learning techniques to find the optimal model for short- and long-term forecasting water quality.

## References

1. MUSCUTT A.D., WITHERS P. J.A. The phosphorus content of rivers in England and Wales. Water Res. **30** (5), 1258, **1996**
2. SANSCHI P. H. Seasonality in nutrient concentrations in Galveston Bay. Marine Environ. Res. **40** (4), 337, **1995**
3. LARSEN S.E., KRONVANG B., WINDOLF J., SVENDSEN L.M. Trends in diffuse nutrient concentrations and loading in Denmark. Wat. Sci. Tech. **39** (12), 197, **1999**
4. JARVIE H. P., WHITTON B.A., NEAL C. Nitrogen and phosphorus in coast British rivers. Sci. Total Environ. **210/211**, 79, **1998**
5. LIQIANG XIE, PING XIE. Long-term (19956-1999) dynamics of phosphorus in a shallow, subtropical Chinese lake with the possible effects of cyanobacterial blooms. Water Res. **36**, 343, **2002**
6. HUANG W., FOO S. Neural network modeling of salinity variation in Apalachicola River. Water Res. **36**, 356, **2002**
7. SCARDI. M., HARDING L. W. JR. Developing an empirical model of phytoplankton primary production: a neural network study. Ecol. Model. **120**, 213, **1999**
8. SCARDI. M. Advances in neural network modeling of phytoplankton primary production. Ecol. Model. **146**, 33, **2001**
9. KWANG–SEUK JEONG, GEA-JAE JOO, HYUN-WOO KIM, KYONG HA, RECKNAGEL F. Prediction and elucidation of phytoplankton dynamics in the Nakdong River (Korea) by means of a recurrent artificial neural network. Ecol. Model. **146**, 115, **2001**
10. WILSON H., RECKNAGEL F. Towards a generic artificial neural network model for dynamic predictions of algal abundance in freshwater lakes. Ecol. Model. **146**, 69, **2001**
11. KARUL C., SOYUPAK S., CILESIZ A. F., AKBAY N., GERMEN E. Case studies on the use of neural network in eutrophication modeling. Ecol. Model. **134**, 145, **2000**
12. WALTER M., RECKNAGEL F., CARPENTER C., BORMANS M. Predicting eutrophication effect in the Burrinjuck Reservoir (Australia) by means of the deterministic model SALMO and the reccuren neural network model ANNA. Ecol. Model. **146**, 97, **2001**
13. JAN-TAI KUO, MING-HAN HSIEH, WU-SENG LUNG, NIAN SHE. Using artificial neural network for reservoir eutrophication prediction. Ecol. Model. **200**, 171, **2007**
14. QING ZHANG, STANLEY S. J. Forecasting raw- water quality parameters for the North Saskatchewan River by network modeling. Water Res. **31** (5) 2340, **1997**
15. MOATAR F., FESSANT F., POIREL A. pH modelling by neural networks. Application of control and validation data series in the Middle Loire River. Ecol. Model. **120**, 141, **1999**
16. DUCH W., KORBICZ J., RUTKOWSKI L., TADEUSIEWICZ R. Biocybernetics and biomedical engineering. Vol. 6. Neural networks. Akademicka Oficyna Wydawnicza EXIT, Warszawa **2000.** [In Polish]
17. WITKOWSKA D. Artificial neural networks in economic analysis. Wydawca „Menadżer", Łódź, **2000.** [In Polish]
18. RABUNAL, JUAN, R.(Editor). Artifical Neural Networks in Real-Life Applications. Hershey, PA, USA: Idea Group Publishing, **2005**
19. BERTHOUEX MAC P., BROWN L.C. Statistics for Environmental Engineers CRC. Press LLC, **2002**
20. Electronic Statistics Textbook. StatSoft. http://www.statsoft.com/textbook/stathome.html