Iranian Journal of Science & Technology, Transaction B, Engineering, Vol. 31, No. B5, pp 501-508
Printed in The Islamic Republic of Iran, 2007

MOST PROBABLE POLLUTION SOURCE IDENTIFICATION IN RIVERS BY NEURAL NETWORKS*

K. NORUZI AND G. R. RAKHSHANDEHROO**

Dept. of Civil Engineering, Shiraz University, Shiraz, I. R. of Iran
Email: rakhshan@shirazu.ac.ir

Abstract– In this study, a certain characteristic of neural networks called Self Organizing Feature Maps (SOFM's) was applied to pollution source identification in the Kor and Sivand Rivers located in Fars Province, Iran. Wastewater quality data from significant industrial pollution sources to these rivers (mainly factories located upstream) were given. Observed sets of water quality data in sampling stations, downstream from the pollution sources, were utilized to identify the most probable pollution source that may have contributed to pollution in these rivers. With the aid of partial semantic maps generated by SOFM's, different patterns with different likelihoods were recognized in the pollution data. Certain patterns matched that of the pollution sources very closely. In other words, the fingerprints of all pollution sources (which were studied) were identified in the pollution data. Therefore, it is possible to use the maps as an aid to the management and decision support system of these rivers.

Keywords– Pollution source identification, river, neural networks, SOFM, partial semantic maps

1. INTRODUCTION

River hydrodynamics, pollution transformation, and relevant processes involved are affected by so many factors that the conventional approach to pollution source identification in rivers is almost impractical [1]. In particular, water quality in rivers is affected by multiple sources that vary constantly in extent and strength due to the weather, season, and level of human activity. Usually, such cases are too complicated to be completely defined by governing equations or other deterministic models. Even if such equations and methods could be derived, their solutions are too expensive in time and computational effort and prove to be of less importance in providing useful information. Considering all the variables and myriad conditions that affect the water quality and transport phenomena in river systems, it becomes difficult, if not impossible, to make precise and accurate predictions. The stochastic nature of rivers makes them more suitable for the application of artificial intelligence methods. As Professor Lotfi Zadeh, one of the pioneers of the fuzzy set theory has pointed out "as the complexity of the system increases, our ability to make precise and significant statements about the behavior of the system diminishes until a threshold is reached, beyond which precision and significance become mutually exclusive characteristics" [1]. To cope with this inherent complexity of such a system that results in extremely erratic and interdependent multiple input variables, Artificial Neural Networks (ANN's) are techniques that are used extensively [1].

ANN's mimic the functionalities of the human brain. Although they are very naïve in comparison to the brain's capability, in many cases they prove superior compared to most other tools. ANN's are mainly used as black box models to find the relationship between output and input data. Traditional applications of ANN's are in a variety of fields such as signal processing, adaptive filter design, system identification, business, medicine, and speech recognition and production [2].

*Received by the editors May 7, 2005; final revised form December 3, 2006.
**Corresponding author

One of the more recent applications of ANN's, widely used for data analysis and clustering, is Self Organizing Feature Maps (SOFM's). In many engineering problems one may not have any prior knowledge of the system. Furthermore, if the input variables of the system are multidimensional, data exploration with traditional techniques like Principal Component Analysis, Empirical Orthogonal Functions, and Correspondence Analysis becomes very difficult [3]. SOFM's are well qualified tools for such purposes because they organize imprecise and multivariate data into useful associative clusters. These clusters are formed without any given rule and self-train themselves to represent the given data set or input space. Unlike other ANN's, which are mostly used for function approximation, SOFM's are qualitative in nature, and in every problem they may be interpreted according to the domain knowledge of that problem. The result of a trained SOFM is a topology-preserved representation of the input space with a user controlled error that controls the overall accuracy of the network. Every cluster in the resulting feature maps may be named accordingly to represent the species of the interest. Such a map is called a partial semantic map and may be regarded as a condensed depiction of the inputs [1].

Researchers have applied SOFM's to the problem of pollution source identification in rivers in recent years. The water quality and its sufficient support for the aquatic ecology of the specific regions have been studied. The application of the SOFM model proved to be useful; the patterns of the pollution and water quality have been extracted from the models appropriately. In this study, SOFM's were applied to pollution source identification in the Kor and Sivand Rivers located in Fars Province, Iran. Resulting maps reveal the state of water quality and pollution patterns in these rivers. They may be used as an aid to the management and decision support system of the rivers.

2. THEORY

An SOFM network is composed of two layers; an input layer with an arbitrary dimension and a clustering layer which manifests the output layer (Fig. 1). Layers form a lattice that is usually one or two-dimensional, with neurons placed at the nodes of the lattice. Neurons are selectively tuned to particular input patterns or classes of input patterns in the course of the competitive learning process. Locations of the neurons are also tuned with respect to each other, so that a meaningful coordinate system for different patterns is created for the input space. Therefore, each neuron in the lattice is fully connected to all source nodes in the input layer. This network represents a feed forward structure with a single computational layer consisting of neurons arranged in rows and columns [4].

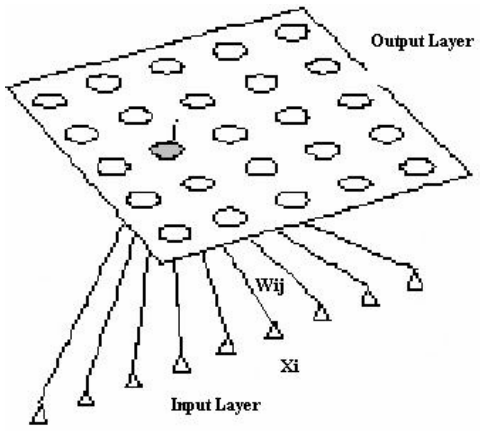

Fig. 1. Structure of an SOFM network

SOFM may be characterized by the formation of a topographic map of input patterns in which the spatial locations or coordinates of the neurons in the lattice are indicative of intrinsic statistical features contained in the input patterns. The input data in an SOFM may consist of K vectors in an N-dimensional space. Every vector corresponds to an individual dataset in the input space which has N properties of interest. Therefore, every component of the input vector reflects a property of interest in the input data. There is no limit to the number of properties or the number of input vectors. However, they are usually fixed for every specific SOFM.

The principal goal of an SOFM is to perform an adaptive transformation of the incoming signal of an arbitrary dimension into a one or two-dimensional discrete map in a topologically ordered fashion. In other words, SOFM's are used to discover clusters, or similar patterns, in data sets without supervision. They try to find the structure of the data set by iteratively exploring the data many times in a competitive learning process. The topology of the network is fully connected so that each input signal completely affects all neurons in the network. The output layer is a discrete map depicting the clusters in a topology preserving fashion [5].

In order to have a measure of similarity in an SOFM network, a "similarity function" is defined which measures the distance between the weight vector $\boldsymbol{w}_m$ and the input vector $\boldsymbol{x}_k$. Several options are available for a definition of the distance between the two vectors, which in turn, will have an important effect on the topology of the resulting map. The most common definition is the dot product of the two vectors that are being compared, which is often called the Euclidean distance. The dot product of two vectors becomes the largest when the two vectors are in the same direction or "most similar" to one another. Non-Euclidean distance measures like the Minkonvsky distance have also been used in measurement of similarity in SOFM networks [6].

3. TRAINING PROCESS OF AN SOFM

Training an SOFM network implies adjusting its weights such that the network performs the desired task optimally. The process starts with exploring the data set and looking for any pattern (cluster) there. Then the network adapts to the input patterns iteratively by adjusting the weights to reflect the topology of the input space. The training process of an SOFM is an unsupervised one since there are no targets associated with the input data.

There are two main parameters involved in the training process. One is the learning rate α, which describes the rate of change of the weights for every input. This parameter changes from its initial value to a small constant value during the course of training and remains constant for the rest of the iterations. Another parameter affecting the performance of an SOFM is the neighborhood distance; R. It encloses the neurons that can be activated within the vicinity of the wining neuron or the neuron that best matches a particular input data. Just like the learning rate, neighborhood distance shrinks to a constant value during the course of training. This value may include only the wining neuron after some iterations.

The training process of an SOFM network is initialized by setting the learning rate and neighborhood distance for the network. The initial weights are usually set to small random numbers. Then for every input vector in the input space, its distance to each neuron is computed and the winning neuron, which is the neuron with the minimum distance, is chosen. The weight of every neuron within the neighborhood distance of the wining neuron is then updated according to the learning rule:

$$\boldsymbol{w}(new) = \boldsymbol{w}(old) + \alpha[\boldsymbol{x}-\boldsymbol{w}(old)]$$

in which $\boldsymbol{w}$ is the weight vector and $\boldsymbol{x}$ is the input vector. By updating the weights for all vectors in the input space, a single iteration is completed. The learning rate and neighborhood distance are reduced and

the second iteration is initialized, and the iterations are continued until the change in the weights of the network becomes negligible. Due to the random initialization of the weights, every run of an SOFM will be different from the previous one, but the relative position of every cluster would be set as the parameters convergence [7, 8]. One recent application includes forecasting salinity in a river demonstrating the outstanding features of the SOFMs with regard to PCA and Genetic algorithms [6]. The topological ordering of the data sets by SOFMs are used extensively in hydrology such as combining the remote sensing and neural network to find the relationship between data and topological ordering of the data sets [7].

4. POLLUTION IN KOR AND SIVAND RIVERS

Kor and Sivand are two important rivers running in the Tashk and Bakhtegan lakes' basin. Kor is a permanent river, ~280 km long, running from northwest of Fars province along the Zagross Mountains to the east of the province ending up at the Bakhtegan and Tashk Lakes. It has two main reaches in the basin; Kor Olya, from its highland origins to the Khan bridge where the Sivand joins it, and Kor Sofla from the Khan bridge to the Tashk lake. Kor Sofla, with an average annual flow of ~30 m^3/s at the bridge, hosts several dykes such as Amir, Tilakan, Mowan, and FaizAbad, where water is diverted for irrigation purposes. The sampling stations used in this study were all located on the Kor Sofla (Figure 2). Sivand, a ~170 km long river, originates from north of Fars and joins the Kor at the Khan Bridge. The lakes basin has an area of 28234 km^2 spanning from 51° 44′ to 54 ° 30′ eastern longitude and from 29 ° 7′ to 31 ° 15′ northern latitude. Bakhtegan Lake, receiving an annual average flow of ~ 19 m^3/s, is one of the most important wetlands in southern Iran, being registered as a natural habitat for certain endangered species in the country [9]. Figure 2 shows a schematic of the study area in northern Fars, Iran, where the two rivers flow and pollution sources and sampling stations are located.

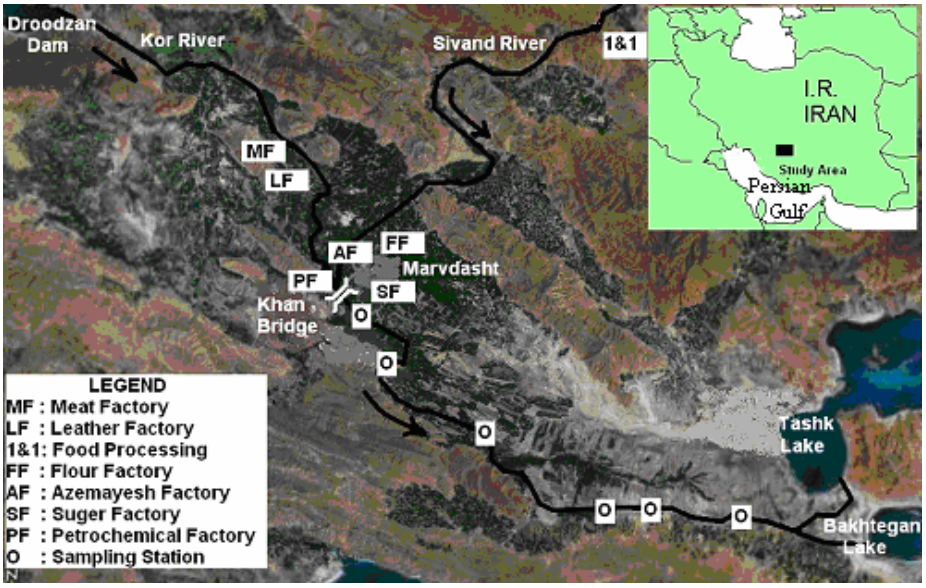

Fig. 2. A schematic of the study area in northern Fars, Iran, where pollution sources and sampling stations are located along Kor and Sivand rivers

Water quality in the Kor and Sivand rivers, being used for drinking, agricultural, domestic and industrial purposes, is essential to the development of the area. Measurements of the water quality were made on a regular basis covering a total time span of three years (from 1997 to 2000) and the data were regarded as batch. A summary of 83 sets of measured water quality data in sampling stations is given in

Table 1. Pollution sources for these rivers include discharges from many factories located upstream of the rivers such as a Meat Factory (MF), a Sugar Factory (SF), Petrochemical Facilities (PF), a Leather Factory (LF), 1 & 1 Factory, Azemayesh Factory (AF), and a Flour Factory (FF). 1 & 1 and Azemayesh are food processing and appliances manufacturing factories, respectively. Wastewater quality patterns (source fingerprints) from these pollution sources to the rivers were known. Non-metal pollutants in these sources include nitrate, phosphate, chloride, and other pollutants mentioned in Table 2. The data in this table is the average of many measurements on the discharge quality from the mentioned factories from 1990 to 1994. The major sources of heavy metals such as lead in the rivers are PF, LF, and AF. While heavy metals were not crucial pollutants in the rivers before 1995, their concentration has increased in amount ever since and they are one of the major concerns in the Kor and Sivand rivers nowadays. Investigations in this paper focused on the sources that had been considered major pollution sources by the Environmental Protection Agency of Fars Province, were studied more extensively before, and their finger prints were available [9,10]. Other pollution sources such as municipal wastewater and drainage from agricultural activities may also exist which were not considered in this study. However, lack of knowledge about some possible sources will not jeopardize identification of the sources where their finger prints are known.

Table 1. A summary of water quality data in Kor and Sivand rivers measured at sampling stations

Statistical parameters	Air Temp. C	Sample Temp. C	pH	EC mS/cm	DO mg/L	BOD mg/L	COD mg/L	Cl- mg/L	Alkalinity mg/L	NH4+ mg/L	NH3 mg/L	NO3- mg/L	PO4--- mg/L	TDS mg/L	TSS mg/L
Minimum	8.0	11.0	7.3	25.0	2.0	0.0	3.0	21.0	110.0	0.0	0.0	0.1	0.0	215.0	0.0
Maximum	36.0	22.0	8.9	9750.0	9.2	29.0	102.6	2655.0	445.0	80.0	0.7	30.0	7.7	7000.0	2022.0
Average	23.2	15.8	8.0	1158.6	6.8	5.1	24.8	227.1	185.7	5.3	0.1	10.3	0.2	809.4	176.7
Standard Dev.	6.4	3.0	0.3	1311.7	1.4	5.9	21.8	346.9	69.6	12.5	0.1	6.6	0.9	914.2	288.6

Table 2. Wastewater quality data from pollution sources to Kor and Sivand rivers [13]

Parameter	Sugar F.	Leather F.	Azemayesh F.	Flour F.	Meat F.	1 & 1 F.	Petrochemical F.
Air Temp. ^{o}C	_	_	_	_	23.5	34	_
Sample Temp. ^{o}C	26	_	16.2	21	21	30	_
pH	11.5	9.5	7.5	6.4	8.0	6.3	8.6
EC (μS/cm)	2980	0	596	2980	2001	2500	11026
DO (mg/L)	0	0	0	0	0	0	0
BOD (mg/L)	1194	1750	103	635.5	75.5	180	112.6
COD (mg/L)	2666	5300	456	921	100	284	196
Cl- (mg/L)	14.3	3000	46	586.6	390	0	820
Alkalinity (mg/L)	175	0	0	0	0	315	0
NH4+ (mg/L)	16.8	0	0.1	13.4	0	0	69.4
NH3 (mg/L)	3.9	0	0	7.62	0	0	0.1
NO3- (mg/L)	5	0	1.4	0	0.5	0	7
PO4--- (mg/L)	0.5	0	2.4	10	0	0	0.1
TDS (mg/L)	2700	45000	340	2740	14.18	275	7560
TSS (mg/L)	5300	500	40	450	902	100	2500

5. PROCEDURE

An SOFM network was constructed having a feed forward structure with a single computational layer. The network was applied to pollution source identification in the Kor and Sivand Rivers. Since there was no initial guess as to the number of clusters present in the data, different map sizes were considered. The criterion used to evaluate the maps were visual inspection, quantization error, and ability of the map to distinguish different pollution patterns. The quality of the maps was also measured by calculating the difference between all patterns in a cluster with the cluster representatives all over the map [11].

While wastewater quality data from significant industrial sources to these rivers (mainly factories located upstream) were given, observed sets of water quality data in sampling stations downstream from the pollution sources were utilized to identify the most probable pollution source that may have contributed to pollution in these rivers. A clustering procedure was performed with the aims of 1) identifying and extracting the existing pollution templates or patterns from the observed pollution records (input data) in the rivers, and 2) labeling the templates according to the best matching known source fingerprints shown in Table 2.

A U-matrix was generated to determine how distinguishable the clusters were in the cluster map. This colored map showed the goodness of the clustering procedure by the intense contrast between neighboring clusters. A higher contrast between two neighboring clusters indicates better differentiation between the two clusters. As the map size increased, the overall differentiation between clusters improved. However, the drawback was that large maps were not useful, as most of the cells remained empty. It is worth mentioning that the process of map size increase may not be reversed because unnecessary clusters, other than the known ones, will show up on the map. Finally, a density and partial semantic map was created and patterns were labeled with the known source fingerprints. Different likelihoods were also associated with each recognized pattern in the pollution data [12].

6. RESULTS

The optimum size and shape of the cluster map found for the data was a 7 by 7 hexagonal grid. The resolution of this map is high enough to differentiate clusters in the pollution data. Figure 3 shows the cluster map for the data. Each histogram on the map is a qualitative depiction of the concentration of different pollutants present alongside the river. In other words, each plot represents a unique cluster of pollution identified in the river monitoring data. As shown by histograms on the figure, 49 different patterns (clusters) were recognized for the data. Clusters located on the right and lower part of the map showed many pollutants, while other clusters showed only few pollutants.

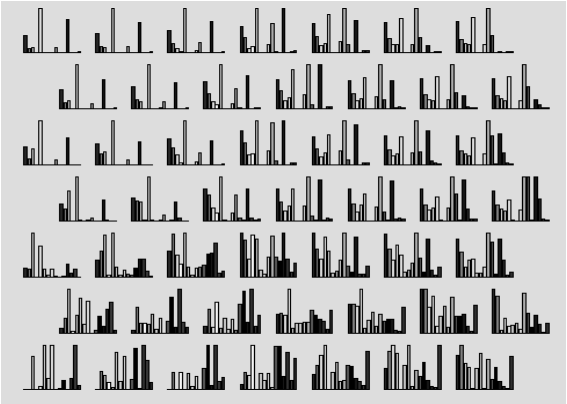

Fig. 3. Cluster map and templates for pollution data

Figure 4, the labeled U-matrix shows how well the recognized clusters in the data and the sources fingerprints are differentiable. As shown on the figure, a good color intensity contrast exists between groups of clusters, meaning that such groups are differentiable with a higher confidence. Such groups exist mainly in the pollution data and represent different clusters occurring in the data set naturally. The clusters depicting sources fingerprints were labeled accordingly. They showed a differentiable cluster in the case of the Leather Factory (LF) on the lower left corner of the matrix. However, groups of differentiable clusters in the data appeared mainly in the upper part of the matrix, while sources fingerprints were generally less differentiable and compressed in the lower part of the matrix. It was postulated that different pollutants in source fingerprints were not strictly preserved as pollutants were discharged into the rivers and transported downstream to the sampling stations.

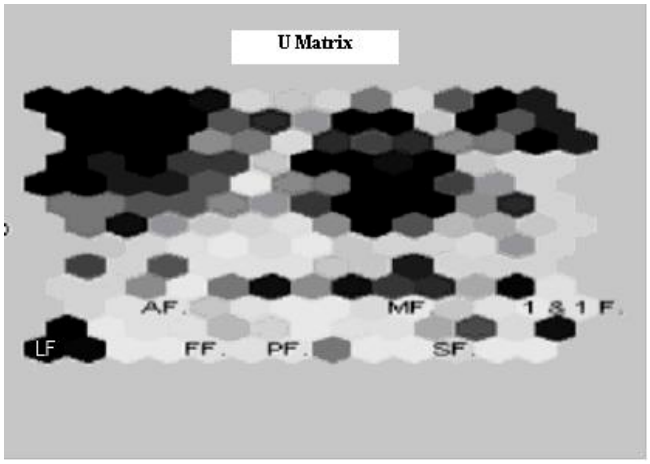

Fig. 4. Labeled U-matrix for pollution data

Figure 5 shows the density and partial semantic map for the pollution data. The relative frequency of samples being located in each cluster (its likelihood) is represented by the area of the white spot in that cluster. As shown, with the aid of the partial semantic map generated by SOFM's, many different patterns with different likelihoods were recognized in the pollution data. Certain patterns matched that of the pollution sources very closely. In other words, the finger prints of all pollution sources (which were studied) were identified in the pollution data. Therefore, it is possible to use the maps as an aid to the management and decision support system of these rivers.

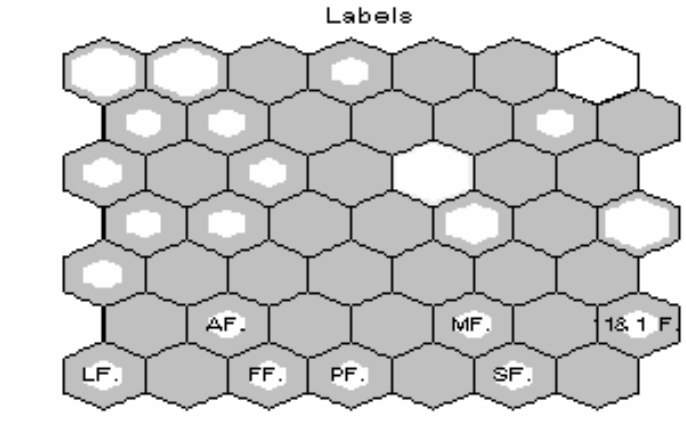

Fig. 5. Density and partial semantic maps for pollution data

7. CONCLUSIONS

SOFM's proved to be an effective inexpensive tool in data set clustering and conceptualization. With the aid of cluster maps, U-matrix, and partial semantic maps generated by SOFM's, clusters (patterns) were identified in the pollution data in the Kor and Sivand rivers. Clusters were then associated with given source fingerprints that may have contributed to them by different likelihoods. The topology preserving capability of SOFM's may be utilized to identify the most probable pollution source in rivers. The flexibility of the SOFMs makes them more desirable in developing parsimonious models of a complicated process such as pollution determination with regard to conventional methods and models discussed extensively in [14].

REFERENCES

1. Brion, G. M. & Lingireddy, S. (2000). *Artificial neural networks in hydrology*. Chapter 9. Identification of pollution sources via neural networks. Govindaraju, R.S., and A. Ramachandra Rao., Eds. Kluwer Academic Publishers, pp. 179-197.
2. Fausett, L. (1994). *Fundumantals of neural networks*. Prentice Hall International, Inc.
3. Walley, W. J., Robotham P. W. J. & O'Connor, M. A. (1999). Use of pattern recognition to identify the source of an oil spill on inland water. Environmental Agency Technical Report E72, Water Research Center.
4. Bowden, G. J., Dandy, G. C. & Maier, H. R. (2005). Maier, input determination for neural network models in water resources applications. Part 1. background and methodology. *Journal of Hydrology*, Vol. 302.
5. Bowden, G. J., Dandy, G. C. & Maier, H. R. (2005). Input determination for neural network models in water resources applications, Part2 – a case study forecasting salinity of a river. *Journal of Hydrology*, Vol. 302.
6. Islam, S. & Kothari, R. (2000). Artificial Neural Networks in the remote sensing of the hydrological process. *Journal of Hydrologic Engineering*, Vol. 5.
7. Haykin, S. (1994). *Neural Networks: A comprehensive foundation*. Macmillan/IEEE Press.
8. Sadreddini, M. H. & Sami, A. (2001). A model for physical suitability evaluation of kor and sivand subbasin. *Iranian Journal of Science and Technology*, Vol. 25, No. B3.
9. Rakhshandehroo, G. & Talebbeydokhti, N. (1998). Assessment of the Kor and Sivand river pollution. Research Report Grant, Shiraz University, Grant No. 77-LB-EN-29-0.
10. Rakhshandehroo, G. & Talebbeydokhti, N. (1999). Assessment of the Kor and Sivand river pollution. Research Report Grant, Shiraz University, Grant No. 78-LB-EN-59-0.
11. Walley, W. J. & Hawkes, H. A. (1996). A computer-based reappraisal of biological monitoring working party scores using data from the 1990 River Quality Survey of England and Wales. *Water Research*, Vol. 30, No. 9, pp. 2086-2094, (1996)
12. Ruck, B. M., Walley, W. J. & Hawkes, H. A. (1996). Biological classification of river water quality using neural networks. In *Applications of Artificial Intelligence VIII*, Vol. 2, *Applications and Techniques*, pp. 361-372. Elsevier/CMP.
13. Karimi, Y. (1994). Study of Kor and Sivand Watersheds, Soil and Water Resources Management Project, Environmental Protection Agency of Fars Province.
14. Sergios, T. & Konstanitinos, K. (1999). *Pattern recognition.* Academic Press.