*10th IMEKO International Symposium*
*June 30–July 2, 2004, Saint-Petersburg, Russia*

# WATER QUALITY DATA PROCESSING USING FUZZY NEURAL NETWORKS AND KOHONEN SELF ORGANIZING MAPS

*O. Postolache[1,2], J.M. Dias Pereira[1,2], P. Girão[2], H.G. Ramos[2]*

1. Escola Superior de Tecnologia de Setúbal, Instituto Politécnico de Setúbal, 2910-761, Setúbal, Portugal
2. Instituto de Telecomunicações, DEEC, IST, Av. Rovisco Pais, 1049-001, Lisboa, Portugal

**Abstract –** This paper is about different techniques for intelligent processing of data obtained with water quality monitoring distributed systems. The techniques include a set of fuzzy neural networks (FuNNs) for modelling measuring channels and Kohonen Self Organizing Maps (K-SOMs) for information classification. Elements of FuNN and K-SOM optimization in terms of architecture and training are presented and briefly discussed.

**Keywords**: neural network, water quality, sensor modelling, self organizing maps

## 1. INTRODUCTION

To quality monitoring and pollution detection of water occupying a wide area (e.g. rivers and lakes), a multi-point measuring system with measuring stations distributed over the area (distributed measuring system) is usually required or at least recommended. The correct interpretation of data obtained with such type of system depends not only on the accuracy of the measured values of the parameters selected for water quality assessment, but also on the way data is presented. Sensor modelling and disturbance factors correction (e.g. water flow variation, temperature variation) are processing components with positive impact on measurements accuracy of water quality (WQ) monitoring systems. Water quality sensors nonlinearity and the susceptibility to different influence factors are well dealt with intelligent algorithms such as neural networks and fuzzy logic systems [1,2,3]. On the other hand, automated assessment of water quality based on large volume of data is only possible with the implementation of adequate processing algorithms such as Kohonen Self Organizing Map (K-SOM) based algorithms [4, 5].

This paper addresses the problem of modelling water quality measurement channels using fuzzy neural network (FuNN) algorithms and data sensor mapping using Kohonen Self Organizing Maps. The intelligent designed structures are implemented on a personal computer that controls a distributed IEEE802.3 or SDI-12 based water quality measuring system.. For exemplification purposes, each field unit incorporates the hardware required to measure temperature, turbidity, pH and conductivity.

## 2. WQ DATA PROCESSING ARQUITECTURES

The pH, conductivity and turbidity measuring channels are characterized by temperature non-linear dependence requiring data processing in order to obtain accurate values of WQ measured parameters. Thus, for each field unit, a set of Fuzzy Neural Network – ANFIS type [6] were designed and implemented. Figure 1 depicts the overall processing scheme that includes also the K-SOM processing.

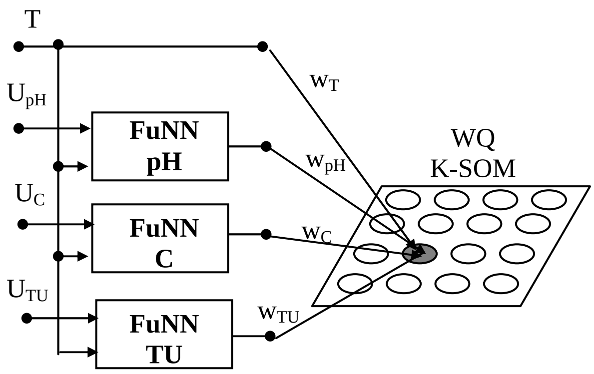


Fig. 1 WQ Field unit intelligent processing scheme

As shown in the figure, parameter processing includes a set of fuzzy neural networks ($FuNN_{pH}$, $FuNN_C$, $FuNN_{TU}$) associated with a digital readout of pH, C and TU values with temperature compensation. $U_{pH}$, $U_C$ and $U_{TU}$ represent the acquired voltages from the WQ sensor channels without temperature compensation. T represents the temperature value obtained from a voltage to temperature conversion block characterized by the following transfer function:

$$T = \sum_{i=0}^{4} a_i \cdot U_T^i \qquad (1)$$

where $a_0$=64.02°C, $a_a$=-27.46°C·V$^{-1}$, $a_2$=5.92°C·V$^{-2}$, $a_3$=-0.71°C·V$^{-3}$, $a_4$=-0.71°C·V$^{-4}$. The T($U_T$) modelling error (Fig. 2) is less than 0.01°C. The temperature sensor used (ON401) provides temperature values with an accuracy of ±0.1°C in the [5, 25] ºC temperature range.

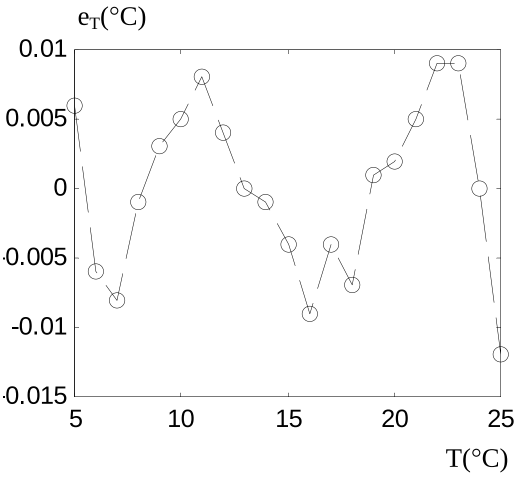


Fig. 2 Temperature measuring channel absolute modelling error.

After fuzzy neural network processing, WQ values are recorded in field unit WQ databases and are mapped into Kohonen self-organizing maps that were previously designed using historical values of the WQ parameters gathered from the same assessed area. The design of the FuNNs and K-SOMs is presented in the following paragraphs.

*2.1. WQ sensor model based on a Fuzzy Neural Network*

Neural network and fuzzy systems can approximate, with success, a large class of non-linear systems with a desired degree of accuracy [7, 8, 9]. In the area of neural network and fuzzy inference system modelling, Adaptive Neuro-Fuzzy Inference System (ANFIS) represents an important solution. The contribution of ANFIS is the idea of expressing a fuzzy inference system as a neural network architecture [10]. The six layers associated to an ANFIS processing structure corresponds to fuzzification, implication and (if needed) defuzzification (Fig. 3).

The first layer of neurons (input layer) receives the input information, the voltage or current associated to the measuring channels.

The second layer, input membership function layer, calculates the fuzzy membership degree to which the input values ($U_{pH}$, $U_C$ ou $U_{TU}$) are mapped from the input voltage intervals to the unit interval through a membership function, mf. This mf can be defined in linguistic terms. For example, the pH of water under test is not defined in a crisp sense as acid or neutral but rather as 0.5 acid and 0.5 neutral. Each node (neuron in the neural network sense) of mf layer includes a mf for one of the inputs ($U_{pH}$ and T , $U_C$ and T, or $U_{TU}$ and T). Figure 3 shows an ANFIS architecture with two inputs and four corresponding membership functions. The mf functions used in the present work are of the triangular, trapezoidal and Gaussian type. As an example, the Gaussian functions are defined by:

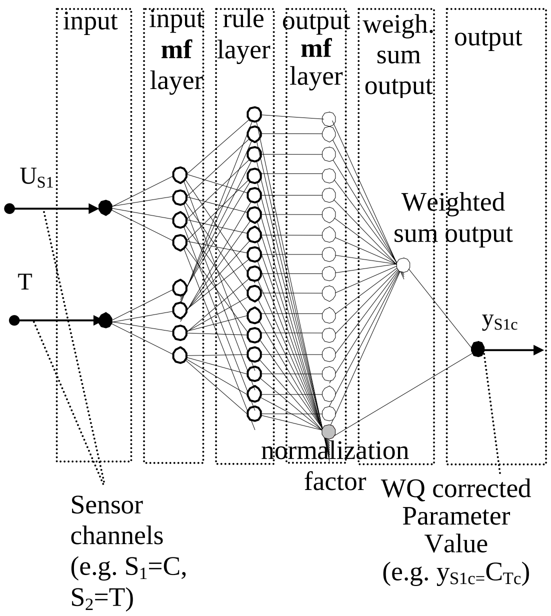


Fig. 3 A FuNN structure (ANFIS) designed to extract the temperature corrected value of conductivity ($C_{Tc}$). $I_C$ and $I_T$ are the 4-20mA currents delivered by the conductivity and temperature transducers

$$\mu_{A_{ij}}\left(x_{s_i}, c_{ij}, \sigma_{ij}\right) = \exp\left(-\left(\frac{x_j - c_{ij}}{2 \cdot \sigma_{ij}}\right)^2\right) \qquad (2)$$

where the $c_{ij}$, $\sigma_{ij}$ internal parameters are adjusted during the FuNN training phase.

The third layer is the rule layer and represents associations between the input and the output variables. The number of rules (n - number of rules layer neurons) is included in the 4 to 100 interval and their syntax has the structure

$$\begin{aligned}
&\text{If } x_{S_1} \text{ is } A_1 \text{ and } x_{S_2} \text{ is } B_1, \text{ then } f_1 = p_1 \cdot x_{S_1} + q_1 \cdot x_{S_2} + r_1 \\
&\text{If } x_{S_1} \text{ is } A_2 \text{ and } x_{S_2} \text{ is } B_2, \text{ then } f_2 = p_2 \cdot x_{S_1} + q_2 \cdot x_{S_2} + r_2 \\
&\ldots \\
&\text{If } x_{S_1} \text{ is } A_n \text{ and } x_{S_2} \text{ is } B_n, \text{ then } f_n = p_n \cdot x_{S_1} + q_n \cdot x_{S_2} + r_n
\end{aligned} \qquad (3)$$

which corresponds to a Sugeno fuzzy inference system [11]. In (3) the $A_i$ and $B_i$ are linguistic terms. For example in the case of $FuNN_{PH}$, $A_1$ represents “high acid”, $B_1$ “low temperature”, $x_{S1}$ and $x_{S2}$ are FuNN input numerical values, and $p_1 \ldots p_n$ and $q_1 \ldots q_n$, $r_1 \ldots r_n$ are consequent parameters.

The fourth layer calculates the degrees to which output membership functions, $o_{Sic}$ are matched by input data:

$$o_{S_i c} = w_i \cdot f_i \qquad (4)$$

where $w_i$ is the firing strength of rule i.

Layer five includes summation of rule outputs and firing strength, the former sum being divided by the latter on the sixth layer to yield the overall output of the system.

In the situation reported here, the FuNN output corresponds to temperature compensated values of a WQ parameter.

ANFIS training is supported on backpropagation algorithm hybrid algorithms [12] implemented using MATLAB fuzzy toolbox functions.

Different scalar parameters have been used to express FuNN model performance. The parameters we selected were the maximum modelling absolute error ($e_{max}$) and root mean square modelling error (rms).

### 2.2. *Kohonen Self Organizing Maps*

A Kohonen self-organizing map (K-SOM) is a network capable of unsupervised clustering of input data [4]. Kohonen self-organizing maps designed and implemented on a PC define mapping from the WQ parameter input space (TU, C, pH, T) into a regular two-dimensional array of nodes (WQ map). The map nodes correspond to different classes of water quality such as very-good, good and bad. The threshold values associated to pH, C, TU and T measured parameters were imposed according to EEA [13] and the method presented by Guilikeng [14]. The used thresholds in K-SOM based classification are presented in Table 1.

Table 1 Classification thresholds.

| WQ parameter | WQ Very Good | WQ Good | WQ Bad |
|---|---|---|---|
| $pH_{min}$ -$pH_{max}$ | 6.5-8 | 6-6.5, 8-8.5 | 5-6, 8.5-9 |
| C | 2000uS/cm | 5000uS/cm | 10000uS/cm |
| TU | 100NTU | 500NTU | 1000NTU |
| T | 15°C | 20°C | 25°C |

The K-SOM training used 624 sets of pH, C, TU and T values previously stored on the memory of a personal computer working as an advanced processing unit. The training was performed using the incremental-learning SOM algorithm where the prototype vector, $m_i$ of unit i of the K-SOM network, randomly initialised, was updated according to the following learning rule [15]:

$$m_i(t+1) = m_i(t) + h_{ci}(t)\cdot\left(x(t) - m_i(t)\right) \quad (5)$$

$$h_{ci} = \alpha(t)\cdot\exp\left(-\frac{\|r_i - r_c\|^2}{2\cdot\sigma^2(t)}\right) \quad (6)$$

where $h_{ci}$ is called the *neighbourhood function*, $0<\alpha(t)<1$ is the learning factor, $r_i$, $r_c \in R^2$ are vectorial locations in the display grid and $\sigma(t)$ corresponds to the width of the neighbourhood function.

The quality of WQ K-SOMs design was evaluated using the following parameters:

- $q_e$ - data quantization error, which expresses the representation accuracy.
- $t_e$ - topographic error, which expresses data set representation accuracy [16].

## 3. TEST AND RESULTS

### 3.1. *FuNN models*

Using the normalized values of $U_{pH}$, $U_C$, $U_{TU}$ and $U_T$, different two-input one-output FuNN modelling structures were implemented. A practical study concerning the relation between the number of membership functions of FuNN mf layer and the FuNN model accuracy was carried out. Table 2 shows results of the accuracy of $FuNN_{pH}$ modelling versus FuNN's architecture and training type (back-propagation and hybrid training algorithms).

Table 2 Accuracy of FuNNpH modelling versus FuNN's architecture and training type – FuNN testing phase.

| $N_{mf1}$, $N_{mf2}$ | mf1,mf2 type | Backpropagation | Hybrid |
|---|---|---|---|
| | | $e_{max}$ | |
| 2 | trimf | 0.7962 | 0.0042 |
| | trapmf. | 0.7959 | 0.0043 |
| | gaussmf | 0.7861 | 0.0041 |
| 3 | trimf | 0.7895 | 0.0047 |
| | trapmf | 0.7894 | 0.0044 |
| | gaussmf | 0.7834 | 0.0042 |
| 10 | trimf | 0.7964 | 0.0090 |
| | tramf | 0.7964 | 0.0011 |
| | gaussmf | 0.7992 | 0.0010 |

As shown in Table 2, the hybrid-training algorithm conducts to better results than the ones obtained with back-propagation algorithm. At the same time, increasing the number of mf does not lead to an increase in model's accuracy. From Table 2 it can be concluded that the best results are obtained when the mf layer functions are of the Gaussian type and the number of mf functions for each input is up to three.

In order to avoid WQ characteristic overfitting problems, different simulation were carried out and the correspondent RMS (root mean square) errors evaluated based on FuNN training, FuNN testing and FuNN output data. The obtained results indicate the $N_{epoch}$=30 as the optimal value associated to FuNN training (Table 3).

Table 3 The RMS training and test errors versus number of epochs.

| Number of epochs | RMS training | RMS test |
|---|---|---|
| 10 | 0.0039 | 0.0046 |
| 20 | 0.0042 | 0.0046 |
| 30 | 0.0045 | 0.0045 |
| 40 | 0.0041 | 0.0047 |
| 50 | 0.0042 | 0.0056 |

For $N_{epoch}$=30, different FuNN architectures characterized by different numbers of mf, for each of WQ parameter channel, were tested. The results represented in Fig. 4 show that modelling errors ($e_{max}$) calculated using the FuNN training set decrease when the number of mf increases (a-trimf and a-gaussmf curves). Calculating the modelling error using the

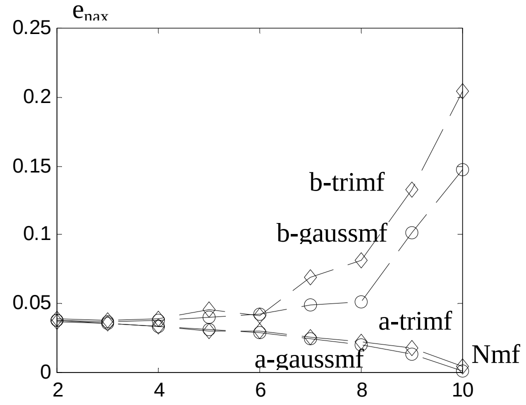


Fig. 4 The evolution of $e_{max}$ associated to $FuNN_{pH}$ model for mf triangular and Gaussian functions when the number of mf varies in 2 to 10 interval.

FuNN testing set, the obtained values increase when the number of mf increases (b-trimf and b-gaussmf curves). This effect, known as over fitting (Fig. 5), underlines the generalization losses of FuNN with excessive number of mf (neurons in neural network language). The over fitting ratio $R_{over}$ is defined as:

$$R_{over} = \frac{e_{max}\big|_{test}}{e_{max}\big|_{training}} \tag{7}$$

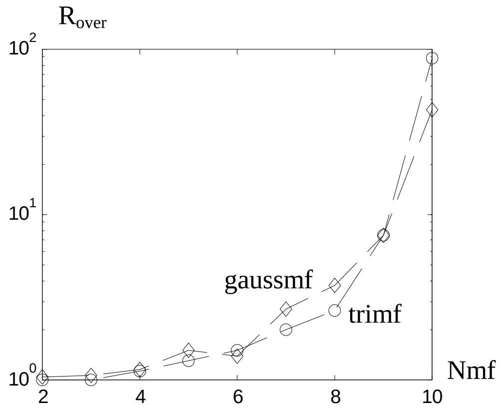


Fig. 5 The evolution of over fitting ratio ($R_{over}$) versus the number of FuNN membership functions ($N_{mf}$).

In Fig. 5 it can be observed that increasing the number of mf, $e_{max}|_{training}$ decreases but the ratio between the maximum errors obtained with training and testing sets increases.

At the same time, increasing the number of mf (number of nodes on the mf layer) the processing complexity will increase (Fig. 6).

*3.2. K-SOMs*

For each field unit, the mapping of pH, C and TU temperature compensated historical values was made using 6×10 K-SOM units (Fig. 7). Three WQ zones, called K-SOM clusters, can be identified in the figure: (Vgood, Good and Bad). Based on the designed map, the global evolution of WQ can be carried out. For example, for the particular case of

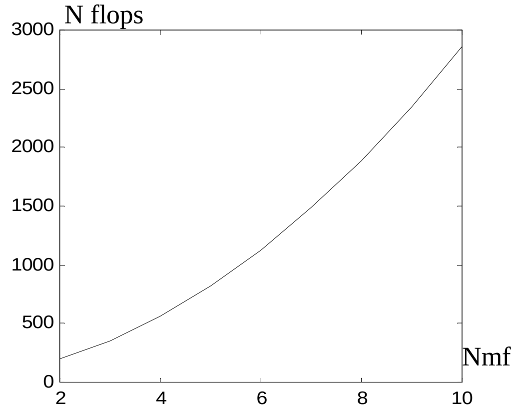


Fig. 6 Number of flops for FuNN with different number of Gaussian mf.

WQ data, pH=6.8, C=500μS/cm, T=15.5°C, and TU=202NTU, the best matching unit cell (BMU) corresponds to Vgood cluster ( $K\text{-}SOMC_{42}$).

To express the pH, C, TU and T data distribution on the K-SOM the multiple hit histogram representation [10] was used. Thus highly concentration of historical data corresponds to $K\text{-}SOMC_{31}$, $K\text{-}SOMC_{62}$ of the "Vgood" cluster, $K\text{-}SOMC_{13}$ and $K\text{-}SOMC_{35}$ of "Good" cluster and $K\text{-}SOMC_{19}$ and $K\text{-}SOMC_{610}$ of the "Bad" cluster.

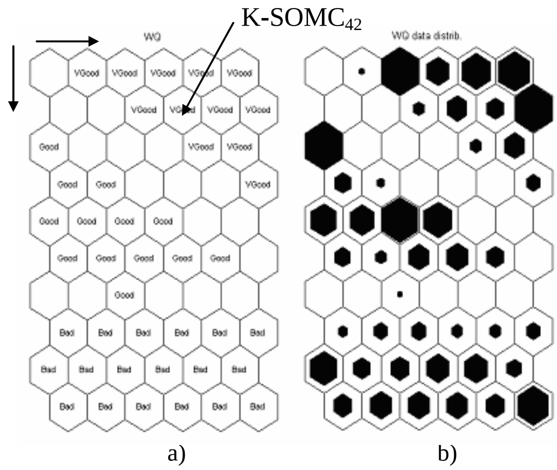


Fig. 7 The K-SOM associated to water quality classification in Tagus Estuary (WQ class1 – very good, WQ class2 – good, WQ class3 –bad) and the corresponding multiple hit histogram of WQ values.

Referring to WQ K-SOM design quality [15][16], the quantization error, $q_e$, is lower than 0.3 and the topology error, $t_e$, is lower than 0.1.

## 4. CONCLUSION

The paper presents two intelligent algorithms applied in water quality data processing and analysis. Based on fuzzy neural network techniques, more accurate models of WQ measurement channels can be carried out leading to accurate measurement of water quality parameters. At the same time, multidimensional data representation of water quality measurement channels, using Kohonen self organizing maps (K-SOM) to express the WQ class, permits a quick identification of pollution events. The designed and implemented intelligent structures provide temperature corrected WQ parameter values readout within ±2% error (full scale) over a 5°C to 25°C temperature range.

## ACKNOWLEDGEMENTS

This work was supported both by Portuguese Science and Technology Foundation POSI SFRH/BPD/11549/2002 and by Project FCT PNAT/1999/EEI/15052. These supports are gratefully acknowledged.

## REFERENCES

[1] M. Heniche, M. Attari, F. Boudjena, "Reducing the effects of temperature and $K^+$ activity in Sodium Measurement by Mean of Fuzzy Logic System", Proceedings IMEKO TC-4, vol. 1, pp. 602-606, Naples, Italy, 1998.

[2] J. Patra, A. Bos, A. Kot, "An ANN-based smart capacitive sensor in dynamic environment", Elsevier, Sensors and Actuators 86 (2000) 26-38, 2000.

[3] O. Postolache, P. Girão, M. Pereira, H. Ramos, "Self Organizing Maps Application in a Remote Water Quality Measurement System" Proc. IEEE IMTC 2003, vol.1, pp. 529-533, Vail, Colorado, USA.

[4] T.,Kohonen, *Self Organizing Maps*, 2nd Ed. Springer Verlag, 1997.

[5] O. Postolache, P. Girão, J.Pereira, H. Ramos, "Wireless Water Quality Monitoring System Based on Field Point Technology and Kohonen Maps" Proc. IEEE Canadian Conf. On Electrical &Computer, vol.3, pp. 1873-1877, Montreal, Canada, May, 2003.

[6] Jang J. S. R., ANFIS: adaptive-network-based-fuzzy-inference-system, IEEE Trans. Syst. Man Cybernet. SMC23, 665-85, 1993.

[7] O. Postolache, M. Pereira, P. Girão, M. Cretu and C. Fosalau, "Application of Neural Structures in Water Quality Measurements", in *Proc. IMEKO World Congress*, 2000, vol. IX, pp.353-358.

[8] C. Temneanu, O. Postolache, P. Girão, J. Pereira, "Fuzzy Modeling of Measurement Data in Water Quality Assessment", Proc. 12th IMEKO TC4 International Symposium, Part.2, pp. 396-400, Zagreb Croatia, September, 2002.

[9] B. Kosko, "Fuzzy systems as universal approximators". IEEE Trans. Computers, 43:1329-1333, 1994.

[10] A. Horikawa, T. Furuhashi, Y. Uchikava, "On Fuzzy Modeling Using Fuzzy Neural Networks with Back-Propagation Algorithm", IEEE Trans. on Neural Net., Vol.3, No.5, pp. 801-806.

[11] R. Babuska, *Fuzzy modelling for control*, Kluwer Academic Publishers, 1998.

[12] J.S. Jang, C. Sun, E. Mizutani, *Neurofuzzy and Soft Computing*, Prentice-Hall 1997.

[13] European Environment Agency, "Water temperature level of pollution in surface waters for quality class I". [Online]. Available: www.star.eu.int

[14] A.M. Guilikeng – Real–time Monitoring and Assessment of River Basin Conditions for Drinking Water Intake protection, Water Quality and Resource Management, REC-Telematics, 1999.

[15] Esa Alhoniemi, Johan Himberg, Juha Parhankangas and Juha Vesanto, "MATLAB-SOM toolbox V2", 2002.

[16] J. S. Kirk and J. M. Zurada. "A two-stage algorithm for improved topography preservation in self-organizing maps" Proc. of IEEE International Conference on Systems, Man, and Cybernetics, IEEE Service Center, 2000, Vol. 4, pp. 2527-2532.

---

**Authors:** Prof. Octavian Postolache
Instituto de Telecomunicações – Pólo de Lisboa
Escola Superior de Tecnologia do Instituto Politécnico de Setúbal
Departamento de Sistemas e Informática
Rua do Vale de Chaves, Estefanilha
2910-761 Setúbal PORTUGAL
Phone: +351.265.790000 Fax: +351.265.721869
Email: poctav@alfa.ist.utl.pt