Application of computational intelligence techniques for monitoring and prediction of biological wastewater treatment systems

A. M. A. Dias, M. M. Alves, E. C. Ferreira

IBB-Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, Universidade do Minho, Campus de Gualtar, 4710-057, Braga, Portugal
(Email: {*adias, madalena.alves, ecferreira*}@*deb.uminho.pt*)

Abstract
Computational intelligence models are being increasingly employed for the supervision and control of biological wastewater treatment systems. These models can be described as mathematical methodologies which explain relations between cause (input data) and effects (output data) irrespective to the process and without the need for making assumptions considering the nature of the relations.
In this work both Artificial Neural Network and Neural Fuzzy models were used for monitoring and prediction of biological wastewater treatment systems. The proposed approaches were tested for their ability to detect external and internal disturbances in data obtained from the IWA/COST Benchmark Simulation Model.
The models were also applied to predict, with one hour is advance, the response of the system to a sequence of two large increases in the influent flow rate. Both models learned well from the training data and exhibited good and fast predictions of the performance of the system submitted to the tested shocks. The results obtained indicate that the Neural Fuzzy model is slightly superior to the Neural Network model being however the correlation coefficients obtained for both models superior to 0.96.

Keywords: adaptive neural fuzzy model, artificial neural network, COST Benchmark, monitoring, prediction

INTRODUCTION

Conventional control methods are adequate when analytical mathematical models are available to support their development and operation however, this situation is uncommon in real processes. Biological wastewater treatment processes for instance have several features that set them apart from other industrial processes. In fact, in biological systems some changes are not so obvious and may gradually grow until they become a serious operational problem. It is difficult to take into account the numerous factors that can influence the bacterial growth rate and metabolic activity making an analytical description of the system's behaviour quite difficult. Moreover, frequently the changes occur slowly being the recovery from failures usually time-consuming and expensive. Biological treatment plants also exhibit strong non-stationary characteristics, with the flow rate and the composition of the feed waste stream presenting strong fluctuations. In these kind of processes an efficient process monitoring scheme can help plant operators to understand the status of the process and thus to take appropriate action when an abnormality is detected before the situation becomes dangerous.

Many of the traditional models used to describe biological wastewater treatment systems are often extremely complex, analytically insolvable and not always useful for control applications (Tay and Zhang, 2000). The idea of looking to the system as a black box and infer about the system's performance using only input and output data seems to be an attractive approach when working with these systems. Intelligent control is a very promising technique. The techniques of Computational Intelligence such as fuzzy logic, neural networks and neuro-fuzzy are powerful tools to overcome the description of the complex biological nonlinear system that occurs in a wastewater treatment plant (WWTP).

In this study, a neural network and a neural fuzzy system, based on the adaptive network-based fuzzy inference system (ANFIS) were employed in an attempt to conclude about the prediction and monitoring capacities of each model when applied to the description of a wastewater treatment plant.

METHODOLOGY

Neural networks (NN) are information processing systems that demonstrate the ability to learn, recall, and generalize from training data. NNs usually have a larger number of highly connected processing elements that can be connected in various topologies (Rumelhart et al, 1986). The way the neurons are connected determines the shape of the network. Generally, a NN has three functional layers namely the input, the hidden layer, and the output. The information enters the NN at the input layer, and then all layers process these signals through the NN until they reach the output layer. Each input is weighted and the sum of the weighted inputs and the bias forms the input to the transfer function. The connection weights of a NN can change in response to the inputs and states of the network. This provides the NNs with the abilities of learning and adaptation. The application of a NN model to modelling purposes involves a sequence of two main phases which are the learning and recall processes. Learning is the process of adapting or modifying the connection weights in response to training data. Recall refers to how the NN performs when new data are fed to the trained net. In this work, the NN was applied as follows. First, the normalized data is subjected to an analysis of principal components (PCA) in order to eliminate redundant information in the input data. The number of principal components retained was fixed by eliminating those principal components that contribute less than 2% to the total variation in the data set. The structure of the NN was them selected by trial and error since there are no simple ways to determine in advance the minimal number of hidden nodes necessary to achieve the desired performance level (Wasserman, 1993). "In this case, a three layer network was selected. An input vector, with the number of PC resulting from the PCA pre-treatment, connects data to a hidden layer with ten neurons followed by the output layer with one neuron. The transfer function (TF) used in the first layer is *tan-sigmoid* and the output layer TF is linear being these the most usual TFs when using feedforward networks. The network was trained by means of a feed-forward resilient backpropagation algorithm."

.

The neural-fuzzy system is a fuzzy system that uses a learning algorithm, inspired by the NN theory, to determine its parameters, namely its fuzzy memberships and fuzzy rules. The neuro-fuzzy model adopted in this study is the Adaptive Neural Fuzzy Inference System or ANFIS, proposed by Jang (1993) which is based on the first-order Sugeno fuzzy model (Takagi and Sugeno, 1985). In a very simplistic analysis, it can be said that the ANFIS is a first-order Sugeno fuzzy system, characterized by a set of structured fuzzy rules of the type `IF` (observation) `THEN` (actuation) that is 'adapted' in order to be able to incorporate the learning capacities of the neural networks as described above. There are four key components to be considered when modelling with ANFIS which are: a reliable database (constituted by a set of input variables and corresponding outputs), a fuzzy system generator, a fuzzy inference system (in this work a Sugeno FIS) and its associated adaptive NN representing the fuzzy system (in this work the ANFIS). A fuzzy system generator was used to create a prototype first-order Sugeno fuzzy system based on subtractive clustering of the data set provided (Lin and Lee 1996). The system generator used in this work is able to extract a set of rules that models the data behaviour (without any further explicit knowledge of the system) and simultaneously cluster the data (using a cluster's centre range of influence was equal to 0.5) in order to eliminate correlated and/or conflicting data information and speed up the computation time. Figure 1 represents the fuzzy set scheme originated when the effluent's total nitrogen concentration, is used as monitoring variable. Once the prototype is available, the ANFIS algorithm will tune the fuzzy system by learning from the training data, and finally produce a Sugeno fuzzy system with the same structure as the prototype (Demuth and Beale, 1994). The MATLAB 7.1 (The MathWorks) computing language was used for algorithm development, data visualization, and data analysis.

RESULTS AND DISCUSSION

The IWA/COST Benchmark Simulator – BSM1 (Copp, 2001) was used to define the plant layout, the simulation model and the influent loads. This simulator uses the Activated Sludge Model No. 1 (Henze et al., 1987) to simulate the biological reactions and a double-exponential settling velocity model (Spanjers et al., 1998) for the settling tank. The plant layout consists of a five-compartment bioreactor (two anoxic reactors followed by three aerated ones) and a secondary settler designated to treat an average flow of 20000 m^3/day with an average biodegradable chemical oxygen demand concentration of 300 mg/L. Influent data and operation parameters developed by IWA task group on benchmarking of control strategies for WWTPs (www.benchmarkwwtp.org) are used in the simulations.

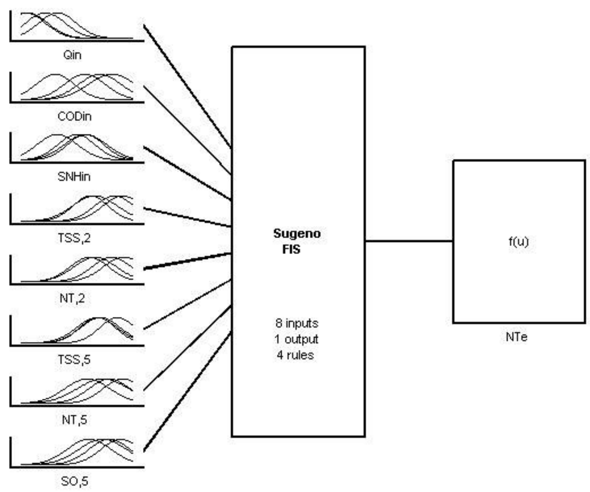

Figure 1. Diagram of the Sugeno FIS generated when the effluent's total nitrogen concentration is used as monitoring variable. Inputs and their membership functions appear on the left of the FIS.

The selection of the input variables to be used for model's training is of extreme importance. They have to be both easy to measure and representative of the system and of the process changes. Table 1 presents three different sets of input parameters that were tested in order to select the set that induces the best model predictions. According to the results given in Table 2, the set of input variables named as Case 3 is the one presenting the lower calibration and prediction errors and the higher correlation coefficients between COST Benchmark data and data given by the ANFIS and NN models. So, Case 3 variables will be used to train and test the models. All the selected input variables are relatively easy to measure in a WWTP and they try to represent three key sampling points on the Benchmark WWTP, namely the influent conditions, the conditions at the end of the anoxic phase and the conditions at the end of the aerobic phase. The output parameters should be variables that are important to control and to access, as fast as possible, giving an idea of the performance conditions of the WWTP. In this work the output parameters are the effluent's total suspended solids, TSS_e, chemical oxygen demand, COD_e, biological oxygen demand, BOD_e, and total nitrogen, NT_e. The models will be able to correctly predict these parameters and also to use them as monitoring parameters to detect perturbations on the system.

The normalized regular dry weather data set corresponding to 14 days of operation was alternately divided into a training set (60% of the input data) and a testing set (40% of the input data). The sampling period is 15 min hence each 24 hr period consists of 96 samples and the total number of samples used for this study is equal to 1344 samples. Data were disturbed with Gaussian noise with zero mean and standard deviation of ±5 %. Having the models calibrated for the regular period data, three different new input data sets are used to test the ability of the models to detect changes in the operation conditions. New data correspond to two external perturbations namely, a rain weather event (characterized by a perturbed interval between sample 800 and 1000) and a storm event (characterized by two individual perturbed instants at samples 850 and 1100) and one internal perturbation corresponding to a step decreasing autotrophic growth rate (characterized by the decrease of the specific growth rate of the autotrophs from 0.5 to 0.3 day^{-1} after the fourth day and until the end of the simulation). The new input variables, corresponding to these new conditions, are first normalized and are then fed to the trained models. The results are then compared with the pre-treated outputs given by the IWA/COST simulator.

Table 1. Different sets of inputs parameters studied.

	Input parameters	
Case 1	**Case 2**	**Case 3**
Q_{in} - influent's flow rate	Case 1 plus:	Q_{in} - Influent's flow rate
TSS_{in} - influent's total suspended solids	SNO_2 - nitrate and nitrite in reactor 2	COD_{in} - influent's chemical oxygen demand
COD_{in} - influent's chemical oxygen demand	Kla_5 - oxygen transfer coefficient	SNH_{in} - influent's $NH_4^+ + NH_3$ nitrogen
BOD_{in} - influent's biochemical oxygen demand		TSS_2 - total suspended solids in reactor 2
NT_{in} - influent's total nitrogen		NT_2 - total nitrogen in reactor 2
		SO_5 - dissolved oxygen in reactor 5
		TSS_5 - total suspended solids in reactor 5
		NT_5 - total nitrogen in reactor 5

It is important to notice that a different model (both NN and ANFIS) is generated for each of the outputs to be predicted (four in this case). A stopping criterion of 250 epochs was chosen to compare the results given by both models. It was decided that the improvement given by using a higher epoch's number does not compensate the slower computational performance resulting from this increment in the stopping criteria (Table 2). The performance of the models were assessed by evaluating the scatter between the observed and correlated/predicted results via correlation coefficients *R* and root mean-squared errors of calibration (RMSEC) and prediction (RMSEP) defined as:

$$RMSE = \sqrt{\frac{(y_1 - y_2)^2}{n}} \qquad (1)$$

where y_1 and y_2 refers to the observed and calibration/prediction data respectively and n refers to the number of samples. Table 2 presents the calibration/prediction errors and the correlation coefficients obtained for each output variable when the dry weather file is used to train and test both models. It seems that both models give good prediction and monitoring results with the ANFIS model to some extent more accurate than the NN model.

As mentioned above, the models were also used as monitoring tools for detection of external and internal perturbations in the process due to changes in the weather conditions and in the autotrophic growth rate respectively.

Microorganisms are sensible to different external and/or internal perturbations that can affect their development and consequently the process for which they are responsible. Figure 2 illustrates how some of the process variables are affected by the internal perturbation simulated. Clearly, the most affected variables are the ones related with nitrogen's concentration in its different forms ($NH4^+$, NH_3, NO^{3-} and NO^{2-}). The effluent's concentrations of ammonium and amoniacal nitrogen increase while the effluent's concentrations of nitrates and nitrites decrease as the nitrification process is being affected by the decrease in the autotrophic growth rate. As mentioned in the introduction, the capacity to identify these changes as soon as possible can be the only mean to avoid the complete failure of the process.

Table 2. Calibration and prediction results obtained with the ANFIS and the Neural Network models.

		ANFIS				Neural Network			
		Training		Testing		Training		Testing	
Case	**Output**	**RMSEC**	**R^2**	**RMSEP**	**R^2**	**RMSEC**	**R^2**	**RMSEP**	**R^2**
1	TSSe	0.151	0.988	0.181	0.987	0.143	0.989	0.148	0.988
	CODe	0.157	0.988	0.166	0.986	0.159	0.987	0.166	0.986
	BODe	0.170	0.985	0.181	0.983	0.170	0.985	0.179	0.983
	NTe	0.357	0.934	0.377	0.926	0.439	0.898	0.448	0.893
	Average (250 epochs)	**0.209**	**0.974**	**0.227**	**0.971**	**0.228**	**0.965**	**0.236**	**0.963**
2	TSSe	0.098	0.995	0.104	0.994	0.149	0.998	0.155	0.988
	CODe	0.105	0.994	0.111	0.993	0.160	0.987	0.164	0.986
	BODe	0.111	0.993	0.124	0.992	0.184	0.982	0.190	0.982
	NTe	0.278	0.960	0.313	0.949	0.758	0.651	0.760	0.649
	Average (250 epochs)	**0.148**	**0.986**	**0.163**	**0.983**	**0.313**	**0.905**	**0.318**	**0.901**
3	TSSe	0.096	0.995	0.102	0.995	0.158	0.988	0.166	0.986
	CODe	0.097	0.995	0.104	0.995	0.162	0.985	0.185	0.983
	BODe	0.116	0.993	0.129	0.992	0.182	0.983	0.201	0.980
	NTe	0.171	0.985	0.203	0.979	0.424	0.906	0.446	0.895
	Average (250 epochs)	**0.120**	**0.992**	**0.135**	**0.990**	**0.232**	**0.965**	**0.250**	**0.961**
	Average (1000 epochs)	**0.108**	**0.994**	**0.130**	**0.990**	**0.214**	**0.972**	**0.237**	**0.990**

Figure 3 presents the detection capacities of both models for the perturbations being tested. The results reveal that for the external perturbations simulated, all the output variables can be used for monitoring purposes, at least for detection of the changes in the weather conditions studied in this work. In fact, all the variables permit to identify and distinguish between the rain and the storm event. However, from the figure it can be observed that the use of TSSe and BODe as monitoring variables will lead to more accurate conclusions since, in both cases, the start and the end of the perturbation events are better defined. With respect to the internal perturbation simulated, it can be observed that the moment were the perturbation occur for the first time (sample 385 corresponding to day 4) is detected by all the variables and by both models. Although it is possible to differentiate the period before and after the perturbation with all the variables it is clear that NTe is the most sensible variable to detect the simulated decrease in the autotrophic growth rate. This is a direct consequence of what was mentioned above.

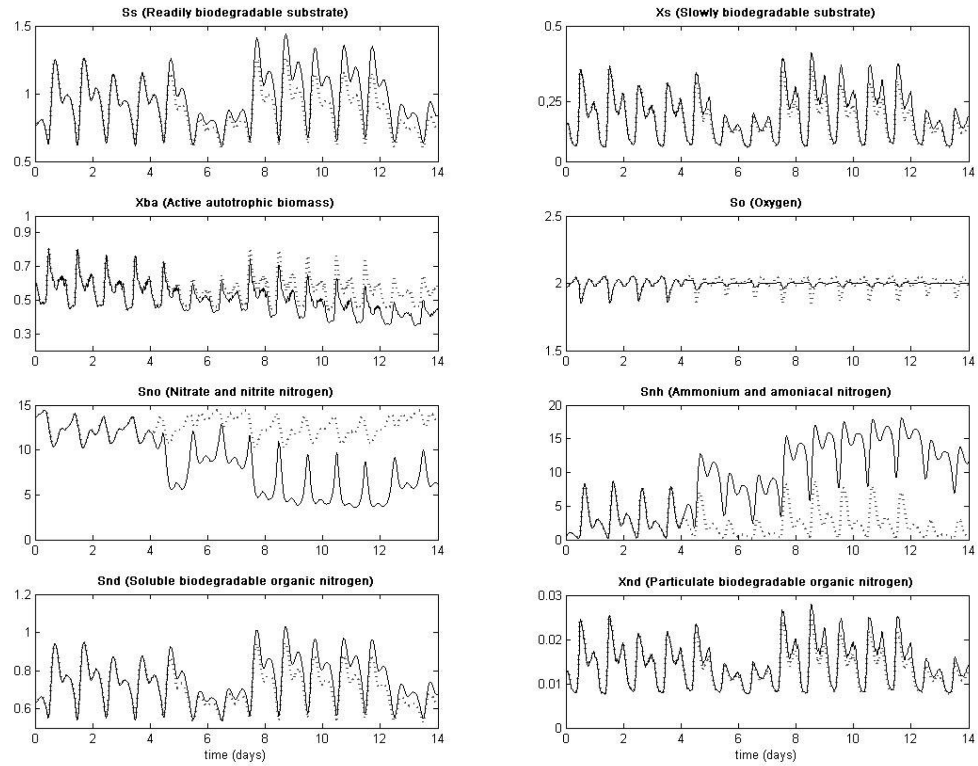

Figure 2. Simulated effluent data for regular dry weather (black line) and for system perturbated with a decrease in the autotrophic growth rate (grey dashed lines).

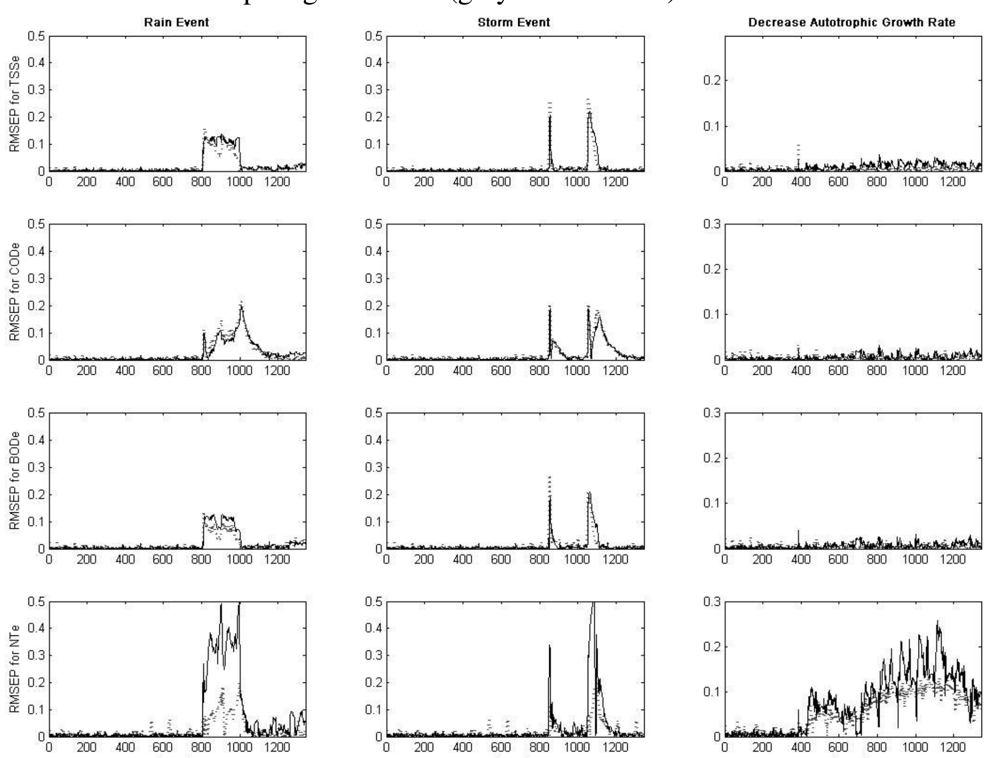

Figure 3. Detection capacities of ANFIS (full black line) and Neural Network (dashed grey line) models for different simulated perturbations using different output variables: TSS_e, COD_e, BOD_e and NT_e.

It is important to remark that the fault is detected with minimum delay which is a very important feature for online control purposes. Although globally both models gave similar results, a detailed observation of the figure permits to conclude that the ANFIS model is more sensible than the NN model.

The above models were also used to predict the response of the system to a large increase in the influent flow rate but this time with an antecipation of one hour. The same input and output variables were used however, in this case, the input parameters correspond to time n while the output parameters correspond to time $n+1$. The dry and rain weather files were used as training data and it was expected that the models would be able to predict the changes in the output variables originated by the induced disturbance with an hour in advance. The results obtained with both models are presented in Figure 3. For each of the output parameters the left figures represent the training data while the right figures represent the predicted data. The results are again very encouraging. Both models permit to detect the increase in TSS, COD and BOD and the decrease in the NT effluent's concentration that characterize the increase in the influent flow rate. Although, here again the most accurate results are obtained when TSSe and BODe are used as monitoring variables. The more accurate calibration results observed for these variables consequently originate more accurate predictions. In these cases, both models are able to detect correctly and in due time the increase in TSS and BOD concentrations resulting from the storm event.

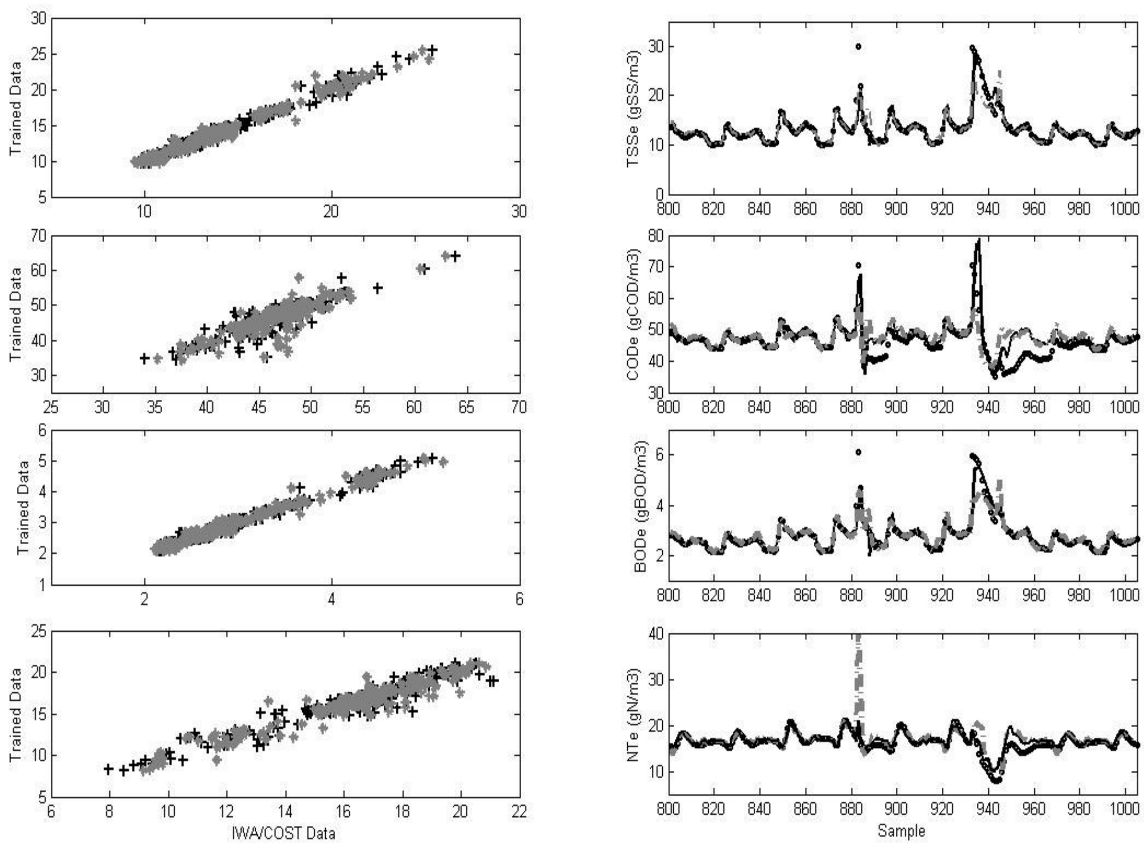

Figure 3. Prediction performance of ANFIS and Neural Network models to detect the response of the system to a large increase in the influent flow rate with one hour in advance. The figure represents the training results, on the left, and the predicted results on the right. Black symbols and lines correspond to the ANFIS's results while grey symbols and lines correspond to NN's results. Circles represent the benchmark calculated data.

CONCLUSIONS

Because biological wastewater treatment systems are too complex to model using conventional methods, advanced artificial intelligence technologies were employed in an attempt to develop adaptive and functional models for monitoring these systems. A neural network and a conceptual neural fuzzy model were used as prediction and monitoring tools of four important output variables for which maximum legal limits have to be guaranteed. Both models gave rise to very good results when predicting dry weather data and detecting external and internal perturbations although it can be said that the results given by the ANFIS model are slightly superior to the NN model. From the results obtained in this work it can be concluded that TSS and BOD effluent's concentrations were more accurate to detect the weather disturbances simulated while the NT effluent's concentration was very sensible to detect the decrease in the autotrophic growth rate. The models were also applied to predict the response of the system, with one hour in advance, to a large increase in the influent flow rate. Here again the ANFIS model originated more accurate training and predictive results. Though being preliminary results they are very encouraging and future work in this area seems valuable since the existence of models capable of predicting in advance is of extreme importance for WW control and treatment purposes.

ACKNOWLEDGEMENTS

The authors grateful acknowledge the financial support from the *Fundação para a Ciência e a Tecnologia* (Portugal) to A.M.A. Dias through a post-doctoral grant and to the project POCTI/AMB/60141/2001.

REFERENCES

Copp J (Ed) (2001) The COST Simulation Benchmark: Description and Simulation In *Manual. Office for Official Publications of the European Community: Luxemburg*

Demuth H., Beale M. In *Neural network toolbox for use with MATLAB*. Mathworks Inc; 1994, p.158

Henze M, Leslie CP, GuierW, Marais GVR, Matsuo T. (1987) A general model for single-sludge wastewater treatment systems. *Water Res*, 21,505–15.

Jang J. S. R. (1993) ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Trans. Systems, Man and Cybernetics*, 23(3), 665-685

Lin C. T., Lee C. S. G. (1996) In *Neural Fuzzy Systems*, Prentice-Hall International, Inc, USA

Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986) Learning representations by back-propagating errors, *Nature*, 323,533-536

Spanjers H, Vanrolleghem PA, Nguyen K, Vanhooren H, Patry GG. (1998) Towards a simulation-benchmark for evaluating respirometry-based control strategies. *Water Sci Technol*; 37(12), 219–226.

Takagi T. and Sugeno M. (1985) Fuzzy identification systems and its application to modelling and control. *IEEE Trans. Systems, Man and Cybernetics* 15, 116-132

Tay J. H, Zhang X. (2000) A fast predicting neural fuzzy model for high-rate anaerobic wastewater treatment systems, Wat. Res. 34(11), 2849-2860

Wasserman, P. D. (1993) In *Advanced Methods in Neural Computing*, New York: Van Nostrand Reinhold