# AEROBIC END-POINT DETECTION IN A SEQUENCING BATCH REACTOR

**Hilario López García** [*,**]
**Iván Machón González** [*,***]

[*] *Universidad de Oviedo. Escuela Politécnica Superior de Ingeniería. Departamento de Ingeniería Eléctrica, Electrónica de Computadores y Sistemas. Edificio Departamental 2. Zona Oeste. Campus de Viesques s/n. 33204 Gijón (Asturias). Spain*
[**] *hilario@isa.uniovi.es*
[***] *machonivan@uniovi.es*

Abstract: The estimation of the aerobic phase end-point is usually used to improve the operating capacity in a sequencing batch reactor. In this paper, a software technique and a configuration of the dissolved oxygen control closed-loop are proposed to achieve the aerobic end-point detection. The proposed software technique consists of Self-Organizing Map and clustering algorithms.

Keywords: Waste treatment, sequencing batch reactor, self-organizing mapping, clustering algorithms.

## 1. PURPOSE OF THE PAPER

This paper is part of the KNOWATER II project "Implementation of a Knowledge Based System for Control of Steelworks Waste Water Treatment Plant", which is sponsored by ECSC and their agreement number is 7210-PR-234. The contractors are Centro Sviluppo Materiali S.p.A., Corus RT&D, Betrieb Forschung Institut (BFI) and Universidad de Oviedo. The main objective of the KNOWATER II project was the development of plant supervision techniques for implementation in wastewater treatment plants.

The present work was focused on the sequencing batch reactor to treat the wastewater effluents of the coke plant of Arcelor in Avilés (Spain). An important task was the development of a software tool that integrates the data acquisition system and the Self-Organizing Map (SOM) algorithm as Artificial Intelligence (AI) technique. Also the supervision of the biological treatment is carried out using K-means as clustering algorithm and Davies-Bouldin index for clustering validation. The estimation of the current process state is calculated by means of the SOM networks that have been validated in function of the topographic error and the quantization error. Moreover, the dissolved oxygen control closed-loop was configured to achieve the aerobic end-point detection using the proposed AI technique. Thus, important on-line knowledge is obtained.

## 2. SEQUENCING BATCH REACTOR

The wastewater is treated biologically in a Sequencing Batch Reactor (SBR). The closed-loop of the oxygen control in the SBR is shown as block diagram in figure 1. The dissolved oxygen concentration is controlled by a PID. Air is pumped into the reactor and a valve is regulated. The set-point is between 6 and 5 $mgO_2/l$. Moreover, a recorder is installed to work as data acquisition interface

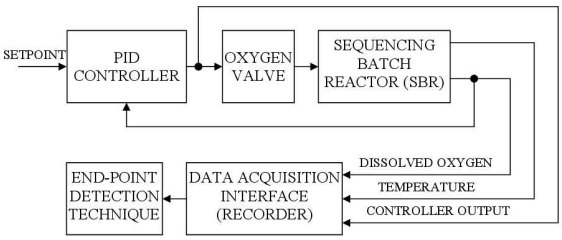


Figure 1. Connection between the oxygen control closed-loop and the end-point detection technique

between the sensors and the developed end-point detection tecnique located on a PC station (figure 1). The interface is able to establish a TCP/IP protocol with the developed software. The dissolved oxygen electrode has a temperature sensor to compensate the measurement deviations, so the temperature can be measured.

Also the measurements can be stored in a floppy disk unit of the recorder. This unit is like a data logger. The data files of the floppy disk make an initial off-line study of the process possible. Taking into account this previous study, the PID controller output of the oxygen closed-loop was connected and registered as one of the process variables to train the SOM network.

## 3. SOM MODEL FORMULATION

Self-Organizing Map (SOM) was used to construct a model that can be used as a pattern (Kohonen *et al.*, 1996). The SOM (Kohonen, 2001) consists of a regular lattice typically defined in a two dimensional space composed of several neurons placed in the nodes of the lattice. SOM training implies assigning a set of coordinates in the input data space (prototype vector) to each neuron. Thus, each neuron is represented by a prototype vector and a correspondence is established between the coordinates of each neuron in the input space (data set) and their coordinates in the 2D-lattice.

The number of neurons determines the accuracy and generalization capability of the SOM and it is determined by equation 1. $M$ is the number of neurons and $N$ is the number of samples of the training data.

$$M = 5\sqrt{N} \tag{1}$$

The next step consists of determining the ratio between the number of rows $n_1$ and the number of columns $n_2$ of the 2D grid or output space. According to equation 2, the ratio between sidelengths of the map is the square root of the ratio between the two biggest eigenvalues of the training data. The highest eigenvalue is $e_1$ and the second highest is $e_2$.

$$\frac{n_1}{n_2} = \sqrt{\frac{e_1}{e_2}} \tag{2}$$

The present study was carried out using the SOM toolbox version 2.0 (Vesanto *et al.*, 1999) developed at the HUT (Helsinki University of Technology). The steps taken to analyze the data are outlined in (López and Machón, 2004*c*). Firstly, the most significant process variables are selected. An off-line study of the process data were carried out to achieve this selection (López and Machón, 2004*a*). These variables are described in table 1. Secondly, the data were normalized to a zero mean value and a unitary variance to make SOM treat them in the same way. After normalizing, the SOM network was trained with these variables using batch training algorithm. Once the SOM has converged, it stores the most relevant information about the process in its prototype vectors. The visualization process allows all this information to be displayed in several ways: Interneuron distance matrix (U-matrix) that shows in gray or color levels the mean distance of each unit to its closest neighbors; the component planes that display the value of a given input variable throughout the whole data set using gray or color levels in the 2D lattice; the best clustering structure that allows the main working zones of the process to be visualized.

Table 1. Training variables.

| Name | Description |
|---|---|
| OXYGEN | Dissolved oxygen concentration ($mgO_2$/litre) |
| CONTROLLER OUTPUT | Output of the PID controller of the oxygen closed-loop (0-100) |
| TEMPERATURE_SBR | Temperature in the SBR (C) |

## 4. VALIDATION METHOD

According to the properties of the SOM, the trained neural network must achieve the topology preservation of the data. Therefore the neighborhood on the model and in the input space must be similar. If two prototype vectors close to each other in the input space are mapped wide apart on the grid, this is signaled by the situation where two closest best matching neurons of an input vector are not adjacent neurons. This kind of fold is considered as an indication of the topographic error in the mapping.

The topographic error (Kiviluoto, 1996) can be calculated by equation 3 as the proportion of sample vectors for which two best matching neurons are not adjacent. $N$ is the number of samples, $x_k$ is the $k$th sample of the data set and $u(x_k)$ is equal

to 1 if the first and second best matching neurons of $x_k$ are not adjacent neurons, otherwise zero.

$$e_t = \frac{1}{N}\sum_{k=1}^{N} u(x_k) \tag{3}$$

Moreover, the prototype vectors try to approximate to the data set. A consequence of this approach is the resolution error or the quantization error. Equation 4 is usually used to measure the resolution of the mapping calculating the average quantization error over the whole testing data set. $N$ is the number of samples, $x_i$ is the $i$th data sample and $m_b$ is the prototype vector of the best matching neuron for $x_i$

$$e_q = \frac{1}{N}\sum_{i=1}^{N} \|x_i - m_b\| \tag{4}$$

Several data sets which correspond to the aerobic phase of the SBR are available to carry out the validation of the model. The objective is to find out the model that minimizes the quantization and topographic errors from several neural networks which have been trained using each of these available patterns and, at the same time, for different map sizes. The method of validation can be summarized in the following steps:

(1) A data set or pattern $p_i$ is chosen to train the network. The data are normalized to a distribution with zero mean value and unitary variance.
(2) Batch training is carried out on the SOM map whose sidelengths are calculated using equations 1 and 2. using pattern $p_i$ as training data.
(3) Once the trained model is obtained, the topographic and quantization errors are calculated for the remaining patterns $p_j$ which have not been used during the training. These patterns must also be previously normalized.
(4) The size of this trained map is increased and reduced respecting the proportionality of its sidelengths (width and length). Once the size has been modified, the neural network is again trained using pattern $p_i$.
(5) The third and fourth steps are repeated for different map sizes.
(6) Steps 1 through 5 are repeated for the remaining patterns $p_j$, assuming each of these the role of pattern $p_i$.

Several map sizes have been trained using the patterns. The mean values of the errors over the test patterns (the remaining patterns which have not been used during the training) in function of the map size are shown in figure 2. It can be seen that the larger the map size the lower the quantization error but the higher the topographic error. This is due to the neural network folds to reduce the quantization error. Moreover, the larger the map size the higher the computational cost. Therefore, there is compromise between the increase of the topographic error and the reduction of the quantization error. A curve, which represents the sum of both errors, has been added to the graphic. The selected model, which is integrated into the software tool, was the neural network that corresponds to the value equal to 6 of the horizontal axis. This horizontal value corresponds to a map where $n_1$ is equal to 14 and $n_2$ is equal to 8. The number of samples $N$ of the training data set was equal to 500.

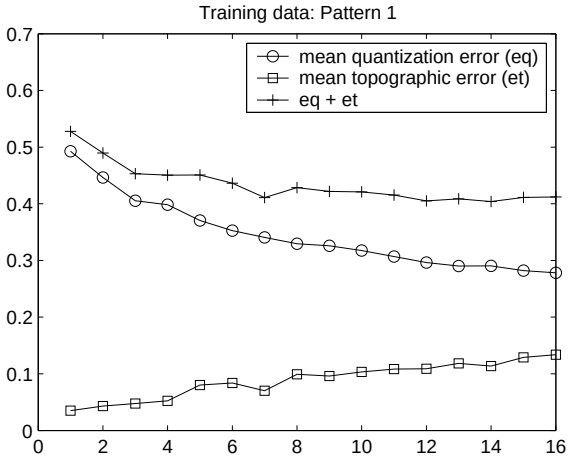


Figure 2. Mean value of errors

## 5. BEST CLUSTERING TECHNIQUE

In this paper, the clustering process consists of a two-stage procedure (Vesanto and Alhoniemi, 2000). Firstly, the prototype vectors are obtained training the data of the aerobic phase using a SOM algorithm and then clustering them using a K-means algorithm, see (McQueen, 1967) and (Dubes and Jain, 1976). Ten clustering structures were obtained varying the predefined number of clusters.

Finally, the best clustering structure among the ten structures, which have been obtained from the K-means algorithm, is selected using the Davies-Bouldin index (Davies and Bouldin, 1979). This index searches the best model that minimizes the within-cluster distance and maximizes the between-clusters distance. The Davies-Bouldin index is suitable for evaluation of k-means partitioning, because it gives low values indicating good clustering results for spherical clusters. Figure 3 shows the Davies-Bouldin index after being applied to the data from the aerobic stage of the treatment. The best clustering corresponds to a number of two clusters and can be seen in figure

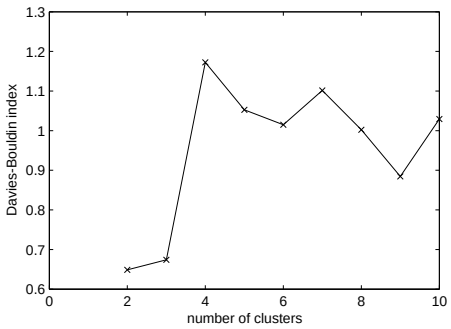


Figure 3. Results of Davies-Bouldin index


Figure 4. Best clustering structure

3. Once the best clustering has been selected, the clustering structure must be visualized in order to label the input data space onto the 2D-lattice. This allows the visualization of the different features of the process.

The best clustering has been projected onto the SOM and a new component plane has been obtained that is displayed in figure 4. Cluster 1 corresponds to the first hours of the aerobic treatment where the values of the controller output are high due to the high chemical oxygen demand (COD). During this period the biological activity is high and the toxic substances are eliminated by means of the cellular metabolism, whereas cluster 2 represents the data collected after this high biological activity where the values of the controller output are lower because the COD has decreased. If this is the state of the treatment plant, the biological treatment of the aerobic stages can be finished improving the capacity of the plant.The results were highly repeatable except in winter time due to the temperature influence.

## 6. SOFTWARE TOOL

The developed software is running in the PC station that is connected to the data acquisition interface (recorder) by means of Ethernet connection and TCP/IP protocol (MODBUS protocol in particular). The proposed AI techniques are integrated into this application to achieve the process monitoring and the process state estimation. Data from the aerobic stage is collected on-line automatically to train a SOM network. The plant operator can visualize the latest SOM network that corresponds to the latest aerobic treatment cycle of the plant, viewing the correlation between the process variables and the data classification can be obtained. The estimation of the current process state is also calculated by the stored patterns (SOM networks that have already been trained and validated).

The program improves the plant operation with the estimation of the oxygen uptake rate by means of the controller output dynamic from the dissolved oxygen control (Cavalcanti *et al.*, 1999). The oxygen uptake rate is one of the most important variables in the wastewater biological treatments and is the associated demand which is necessary to oxidize the organic substrate by the heterotrophic biomass. The estimation of the time of the main activity of the treatment (aerobic phase) achieves operating cost savings and increases the plant performance, see (Paul *et al.*, 1998), (Cho *et al.*, 2001) and (Andreottola *et al.*, 2001).

In order to achieve the process monitoring of the aerobic treatment, the process variables of the table 1 were chosen for SOM training. A key aspect is: how can the training data set be determined? The training data set must only contain the samples of the aerobic stage. The mean value of the controller output was used to obtain these samples because this signal can be considered as a key variable to estimate the states of the treatment, see (López and Machón, 2004*a*) and (López and Machón, 2004*b*). The user can change the values for starting and finalizing the training. Also the sample rate can be changed.

The results of the latest SOM network can be visualized in figure 5 and they correspond to the latest aerobic stage of the wastewater treatment. They are the U-matrix, the component planes and the best clustering structure. The U-matrix shows the distances between neighboring prototype vectors of neighboring neurons. Thus, a group of neurons, which is too concentrated, implies a cluster. Distanced zones of neurons must separate several clusters. Each component plane shows the value of each neuron to estimate the data variable of the input space. It is useful to determine the several zones where the variable value is high or low and to observe any correlation or relationship between process variables. The correlations among variables can be observed, for example, between the controller output and the oxygen concentration. Also there is a correlation between the controller output and the temperature in the reactor due to the higher the temperature the lower the dissolved oxygen concentration and the controller must compensate this effect. As mentioned above, the best clustering structure is composed of 2 clusters and is calculated by means of the Davies-Bouldin index. A cluster corresponds to HIGH COD and the other is the LOW COD.

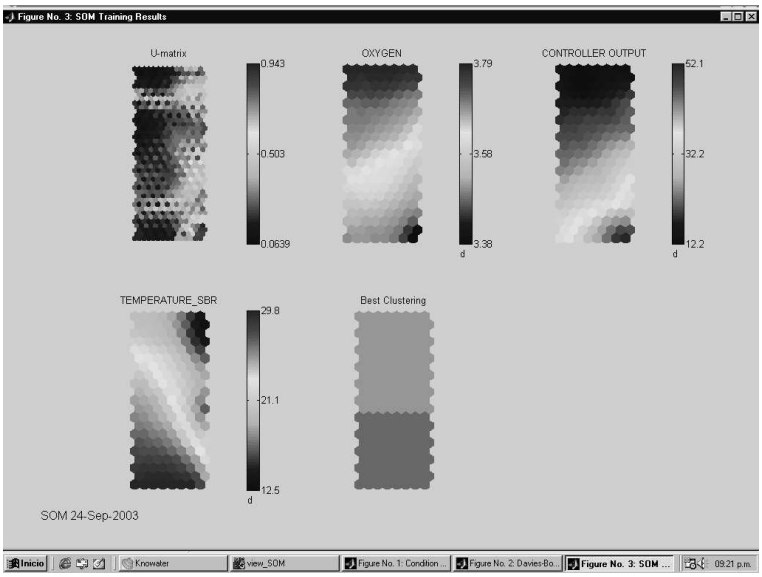


Figure 5. SOM training results

The sample rate of the data acquisition system has been increased in order to obtain a better definition of the sensor signals. The dissolved oxygen concentration and its controller output are signals with higher component frequencies than the temperature, i.e., the values of some variables are changing faster than others. Thus different sample times have been assigned to each variable.

The cycles of the biological treatment at the sequential batch reactor can be clearly observed in figures 6 and 7. The higher values correspond to the anoxic stage when the controller output is saturated and equal to 100%. The rest of the data corresponds to the aerobic stage (including sedimentation).

An important aspect appears: the end-point of the aerobic reaction. This end-point detection can be used to finalize the aerobic stage and in this way the duration of the cycle is shorter increasing the operating capacity of the plant. Moreover, the oxygen consumption is high, influenced by the temperature in the reactor as could have been expected. The duration of the cycle was initially 48-72 hours one year ago, see figure 6, and it has been reduced to 24 hours as is showed in figure 7. In this way the operating capacity of the plant has been increased by reducing the retention time.

In figure 8 the process state is estimated projecting the current values onto a SOM network by means of standing out the best matching neuron from the rest of the neurons. This SOM network is used as a pattern and is previously stored and validated using the validation method explained above. The projection is carried out onto the component planes and the best clustering structure. The best clustering structure is composed of two clusters. The first one corresponds to the first hours of the treatment. During this phase high

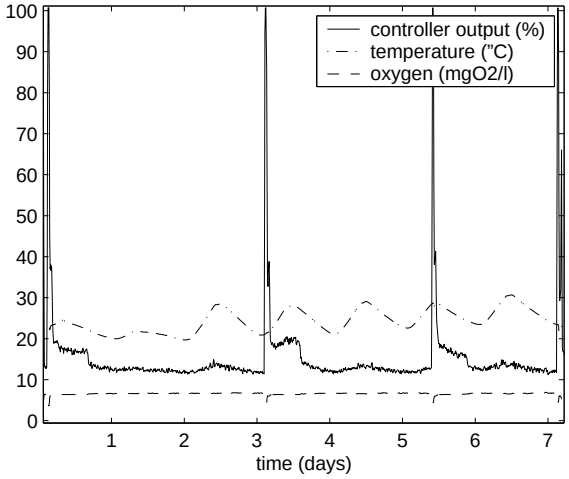


Figure 6. Process values of one year ago

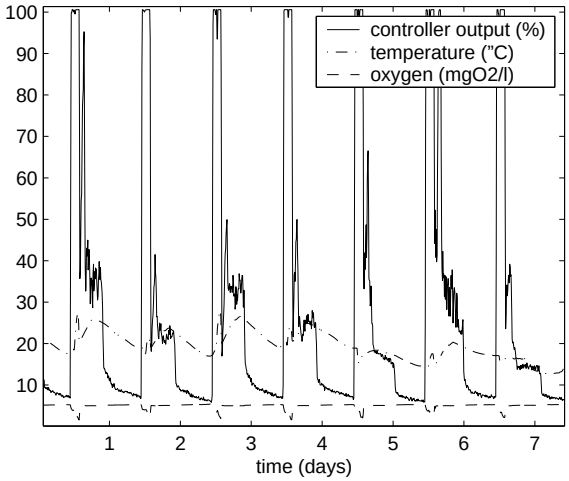


Figure 7. Current process values

oxygen uptake rates take place and the aerobic activity of the biological process is high. The second cluster represents the samples which have been obtained after the biological activity has

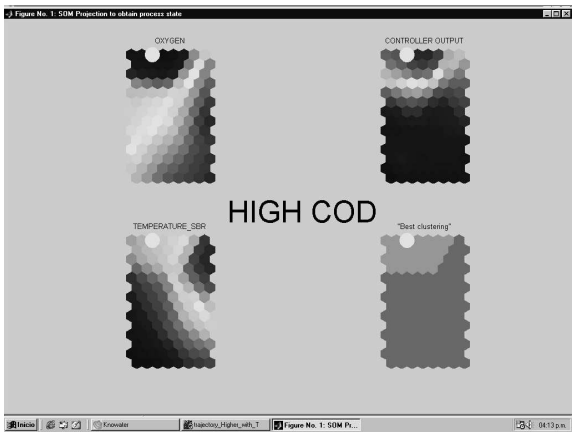


Figure 8. Process state estimation

decreased. The implemented application names them as HIGH COD and LOW COD, respectively. Thus, important on-line knowledge is obtained and the end of the main biological activity can be identified.

## 7. CONCLUSIONS

An oxygen control closed-loop was configured to achieve the end-point detection of the aerobic phase in a sequencing batch reactor (SBR) using an AI software. This software tool was developed to supervise the treatment and is running, connected to the SBR. It is a stand-alone application which is composed of the data acquisition system from the SBR and the SOM as proposed AI technique. It is also a remote supervision application due to the use of a TCP/IP connection. The data set of the aerobic stage is collected to train automatically a SOM network. In this way, the correlation among the process variables can be observed by visualizing the latest SOM network that corresponds to the latest aerobic treatment cycle of the plant. The data classification is obtained using K-means algorithm as partitive clustering algorithm and Davies-Bouldin index for clustering validation. The estimation of the current process state can be assigned calculating the best matching neuron that corresponds to the current process values. The end-point of the aerobic reaction can be detected by this AI technique. So, operating cost savings are achieved and the plant performance is increased. In this way, total retention time was reduced from 48-72 hours to 24 hours.The results were highly repeatable except in winter time due to the temperature influence.

## REFERENCES

Andreottola, G., P. Foladori and M. Ragazzi (2001). On-line control of a sbr system for nitrogen removal from industrial wastewater. *Water Science and Technology* **43**(3), 93–100.

Cavalcanti, S.Y., G. Singh, A. Cornelius and R.C. Silvrio (1999). Feedback control method for estimating the oxygen uptake rate in activated sludge systems. *IEEE Transactions on Instrumentation and Measurement* **48**(4), 864–869.

Cho, B.C., S.L. Liaw, C.N. Chang, R.F. Yu, S.J. Yang and B.R. Chiou (2001). Development of a real-time control strategy with artificial neural network for automatic control of a continuous-flow sequencing batch reactor. *Water Science and Technology* **44**(1), 95–104.

Davies, D.L. and D.W. Bouldin (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1**(2), 224–227.

Dubes, R. and A.K. Jain (1976). Clustering techniques: The user's dilemma. *Pattern Recognition* **8**, 247–260.

Kiviluoto, K. (1996). Topology preservation in self-organizing maps. *IEEE International Conference on Neural Networks* **1**, 294–299.

Kohonen, T. (2001). *Self-Organizing Maps*. Springer-Verlag. New York.

Kohonen, T., E. Oja, O. Simula, A. Visa and J. Kangas (1996). Engineering applications of the self organizing mapping. *IEEE Proceedings* pp. 1358–1384.

López, H. and I. Machón (2004*a*). Biological wastewater treatment analysis using som and clustering algorithms. *12th Mediterranean Conference on Control and Automation*.

López, H. and I. Machón (2004*b*). An introduction to biological wastewater treatment explained by som and clustering algorithms. *IEEE International Symposium on Industrial Electronics*.

López, H. and I. Machón (2004*c*). Self-organizing map and clustering for wastewater treatment monitoring. *Engineering Applications of Artificial Intelligence* **17**(3), 215–225.

McQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *5-th Berkeley Symposium on mathematics, Statistics and Probability* (1), 281–298.

Paul, E., S. Plisson-Saune, M. Mauret and J. Cantet (1998). Process state evaluation of alternating oxic-anoxic activated sludge using orp, ph and do. *Water Science and Technology* **38**(3), 299–306.

Vesanto, J. and E. Alhoniemi (2000). Clustering of the self-organizing map. *IEEE Transactions Transactions on Neural Networks* **11**(3), 586–600.

Vesanto, J., E. Alhoniemi, J. Himberg, K. Kiviluoto and J. Parviainen (1999). Self-organizing map for data mining in matlab: the som toolbox. *Simulation News Europe* pp. 25–54.