WATER QUALITY MODELING USING ARTIFICIAL INTELLIGENCE-BASED TOOLS

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

ABSTRACT

Water, like any other biosphere natural resource, is scarce, and its judicious use includes its quality safeguarding. Indeed, there is a wide concern to the fact that an inefficient water management system may become one of the major drawbacks for a human-centered sustainable development process. The assessment of reservoir water quality is constrained due to geographic considerations, the number of parameters to be considered and the huge financial resources needed to get such data. Under these circumstances, the modeling of water quality in reservoirs is essential in the resolution of environmental problems and has lately been asserting itself as a relevant tool for a sustainable and harmonious progress of the populations. The analysis and development of forecast models, based on Artificial Intelligence-based tools and the new methodologies for problem solving, has proven to be an alternative, having in mind a pro-active behavior that may contribute decisively to diagnose, preserve, and rehabilitate the reservoirs. In particular, this work describes the training, validation and application of Artificial Neural Networks (ANNs) and Decision Trees (DTs) to forecast the water quality of the Odivelas reservoir, in Portugal, over a period of 10 years. The input variables of the ANN model are chemical oxygen demand (COD), dissolved oxygen (DO), and oxidability and total suspended solids (TSS), while for the DT the inputs are, in addition to those used by ANN, the Water Conductivity and the Temperature. The performance of the models, evaluated in terms of the coincidence matrix, created by matching the predicted and actual values, are very similar for both models; the percentage of adjustments relative to the number of presented cases is 98.8% for the training set and 97.4% for the testing one.
Keywords: Artificial Neuronal Networks, data mining, Decision Trees, water quality.

1 INTRODUCTION

The water quality at ground zero in a given region largely depends on the nature and the extent of the industrial, agricultural, and other anthropogenic activities in the catchments. Undeniably, ensuring an efficient water management system is a major goal in contemporary societies, taking into account the importance to health and the need to safeguard and promote its sustainable use. However, the assessment of a dam water quality is being done through analytical methods, which may not be a good way of doing the assessment, due to the distances to be covered, the number of parameters to be considered, and the financial resources spent to obtain such data, that is, being what it is and no more. To this picture, the latency times between the sampling moment and the one of the outcome in terms of the laboratory analyses should be added. Due to these constraints, the development of Data Mining (DM)-based models [1] in conjunction with the development of Decision Support Systems [2], seems to be a better alternative for the quality management process of water resources.

New technological breakthroughs provided new ways to create and store information. Indeed, organizations accumulate huge amounts of information on a daily basis according to their specificities and processes, based on the assumption that large volume of data may be a source of knowledge which may be used to improve their performance and behavior, either by discovering trends or accelerating the course of efficient decision-making. However, the conventional tools for data

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://journals.witpress.com
DOI: 10.2495/DNE-V7-N3-300-309

analysis have a great number of drawbacks, since they do not allow the detection of singularities inside such massive facts. Definitely, having in mind a response to a given number of difficulties (e.g. those resulting from the use of large amount of data, multiple sources of data, or several application domains), a new area of Knowledge Discovery from Databases (KDD) was brought to life, and its tools and techniques to problem-solving have been since then enforced. The designation KDD was formally adopted in 1989 and refers to a process that involves the identification and recognition of patterns in a database, in an automatic process, that is, obtaining relevant, unknown information that can be useful in a decision-making process, without a previous formulation of hypothesis [1,3].

The interest on ecological mining has been growing substantially in recent years. Indeed, several innovative computational intelligence approaches have been used to find patterns in water quality databases, such as Artificial Neural Networks (ANNs) and Decision Trees (DTs) [4-7]. Although the ANNs have been used more extensively in ecological modeling than the DTs, these have the advantage of expressing regularities explicitly and thus being easy to inspect for ecological validity.

This study took place in the Odivelas reservoir, which is located 50 km southwest of the Portuguese city of Évora.

2 MATERIALS AND METHODS

The water samples used for the development of the models were collected in the Odivelas reservoir from January 2001 to December 2010. The parameters analyzed were water temperature, pH, dissolved oxygen (DO), conductivity, ammonium, iron, nitrate, orthophosphate, cadmium, chromium, copper, lead, manganese, nickel, total suspended solids (TSS), chemical oxygen demand (COD), and oxidability.

2.1 Sample collection and preservation

Sample collection and sample preservation followed procedures described in Standard Methods for the Examination of Water and Wastewater (SMEWW) [8]. For water temperature, pH, DO, and conductivity, the samples were collected in wide-mouth polyethylene bottles of 50 mL and analyzed immediately; for ammonium and COD analysis, the samples were collected in polyethylene bottle of 500 mL, preserved with sulfuric acid, $pH \leq 2$, and kept refrigerated; for iron analysis, the samples were collected in a glass bottle of 100 mL, preserved with sulfuric acid 4.5 M, $pH \leq 2$; for nitrate, orthophosphate, and TSS analysis, the samples were collected in polyethylene bottles of 100 mL and kept refrigerated; for oxidability analysis the samples were collected in polyethylene bottles of 100 mL, stored in dark, and kept refrigerated; finally, for remaining metals analysis, the samples were collected in polyethylene bottle of 1000 mL rinsed with nitric acid and preserved with nitric acid, $pH \leq 2$.

2.2 Analytical procedures

The analyses of water quality parameters followed the SMEWW [8] or International Standard Organization (ISO) or European Standards or Portuguese Standards (Table 1).

The water temperature measurements were carried out in field using SLW N16B Glas (−10 +50°C, 0.1°C) thermometer. The determination of pH was executed using a Sherwood SCI Delta 345 pH meter equipped with a Mettler Toledo Inlab 412 electrode. The DO was determined in field with a Crison OXI 45 oxymeter equipped with a DurOx 325 electrode. The conductivity measurements were carried out on a Crison 2202 micro CM conductivity meter equipped with a Crison ACC 5292 cell.

Table 1: Analytical techniques and test methods.

Parameter	Analytical technique	Test method
Water temperature	-------	SMEWW 2550 B
pH	Potentiometry	SMEWW 4500-H^+
DO		SMEWW 4500-O G
Conductivity	Conductimetry	NP EN 27888:1996 (*)
Ammonium	Molecular absorption spectrometry	ISO 7150-1:1984
Iron		NP 2202:1996 (**)
Nitrate		SMEWW 4500-NO_3^-
Orthophosphate		SMEWW 4500-P E
Cadmium	Atomic absorption spectrometry	SMEWW 3113-B
Chromium		SMEWW 3113-B
Copper		SMEWW 3113-B
Lead		SMEWW 3113-B
Manganese		SMEWW 3113-B
Nickel		SMEWW 3113-B
TSS	Gravimetry	SMEWW 2540 B
COD	Volumetry	SMEWW 5220 B
Oxidability		NP 731:1969 (**)

(*) NP EN – Portuguese version of European Standard; (**) NP – Portuguese Standard.

The molecular absorption spectrometry measurements were carried out on a Thermo Electron spectrometer model Nicolet Evolution 300 LC. Finally, the atomic absorption spectrometry measurements were carried out on a Perkin Elmer 3110 spectrometer equipped with a HGA-600 graphite furnace.

2.3 Artificial Neural Networks

The most common ANN type, the multilayer perceptron, was adopted. In this network, neurons are grouped into layers and connected by feed-forward links [9, 10]. In the training phase, the back-propagation algorithm [11] was applied. In all experiments, the sigmoid activation function was used:

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

where u_j designates the weighted sum of the jth neuron for the input received from a former layer with n neurons, calculated as

$$u_j = \sum_{i=1}^{n} w_{ij} x_i + bias_j \tag{2}$$

where w_{ij} denotes the weight between the jth neuron and the ith neuron in the preceding layer, x_i stands for the output of the ith neuron in the preceding layer and $bias_j$ refers to the weight between the jth neuron and the bias neuron in the preceding layer.

2.4 Decision Trees

DTs stand for one of the most efficient data mining classification methods. DTs have many attractive features, such as allowing human interpretation and hence making it possible for a decisionmaker to gain insights into what factors are important for a particular classification. DTs adopt a branching structure of nodes and leaves, where the knowledge is hierarchically organized. Each node tests the value of a feature, while to each leaf is assigned a class label. The basic strategy that is employed when generating DTs is called recursive partitioning or divide-and-conquer. It works by partitioning the examples by choosing a set of conditions on an independent variable, and the choice is usually made such that errors on dependent variables are minimized within each group. The process continues recursively with each subgroup until definite conditions are met, such as looking to an error that cannot be further reduced (e.g. all examples in a group belong to the same class) [3]. Early systems for generating DTs include CART [12] and ID3 [13], the later being followed by the version C4.5 and C5.0. The C4.5 version was an improvement of the ID3 algorithm that allows the use of continuous values, support omitted values, tree pruning, and rules extraction [14].

2.5 Tests

The Waikato Environment for Knowledge Analysis (WEKA) [15] was used to implement ANNs and DTs, keeping the default software parameters. The algorithm used to induce DTs was the J.48 algorithm that implements the 8th revision of the universally known C4.5 algorithm [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 two mutually exclusive partitions: the training set, with two-third of the available data was used during the modeling phase and the test set, with the remaining one-third of the examples, was used after training to evaluate the performance of the models [17].

To improve the performance of the learning algorithms and to 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 [3]:

$$\overline{X} = \frac{X - X_{min}}{X_{max} - X_{\min}} \tag{3}$$

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.

A common tool for classification analysis to evaluate the performances of classification models is the coincidence matrix [18], a matrix of size $L \times L$, where L denotes the number of possible classes. This matrix is created by matching the predicted (test result) and actual (water quality class) values. L was set to 2 in the present case.

3 RESULTS AND DISCUSSION

3.1 Database

The data used in this study covered the time period from January 2001 to December 2010. However, in most of the cases, the parameters ammonium, iron, cadmium, chromium, copper, lead, manganese, and nickel exhibited values below the quantification limit of the analytical methods being, therefore, excluded. The database used in this study contained a total of 120 records with 9 numeric

fields. The numerical fields were pH, conductivity, DO, water temperature, orthophosphate, oxidability, TSS, nitrate, and COD. Table 2 shows the statistical characterization of the numerical fields included in the database.

Besides the numerical variables presented in Table 2, the data base includes the classification of the water quality body at the ground zero level. The criterion used in this study to classify the quality of the water was adopted by INAG, the Portuguese water management service. Therefore, water will be classified in the nonlinear scale A, B, C, D, or E [19], where A denotes no pollution and E alludes to extreme pollution, which represents serious risks in terms of public health and the environment (Fig. 1). The original dataset presented biased distributions: in 52.2% of the observations, the water quality of the Odivelas reservoir is polluted (**C**); 47.8% is weakly polluted (**B**). The classifications extremely polluted (**E**), very polluted (**D**), or non polluted (**A**), were not found.

3.2 DTs model

The DTs model obtained to predict water quality of the Odivelas reservoir is showed in Fig. 2. It should be noted that the algorithm for induction of DTs did not use the data related to pH, nitrate,

Table 2: Statistical characterization of the numerical variables used in the study.

Variable	Minimum	Maximum	Mean	Standard deviation
Water temperature (ºC)	11	29.1	20.1	4.9
pH (Sørensen scale)	7.3	9.09	8.2	0.32
DO (% sat)	60.5	113.8	81.3	11.7
Conductivity (μS/cm)	280	494	392	54
Nitrate (mg/dm^3)	0.13	4.6	1.09	1
Orthophosphate (mg/dm^3)	0.05	0.182	0.45	0.034
TSS (mg/dm^3)	1.3	70	8.92	11.03
COD (mg/dm^3)	1	51	22	8.25
Oxidability (mg/dm^3)	2.4	23.2	5.8	2.23

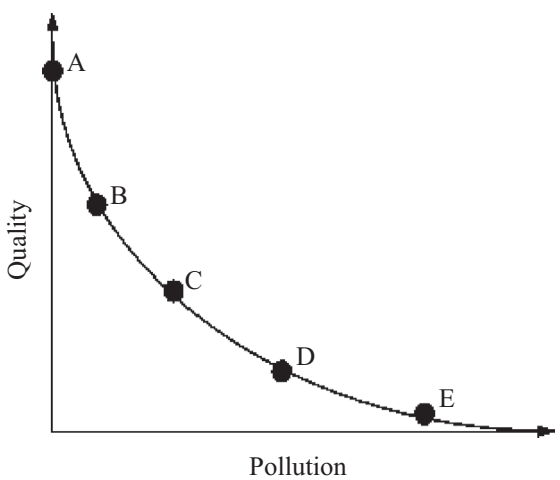

Figure 1: The water quality classes versus the pollution factor.

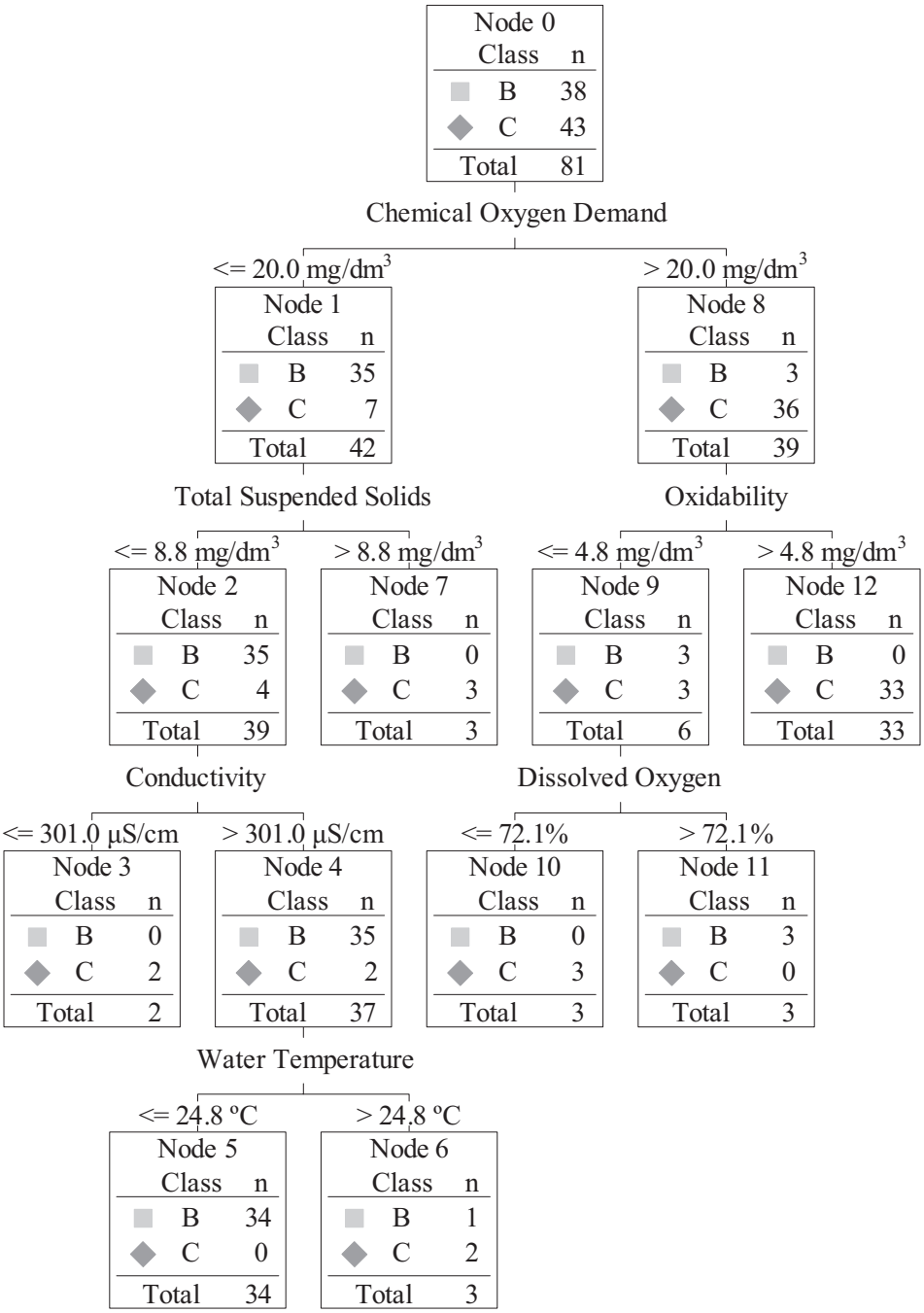

Figure 2: DT model to predict the water quality of the Odivelas reservoir.

and orthophosphate, despite being available, having chosen COD, oxidability, water temperature, DO, conductivity, and TSS. To evaluate the model output sensitivity to changes in its input variables, the sensitivity, according to the variance [20], to compute the relative importance of the input variables is used. The results are presented in Fig. 3, and reveals that the most informative variable is COD followed by oxidability and water temperature. These results seem to suggest that these three variables have direct relevance and play a significant role in preserving of water quality of the Odivelas reservoir. Table 3 presents the coincidence matrix for the DTs model. The values denote the average of the 20 runs. The results reveal that the model exhibits 100% accuracy in predicting polluted cases (C) and shows 96.5% accuracy in predicting the weakly polluted examples (B).

3.3 ANNs model

The ANN model obtained to predict the water quality of the Odivelas reservoir is showed in Fig. 4. The architecture of the model consists in an input layer with 4 nodes, 2 hidden layers with 14 and 5 nodes, respectively, and a 2-nodes output layer. It should be emphasized that the algorithm uses only 4 (four) variables (namely COD, oxidability, DO and TSS), even though all the variables presented in Table 2 were available for utilization. As before, the relative importance of the input variables was

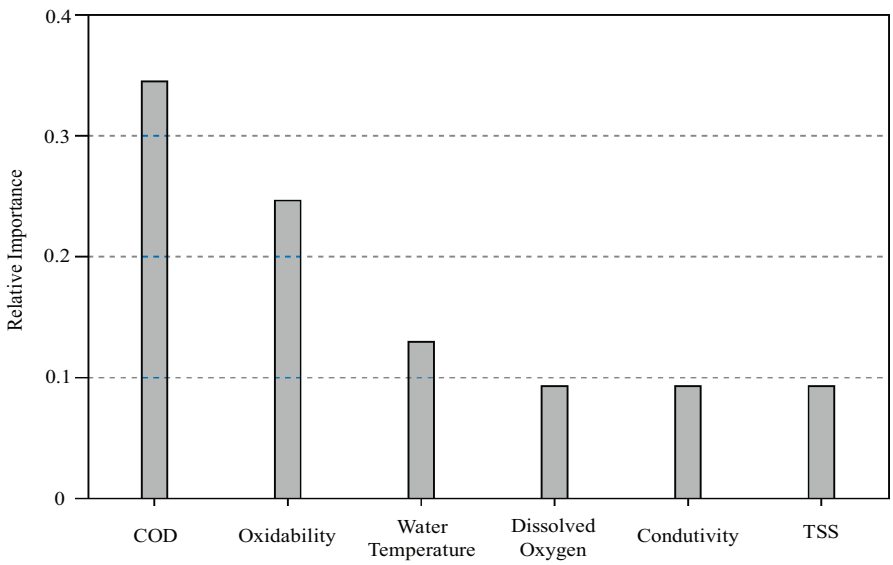

Figure 3: Relative importance of the input variables for DTs model.

Table 3: The coincidence matrix for DT model.

	Training set		Test set	
Class	B	C	B	C
B	37	1	18	1
C	0	43	0	20

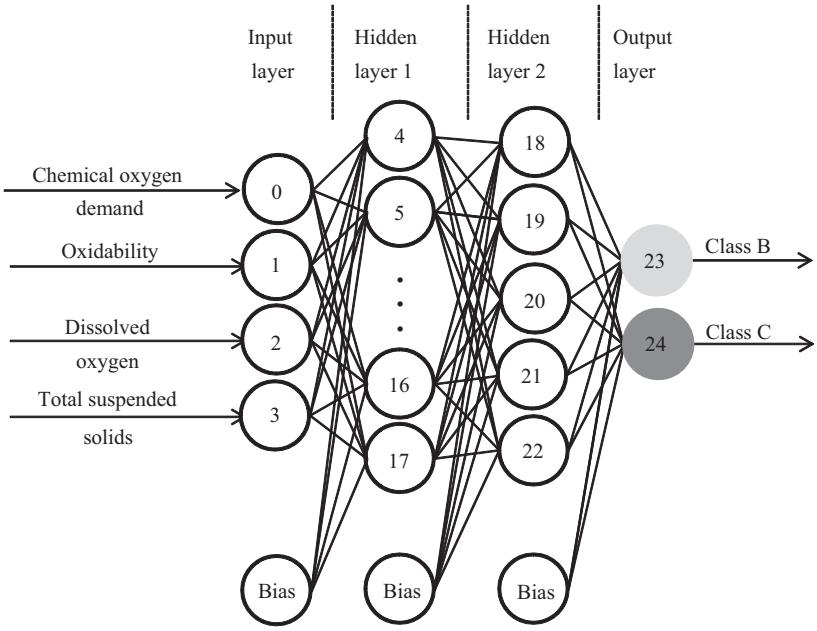

Figure 4: The ANN model to predict water quality of the Odivelas reservoir.

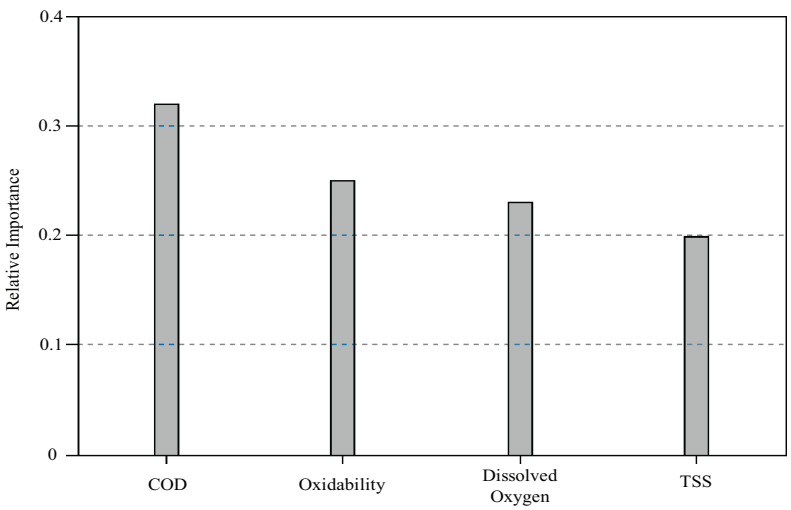

Figure 5: Relative importance of the input variables for the ANN model.

computed (Fig. 5). The analysis of Fig. 5 shows that all input variables contribute significantly to the network, although COD provide a relatively higher contribution. Table 4 presents the coincidence matrix for the ANN model. The values denote the average of the 20 runs. The results reveal that the model exhibits 100% accuracy in predicting polluted cases (C) and shows 96.5% accuracy in predicting of weakly polluted examples (B).

Table 4: The coincidence matrix for ANN model.

	Training set		Test set	
Class	B	C	B	C
B	38	1	17	1
C	0	44	0	19

4 CONCLUSIONS

The use of data mining techniques can solve complex problems in environmental applications, such as the prediction of water quality in reservoirs. In this work, two classification models were presented and tested, using DTs and ANNs. The former adopted six input variables, while the latter only considered four input variables. The feeling shows, according to the variance of both models, that COD and oxidability provide a relatively higher contribution for the results of the models and seem to suggest that these variables play a significant role in prediction of water quality of the Odivelas reservoir. On the other hand, both models presented a classification rate of 96.5% for weakly polluted class (B) and 100% for polluted class (C). The on hand models have advantages and disadvantages. In fact, the DTs-based model is easy to interpret and can be validated by experts in contrast to the ANN model. Conversely, the network model requires less input variables, which constitute an effective advantage. The encouraging results obtained in this work show that the DTs and ANNs can be very useful as tools to predict water quality and can contribute significantly to the effort that is needed for constant improvement of the quality of the water resources.

REFERENCES

[1] Fayyad, U., Piatetshy-Shapiro, G., Smith, P. & Uthurusamy, R., *Advances in Knowledge Discovery and Data Mining*, MIT Press: Massachusetts, USA, 1996.

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

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

[4] 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.

[5] 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.

[6] 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

[7] 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

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

[9] Galushkin, A.I., *Neural Networks Theory*, Springer: New York, USA, 2007.
[10] Haykin, S., *Neural Networks and learning machines*, Prentice Hall: New Jersey, USA, 2008.
[11] 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.
[12] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J., *Classification and Regression Trees*, Chapman & Hall/ CRC Press: Boca Raton, U.S.A., 1984.
[13] Quinlan, J.R., Induction of decision trees. *Machine Learning*, **1**, pp. 81–106, 1986. doi: http://dx.doi.org/10.1007/BF00116251
[14] Quinlan, J., *C4.5 Programs for Machine Learning*, Morgan Kaufmann Publishers Inc: USA, 1993.
[15] 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
[16] Witten, I.H. & Frank, E., *Data Mining – Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers: San Francisco, USA, 2005.
[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] Kohavi, R. & Provost, F., Glossary of Terms. *Machine Learning*, **30**, pp. 271–274, 1998. doi: http://dx.doi.org/10.1023/A:1017181826899
[19] CCDRA, *Anuário de Qualidade da Água da Região Alentejo 2006-2007*, Comissão de Coordenação e Desenvolvimento Regional do Alentejo: Evora, Portugal, 2008.
[20] 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