Comparative use of artificial neural networks for the quality assessment of the water reservoirs of Athens

Eleni G. Farmaki, Nikolaos S. Thomaidis, Vasil Simeonov and Constantinos E. Efstathiou

ABSTRACT

Neural networks are powerful tools that could explore the basic structure of environmental data. In this work, the most common artificial neural network (ANN) architectures, multi-layer perceptrons (MLPs), radial basis function (RBF) and Kohonen's self-organizing maps (SOM), are applied in order to assess the quality of the water reservoirs used for the domestic and industrial water supply of the city of Athens, Greece. In parallel, ANN models are optimized and their recognition and predictive accuracy is tested. The data set consisted of 89 samples collected from the three Athenian water reservoirs during a period of 6 months (October 2006 to April 2007). Thirteen metals and metalloids, Fe, B, Al, V, Cr, Mn, Ni, Cu, Zn, As, Cd, Ba, Pb, were determined. For the validation of the optimized ANN models, new data from subsequent sampling campaigns (December 2007) were used. The constructed classification models predicted successfully the origin of the new posterior samples and simultaneously revealed the differences in sample compositions that occurred in that period. Critical comparison of the different architectures in site classification and modeling verified the validity and usefulness of ANNs, as a powerful and effective tool for water quality assessment.

Key words | artificial neural networks, chemometrics, classification, Kohonen, prediction, water quality

Eleni G. Farmaki
Nikolaos S. Thomaidis (corresponding author)
Constantinos E. Efstathiou
Laboratory of Analytical Chemistry,
Department of Chemistry,
National and Kapodistrian University of Athens,
Panepistimioupolis Zografou, 15771 Athens,
Greece
E-mail: *ntho@chem.uoa.gr*

Eleni G. Farmaki
Athens Water Supply and Sewerage Company (EYDAP SA),
Quality Control Division,
Acharnes Attikis,
Greece

Vasil Simeonov
Chair of Analytical Chemistry,
Faculty of Chemistry,
University of Sofia 'St. Kl. Okhridski',
J. Bourchier Blvd 1,
Sofia 1164,
Bulgaria

INTRODUCTION

Authorities and research bodies all over the world perform regular monitoring studies which determine a high number of parameters in order to ensure a high quality of the supplied water. These studies often produce large data sets from which only a small amount is really relevant to the problem (Zupan & Gasteiger 1993). Indeed, the valuable information that scientists seek can often be very difficult to extract from the abundant data obtained. Thus, although some decades ago, scientists had to carry out a great deal of hard routine work to obtain just a few numbers, nowadays it can be arduous to explore the available information to find what is valuable. To this end, and in order to be prepared to handle large quantities of data, artificial neural networks (ANNs) seem to be a promising and effective tool.

Inspired initially from biological systems, ANN models are capable of gradual learning over time and modeling extremely complex functions. Their contribution in handling large data sets of results is very important. In addition to the traditional multivariate chemometric techniques, ANNs are often applied for function approximation, or regression analysis, time series prediction, fitness approximation and modeling, clustering, classification, novelty detection, data and pattern recognition (Farmaki *et al.* 2010). ANNs can detect complex relationships between inputs and outputs or recognize patterns in data structure. The success key is always an appropriate training data.

ANNs have seen an explosion of interest over the last two decades and have been successfully applied in all fields of chemistry and particularly in analytical chemistry.

doi: 10.2166/aqua.2013.108

In water analysis particularly, a relatively high number of articles concerning ANN applications have been published during the last decade. Optimized models, mainly multilayer perceptrons (MLPs), radial basis function (RBF) and Kohonen, are often used for modeling and prediction (Brodnjak-Vončina *et al.* 2002; Huang & Foo 2002; Sharma *et al.* 2003; Fernández-Sánchez *et al.* 2004; Sahoo *et al.* 2005; Kim & Kim 2007; Elhatip & Kömür 2008; Rene & Saidutta 2008), water quality assessment (Brodnjak-Vončina *et al.* 2002; Tutu *et al.* 2005; Astel *et al.* 2007; Tobiszewski *et al.* 2010; Tsakovski *et al.* 2010; Bieroza *et al.* 2011), sample classification (Astel *et al.* 2007; Çinar & Merdun 2009; Yan *et al.* 2010; Jin *et al.* 2011), or for exploring correlation and significance among variables (Brodnjak-Vončina *et al.* 2002; Astel *et al.* 2007; Çinar & Merdun 2009; Jin *et al.* 2011), and are frequently compared to more conventional statistical techniques (Brodnjak-Vončina *et al.* 2002; Bieroza *et al.* 2011). In all works, ANNs seem to be very effective and produce better results than the traditional chemometric techniques during comparison studies (Brodnjak-Vončina *et al.* 2002; Bieroza *et al.* 2011).

In this work, water quality with respect to the metal and metalloid content in samples collected from the three main water reservoirs of Athens (Iliki, Mornos and Marathon) is evaluated by ANNs. Moreover, the comparative application of different architectures for the evaluation of surface water quality is comprehensively presented, showing that each of them could reveal and verify hidden intrinsic characteristics and develop simple classification rules that could be used, either to identify potential changes in a posterior samples' composition, or to prove homogeneity of a lake, thus facilitating the sampling authority (EYDAP company) in sampling and analysis management.

Specifically, data concerning metal and metalloid concentrations from a total of 15 sampling sites in Iliki (three sites), Mornos (seven sites) and Marathon (five sites) were subjected to different ANN algorithms in order to classify the sampling sites, find the critical variables that are responsible for this spatial variability and construct models that characterize each individual water reservoir. Particularly, the objectives of the data processing were:

1. to apply and compare different ANN architectures through their predictive ability using 'unknown' samples;
2. to construct models for each lake containing the critical parameters that define their water quality; and
3. to reveal the similarities and dissimilarities between sampling points of a reservoir and between Athenian reservoirs. In that way, either pollution can be identified, or the homogeneity of the lakes can be examined.

For these purposes, the ANN models have been successfully optimized, while validation and external test samples verified their efficiency.

EXPERIMENTAL

Monitoring sites and sampling campaigns

Monitoring sites are thoroughly presented in our previous work (Farmaki *et al.* 2012). Three sites have been selected for Iliki, seven for Mornos and five in Marathon lake. The three water Athenian reservoirs are depicted in Figure 1, Iliki (Y), Mornos (MO) and Marathon (MA).

Water quality of Iliki (marked as Y in Figure 1) is dictated by the Kifisos River that crosses all the Kopaida plain and discharges in the northwest side of the lake. The plain is cultivated intensively; the main products being cotton, tobacco, olives, cereals, legumes, vegetables and animal products. Big municipalities dominate in the surrounding area, while cotton ginning, textile and agrochemical factories and building materials production units are active. Mining industries (bauxite, iron ore) are also well developed.

Mornos (marked as MO) accepts the water bodies of all the rivers, tributaries, and streams of the region, while Marathon (marked as MA) basin is surrounded by five municipalities. The river water of the area determines the quality of Marathon water.

Fifteen sampling sites were selected on the three lakes under the quality control monitoring program of EYDAP SA. A description of the sampling sites is given in Table 1.

Six monthly sampling campaigns were conducted, between October 2006 and April 2007. Posterior samples for validation purposes were collected in December 2007. Details about the collection and preservation of samples can be found elsewhere (Farmaki *et al.* 2012).

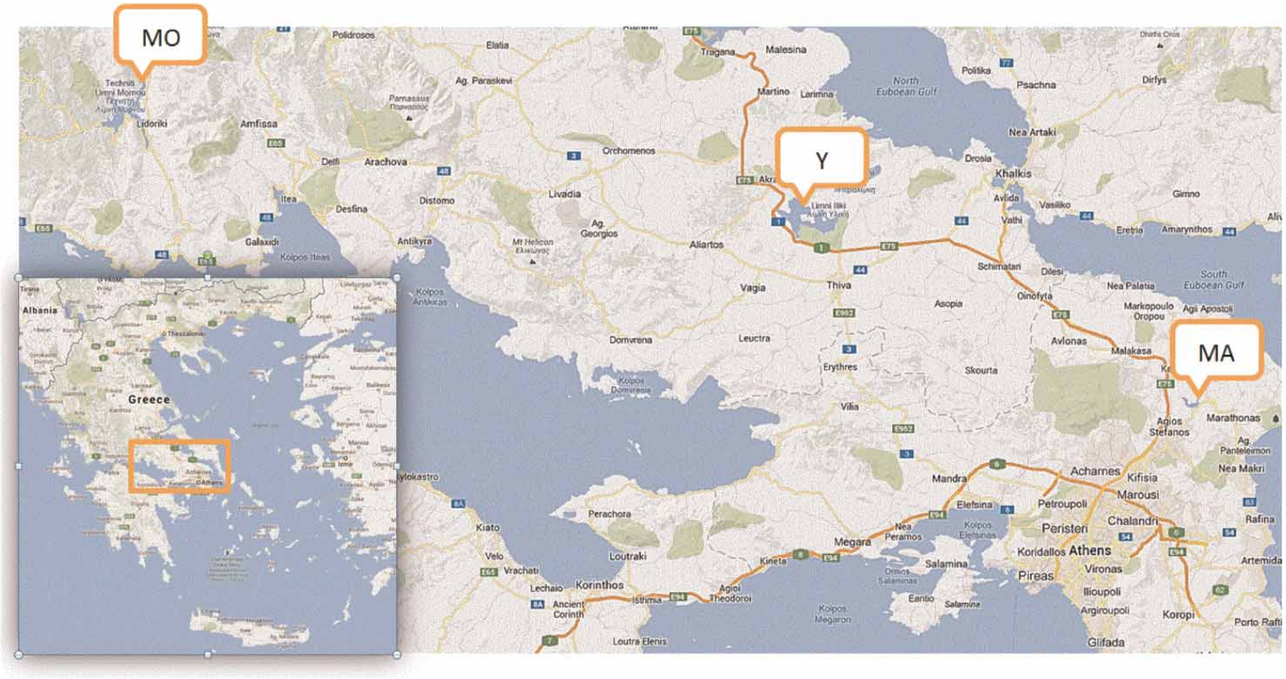

Figure 1 | Map of water reservoirs of Athens.

Instruments and methods

Fe was determined by Electrothermal Atomic Absorbance Spectrometry (ETAAS) (Perkin-Elmer, model AAnalyst 800, Bodenseewerk, Germany). For the rest of the elements, B, Al, V, Cr, Mn, Ni, Cu, Zn, As, Cd, Ba, Pb, Inductively Coupled Plasma Mass Spectrometry (ICP-MS) (Agilent, model 7500e, Santa Clara, California, USA) was used.

Table 1 | Sampling sites description

Sampling site/No	Reservoir	Description
1	Iliki	River Kifisos estuary
2	Iliki	Center of the lake (shore side)
3	Iliki	Mouriki
4	Mornos	River Mornos estuary
5	Mornos	River Avoros estuary
6	Mornos	Center of the lake (shore side)
7	Mornos	City of Lidoriki
8	Mornos	Pump-station
9	Mornos	Katadi (stagnant water)
10	Mornos	River Kokinos estuary
11	Marathon	Pump-station
12	Marathon	Inflow of stream 1 (from Mornos)
13	Marathon	Inflow of stream 2
14	Marathon	Inflow of stream 3
15	Marathon	Inflow of stream 4

A detailed description of the analytical data quality (calibration data, limit of detection (LOD) and limit of quantification (LOQ) values, precision and trueness data) is presented in our previous work (Farmaki *et al.* 2012). All measurements were performed in triplicate and the average was used.

Data analysis and statistical methods

Thirteen metals and metalloids were initially determined in a total of 89 samples. Finally, 11 variables were retained. Cd and Pb measurements have been omitted, since the majority of the results were below or around LOD level, respectively.

The results of all measurements were analyzed by different ANN algorithms. The whole raw data set was used with lake-marked sampling sites, after substituting the respective numbers of the sampling sites with the respective reservoir name.

In the following sections, a comprehensive summary of the theory of the ANN architectures used is presented. More details, along with a review of applications in water analysis, can be found in a previous work (Farmaki *et al.* 2010). MLP and RBF models were applied on the same data set and their predictive assessment is thoroughly discussed, while Kohonen networks succeeded in revealing site similarities (within the same lake) or differences (between different lakes).

THEORY

General

A 'primitive' unit called *perceptron* based on an innovative idea of Rosenblatt's (Rosenblatt 1958) was the beginning for the development of ANNs. *Elementary Perceptron* is a binary classifier that combines inputs to binary outputs. For every input vector x_i, a weight w_i is applied, multiplied with the corresponding value, a constant term (bias b) is added, and the sum y is finally calculated by the formula (Farmaki *et al.* 2010):

$$y = \sum_{i=1}^{n} x_i w_i + b$$

A threshold function f is then applied and a final threshold value θ determines the output:

$$f(x) = \begin{cases} 1 & \text{if } y > \theta \\ 0 & \text{otherwise} \end{cases}$$

This type of network, however, could only respond in limited linear problems. A large class of problems described by non-linear functions could not be resolved; the solution was given by adding an intermediate (hidden) layer and new multi-layer networks (MLPs) were developed. The last ones are non-linear statistical data modeling tools that can handle successfully non-linear problems that a simple perceptron cannot (Farmaki *et al.* 2010).

The relationship between inputs and outputs in a network is established through the process of 'training' or 'learning'. In supervised learning, a set of example pairs (inputs X combined to outputs Y) is given. This comprises the '*training*' data set and it is used for 'modeling' the relationship between inputs and outputs. In other words, the aim is to find a function (X → Y) that matches the examples. For this purpose, the initial weights are adjusted so that the calculated error (difference between outputs and desired targets) is as low as possible. Training is usually terminated when the error is lower than a predetermined value or a predefined number of epochs are completed. The model is then checked for its generalization ability. A special sample set is used for this purpose called '*validation*' or '*selection*' set. This sample set can check the network performance and control the 'overfitting' problem. An overfitted model is a complicated model that generally has too many parameters relative to the number of samples. Overfitting occurs when a model describes random error or data noise, instead of approximating the underlying function. Thus, the model function is too closely fitted to a limited set of data points (the training data); as a result, it memorizes the idiosyncrasies in these and it is 'built' upon them. An overfitted model is a model based on limited data. So, when it will be asked inevitably to judge over other unknown pairs of inputs and outputs, it will fail. This means poor generalization (less predictive power as the model exaggerates minor fluctuations in the data). A simpler model can generalize over unknown samples, but it may not be powerful to model the data. Thus, a balance between model complication and generalization ability is required. The efficiency of the trained network finally, can be verified by a new unknown sample set called '*test*' set. If the network is properly trained, the model can also predict correctly the outputs for this unknown set.

Unsupervised training (represented here by the Kohonen technique) refers to the problems of trying to find hidden structure in unlabeled data. The network learns to represent particular input patterns in a way that reflects the initial data. The system is provided with a sample set and is left to settle down (or not) without a known desired output (Svozil *et al.* 1997).

Back propagation algorithm

MLP network is the most frequently used type of ANN for approximating general relationships. The basic structure is comprised generally by three layers of units (neurons): the input that does not perform any calculations; the intermediate (hidden) that performs the computation of the weighted sum of the inputs and the implementation of the activation function; and the output layer that produces the final results by treating the outputs of the hidden layer, with the same way. Less frequently, more than one hidden layer can be used.

The most popular learning algorithm used here is referred to as the back-propagation algorithm (BP), as the error is propagated (distributed) from the output to the input layer. When a sample is entered into the network, the initial input values are propagated (through the hidden layer) forward to the output units. This is where the final error, i.e. the differences between the computed and required (theoretical) values, can be calculated. Thus, based on these differences, the adjustment of the output weights is initially performed. This adjustment is propagated backwards to the weights of the hidden layer or finally the input one, so that all the weights are adjusted. Many iterations are performed for an effective error reduce. More details about MLPs and BP algorithms are presented elsewhere (Nguyen *et al.* 2003; Farmaki *et al.* 2010).

Radial basis function networks

The RBF network also consists of three layers: an input, a hidden and an output. Each input neuron is connected to all the hidden ones, while hidden and outputs are interconnected to each other by a set of weights (Sharma *et al.* 2003). The neurons in the hidden layer usually contain Gaussian activation functions whose outputs are inversely proportional to the distance from the center of the neuron. This means that every time an input enters the network, the Euclidean distance is computed between the center of every neuron and the input vector. Then a Gaussian function is applied. When an input is far from the specific neuron, the Gaussian response is small. On the contrary, for each input sample that is closer to the center, the response is significant. As a result, there is a predefined range for the inputs that corresponds to a response field (center ± width) for every neuron. Thus, the key for a successful RBF network is the appropriate choice of the centers and widths of the neurons (Vandeginste *et al.* 1998), while it is evident that extrapolation in this case is prohibited.

Finally, in the output layer, each unit makes a linear transformation to the data of the hidden layer. In other words, the final predicted value for the input vector is computed by summing the output values of all the RBF functions in the hidden neurons.

Kohonen neural networks

Kohonen neural networks differ from the other supervised learning architectures. Self-organizing map (SOM) algorithm has been proposed by Kohonen (Kohonen 1982), and is a neural network model that implements a characteristic non-linear projection from the high dimensional space of input signals onto a low-dimensional array of neurons (Kohonen *et al.* 1996). The term 'self-organizing' refers to the unsupervised ability to learn and organize information without being given combined output values for the input pattern (Mukherjee 1997). A Kohonen network consists of two levels. The first one is the input level. The second is usually organized on a regular two-dimensional grid. This is usually a rectangular surface with $m \times m$ neurons (units). The two levels are fully interconnected; so each input unit is connected to all other neurons of the second level. If the input vector has n dimensions, we have a number of $n \times m \times m$ connections. Kohonen neural networks resemble most the biological networks due to the correction implementation that seems to be a rather 'local' procedure. The weights correction affects only some of the neurons of the network that resemble the 'winner' or the 'best matching unit' (BMU). This procedure finally dictates the topology of the Kohonen 'map' and thus similar objects (in our case sampling points) are mapped close together on the grid.

The first step in training a Kohonen network is to initialize the weight vectors. Each weight vector has two components: the first part of a weight vector is its data (input sample vector), while the second part of a weight vector is its natural location (a neuron in the Kohonen map). When an input vector enters the network, the map is searched for weight vectors (of the winner neuron) that best represent that input. Distances between the input and all the neurons are calculated, but the winner represents the minimum one. Thus, the winner is rewarded by being able to adjust its weights to be closer to the first entered input vector. The new weights are calculated from the old ones through a learning rate, initially defined and are updated after each iteration. In parallel, the neighbors of that neuron are also rewarded by being able to become closer to the chosen input vector. The number of neighbors selected to update their weights is defined by an appropriate neighborhood function that ensures their decrement over time. The input vector is finally attributed to its winner neuron, while the whole process is repeated for the next sample. In this way, the Kohonen algorithm constructs a neuron map that represents the whole data set and reflects their initial topology.

All the calculations and plots of this work were made using Excel 2003 by MicroSoft, Statistica for Windows, Version 7.0 by StatSoft Inc., 2004 and Matlab 6.5 software.

RESULTS AND DISCUSSION

Results of the sampling campaigns

The overall results are presented in our previous work (Farmaki *et al.* 2012). Box plots for some critical variables (Fe, Al, V, Mn, Ni, As) for the three reservoirs are presented in Figure 2.

As expected, due to the intense anthropogenic (agricultural and industrial) activities, Iliki seemed to be the most polluted with regard to the elemental concentrations. Thus, in the three sampling sites of Iliki (numbers 1, 2, 3 in Table 1) slightly higher values of Fe, B, Al, V, Mn, Ni and Zn were determined (however, below the regulated limits). On the contrary, Mornos, which is the major water supply reservoir of Athens, showed ‘background level’ values for all the determined elements. Marathon lake, due to the surrounding agricultural activities and natural processes in the wider area, gave higher values for Fe, Al, Mn, Ni and As (Figure 2).

MLP-BP network

MLP networks used in this work were optimized through the parameters of the number of the hidden units and the inputs (metal and metalloids), the learning rate and the size of the training set. The criterion used was the RMS (Root Mean Square) error for the validation sample set. For each trial, 20 different networks were tested. Thus, the initial parameters for the model construction are different and independent in order to validate the final result and avoid local minima and paralyzed networks (Kröse & Van der Smagt 1996; Vandeginste *et al.* 1998; Hernández-Caraballo *et al.* 2005; Carlucci *et al.* 2007).

Moreover, in order to optimize the number of the inputs, discriminant analysis (DA) was used as a preliminary tool (Farmaki *et al.* 2012). Thus, variables sets of three (V, Ni, As), four (V, Ni, As, B), six (V, Ni, As, B, Cu, Mn), eight (V, Ni, As, B, Cu, Mn, Cr, Fe) and 11 (V, Ni, As, B, Cu, Mn, Cr, Fe, Ba, Al, Zn) were tested according to the importance order that had derived from an initial application of the standard approach of the DA method. The number of hidden units fluctuated from two to 18, while the learning rate and the size of the training set were tested from an initial value of 0.01–0.20 and from an initial value of 35–65, respectively. Finally, the best model was chosen, with RMS error = 0.26 for the validation set, 12 units in the hidden layer and only three inputs (V, Ni, As) (Figure 3(a)). Their contribution in the discrimination is explained due to the high V and As values in Iliki and Marathon, respectively, while Mornos is characterized by nearly background values for all three descriptors (Farmaki *et al.* 2012). Figure 4 is indicative for this differentiation. More specifically, vanadium seems to be the discriminating element for Iliki (Y), as the red color of ‘V’ and ‘Y’ nodes shows in Figure 3(a) (the full color version of Figure 3 is available online at http://www.iwaponline.com/jws/toc.htm). Indeed, the colors represent the distribution ratio of each variable in the final result (Galão *et al.* 2011). Even the red hidden nodes are characterized by the higher weight values for vanadium. The final outputs for the MLP network are the three lakes (marked in Figure 3(a)). The learning rate and the size of the training set were 0.01 and 65, respectively, during the learning phase.

The accuracy of the constructed MLP model was assessed with a series of new results received during a posterior time period (December 2007) from the same sampling sites. Table 2 summarizes the results. The first nine columns show the measured concentrations of V, Ni and As of the new samples, while the real origin of the samples (MO, MA or Y) is provided in the first line of the table. The predicted results according to the models used

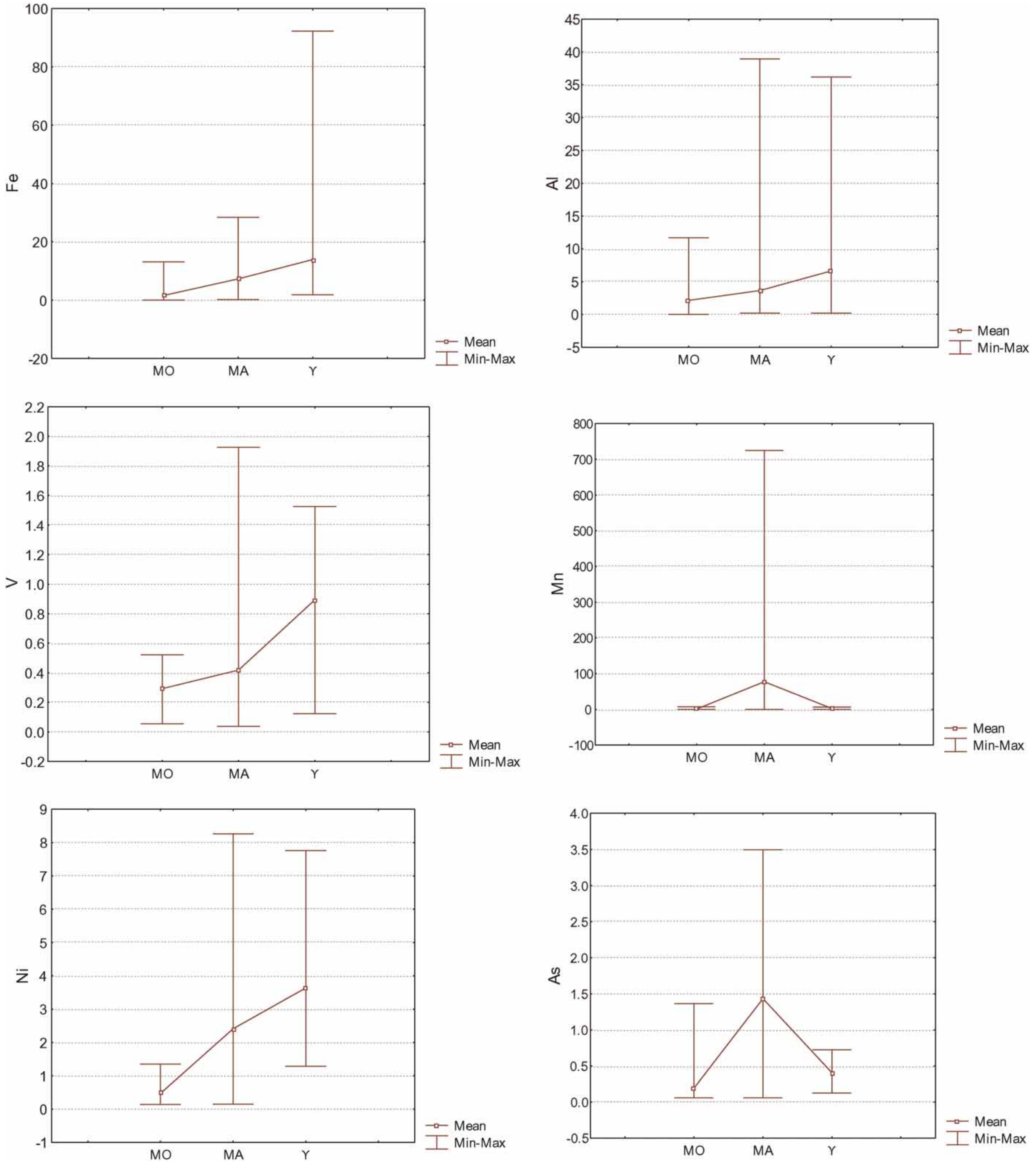

Figure 2 | Box plots of some critical parameters of the three lakes: Fe, Al, V, Mn, Ni, As (in μg L^{-1}; mean and minimum/maximum values are presented).

are presented in the last two columns. There was only one error in a set of 14 samples. This mistakenly predicted site (no. 12 of Marathon, Table 1) contains water from a stream coming from Mornos. However, after October 2007, due to water shortage, the supply of Mornos was completed (half amount) with Iliki water. As a result, the water quality in this site represented equally Mornos and Iliki water. The 'mixing' of sites is thus expected. So, with the

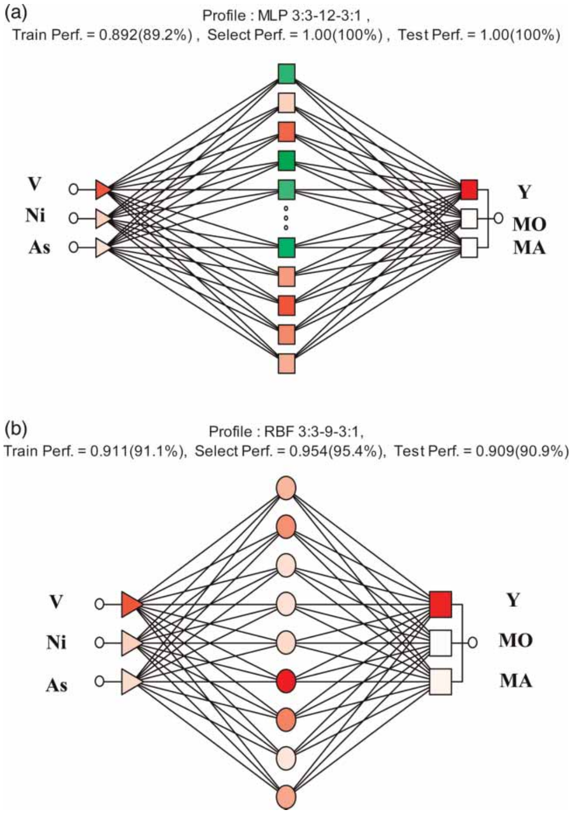
(a)
Profile : MLP 3:3-12-3:1 ,
Train Perf. = 0.892(89.2%) , Select Perf. = 1.00(100%) , Test Perf. = 1.00(100%)
V
Ni
As
Y
MO
MA
(b)
Profile : RBF 3:3-9-3:1,
Train Perf. = 0.911(91.1%), Select Perf. = 0.954(95.4%), Test Perf. = 0.909(90.9%)
V
Ni
As
Y
MO
MA

Figure 3 | Architecture of the final optimized networks: (a) MLPs and (b) RBF. The full color version of this figure is available online at http://www.iwaponline.com/jws/toc.htm.

use of the constructed model, one could predict the origin of the water in the canal.

RBF network

RBF networks were optimized through the parameters of the number of the hidden units and the width. The criterion used was again the RMS error for the validation sample set. The variables used were V, Ni, As, as they had already been evaluated in the previous ANN technique, in order that comparisons were feasible. Figure 3(b) depicts the three colored inputs (V, Ni and As) and the corresponding outputs (MO, MA or Y). For each trial, different networks were also tested (the number of hidden units fluctuated from two to 22), and the best model was chosen with RMS error = 0.23 for the validation set and nine units in the hidden layer (Figure 3(b)). Its accuracy was also confirmed with the same series of new data set of December 2007. The results were exactly the same with the MLPs model (Table 2): the same controversial site gives the only wrong prediction. However, fewer hidden units have been used in the optimized model.

In both supervised ANN models, the differentiation between the three lakes was proved through two successful models. The critical variables (responsible for this

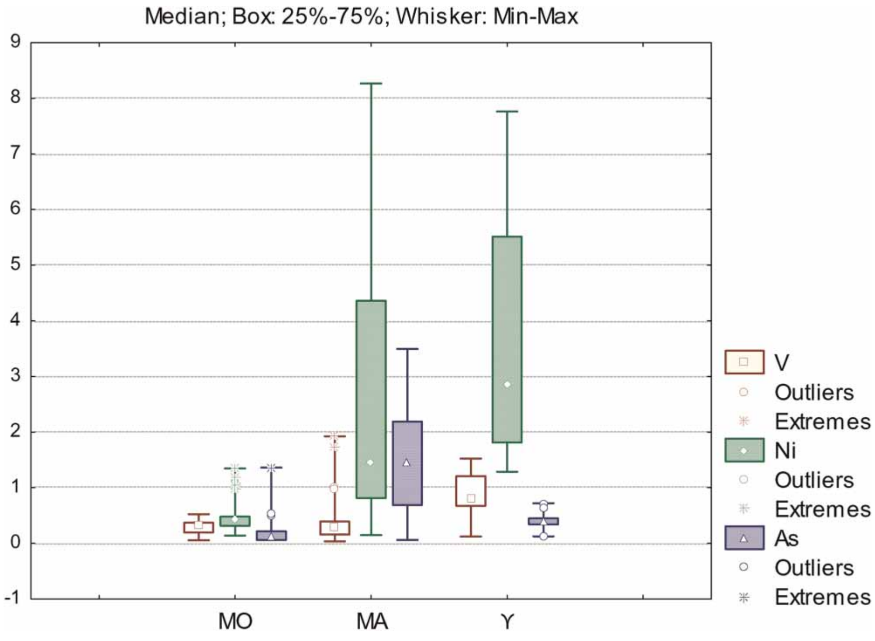

Figure 4 | Box-and-whisker plots of the most critical values: V, Ni, As (in μg L^{-1}) used for the discrimination of the three lakes (median and minimum/maximum values are given).

Table 2 | Predictions in new samples based on the optimized MLP and RBF model. The variables values are recorded in $\mu g\, L^{-1}$

Mornos (MO)			Marathon (MA)			Iliki (Y)				
V	Ni	As	V	Ni	As	V	Ni	As	Prediction	
0.42	1.16	0.25							MO	√
0.20	0.37	0.16							MO	√
0.70	0.43	0.41							MO	√
0.78	0.41	0.42							MO	√
0.66	0.33	0.33							MO	√
0.42	0.35	0.19							MO	√
0.39	0.33	0.22							MO	√
			1.22	2.02	2.46				MA	√
			1.13	3.07	1.04				Y	X
			0.56	0.73	1.96				MA	√
			0.61	1.87	2.34				MA	√
						1.41	2.38	0.62	Y	√
						0.91	3.38	0.80	Y	√
						0.67	6.83	0.59	Y	√

differentiation) were identified, while the homogeneity of each lake was obvious, especially that of Mornos lake. This can justify fewer sampling points per lake and consequently less laboratory and statistical work.

Kohonen network

The unsupervised Kohonen technique was applied to the data set by using all the initial variables as this technique is different from the aforementioned ones (it is an unsupervised ANN technique).

The Kohonen map has been chosen as a rectangular grid with number of neurons (n) determined using the following formula (Vesanto 2000):

$$n = 5 \times \sqrt{\text{number of samples}} = 5 \times \sqrt{89} \approx 48$$

In Figure 5 (component planes), the general Kohonen classification of all samples (89) for all chemical variables (11) is presented. It is readily seen that similar distribution patterns are formed for the variables Fe and Al with slight resemblance with Cu and B, then V and Cr, Ni and As, Mn and Zn. The distribution for Ba is different from the other patterns. Moreover, the discriminating power of some variables mentioned above in the previous techniques is confirmed here. Thus, vanadium component plane gives dark red neurons in the left-bottom area that represents Iliki (see V planes in Figure 5), arsenic component plane gives dark red neurons in the right-bottom area that represents Marathon (see As planes in Figure 5), while the neurons that depict Mornos are blue (no particular contaminating element; background levels). (The full color version of Figure 5 is available online at http://www.iwaponline.com/jws/toc.htm.)

Then, similarity between the sampling sites was investigated by using the Kohonen network. The dimensionality of the Kohonen's map was 6 × 8 (see 'hit diagram' in Figure 6). The Kohonen classification obtained was then compared with the real data from the initial data set (all 89 samples of the three lakes, Figure 6). The classification indicates that three major groups of objects are formed.

1. In Group 1 (total 20 objects), Marathon lake is represented by 18 objects and lake Iliki by two objects. This group is characterized by elevated concentrations of Mn, Ni, Zn, As and Ba (anthropogenic origin for the first four of them, while Ba is attributed to the geological background).
2. In Group 2 (total 17 objects), again Marathon and Iliki are grouped together; however, only one object is

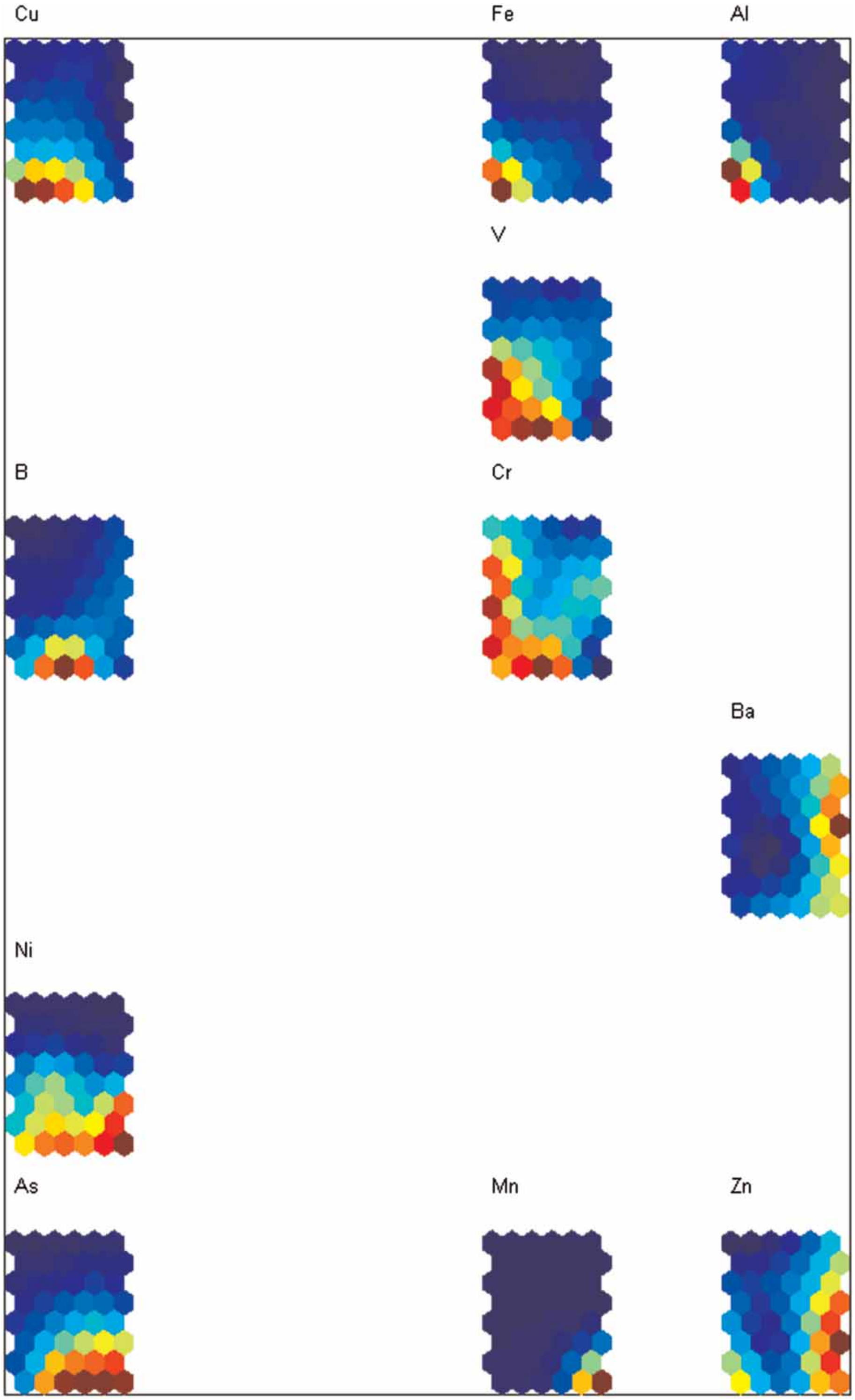

Figure 5 | Kohonen network: component planes for all variables (11) and all samples (89). The full color version of this figure is available online at http://www.iwaponline.com/jws/toc.htm.

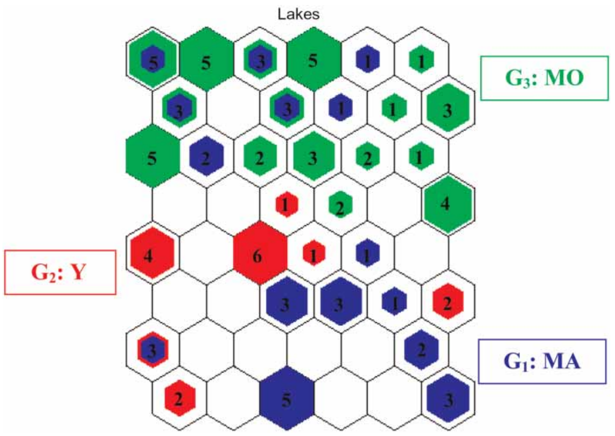

Figure 6 | Kohonen network: hit diagram with all samples (89) and separation by sampling sites, all sites (Iliki – RED, Mornos – GREEN, Marathon – BLUE). The full color version of this figure is available online at http://www.iwaponline.com/jws/toc.htm.

assigned in Marathon and 16 objects are assigned to Iliki. This group is characterized by elevated concentrations of Fe, B, Al, V, Cr and Cu (anthropogenic inputs due to agricultural and/or industrial activities).

3. In Group 3 (total 52 objects), one could find again objects from two lakes (Mornos and Marathon). However, here, Mornos objects are in the majority (42) and the rest are derived from Marathon (10). These are background objects with lowest concentrations of the metals.

More details about the origin of the metals and metalloids (source identification through factor analysis can be found in previous work; Farmaki *et al.* (2012)).

It is apparent that the three lakes are mixed within the groups. The majority of the neurons seem 'pure', as they accept hits by one lake. Thus, only five neurons from a total of 33 populated neurons (Figure 6) accept hits from two lakes (four neurons are hit by Marathon and Mornos and one from Marathon and Iliki). The first four of them are hit by the sample from site no. 12 of Marathon (Table 1) that during the sampling period contained water from a stream coming from Mornos. As a result, the water quality in this site represented Mornos (although nominally is recorded as Marathon). The 'mixing' of sites is thus expected. So, Group 3 seems to contain only Mornos samples.

Generally, Mornos differs in quality from the other two lakes, being less polluted. The other two lakes are influenced by natural (geological) and anthropogenic (industrial and agricultural) sources. Indeed, Mornos accepts mainly rainfall and water from all the surrounding water bodies (rivers, tributaries and streams). On the contrary, Iliki is the major receiving body of all the run-off waters in the Kopaida plain that is intensively cultivated. Additionally, adjacent mining industries affect the water quality of the lake. Marathon is surrounded by small municipalities and agricultural areas are moderately cultivated.

CONCLUSIONS

ANNs (three different architectures) have been applied in the same data set concerning metal and metalloids from the three water reservoirs of Athens. Optimized BP-MLP, RBF and Kohonen models allowed the discrimination of surface water samples from the reservoirs of Athens, according their origin, using only three parameters (V, Ni and As) or all of them (Kohonen model).

Particularly, for the BP-MLP and RBF models, only one failure out of the 14 new samples (not used in the models' construction) was recorded. This deviation was absolutely justified due to the composition change of a specific sample from Marathon lake. The ANN models gave excellent results having exploited only three variables. Obtaining results for models with only three variables (compared to 11 or initially 13 elements) is economically favored, especially when routine monitoring is considered.

Concluding, supervised ANN techniques succeeded in: (1) successfully assigning the sampling sites into groups; (2) classifying new samples by constructing a robust and accurate model; and (3) evaluating the critical variables. The optimization processes on ANN models can assess the available variables, eliminating the abundant ones.

The Kohonen unsupervised technique seemed to surpass the other tested ANN techniques, having excellent visualization abilities and better interpreting the initial information. Generally, Kohonen technique may project variables and objects in a two-dimensional space preserving the topological structure, while in the case of linear principal component analysis (PCA), this does not always prove to be successful (Brodnjak-Vončina *et al.* 2002; Marini 2009). Since Kohonen networks are a non-linear mapping technique, there can be cases where they provide a clearer separation among the samples than PCA. Compared with CA, Kohonen networks can be applied as a supervised technique and classify new samples. Kohonen networks can model and consequently classify an unknown sample in the area of the already designed map.

Concluding, ANN models, free of traditional assumptions (like normal distribution or an abundant number of variables), seem to be suitable for complex non-linear problems concerning prediction, modeling and classification.

REFERENCES

Astel, A., Tsakovski, S., Barbieri, P. & Simeonov, V. 2007 Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Res.* **41**, 4566–4578.

Bieroza, M., Baker, A. & Bridgeman, J. 2011 Classification and calibration of organic matter fluorescence data with multiway analysis methods and artificial neural networks: an operational tool for improved drinking water treatment. *Environmetrics* **22**, 256–270.

Brodnjak-Vončina, D., Dobčnik, D., Novič, M. & Zupan, J. 2002 Chemometrics characterisation of the quality of river water. *Analyt. Chim. Acta* **462**, 87–100.

Carlucci, G., D'Archivio, A. A., Maggi, M. A., Mazzeo, P. & Ruggieri, F. 2007 Investigation of retention behaviour of non-steroidal anti-inflammatory drugs in high-performance liquid chromatography by using quantitative structure–retention relationships. *Analyt. Chim. Acta* **601**, 68–76.

Çinar, Ö. & Merdun, H. 2009 Application of an unsupervised artificial neural network technique to multivariant surface water quality data. *Ecol. Res.* **24**, 163–173.

Elhatip, H. & Kömür, M. A. 2008 Evaluation of water quality parameters for the Mamasin dam in Aksaray City in the central Anatolian part of Turkey by means of artificial neural networks. *Environ. Geol.* **53**, 1157–1164.

Farmaki, E. G., Thomaidis, N. S. & Efstathiou, C. E. 2010 Artificial neural networks in water analysis: theory and applications. *Int. J. Environ. An. Ch.* **90**, 85–105.

Farmaki, E. G., Thomaidis, N. S., Simeonov, V. & Efstathiou, C. E. 2012 A comparative chemometric study for water quality expertise of the Athenian water reservoirs. *Environ. Monit. Assess.* **184**, 7635–7652.

Fernández-Sánchez, J. F., Carretero, A. S., Benítez-Sánchez, J. M., Cruses-Blanco, C. & Fernández-Gutiérrez, A. 2004 Fluorescence optosensor using an artificial neural network for screening of polycyclic aromatic hydrocarbons. *Analyt. Chim. Acta* **510**, 183–187.

Galão, O. F., Borsato, D., Pinto, J. P., Visentainer, J. V. & Carrão-Panizzi, M. C. 2011 Artificial neural networks in the classification and identification of soybean cultivars by planting region. *J. Brazil Chem. Soc.* **22**, 142–147.

Hernández-Caraballo, E. A., Rivas, F., Pérez, A. G. & Marcó-Parra, L. M. 2005 Evaluation of chemometric techniques and artificial neural networks for cancer screening using Cu, Fe, Se and Zn concentrations in blood serum. *Analyt. Chim. Acta* **533**, 161–168.

Huang, W. & Foo, S. 2002 Neural network modeling of salinity variation in Apalachicola River. *Water Res.* **36**, 356–362.

Jin, Y.-H., Kawamura, A., Park, S.-C., Nakagawa, N., Amaguchi, H. & Olsson, J. 2011 Spatiotemporal classification of environmental monitoring data in the Yeongsan River basin, Korea, using self-organizing maps. *J. Environ. Monit.* **13**, 2886–2894.

Kim, M.-Y. & Kim, M.-K. 2007 Dynamics of surface runoff and its influence on the water quality using competitive algorithms in artificial neural networks. *J. Environ. Sci. Heal. A.* **42**, 1057–1064.

Kohonen, T. 1982 Self-organizing formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59–69.

Kohonen, T., Oja, E., Simula, O., Visa, A. & Kangas, A. J. 1996 Engineering applications of the self-organizing map. *Proc. IEEE* **84**, 1358–1384.

Kröse, B. & Van der Smagt, P. 1996 *An Introduction to Neural Network*. The University of Amsterdam, Amsterdam.

Marini, F. 2009 Artificial neural networks in foodstuff analyses: trends and perspectives. A review. *Analyt. Chim. Acta* **635**, 121–131.

Mukherjee, A. 1997 Self-organizing neural network for identification of natural modes. *J. Comput. Civil. Eng.* **11**, 74–77.

Nguyen, H. T., Prasad, N. R., Walker, C. L. & Walker, E. A. 2003 *A First Course in Fuzzy and Neural Control*. Chapman & Hall/CRC, Boca Raton, Florida.

Rene, E. R. & Saidutta, M. B. 2008 Prediction of water quality indices by regression analysis and Artificial Neural Networks. *Int. J. Environ. Res.* **2**, 183–188.

Rosenblatt, F. 1958 The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386–408.

Sahoo, G. B., Ray, C. & Wade, H. F. 2005 Pesticide prediction in ground water in North Carolina domestic wells using artificial neural networks. *Ecol. Model.* **183**, 29–46.

Sharma, V., Negi, S. C., Rudra, R. P. & Yang, S. 2003 Neural networks for predicting nitrate-nitrogen in drainage water. *Agr. Water Manage.* **63**, 169–183.

Svozil, D., Kvasnička, V. & Pospíchal, J. 1997 Introduction to multi-layer feed-forward neural networks. *Chemometr. Intell. Lab.* **39**, 43–62.

Tsakovski, S., Tobiszewski, M., Simeonov, V., Polkowska, Z. & Namieśnik, J. 2010 Chemical composition of water from roofs in Gdansk, Poland. *Environ. Pollut.* **158**, 84–91.

Tobiszewski, M., Tsakovski, S., Simeonov, V. & Namieśnik, J. 2010 Surface water quality assessment by the use of combination of multivariate statistical classification and expert information. *Chemosphere* **80**, 740–746.

Tutu, H., Cukrowska, E. M., Dohnal, V. & Havel, J. 2005 Application of artificial neural networks for classification of uranium distribution in the Central Rand goldfield, South Africa. *Environ. Model. Assess.* **10**, 143–152.

Vandeginste, B. G. M., Massart, D. L., Buydens, L. M. C., De Jong, S., Lewi, P. J. & Smeyers-Verbeke, J. 1998 *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier, Amsterdam.

Vesanto, J. 2000 Neural network tool data mining: SOM Toolbox. Proceedings of Symposium on Tool Environments and Development Methods for Intelligent Systems. (TOOLMET2000), Oulun yliopistopaino, Oulu, Finland, pp. 184–196.

Yan, H., Zou, Z. & Wang, H. 2010 Adaptive neuro fuzzy inference system for classification of water quality status. *J. Environ. Sci.* **22**, 1891–1896.

Zupan, J. & Gasteiger, J. 1993 *Neural Networks for Chemists; An Introduction*. VCH Verlagsgesellschaft, Weinheim, Germany and VCH Publishers, New York.

First received 15 September 2012; accepted in revised form 12 April 2013