NEURAL NETWORK APPROACH IN RISK ASSESSMENT OF PHOSPHORUS LOSS

Laima Berzina, Andrejs Zujevs, Ritvars Sudars
Latvia University of Agriculture
e-mail: l*aima.berzina@llu.lv; azujevs@llu.lv; ritvars.sudars@llu.lv*

Abstract

The main objective of this study is to demonstrate the use of artificial neural network (ANN) modeling tool to predict the risk of phosphorus (P) loss from the fields to nearest water body. The attention is drawn to ANN as an alternative approach to the P index calculation for prediction of the P losses. The specific tasks of this study were to determine risk classes of P loss by linking together source and transport factors that accelerate P losses and to evaluate ANN model performance for predicting risk classes via nutrient transport. ANN was trained with a Levenberg-Marquardt algorithm, and Scaled Conjugate Gradient algorithm was used to estimate the possible risk of P losses from agricultural land. Two small agricultural watersheds in Auce and Bauska were chosen to determine field parameters, and expert's evaluation was used for description of the risk classes' of P loss. Finally these values were used as inputs for the neural network model. The model was trained and validated by assessing its predictive performance on a testing set of data excluded from the training set. The research results highlight the capabilities of ANN to predict risk for a particular field and suggest that future research on application of other algorithms is required.

Key words: neural network, P loss prediction, risk assessment.

Introduction

The problem of phosphorus (P) loss in environmental science is well studied (Buczko and Kuchenbuch, 2007). Widely used approach for control of the P loss is designation of Phosphorus Index (P Index). The P Index (Sharpley et al., 2003; Heathwaite et al., 2000) is a tool that combines indicators of P source and of P transport as well as management factors to get qualitative risk characteristics of the site. P Index ranks fields according to risk of P loss in categories such as low, medium, high, and very high risk. General approach of P Index is to access the potential risk of P transport to surface waters from various fields by weighted parameters that promote risk of the P movement. Parameters values usually are rated (low = 0, medium = 2, high = 4, very high = 8) and rates for each level are summed. The original P Index uses a technique, which multiplies the site characteristics weighting factor with the phosphorus loss rating value to calculate the vulnerability of each site, but a numerous of modified techniques have been derived from the original version (Buczko and Kuchenbuch, 2007). Full understanding of the nutrient transport process is still difficult. Development of advanced tools is often restricted by large data input requirements and this limits the accuracy and reliability of many models. However, it is essential for good index to get appropriate index parameters ranks or weights and scale range boundaries for P index outcome in specific region (Kim et al., 2008). Since the estimation of nutrient losses fills an evident part of environmental studies, a number of computer-based models have been developed to enhance prediction of nutrient losses. Examples of computer-based techniques for studying of the water-quality-management systems include artificial intelligence, expert systems, neural networks, genetic algorithms, and other (Huang and Xia, 2001). Recently, one of the more popular and widely applied computational approaches is the artificial neural network approach. In comparison to traditional statistical methods, ANN is presented as a powerful data-modelling tool that is able to capture and represent complex input-output relationships (Govindaraju and Rao, 2000). Basically, the advantages of neural networks are ability to represent both linear and non linear relationships and to learn these relationships directly from data. For example, comparing ANNs with traditional multiple regression, ANN is found more flexible, hence more suitable and accurate for prediction (Talib et al., 2008). A set of inputs and output responses, representing a variety of simulation scenarios is sampled at random, and a particulate technique to allocate this set into training and testing subsets, is developed to obtain the best performance of network for the smallest error between observed and calculated data sets (Kim et al., 2006). Like biological neurons, ANN models contain multiple layers of simple computing nodes (neurons) that operate as summing devices. Weighted links interconnect these nodes. Each weight is adjusted when measured data are presented to the network during a 'training' process. The artificial

neuron which is given in Figure 1 has *N* input denoted as *uj*, for *j* = 1...N and each line connecting these inputs to the neuron is assigned a weight, which are denoted as *wj* respectively and corresponds to the connection between neurons. While a single artificial neuron may not be able to implement some functions, the problem is solved by connecting the outputs of some neurons as input to the others, so constituting a neural network (Gümrah et al., 2000). Successful training can result in an ANN model that performs tasks such as predicting an output value, classifying an object, approximating a function, and others (Kim, Gilley, 2008). Regarding variable prediction as one of the artificial neural network technology broad categories, it comes useful to test how accurately ANN learns to predict the value of an output variable (P loss risk class for a field) by giving input variable information (evaluation of P source and P transport factors that promote P loss from a field).

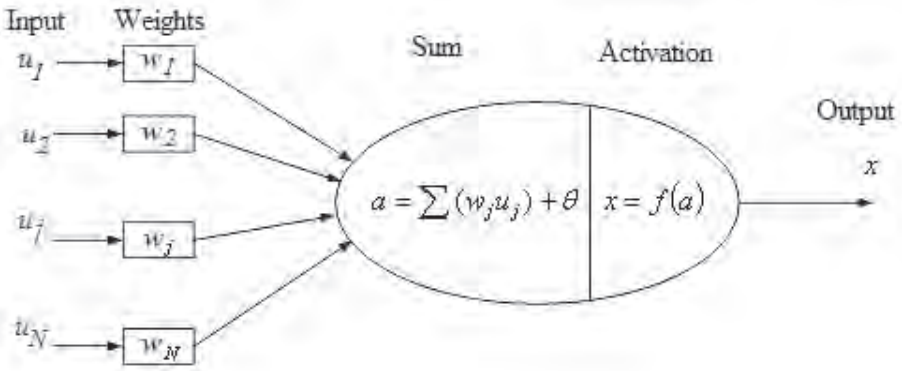

Figure 1. Artificial neuron and its structure (Gümrah et al., 2000).

The most widespread ANN design consists of an input layer, hidden layer(s), and an output layer of processing units (neurons). These are key components of artificial neural network models. The input layer introduces inputs to the network, or in other words, serves as an interface between the input variable data and the ANN model. Most of models also contain one, two or more hidden layers that transform inputs by adding them and applying linear or non-linear activation function(s) thus performing most of the calculations within the network (Nour et al., 2006). The output layer represents the response of the network. The goal of artificial neural network learning is to minimize the error between the models predicted value and the actual value of the output variable(s). According to Nour et al. (2006), the error minimization takes place by modifying the weights between neurons by a learning rule. As training progresses, the mean squared error (MSE) between the target output and the network output is calculated, and the weights are updated systematically. Weight adjustments are made based on an objective function that reduces MSE. Training proceeds until the prediction error is sufficiently small or until a maximum number of iterations have been reached (Nour et al., 2006; Baxter et al., 2002).

ANN modeling suggests that subject to data should be divided into three sets in the ratio 3:1:1 for training, testing, and validating the model, respectively. The training data set is used to adjust the connection weights. The validation data set measures network generalization to halt training when generalization stops improving, but testing data set measures of network performance during and after training, but does not affect the training. Advantages of artificial neural network modeling include handling of nonlinear relationships and providing of output variables in response to simultaneous and independent fluctuations of the values of model input variables Also data patterns with missing values of input variables can be incorporated into model building (Govindaraju and Rao, 2000). Besides, ANN does not require complicated programming, several user-friendly ANN software packages exist. Challenges of artificial neural network modeling show that model predictions are more accurate if only large and complete training data sets are used and extremes of possible values are present. Consequently, ANNs will almost never perfectly predict all values, so a reasonable error must be used for training and testing of networks (Govindaraju and Rao, 2000). The key to a good network is the appropriate training data; consequently artificial neural network models can be developed only where sufficient historical data for each of the process variables exists (Baxter et al., 2002).

Artificial neural networks (ANNs) have found wide applications in recent years. ANNs capabilities have been successfully used and proved through many water resource applications (Govindaraju and Rao, 2000). Studies of ANN include chemical composition of surface

waters and water quality prediction (Maier, Dandy, 1996), water quality modeling (Gümrah et al., 2000), prediction of eutrophication (Kuo et al., 2007), estimation of soil erosion and nutrient concentrations in runoff (Kim and Gilley, 2008), prediction of nutrient transport in runoff (Kim et al., 2006), phosphorus dynamics in small streams (Nour et al., 2006), and others (Talib et al., 2008). This study aims to test an ANN modeling tool that can predict agriculture field vulnerability to P loss risk.

Materials and Methods

Field tests for experimental data of P loss risk were conducted at Auce and Bauska (central part of Latvia). The individual risk indices were evaluated for 30 fields in Vecauce farm and 41 fields in Bauska farm. The following information was available for index calculation: soil P contents, land use (crop rotation), inputs of P in fertilizers and manures, soil type, field slope, and drainage. Data on land use and inputs of P were collected from farmers and field observations. Soil types, field slopes and location of drainage were derived from land amelioration maps developed by Department of Environment and Water Management (Latvia University of Agriculture). Knowledge on P input time and methods made the greatest uncertainty. Uncertainty in fertilizer application rates consequently contributed most to the output uncertainty.

The MathLab software was used to create neural network. The architecture of network is organized as a set of interconnected layers of artificial neurons – input, hidden and output layers (Fig. 2) – trained by Levenberg-Marquardt algorithm. Levenberg-Marquardt learning algorithm as improved Guass-Newton method is mentioned as one of the popular methods to speed up the learning process; other characteristic of this method is to deal with the small residual problems in learning (Chan, 1996). Detailed information about the algorithm is covered by R.M. Hristev (1998) and A.A. Suratgar et al. (2005).

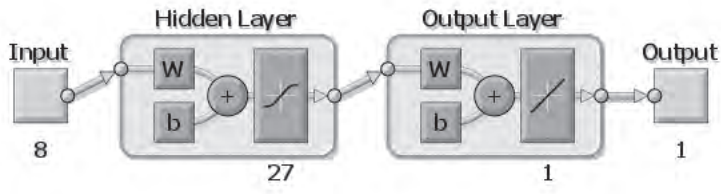

Figure 2. Arhitecture of the neural network used for P loss risk estimation.

Eight variables were selected as the inputs: results of soil P test, P fertilizer rate and P transport factors – erosion, runoff, leaching, drainage, surface run-off inlets and buffers for training of the neural network. All transport factors were calculated based on soil properties and evaluated by direct observations of fields. Details about P loss identification variables are covered by L. Berzina and A. Zujevs, 2008. P loss risk class was provided as an output variable. The input and output process elements (PEs) are fixed by the particular user application, but the number of hidden PEs must be specified. Hidden layer includes 27 hidden neurons that gave the best results. The weights (w) and biases (b) are iteratively adjusted during training to minimize network error. Networks were trained with experimental data that represent the characteristics of the process of risk of P loss identification. 71 data point was used in this study. For this dataset, each data points of P loss risks parameters were randomly divided into three subsets: a training set (70% of the total), a validation set (15% of the total), and a test set (15% of the total). Training data set was used for ANN prediction model development, validation set – for ANN performance evaluation, but the test set was used to guide the fitting of ANN.

Mean squared error algorithm was used for performance, and random algorithm was used for data division. The ANN modeling approach conducted in this study can be divided into three phases: data pre-processing, model building, and model evaluation.

Basically, the four main steps were taken in this forecasting study:

1) model design: choose a suitable model;
2) training: estimate the parameters of the model;
3) validation: test the model on data sets to determine its validity;
4) interpretation: explain results.

Results and Discussion

ANN was trained in 7 epochs that gave the best overall results for prediction of P loss. Model evaluation was based on the correlation coefficient and graphical examination of both measured and predicted values; however, residuals analysis and model stability also are suggested and can be used in further analysis of the

results (Nour et al., 2006). The training process is plotted in Figure 3. It shows on logarithmic scale the precision of response of the network to validation and test data sets explicated by mean square error. The graph displays that neural network is able to predict targets from training set with reasonable accuracy already at epochs 4 to 5. At epochs 6 to 7, the accuracy of prediction tends to be almost absolutely correct. Meanwhile the response to validation and test data sets reached a stable unchanging level of mean square error of 0.75 for validation and 1.20 for testing data sets, which is number of times greater in comparison to training samples. It can be explained with over-learning characteristic of AAN's, when the network adopts to all input vectors of training data only, while improvement in response to other data cannot be observed.

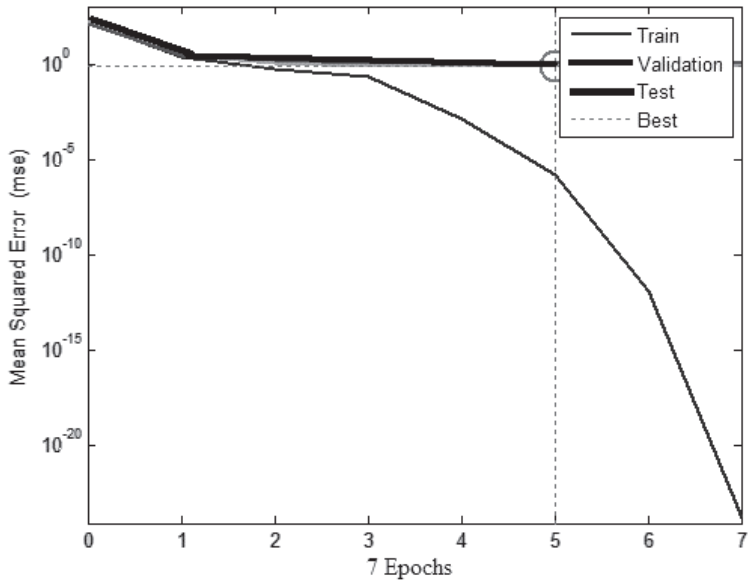

Figure 3. Training process of ANN.

The correlation of AAN response with expert evaluations in all data sets is shown in Figure 4. It also highlights that AAN used and trained in the study shows the strongest correlation with training data (R = 1). The correlation with validating and test data sets is also strong, respectively 0.96 and 0.89, but considering the mean square error for each data set mentioned above, the architecture and learning parameters of the network should be adjusted in order to lower it.

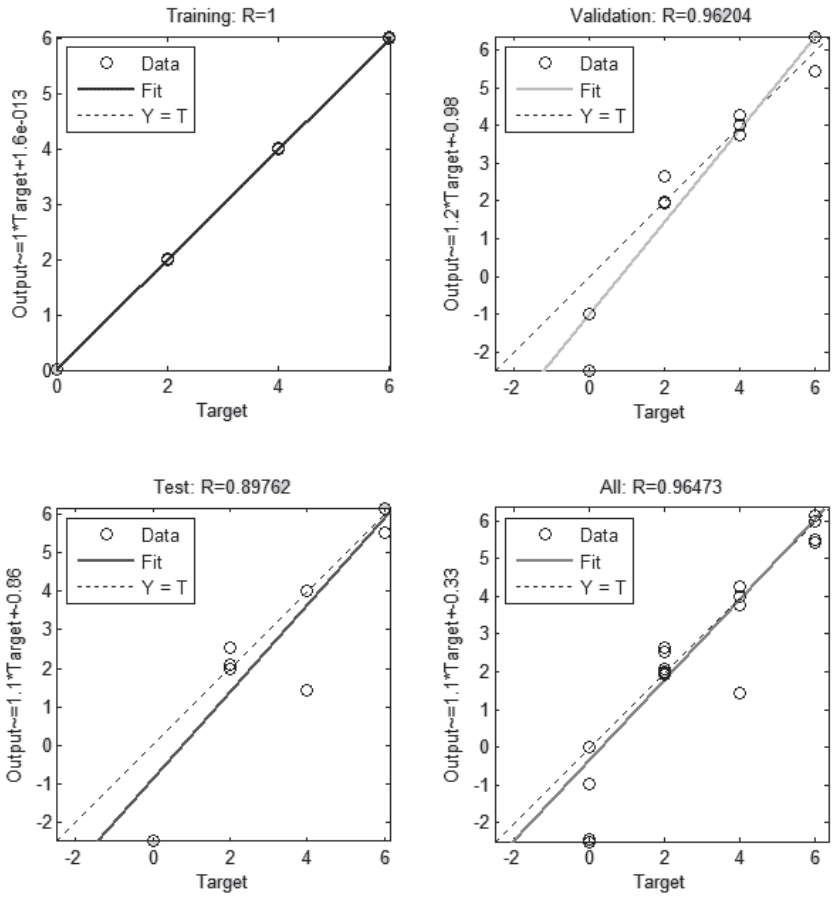

Figure 4. Correlation of expert's evaluations to AAN's predictions.

Studies have shown that a neural network with one hidden layer is capable with very high accuracy (Kim and Gilley, 2008) and this is consistent with the present study. Typically, the increased number of neurons enhanced the training-set performance. The testing-set performance increased whilst the additional neurons help to correctly predict outputs from inputs, and decreased when the network started to memorize the data due to too many neurons. However results indicate potential of network to predict P loss risk class, the truth of results still depends on expert judgment about output variable.

The network with two hidden layers was also trained with Scaled Conjugate Gradient algorithm described by M.T. Hagan and others (1996). The architecture of the network is shown in Figure 5: first layer includes 20 neurons, seconds 45 neurons. Input layer consists of 8 neurons, and output layer of one neuron. Consequently, the network structure is 8-20-45-1.

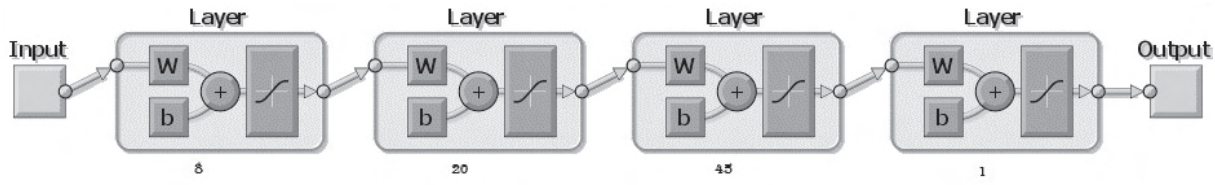

Figure 5. Comparing architecture of the neural network used for P loss risk estimation.

Conjugate Gradient algorithm network gave the best results from other 23 experimental networks and was chosen for ANN training. ANN was trained in 5 epochs, and Figure 6 displays that neural network is able to predict targets from training set with reasonable accuracy already at epoch 0.

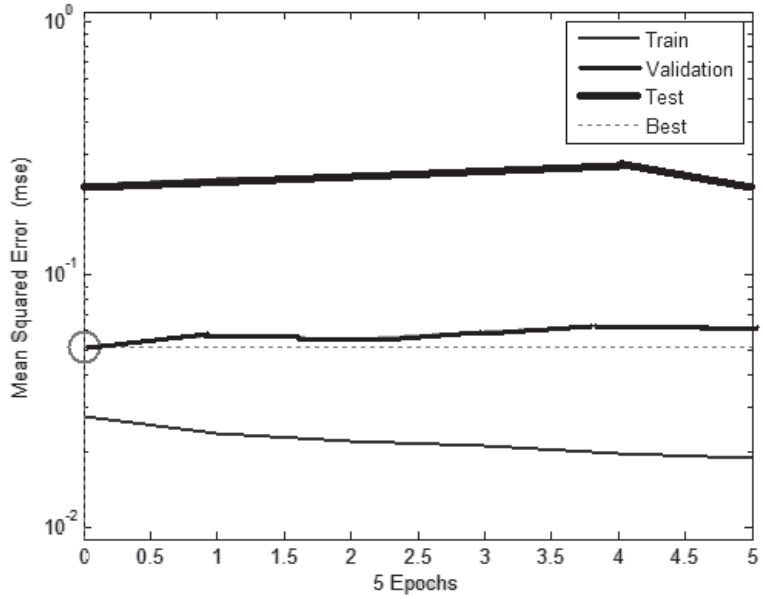

Figure 6. Training process of ANN with Conjugate Gradient algorithm.

In validation and testing of ANN, measured correlation coefficients between observed and predicted P loss risk classes were more than 0.99 for validation data and 0.97 for testing data. The maximum mean squared error for validation data set observed was 0.0276. Also several statistical methods can be used to solve a range of problems in forecasting and data classification. Since each statistical method uses different data assumptions, relationship between the variables being forecasted and the variables used to produce the forecast, as well as the distribution of forecast errors must be considered before applying statistical methods. As a result, there are certain instances where traditional statistical methods are unsuitable. ANN training algorithms help learn the structure of the data, consequently neural networks learn by example, witch is very useful when there is no idea of the functional relationship between the dependent and independent variables. The most evident advantage of ANN is the use of very sophisticated modelling techniques capable of modelling extremely complex functions, at the same time ANN requires fewer statistical assumptions. This is also reason why ANN could be valuable alternative approach to P Index modelling by considering assumption that P loss is extremely difficult to predict via complicated relationships intermediary factors that accelerate P loss. The basis of the power of the neural networks in P Index calculation is to let to define the input-output relationship functional form using training data.

Conclusions

ANN model with Levenberg-Marquardt training algorithm was developed and used for forecasting the risk class of P loss for agriculture fields. In all, 70% of data observed in field experiments in the central part of Latvia have been used for training, and 30% of data have been used for validation and testing of ANN performance. In validation and testing of ANN measured correlation coefficient between observed and predicted P loss risk classes was more than 0.96 for validation data and 0.89 for testing data, which shows the ability of ANN in acceptable forecasting of risk class for selected fields. The maximum mean squared error for validation data set was 0.75, and for testing data set was 1.2, witch is still acceptable for P risk classes' prediction that varies from 0 to 8 corresponding to good model performance. However, future research on the application of other algorithms is required by considering the amount of squared mean error, for example, the use of Conjugate Gradient algorithm that gave correlation coefficients between observed and predicted P loss risk classes with values 0.99 for validation data and 0.97 for testing data. The survey results confirm high capabilities of ANN to predict risk of P loss and suggest future research on application of other algorithms.

References

1. Baxter C.W., Stanley S.J., Zhang Q., Smith D.W. (2002) Developing artificial neural network models of water treatment processes: a guide for utilities. *Journal of Environmental Engineering Science*. 1, pp. 201-211.
2. Berzina L., Zujevs A. (2008) Design of Phosphorous Index Model as Environmental Risk Assessing Tool. In: *HAICTA 2008 4th International Conference on Information & Communication Technologies in Bio & Earth Sciences Proceedings*. pp. 70-77.
3. Buczko U., Kuchenbuch R.O. (2007) Phosphorus indices as risk-assessment tools in the U.S.A. and Europe – a review. *Journal of Plant Nutrition and Soil Science*. 170, pp. 445-460.
4. Chan L. (1996) Levenberg-Marquardt Learning and Regularization. In: Amari S., Xu L., King I., Leung K. S., Verlag S. (eds) *Progress in Neural Information Processing*. Springer Verlag, pp. 139-144.
5. Gümrah F., Öz B., Gülerand B., Evin S. (2000) The application of artificial neural networks for the prediction of water quality of polluted aquifer. *Water, air, and soil pollution*. 119, pp. 275-294.
6. Govindaraju R.S., Rao A.R. (2000) *Artificial Neural Networks in Hydrology*. Springer Kluwer Academic Publishers, 348 p.
7. Hagan M.T., Demuth H.B, Beale M.H. (1996) *Neural Network Design*. MA: PWS Publishing, Boston, 736 p.
8. Heathwaite A.L., Sharpley A.N., Bechmann M. (2000) The conceptual basis for a decision support framework to assess the risk of phosphorus loss at the field scale across Europe. *Journal of Plant Nutrition and Soil Science*. 166, pp. 1-12.
9. Maier H.R., Dandy G.C. (1996) The use of artificial neural networks for the prediction of water quality parameters. *Water Resources Research*. 32, pp. 1013-1022.
10. Maier R.H., Dandy G.C. (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software*. 15. pp. 101-124.
11. Hristev R.M. (1998) *The ANN Book*. GNU Public Licence. Available at: ftp://ftp.informatik.uni-freiburg.de/papers/neuro/ANN.ps.gz/*, 28.02.2009*.
12. Huang G.H., Xia J. (2001) Barriers to sustainable water-quality management. *Journal of Environmental Management*. 61, pp. 1-23.
13. Kim M., Gilley J.E. (2008) Artificial Neural Network estimation of soil erosion and nutrient concentrations in runoff from land application areas. *Computers and electronics in agriculture*. 64, pp. 268-275.
14. Kim M.Y., Hong K.J., Lee S.T., Kim M.K. (2006) Prediction of nitrogen and phosphorus transport in surface runoff from agricultural watersheds. *KSCE Journal of Civil Engineering*. 10, pp. 53-58.
15. Kuo J.T., Hsieh M.H., Lung W.S., She N. (2007) Using artificial neural network for reservoir eutrophication prediction. *Ecological modeling*. 200, pp. 171-177.
16. Nour M.H., Smith D.W., El-Din M.G., Prepas E.E. (2006) The application of artificial neural networks to flow and phosphorus dynamics in small streams on the Boreal Plain, with emphasis on the role of wetlands. *Ecological Modeling*. 191, pp. 19-32.
17. Sharpley A.N., Weld J.L., Beegle D.B., Kleinman P.J.A., Gburek W.L., Moore P.A., Mullins G. (2003) Development of phosphorus indices for nutrient management planning strategies in the U.S. *Journal of Soil and Water Conservation*. 58, pp. 137-152.
18. Suratgar A.A., Tavakoli M.B., Hoseinabadi A. (2005) Modified Levenberg-Marquardt Method for Neural Networks Training. Available at: http://www.waset.org/pwaset/v6/v6-10.pdf/, 01.03.2009.
19. Talib A., Hasan Y.A., Varis O. (2008) Application of Machine Learning Techniques in Data Mining of Ecological Datasets. In: *International Conference on Environmental Research and Technology Proceedings*, pp. 677-681.