Copyright © 2002 IFAC
15th Triennial World Congress, Barcelona, Spain

FUZZY IDENTIFICATION OF BIOPROCESSES APPLYING TSK-TYPE MODELS WITH CONSEQUENT PARAMETER ESTIMATION THROUGH ORTHOGONAL ESTIMATOR

Rosimeire Aparecida Jerônimo [*,1], **Claudio Garcia** [*,1], **Oscar A. Z. Sotomayor** [**]

[*] *Department of Telecommunications and Control Engineering, Laboratory of Automation and Control (LAC)*
[**] *Department of Chemical Engineering, Laboratory of Simulation and Process Control (LSCP)*

Polytechnic School of the University of São Paulo
Av. Prof. Luciano Gualberto, Trav. 3, n. 158, Cidade Universitária, CEP 05508-900, São Paulo - SP, Brazil

Abstract: This paper presents fuzzy identification of two bioprocesses employing TSK-type models. A "*Modified Gram-Schmidt*" (*MGS*) orthogonal estimator is used to estimate the consequent parameters. This approach is then applied to identify two distinct cases involving dissolved oxygen concentration: one related to a bioreactor and the other one related to an activated sludge process. The obtained models are then cross-validated.

Keywords: System identification, Bioprocesses, Nonlinear models, Fuzzy modeling, TSK-type models, Modified Gram-Schmidt algorithm, Parameter estimation.

1. INTRODUCTION

Fuzzy systems have been employed in many applications. A central characteristic of fuzzy systems is that they are based on the concept of fuzzy coding of information and operating with fuzzy sets instead of numbers. In essence, the representation of information in fuzzy systems imitates the mechanism of approximate reasoning performed in the human mind.

A common practice to model a complex non-linear system is by decomposing it in smaller parts, around given operating points. Thus, for non-linear, partially known high-order systems, the fuzzy modeling approach may be very adequate. As it divides the input and output spaces in subspaces, it means that local models are created which cope with the complexity of nonlinear systems.

A fuzzy modeling technique is here applied to estimate the dissolved oxygen concentration in two distinct cases: a bioreactor and an activated sludge process. In both cases, the TSK-type model developed by Takagi, Sugeno and Kang is used (Takagi and Sugeno, 1985), (Sugeno and Kang, 1986). They are formed by logical rules that have a fuzzy antecedent part and a functional consequent one. NARX ("**N**on-**L**inear **A**uto-**R**egressive with e**X**ogenous inputs") structures are used in the consequent part. The consequent parameters are estimated using an advance regression orthogonal estimator based on the Modified Gram-Schmidt

[1] * Corresponding authors. Tel: +55-11-3818-5648; Fax: +55-11-3818-5718
E-mail address: rosi@lac.usp.br, clgarcia@lac.usp.br (R.A. Jerônimo), (C. Garcia)

estimation method proposed in (Korenberg *et al.*, 1988) and (Chen *et al.*, 1989).

This paper is organized as follows: section 2 presents the description of TSK-type fuzzy models for dynamic systems. Section 3 presents the 2 bioprocesses to be identified. Section 4 presents the fuzzy identification method applied to obtain the models of the bioprocesses and their behavior. Section 5 shows the conclusions of this work.

2. FUZZY MODELS FOR DYNAMIC NON-LINEAR SYSTEMS AND TSK MODELS

To build fuzzy models from data generated by poorly understood dynamic systems, the input-output representation is often applied. A very common structure is the NARX model ("**N**on-**L**inear **A**uto-**R**egressive with e**X**ogenous inputs"), which can represent the observable and controllable models of a large class of discrete-time non-linear systems. It establishes a relationship between the collection of past input-output data and the predicted output.

$$\hat{y}(k+1) = F(y(k), \dots, y(k-n_y+1), u(k), \dots, u(k-n_u+1)) \quad (1)$$

where k denotes discrete time samples, n_y and n_u are integers related to the system order.

2.1 *TSK Models*

Consider the nonlinear multi-input single-output system

$$y = f(x_1, x_2, \dots, x_n) \quad (2)$$

which has known m operating points $(x_{i,1},\ x_{i,2}, \dots, x_{i,n}),\ i = 1, \dots, m$, around which the system can be decomposed into m linearized subsystems: $y_i = f_i(x_1, x_2, \dots, x_n)$. The TSK model for the system can be generically represented as

$$R_i: \text{ if } x_1 \text{ is } A_{i,1} \text{ and} \dots \text{and } x_n \text{ is } A_{i,n} \text{ then } y_i = f_i(x_1, x_2, \dots, x_n) \quad (3)$$

or, more specifically:

R_i: **if** $y(k)$ **is** $A_{i,1}$ **and** $y(k-1)$ **is** $A_{i,2}$ **and**... **and** $y(k-n_y+1)$ **is** A_{i,n_y} **and** $u(k)$ **is** $B_{i,1}$ **and** $u(k-1)$ **is** $B_{i,2}$ **and**...**and** $u(k-n_u+1)$ **is** B_{i,n_u} **then**

$$\hat{y}_i(k+1) = \sum_{j=1}^{n_y} a_{i,j} y(k-j+1) + \sum_{j=1}^{n_u} b_{i,j} u(k-j+1) + c_i \quad (4)$$

where $a_{i,j}$, $b_{i,j}$ and c_i are the consequent parameters. The NARX local model can represent MISO systems directly and MIMO systems in a decomposed form as a set of coupled MISO models. It should be noted that the dimension of the regression problem in input-output modeling is often larger than the state-space models, since the state of the system can usually be represented by a vector of a lower dimension than, for instance, in the NARX model given by equation (1).

3. CASES OF STUDY

Two distinct cases of identification are considered here. The first one involves the identification of dissolved oxygen dynamics, considering just one bioreactor, which is based on a model proposed in (Nakajima *et al.*, 1996). The second case is based on identification of the dissolved oxygen dynamics of a benchmark simulator as presented in (Sotomayor *et al.*, 2001*b*). A detailed description of this process can be found in (Sotomayor *et al.*, 2001*a*). In both cases, the consequent part of parameter identification was made "*off-line*", combining the "*Modified Gram-Schmidt*" method (its description can be found in (Chen *et al.*, 1989)) with the method presented in (Wang and Langari, 1995)). The model significant terms could be selected based on the "*error reduction ratio*" (ERR).

It is next presented, in a condensed form, the description of the parameter estimation procedure of the TSK model consequent, based on (Wang and Langari, 1995).

$$y_k = \phi_k^T \theta \quad (5)$$

where

$$\phi_k^T \triangleq [f_{1,k} f_{2,k} \cdots f_{(r+1)*q,k}] \quad (6)$$

$$\theta \triangleq [\theta_1 \theta_2 \cdots \theta_{(r+1)*q,k}]^T \quad (7)$$

where r and q are, respectively, the number of inputs and the number of rules.

Equation (5) can then be represented as:

$$y_k = \sum_{i=1}^{(r+1)*q} f_{i,k} \theta_i \quad (8)$$

Transforming equation (8) into an equivalent orthogonal equation, results:

$$y_k = \sum_{i=1}^{(r+1)*q} w_{ik} g_i \quad (9)$$

where

$$w_{1k} = f_{1k} \quad (10)$$

$$w_{mk} = f_{mk} - \sum_{i=1}^{m-1} \alpha_{im} w_{ik},$$
$$m = 2, 3, \cdots, (r+1) \star q \qquad (11)$$

and

$$\alpha_{ij} = \frac{\sum_{k=1}^{N} w_{ik} f_{ik}}{\sum_{k=1}^{N} w_{ik}^2},$$
$$i < j \quad j = 2, 3, \cdots, (r+1) \star q \qquad (12)$$

where N corresponds to the number of sampled data.

The estimated coefficients g_i are calculated by:

$$\hat{g_i} = \frac{\sum_{k=1}^{N} w_{ik} y_k}{\sum_{k=1}^{N} w_{ik}^2},$$
$$i = 1, 2, \cdots, (r+1) \star q \qquad (13)$$

3.1 *O_2 Dynamics - Bioreactor (Case 1)*

The DO concentration dynamics $y(t)$ is obtained from a mass balance (Bastin and Dochain, 1990).

$$\frac{dy(t)}{dt} = \frac{Q(t)}{V}(y_{in}(t) - y(t)) + OTR(t) - OUR(t) \qquad (14)$$

The first term of the right hand describes the transport of the DO concentration, where $Q(t)$ is the waste water flow rate, V is the volume of the waste water, $y_{in}(t)$ is the DO concentration in the influent flow and $y(t)$ is the DO concentration in the effluent flow. The second term, OTR, is called the oxygen transfer rate. A common way to formulate OTR is

$$K_L a(u(t)) \star (y_{sat}(t) - y(t)) \qquad (15)$$

where $K_L a$ is the oxygen transfer function, $u(t)$ is the air flow rate and $y_{sat}(t)$ is the oxygen saturation concentration. OTR describes the rate at which oxygen is transferred into the waste water when air bubbles pass upwards. This function is expected to be non-linear and dependent on several factors: the geometry of air diffusers, waste water composition, temperature, etc. The third term in 14), OUR, is the respiration rate, denoted by $R(t)$. It can suffer some abrupt disturbances caused by several factors such as DO concentration, biomass concentration, substrate concentration, pH and temperature. From equations (14) and (15), the continuous-time model of the DO dynamics is derived:

$$\frac{dy(t)}{dt} = D(t)(y_{in}(t) - y(t)) + K_L a(u(t)) * (y_{sat} - y(t)) - R(t) \qquad (16)$$

where $D(t)$ denotes $\frac{Q(t)}{V}$, the dilution rate.

The conditions for the simulation of DO dynamics are presented in table 1, where T_s means sampling time.

Table 1. Parameters of Bioreactor Process - Case 1

Bioreactor - Case 1
$T_s = 2.8 * 10^{-3}$ $[h]$, $\frac{Q}{V} = 1.7$ $[1/h]$
$y_{sat} = 10$ $[mg/l]$, $y_{in} = 0$ $[mg/l]$
$R(t) = 12 + 8\sin(1.5t)$ $[mg/l/h]$
$K_L a(u) = 5\arctan(4\pi u/1000)$ $[1/h]$

3.2 *O_2 Dynamics - Activated Sludge Process (Case 2)*

The Activated Sludge Process (ASP) is the most widespread process used for biological waste water treatment. Bioactivities in the ASP are intimately related to the dissolved oxygen concentration (DOC). Two factors that affect the dynamics of the dissolved oxygen are the respiration rate or the oxygen uptake rate (OUR) and the oxygen transfer function ($K_L a$).

A schematic diagram of an activated sludge process in a configuration with pre-denitrification for organic matter and nitrogen removal of domestic effluent is shown in figure 1.

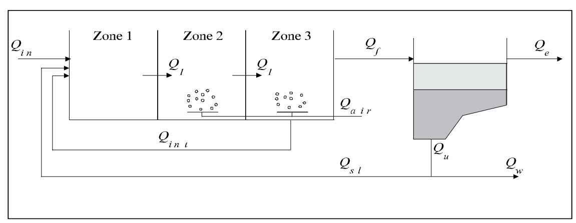

Figure 1. Layout of the Activated Sludge Process

The process configuration is formed by a bioreactor composed of an anoxic zone, two aerobic zones and a settler. The compartments of the bioreactor are considered to have constant volume (13 m^3, 18 m^3 and 20 m^3, respectively) and to be ideally mixed whereas the secondary settler (20 m^3) is modeled employing a series of 10 layers (one-dimensional model, where it is assumed that no biological reaction occur). The influent flow Q_{in} is 4.17 m^3/h with a proportion of biodegradable COD of 224 mg/l and a hydraulic retention time of 17.0 h. The internal recycle flow is $Q_{int} = 2Q_{in}$, the external sludge recycle flow is $Q_{sl} = 0.5Q_{in}$, the wastage flow rate is $Q_w = 0.0258$ m^3/h, the external carbon source flow rate is $Q_{ext} = 0$ m^3/h, the air flow rates are $Q_{air\_2} = 0.044$ m^3/h and $Q_{air\_3} = 0.033$ m^3/h, for the second and third zones, respectively. In the anoxic zone, no airflow rate is considered. For further details about this benchmark simulator, see (Sotomayor *et al.*, 2001*a*).

4. FUZZY IDENTIFICATION AND SIMULATION RESULTS

Data for cases 1 and 2 were collected in open-loop.

In case 1, the process has been simulated for 6 hours, being collected 925 samples (air flow rate as input and Dissolved Oxygen Concentration as output). The first 463 collected input and output samples were considered in the identification procedure, which basically corresponded to a period of 3 hours. The remaining 462 input and output samples were considered for the process validation. Figure 2 shows both: the identification and validation input and output data.

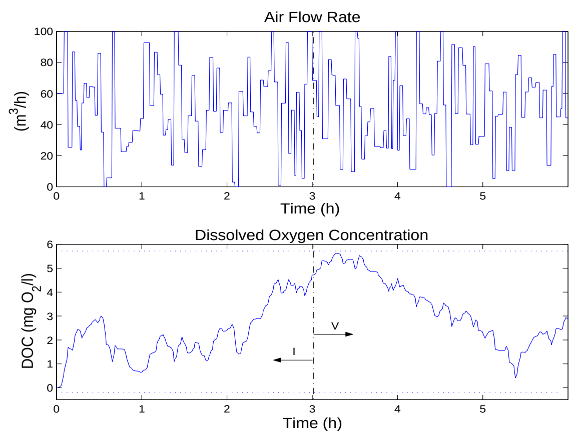

Figure 2. Collected input and output signals - Case 1

In case 2, the input and output data used in the identification of DOC dynamics, namely the air flow rate and the corresponding dissolved oxygen concentration (the other variables were considered to be in the nominal state of operation), are displayed in figure 3.

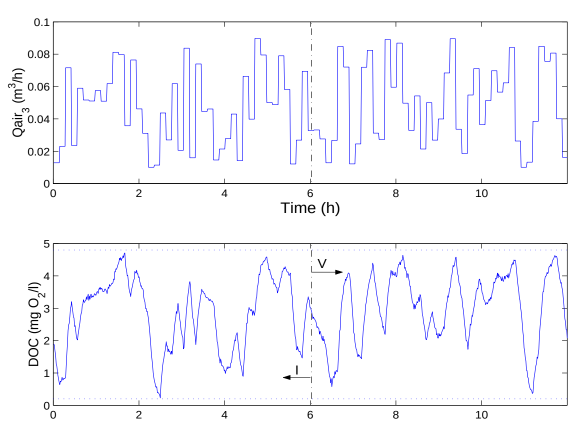

Figure 3. Collected input and output signals - Case 2

The data were collected during a period of 12 hours, with a sampling time of $T_s = 49.8\ s$, representing a set of 869 input and output points. The first 435 points were used in the identification process, corresponding to 6 hours of operation, whereas the 434 remaining points were employed to validate the model.

4.1 *Universe of Discourse*

Being a fuzzy system, one of the first steps is related to defining the Universe of Discourse, based on prior knowledge of the process. In both cases, the definition of the universe of discourse for the input and output variables was based on the minimum and maximum values of input and output data. For case 1, the air flow rate varies between $(0 - 100)\ m^3/h$, whereas the oxygen concentration varies between $(0 - 5.4958)\ mg\ O_2/l$. For case 2, the air flow rate varies between $(0.0101 - 0.0898)\ m^3/h$, whereas the oxygen concentration varies between $(0.2285 - 4.7101)\ mg\ O_2/l$.

4.2 *Fuzzy Linguistic Attributes and Membership Functions*

Both input variables for cases 1 and 2 were restricted to positive values. Therefore, in both cases, the fuzzy linguistic attributes for the input variable were defined as low, below medium, above medium and high. For the output variable, the fuzzy linguistic attributes for the dissolved oxygen concentration are low, medium, high and very high. For both cases herein presented, the linguistic attributes of membership functions were based on the human expert experience, considering the discourse intervals of each studied process, as well as the premises of the fuzzy inference.

Figure 4 illustrates the membership functions of the input and output variables for case 1.

Figure 5 illustrates the membership functions of input and output variables for case 2.

In both cases, the process is represented as a first order discrete-time NARX model

$$y(k+1) = f(y(k), u(k)) \qquad (17)$$

where k denotes the sampling instant, f is an unknown relationship approximated by the TSK fuzzy model, $u(k)$ represents the air flow for both cases and $y(t)$ represents the dissolved oxygen concentration for both cases. Based on prior knowledge, this structure is considered adequate for approximation of the activated sludge process dynamics (Jerônimo *et al.*, 2000).

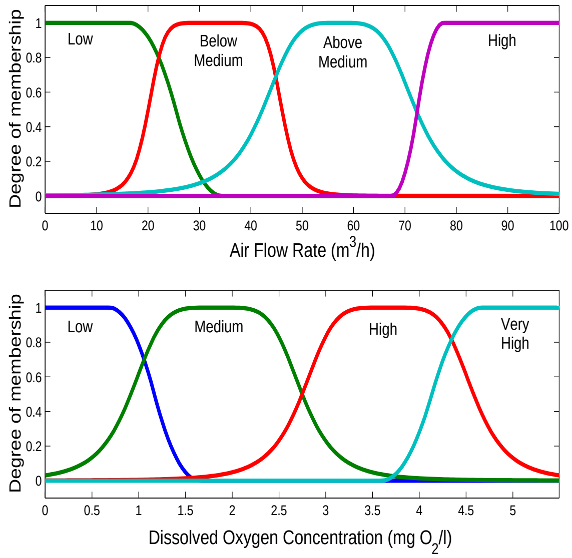

Figure 4. Membership functions of input (air flow rate) and output (DOC) variables - Case 1

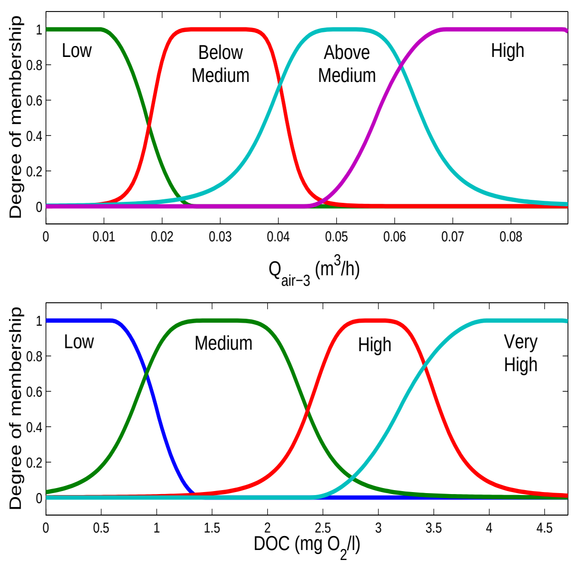

Figure 5. Membership functions of input (Q_{air-3} and output (DOC) variables - Case 2

For both cases, the fuzzy rule base was derived from the human expert knowledge. In a first approach, the idea would be to have 16 rules, since the number of linguistic attributes of each model input is 4. But it is known that the dissolved oxygen concentration is directly proportional to the air flow rate and there is no significant delay between them. So, based on that, it is possible to infer that for low air flow rates it will result a low DOC and so on, resulting in just 4 rules for each case.

For case 1, the rule base is:

R_1 : if $y(k)$ is low and $u(k)$ is low then $y(k+1) = 0.011603 - 0.321802y(k) - 0.109168u(k)$

R_2 : If $y(k)$ is medium and $u(k)$ is below medium then $y(k+1) = 0.591093 - 0.060413y(k) + 0.923741u(k)$

R_3 : If $y(k)$ is high and $u(k)$ is above medium then $y(k+1) = 0.103973 + 0.053217y(k) + 0.103745u(k)$

R_4 : if $y(k)$ is very high and $u(k)$ is high then $y(k+1) = 0.259074 - 0.692062y(k) + 0.187504u(k)$

For case 2, the rule base is:

R_1 : if $y(k)$ is low and $u(k)$ is low then $y(k+1) = 0.010673 - 0.014021y(k) + 0.094528u(k)$

R_2 : if $y(k)$ is medium and $u(k)$ is below medium then $y(k+1) = 0.006507 + 1.674532y(k) + 0.026547u(k)$

R_3 : if $y(k)$ is high and $u(k)$ is above medium then $y(k+1) = 0.284062 - 0.056435y(k) + 0.073404u(k)$

R_4 : if $y(k)$ is very high and $u(k)$ is high then $y(k+1) = 1.678352 + 0.724533y(k) + 0.224381u(k)$

The validation of the fuzzy models was performed using new data sets different from the ones employed in the identification. Figures 6 and 7 show the cross-validation graphics for both cases.

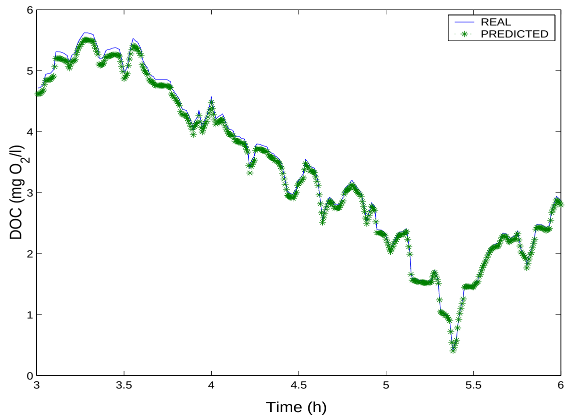

Figure 6. Model validation - comparison of the process output and model predictions - Case 1

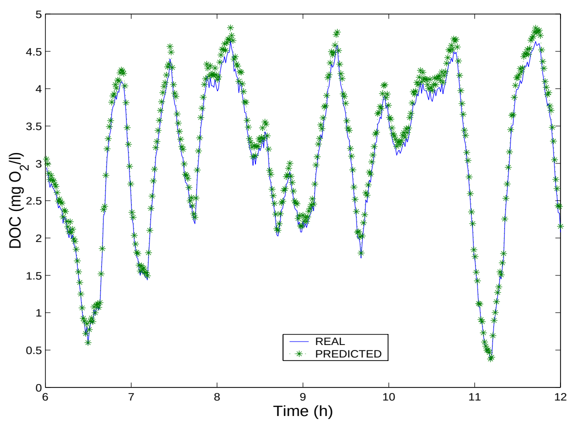

Figure 7. Model validation - comparison of the process output and model predictions - Case 2

5. CONCLUSIONS

This modeling approach has shown satisfactory results for both studied cases, being able to reproduce the dynamics of highly non-linear processes with a very simple model constituted by just a few fuzzy rules.

Some proposals for further works include the definition of the number of membership functions and the minimization of the number of fuzzy rules, independent of human expert knowledge but based on the collected data set.

Acknowledgement: This work was sponsored by CAPES and FAPESP(process no. 98/12375-7).

6. REFERENCES

Bastin, G. and D. Dochain (1990). *On-line estimation and adaptive control of bioreactors.* Prentice Hall.

Chen, S., S. A. Billings and W. Luo (1989). Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control* **50**(3), 8873–1896.

Jerônimo, R. A., O. A. Z. Sotomayor, C. Garcia and S. W. Park (2000). Polynomial nonlinear modelling of biological nutrients removal (BNR) activated sludge processes. XIII Brazilian Congress of Automatic - CBA-2000, Florianópolis, SC, Brazil.

Korenberg, M., S. A. Billings, Y. P. Liu and P. J. McILROY (1988). Orthogonal parameter estimation algorithm for non-linear Stochastic Systems. *International Journal of Control* **48**(1), 193–210.

Nakajima, S., C. F. Lindberg and B. Carlsson (1996). On-line estimation of the respiration rate and the oxygen tranfer function using an extended Kalman Filter. *Technical Report IR-S3-REG-9613*, Royal Institute of Techonology, Sweden.

Sotomayor, O. A. Z., S. W. Park and C. Garcia (2001*a*). A simulation benchmark to evaluate the performance of advanced control techniques in biological wastewater treatment plants. *Brazilian Journal of Chemical Engineering* **18**(1), 81–101.

Sotomayor, O. A. Z., S. W. Park and C. Garcia (2001*b*). Software sensor for on-line estimation of the microbial activity in activated sludge systems. *ISA Transactions.* (in press).

Sugeno, M. and G. T. Kang (1986). Fuzzy modelling and control of multilayer incinerator. *Fuzzy Sets and Systems* **18**, 329–346.

Takagi, T. and M. Sugeno (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics* **15**(1), 116–132.

Wang, L. and R. Langari (1995). Building Sugeno-type models using fuzzy discretization and orthogonal parameter estimation techiques. *IEEE Transactions on Fuzzy Systems* **3**(4), 454–458.