COMPARATIVE USE OF ARTIFICIAL NEURAL NETWORKS FOR THE DISCRIMINATION OF THE WATER RESERVOIRS OF ATHENS ACCORDING TO THEIR ELEMENTAL CONTENT

Eleni G. FARMAKI1,2, Nikolaos S. THOMAIDIS1,* and Constantinos E. EFSTATHIOU1

1Laboratory of Analytical Chemistry, Department of Chemistry, University of Athens, Panepistimioupolis Zografou, 15771 Athens, Greece.
2Athens Water Supply and Sewerage Company (EYDAP SA), Water Quality Control and Protection Division, Polydendri Attikis, Greece.
*e-mail: ntho@chem.uoa.gr

EXTENDED ABSTRACT

This work aims at the comparative use of three different Artificial Neural Networks (ANNs) techniques for the water quality assessment and classification of the water reservoirs of Athens, Greece. During the period of October 2006 to April 2007, 89 samples were collected from the three lakes (Iliki, Mornos and Marathon) used as source for the domestic and industrial water supply of the city of Athens. The determination of 13 parameters (metals and metalloids) was carried out by ICP-MS and ETAAS. The most common network architectures, namely Multi-layer Perceptrons (MLPs), Radial Basis Function (RBF) and Kohonen’s self-organizations maps (SOM) were successfully applied to the data. Initially, using discriminant analysis, only three variables (V, Ni, As) were proved to be able to describe the intrinsic characteristics of the three lakes and used for the construction of the ANNs. Optimized models (through the Root Mean Square criterion of the validation sample set), describing these characteristics, were constructed and compared. Moreover, unsupervised Kohonen’s technique achieved samples’ visualized clustering.

Targeting also in the prediction of source apportionment of unknown samples through robust models, newer samples were tested. Thus, from December 2007, water samples characterized only from the variables V, Ni and As were applied to the optimized ANNs models. The prediction accuracy concerning the classification of the “unknown” samples in the three lakes was impressive (92.8 % success for all the models). Thus, ANNs proved to be a powerful tool for water quality assessment, prediction and classification.

KEYWORDS: Artificial Neural Networks (ANNs), chemometrics, classification, pattern recognition techniques, water quality.

1. INTRODUCTION

Metals and metalloids are important parameters for water quality and safety. Due to their extensive use represent an important fraction of the pollutants released in air, soil and water. They really seem to be ubiquitous today [1].

In order to ensure the high quality of the supplied water, authorities and research bodies often conduct monitoring studies, determining a high number of parameters (explanatory variables) in a high number of sampling sites. These studies produce large data sets, which often hide important information and need further statistical treatment for their interpretation.
For this purpose, in addition to the “traditional” multivariate chemometric techniques, Artificial Neural Networks (ANNs) are often applied for prediction, clustering, classification, modeling of a property, process control, procedural optimization and/or regression of the obtained data.

In the present work, data concerning metals and metalloids concentrations from a total of 15 sampling sites in the three water reservoirs of Athens, Iliki, Mornos and Marathon were subjected to different ANNs techniques. Three anticipated goals were successfully fulfilled by the data processing with multivariate ANNs techniques:
1. to construct robust models concerning the three Athenian lakes, so that new samples could automatically be recognized,
2. to emphasize the critical variables (metals or metalloids), that lead to right predictions of sites and effectively contribute to sites’ discrimination,
3. to extract conclusions for the efficiency of the ANNs models used.

To the authors' best knowledge, the water quality, concerning the metal and metalloid content, of the three main water reservoirs of Athens has not been evaluated in any previous study. The same data have been explored with more convectional techniques like Cluster Analysis (CA), Principal Components Analysis (PCA), Discrimination Analysis (DA) and Classification Trees (CTs) [2]. Thus, the comparative application of ANNs techniques for the evaluation of surface water quality is comprehensively presented, showing that the use of different pattern recognition techniques could reveal and verify hidden intrinsic characteristics and develop simple classification rules that could be used to determine the assignation of unknown samples to the studied groups.

2. EXPERIMENTAL

Athens, a city of nearly 4 million people, is mainly watered by Mornos (Fig. 1), an artificial lake that accepts the water bodies of rivers, tributaries and streams of the surrounding region. Iliki and Marathon are used as alternative water reservoirs for the city of Athens in cases of water shortage. Fifteen (15) sampling sites were selected during the period of October 2006 to April 2007 from these lakes under the quality control monitoring program of EYDAP Company (six sampling campaigns). A detailed description of the sampling sites is given in Table 1. Water was collected in polyethylene bottles from a depth of about 20 – 30 cm. Filtration was made through 0.45 µm-pore filter (Minisart-plus, Sartorius) and the samples were acidified with *supra pur* grade concentrated nitric acid (Merck, Darmstadt, Germany) at pH<1 and kept at 4 °C until the analysis [3].

Electrothermal Atomic Absorbance Spectrometry (Perkin-Elmer, model AAnalyst 800, Bodenseewerk, Germany) was used for the determination of Fe. Inductively Coupled Plasma Mass Spectrometry (Agilent, model 7500e, Santa Clara, California, USA) was used for the determination of the rest of the elements: B, Al, V, Cr, Mn, Ni, Cu, Zn, As, Cd, Ba. All the reagents were of analytical grade. The analytical data quality was ensured through blank measurements, analysis of CRMs (certified references materials) and duplicate measurements.

3. DATA ANALYSIS AND STATISTICAL METHODS

The 89 samples were characterized by 11 variables. The results of all measurements were analyzed by three different ANNs models. The approach used, concerned the whole data set (lake-marked sampling sites, after substituting the respective numbers of the sampling sites with the respective reservoir name) of the total six campaigns. In the following sections a comprehensive summary of the theory of the used ANNs techniques is presented.
Figure 1: Map of water reservoirs of Athens

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

4. THEORY

4.1. Fundamentals

Artificial Neural Networks receive a number of inputs in the processing units which are capable to communicate by sending signals to each other over a large number of weighted connections. Their basic features should be initially emphasized [5]:

1. A set of input signals $X_1, X_2, \ldots, X_i$;
2. Connections for each unit. Each connection is defined by a weight $w_{ij}$ between the unit $i$ and $j$;
3. An output $y_j$ for each unit;
4. An external input $b$ (bias) for each unit;
5. A propagation rule, which determines the effective input $\Sigma$ from the external inputs $x_i$;
6. An activation function $f$ (usually sigmoid), which determines the correlation between the sum input $\Sigma = \sum_i x_i w_{ij} + b$ and the output $y_j$ of this unit;
7. A method (algorithm) for updating the information; Figure 2 summarizes the aforementioned features. The bias of every unit provides a term in the weight sum, so that the flexibility of the net is increased. The central idea behind the algorithm, is that the generated errors (differences between each output and the theoretical response) are propagated (distributed) along the units of the hidden layers. Thus, the initial applied weights are corrected due to the calculated error.

![Figure 2: Scheme of an artificial neuron](image)

4.2. Training a network

Working with a network means that the exact nature of the relationship is unknown. This relationship is established through the process of “training” or “learning”. There are two types of training used in ANNs: supervised and unsupervised [6]. In supervised learning, a set of training data is used. This set contains input examples combined with outputs, in order the network to “construct” the relationship between them. The outputs are a-priori known and the network calculates and adjusts the weights in such a way, that the calculated and desired outputs are as close as possible [7]. If the network is properly trained, it has then learned and the model can predict the outputs for unknown samples for given inputs. A network with more weights than appropriate models a complex function and is prone to “over-fitting” or “over-learning” [6]. This means that the network can memorize the training data and therefore be less able to generalize between other inputs and outputs. On the contrary, a network with fewer weights may not be powerful to model the data. As a golden rule, the use of fewer weights is encouraged in order to avoid over-fitting problems [8-11].

As a solution to the problem, the use of another, independent, sample set, the “validation” or “selection” set, is also suggested. This checks the network performance from an independent view: as during the training processes, the training and validation error are calculated. If the first is decreased and the second is increased, this indicates that the network starts to over-fit. If, however, both training and validation errors drop, the network is trained properly. Finally, the network is tested for its prediction accuracy by another (third) sample set, the “test” or “unknown” sample set. These samples are completely independent and do not participate in the optimization of the network parameters.

Unsupervised training (see below the Kohonen technique, section 4.5) means that the network is provided with a sample set and is left to settle down (or not) without a known desired output [7].

4.3. Back propagation algorithm

In the most problems, multi-layer perceptrons networks (MLPs) are needed, for approximating general relationships. Algorithms for these more complicated neural
networks are also necessary [12]. The most popular learning algorithm, referred as the back-propagation algorithm (BP), as the error (the calculated difference between the computed and theoretical output values) is propagated (distributed) from the output to the input layer. When a sample is introduced in the network, the activated values (through weights, bias and an activated function) are propagated to the output units. Then, the squared differences between the computed and required (theoretical) values are calculated. In order for these values to be set to zero, the method requests the adjustment of all the weights. This adjustment is occurred by the means of the calculated error, with the use of appropriate coefficients like the learning rate and the momentum.

The momentum is a coefficient used to make the new weight dependent to the previous one [5,13,14]. Thus, sudden changes in the direction that corrections are made, are aborted [14], while stuck in local minimum is avoided [15]. When for example, the learning rate is large, the oscillations can be avoided and the minimum is reached. The use of momentum may also accelerate the final result [5].

4.4. Radial Basis Function Networks

Like the MLPs network, the Radial Basis Function (RBF) network consists of three layers: input, hidden and 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 [16]. The hidden layers transform the input data using a function that is usually a Gaussian. This symmetric function has a center $\mu$ and a “spread” or “peak width” or “width factor” $\sigma$. For each input that is distant from the center (due to the Euclidean distance [17]), the Gaussian function decays the result to zero. For each sample that is closer to the center, the response is significant. Thus, a region is defined which comprises the respective field ($c \pm \sigma$) for every neuron. The key for a successful RBF network is the appropriate choice of the centres and spreads [13,18]. Finally, in the output layer, each unit makes a linear transformation to the data of the hidden layer.

4.5. Kohonen Neural Networks

Kohonen neural networks (also known as self-organizing maps, SOM), resemble most to the biological networks due to the correction implementation. These corrections do not cover the whole network, but training is strictly a “local” procedure. The weights correction affects only some the neurons of the network: the ones that adjoin with the "winner". Kohonen is considered as an unsupervised technique that attempts to recognize clusters within the training cases. It is a learning procedure that identifies some pivotal points in the space (characterized by weights) and maps the samples onto a two-dimensional layer according to the proximity to these points. Each point is characterized by a weight vector which is associated with the respective vector of the samples. The point that is closer to the sample is the winner: its weights are updated and the sample is considered to “belong” in its cluster. The new weights are calculated from the old ones through a learning rate, initially defined and updated after each iteration. When the Kohonen network has been trained, the topological map can be derived. This map performs clustering of the initial data, as a genuine unsupervised technique or classification of them, as a supervised technique by labeling (characterization) of the groups, if the outputs are provided and the direct visualization of the data [19].

5. RESULTS AND DISCUSSION

5.1. MLPs-BP network

Multi-layer perceptrons 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 [5, 13, 20, 21].

Moreover, in order to optimize the number of the inputs, the discriminant analysis (DA) results from a previous work were used [2]. Thus, variables sets of 3 (V, Ni, As), 4 (V, Ni, As, B), 6 (V, Ni, As, B, Cu, Mn), 8 (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 the standard approach of DA method. Finally, the best model was chosen, with RMS error = 0.26 for the validation set, 12 units in the hidden layer and 3 only inputs (V, Ni, As) (Figure 3(a)).

![Figure 3: Architecture of the final optimized networks (a) MLP and (b) Kohonen.](image)

The accuracy of the constructed MLP model was assessed with a series of new results received from a posterior time period (December 2007) from the same sampling sites. Table 2 summarizes the results. There was only one error in a set of fourteen 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 “confusion” of sites is thus expected. Thus, with the use of the constructed model, one could predict the water origin of the canal.

5.2. RBF network

Radial Basis Functions networks were optimized through the parameters of the number of the hidden units and the spread. The criterion used was again the RMS error for the validation sample set. The variables used were V, Ni, As, as they have already been evaluated in the previous ANNs technique, in order that comparisons were feasible. For each trial, different networks were also tested, and the best model was chosen with RMS error = 0.23 for the validation set and 9 units in the hidden layer. The accuracy of it was also confirmed with the same series of new data set of December 2007. The results were exactly the same with MLP (Table 2): the same controversial site gives the only wrong prediction.
Table 2: Predictions in new samples based on the optimized MLP model. The variables values are recorded in µg L⁻¹.

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

5.3. Kohonen Neural network

The unsupervised Kohonen technique was applied to the data set by using the already mentioned three variables. The optimization of SOM networks usually includes the number of the neurons, the dimensions of the topological map [13-15,22,23], the neighborhood radius [14], the number of iterations and the learning rate [15]. The resulting maps identify the clustering of similar groups and label all the neurons. There must be no gaps (non-recognizable units) [6,14].

Kohonen networks are considered as an unsupervised technique, but in the examined case, the final sample groups were already known (the three Athenian reservoirs) and therefore, the method could also be evaluated as a supervised one, according the classification results. Thus, in this case, Kohonen technique includes two steps [6]:

1. The unsupervised process, as it has already been described in the theoretical section of this work.
2. The labeling (characterization) of the groups according the a-priori known training set.

Many topological structures were tested and the same criterion of RMSE was used. Figure 3(b) depicts the best (with the lower RMSE) two-dimension topology (3:3-8→1). The validation error was 0.15 for this 4x2 neurons structure. It was simple and consequently less prone to over-fitting. Moreover, there were no gaps, meaning no nominated as “winners” neurons (Figure 4). Numbers in the brackets represent the win frequencies (the respective times that the neuron was nominated as the “winner”). Lakes classification in the map seems to be confused and the groups are not so discernible.
Figure 4: Topological map of the optimized Kohonen network (MO: Mornos, MA: Marathon and Y: Iliki).

Nevertheless, the accuracy of the first map was proved to be high: the same as the previous MLPs and RBF models for the new data from December 2007 (Table 2). One error was also observed in the Kohonen model 3:3→1.

5. CONCLUSIONS

In this work we have confirmed that the application of supervised ANNs techniques allowed the discrimination of surface water samples from reservoirs of Athens, according their origin, using only three parameters (V, Ni and As). Optimized BP-MLPs, RBF and Kohonen models were constructed based on the initial data set, while the same excellent results were derived for new, “unknown” samples, not used in the models configuration. Particularly, only one failure out of the 14 new samples was recorded. This deviation was absolutely justified due to the composition change of the specific sample.

Kohonen technique seemed to surpass the aforementioned ANNs techniques combining the ability of performing simultaneously as a supervised and unsupervised technique. Generally, Kohonen technique can always project data in a two-dimensional space, (while for example traditional linear PCA cannot). Moreover, Kohonen networks have excellent visualization abilities and interpret better the initial information. In addition, this technique can operate as a modeling “device” on the condition that adequate number of samples is provided for the learning procedure. Thus, compared with the classical CA that it can also classify unknown samples, Kohonen network can model and classify an unknown sample in the area of the already designed map.

Concluding, ANNs 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

2. Farmaki E. (2007) “Chemometric techniques used for the determination of the pollution sources of aquatic resources from metals”, **MSc Thesis, Department of Chemistry, National & Kapodistrian University of Athens.**
4. [http://www.generation5.org/content/1999/perceptron.asp](http://www.generation5.org/content/1999/perceptron.asp) (accessed 03/02/08).


