PREDICTION OF WATER QUALITY PARAMETERS IN A RESERVOIR USING ARTIFICIAL NEURAL NETWORKS

H. VICENTE[1], C. COUTO[2], J. MACHADO[3], A. ABELHA[3] & J. NEVES[3]
[1]Department of Chemistry and Chemistry Centre of Évora, University of Évora,
Portugal (e-mail: hvicente@uevora.pt).
[2]Department of Chemistry, University of Évora,
Portugal (e-mail: horbite@gmail.com).
[3]Department of Informatics, University of Minho, Braga,
Portugal (e-mail: {jmac, abelha, jneves}@di.uminho.pt).

ABSTRACT

Water quality brings to the ground the discussion on water utilization once the consumption, of degraded water, is not possible or safe. On the other hand, the assessment of the water quality in a reservoir is constrained due to geographic considerations, the number of parameters to be studied, and the huge financial resources needed to get the necessary data. To this picture it should be added the latency times between the sampling moment and the instant that portrait the results of the laboratory analyses. However, new approaches to problem solving, namely those borrowed from the Artificial Intelligence arena have proven their ability and applicability in terms of simulation and modeling of the physical phenomena. Indeed, Artificial Neural Networks (ANNs) capture the embedded spatial and unsteady behavior in the investigated problem, using its architecture and nonlinearity nature, when compared with the other classical modeling techniques. This work describes the training, validation, and application of ANNs models for computing the oxidability and total suspended solids (TSS) levels in the Monte Novo reservoir, in Portugal, over a period of 15 years. Different network structures have been elaborated and evaluated. The performance of the ANNs models was assessed through the coefficient of determination (R^2), mean absolute deviation, mean squared error, and bias computed from the measured and model calculated values of the dependent variables. Goodness of the model fit to the data was also evaluated through the relationship between the errors and model computed values of oxidability and TSS. The ANNs selected to predict the oxidability from pH, conductivity, dissolved oxygen (DO), water temperature, and volume of water stored in reservoir has a 4-11-5-1 topology, while the network selected to predict the TSS has a 5-6-5-1 topology. A good match between the observed and predicted values was observed with the R^2 values varying in the range 0.995–0.998 for the training set, and 0.994–0.996 for the test set.

Keywords: Artificial Neural Networks, water quality, water reservoirs.

1 INTRODUCTION

Water quality is a term usually used to express the suitability of water to sustain various uses or processes. Quantity and quality demands from different users will not always fit together, as the activities of one user may restrict the activities of others, either by demanding water of a quality outside the range required by the other users or by lowering the water quality when using it. Efforts to improve or maintain a certain water quality often requires a middle ground between the quality and quantity demands of different users. Water quality management usually involves the monitoring of a series of key pollutants that serve as indicators of its acceptability for a specific use. The quality of water may be described in terms of the concentration and state (dissolved or particulate) of some or all of the organic and inorganic material present in it, together with certain water physical characteristics. It is determined by *in situ* measurements and by examination of water samples on site or in the laboratory. Indeed, this is a very restricted approach due to the distances, the number of parameters to be considered, and the financial resources spent to get their values. Moreover, under this context, the latency times between the sampling moment and the instant that points the conclusion of the laboratory analyses should be added. Due to these constraints, the development of computational

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://journals.witpress.com
DOI: 10.2495/DNE-V7-N3-310-319

models based on Artificial Intelligence tools and techniques for problem solving, in conjunction with the analysis and progresses in Decision Support Systems [1], fits as an alternative for the quality management of water resources.

The presence of solids and organic matter in water may affect water quality adversely in a number of different ways and may jeopardize its use for different purposes, such as the production of water for public supply. Although parametric statistical and deterministic models have been the traditional approaches for modeling water quality, these require vast information on various hydrological subprocesses to achieve the expected results. However, since a large number of factors affecting the water quality have a complicated nonlinear relation with the variables, traditional data processing methods are no longer good enough for solving the problem [2]. In recent years, some Artificial Intelligence based tools, namely Artificial Neural Networks (ANNs) and Decision Trees (DTs) have been applied for water quality assessment [3–6]. However, the prediction of TSS and oxidability is a complex and highly nonlinear problem for which, to our knowledge, no methods have been reported in the literature.

The aim of the current study was to use Artificial Intelligence-based tools, particularly ANNs to address this problem. ANNs can learn from examples, are fault tolerant in the sense that they can handle noisy and incomplete data, can deal with nonlinear problems and, once trained, can perform prediction and generalization [7,8]. This study took place in Monte Novo reservoir, which is located 20 km southwest of the Portuguese city of Évora, considered by UNESCO as World Heritage. The raw water of the Monte Novo reservoir is used to produce drinking water, supplying 70,000 inhabitants currently.

2 MATERIALS AND METHODS

The water samples used for the development of the models were collected during a given time period, from August 1995 to December 2010. The parameters analyzed were pH, conductivity, dissolved oxygen (DO), water temperature, volume of water stored in reservoir, oxidability, and total suspended solids (TSS).

2.1 Sample collection and preservation

Sample collection and sample preservation makes use of procedures described in Standard Methods for the Examination of Water and Wastewater (SMEWW) [9].

For pH, conductivity, DO and water temperature, the samples were collected in wide-mouth polyethylene bottles of 50 mL and analyzed immediately; for oxidability analysis, the samples were collected in polyethylene bottles of 100 mL, stored in dark, and kept refrigerated; finally, for TSS analysis, the samples were collected in polyethylene bottles of 100 mL and kept refrigerated.

2.2 Analytical procedures

The determination of pH was executed according to SMEWW 4500-H+ B using a Crison GLP 22 pH meter equipped with a Crisolyt 50 14 electrode. The conductivity was evaluated according to the Portuguese version of the European Standard 27888:1996 using a WTW InoLab cond 720 conductivity meter. The water temperature was determined according to SMEWW 2550 B. Measurements were carried out in field using SLW N16B Glas (–10 +50°C, 0,1°C) thermometer. The DO was determined in field with a Crison OXI 45 oxymeter equipped with a DurOx 325 electrode according to SMEWW 4500-O B. The TSS were evaluated according to SMEWW 2540 D. Finally, the oxidability was determined according to the Portuguese Standard 731:1969.

2.3 Artificial Neural Networks

The ANNs are computational tools that attempt to simulate the architecture and internal operational features of the human brain and nervous system. In this study, the most common neural network type, the multilayer perceptron, was adopted. This type of networks are formed by three or more layers of artificial neurons or nodes, the basic computing units, which include an input layer, an output layer, and a number of hidden layers with a certain number of active neurons connected by feed-forward links, to which are associated modifiable weights. In addition, there is also a bias, which are connected to neurons in the hidden and output layers. The number of nodes in the input layer denotes the number of independent variables and the number of nodes in the output layer stands for the number of dependent variables [8].

Although it has been proven that a network with one hidden layer can approximate any continuous function, given sufficient degrees of freedom [10], other studies have shown that, in practice, many functions are difficult to approximate with one hidden layer [11,12]. Indeed, there are no clear rules as to the ‘best’ number of hidden layer units. Network design is a trial-and-error process and may affect the accuracy of the resulting trained network. A number of automated techniques have been proposed to search for a ‘good’ network structure. These typically use a hill-climbing approach that starts with an initial structure that is selectively modified to improve performance, that is, to minimize an error metric [13,14]. In the present work, the error metric used was the mean squared error (MSE).

In the training phase, the back-propagation algorithm (BP) [15] was applied. This is the most widely used training algorithm for multilayered perceptron and basically involves two phases. One is the forward phase where the information is propagated from the input to the output layer. The second is the backward phase where an error, defined as the discrepancy between the observed value and the desired nominal value in the output layer, is propagated backwards to adjust the weightings and bias values. In the forward phase, the weighted sum of input components, u_j, is calculated as

$$u_j = \sum_{i=1}^{n} w_{ij} x_i + bias_j \tag{1}$$

where w_{ij} denotes the weight between the *j*th neuron and the *i*th neuron in the preceding layer, x_i denotes the output of the *i*th neuron in the preceding layer, and $bias_j$ denote the weight between the *j*th neuron and the bias neuron in the preceding layer.

The output of the *j*th neuron in any layer, y_j, is calculated as

$$y_j = f(u_j) \tag{2}$$

where *f* denotes the activation function. In all experiments the sigmoid activation function was used as given below:

$$\varphi(u_j) = \frac{1}{1+e^{-u_j}} \tag{3}$$

The BP algorithm is controlled by two parameters, the momentum coefficient and the learning rate, ranging between 0 and 1. The momentum coefficient is used in updating weights stage and tends to keep the weight changes in a consistent direction. Learning rate controls how much the weights are adjusted at each update. The Waikato Environment for Knowledge Analysis (WEKA) was used to implement ANNs, keeping the default software parameters [16].

To ensure statistical significance of the attained results, 20 runs were applied in all tests. In each simulation, the available data was randomly divided into three mutually exclusive partitions: the training set, with 60% of the available data, used during the modeling phase; the test set, with 25% of the examples, used after training to evaluate the model performance; and the validation set, with the remaining 15% of data to validate the models [17]. To improve the performance of the learning algorithm and avoid the overvaluation of the attributes with larger intervals at the expense of the attributes with smaller ones, the data was normalized to the interval [0, 1], using the equation depicted below [14]:

$$\overline{X} = \frac{X - X_{min}}{X_{max} - X_{\text{min}}} \tag{4}$$

where $\overline{X}$ denotes the normalized value, X denotes the attribute value and X_{min} and X_{max} denote, respectively, the minimum and the maximum values for the attribute.

3 RESULTS AND DISCUSSION

3.1 Database

The data used in this study covered the period from August 1995 to December 2010, containing a total of 184 records with 7 fields. The fields were pH, conductivity, DO, water temperature and volume of water stored in reservoir, oxidability, and TSS. Table 1 shows the statistical characterization of the fields included in the database.

Excluding pH, Table 1 shows large dispersion of the data with high coefficient of variation, ranging from 22.5 to 74.9%. The coefficient of variation (CV) is a measure of dispersion of data and it is calculated as *(standard deviation/mean)* × *100*. Such variability may be attributed to the large geographical variations in climate and seasonal influences in the region object of study. The local climate is Mediterranean-type (Csa according to Köppen), characterized by winter-wet and summer-dry pattern. Mean annual rainfall is 665 mm, most of which falling from autumn to early spring (90%) in < 75 days of rain per year [18]. Mean annual air temperature is about 15.4°C, ranging from 8.6°C in January to 23.1°C in August. Air relative humidity is about 70%. The dry period is up to 5 months. The pH shows the CV lowest variation, and it may be due to the buffering capacity of the reservoir. Nevertheless, these results are in agreement with results presented by other authors for similar systems [19,20].

Table 1: Statistical characterization of the numerical variables used in the study.

Variable	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation (%)
pH (Sørensen scale)	7.2	9.1	7.9	0.5	6.3
Conductivity (μS/cm)	117	667	287	119	41.5
Water temperature (°C)	7.6	26.9	17.1	5.6	32.7
DO (% sat)	43.3	170.9	87.1	22.8	26.2
Volume stored (dam^3)	6900	15277	12668	2847	22.5
TSS (mg/dm^3)	5.0	116.9	34.7	26.0	74.9
Oxidability (mg/dm^3)	0.3	16.0	7.1	3.8	53.5

3.2 ANNs models

To obtain the best prediction of the output parameters (oxidability and TSS), different network structures and architectures were elaborated and evaluated. The optimum number of hidden layers and the optimum number of nodes in each of these was found through a process of trial and error. Common tools to compare the performances of regression models are the mean absolute deviation (MAD), and the MSE. According to Torgo [21], these tools, when applied to the evaluation of regression models, serve different purposes. If the goal is a model with good fit for most cases even though allowing some higher deviations, MAD should be minimized. Conversely, if the objective is not committing large deviations, although frequent small errors can be allowed, MSE should be minimized, once this measure reflects the large deviations in the final result. These two measures of goodness-of-fit are related to the average prediction error. Nevertheless, they do not provide any information on the nature of the errors. According to Chenard and Caissie [22], the average of all individual errors, named bias, can be calculated indicating whether the model overestimates or underestimates the output variables. Table 2 presents the values of MAD, MSE and bias for some of the topologies tested.

Concerning the prediction of oxidability, Table 2 shows that 4-11-5-1 ANN minimizes MAD and MSE and exhibits a bias value closer to zero for the training set and for the test set. Regarding the prediction of TSS, Table 2 shows that 5-6-5-1 network minimizes MAD and MSE and exhibits a bias value closer to zero for both data sets. The architecture of the best ANN for modeling the oxidability

Table 2: MAD, MSE, and bias for some ANN topologies tested.

	ANN topology	MAD*		MSE*		Bias*	
		Training set	Test set	Training set	Test set	Training set	Test set
Oxidability	4-6-3-1	1.572	1.741	6.901	6.230	–0.235	–0.571
	4-7-5-1	1.794	1.821	8.221	8.615	0.374	0.295
	4-9-4-1	0.829	0.813	0.375	0.397	–0.051	0.065
	4-10-6-1	1.122	1.099	2.253	2.654	–0.381	0.222
	4-11-5-1	0.209	0.253	0.060	0.073	–0.004	0.028
	4-11-7-1	0.841	1.036	1.113	1.541	–0.658	0.825
	4-12-10-1	1.222	1.632	2.478	2.616	0.845	–0.755
TSS	5-4-1-1	2.033	2.211	4.233	5.101	–0.326	0.445
	5-5-3-1	1.154	1.504	3.123	3.621	0.472	0.581
	5-6-4-1	0.742	0.971	1.026	1.741	–0.102	–0.233
	5-6-5-1	0.339	0.401	0.175	0.257	0.002	–0.007
	5-7-2-1	0.392	0.471	0.309	0.724	–0.013	0.075
	5-8-5-1	0.542	0.847	0.846	0.973	0.056	–0.181
	5-10-3-1	2.045	2.256	3.969	3.924	0.418	–0.473

$*$ $\mathrm{MAD} = \frac{\sum_{i=1}^{N}\left|Y_i' - Y_i\right|}{N}$; $\mathrm{MSE} = \frac{\sum_{i=1}^{N}(Y_i' - Y_i)^2}{N}$; $\mathrm{bias} = \frac{\sum_{i=1}^{N}(Y_i' - Y_i)}{N}$; Y denotes an experimental value, Y' stands for a predicted value, and N indicates the number of observations.

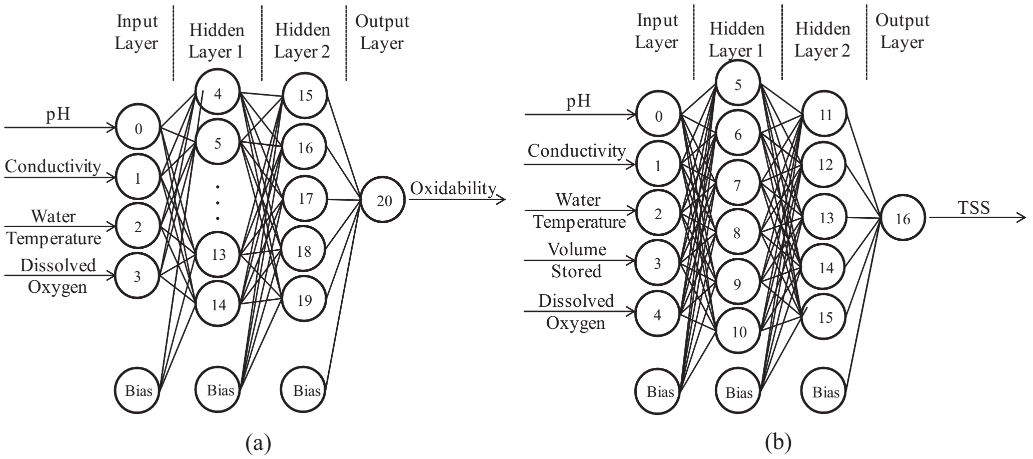

Figure1: ANN structure for modeling oxidability (a) and TSS (b).

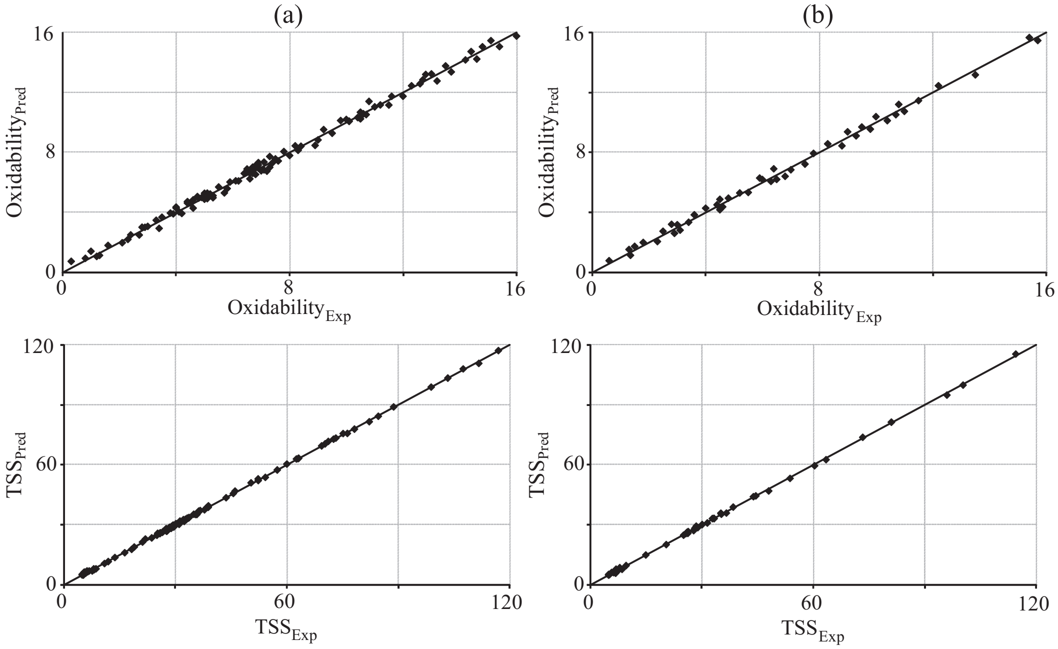

Figure 2: Plot of predicted response by ANN models for oxidability and for TSS versus experimental values for training (a) and test (b) sets.

in Monte Novo reservoir is shown in Fig. 1a. This model consists in an input layer with 4 nodes, 2 hidden layers with 11 and 5 nodes, and 1 node output layer. Figure 1b shows the best network architecture for modeling TSS that consists of an input layer with five nodes, two hidden layers with six and five nodes, and one node output layer.

Figure 2a,b shows the plots between experimental and predicted values of oxidability and TSS for training and test sets. The values of the coefficient of determination ($p < 0.001$) for the training and test sets were 0.995 and 0.994 for the ANN model to predict oxidability, and 0.998 and 0.996 for the ANN model to predict the TSS. The agreement between the experimental and predicted values for both parameters of water quality, R^2, MAD, and MSE (Table 2) seems to suggest a good-fit of both models to the data set.

In addition to what was stated above, Fig. 3 shows the plots between errors and predicted values of oxidability and TSS for training and test sets. The observed relationship between errors and predicted values for the two water quality parameters for training and test sets shows complete independence and random distribution. Indeed, the determination coefficients are negligibly small (0.001 and 0.007 for training set and 0.007 and 0.006 for test set, respectively, for oxidability and TSS). Figure 3 shows that the points are well distributed on both sides of the horizontal line of zero ordinate, corresponding to the correct prediction. Plots of the errors versus predicted values can be more informative regarding model fitting to a data set. If the errors appear to behave randomly, it suggests that the model fits the data quite well. On the other hand, if nonrandom distribution is evident in the errors, the model does not fit the data satisfactorily [2,23].

3.3 Validation of the ANN models

To validate the models, a set of independent data was used and computed the values of oxidability and TSS. Figure 4 shows the plots between experimental and predicted values of oxidability and TSS for the validation set. The coefficient of determination (R^2), MAD, MSE, and bias were

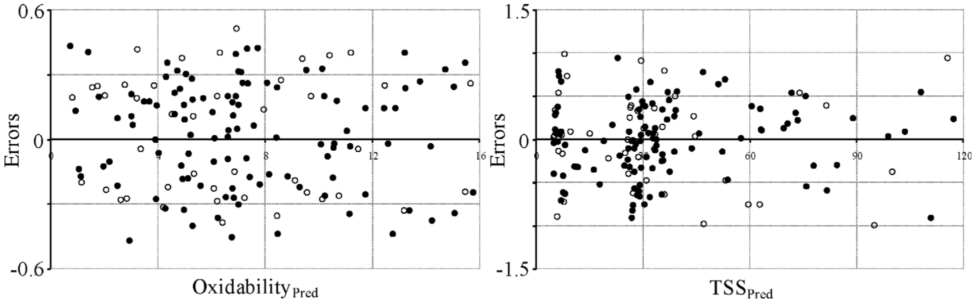

Figure 3: Plot of the errors versus predicted response by ANN models for oxidability and TSS for training (•) and test (◦) sets.

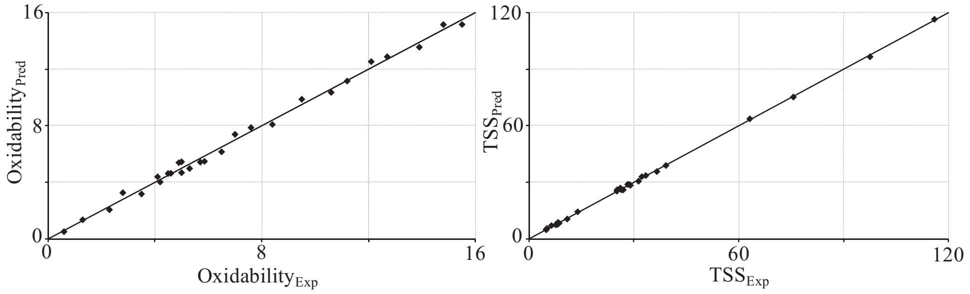

Figure 4: Plot of predicted response by ANN models for oxidability and for TSS versus experimental values for the validation set.

Table 3: Comparison between measured and computed values by the selected ANN models for the validation set.

ANN model	Output variable	R^2	MAD	MSE	Bias
4-11-5-1	Oxidability	0.992	0.284	0. 097	0.097
5-6-5-1	TSS	0.994	0.478	0.300	–0.068

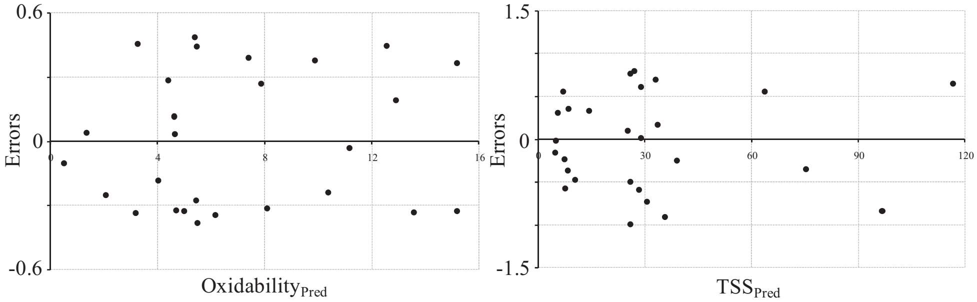

Figure 5: Plot of the errors versus predicted response by ANN models for oxidability and TSS for the validation sets.

computed for both models and are presented in Table 3. The obtained values are similar to those presented earlier, for both models, for training and test sets. Furthermore, Fig. 5 shows the plots between errors and predicted values of oxidability and TSS for the validation test. The relationships observed for both models show complete independence, random distribution and are well distributed on both sides of the line of correct prediction. These results demonstrate that both models performed well for an independent set of data and, therefore, do not show overfitting.

3.4 Sensitivity analysis of the ANN models

Sensitivity analysis is the process of defining model output sensitivity to changes in its input variables. Typically, the efforts in data acquisition will be focused on the more relevant variables for the model accuracy and dropping or ignoring those that matter least. Sensitivity analysis is a simple procedure that is applied after the modeling phase and analyzes the model responses when the inputs are changed. In this work the sensitivity according variance [24] was used to compute the relative importance of the input variables for selected models. The results are presented in Fig. 6 and reveal that the most informative variable for the model to predict oxidability is conductivity, followed by water temperature and pH, suggesting their direct influence on the oxidability level in the water. Regarding the model to forecast the TSS, Fig. 6 suggests that all input variables contribute significantly to the network, although DO and pH provide a relatively higher contribution.

4 CONCLUSIONS

In this study, two models based on ANNs were developed to predict the oxidability and the TSS in water. The models were trained, tested, and validated using monthly data measured over a period of 15 years. The feed-forward network with the BP learning algorithm was employed. The selected

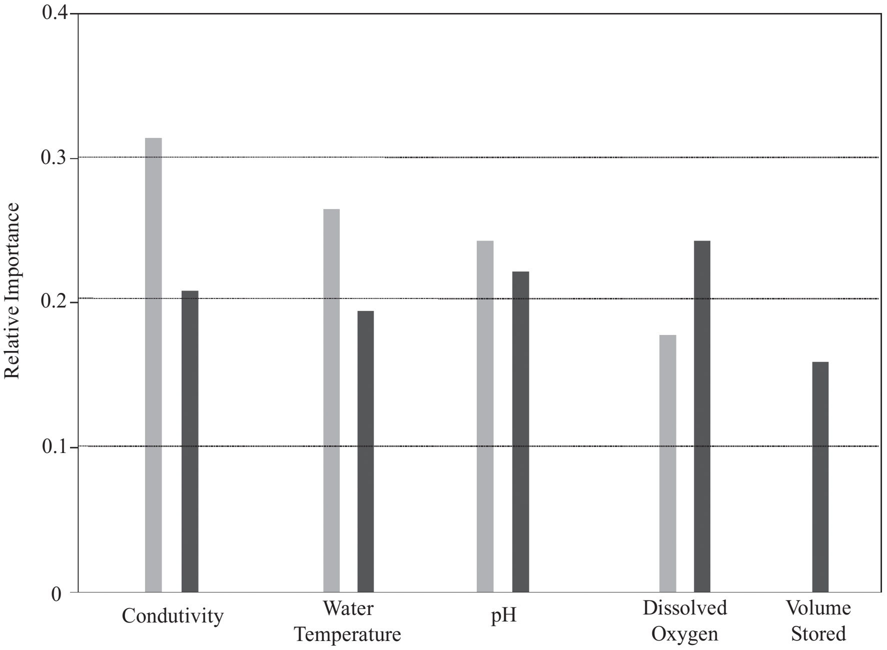

Figure 6: Relative importance of the input variables for selected ANNs to model oxidability (□) and to model TSS (■).

models performed well in prediction of the output variables based on pH, conductivity, water temperature, DO, and volume of water stored in the reservoir. These encouraging results obtained in this work show that ANNs can be very useful as tools to predict water quality and can contribute significantly to the effort that is needed for a constant improvement not only on the management of the reservoirs but also on the preservation of the quality of the water.

REFERENCES

[1] Turban, E., Aronson, J.E. & Liang, T.-P., *Decision Support Systems and Intelligent Systems*, Prentice Hall: New Jersey, USA, 2004.

[2] McBride, G.B., *Using Statistical Methods for Water Quality Management Issues, Problems and Solutions*, John Wiley & Sons: Hoboken, U.S.A., 2005.

[3] Santos, M.F., Cortez, P., Quintela, H., Neves, J., Vicente, H. & Arteiro, J., Ecological Mining - A Case Study on Dam Water Quality. *Data Mining VI - Data Mining, Text Mining and their Business Applications*, eds. A. Zanasi, C.A. Brebbia & N.F.F. Ebecken, WIT Press: Southampton, UK, pp. 523–531, 2005.

[4] Pinto, A., Fernandes, A.V., Vicente, H. & Neves, J., Optimizing water treatment systems using artificial intelligence based tools. *Water Resourse Management V*, eds. C.A. Brebbia & V. Popov, WIT Press: Southampton, UK, pp. 185–194, 2009.

[5] Maier, H., Jain, A., Dandy, G. & Sudheer, K., Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling & Software*, **25**, pp. 891–909, 2010. doi: http://dx.doi.org/10.1016/j.envsoft.2010.02.003

[6] West, D. & Dellana, S., An empirical analysis of neural network memory structures for basin water quality forecasting. *International Journal of Forecasting*, **27**, pp. 777–803, 2011. doi: http://dx.doi.org/10.1016/j.ijforecast.2010.09.003

[7] Galushkin, A.I., *Neural Networks Theory*, Springer: New York, USA, 2007.

[8] Haykin, S., *Neural Networks and learning machines*, Prentice Hall: New Jersey, USA, 2008.

[9] Eaton, A., Clesceri, L., Rice, E. & Greenberg, A., (eds). *Standard Methods for the Examination of Water and Wastewater*, American Public Health Association: USA, 2005.

[10] Hornik, K., Stinchcombe, M. & White, H., Multilayer feed-forward networks are universal approximators. *Neural Networks*, **2**, pp. 359–366, 1989. doi: http://dx.doi.org/10.1016/0893-6080(89)90020-8

[11] Cheng, B. & Titterington, D., Neural networks: a review from a statistical perspective, *Statistical Science*, **9**, pp. 2–30, 1994. doi: http://dx.doi.org/10.1214/ss/1177010638

[12] Flood, I. & Kartam, N., Neural network in civil engineering: I. Principles and understanding. *Journal of Computational in Civil Engineering*, **8**, pp. 131–148, 1994. doi: http://dx.doi.org/10.1061/(ASCE)0887-3801(1994)8:2(131)

[13] Cortez, P., Rocha, M. & Neves, J., Evolving time series forecasting ARMA models. *Journal of Heuristics*, **10**, pp. 415–429, 2004. doi: http://dx.doi.org/10.1023/B:HEUR.0000034714.09838.1e

[14] Han, J. & Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kauffmann Publishers: San Francisco, USA, 2006.

[15] Rumelhart, D., Hinton, G. & Williams, R., Learning Internal Representation by Error Propagation. *Parallel Distributed Processing*, eds. D.E. Rumelhart & J.L. McCleland, MIT Press: Massachusetts, U.S.A., pp. 318–362, 1986.

[16] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H., The WEKA Data Mining Software: An Update. *SIGKDD Exploration*, **11**, pp. 10–18, 2009. doi: http://dx.doi.org/10.1145/1656274.1656278

[17] Souza, J., Matwin, S. & Japkowicz, N., Evaluating data mining models: a pattern language. *Proc. of the 9th Conference on Pattern Language of Programs*, pp.11–23, 2002.

[18] INMG, *O clima de Portugal. Normais climatológicas da região de Alentejo e Algarve, Correspondentes a 1951-1980*. Fasciculo XLIX, Vol. 4 – 4ª região, Instituto Nacional de Meteorologia e Geofísica: Lisboa, Portugal, 1991.

[19] Palani, S., Liong, S.-Y. & Tkalich, P., An ANN application for water quality forecasting. *Marine Pollution Bulletin*, **56**, pp. 1586–1597, 2008. doi: http://dx.doi.org/10.1016/j.marpolbul.2008.05.021

[20] Singh, K., Basant, A., Malik, A. & Jain, G., Artificial neural network modeling of the river water quality - A case study. *Ecological Modelling*, **220**, pp. 888–895, 2009. doi: http://dx.doi.org/10.1016/j.ecolmodel.2009.01.004

[21] Torgo, L., *Inductive Learning of Tree-Based Regression Models*, University of Oporto: Oporto, Portugal, 1999.

[22] Chenard, J.-F. & Caissie, D., Stream temperature modelling using artificial neural networks: application on catamaran brook. *Hydrological Processes*, **22**, pp. 3361–3372, 2008. doi: http://dx.doi.org/10.1002/hyp.6928

[23] e-Handbook of Statistical Methods; NIST/SEMATECH Online. http://www.itl.nist.gov/div898/handbook

[24] Kewley, R., Embrechts, M. & Breneman, C., Data strip mining for the virtual design of pharmaceuticals with neural networks. *IEEE Transactions on Neural Networks*, **11**, pp. 668–679, 2000. doi: http://dx.doi.org/10.1109/72.846738